Tagging of Icelandic Text – ongoing work Background

- November 2000, steering committee in Language Technology appointed
- The Icelandic NLP group (*Auður Þ. Rögnvaldsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir, Sigrún Helgadóttir*) experimented with tagging Icelandic text (µ-TBL Dec 2002, TnT Apr 2002)
- Icelandic NLP group together with The Institute of Lexicography in Reykjavík, applied for and was given a grant in April 2002 to develop a tagger that would tag Icelandic text with 92-97% accuracy
- The project started in October 2002.
- The present paper reports the status of the project in May 2003.

Tagging of Icelandic Text – ongoing work Icelandic Language

- Icelandic is a North Germanic or Nordic language, most closely related to Faroese and the dialects of Western Norway.
- Icelandic is a synthetic language with **rich inflections** in all major word classes.
- *Nouns* have **16** grammatical forms
- *Pronouns* have same basic declensional paradigm as nouns
- *Adjectives* have up to **120** grammatical forms, though the number of actual word forms is greatly reduced by homophony and gaps.

Tagging of Icelandic Text – ongoing work Icelandic Language

- *Verbs* have present and past tense, inflect for person and number and, have a fully inflected subjunctive in both tenses, etc.
- Inflection primarily marked by endings, often accompanied by morphophonemic alterations of the stem
- Most productive word formation process is *composition*
- *Derivation* is productive in word formation, both suffixation and prefixation
- Icelandic has a basic SVO order (subject – verb – object), in subordinate as well as in main clauses
- Inflection takes much of the load of showing the internal structure of sentences so word order is relatively free

Tagging of Icelandic Text – ongoing work
The Data

- Icelandic Frequency Dictionary, published in 1991 (IFD) a carefully balanced corpus consisting of just over half a million running words
- Criteria for Choosing the material
  - 100 fragments of texts, approximately 5,000 running words each.
  - All texts are published for the first time in 1980–1989.
  - Five categories of texts:

a) Icelandic fiction
b) Translated fiction
c) Biographies and memoirs
d) Non-fiction (evenly divided between science & humanities)
e) Books for children and youngsters (original Icelandic and translations)

No two texts can be attributed to the same person, i.e., as author or translator

(a

Tagging of Icelandic Text – ongoing work
The Data

- Tagging of the IFD
  - (1) 54.000 running words were handtagged in a pilot project
  - (2) Automatic tagging, based on the handtagged material in (1); 50 texts, 250.000 running words
  - (3) Automatic tagging, with improvements to the tagger based on (2); 50 texts, 250.000 running words
- The tagger developed used a mix of morphological features, syntax rules and a statistical approach
- About **80**% of words were correctly analyzed by the program. Subsequent work on the program improved its performance on ordinary texts to just below **90**%

Tagging of Icelandic Text – ongoing work
The Data

- Tagset of the IFD
  - Based on the traditional Icelandic analysis of word classes and grammatical categories, with some exceptions where that classification has been rationalized
  - In the original computer files case-assignment of verbs was included in the tags. The case-assignment was removed in the analysis presented here and the tagset used is the same as was used for the printed version of the IFD
  - There are **621** tags in the tagset of the printed IFD even though there are gaps in the set which would probably be filled with a larger corpus

Tagging of Icelandic Text – ongoing work
The Data

- 59,343 word forms in the entire IFD
- 15.9% of the word froms are ambiguous as to the tagset within the IFD
- The most ambiguous of word forms in the IFD, *minni*, has 24 tags in the coprus, and has not exhausted its possibilities

Tagging of Icelandic Text – ongoing work
The Data
Ambiguity of word form *minni*

Lexeme:

lítill adjective

minni noun, neut.

minn possess. pron.

minna verb

Tags grammatical features:

minni ‘smaller’, comparative, weak declension:
4 tags: LMVKEN LMVKEO LMVKEÞ LMVKEE
(masc.sg. nom., acc., dat., gen.)
3 tags: LMVKFO LMVKFÞ LMVKFE
(masc..pl. acc., dat., gen.)
4 tags: LMVVEN LMVVEO LMVVEÞ LMVVEE
(fem..sg. nom., acc., dat., gen.)
3 tags: LMVHFN LMVHFO LMVHFE
(neut.pl. nom., acc., gen.)
(gaps: masc..pl..nom. and neut..pl.dat.)
‘memory’:
5 tags: NHEN NHEO NHEÞ NHFN NHFO
(sg..nom., acc., dat.;pl..nom., acc.)
‘mine’
1 tag: FEVEÞ (fem.sg.dat.)
minni ‘remind’ (sby of sth.), ‘remember’
1 tag: SGVNE3
(active voice, subjunct., pres., sing. 3rd pers.)
(a number of gaps in the paradigm)

Tagging of Icelandic Text – ongoing work
Aim of project

- Use data-driven methods to develop a tagger for Icelandic in as short a time as possible with at least 92% accuracy
- The tagger should be able to tag running words in a variety of Icelandic text according to word classes and morphological features
- The tagger should be able to tag previously unseen words and disambiguate multiple tags from their context

Tagging of Icelandic Text – ongoing work

Methods and taggers used in the Icelandic experiment

- **Hidden Markov Models**
  - TRIGRAMS'N'TAGS (TnT) tagger of Thorsten Brants (Brants 2000). Uses second order Markov models for part-of-speech tagging. Unknown words handled with suffix analysis, the longest suffix used in TnT is 10 characters long

Tagging of Icelandic Text – ongoing work

Methods and taggers used in the Icelandic experiment

- **Maximum Entropy Learning**
  - **MXPOST, developed for part-of-speech tagging by Ratnaparkhi was used in the Icelandic experiment (Ratnaparkhi 1996).**
  - **Training data are described as a large number of features which are binary valued functions of histories (word and tag context) and tags.**
  - **In Ratnaparkhi's implementation the features include the current word, the preceding two words, the following two words and the preceding two tags.**
  - **For rare and unknown words the features also include the first and last four characters and information about whether the word contains an uppercase letter, hyphens or numbers**

Tagging of Icelandic Text – ongoing work

Methods and taggers used in the Icelandic experiment

- **Memory-Based Learning**
  - **Method that stores the training data in memory**
  - **Similarity of new situations is compared to stored representations of earlier experiences**
  - **The similarity between a new instance and all examples in memory is computed using a distance metric.**
  - **Tilburg learner, TiMBL, several memory-based learning approaches with different types of distance metrics are implemented**

Tagging of Icelandic Text – ongoing work

Methods and taggers used in the Icelandic experiment

- **Memory-Based Learning**
  - **Use previous tagger decisions as input for current decisions, build separate case bases for known and unknown words, allow global sentence-level optimization, etc**
  - **MBT software (Daelemans *et al*. 2002a), supplied by the same group as TiMBL, implements this specific tagging functionality by wrapping software around TiMBL while keeping most of the flexibility of TiMBL intact.**

Tagging of Icelandic Text – ongoing work

Methods and taggers used in the Icelandic experiment

- **Memory-Based Learning**
  - **Define feature patterns for both known and unknown words. Used in experiment:**
- **For known words:**
  - **ddwdwfWawawa** **focus ambitag and the focus word with three disambiguated tags on the left and three ambitags on the right plus the corresponding word for two of the contexts on the left and two on the right**
- **For unknown words:**
  - **chnssssdwFaw** **the focus contains capitalized letters, the focus contains a hyphen, the focus contains numerical characters, four last characters of the word, one disambiguated tag on the left and one ambitag on the right, the left and right neighboring words**

Tagging of Icelandic Text – ongoing work

Transformation-Based Error-Driven Learning (Brill-tagging)

- **Tagger chosen: fnTBL**
- **Corpus-based, relies on accurately tagged corpus**
- **Initial-state annotator, assign most likely tag to each word**
- **Compare output to true tags**
- **Learn transformation rules that applied to output of initial-state annotator make it more like the truth**
- **The rules are learned from templates that describe an action**
  - **Action (change tag A to tag B)**
  - **Triggering environment**
    - **Only tags (non-lexicalized)**
    - **Tags and words (lexicalized)**
- **Apply transformation and count errors**
- **Choose transformation that reduces errors most**

Tagging of Icelandic Text – ongoing work

Transformation-Based Error-Driven Learning (Brill-tagging)

- **Unknown word guesser of Brill**
  - **Initially tagging capitalized words as proper nouns, common nouns otherwise**
  - **Transformation-based learner to learn rules for accurately guessing PoS of words not seen in the training corpus**
  - **Allowable transformations:**
    - **Look at 1–4 first or last letters of a word**
    - **Adding or deleting a suffix or prefix of 1–4 letters results in a word**
    - **Appears a particular word immediately to left or right of the word?**
    - **Contains the word a specified character?**

Tagging of Icelandic Text – ongoing work

Transformation-Based Error-Driven Learning (Brill-tagging)

- **Implementation chosen: fnTBL by Radu Florian an Grace Ngai (Florian and Ngai 2002).**
  - **Contextual templates. Use set of templates provided with the program, contains 40 templates as opposed to the 26 original Brill templates. The extra templates mostly extend the triggering environment to the identity of a combination of words and part-of-speech tags.**
  - **Lexical templates. Set contains all the templates suggested by Brill and additionally templates for a prefix and suffix of 5 characters. The set contains templates both with and without conditioning on the part-of-speech tag of the word being considered.**

Tagging of Icelandic Text – ongoing work

10-fold cross-validation

- Divide each text into 10 approximately equal parts
- Make from these ten different disjoint pairs of files
- Each pair contains a training set containing about 90% of running words from the corpus and a test set containing about 10% of running words from the corpus.
- The test sets are independent of each other
- Training sets overlap and share about 80% of the examples

Tagging of Icelandic Text – ongoing work

Baseline

- Baseline, i.e. the lowest accuracy that a tagger should be able to achieve on the test set
  - Use method reported in (Megyesi 2002:55).
- Derive a lexicon from the training set where all word forms are given all possible tags that appear for that word form in the training set. Words in the test set are then given the tag that occurs most frequently for that word form in the training set.
- Unknown (do not occur in the training set) words are treated in three different ways:
  - Unknown words are considered incorrectly tagged.
  - Unknown words get the tag that occurs most frequently in the training set.
  - Unknown words that are not capitalized are tagged with the most frequently occurring tag for common nouns. Unknown capitalized words are tagged with the most frequently occurring tag for proper nouns

.

Tagging of Icelandic Text – ongoing work

Baseline

- Calculated for the first pair of training and test data.
- The percentage of unknown words in that test set with respect to the corresponding training set is 7.57
- Unknown words treated as incorrectly tagged, baseline = 75.43%
- Unknown words tagged with the most frequently occurring tag (aa, adverbs that do not govern case), baseline = 75.50%
- Unknown words that are not capitalized are tagged with the most frequently occurring tag for common nouns (*nken*, noun, masculine, singular, nominative) and unknown capitalized words are tagged with the most frequently occurring tag for proper nouns (*nken-m*, noun, masculine, singular, nominative, name of a person), baseline = 75.81%.

Tagging of Icelandic Text – ongoing work
Evaluation of results

- *Accuracy=(number of correctly tagged tokens)/(total number of tokens in test set)*
  - Accuracy figures are usually given both for known and unknown words.
- Evaluate the performance for each individual tag
  - *precision=(number of correctly tagged tokens with tag X)/(total number of tokens with tag X given by tagger)*
  - *recall=(number of correctly tagged tokens with tag X)/(total number of tokens with tag X in test set)*
- *F-score* is calculated to measure the balance between precision and recall. The F-score is defined as the harmonic mean of precision and recall.
  - Parameter β. When $\beta = 1$ precision and recall have equal weight, when $\beta > 1$ precision gets higher weight, when $\beta < 1$ recall gets more weight. The F-score is defined as:
  - $F=(\beta^2+1)*P*R/(\beta^2*P+R)$
  - When $\beta = 1$ F-score is defined as:
  - $F=2*P*R/(P+R)$

Tagging of Icelandic Text – ongoing work

Evaluation of taggers

Table 1. Some statistics for the first pair of training and test sets used for evaluation of taggers

Number of	Training set	Test set
Tokens	531.128	59.169
Types	55.188	13.279
Tags *	639	552

* Including punctuation tags

Tagging of Icelandic Text – ongoing work

Evaluation of results

Table 2. The tagging accuracy for all words, known words and unknown words for four taggers

	MBT	MXPOST	fnTBL	TnT
Accuracy	%	%	%	%
All words	79,16	88,39	88,28	89,74
Known words	81,63	90,54	91,09	91,43
Unkown words	49,11	62,23	54,08	69,23

Tagging of Icelandic Text – ongoing work

Evaluation of results

Difference between performance of taggers

McNemar's chi-squared test:

	McNemar's test	p
TnT vs fnTBL	151,13	<0.001
TnT vs. MXPOST	110.41	<0.001
fnTBL vs MXPOST	0.71	n.s.

Tagging of Icelandic Text – ongoing work
Evaluation of results

- Major difficulty in annotating Icelandic words stems from the difficulty in finding the correct tag for unknown words
- Test set
  - open word classes (nouns, adjectives and verbs) 51%
  - unknown words, open word classes (nouns, adjectives and verbs) 96%
- The three taggers have different procedures for annotating unknown words and this is reflected in the difference in performance

Tagging of Icelandic Text – ongoing work

Evaluation of results

Table 3. The tagging accuracy for all words, for known and unknown words for three POS taggers when tagging and testing on the original tagset but only considering the correctness of the 10 word classes.

	MXPOST	fnTBL	TnT
Accuracy	%	%	%
All words	96,94	97,11	97,94
Known words	97,68	98,34	98,47
Unkown words	87,93	82,11	91,43

Tagging of Icelandic Text – ongoing work

Evaluation of results

Table 4. The tagging accuracy for all words, for known and unknown words for three POS taggers when tagging and testing on the original tagset and not considering morphological features of adverbs.

	MXPOST	fnTBL	TnT
Accuracy	%	%	%
All words	89,45	89,22	90,88
Known words	91,68	92,10	92,66
Unkown words	62,27	54,11	69,23

Tagging of Icelandic Text – ongoing work

Evaluation of results

Table 5. Comparison of tagging accuracy for morphological tags and word classes

	MXPOST	fnTBL	TnT
Accuracy	%	%	%
Correctly tagged	88,39	88,28	89,74
Correct word class, morphology incorrect	8,55	8,83	8,20
Wrong word class	3,06	2,89	2,06

Tagging of Icelandic Text – ongoing work

Evaluation of results

Table 6. P: precision; R: recall; F: F-score for three taggers

Word class	No. in test set	MXPOST			fnTBL			TnT		
		P	R	F (β=1)	P	R	F (β=1)	P	R	F (β=1)
a (adverb)	11.660	97,42	97,98	97,70	97,82	98,32	98,07	97,74	98,21	97,97
c (conjunction)	5.959	97,96	98,94	98,45	98,42	99,01	98,71	98,15	98,86	98,50
e (foreign word)	42	55,00	26,19	35,48	34,62	21,43	26,47	59,26	38,10	46,38
f (pronoun)	7.035	99,02	97,63	98,32	99,10	98,71	98,90	98,75	98,65	98,70
g (article)	67	70,24	88,06	78,15	83,56	91,04	87,14	95,74	67,16	78,95
l (adjective)	3.914	89,22	85,03	87,07	90,75	85,26	87,92	93,59	91,08	92,32
n (noun)	13.129	96,22	96,96	96,59	96,46	96,76	96,61	98,45	98,49	98,47
s (verb)	9.793	96,70	97,30	97,00	95,92	97,13	96,52	97,41	98,05	97,73
t (numeral)	718	94,41	94,15	94,28	93,54	94,85	94,19	95,81	92,34	94,04
x (not analysed)	14	37,50	21,43	27,27	55,56	71,43	62,50	47,37	64,29	54,55

Tagging of Icelandic Text – ongoing work

Evaluation of results

Table 7. Comparison of the performance of three taggers
Percentages computed of total number of words in each word class in test set

		Recall for MXPOST			Recall for fnTBL			Recall TnT		
		Corr. anal.	Corr. w.class	Wrong w. class	Corr. anal.	Corr. w.class	Wrong w. class	Corr. anal.	Corr. w.class	Wrong w. class
Word class	Freq.	%	%	%	%	%	%	%	%	%
a (adverb)	11.660	92,62	5,37	2,59	93,58	4,73	2,20	92,44	5,77	2,27
c (conjunction)	5.959	97,33	1,61	2,06	97,82	1,19	1,59	97,10	1,76	1,86
e (foreign word)	42	26,19	0,00	21,43	21,43	0,00	40,48	38,10	0,00	26,19
f (pronoun)	7.035	87,31	10,32	0,97	89,35	9,35	0,90	89,18	9,47	1,25
g (article)	67	62,69	25,37	37,31	76,12	14,93	17,91	56,72	10,45	2,99
l (adjective)	3.914	65,02	20,01	10,27	61,60	23,66	8,69	70,82	20,26	6,23
n (noun)	13.129	79,79	17,17	3,81	78,60	18,16	3,55	83,74	14,75	1,55
s (verb)	9.793	93,10	4,21	3,32	92,52	4,62	4,14	92,39	5,66	2,60
t (numeral)	718	73,96	20,19	5,57	71,17	23,68	6,55	76,32	16,02	4,04
x (not analysed)	14	21,43	0,00	35,71	71,43	0,00	57,14	64,29	0,00	71,43

Tagging of Icelandic Text – ongoing work

Evaluation of results

Table 8. Twenty most common errors made by each tagger

MXPOST			fnTBL			TnT		
tagged> correct	Freq.	%	tagged> correct	Freq.	%	tagged> correct	Freq.	%
aþ>ao	232	3,38	ao>aþ	191	2,76	aþ>ao	177	2,92
ao>aþ	139	2,02	aþ>ao	142	2,05	ao>aþ	153	2,52
ao>aa	63	0,92	nveo>nveþ	94	1,36	ao>aa	107	1,76
nveþ>nveo	59	0,86	nveþ>nveo	74	1,07	nveo>nveþ	89	1,47
aa>ao	57	0,83	sng>sfg3fn	69	1,00	nhen>nheo	76	1,25
nveo>nveþ	56	0,82	nheo>nhen	68	0,98	nveþ>nveo	76	1,25
c>aþ	49	0,71	nhen>nheo	67	0,97	nheo>nhen	71	1,17
nhen>nheo	49	0,71	lhensf>lheosf	59	0,85	aþ>aa	70	1,15
lvensf>lhfnsf	46	0,67	sfg3eþ>sfg1eþ	56	0,81	lhensf>lheosf	69	1,14
sfg3eþ>sfg1eþ	46	0,67	nkeo>nkeþ	54	0,78	ssg>sþghen	67	1,10
aa>aþ	45	0,66	aa>aþ	50	0,72	aa>ao	65	1,07
lhensf>lheosf	44	0,64	ao>aa	46	0,66	nkeþ>nkeo	63	1,04
nkeþ>nkeo	43	0,63	nvfn>nvfo	43	0,62	sng>sfg3fn	62	1,02
nvfn>nvfo	42	0,61	aa>ao	42	0,61	sfg3eþ>sfg1eþ	58	0,96
nheo>nhen	39	0,57	nheog>nheng	42	0,61	ct>c	52	0,86
sþghen>ssg	39	0,57	nvfo>nvfn	42	0,61	lvensf>lhfnsf	49	0,81
sfg3fn>sng	38	0,55	nkeþ>nkeo	41	0,59	sþghen>ssg	48	0,79
nkeo>nkeþ	35	0,51	aa>lhensf	40	0,58	nhfo>nhfn	47	0,77
aa>ae	34	0,50	ct>c	39	0,56	nkeo>nkeþ	47	0,77
aa>fpheþ	32	0,47	nhfo>nhfn	38	0,55	nvfo>nvfn	45	0,74

Tagging of Icelandic Text – ongoing work

Evaluation of results

Table 9. Tagger agreement on the correct tag

Vote	Correct	%	Cum. %
3 taggers	47.978	81,09	81,09
2 taggers	5.387	9,10	90,19
1 tagger	2.931	4,95	95,14
none	2.873	4,86	100,00

Tagging of Icelandic Text – ongoing work

Evaluation of results

Table 10. Tagger aggreement on icorrect tag

Vote	Incorrect	%	Cum. %
3 taggers	1.297	2,19	2,19
2 taggers	1.938	3,28	5,47

Tagging of Icelandic Text – ongoing work

Evaluation of results

Table 11. Pairwise comparison of taggers

Pair	Agreement correct %	Agreement incorrect %	Total %
TnT and MXPOST	84,18	3,53	87,71
TnT and fnTBL	**84,85**	4,7	89,55
MXPOST and fnTBL	83,33	3,6	86,93

Tagging of Icelandic Text – ongoing work

Evaluation of results

Table 12. The 35 most common tokens that all taggers agreed upon and were correct

Word	Tag	Freq.	%	English
.	.	3.249	6,77	(.)
og	c	2.323	4,84	(and)
,	,	2.313	4,82	(,)
að	cn	972	2,03	(to)
í	aþ	962	2,01	(in)
var	sfg3eþ	803	1,67	(was)
hann	fpken	638	1,33	(he)
á	aþ	632	1,32	(on)
en	c	543	1,13	(but)
sem	ct	540	1,13	(that)
ekki	aa	530	1,10	(not)
að	c	511	1,07	(that)
er	sfg3en	506	1,05	(is)
ég	fp1en	502	1,05	(I)
hún	fpven	429	0,89	(she)
um	ao	411	0,86	(about)
til	ae	372	0,78	(to)

Tagging of Icelandic Text – ongoing work

Evaluation of results

Table 13. The 35 most common tokens that all taggers classified with the same but erreneous tag.

Word	Correct tag	Incorrect tag	Freq.	%	English
í	ao	aþ	24	1,85	(in)
sem	c	ct	18	1,39	(that)
sér	fpveþ	fpkeþ	14	1,08	
þá	fpkfo	aa	13	1,00	(them)
sig	fpveo	fpkeo	12	0,93	
á	ao	aþ	11	0,85	(on)
hann	fpkeo	fpken	11	0,85	(him)
um	aa	ao	10	0,77	(about)
þeirra	fphfe	fpkfe	9	0,69	(them)
í	aþ	ao	8	0,62	(in)
það	fahen	fphen	8	0,62	(it)
það	fpheo	fphen	8	0,62	(it)
á	aþ	ao	7	0,54	(on)
sér	fpheþ	fpkeþ	7	0,54	
til	ae	aa	7	0,54	(to)
þeim	fpvfþ	fphfþ	7	0,54	(them)
neitt	fohen	foheo	6	0,46	(none)

Tagging of Icelandic Text – ongoing work

Evaluation of results

Table 14. Tag confusion by all three taggers
35 most common cases

Correct tag	Wrong tag	Freq	%	Cum. %
ao	aþ	138	2,89	2,89
aþ	ao	122	2,56	5,45
aa	ao	90	1,89	7,34
nveþ	nveo	68	1,43	8,76
nveo	nveþ	66	1,38	10,14
aa	aþ	65	1,36	11,51
sþghen	ssg	61	1,28	12,79
nhen	nheo	59	1,24	14,02
ao	aa	55	1,15	15,18
nheo	nhen	55	1,15	16,33
nkeo	nkeþ	48	1,01	17,33
sfg3fn	sng	42	0,88	18,21
ssg	sþghen	42	0,88	19,09
nvfn	nvfo	41	0,86	19,95
nhfn	nhfo	40	0,84	20,79
lheosf	lhensf	38	0,80	21,59
nkeþ	nkeo	36	0,75	22,34

Tagging of Icelandic Text – ongoing work

Evaluation of results

Table 15. 10-fold cross-validation test for TnT and MXP

	No. of running words			Unknown	Accuracy for MXPOST %			Accuracy for TnT %		
Pair	All	Known	Unknown	%	All	Known	Unknown	All	Known	Unknown
01	59.169	54.687	4.482	7,57	88,39	90,54	62,23	89,74	91,43	69,23
02	58.967	54.961	4.006	6,79	88,93	91,00	60,58	90,12	91,64	69,35
03	59.077	55.014	4.063	6,88	89,31	91,27	62,81	90,31	91,73	71,13
04	59.067	55.113	3.954	6,69	89,25	91,07	63,96	90,53	91,85	72,05
05	59.075	55.227	3.848	6,51	89,47	91,31	63,12	**90,69**	92,00	71,80
06	59.136	55.172	3.964	6,70	88,91	90,80	62,66	90,32	91,70	71,14
07	59.109	55.148	3.961	6,70	89,33	91,30	62,03	90,59	92,02	70,61
08	58.981	54.891	4.090	6,93	88,95	90,88	63,03	90,19	91,59	71,44
09	59.143	55.122	4.021	6,80	88,80	90,88	60,26	90,05	91,50	70,08
10	58.573	54.570	4.003	6,83	**89,49**	91,33	64,40	90,49	91,88	71,62
Mean				6,84	89,08	91,04	62,51	90,30	91,73	70,85

Tagging of Icelandic Text – ongoing work

Evaluation of results

Performance of TnT in training and testing on reduced data sets

Mean accuracy %	Remove chemical text Reduction in size of 1%	Remove scientific texts Reduction in size of 11%
Unknown words %	6,84	6,81
All words	90,33	90,45
Known words	91,76	91,91
Unkown words	70,86	70,50

Tagging of Icelandic Text – ongoing work

Evaluation of results

Table 16. Tagging accuracy found in experiments with two different taggers applied to Icelandic text. Size of training corpus 59k

	Tagset			
	1.2	2	3	9
No. of tags (excluding punctuation tags)	602	473	191	10
Total accuracy with TnT (%)	84.51	86.53	86.75	95.07
Accuracy for known words with TnT (%)	89.63	91.17	90.74	97.98
Accuracy for unknown words with TnT (%)	50.67	55.9	60.43	74.88
Total accuracy with µTBL (%)	92.1	94.5	95.2	98.8

Version 1.2
Tagset in the prototype of the IFD, including case assignment of verbs.
Version 2
The full tagset for the IFD
Version 3
Gender classification was removed from the tagset for nouns, adjectives, pronouns, numerals and past participles and the classification of pronouns.
Version 9
Tags for word classes

Tagging of Icelandic Text – ongoing work

Evaluation of results

Performance of TnT when training and testing on different sizes of data sets (520k mean accuracy)

Size of training corpus	59k	520k
Unknown words %	13,16	6,84
All words, accuracy %	86,53	90,30
Known words, accuracy %	91,17	91,73
Unkown words, accuracy %	55,90	70,85

Tagging of Icelandic Text – ongoing work
Conclusion

- TnT gives best results
- Accuracy increases with smaller in tagset
- Accuracy increases with increase in size of training corpus
- Type of text matters
- Percentage of unknown words is important
- Unknown word handling is important
- Data-driven part-of-speech tagging of Icelandic is feasible

Tagging of Icelandic Text – ongoing work

Future work

- Complete 10-fold cross-validation test and any necessary statistical analysis
- Examine the possibility of combining the results of two or more taggers (simple averaging, majority voting, training a new classifier, linguistically motivated rules)
- Improve handling of unknown words, e.g. by providing a large lexicon