The Icelandic tagset

Hrafn Loftsson
Department of Computer Science
Reykjavik University
e-mail: hrafn@ru.is

January 2007

1 Introduction

The main Icelandic tagset was created in the making of the Icelandic Frequency Dictionary (*IFD*) corpus (Pind et al. 1991). Due to the morphological richness of the Icelandic language, the main tagset is large and makes fine distinctions compared to related languages. The rich inflections of an Icelandic word contribute more information about POS of surrounding words than is the case, for example, for English where word order is not as free. The tagset consists of 700 possible tags, 639 tags of which appear in the *IFD* corpus.

We can illustrate the preciseness of the tags by examining the semantics of a tag. Each character in the tag has a particular function. The first character denotes the word class. For each word class there is a predefined number of additional characters (at most six) which describe morphological features, like *gender*, *number* and *case* for nouns; *degree* and *declension* for adjectives; *voice*, *mood* and *tense* for verbs, etc.

Tables 1 and 2 show the semantics of the noun and the adjective tags, and the semantics of the verb tags, respectively. Consider, for example, the tag "*nken*". The first letter, "*n*", denotes the word class "*nafnorð*" (noun), the second letter, "*k*", denotes the gender "*karlkyn*" (masculine), the third letter, "*e*", denotes the number "*eintala*" (singular) and the last letter, "*n*", denotes the case "*nefnifall*" (nominative case).

To give another example, consider the phrase "*fallegu hestarnir hoppuðu*" (the beautiful horses jumped). The corresponding tag for "*fallegu*" is "*lkenvf*", denoting adjective, masculine, singular, nominative, weak declension, positive; the tag for "*hestarnir*" is "*nkfng*",

Char #	Category/ Feature	Symbol – semantics
1	Word class	**n**–noun, **l**–adjective
2	Gender	**k**–masculine, **v**–feminine, **h**–neuter, **x**–unspecified
3	Number	**e**–singular, **f**–plural
4	Case	**n**–nominative, **o**–accusative, **þ**–dative, **e**–genitive
5	Article	**g**–with suffixed article
5	Declension	**v**–strong, **s**–weak
6	Proper noun	**m**–person, **ö**–place, **s**–other
6	Degree	**f**–positive, **m**–comparative, **e**–superlative

Table 1: The semantics of the tags for nouns and adjectives.

Char #	Category/ Feature	Symbol – semantics
1	Word class	**s**–verb (except for past participle)
2	Mood	**n**–infinitive, **b**–imperative, **f**–indicative, **v**–subjunctive, **s**–supine, **l**–persent participle
3	Voice	**g**–active, **m**–middle
4	Person	**1**–1st person, **2**–2nd person, **3**–3rd person,
5	Number	**e**–singular, **f**–plural
6	Tense	**n**–present, **þ**–past

Table 2: The semantics of the tags for verbs.

denoting noun, masculine, plural, nominative with suffixed definite article, and the tag for "*hoppuðu*" is "*sfg3fþ*", denoting verb, indicative mood, active voice, 3-rd person, plural and past tense. Note the agreement in gender, number and case between the adjective and the noun, and the agreement in person and number between the adjective/noun and the verb.

A complete description of the tagset can be found in the Appendix A.

References

J. Pind, F. Magnússon, and S. Briem. *Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]*. The Institute of Lexicography, University of Iceland, Reykjavik, Iceland, 1991.

A The Icelandic tagset

Table 3: The Icelandic tagset

Char#	Category/Feature	Symbol – semantics
1	Word class	**n**–noun
2	Gender	**k**–masculine, **v**–feminine, **h**–neuter, **x**–unspecified
3	Number	**e**–singular, **f**–plural
4	Case	**n**–nominative, **o**–accusative, **þ**–dative, **e**–genitive
5	Article	**g**–with suffixed definite article
6	Proper noun	**m**–person name, **ö**–place name, **s**–other proper name
1	Word class	**l**–adjective
2	Gender	**k**–masculine, **v**–feminine, **h**–neuter
3	Number	**e**–singular, **f**–plural
4	Case	**n**–nominative, **o**–accusative, **þ**–dative, **e**–genitive
5	Declension	**s**–strong declension, **v**–weak declension, **o**–indeclineable
6	Degree	**f**–positive, **m**–comparative, **e**–superlative
1	Word class	**f**–pronoun
2	Subcategory	**a**–demonstrative, **b**–reflexive, **e**–possessive, **o**–indefinite, **e**–personal, **s**–interrogative, **t**–relative
3	Gender/Person	**k**–masculine, **v**–feminine, **h**–neuter/**1**–1st person, **2**–2nd person
4	Number	**e**–singular, **f**–plural
5	Case	**n**–nominative, **o**–accusative, **þ**–dative, **e**–genitive
1	Word class	**g**–article
2	Gender	**k**–masculine, **v**–feminine, **h**–neuter
3	Number	**e**–singular, **f**–plural
4	Case	**n**–nominative, **o**–accusative, **þ**–dative, **e**–genitive
1	Word class	**t**–numeral
2	Category	**f**–alpha, **o**–numeric, **p**–percentage
3	Gender	**k**–masculine, **v**–feminine, **h**–neuter
4	Number	**e**–singular, **f**–plural
5	Case	**n**–nominative, **o**–accusative, **þ**–dative, **e**–genitive
1	Word class	**s**–verb (except for past participle)
2	Mood	**n**–infinitive, **b**–imperative, **f**–indicative, **v**–subjunctive, **s**–supine, **l**–persent participle
3	Voice	**g**–active, **m**–middle
4	Person	**1**–1st person, **2**–2nd person, **3**–3rd person,
5	Number	**e**–singular, **f**–plural
6	Tense	**n**–present, **þ**–past
1	Word class	**s**–verb (past participle)
2	Mood	**þ**–past participle
3	Voice	**g**–active, **m**–middle
4	Gender	**k**–masculine, **v**–feminine, **h**–neuter
5	Number	**e**–singular, **f**–plural
6	Case	**n**–nominative, **o**–accusative, **þ**–dative, **e**–genitive
1	Word class	**a**–adverb and preposition
2	Category	**a**–does not govern case, **u**–exclamation, **o**–governs accusative, **þ**–governs dative, **e**–governs genitive
3	Degree	**m**–comparative, **e**–superlative
1	Word class	**c**–conjunction
2	Category	**n**–sign of infinitive, **t**–relative conjunction,
1	Word class	**e**–foreign word
1	Word class	**x**–unanalyzed word