T-(538|725)-MALV, Natural Language Processing
Part-of-speech and morphology

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]School of Computer Science, Reykjavik University

September 2010

Outline

Outline

Part-of-speech (PoS) (í. orðflokkar)

Definition

- Word classes, whose words share common grammatical properties.
  - Morphological proporties (e.g. inflection vs. no inflection)
  - Syntactic properties (e.g. adjectives often modify nouns)
  - Semantic properties (e.g. nouns "name" things, adjectives describe attributes, verbs describe actions)
- Also called *lexical catagories*.

Two main classes

- Closed class
- Open class

Part-of-speech

Closed class

- Relatively stable over time, new words are not added to this class.
- Function words:
  - Articles/Determiners (í. ákvæðisorð): **the**, **several** (English), **der** (German)
  - Prepositions (í. forsetningar): **to**, **of** (English), **í**, **á** (Icelandic)
  - Conjunctions (í. samtengingar): **and**, **or** (English), **und**, **oder** (German)
  - Auxiliaries/modals (í. hjálparsagnir): **be**, **have** (English), **vera**, **hafa** (Icelandic)
  - Adverbs (í. atviksorð), but not all of them.

Part-of-speech

Open class

- Forms the bulk of a vocabulary.
- Words get constantly added to this class.
- **Nouns** (í. nafnorð), **adjectives** (í. lýsingarorð), **verbs** (í. sagnir) (except auxiliaries/modals), **adverbs** (í. atviksorð; eingöngu háttaratviksorð í íslensku).

11 word classes in Icelandic

Fallorð Words which inflect	Sagnorð Verbs	Smáorð Function words
Nafnorð (e. nouns)	**Sagnorð** (e. verbs)	**Forsetningar** (e. prepositions)
Lýsingarorð (e. adjectives)		**Atviksorð** (e. adverbs)
Fornöfn (e. pronouns)		**Samtengingar** (e. conjunctions)
Töluorð (e. numerals)		**Upphrópanir** (e. interjections)
Greinir (e. article)		**Nafnháttarmerki** (e. infinitive marker)

Words which inflect (í. fallorð)

Features (í. beygingarleg einkenni)
(the bold font refers to the letters in the Icelandic PoS tagset)

- Kyn (e. gender)
  - **k**arlkyn (e. masculine)
  - k**v**enkyn (e. feminine)
  - **h**vorugkyn (e. neuter)
- Tala (e. number)
  - **e**intala (e. singular)
  - **f**leirtala (e. plural)
- Fall (e. case)
  - **n**efnifall (e. nominative)
  - þ**o**lfall (e. accusative)
  - **þ**águfall (e. dative)
  - **e**ignarfall (e. genitive)

Verbs

Features

- Háttur (e. mood)
  - **f**ramsöguháttur (e. indicative mood)
  - **v**iðtengingarháttur (e. subjunctive mood)
  - **b**oðháttur (e. imperative mood)
  - **n**afnháttur (e. infinitive mood)
  - **l**ýsingarháttur nútíðar (e. present participle)
  - lýsingarháttur **þ**átíðar (e. past participle)
- Mynd (e. voice)
  - **g**ermynd (e. active)
  - **m**iðmynd (e. middle)
  - þolmynd (e. passive) - generated using past participle (í. lýsingarháttur þátíðar).

Verbs

Features

- Persóna (e. person): **1**., **2**., **3**.
- Tala (e. number): **e**intala, **f**leirtala
- Tíð (e. tense)
  - **n**útíð (e. present)
  - **þ**átíð (e. past)

The Icelandic PoS tagset

http://cadia.ru.is/wiki/_media/public:t-malv-10-3:icelandictagset.pdf

Outline

A lexicon (í. orðasafn)

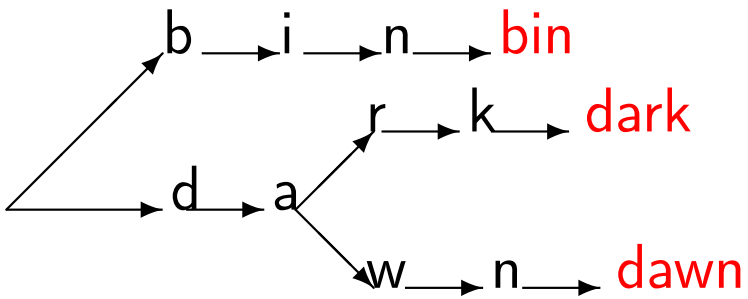
Contains what?

- A list of words, lexical entries (lexemes), with or without **annotations**.
  - Pronunciation
  - Morphology
  - PoS tags
  - Syntactic and semantic labels.
- A lexicon with annotation is sometimes called a *dictionary*.
- Often generated to cover a particular domain, e.g. technology, science, finance.
- Generally the first unit needed for a Language Technology System.

A lexicon/dictionary

Data structure

- Hash table (key–value): each word is a key and further information about the word is stored in the value.
  - Memory intensive for large lexicons.
- *Letter trees* (tries) `http://en.wikipedia.org/wiki/Trie`
  - Words are stored as trees of characters and they share branches as far as the letters of two words are identical.



Outline

Morphology (í. orðhlutafræði)

Morpheme (í. myndan/morfem)

- The smallest meaningful unit of a word
- Divide into **stems** (í. stofn) and **affixes** (í. aðskeyti)
- The stem is the common part shared by all word forms
- Example: The English word form "churches" consists of two morphemes:
  - The stem "church"
  - The plural suffix/ending "es"
- Example: The Icelandic word form "orði" consists of two morphemes:
  - The stem "orð"
  - The inflectional ending "i" (denoting the dative case (í. þágufall)).

Morphology

Affixes

- Prefix (í. forskeyti)
  - Appears before the stem
  - Example: rewrite, örfínn
- Suffix (í. viðskeyti)
  - Appears after the stem
  - Example: teacher, kennari
- Infix (í. innskeyti)
  - Squeezed into the stem
  - Example: humingi in tagalog; hingi: to lend, humingi: the person who gets the loan
- Circumfix (í. umskeyti)
  - Appears both before and after the stem
  - Example: gesagt in German past participle

Morphology

- http://en.wikipedia.org/wiki/Morphology_(linguistics)
- The theory of how words are built from morphemes
- Two kinds of systems:
  - Concatenative morphology
    - For example, Germanic languages
    - http://en.wikipedia.org/wiki/Germanic_languages
    - Prefix + Stem + Suffix
  - Non-concatenative (templatic) morphology
  - http://en.wikipedia.org/wiki/Nonconcatenative_morphology
    - For example, Arabic or Hebrew

Icelandic morphology

Words are generated from morphemes in two ways:

Using inflection (í. beyging)

- Prefix + Stem + inflectional ending
- The new word belongs to the same word class as the stem.
- Example: hest<u>ur</u> (e. a horse), kon<u>a</u> (e. a woman)

Using derivation (í. afleiðsla)

- Prefix + Stem + Suffix
- The new word can belong to a different word class
- Example: <u>mis</u>jafn (e. unequal), klofn<u>ingur</u> (e. a split), kenn<u>ari</u> (e. a teacher)

Morphology

Morphological analysis/parsing

- í. orðhlutafræðileg greining/þáttun
- The task of breaking a word into morphemes.
- A necessary unit in many LT systems:
  - **Lemmatisation** (í. lemmun): The task of finding the lemma of a word – the dictionary form (í. flettimynd).
    - An Icelandic lemmatiser: `http://www.springerlink.com/content/h530q7157285563u/`
  - **Stemming**: The task of finding the stem of a word. For example, important in Information Retrieval.
  - Words not found in a lexicon (unknown words) need analysis, for example in PoS tagging

Morphological generation = Generating words given morphemes.

Lemmatisation

Lemmatisation vs. Stemming

- Example: “hesti”. Lemma = “hestur”. Stem = “hest”

Ambiguity in lemmatisation.

1. A **run** in the forest. Lemma: **run**, a singular noun.
2. The sportsmen **ran** everyday. Lemma: **run**, a verb in past tense, third person, plural.
3. **Rétturinn** var fullskipaður. Lemma: **réttur**, a noun in masculine, singular, nominative, suffixed definite article.
4. Dæmið var **rétt**. Lemma: **réttur**, an adjective in neuter, singular, nominative, strong declension, positive

Outline

What should be in the lexicon?

- Why don't we list all word forms in the lexicon? For example, all inflectional endings?
- Many processes are predictable.
  - It is therefore inefficient to list all word forms.
- Many processes are productive
  - It is therefore not possible to list all word forms.
- Thus, we may need to separate the lexicon from the rules for generating words.

The slide is taken from the course: Inngangur að tungutækni,

Eiríkur Rögnvaldsson, HÍ, 2002.

Two-level morphology (í. tveggja laga orðhlutagreining)

Purpose

- Kimmo Koskenniemi, 1983.
  - `http://www.ling.helsinki.fi/~koskenni/doc/Two-LevelMorphology.pdf`
- Links the *surface form* of a word – the word as it is actually in a text – to its *lexical form* (underlying form) – its sequence of morphemes.
- The mapping between the surface form and the lexical form (in both directions) is implemented using a *finite-state transducer* (í. stöðuferjald).

Two-level morphology (í. tveggja laga orðhlutagreining)

Examples of lexical and surface forms in English

`dis+en+tangle+ed`	`disentangled`
`happy+er`	`happier`
`move+ed`	`moved`

Two-level morphology

Correspondence between lexical and surface forms

- Example for Icelandic (0 denotes the empty string)

```
Lexical form: hest+ur    (a morpheme used)
Surface form: hest0ur

Lexical form: hest +n +k +e +n  (morphological features used)
Surface form: hest 0  0  u  r
```

Two-level morphology

A finite-state transducer

- A finite-state automaton recognising or generating **a pair of strings**.
- Arcs are labeled with two symbols: the first is the input, the second is the output
- The machine transduces the input symbol into the output symbol as a transition occurs on the arc.
- See Fig. 5.6. page 132 in the textbook.

A finite-state transducer (FST)

A mathematical definition

A FST consists of five components $(Q, \Sigma, q_0, F, \delta)$:

1. Q is a finite set of states, $q_0, q_1 \ldots q_n$.
2. Σ is a finite set of symbol pairs $i : o$, i is from the input alphabet, o is from the output alphabet.
3. q_0 is the start state, $q_0 \in Q$.
4. F is a set of final states, $F \subseteq Q$.
5. δ is the transition function $Q \times \Sigma \rightarrow Q$, $\delta(q, i, o)$ returns the state where the automaton moves when it is in state q and consumes the input symbol pair $i : o$.

A finite-state transducer: Example

hestur

Singular					Plural				
Lexical:	hest	+n	+k	+e	+n	hest	+n	+k	+f	+n
Surface:	hest	0	0	u	r	hest	0	0	a	r
Lexical:	hest	+n	+k	+e	+o	hest	+n	+k	+f	+o
Surface:	hest	0	0	0	0	hest	0	0	0	a
Lexical:	hest	+n	+k	+e	+þ	hest	+n	+k	+f	+þ
Surface:	hest	0	0	0	i	hest	0	0	u	m
Lexical:	hest	+n	+k	+e	+e	hest	+n	+k	+f	+e
Surface:	hest	0	0	0	s	hest	0	0	0	a