T-(538|725)-MALV, Málvinnsla
Orðflokkar og orðhlutafræði

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]Tölvunarfræðideild, Háskólinn í Reykjavík

September 2007

Outline

Outline

Orðflokkar

Skilgreining

- Orðflokkur er flokkur orða sem hafa sameiginleg málfræðileg einkenni:
  - Merkingarleg einkenni.
  - Beygingarleg einkenni.
  - Setningarleg einkenni.
  - `http://is.wikibooks.org/wiki/Or%C3%B0flokkagreining`
- e. Part-of-speech, e. lexical categories.

Yfirflokkun

- Lokaður flokkur (e. closed class).
- Opinn flokkur (e. open class).

Orðflokkar

Skilgreining

- Orðflokkur er flokkur orða sem hafa sameiginleg málfræðileg einkenni:
  - Merkingarleg einkenni.
  - Beygingarleg einkenni.
  - Setningarleg einkenni.
  - `http://is.wikibooks.org/wiki/Or%C3%B0flokkagreining`
- e. Part-of-speech, e. lexical categories.

Yfirflokkun

- Lokaður flokkur (e. closed class).
- Opinn flokkur (e. open class).

Orðflokkar

Lokaður flokkur

- Ný orð bætast ekki við þennan flokk.
- Kerfisorð.
- Forsetningar, samtengingar, hjálparsagnir, samtengingar, atviksorð (nema háttaratviksorð).

Opinn flokkur

- Getur bætt við sig nýjum orðum.
- Orð geta einnig tapast!
- Inntaksorð: gegna ákveðnu merkingarhlutverki.
- Nafnorð, lýsingarorð, sagnir (nema hjálparsagnir), háttaratviksorð.

Orðflokkar

Lokaður flokkur

- Ný orð bætast ekki við þennan flokk.
- Kerfisorð.
- Forsetningar, samtengingar, hjálparsagnir, samtengingar, atviksorð (nema háttaratviksorð).

Opinn flokkur

- Getur bætt við sig nýjum orðum.
- Orð geta einnig tapast!
- Inntaksorð: gegna ákveðnu merkingarhlutverki.
- Nafnorð, lýsingarorð, sagnir (nema hjálparsagnir), háttaratviksorð.

11 orðflokkar í íslensku

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

Fallorð fallbeygjast. Sagnorð tíðbeygjast. Smáorð beygjast ekki.

Fallorð

Beygingarleg einkenni (e. features)
(feitletrun vísar í skammstafanir í íslenska markamenginu)

- 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)

Sagnorð

Beygingarleg einkenni (e. features)

- Háttur (e. mood) `http://is.wikipedia.org/wiki/H%C3%A6ttir_sagna`
  - **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) `http://is.wikipedia.org/wiki/Sagnmyndir`
  - **g**ermynd (e. active)
  - **m**iðmynd (e. middle)
  - þolmynd (e. passive) - mynduð með lýsingarhætti þátíðar
- Persóna (e. person): **1**., **2**., **3**.
- Tala (e. number): **e**intala, **f**leirtala
- Tíð (e. tense)
  - **n**útíð (e. present)
  - **þ**átíð (e. past)

Outline

Orðasafn (e. lexicon)

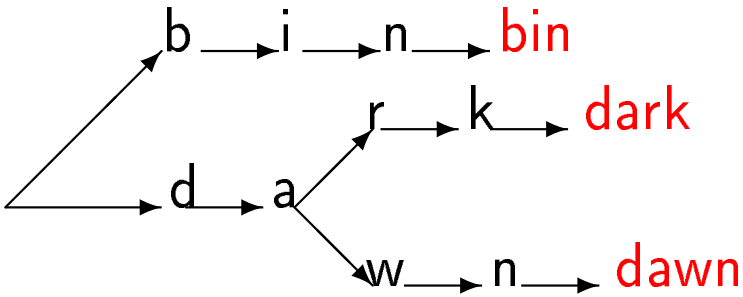
Geymir hvað?

- Listi af (lesmáls)orðum með eða án frekar upplýsinga.
  - Beygingarupplýsingar.
  - Mörk.
  - Setningafræðilegar upplýsingar.
- Oft búið til fyrir tiltekið svið, t.d. tækni, vísindi, fjármál.
- Yfirleitt það fyrsta sem búa þarf til fyrir máltæknikerfi.

Orðasafn

Gagnaskipan

- Tætitafla (lykill–gildi): sérhvert orð er lykill og frekari upplýsingar um orðið er gildi þess.
  - Getur tekið mikið minni fyrir stór orðasöfn.
- *Letter trees* (tries) `http://en.wikipedia.org/wiki/Trie`
  - Orð eru geymd sem stafatré (e. tree of characters) sem deila kvíslum eins lengi og einstakir stafir orða eru þeir sömu.



Outline

Orðhlutafræði

Myndan (e. morpheme)

- Myndan (morfem) er minnsta afmörkuð eind í byggingu orðsins.
- Afmarkaður hluti úr orði sem hefur einhverja ákveðna merkingu eða hlutverk.
- Myndön skiptast í stofn (e. stem) og aðskeyti (e. affix) eða beygingarendingu.
- Stofn er sú mynd orðs sem helst eins í öllum beygingarmyndum þess.
- Dæmi: Orðið "orði" samanstendur af tveimur myndönum:
  - "orð" sem er stofn.
  - "i" sem er beygingarending.

Orðhlutafræði

Aðskeyti

- Forskeyti (e. prefix)
  - Fer á undan stofni.
  - Dæmi: örfínn
- Viðskeyti (e. suffix)
  - Kemur á eftir stofni.
  - Dæmi: kennari
- Innskeyti (e. infix)
  - Er sett inn í stofn.
  - Dæmi: humingi í tagalog; hingi: að lána, humingi: sá sem fær lánað
- Umskeyti (e. circumfix)
  - Bæði fer á undan og kemur á eftir stofni.
  - Dæmi: gesagt í þýsku (lhþt. af sagen).

Orðhlutafræði

- e. morphology

  `http://en.wikipedia.org/wiki/Morphology_(linguistics)`
- Orðhlutafræði – Orðmyndunarfræði
- Fjallar um hvernig orð eru sett saman úr myndönum.
- Tvenns konar orðhlutakerfi:
  - Concatenative morphology
    - T.d. Germönsk mál
    - `http://en.wikipedia.org/wiki/Germanic_languages`
    - Forskeyti Stofn Viðskeyti
  - Non-concatenative (templatic) morphology

    `http://en.wikipedia.org/wiki/Nonconcatenative_morphology`
    - T.d. Arabíska, Hebreska

Orðmyndun í íslensku

Orð eru mynduð á tvennan hátt úr myndönum:

Með beygingu (e. inflection)

- Forskeyti + Stofn + beygingarending
- Nýja orðið tilheyrir sama orðflokki og stofninn.
- Dæmi: hest<u>ur</u>, málvinnsl<u>a</u>

Með afleiðslu (e. derivation)

- Forskeyti + Stofn + viðskeyti
- Nýja orðið getur tilheyrt öðrum orðflokki en stofninn.
- Dæmi: <u>ó</u>ljós, klofn<u>ingur</u>, kenn<u>ari</u>

Orðmyndun í íslensku

Orð eru mynduð á tvennan hátt úr myndönum:

Með beygingu (e. inflection)

- Forskeyti + Stofn + beygingarending
- Nýja orðið tilheyrir sama orðflokki og stofninn.
- Dæmi: hest<u>ur</u>, málvinnsl<u>a</u>

Með afleiðslu (e. derivation)

- Forskeyti + Stofn + viðskeyti
- Nýja orðið getur tilheyrt öðrum orðflokki en stofninn.
- Dæmi: <u>ó</u>ljós, klofn<u>ingur</u>, kenn<u>ari</u>

Orðmyndun í íslensku

Samsett orð (e. compound words)

- Tvö (eða fleiri) orð sett saman til að mynda nýtt orð.
- Algengt að setja saman tvö nafnorð.
- `http://rettritun.is/?id=namsefni_k9_r1`

Orðhlutafræði

Orðhlutafræðileg greining/þáttun

- e. morphological analysis/parsing
- Sú aðgerð að brjóta orð upp í einstök myndön.
- Nauðsynleg í mörkum máltæknikerfum:
  - "Lemmatization". Það að finna uppflettimynd (lemmu, flettu) orðsins (e. dictionary form).
  - "Stemming". Það að finna stofn, notað t.d. í upplýsingaheimt.
  - Orð sem ekki finnast í orðasafni þurfa greiningu (t.d. í mörkun).

Orðhlutafræðileg myndun

- Að mynda orð að gefnum myndönum.

Orðhlutafræði

Orðhlutafræðileg greining/þáttun

- e. morphological analysis/parsing
- Sú aðgerð að brjóta orð upp í einstök myndön.
- Nauðsynleg í mörkum máltæknikerfum:
  - "Lemmatization". Það að finna uppflettimynd (lemmu, flettu) orðsins (e. dictionary form).
  - "Stemming". Það að finna stofn, notað t.d. í upplýsingaheimt.
  - Orð sem ekki finnast í orðasafni þurfa greiningu (t.d. í mörkun).

Orðhlutafræðileg myndun

- Að mynda orð að gefnum myndönum.

Lemmatization

Lemmatization vs. Stemming

- Dæmi: “hesti”. Lemma = “hestur”. Stofn = “hest”

Margræðni í lemmun

1. A **run** in the forest. Lemma: **run**, nafnorð í eintölu.
2. The sportsmen **ran** everyday. Lemma: **run**, sögn í þátíð, þriðju persónu, fleirtölu.
3. **Rétturinn** var fullskipaður. Lemma: **réttur**, nafnorð í karlkyni, eintölu, nefnifalli, með greini.
4. Dæmið var **rétt**. Lemma: **réttur**, lýsingarorð í hvorugkyni, eintölu, nefnifalli, sterk beyging, frumstig.

Lemmatization

Lemmatization vs. Stemming

- Dæmi: “hesti”. Lemma = “hestur”. Stofn = “hest”

Margræðni í lemmun

1. A **run** in the forest. Lemma: **run**, nafnorð í eintölu.
2. The sportsmen **ran** everyday. Lemma: **run**, sögn í þátíð, þriðju persónu, fleirtölu.
3. **Rétturinn** var fullskipaður. Lemma: **réttur**, nafnorð í karlkyni, eintölu, nefnifalli, með greini.
4. Dæmið var **rétt**. Lemma: **réttur**, lýsingarorð í hvorugkyni, eintölu, nefnifalli, sterk beyging, frumstig.

Outline

Hvað á að vera í orðasafninu?

- Því ekki hafa allar orðmyndir í orðasafni?
  - T.d. fall- og töluendingar?
- Mörg ferli eru fyrirsegjanleg (e .predictable)
  - Þess vegna er óhagkvæmt að telja upp myndir.
- Mörg ferli eru frjó (e. productive)
  - Þess vegna útilokað að telja upp myndir.
- Þess vegna er betra að hafa aðskilin söfn og reglur sem raða orðhlutum saman.

Glæra fengin að láni úr námskeiðinu: Inngangur að tungutækni,

Eiríkur Rögnvaldsson, HÍ, 2002.

Tveggja laga orðhlutagreining

Tilgangur

- e. two-level morphology (Kimmo Koskenniemi 1983).
- Tengir saman "surface form" orðs – orðið eins og það kemur fyrir í texta – við "lexical form" (underlying form) – þ.e. röð myndana orðsins.
- Vörpunin á milli "surface form" og "lexical form" á sér stað með *stöðuferjaldi* í báðar áttir.

Dæmi (0 merkir tómi strengurinn)

```
Lexical form: hest+ur   (myndan notað)
Surface form: hest0ur
Lexical form: hest +n +k +e +n  (beygingarleg einkenni notuð)
Surface form: hest 0  0  u  r
```

Tveggja laga orðhlutagreining

Tilgangur

- e. two-level morphology (Kimmo Koskenniemi 1983).
- Tengir saman "surface form" orðs – orðið eins og það kemur fyrir í texta – við "lexical form" (underlying form) – þ.e. röð myndana orðsins.
- Vörpunin á milli "surface form" og "lexical form" á sér stað með *stöðuferjaldi* í báðar áttir.

Dæmi (0 merkir tómi strengurinn)

```
Lexical form: hest+ur   (myndan notað)
Surface form: hest0ur
Lexical form: hest +n +k +e +n  (beygingarleg einkenni notuð)
Surface form: hest 0  0  u  r
```

Tveggja laga orðhlutagreining

Stöðuferjald

- Endanleg stöðuvél sem ber kennsl á eða myndar **par af strengjum**.
- Leggir merktir með tveimur stöfum, sá fyrsti er inntaksstafur, sá seinni er úttaksstafur.
- Vélin breytir (e. transduces) inntaksstaf í úttaksstaf þegar færsla á sér stað eftir legg.
- Sjá Fig. 5.6. bls. 132 í kennslubók.

Stöðuferjald (e. finite-state transducer)

Stærðfræðileg skilgreining

Stöðuferjald samanstendur af fimm hlutum $(Q, \Sigma, q_0, F, \delta)$:

1. Q er mengi af endanlegum stöðum, q_0, q_1 ... q_n.
2. Σ er endanlegt mengi af inntakspörum $i : o$, i er úr inntaksstafrófi, o úr úttaksstafrófi.
3. q_0 er upphafsstaða, $q_0 \in Q$.
4. F er mengi lokastaða, $F \subseteq Q$.
5. δ er breytingafall (e. transition function) $Q \times \Sigma \rightarrow Q$, $\delta(q, i, o)$ skilar þeirri stöðu sem vélin fer í úr stöðu q miðað við inntakspari $i : o$.

Stöðuferjald: Dæmi

hestur

Eintala		Fleirtala	
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

Prolog útfærsla

- Prolog: Sjá Appendix A í kennslubók.
- SWI-Prolog: `http://www.swi-prolog.org/`
- Útfærslan á stöðuferjaldinu (Prolog-kóði): Sjá undir "Annað efni".
- Lesa forritskóða inn í Prolog: consult('transduce.pro').
- Prófa ferjaldið: transduce(1, Final, Lexical, [h,e,s,t,u,r]).
- Prófa ferjaldið: transduce(1, Final, [h,e,s,t,+n,+k,+e,+n], Surface).