# Málvinnsla: Forritunarverkefni III - Hlutaþáttun

Háskólinn í Reykjavík - Tölvunarfræðideild

Október 2007

## 1 Lýsing

Í þessu verkefni eigið þið að búa til lagskiptan (e. cascading/incremental) hlutaþáttara fyrir íslenskan texta. Þið eigið að nota JFlex (eða sambærilegt tól) til að forrita sérhvert stöðuferjald (e. transducer). Úttakið úr fyrsta stöðuferjaldinu er inntak í stöðuferjald númer tvö, og svo koll af kolli. Inntakið í fyrsta stöðuferjaldið er markaður texti og úttakið úr síðasta stöðuferjaldinu er upphaflegi markaði textinn ásamt ýmiss konar merkjum sem sýna klumpa (e. chunks). Þessum merkjum er lýst í næstu köflum.

### 1.1 Fleiryrtar segðir (e. multiword expressions (MWE))

Fyrsta stöðuferjaldið, *Chunk_MWE.flex*, á að merkja eftirfarandi fleiryrtar segðir sem samanstanda af atviksorði og forsetningu:

- á eftir, á milli
- niður á
- út í
- yfir að
- þrátt fyrir

Fyrir sérhverja fleiryrta segð sem stöðuferjaldið finnur á það að setja [MWE_PP ... MWE_PP] utan um segðina (PP stendur fyrir "preposition phrase"), t.d.:

```
[MWE_PP þrátt aa fyrir ao MWE_PP]
```

## 1.2 Sagnorðsklumpar (e. verb chunks)

Þetta ferjald, *Chunk_ Verb.flex*, á að merkja tiltekna sagnorðsklumpa með [VP ... VP] (VP stendur fyrir "verb phrase"):

- Sögn í persónuhætti (framsöguhætti, viðtengingarhætti eða boðhætti). Dæmi:

```
[VP stökk sfg1eþ VP]
```

- Rununa sögn í persónuhætti og sagnbót. Dæmi:

```
[VP hafði sfg3eþ sofið ssg VP]
```

- Rununa sögn í persónuhætti, atviksorð og sagnbót. Dæmi:

```
[VP hefði svg3eþ ekki aa séð ssg VP]
```

- Rununa nafnháttarmerki og sögn í nafnhætti. Dæmi:

```
[VP að cn strjúka sng VP]
```

## 1.3 Nafnorðsklumpar (e. noun chunks)

Þetta ferjald, *Chunk_ Noun.flex*, skal merkja tiltekna nafnorðsklumpa með [NP ... NP] (NP stendur fyrir "noun phrase"):

- Rununa atviksorð (valfrjálst), lýsingarorð (valfrjálst), nafnorð. Dæmi:

```
[NP kirkjugarðinn nkeog  NP]
[NP strjála lveosf byggð nveo  NP]
[NP mjög aa ánægjuleg lvensf ferð nven  NP]
```

- Rununa óákveðinn greinir eða atviksorð, lýsingarorð, nafnorð (valfrjálst). Dæmi:

```
[NP hinn gken stóri lkensf strákur nken NP]
[NP verulega aa óþekkur lkensf NP]
```

- Stakt persónufornafn eða óákveðið fornafn. Dæmi:

```
[NP ég fp1en NP]
[NP enginn foken NP]
```

### 1.4 Forsetningaklumpar (e. preposition chunks)

Þetta ferjald, *Chunk_ Prep.flex*, skal merkja tiltekna forsetningaklumpa með [PP ... PP]:

- Rununa forsetning og nafnorðsklumpur. Dæmi:

  ```
  [PP með aþ [NP auglýsingum nvfþ NP] PP]
  ```

- Rununa forsetning og fleiyrt segð. Dæmi:

  ```
  [PP [MWE_PP niður aa á ao MWE_PP] [NP strjála lveosf byggð nveo NP] PP]
  ```

### 1.5 Hreinsun

Þetta ferjald, *Chunk_ Clean.flex*, skal eingöngu hreinsa auka bil út úr úttakinu sem *Chunk_ Prep.flex* skilar.

### 1.6 Allt sett saman

Þið þurfið að skrifa eina .bat (skel) skrá sem heitir *chunk.bat* og tekur inn eitt viðfang, skrána sem á að "klumpa". Þess konar skrá, *Chunk_ Test*, er aðgengileg á vef námskeiðsins. Þessi skel á að kalla á stöðuferjöldin í þeirri röð sem þau eru talin upp að ofan og skila í lokin skrá sem ber sama heiti og inntaksskráin nema með ".out" skeytt við nafnið. Sú skrá inniheldur upphaflegu setningarnar með klumpum. Dæmi:

```
chunk ChunkTest.txt
...
býr til skrá með nafnið ChunkTest.txt.out
```

Sýnishorn af *ChunkTest.txt.out* má sjá á vef námskeiðsins.

## 2 Skilafrestur

Forritskóða (allar .flex skrár + .bat skrá), ásamt úttaki fyrir skrá *ChunkTest.txt* (þ.e. *ChunkTest.txt.out*), skal skila í síðasta lagi þriðjudaginn 16. október, kl. 23:59.