Natural Language Processing: Programming Project III Partial Parsing

Reykjavik University – School of Computer Science

November 2009

1 Description

In this project you will develop an *incremental finite-state parser* for Icelandic text. Use JFlex (or a comparable tool) to program each transducer. The output of one transducer is used as the input to the next transducer in the sequence. The input to the first transducer is POS tagged text[1] and the output of the last transducer is the original text annotated with various labels denoting chunks. These labels and the corresponding transducers are described in the following sections.

1.1 Multiword Expressions (MWE)

The first transducer, *Chunk_MWE.flex*, labels the following MWE pairs, consiting of an andverb and a preposition.

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

For each MWE found the transducer puts [MWE_PP . . . MWE_PP] around the expression, e.g.:

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

[1] Using the Icelandic Frequency Dictionary tagset.

1.2 Verb chunks

This transducer, *Chunk_ Verb.flex*, labels specific verb chunks with [VP ... VP].

- A finite verb (in indicative, subjunctive or imperative mood) (sögn í persónuhætti (framsöguhætti, viðtengingarhætti eða boðhætti)). Example:

  ```
  [VP stökk sfg1eþ VP]
  ```

- The sequence: a finite verb, an adverb (optinal), and a supine verb (sögn í persónuhætti og sagnbót). Example:

  ```
  [VP hafði sfg3eþ sofið ssg VP]
  [VP hefði svg3eþ ekki aa séð ssg VP]
  ```

- The sequence: the infinitive marker and an infinitive verb (nafnháttarmerki og sögn í nafnhætti). Example:

  ```
  [VP að cn strjúka sng VP]
  ```

1.3 Noun chunks

This transducer, *Chunk_Noun.flex*, labels specific noun chunks with [NP ... NP].

- The sequence an adverb (optional), an adjective (optional), a noun. Example:

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

- The sequence an indefinite article (óákveðinn greinir) or an adverb, an adjective, a noun (optional). Example:

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

- A personal pronoun or an indefinite pronoun (óákveðið fornafn). Example:

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

1.4 Preposition chunks

This transducer, *Chunk_ Prep.flex*, labels preposition chunks with [PP . . . PP].

- The sequence a preposition and a noun chunk. Example:

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

- The sequence a MWE_PP chunk and a noun chunk. Example:

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

1.5 Cleaning

This transducer, *Chunk_ Clean.flex*, removes multiple occurrences of white space from the output generated by any of the preceeding transducers.

1.6 Everything put together

You need to write a shell with the name *chunk.bat* or *chunk.sh* which accepts one command line parameter, i.e. the file which should be parsed. The input file, *ChunkTest.txt*, is attached to the description of the project in MySchool. Your shell should call the transducer in the exact order as described above and return a file with the same name as the input file name but with ".out" appended. The output file thus contains the original input sentences annotated with chunks. Example:

```
chunk ChunkTest.txt
...
```

generates the file ChunkTest.txt.out

2 What to hand in

All program code (all .flex files and shell files), along with the output file *ChunkTest.txt.out*.