

T-(538|725)-MALV, Natural Language Processing lexc - **Lexicon Compiler**

Hrafn Loftsson¹ Hannes Högni Vilhjálmsson¹

¹School of Computer Science, Reykjavik University

October 2010

1 Lexicon compiler

2 hfst-lexc

1 Lexicon compiler

2 hfst-lexc

What is it?

- A program that reads sets of morphemes and their morphotactics combinations (syntax of morphemes) in order to create a finite-state transducer of a lexicon.
- A language (and a compiler) for defining automata and transducers.
- Specifically suitable for defining lexicons.
- Based on two-level morphology.

Examples

- A part of **XFST** (Xerox Finite-State tools).
- A part of **HFST** (Helsinki Finite-state tools)
 - <http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/>
 - Free and open source:
<http://sourceforge.net/projects/hfst/files/>

1 Lexicon compiler

2 hfst-lexc

The format of a hfst-lexc file

- Multichar_Symbols declaration
- Declarations section
- Lexicon Root
- Lexicon X
- Lexicon Y
- ...
- END

A simple example

```
! engSimple.lexc (this line is a comment)
LEXICON Root
dog      # ;
cat      # ;
bird     # ;
END
```

Compile into a finite-state transducer

```
hfst-lexc -o engSimple.hfst engSimple.lexc
```


Work with the transducer

- Display the recognized strings:

```
hfst-fst2strings engSimple.hfst
```

- Lookup a lexical form:

```
echo ‘‘dog’’ | hfst-lookup engSimple.hfst
```

lexc – Continuation classes

LEXICON Root

walk # ;

walks # ;

walked # ;

walking # ;

talk # ;

talks # ;

talked # ;

talking # ;

pack # ;

packs # ;

packed # ;

packing # ;

lexc – Continuation classes

```
LEXICON Root
```

```
walk    V ;
```

```
talk    V ;
```

```
pack    V ;
```

```
LEXICON V
```

```
s      # ;
```

```
ed     # ;
```

```
ing    # ;
```

```
      #;  ! <- an empty-string entry
```

Lexical form to Surface form

LEXICON Root

swim:swam # ;

fi0ght:fought # ;

Compile and run

```
hfst-lexc -o engSimple3.hfst engSimple3.lexc
```

```
echo "swim" | hfst-lookup engSimple3.hfst
```

```
echo "fight" | hfst-lookup engSimple3.hfst
```

Surface form to lexical form

```
LEXICON Root
swim:swam      # ;
fi0ght:fought # ;
```

Compile and run

```
hfst-lexc -o engSimple3.hfst engSimple3.lexc
hfst-invert -i engSimple3.hfst -o engSimple3Invert.hfst
echo "swam" | hfst-lookup engSimple3Invert.hfst
echo "fought" | hfst-lookup engSimple3Invert.hfst
```

An Icelandic example

See example under "Other material" on course web page.