



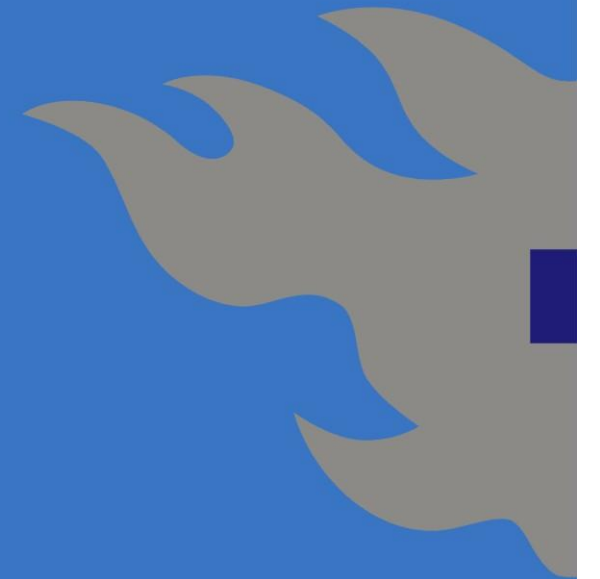
HELSINGIN YLIOPISTO  
HELSINGFORS UNIVERSITET  
UNIVERSITY OF HELSINKI

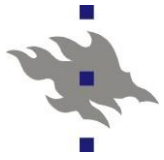
# OMorFi

Open Source Lexicons and Applications

Krister Lindén / Language Technology

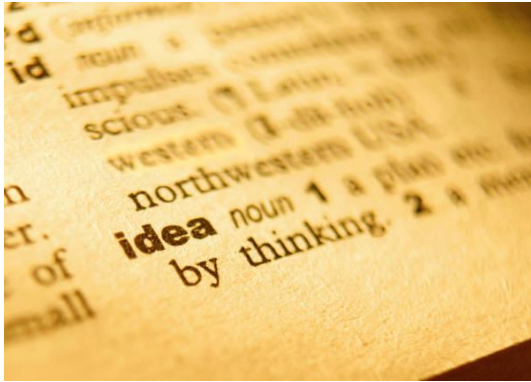
April 2, 2008



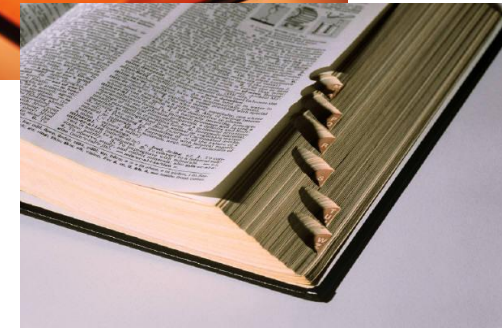


# Usage Environment for Lexical Data

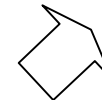
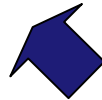
OMorFi  
(what)

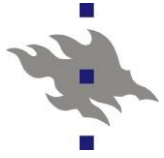


Agents and Applications  
(how & why)



External world  
(where)





## Lexicon Producers and Applications

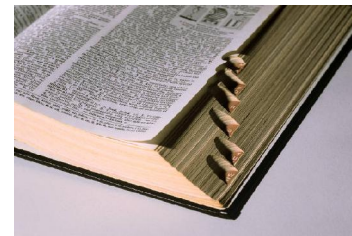
### ■ Production (how)

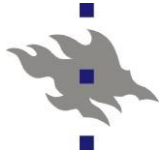
- KOTUS – Research Institute for the Languages of Finland
- Voikko / COSS, Ispell / Gnu
- Publishing Companies (WSOY, ...)
- Translation Companies
- Organization-Specific Terminologies (Semantic Web)
- Language Industry (Lingsoft, Connexor, Kielikone, BitLips, ...)



### ■ Usage (why)

- Spelling
- Search
- Translation
- Language Modeling, ...





## The External World Offers Usage Examples

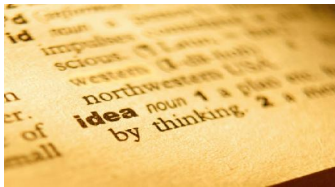
- Material and Corpora (where)
  - Language Bank / CSC
  - CLARIN / EU
  - Internet / WWW
  - Speech Archive of YLE
  - ...



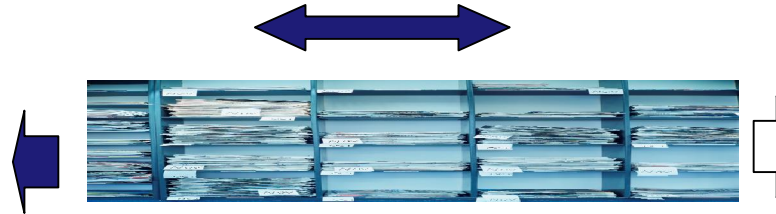


## Traditionally the Lexicons in LT applications have also been hand-made

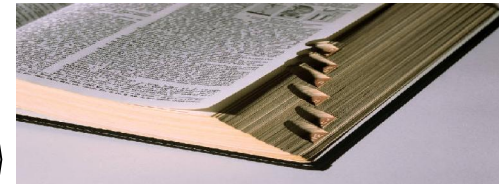
What



By whom



Why



morphological analysis:  
word list + inflections

lexicographer

spelling / Lingsoft

syntactic analysis:  
word classes + rules

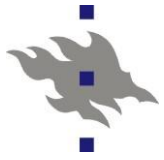
linguist

syntax / Connexor

translation dictionary:  
translation + context

translator

reading assistance /  
Kielikone

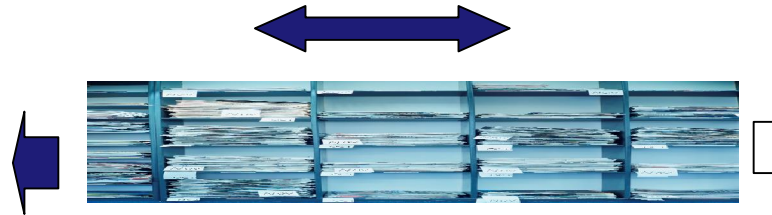


# Current Aims are in Processing and Extending Existing Open Sources

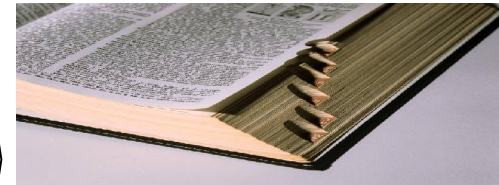
What



How / By whom



Why

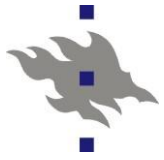


- |   |              |                      |                          |
|---|--------------|----------------------|--------------------------|
| 1 | Morphology   | ← automatically      | (new) word list / Kotus  |
| 2 | Morphology   | ↔ automatically      | spelling / Voikko        |
| 3 | Morphosyntax | ← semi-automatically | morphosyntactic analysis |

...

April 2, 2008

Krister Lindén / Language Technology

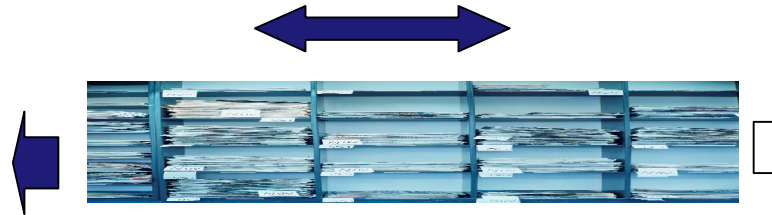


## Curent Aims (cont.)

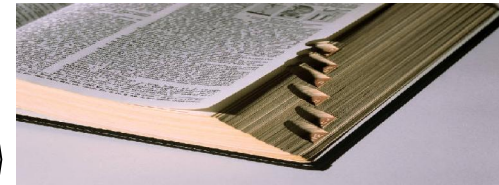
What



How / By whom

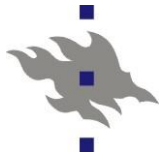


Why

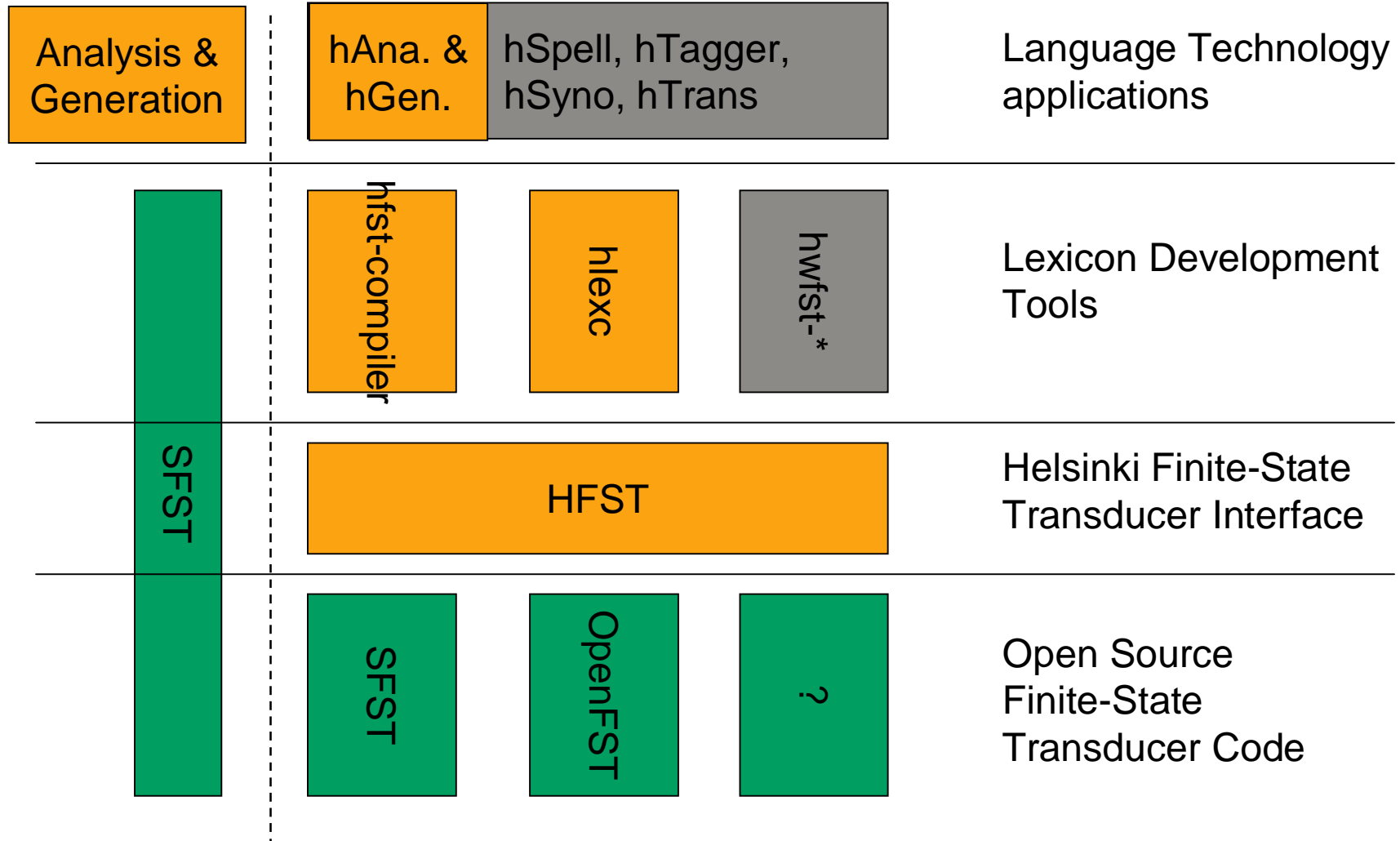


n	Translation Lexicon	semi-automatically	translation assistance
n+1	Thesaurus	semi-automatically	search expansion
n+2	Valency Lexicon	← semi-automatically	syntactic analysis

...

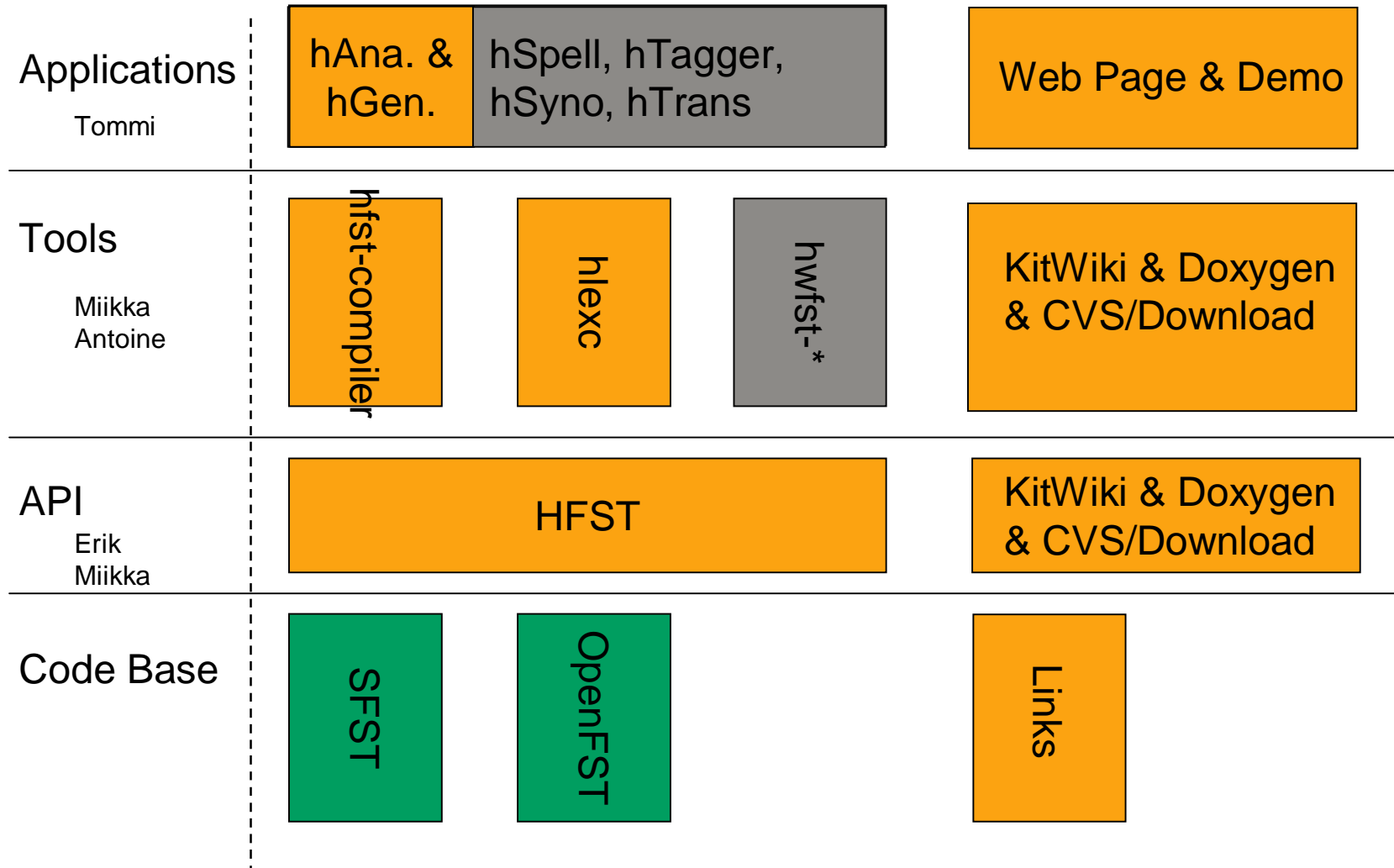


# Open Source Lexicon Development Tools and Lexicon-based Tools





## Code and Documentation





## HFST structure

Tools

- hfst-compiler.C:
  - include HFST.h
  - ...
  - using HFST::

■ Miikka

- hlexc.C:
  - include HFST.h
  - ...
  - using HFST::

■ Antoine

API

- HFST.C:
  - SFST objects
  - Intersecting Composition  
using SFST

■ Miikka

- HFST.C:
  - Objects using OpenFST
  - Character Pair Alphabet
  - Intersecting Composition  
using OpenFST objects

■ Erik