Ruth Vatvedt Fjeld & Rune Lain Knudsen

LBK2013 – a lexicographic corpus for modern Norwegian bokmål
Purpose:

• **Lemma selection**

  *Frequency based lemma selection*
  
  Neologisms/obsolete words
  
  – **Singleword lemmas**
    
    » *Mus (mouse)* (meaning change)
    
    » *Tastafon* obsolete words
  
  – **Multiword lemmas**
    
    » *Være lutter øre (be all ears)* obsolete
LBK – Lexciographical Bokmål Corpus

• The documents in LBK2013 is restricted to the timespan of 1985-2013.
  – Availability
  – Modern language
  – Changes in lexicon related to existing dictionaries (built from excerpts of old language)
  – a balanced corpus of 100 mill. tokens
Selection of text types

- Modern fiction
- Text books
- Blogs
- Factual prose
- Law
- Medicin
- Natural sciences
- Humaniora
- Sports ...
Demography markers

- Age
- Sex
- Place of birth and youth
- Year of birth
- Publisher
- Year of publication
- Such metadata makes it easy to construct subcorpora for comparative investigations and a wide range of queries
Text categories LBK2013

- Skjønnlitteratur: 35%
- Sakprosa: 5%
- Unormert: 5%
- Aviser og kulørte ukeblader: 6%
- TV-tekster: 49%
How?

LBK makes use of the IMS Corpus Workbench, a widely used tool set for managing and querying large text corpora. It is made available for researchers through Glossa, a web based interface for corpora developed at the Text Laboratory, ILN at the University of Oslo. Every document is POS-tagged with the Oslo-Bergen tagger. Additional metadata such as bibliographic and ethnographic information is manually annotated and stored as TEI headers.
Staff

![Staff Chart]

- **01.01.2000**
  - Project Leader: 0
  - Assistant: 0
  - Engineer: 0

- **01.01.2001**
  - Project Leader: 0
  - Assistant: 0
  - Engineer: 0

- **01.01.2002**
  - Project Leader: 0
  - Assistant: 0
  - Engineer: 0

- **01.01.2003**
  - Project Leader: 0
  - Assistant: 0
  - Engineer: 0

- **01.01.2004**
  - Project Leader: 0
  - Assistant: 0
  - Engineer: 0

- **01.01.2005**
  - Project Leader: 0
  - Assistant: 0
  - Engineer: 0

- **01.01.2006**
  - Project Leader: 0
  - Assistant: 0
  - Engineer: 0

- **01.01.2007**
  - Project Leader: 0
  - Assistant: 0
  - Engineer: 0

- **01.01.2008**
  - Project Leader: 0
  - Assistant: 0
  - Engineer: 0

- **01.01.2009**
  - Project Leader: 0
  - Assistant: 0
  - Engineer: 0

- **01.01.2010**
  - Project Leader: 0
  - Assistant: 0
  - Engineer: 0

- **01.01.2011**
  - Project Leader: 0
  - Assistant: 0
  - Engineer: 0

- **01.01.2012**
  - Project Leader: 0
  - Assistant: 0
  - Engineer: 0

- **01.01.2013**
  - Project Leader: 0
  - Assistant: 0
  - Engineer: 0

Legend:
- **PROJECT LEADER**
- **ASSISTANT**
- **ENGINEER**
New statistical tools

• Frequency counts
• Concordances
• DeepDict analysis (Bick)
• Word Sketch Engine (Kilgarriff)
Why compile a balanced corpus

Statistical analysis of interesting subcorpora for

– Actual use of recommended morphology
  • (standardisation and documentation)

<table>
<thead>
<tr>
<th>wordform</th>
<th>TV-text</th>
<th>Total korpus</th>
<th>NoTa</th>
</tr>
</thead>
<tbody>
<tr>
<td>tiden/tida (time)</td>
<td>72/28</td>
<td>92/8</td>
<td>60/40</td>
</tr>
<tr>
<td>takken/takka (thanks)</td>
<td>100/0</td>
<td>100/0</td>
<td>-</td>
</tr>
<tr>
<td>hjelpe/hjelpa (help)</td>
<td>91/9</td>
<td>95/5</td>
<td>50/50</td>
</tr>
<tr>
<td>lysten/lysta (desire)</td>
<td>100/0</td>
<td>100/0</td>
<td>100/0</td>
</tr>
<tr>
<td>moren/mora (mother)</td>
<td>81/19</td>
<td>91/9</td>
<td>79/11</td>
</tr>
<tr>
<td>kvinner/kvinna (woman)</td>
<td>100/0</td>
<td>99/1</td>
<td>100/0</td>
</tr>
<tr>
<td>uken/uka (week)</td>
<td>42/58</td>
<td>63/37</td>
<td>21/79</td>
</tr>
</tbody>
</table>
How to mark up a corpus

- PoS-tagging by automatic analysis
- Grammar: valency/argument structure etc.
  - Jeg har tenkt til å gjøre det (I intend to do it)
  - Flaska knust (the bottle broke)
**Muslim** as first part of composita (1985-2000)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>215</td>
<td>164</td>
</tr>
<tr>
<td>muslim</td>
<td></td>
<td>muslim</td>
<td>muslim</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>176</td>
<td>139</td>
</tr>
<tr>
<td>muslimsk</td>
<td></td>
<td>muslimsk</td>
<td>muslimsk</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>muslimsk-kroatisk</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>muslimsk-dominert</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>muslimsk-dominert</td>
</tr>
</tbody>
</table>

11. april 2011
## Muslim in composita (2001-2013)

<table>
<thead>
<tr>
<th>Year Range</th>
<th>2001-2005</th>
<th>2006-2010</th>
<th>2011-2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001-2005</td>
<td>1073</td>
<td>1217</td>
<td>948</td>
</tr>
<tr>
<td>2006-2010</td>
<td>987</td>
<td>923</td>
<td>499</td>
</tr>
<tr>
<td>2011-2013</td>
<td>14</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>muslim</td>
<td>1217</td>
<td>1073</td>
<td>1217</td>
</tr>
<tr>
<td>muslimsk</td>
<td>923</td>
<td>987</td>
<td>923</td>
</tr>
<tr>
<td>muslimbrødrene</td>
<td>4</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>muslimbror</td>
<td>1</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>muslimskføde</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>muslimsk-arabisk</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>muslimhater</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>muslimskdominert</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>muslimsk-jødisk</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>muslimskdominert</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>muslimskhaten</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>muslimskheter</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>muslimfrykten</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>muslimism</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>muslimhets</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>