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

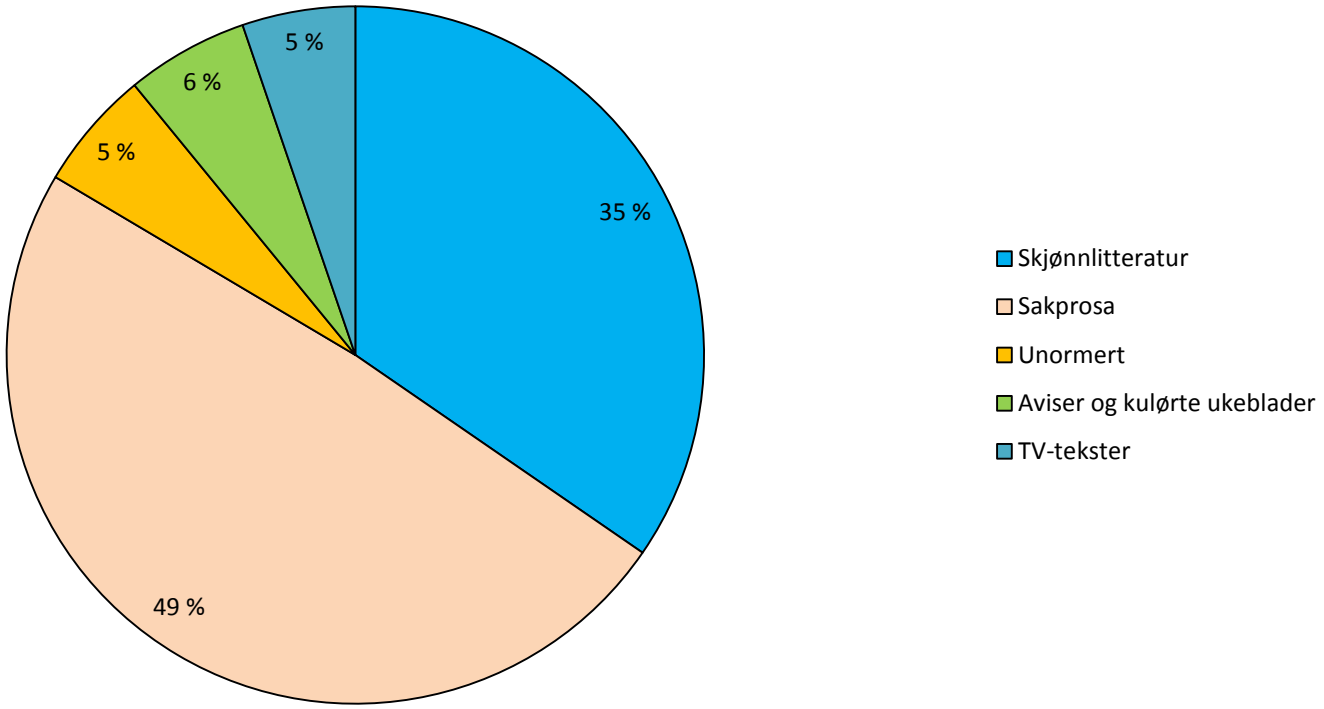
Selection of text types

- Modern fiction
- Text books
- Blogs
- Factual prose
- Law
- Medicin
- Natural siences
- 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

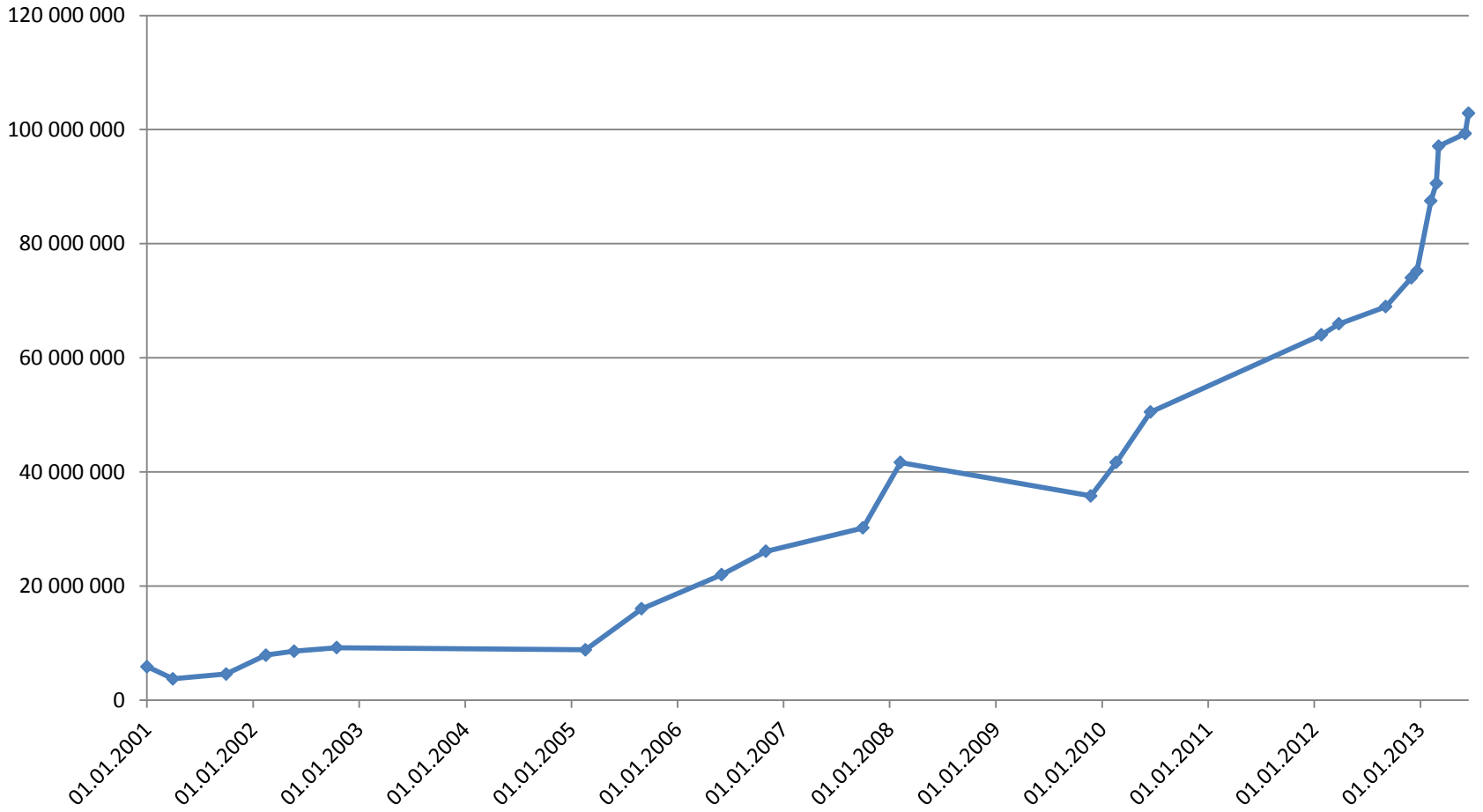
Text categories LBK2013



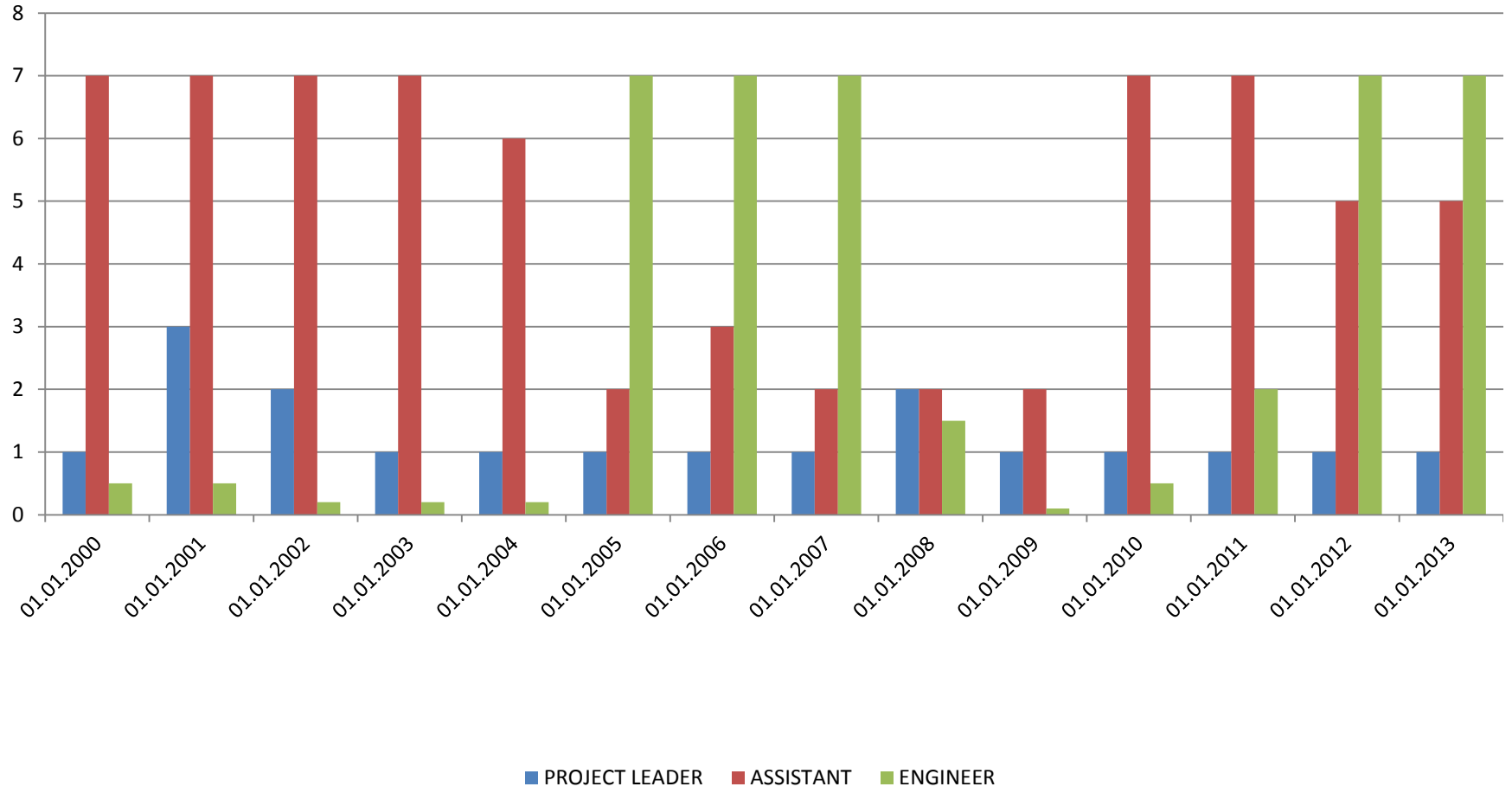
How?

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.

Resources


120 000 000
100 000 000
80 000 000
60 000 000
40 000 000
20 000 000
0
01.01.2001
01.01.2002
01.01.2003
01.01.2004
01.01.2005
01.01.2006
01.01.2007
01.01.2008
01.01.2009
01.01.2010
01.01.2011
01.01.2012
01.01.2013

Staff


8
7
6
5
4
3
2
1
0
01.01.2000
01.01.2001
01.01.2002
01.01.2003
01.01.2004
01.01.2005
01.01.2006
01.01.2007
01.01.2008
01.01.2009
01.01.2010
01.01.2011
01.01.2012
01.01.2013
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)**

wordform	TV-text	Total korpus	NoTa
tiden/tida (time)	72/28	92/8	60/40
takken/takka (thanks)	100/0	100/0	-
hjelpen/hjelpa (help)	91/9	95/5	50/50
lysten/lysta (desire)	100/0	100/0	100/0
moren/mora (mother)	81/19	91/9	79/11
kvinnen/kvinna (woman)	100/0	99/1	100/0
uken/uka (week)	42/58	63/37	21/79

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 knuste** (the bottle broke)

Muslim as first part of composita(1985-2000)

1985-1990		1991-1995		1996-2000	
5	muslim	215	muslim	164	muslim
1	muslimsk	176	muslimsk	139	muslimsk
		8	muslimsk-kroatisk	2	muslimsk-kroatiske
		6	muslimsk-dominert	2	muslimske
				2	muslimsk-dominert

Muslim in composita (2001-2013)

2001-2005		2006-2010		2011-2013	
1073	muslim	1217	muslim	948	muslim
987	muslimsk	923	muslimsk	499	muslimsk
14	muslimbrødrene	4	muslimene	3	muslimhat
5	muslimbror	1	muslimhets	2	muslimskdominert
3	muslimskføde	1	muslimsirkel	1	muslimhater
2	muslimsk-arabisk	1	muslimdominert	1	muslimhatende
1	muslimhater	1	muslimskhet	1	muslimvennlig
1	muslimsk-jødisk	1	muslimfrykten	1	muslimisme
1	muslimskdominert	1	muslimdebatt	1	muslimhets