NB N-gram

Magnus Breder Birkenes
Språkbanken, National Library of Norway

National Library of Norway

NB N-gram

- Explore historical trends, the emergence of specific phenomena, the structure and development of Norwegian in the last 200 years
- A service under active development at the National Library of Norway, Språkbanken
- Currently available in beta
- Both for researchers and the general public
- Similar to Google Books Ngram Viewer for English and other languages

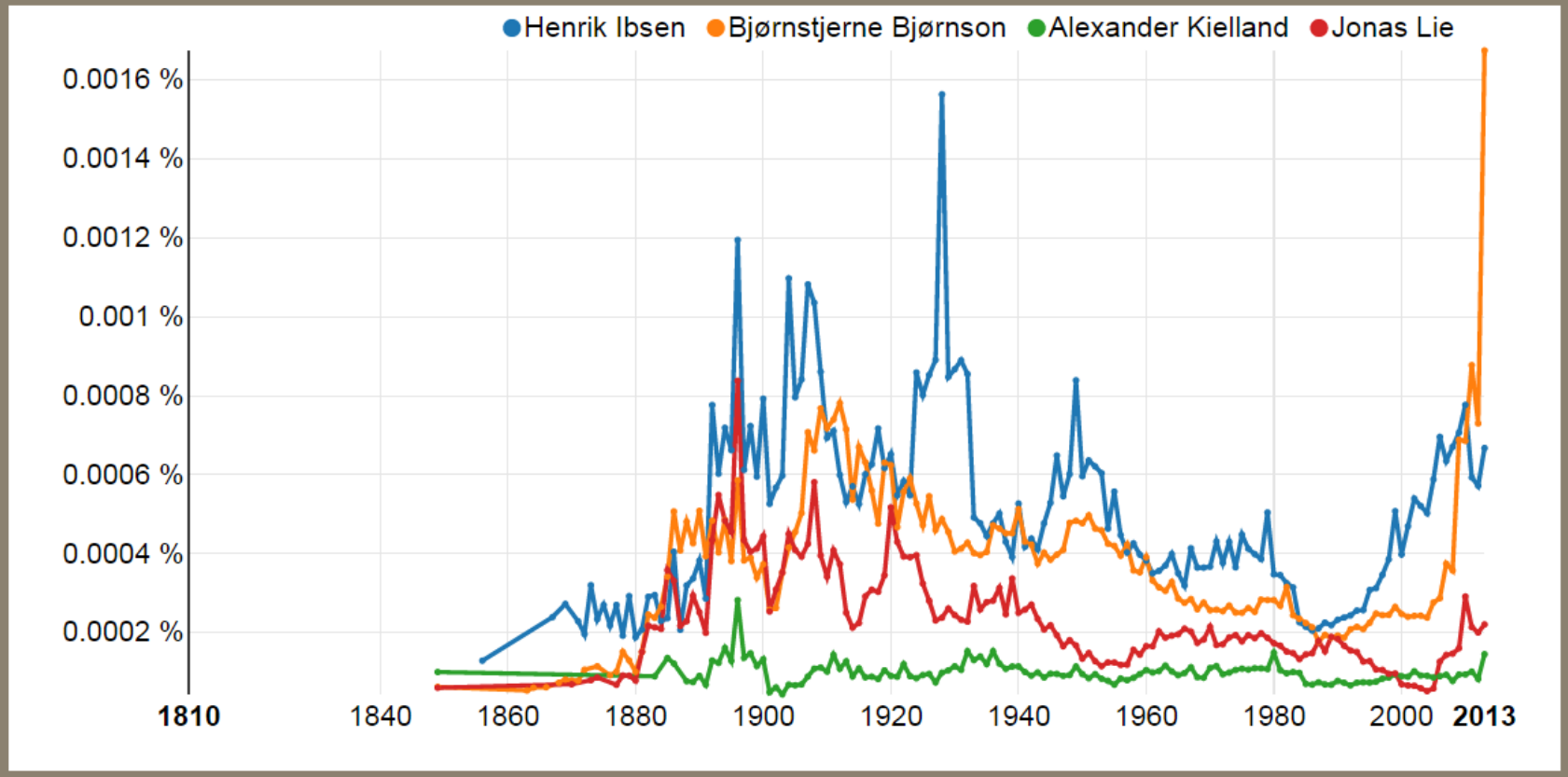
Nasjonalbiblioteket

NB N-gram BETA

Fyll inn søkeuttrykk, adskilt med komma, f.eks. forskning, formidling

Søk

Om tjenesten



1810

2013

1810 1850 1890 1930 1970 2010

Periode

0 5 10 15

Glatting

National Library of Norway

Background: NBdigital

- Large-scale digitization project (started 2006)
- Digitization of the entire collection of the National Library of Norway (subject to the so-called “legal deposit act”)
- All books published until 2001 available online to Norwegian IPs (Bokhylla)
- Full-text search possible, but copyright-protected texts are not downloadable

The NB Corpus

- Current base: All scanned material as of late 2013
- 11 billion words in 230.000 books (+ 4 billion words in other languages, not a part of the service yet) – 4,6 million types
- 23 billion words in 540.000 newspapers – 8,3 milion types

= **34 billion words**

Comparison

- Google Books Corpus (English)
  - 500 billion words in 8.1 million books (cf. Lin et al. 2012: 170)
- The NBdigital corpus
  - 38 billion words in 250.000 books and 500.000 newspapers
- German Reference Corpus
  - 25 billion words (books and newspapers)
- British National Corpus
  - 100 million words

What is a word?

- Important linguistic question
- Not as trivial as it seems
- Basic rule: A word is defined by surrounding whitespace
  - Compounds
  - Abbreviations
  - Numbers
  - Special characters (?,! etc.)

Counting words

- Tokenization
  - Extraction of words from a text using rules
  - Based on regular expressions
- Counting
  - Each unique word is counted (frequency list)
  - Done in Python (using the module collections.counter)

The notion of *n*-gram

- Sequences of words (*n* = any positive number)
- 1-gram = unigram
  - *to*, *be*, *or*, *not*, *to*, *be*
- 2-gram = bigram
  - *to be*, *be* or, *or not*, *not to*, *to be*
- 3-gram = trigram
  - *to be or*, *be or not*, *or not to*, *not to be*

  (example from: http://en.wikipedia.org/wiki/N-gram)

Why *n*-grams?

- Frequently occurring patterns
- Not only single words, also sets of words ("phrases")
- Statistically interesting
- Probability models:
  - With what probability will the word *thank* be preceded by *you*?
  - Application: when typing on a smartphone (word suggestions)

Why not higher than 3-gram?

- Statistically uninteresting
- Hard to find non-unique 5-grams in one single book
- Copyright problems
- Technical challenges (storage, RAM)

Creating *n*-grams

On-screen viewing

- *Aftenposten*, 2014-02-02, part 1, p. 28

Smørebom

Men det var andre som ikke fikk like gode svar. Hverken Petter Northug (beste norske på en 6. plass) eller Martin Johnsrud Sundby (10. plass) blandet seg helt inn i teten. De hadde trolig ikke gode nok ski til å hevde seg.

OCR

- *Aftenposten*, 2014-02-02, part 1, p. 28


```
<String ID="P28_ST00150"
HPOS="82" VPOS="862"
WIDTH="146" HEIGHT="28"
CONTENT="smørebom" WC="1.00"
CC="11111111"/>
```

Extract information

- *Aftenposten*, 2014-02-02, part 1, p. 28

<String ID="P28_ST00150" HPOS="82" VPOS="862" WIDTH="146" HEIGHT="28" CONTENT="smørebom" WC="1.00" CC="11111111"/>

Text

Smørebom

Men det var andre som ikke fikk like gode svar. Hverken Petter Northug (beste norske på en 6. plass) eller Martin Johnsrud Sundby (10. plass) blandet seg helt inn i teten. De hadde trolig ikke gode nok ski til å hevde seg.

Text

Smørebom

Men det var andre som **ikke** fikk like gode svar. Hverken Petter Northug (beste norske på en 6. **plass**) eller Martin Johnsrud Sundby (10. **plass**) blandet **seg** helt inn i teten. De hadde trolig **ikke** gode nok ski til **å** hevde **seg**.

Counting words

```
ikke, 2, 2014
smørebom, 1, 2014
plass, 2, 2014
seg, 2, 2014
å, 1, 2014
```

Storage: Database

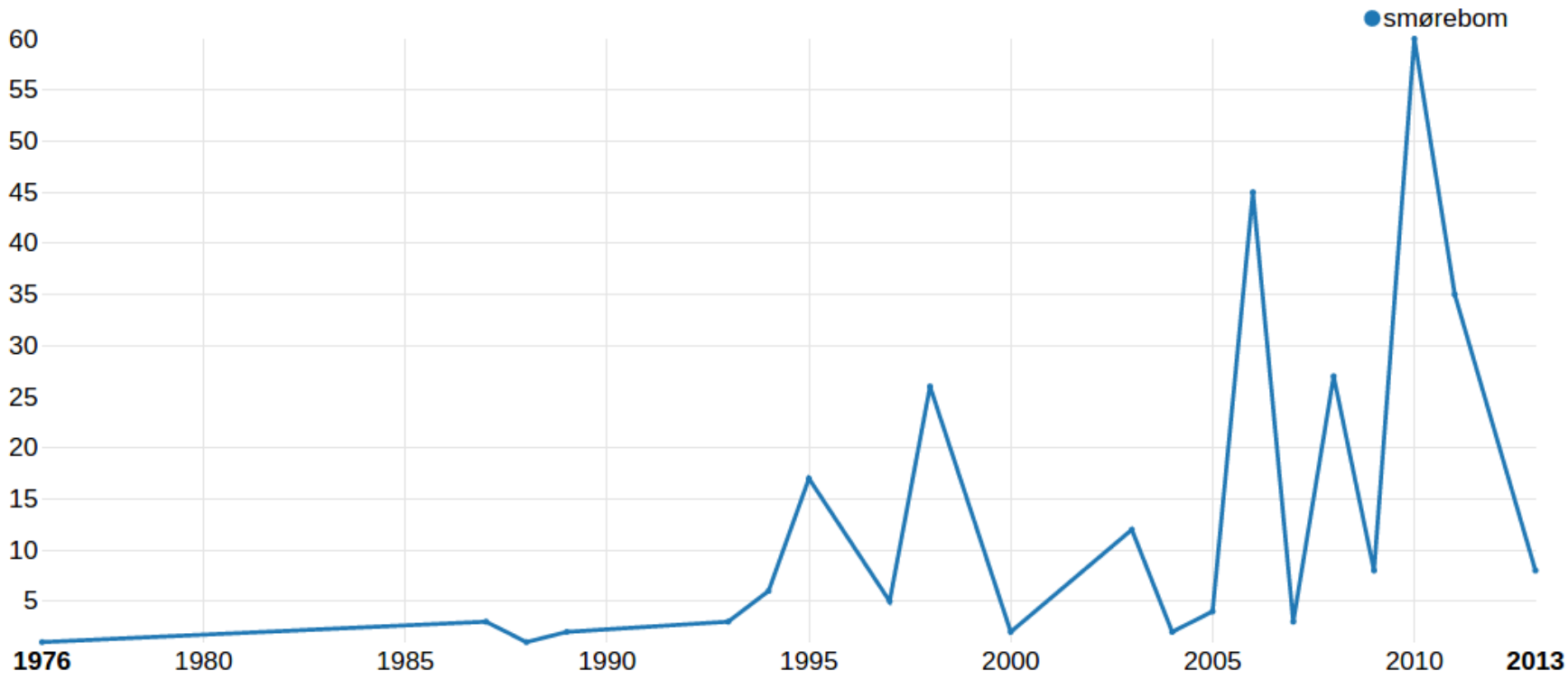
- Database system: sqlite3
  - pros: super fast, portable
  - cons: slow inserts, bad concurrency
- All ngrams stored in separate tables (unigram, bigram, trigram)
- Heavy use of indexes (especially important for wildcard search)

first	json	lang	freq
smørebom	{"1976": 1, "1987": 3, "1988": 1, "1989": 2, "1993": 3, "1994": 6, "1995": 17, "1997": 5, "1998": 26, "2000": 2, "2003": 12, "2004": 2, "2005": 4, "2006": 45, "2007": 3, "2008": 27, "2009": 8, "2010": 60, "2011": 35, "2013": 8}	nob	270

Web application

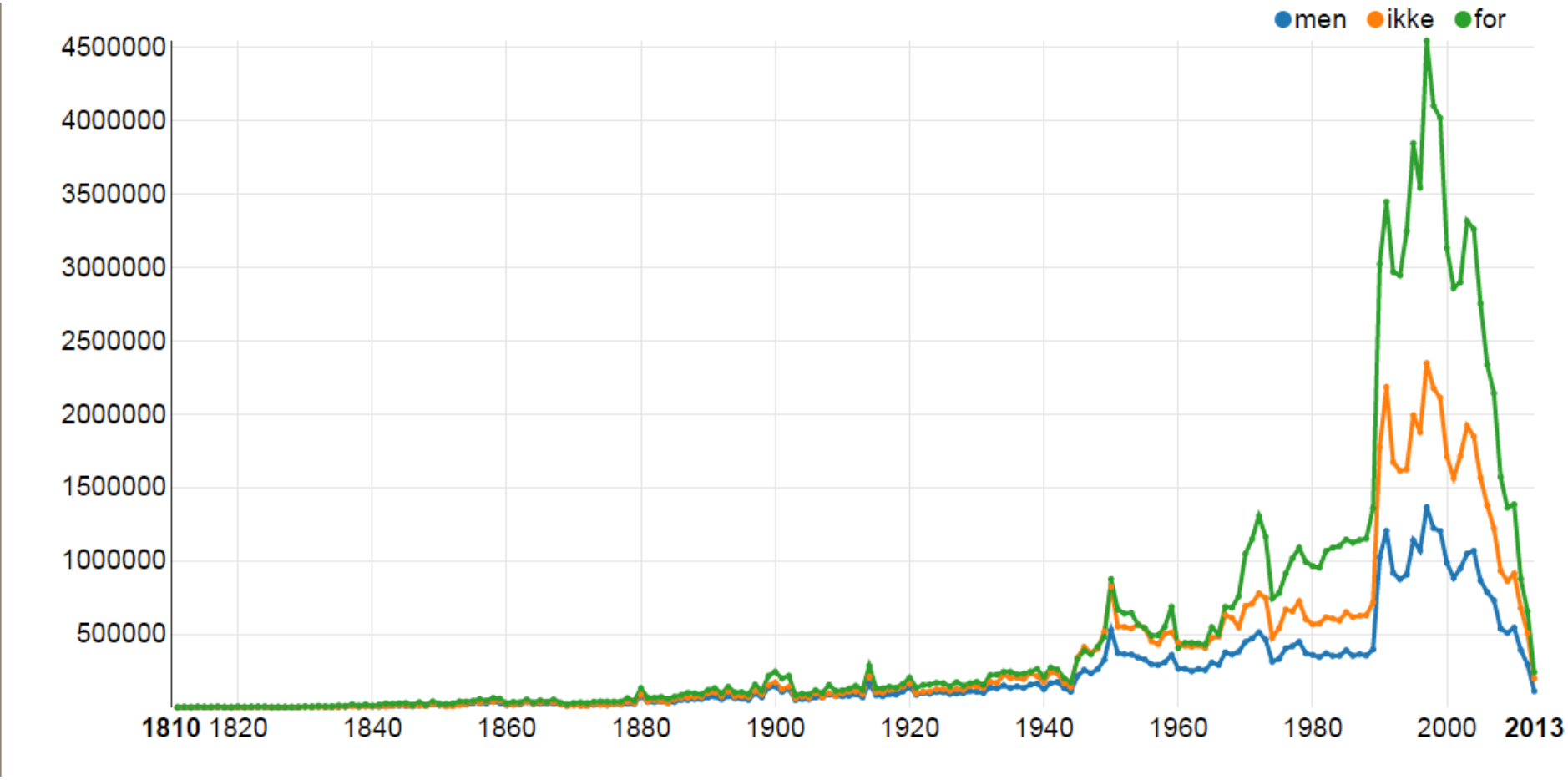
- Coded in Python/Flask and JS
- Three components:
  - N-gram database (sqlite)
  - Front-end (HTML and JS) with chart (using nvd3)
  - Layer between the database and the front-end (providing JSON output)

first	json	lang	freq
smørebom	{"1976": 1, "1987": 3, "1988": 1, "1989": 2, "1993": 3, "1994": 6, "1995": 17, "1997": 5, "1998": 26, "2000": 2, "2003": 12, "2004": 2, "2005": 4, "2006": 45, "2007": 3, "2008": 27, "2009": 8, "2010": 60, "2011": 35, "2013": 8}	nob	270



Absolute frequency

- Absolute frequencies often not very interesting, especially in large corpora
- The number of books and newspapers has risen substantially within the last 200 years

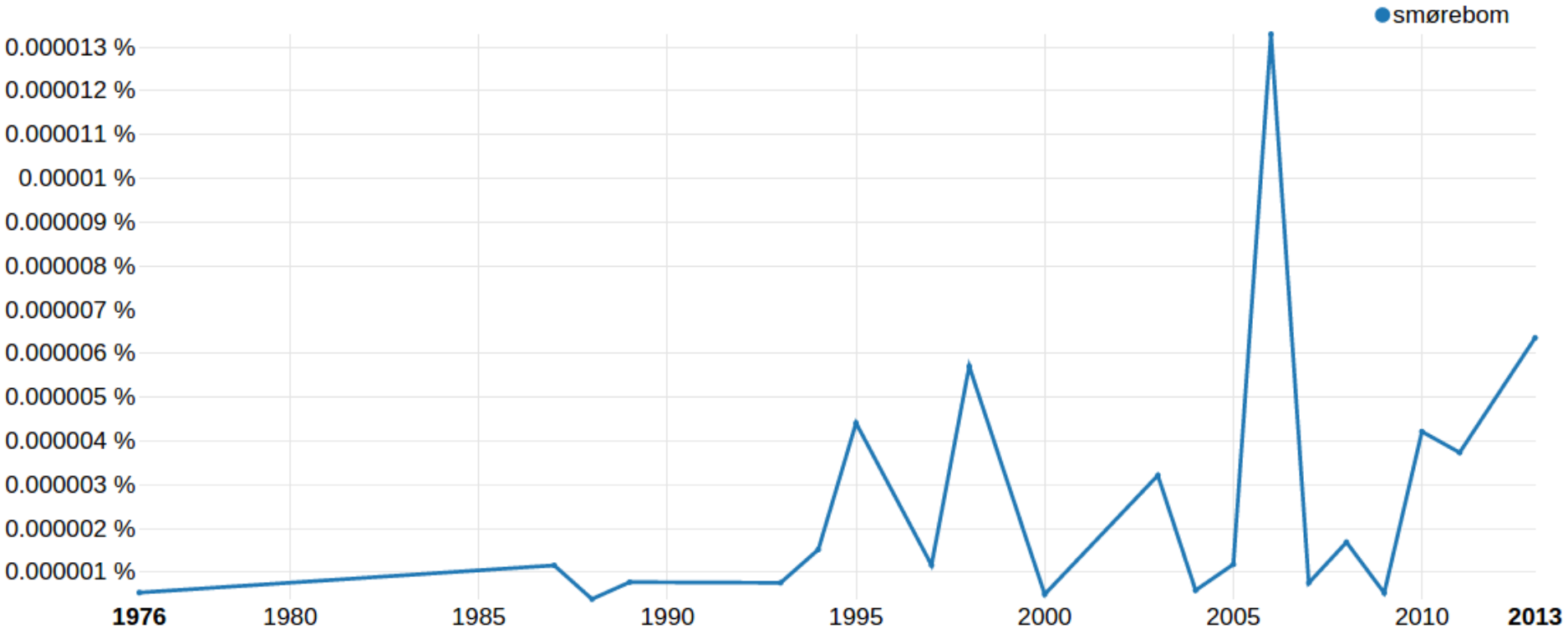

men
ikke
for
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
1810
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
2013

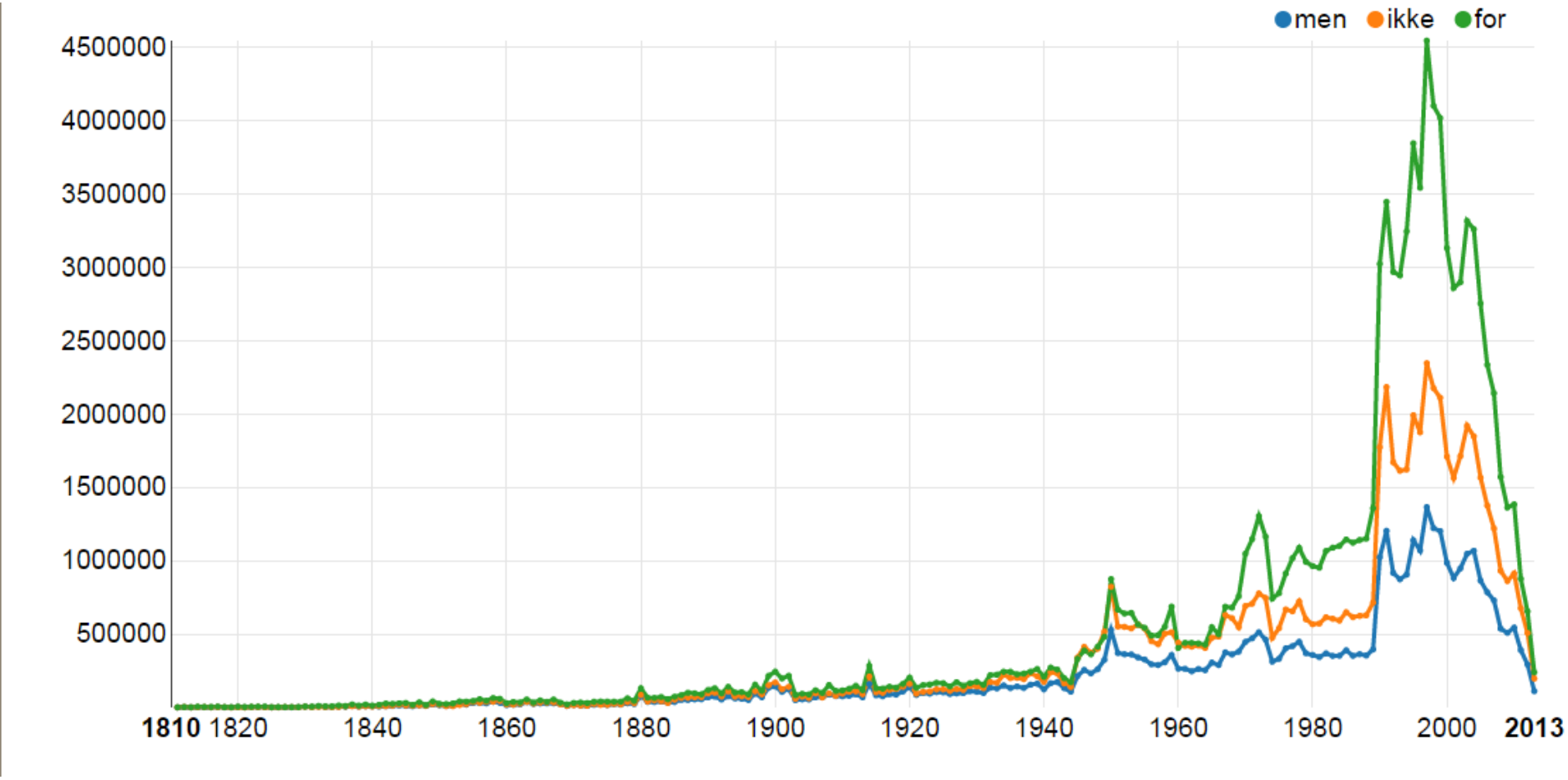
Relative frequency

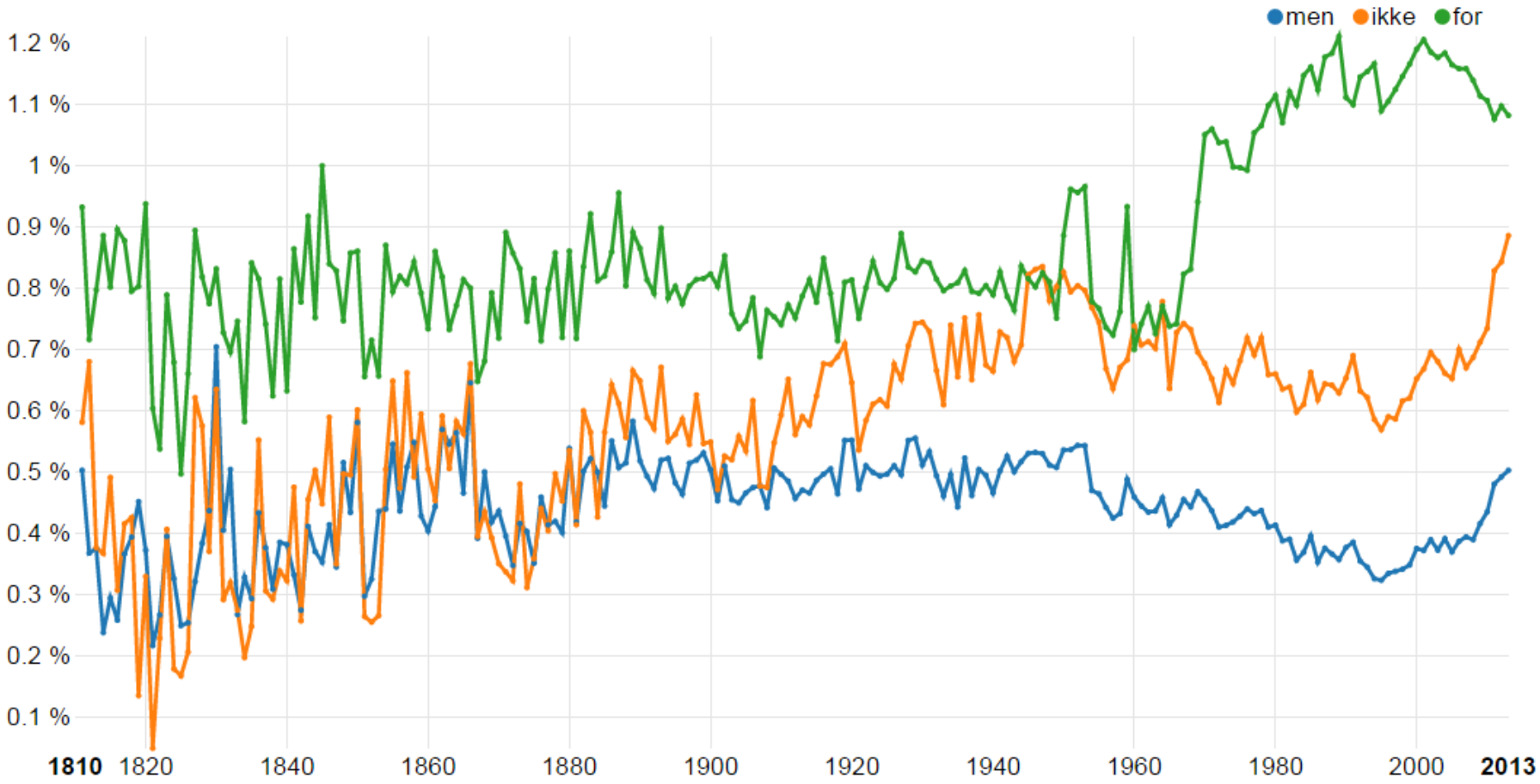
$$relative\ frequency = \frac{word\ frequency\ in\ one\ year}{frequency\ of\ all\ words\ in\ that\ year}$$

$$2006: \frac{46}{338477177} = 1{,}32948402603818 7e\text{-}7$$

$$2010: \frac{60}{1424108082} = 4{,}21316336578448e\text{-}8$$

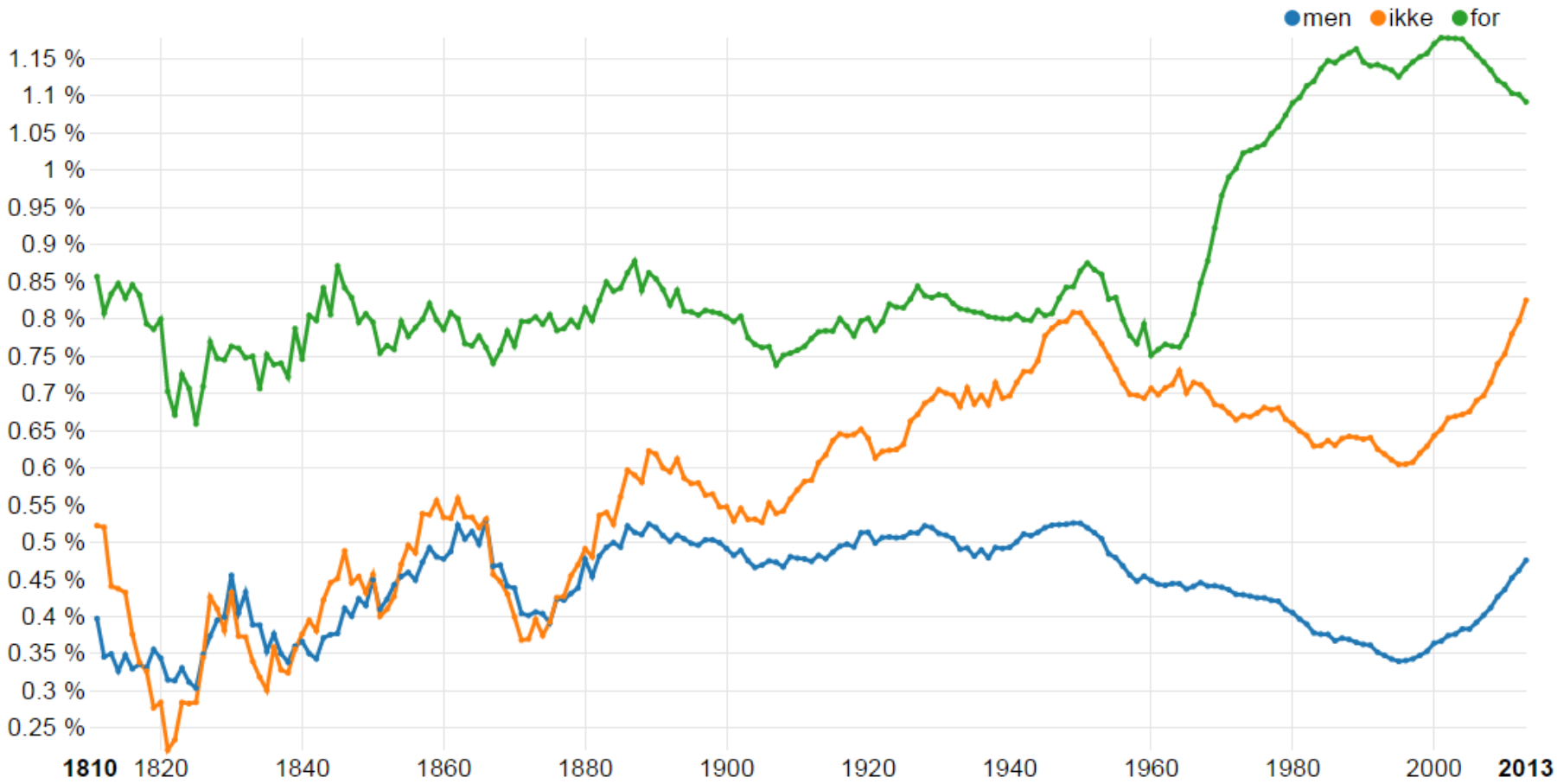



men
ikke
for
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
1810
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
2013



Smoothing

- Spikes in the chart
- Usage of a word to a certain degree random
- “Smoothed out” in the trend viewer
- The relative frequency of one year is the average of that year and a certain number of years before and after that (default: 4), so-called “moving average”.



What can you do with NB N-gram?

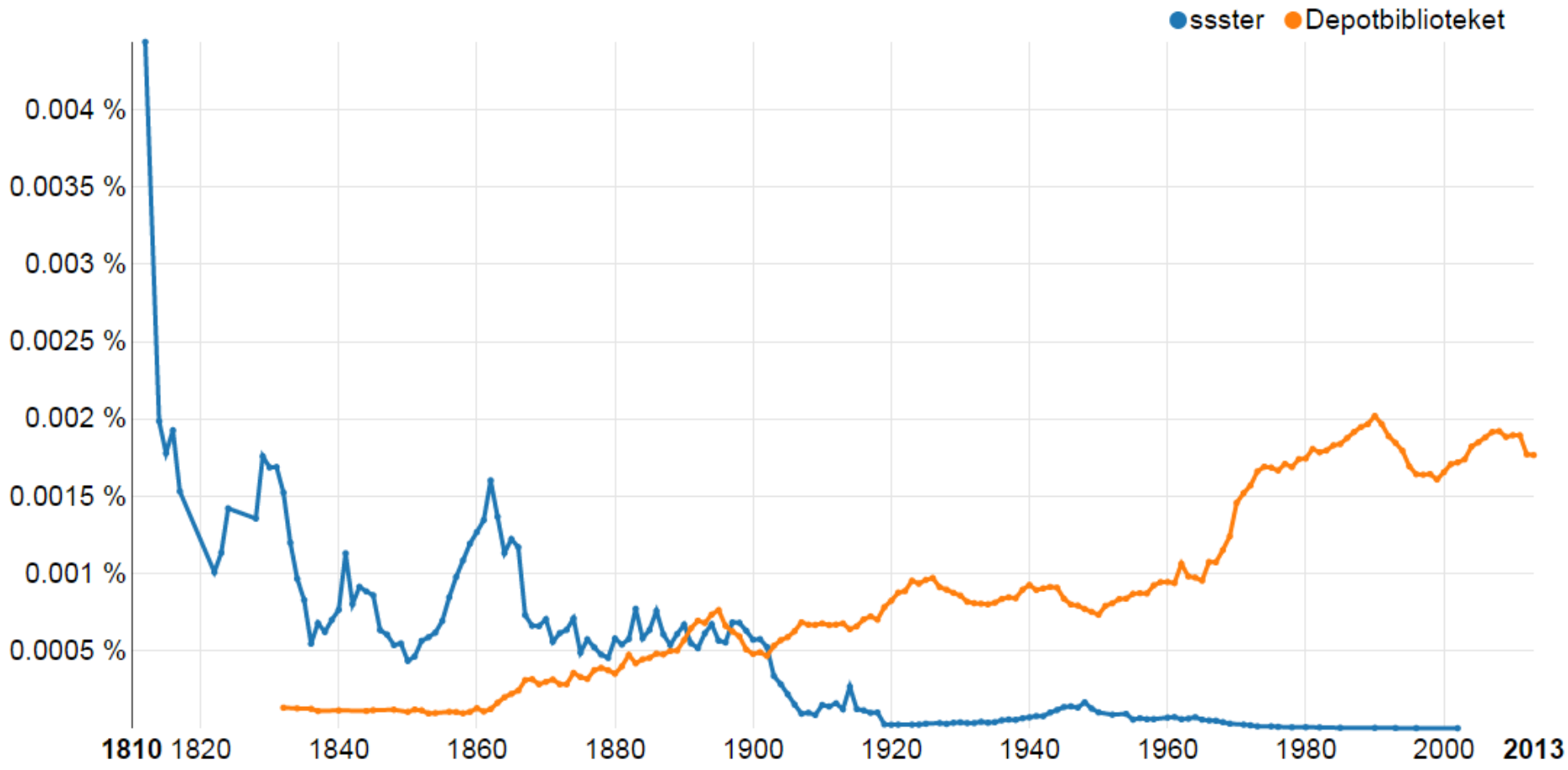
- Trends
- Language change
- History

Demo

URL:
http://www.nb.no/sp_tjenester/beta/ngram_1/

Problems

- OCR errors
- other



National Library of Norway

Future plans

- Morphological tagging and syntactic parsing
  - Search only for nouns, verbs, specific phrases
- Lemmatization
  - Searching for a word will also include all inflectional forms
- Categorization based on Dewey
  - Search for and compare words according to the type of book they occur in

Thanks

- Our colleagues at the National Library of Norway