From digital library to n-grams: NB N-gram

Magnus Breder Birkenes **Lars G. Johnsen**
Arne Martinus Lindstad **Johanne Ostad**
The National Library of Norway
P.O.Box 2674 Solli
NO – 0203 Oslo, Norway
{magnus.birkenes, lars.johnsen, arne.lindstad,
johanne.ostad}@nb.no

Abstract

At the National Library of Norway, we are currently developing a service comparable to the Google Ngram Viewer (Michel et al., 2010; Lin et al., 2012; Aiden and Michel, 2013) called NB N-gram. It is based on all books and newspapers digitized up to and including 2013, as part of the large scale digitization project at the National Library of Norway. Uni-, bi- and trigams have been generated on the basis of this text corpus containing some 34 billion words. In this paper, we sketch the background of NB N-gram and illustrate some applications of it.

1 Background

In 2006, the National Library of Norway initiated an ambitious digitization program, with the goal of digitizing its entire collection. The collection contains all material collected under the legal deposit act, and includes among other things books, newspapers, periodicals, magazines, journals, music, films, posters and maps; basically anything published in the public domain in more than 50 copies. The collection contains material in many different languages.

The aim of the digitization program is to make the entire collection available for viewing purposes in a way that does not challenge intellectual property rights. To this end, agreements have been made, that make it possible to give access to the digitized content. *Bokhylleavtalen* ("The Bookshelf agreement") from 2012 gives the National Library right to make all books published up to and including the year 2000, available to users with a Norwegian IP-address.

The National Library has also made agreements with newspapers that make it possible to give access to a number of newspaper titles in digital format in Norwegian libraries. Some titles are also available outside public libraries to all users. Another example is an agreement made with the major public broadcaster in Norway (NRK), where open and free access for everyone is given to more than 36,000 radio programs. This includes radio broadcast news from the 1930s onward.

2 NB N-gram

In order to present an alternative and linguistically and historically more interesting take on the material, the thought developed that a statistical approach to the contents in the Digital National Library would be interesting.

NB N-gram gives both researchers and the general public the possibility to look at linguistic and cultural trends in this material, by connecting text and metadata (year, language information).

2.1 Generating n-grams

For copyright reasons, it is impossible to give access to full text versions of the material in the collection, but from the material underlying the n-gram viewer, the digitized text has been extracted, and uni-, bi- and trigrams have been generated from a base consisting of 34 billion words (11 billion words from 230,000 books, and 23 billion words from some 540,000 newspapers, spanning the period 1810-2013).

The texts in the Digital National Library are stored in XML format (ALTO XML) and were converted to plain text and then tokenized. Frequencies were counted for each single unique n-gram, but only texts in Norwegian Bokmål and Norwegian Nynorsk were considered in the first revision. The language classification relies upon

information from the national bibliographical system BIBSYS, which is mostly, but not always, correct (an automatic detection using character n-grams would probably provide more exact results). The resulting data set consists of a collection of n-tuples on the form (n-gram, year, language, frequency).

2.2 Technical Implementation of NB N-gram

Building upon this material, NB N-gram consists of three components: a frontend, a backend and an n-gram database. Essentially, it is a web application written in Python/Flask.[1] The user enters search terms that the backend converts into valid SQL statements. The backend then returns the results from the database as a JSON object. The chart is drawn entirely on the client-side, using nvd3.[2]

The database is the single-most important component: We chose sqlite3 for retrieval speed and portability.[3] The database contains tables for each unique n-gram (unigram, bigram and trigram), which are connected to tables containing frequency information on these n-grams for both languages covered (Norwegian Bokmål and Norwegian Nynorsk), as shown in Figure 1 below.

freq	year	lang	first
60	2008	nno	lingvistikk
60	2008	nob	lingvistikk
120	2008	all	lingvistikk

Figure 1: Sample from the database

One frequency table holds all absolute frequencies for a particular n-gram for each year (from 1810 to 2013). Another table, used for wildcard-search (more on that in section 3.2), contains the summed frequencies for all years. We have chosen to store as much information as possible: Since we are dealing with tables containing several hundred millions of entries, doing some operations on the fly (like aggregating numbers for the two languages and then sorting on them) has proven to be way too time-consuming. As a result, we store most of the numbers (only the relative frequency is computed). With the help of indices, a query is very fast, even on slower HDDs (ca. 0.1 seconds).

3 User Interface

3.1 Basics

NB N-gram has a simple user interface that was created also with visually impaired users in mind. Thus all elements scale well on all devices and graphs can be shown either in colors or in grayscale.

The most central element to the user interface is the chart. By default, the chart shows the frequency representation of the four classic authors in the Norwegian literary canon. Frequencies are given as relative frequencies, but an option for showing absolute frequencies is also included.

3.2 Search

Above the chart the user will find a search box, where search terms may be entered, separated by commas (like in the Google Ngram Viewer), each search term resulting in a separate graph. Figure 2 shows a sample search in the newspaper material using the three search terms “EEC”, “EF” and “EU” (different abbreviations for the political institution now known as the European Union) with spikes in 1972 and 1994, when the two referendums on Norwegian EU membership were held.

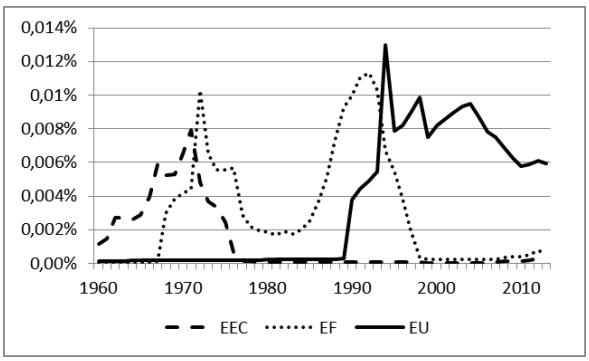

Figure 2: Trend lines for ”EEC”, ”EF”, ”EU”.

A search term may contain one single *n*-gram (from unigram up to the size of a trigram) or the combination of many n-grams, using the + operator. In this way, a search for *hest+hesten+hester+hestene* ‘horse, the horse, horses, the horses’ results in a graph with all inflectional forms of the word *hest* ‘horse’.

A wildcard search is also possible: Similar to the Google Ngram Viewer, using a wildcard in a

[1] http://flask.pocoo.org/
[2] http://nvd3.org/
[3] https://sqlite.org/

search term will plot the ten most frequent n-grams matching that criterion. Technically speaking, "most frequent" is here defined as the n-gram having the highest total frequency across the whole period. Unlike the Google Ngram Viewer, NB N-gram also allows wildcards inside words: for example, the search term **else* will result in ten graphs showing the most frequent words ending in the derivational suffix *–else* within the period 1810-2013.

3.3 Smoothing

In order to compensate for variation from year to year, NB N-gram uses a smoothing algorithm similar to that of the Google Ngram Viewer. Thus, a smoothing of 4 (which is the default) means that the frequency of a particular year is computed as the average of the relative frequency in the four years before and the four years after, divided by nine.[4]

3.4 Customized views

The user interface itself allows for certain customizations: The default range (1810 to 2013) may be decreased in order to focus on a special period (for example a decade). Also, clicking on the bullet points in the legend allows for blending out (and in) individual graphs.

3.5 Download of Raw Data

The statistical data underlying the graphs – both relative and absolute frequencies – can be downloaded as .csv-files (comma-separated text). Also the graphics can be downloaded, as scalable high-quality .svg-files.

3.6 Inspecting the underlying material in NB Bokhylla

Clicking on a graph gives the user the possibility to show examples of the search terms in context through NB Bokhylla. If you are in Norway, you get access to all material published before 2001. If you are outside of Norway, only sources not protected by copyright are shown.

4 Further perspectives

In this paper, we outlined the background of NB N-gram and showed some of its applications.[5] In the future, we hope to provide additional functionality such as linguistic annotation and genre-based search (based on Dewey classification). We also want to look at the possibility of including other languages from our material, such as Sami and Kven.

Acknowledgments

We would like to thank our colleagues at the National Library of Norway for input and technical assistance.

References

Erez Aiden and Jean-Baptiste Michel. 2013. *Uncharted: Big Data as a Lens on Human Culture*. Penguin, New York.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, William Brockman and Slav Petrov. Syntactic Annotations for the Google Books Ngram Corpus. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics Volume 2: Demo Papers (ACL '12) (2012)

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative Analysis of Culture Using Millions of Digitized Books. Science (Published online ahead of print: 12/16/2010).

Google Ngram Viewer Documentation: `https://books.google.com/ngrams/info`

[4] Google Ngram Viewer Documentation: `https://books.google.com/ngrams/info`

[5] A beta version of NB N-gram is available via the following link: `http://www.nb.no/sp_tjenester/beta/ngram_1`