The Swedish Culturomics Gigaword Corpus: A One Billion Word Swedish Reference Dataset for NLP

Stian Rødven Eide
Språkbanken
Dept. of Swedish
University of Gothenburg
stian@fripost.org

Nina Tahmasebi
Språkbanken
Dept. of Swedish
University of Gothenburg
nina.tahmasebi@svenska.gu.se

Lars Borin
Språkbanken
Dept. of Swedish
University of Gothenburg
lars.borin@svenska.gu.se

Abstract

In this paper we present a dataset of contemporary Swedish containing one billion words. The dataset consists of a wide range of sources, all annotated using a state-of-the-art corpus annotation pipeline, and is intended to be a static and clearly versioned dataset. This will facilitate reproducibility of experiments across institutions and make it easier to compare NLP algorithms on contemporary Swedish. The dataset contains sentences from 1950 to 2015 and has been carefully designed to feature a good mix of genres balanced over each included decade. The sources include literary, journalistic, academic and legal texts, as well as blogs and web forum entries.

1 Introduction

Having openly available standard datasets for a language is of great benefit for researchers as experiments can be reproduced and algorithms can be compared. A major effort aimed at sharing linguistically annotated Swedish corpora in order to facilitate research in language technology as well as other fields, is ongoing at Språkbanken (the Swedish Language Bank), a language technology research unit and research infrastructure at the University of Gothenburg. However, these corpora typically represent one type of text each, a book, an online forum or governmental reports to give some examples, much like existing datasets for other languages (Sandhaus, 2008; Ferraresi et al., 2008). Following the work of Schäfer and Bildhauer (2012) and others, this paper presents an effort to create a dataset, the Swedish Culturomics Gigaword Word Corpus, where the corpora available for downloading in sentence-scrambled versions from Språkbanken have been sampled to create a large representative contemporary Swedish dataset, from 1950 and onwards. Like the BNC corpus (BNC Consortium, 2007) the aim is to cover different domains and media to offer a balanced dataset. The dataset will be released in clearly indicated static versions to facilitate reproducibility of experiments, comparison of algorithms as well as referencing of data. Each sentence is marked with the year of publication and a genre to help filter the data for specific purposes. It will therefore be possible to use only a portion consisting of, e.g., social media data for a given year.

To assist with common usage scenarios, we will, in addition to the dataset, release code to help extract the text in a desired format; plain text or with different levels of annotation such as part-of-speech tags or multi-word expressions.

2 Contents of the Swedish Culturomics Gigaword Corpus

The dataset contains just over one billion words, sampled from a variety of sources dating from 1950 and onwards. It is designed to be representative of contemporary Swedish, by which we mean texts published from the 1950s until the present day. The last major change in written Swedish was the gradual phasing out of subject–verb agreement. The written standard prescribed the use of distinct singular and plural forms of verbs, even though most spoken varieties had lacked this distinction for centuries. Most authors conformed to the norm until the 1930s, when it became fashionable to switch to using only one form, corresponding to the earlier singular form. Newspapers followed suit in the 1940s, and by 1945,

after a recommendation from the newly formed *Nämnden för svensk språkvård* (Swedish Language Cultivation Board), all major newspapers had abandoned the plural form (Pettersson, 1996). By choosing texts published after 1950 we ensure that the grammar of the language is identical to the contemporary form for which Språkbanken's language processing tools are adapted, and hence produce higher-quality annotations.

Each sentence in the dataset is analysed using the corpus import pipeline of Språkbanken's corpus infrastructure *Korp* (Borin et al., 2012). The NLP tools used by the pipeline are state of the art, but their performance is unequal and heavily dependent on text type, genre, etc. Adesam et al. (2014) describe ongoing work on building an evaluation dataset which will more faithfully reflect the variety of text types and genres found in our corpus collection, and which consequently will allow us to reach a better estimation of the accuracy of the NLP tools that we use for corpus annotation. Our dataset becomes a snapshot of all sub-corpora from their creation date and thus constitutes a static dataset which facilitates reproducibility of experiments as well as comparison of algorithms. We trust that this dataset can cover the needs of most NLP research working with contemporary Swedish.

2.1 An overview of our sources

In Table 1 we provide an overview of the various sources included in our dataset, listed with their respective genres and time periods, as well as the numbers of tokens and sentences. We aim to balance the dataset with regards to genre and the date of publication, subject to availability. The fiction genre is particularly small compared to the others because publishing houses have generally been reluctant to release works of fiction for inclusion in corpora. In most of the corpora which can be downloaded from Språkbanken, in order not to infringe copyright the order of the sentences within each corpus have been randomised to limit the possibilities to reconstruct the original text, the freely licensed Wikipedia and official government texts being the notable exceptions.

Source	Genre	Time period	No. of tokens	No. of sentences
Bonniersromaner	Fiction	1976-1981	10,884,795	806,627
Norstedtsromaner	Fiction	1999	2,534,307	194,699
SALT svenska-nederländska	Fiction	1980-1989	1,335,455	96,995
SUC-romaner	Fiction	1990-1999	4,653,784	330,127
Smittskydd	Government	2000-2009	691,716	41,066
Statens offentliga utredningar	Government	1950-1999	50,000,071	2,391,382
Svensk författningssamling	Government	1990-1999	8,335,298	277,030
Svenska partiprogram och valmanifest	Government	2000-2009	821,777	50,684
8 Sidor	News	2000-2009	678,766	59,236
Dagens Nyheter	News	1987	5,122,237	364,226
Göteborgsposten	News	1994-2013	271,239,984	18,935,974
Press 65-98	News	1965-1998	41,177,162	2,891,152
Webbnyheter	News	2001-2013	271,806,921	15,112,300
DiabetologNytt	Science	1996-1999	228,398	14,129
Forskning & Framsteg	Science	1990-1999	744,000	44,538
Humaniora	Science	2010-2015	14,437,043	673,820
Läkartidningen	Science	1996-2005	19,471,910	1,085,785
Samhällsvetenskap	Science	2000-2009	10,873,267	523,102
Svenska Wikipedia	Science	2015	152,333,391	5,972,649
Bloggmix	Social media	1998-2015	35,253,548	2,254,343
Familjeliv	Social media	2000-2015	68,011,169	4,521,566
Flashback	Social media	2000-2015	45,000,152	3,095,212

Table 1: An overview of the sources on which the Swedish Culturomics Dataset is based.

3 Creating the Dataset

For our dataset, we have chosen to keep the XML format inherited from Korp, with some modifications described later in this section. All source texts have been downloaded from Språkbanken's resource pages.[1] The texts have been annotated with the Korp pipeline and the resulting output file is in a simple XML format, distingushing the hierarchical levels *corpus*, *text*, *sentence*, and *w(ord)*.

[1] http://spraakbanken.gu.se/eng/resources

Each word is enclosed in <*w*> tags, which contain syntactical, morphological and semantic annotation from Korp. Much of Korp's word-level analysis is done using SALDO, a lexical-semantic network linking words by their associations, as well as providing information on inflectional morphology and compounding behaviour of lexical items. Although different in several respects, SALDO can be described as a Swedish alternative to WordNet (Borin et al., 2013). The annotations that we extract through the annexed Python code (see section 4) are all derived from SALDO.

While we have kept all of Korp's annotation in our dataset, the code we provide to extract data from it uses either the word itself (for plain text output), the word's *lemma* attribute, the *saldo* attribute (signifying word sense) or the *lex* attribute (the lemgram – a combination of a *lem*ma and *gram*matical information – part-of-speech, inflectional paradigm and compounding behaviour).

In addition to the linguistic annotation, we have added two attributes to each *text* tag; a *year* attribute corresponding to the year of publication and a *genre* attribute chosen from one of the following:

- fiction
- government
- news
- science
- socialmedia

The dataset is structured by decades where each decade consists of several files and each file contains up to one million sentences. This structure, as well as the possibility to filter on genre and year, will allow users to easily choose a subset that suits their purpose, as well as keep the processing requirements low.

4 Using the dataset

The dataset is provided as a series of XML files in UTF-8 encoding, compressed with bzip2 to preserve space and bandwidth. The structure of the XML files is identical to that produced by Korp as described above, with the addition of *year* and *genre* attributes in each *text* tag.

Note that while the order of the sentences is mostly random within each source, we have not performed any additional randomisation when creating the dataset. This means that certain genres may only be found at the beginning or end of the dataset. If randomisation is important for the application, then this must be performed on the dataset. Additionally, some time periods may be dominated by a specific genre, especially the period 1950–1960 for which we only have government-produced text. More recent decades are, however, better balanced.

Distributed with the dataset are two files with Python code[2] that can be used to extract data to a text file also encoded in UTF-8. The code can output any of the following:

- Plain (the original words from the source without any formatting)
- Lemma (each word is replaced by its lemma where Korp has found one)
- Word sense (each word is replaced by its word sense as classified by SALDO)
- Lemgram (each lemgram contains the part-of-speech tag as well as a number signifying the inflectional paradigm)

The output is determined by the `--mode` flag. If this flag is not given by the user, the program will default to plain mode. An overview of the basic usage is provided in Table 2.

For extracting to all outputs except plain, a flag `--mwe` can be used that contracts multi-word expressions (MWEs). In practice, this means that the lemma for the first word in an MWE will contain the whole expression, while the lemmas for the rest of the words are removed from their respective positions in the sentence. The POS is also updated to reflect that it is an MWE. E.g., an adverbial expression such as *en gång till* 'one more time' will receive "abm" as its POS, the final "m" signifying an MWE.

In addition, a flag `--first-only` can be used to only output the first *lemma*, *saldo* or *lex* attribute, respectively, where more than one option is possible. This can be particularly useful for *saldo* output, where the first sense is more likely (Nieto Piña and Johansson, 2016), but less so for *lemma* and *lex*, where any one of the options should be considered equally likely to be correct.

[2]Python 3 is currently required to run the code.

Output	Flag	Attribute	Optional flags
Plain	`--mode`	`plain`	
Lemma	`--mode`	`lemma`	`--mwe --first-only`
Word sense	`--mode`	`saldo`	`--mwe --first-only`
Lemgram	`--mode`	`lex`	`--mwe --first-only`

Table 2: Basic usage of the code to extract data from our dataset.

It is also possible to filter on genre using the `--genre [GENRE]` flag, where `[GENRE]` is one of the above listed genres, written in lowercase. If omitted, the genre flag defaults to all. It is currently not possible to filter on time period without adapting the code, though as the XML files reside in different subfolders from each decade, that should not be necessary in most cases.

4.1 A usage example

A short example that shows how our XML files are annotated is listed here, using the sentence *Hönan lade sina ägg i gräset* 'The hen laid her eggs in the grass'.

```
<sentence id="8f7-8ee">
  <w pos="NN" msd="NN.UTR.SIN.DEF.NOM" lemma="|höna|" lex="|höna..nn.1|" saldo="|höna..1|" prefix="|" suffix
      ="|" ref="1" dephead="2" deprel="SS">Hönan</w>
  <w pos="VB" msd="VB.PRT.AKT" lemma="|lägga|lägga ägg|" lex="|lägga..vb.1|lägga_ägg..vbm.1|" saldo="|lägga
      ..1|lägga..2|lägga..3|lägga_ägg..1|" prefix="|" suffix="|" ref="2" dephead="" deprel="ROOT">lade</w>
  <w pos="PS" msd="PS.UTR+NEU.PLU.DEF" lemma="|sig|" lex="|sig..pn.1|" saldo="|sig..1|" prefix="|" suffix
      ="|" ref="3" dephead="4" deprel="DT">sina</w>
  <w pos="NN" msd="NN.NEU.PLU.IND.NOM" lemma="|ägg|lägga ägg:2|" lex="|ägg..nn.1|lägga_ägg..vbm.1:2|" saldo
      ="|ägg..1|ägg..2|ägg..3|ägg..4|lägga_ägg..1:2|" prefix="|" suffix="|" ref="4" dephead="2" deprel="OO
      ">ägg</w>
  <w pos="PP" msd="PP" lemma="|i|" lex="|i..pp.1|" saldo="|i..2|" prefix="|" suffix="|" ref="5" dephead="2"
      deprel="RA">i</w>
  <w pos="NN" msd="NN.NEU.SIN.DEF.NOM" lemma="|gräs|" lex="|gräs..nn.1|" saldo="|gräs..1|gräs..2|" prefix
      ="|" suffix="|" ref="6" dephead="5" deprel="PA">gräset</w>
  <w pos="MAD" msd="MAD" lemma="|" lex="|" saldo="|" prefix="|" suffix="|" ref="7" dephead="2" deprel="IP
      ">.</w>
</sentence>
```

Using our extraction code, the plain mode would generate the exact sentence as above. The output of the other modes are as follows:

```
$ bw_extract.py --mode lemma
höna lägga sig ägg i gräs .

$ bw_extract.py --mode saldo
höna..1 lägga..1|lägga..2|lägga..3 sig..1 ägg..1|ägg..2|ägg..3|ägg..4 i..2 gräs..1|gräs..2 .

$ bw_extract.py --mode saldo --first-only
höna..1 lägga..1 sig..1 ägg..1 i..2 gräs..1 .

$ bw_extract.py --mode lex
höna..nn.1 lägga..vb.1 sig..pn.1 ägg..nn.1 i..pp.1 gräs..nn.1 .

$ bw_extract.py --mode lemma --mwe
höna lägga ägg sig i gräs .

$ bw_extract.py --mode saldo --mwe
höna..1 lägga_ägg..1 sig..1 i..2 gräs..1 .

$ bw_extract.py --mode lex --mwe
höna..nn.1 lägga_ägg..vbm.1 sig..pn.1 i..pp.1 gräs..nn.1 .
```

4.2 Resource and licence

The dataset described in this article, as well as the annexed code files can be found at `https://spraakbanken.gu.se/eng/resource/gigaword`. They are licensed under the Creative Commons Attribution 4.0 International Licence: `http://creativecommons.org/licenses/by/4.0/`.

4.3 Use cases

A dataset like the gigaword corpus can be highly beneficial not only for language technology in general, as we mentioned in the introduction, but also for Culturomics in particular (Michel et al., 2011), seeing

as it makes it possible to track cultural changes as reflected in Swedish texts over time. Examples of use cases for this would be to analyse how attitudes have changed, the emergence of new technologies or to detect shifts in importance for any given topic.

5 Conclusions and future work

In this paper we have described a one billion word corpus of contemporary Swedish, containing sentences from 1950 to 2015. The sentences were chosen to feature a good mix of sources and to be balanced over each decade. The dataset is released with code to help users retain the text in a desired format, with or without annotations. In the future, we plan to release updated versions of the dataset to contain up-to-date texts as well as improved and additional Korp annotations with, for example, word sense disambiguation. We also intend to create embeddings with Word2Vec (Goldberg and Levy, 2014) and make them available together with the corpus.

References

Yvonne Adesam, Lars Borin, Gerlof Bouma, Markus Forsberg, and Richard Johansson. 2014. Koala – korp's linguistic annotations developing an infrastructure for text-based research with high-quality annotations.

BNC Consortium. 2007. The British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, page 474–478, Istanbul. ELRA.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *In Proceedings of the 4th Web as Corpus Workshop (WAC-4*.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Luis Nieto Piña and Richard Johansson. 2016. Embedding senses for efficient graph-based word sense disambiguation. In *Proceedings of TextGraphs-10*, San Diego, United States.

Gertrud Pettersson. 1996. *Svenska språket under sjuhundra år*. Studentlitteratur, Lund.

E. Sandhaus. 2008. The New York Times Annotated Corpus. *Linguistic Data Consortium, Philadelphia*, 6(12).

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul, Turkey. European Language Resources Association (ELRA).