Proceedings of ParlaCLARIN II Workshop, pages 5–10
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Anföranden: Annotated and Augmented Parliamentary Debates from Sweden

Stian Rødven Eide
Språkbanken Text, Department of Swedish
University of Gothenburg
stian.rodven.eide@svenska.gu.se

Abstract

The Swedish parliamentary debates have been available since 2010 through the parliament's open data web site *Riksdagens öppna data*. While fairly comprehensive, the structure of the data can be hard to understand and its content is somewhat noisy for use as a quality language resource. In order to make it easier to use and process – in particular for language technology research, but also for political science and other fields with an interest in parliamentary data – we have published a large selection of the debates in a cleaned and structured format, annotated with linguistic information and augmented with semantic links. Especially prevalent in the parliament's data were end-line hyphenations – something that tokenisers generally are not equipped for – and a lot of the effort went into resolving these. In this paper, we provide detailed descriptions of the structure and contents of the resource, and explain how it differs from the parliament's own version.

Keywords: parliamentary data, Swedish, NLP, speech, politics, language resource

1. Introduction

Since the freedom of information acts started becoming implemented in various countries, we have seen a plethora of parliamentary corpora being released and enhanced, by governments as well as researchers. Significant corpora have been published e.g. from the parliaments of Norway (Lapponi et al., 2018), Slovenia (Pančur et al., 2018) and the UK (Nanni et al., 2018), to name but a few.

This paper presents and describes a corpus of Swedish parliamentary debates that has been adapted from the parliament's data. In order to make it easier for further research on this data – the government's own version has also been somewhat underdocumented – we have devoted section 2 to a detailed description of the content and structure of the corpus and the accompanying metadata. In section 3 we present our improvements to the resource, in particular the handling of prevalent end-line hyphenations.

The word *anförande* (plural: *anföranden*) refers to any entry in the Swedish parliamentary debates. While the most reasonable translation into English is *speech*, an *anförande* in this context can also be a short reply to a previous speech. For the remainder of this article, however, we will use the term *speech* for all debate entries, and *anföranden* only when referring to the resource as a whole.

2. Content and structure of the corpus

The Swedish parliament has published minutes for all parliamentary debates from 1971 and onward.[1] These files are derived from scans of printed or typed documents and the large amount of HTML formatting present in the files are only for preserving layout; it does not generally segment the text in a way that helps with parsing. Metadata is restricted to document level information, and as such does not say anything about which speakers participate or which topics are being discussed.

However, all debates from 1993 and onward are also available in a separate dataset aptly named *anföranden*, where each speech is complemented with appropriate metadata such as speaker, party, topic and speech order.[2] This is the resource that we have enhanced.

2.1. Resource size and contents

After removing 20 empty documents from the parliament's data, we have 325,202 speeches, the speech texts of our cleaned version containing 122,079,937 tokens as measured with the Spacy tokeniser.[3] This gives an average of 375.4 tokens per speech.

To get a better sense of the contents of the resource, we refer to table 1. The property *kammaraktivitet* (chamber activity) in each document provides an indication to the context of the text. Unfortunately, this is not applied entirely consistently across all documents. For instance, questions to the prime minister can be found under both *statsministerns frågestund* and *frågestund med statsministern*. More importantly, however, most of the regular debates have no value for this property; they are in the table listed as *None*. On the other hand, some of them do have a label; most of the categories whose descriptions contain the word *debatt* are the types of regular debates that also dominate the category *None*. For any research pertaining strictly to the debates, our recommendation is therefore to exclude the categories we know are not debates rather than vice versa.

2.2. Document structure

In table 2, we show the complete structure of a typical speech document. In our version of the corpus, all properties except for *anförandetext* (speech text) are XML attributes of the speech as a whole. These attributes have been transferred directly from the parliament's data, with the exception of *dok_datum* which erroneously listed all parliamentary sessions as having taken place at midnight; for this reason, we removed the time stamp from the data, leaving only the dates, which are correct.

[1] `http://data.riksdagen.se/data/dokument/`

[2] `http://data.riksdagen.se/data/anforanden/`

[3] `https://spacy.io/`

Type	Amount
None	139,446
interpellationsdebatt *interpellation debate*	61,781
föredragning av utskottsärende *presentation of committee report*	58,381
frågestund *question time*	20,975
ärendedebatt *legislative debate*	16,947
allmänpolitisk debatt *general policy debate*	7,906
partiledardebatt *party leader debate*	3,616
frågestund med statsministern *Prime Minister's question time*	2,878
aktuell debatt *topcial debate*	2,601
information från regeringen *information from the government*	2,411
bordläggning *tabling*	1,441
val *election*	1,306
utrikespolitisk debatt *foreign policy debate*	1,241
statsministerns frågestund *Prime Minister's question time*	1,098
debatt vid allmän debattimme *hour of general debate*	858
särskild debatt *special debate*	536
avgörande av utskottsärende *decision on committee proposal*	512
budgetdebatt *budgetary debate*	401
meddelande *message*	323
hänvisning till utskott *referral to committee*	236
avlämnande av regeringsförklaring *submission of government declaration*	72
återupptagning av förhandlingarna *resumption of negotiations*	67
ceremoni *ceremony*	47
beslutsfattande om uppdrag *assignment decision*	46
återrapportering *report*	36
anmälan *notification*	31
riksmötets öppnande *parliamentary opening*	6
regeringsförklaring *declaration of government*	2
hälsningsanförande *welcoming speech*	1

Table 1: Types of parliamentary activity.

Property	Description
dok_hangar_id	Internal document ID
dok_id	Meeting + speech no.
dok_titel	Protocol title
dok_rm	Parliamentary year
dok_nummer	Number of meeting
dok_datum	Date of speech
avsnittsrubrik	Topic title
kammaraktivitet	Type of debate
anforande_id	Unique speech ID
anforande_nummer	Speech number in debate
talare	Speaker name
parti	Speaker party
anforandetext	Full speech text
intressent_id	Speaker's ID
rel_dok_id	Document being debated
replik	Speech type
systemdatum	Date of publishing

Table 2: A typical speech document.

- **dok_hangar_id** is a unique and strictly numerical ID which is assigned to every document in the parliament's database. It is not referenced in other documents, however, and can normally be safely ignored.

- **dok_id** is a unique ID (different from dok_hangar_id) assigned to every document in the parliament's database. In contrast to the above, dok_id is alphanumeric and referenced by other documents. Its form is derived from a set of codes that signify the time and type of the document. The two first characters refer to the parliamentary period in which the document was created, the third and fourth characters refer to the type category to which the document belongs, while the remaining characters signify a category subtype and/or number within its category. In this dataset, the category is consistently 09, meaning *minutes from the chamber*, with the subsequent digits representing the chronological number of the meeting within the parliamentary year, corresponding to dok_nummer below. A more detailed description of the dok_id format is available on the Swedish parliament website.[4]

- **dok_titel** is a human readable label that for this dataset consistently states that it is the minutes from a given parliamentary session. While it does contain an hour / minute time reference, this refers to the time of the session and not of individual speeches during the session.

- **dok_rm** refers to the parliamentary period. Since the autumn of 1975, a parliamentary period lasts from the beginning of an autumn until the end of spring the following year. The format used here is e.g. 2015/16.

- **dok_nummer** is the chronological number of the parliamentary session within a parliamentary year.

[4] http://data.riksdagen.se/dokumentation/sa-funkar-dokument-id/

- **dok_datum** refers to the date of the parliamentary session, using the format YYYY-MM-DD.
- **avsnittsrubrik** is a text label that for debates generally is informative, describing what is being debated. During a parliamentary session, it is common that several topics are debated, each usually from the premise of a proposal pertaining to legal or budgetary matters. The exact proposal being discussed is referenced by rel_dok_id below, while this label ranges from general topics such as 'climate politics' to rather specific ones such as 'increased possibilities of travelling within the European Union using national identity cards'. Not all categories of parliamentary activity feature an informative label, however; e.g. question time or debates between party leaders are only labelled with their respective categories as listed in table 1.
- **kammaraktivitet** refers to the type of parliamentary activity, as we described above in section 2.1 and listed in table 1.
- **anforande_id** is another unique alphanumeric ID assigned to each speech. As with dok_hangar_id, this is currently not referenced by other documents in the parliamentary database.
- **talare** is a string containing the name and party affiliation of the current speaker. For acting ministers, their title is usually also included, e.g. 'Finansministern Magdalena Andersson (S)'
- **parti** is a string containing only the party affiliation of the current speaker. This is listed using the common abbreviations for Swedish political parties, all currently with one or two letters.
- **anforandetext** is the transcribed speech.
- **intressent_id** is a unique ID number for the speaker. Each member of parliament since 1990 (as well as some before that) is assigned an ID of this type. This can be used to cross-reference with other data sources, as we will demonstrate later.
- **rel_dok_id** is a reference to the dok_id of whatever document is being discussed. Usually, what is being debated is some kind of proposal, from the parliament, the government, or from a commission. The formal document detailing this proposal features the dok_id referenced here. As such, it can be cross-referenced with a database containing proposals. Also, for many purposes of linguistic mining or classification, it can be more reliable as a topic than the avsnittsrubrik mentioned above.
- **replik** is a binary string, 'Y' if the speech is of the type *replik* (reply), 'N' if not. While many of the speeches not marked as *replik* may also contain or be regarded as replies to previous speeches, a *replik* is subject to slightly different rules than other speeches, the most significant being that they are much shorter.
- **systemdatum** refers to the date and time when the document was published to the parliament's database.

3. Processing the corpus

In this section, we detail our effort to improve the resource.

3.1. Cleaning

Although the digitisation of the Swedish parliamentary debates has involved optical character recognition (OCR) as part of the process, our relatively thorough manual investigation found that the result is, for the most part, excellent. There are very few typos or other indications of OCR errors. However, one particularly visible result of this process is the abundant prevalence of end-line hyphenations.

Generally, end-line hyphenation has been ignored by tokenisers, as they do not know whether to join the tokens together as a single word, join them as a hyphenated compound, or leave it as a hanging hyphen (used in elliptical constructions of a conjunction of several terms) (Grefenstette and Tapanainen, 1994; Frunza, 2008).

The commonly used tokenisers, most notably the widely used Stanford tokeniser, ignore this problem (Manning et al., 2014), and while projects such as Dridan and Oepen (2012) and Graën et al. (2018) suggest useful improvements in the area, the focus is on multi-lingual approaches which would have a hard time capturing the variety of Swedish compounds.

Due to Swedish compounding rules, where basically any number of nouns can be joined together, a pure dictionary approach is insufficient, and parliamentary debates in particular do contain a lot of hanging hyphens. This means that from the outset, a rule based approach to fixing end-line hyphenation needs to account for language specific features and preferably be complemented by manual corrections in order to reach a high accuracy.

One solution is of course to ignore them and treat them as noise, which often makes sense for large corpora where the amount of end-line hyphenation is negligible. For our anföranden, however, we found that not only were they especially prevalent, but that it often is longer low-frequency words that have been split. Such words can make a significant difference in several methods for information retrieval, text mining, and user modelling, which often use term frequency–inverse document frequency (tf-idf) or similar term weighting systems (Beel et al., 2016).

We therefore devised a rule-based method, which combined corpus look-up with hand-crafted rules and an interactive query allowing for simple manual correction of those cases that could not be resolved automatically. The procedure was as follows:

1. Generate a word frequency list from the resource. This will be used to decide whether line-end hyphenations should be kept or joined, with or without a hyphen.
2. Remove all line breaks. The reason for doing this instead of keeping the line break as a signifying feature is that there were several cases of end-line hyphenation in-line, indicating either OCR errors or several layers of OCR processing having been done.
3. Filter out all cases where the word after the hyphen is a conjunction. These cases are almost certainly part of an elliptical construction and should be kept as is.

Conjunction	English	Amount
och	and	90,025
eller	or	3,225
som	as	1,848
men	but	1,379
samt	and	744
till	to	186
respektive	respectively	172
än	yet	35
utan	without	6
såväl	as well as	7
og	and (Norwegian)	4
und	and (German)	4
kontra	versus	3
framför	before	2
liksom	as	2
snart	soon	1
inklusive	inclusive	1
o	and (shortened)	1
SUM		97,645

Table 3: Conjunctions in elliptical componds.

An overview of the frequency of the different conjunctions in elliptical compounds of several terms in the resource can be found in table 3.

4. Use regular expression matching to identify structures that almost certainly should be hyphenated compounds. These are:

   (a) All characters before the hyphen are upper-case and all characters after the hyphen are lower-case. This indicates an acronym used as a semantic qualifier.

   (b) The words before and after the hyphen are both capitalised. This indicates a proper name, which for some people and organisations is hyphenated in Swedish.

   (c) All characters before the hyphen are numerals, while the characters after the hyphen are not. This is common in Swedish, e.g. for time references such as *1990-talet*, 'the 1990s'.

   (d) The word *icke*, 'not', is particular to Swedish for requiring a hyphen when used as a prefix.

   An overview of these can be seen in table 4.

5. Generate two word forms comprising all characters before and after the hyphen, one joined with hyphen and one joined without. Check whether any or both of these are present in the word frequency list. If only one is present, choose that. If both are present, choose the one that is most frequent. If both are either missing or equally frequent, ask the user what to do.

6. Whenever a selection has been made, either by the heuristics or the user, save that selection and apply it to subsequent identical cases.

Regular expression	Unique	Total
(a) [A-ZÅÄÖ]+- [a-zåäö]+	2,527	7,740
(b) [A-ZÅÄÖ][a-zåäö]+- [A-ZÅÄÖ][a-zåäö]+	338	1,560
(c) \d+- \w+	949	2,802
(d) icke- \w+	162	283
SUM	3,976	12,385

Table 4: Hyphenated compounds matched with regular expressions.

The overall statistics are presented in table 5. Please also note that we have no way of distinguishing between end-line hyphenations and elliptical compound constructions with a hanging hyphen prior to processing. The latter are therefore included in the number of end-line hyphenations in the table.

Property	Unique	Total
Files	325,202	
Tokens (before processing)	123,261,960	
Tokens (after processing)	122,079,937	
Files with no ELH	180,350	
Number of ELH	1,080,471	
Ignored ELH due to conjunctions	97,645	
H from regular expressions	3,976	12,385
Only J in WF	97,698	904,172
Only H in WF	519	1,091
J more frequent in WF	604	44,887
H more frequent in WF	124	971
J manually selected	8,509	8,908
H manually selected	397	433
Keep manually selected	111	116

Table 5: Statistics of the end-line hyphenation processing. For the purposes of fitting the table into one column we have abbreviated *end-line hyphenation* (ELH), *hyphenated compound* (H), *compound without hyphen* (J) and *word frequency list* (WF).

As we can see, even after subtracting the elliptical compound constructions, we end up with 982,826 end-line hyphenations, comprising 0.8% of the tokens. This puts them in line with frequent prepositions; the word *med*, 'with', occurs 1,090,275 times in the data. We can also see that the strategy of looking up in the word frequency list was very effective, capturing 96.77% of the remaining end-line hyphenations.
In order to test the accuracy of this process, we chose 1,000 random items from the set of selections that were made and assessed them manually. Of the 1,000 choices our system made, we only found a single error, indicating an accurracy of 99.9%.
Our de-hyphenator has been published on GitLab under the GNU GPLv3.[5]

[5] https://gitlab.com/Julipan/swedish-de-hyphenator

3.2. Annotating

After cleaning the end-line hyphenations, we imported the resulting files into Korp, via the Sparv pipeline. Korp is a tool for searching and exploring corpora (Borin et al., 2012), while Sparv is the annotation pipeline through which most of the corpora in Korp are processed (Borin et al., 2016). Both of the tools are developed and maintained by Språkbanken Text, a language technology research unit under the department of Swedish at the University of Gothenburg.[6]

The linguistic annotation provided by Sparv is thorough and multifaceted, ranging from part-of-speech and word sense to compound and dependency analyses. A complete list of the available annotations can be found on the Sparv web page and its user manual.[7][8] The annotated anföranden can be explored at https://spraakbanken.gu.se/korp/?mode=default#?corpus=rd-anf and XML files can be downloaded from https://spraakbanken.gu.se/en/resources/rd-anf.

3.3. Augmenting

For use with the annotated anföranden, we previously created the Swedish PoliGraph, a Prolog application designed for querying and exploring Swedish members of parliament, along with their roles and activity in parliament and government (Rødven Eide, 2019).

One of the use-cases we envision is to explore speeches based on speaker metadata. Combining anföranden with the Swedish PoliGraph, we can examine questions such as which linguistic features are more common among which speakers or parties, who speaks more or less on which topics, or how commission work affects the speeches of members of parliament.

Seeing as we have exact temporal metadata for both speakers and speeches, the corpus can also be examined diachronically. We can examine how speeches change over time, for instance in the context of an individual speaker from newly elected to established, of a party changing their rhetoric in response to external events or internal conditions, or of changing attitudes as the years go by.

For further augmentation, we have also matched the internal parliamentary ID for each politician with their respective Wiki-ID's in the Swedish PoliGraph. This enables exploration of connections from politicians and speeches with data that is not part of the parliament's database, but can be found on Wikipedia or Wikidata, or other resources that use the same references.

4. Conclusion and future work

Considering the importance and availability of parliamentary data in Swedish, as well as its practical advantages for natural language processing methods – in particular the standardised language and precise metadata – very little research has taken full advantage of these resources. We hope that the publication of *anföranden*, in a cleaned, annotated and augmented form, will be a step towards further investigation of parliamentary speech in Swedish.

As part of a Swe-Clarin project on named entity recognition (NER), our next step is to manually annotate named entities in speeches from the anföranden corpus. We will then apply and evaluate various algorithms to find the current state of NER on Swedish parliamentary debates, and see if we can improve the current state of the art further.

After that, we plan to perform named entity resolution to the recognised entities, automatically linking names of politicians found in the text to their respective ID in the Swedish PoliGraph. The aim is to be able to model a complete parliamentary debate; to understand and visualise who is replying to whom.

Following the 2019 ParlaFormat Workshop in Amersfoort,[9] we will also implement export to the Parla-CLARIN XML format from Korp, after a planned upgrade of the export pipeline of Språkbanken Text is in place.

As our de-hyphenator turned out to be successful, we also plan to incorporate it in Språkbanken Texts import pipeline as an optional pre-processing step.

5. Acknowledgements

The work presented here has been partly supported by an infrastructure grant to Språkbanken Text, University of Gothenburg, for contributing to building and operating a national e-infrastructure funded jointly by the Swedish Research Council (under contract no. 2017-00626) and the participating institutions.

We also give a heartfelt thanks to the reviewers, who provided us with a lot of constructive suggestions for improvement.

6. Bibliographical References

Beel, J., Gipp, B., Langer, S., and Breitinger, C. (2016). Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17(4):305–338, Nov.

Borin, L., Forsberg, M., and Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, page 474–478, Istanbul, Turkey. ELRA.

Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schäfer, R., and Schumacher, A. (2016). Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *SLTC 2016. The Sixth Swedish Language Technology Conference, Umeå University, 17-18 November, 2016.*

Dridan, R. and Oepen, S. (2012). Tokenization: Returning to a long solved problem — A survey, contrastive experiment, recommendations, and toolkit. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 378–382, Jeju Island, Korea. Association for Computational Linguistics.

Frunza, O. (2008). A trainable tokenizer, solution for multilingual texts and compound expression tokenization.

[6] https://spraakbanken.gu.se/

[7] https://spraakbanken.gu.se/en/tools/sparv/annotations

[8] https://spraakbanken.gu.se/en/tools/sparv/usermanual

[9] https://www.clarin.eu/event/2019/parlaformat-workshop

In *Proceedings of LREC 2008*, Marrakech, Morocco. ELRA. http://www.lrec-conf.org/proceedings/lrec2008/.

Graën, J., Bertamini, M., and Volk, M. (2018). Cutter – a universal multilingual tokenizer. In *Proceedings of the 3rd Swiss Text Analytics Conference – SwissText 2018*, pages 75–81, Winterthur, Switzerland.

Grefenstette, G. and Tapanainen, P. (1994). What is a word, what is a sentence? Problems of tokenization. In *The 3rd International Conference on Computational Lexicography*, pages 79–87, Budapest, Hungary.

Lapponi, E., Søyland, M. G., Velldal, E., and Oepen, S. (2018). The talk of Norway: a richly annotated corpus of the Norwegian parliament, 1998–2016. *Language Resources and Evaluation*, 52(3):873–893.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, Baltimore, Maryland. ACL.

Nanni, F., Osman, M., Cheng, Y.-R., Ponzetto, S. P., and Dietz, L. (2018). UKParl: A semantified and topically organized corpus of political speeches. In Darja Fišer, et al., editors, *Proceedings of LREC 2018*, Miyazaki, Japan. ELRA.

Pančur, A., Šorn, M., and Erjavec, T. (2018). SlovParl 2.0: The collection of Slovene parliamentary debates from the period of secession. In Darja Fišer, et al., editors, *Proceedings of LREC 2018*, Miyazaki, Japan. ELRA.

Rødven Eide, S. (2019). The Swedish PoliGraph. In *Proceedings of the 6th Workshop on Argument Mining*, Florence, Italy. Association for Computational Linguistics.