Anna Lindahl and Stian Rødven-Eide

Argumentative Language Resources at Språkbanken Text

Abstract: Språkbanken Text at the University of Gothenburg is a CLARIN B-centre providing language resources in Swedish, as well as tools to use them, for a wide range of disciplines. In 2017, we began exploring the field of argument mining – the process of automatically identifying and classifying arguments in text – partly aimed at establishing language resources and tools for argument analysis and mining in Swedish.

Keywords: argumentation, language resources, argumentation mining

1 Introduction

Språkbanken Text at the University of Gothenburg is a CLARIN B-centre providing language resources in Swedish, as well as tools to use them, for a wide range of disciplines. In 2017, we began exploring the field of argument mining – the process of automatically identifying and classifying arguments in text – partly aimed at establishing language resources and tools for argument analysis and mining in Swedish. Depending on the context, different definitions of argumentation are applicable. For our resources, we have focused on three ways of approaching argumentation in text:

1. We have devised a set of preliminary guidelines for the annotation of argumentation in text.
2. We have looked at classifying arguments into various types of inference, in accordance with Walton's argument schemes (Walton, Reed, and Macagno 2008).

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

Note: The authors contributed equally.

Anna Lindahl, University of Gothenburg, Gothenburg, Sweden, e-mail: anna.lindahl@svenska.gu.se
Stian Rødven-Eide, University of Gothenburg, Gothenburg, Sweden,
e-mail: stian.rodven.eide@svenska.gu.se

https://doi.org/10.1515/9783110767377-026

3. With Inference Anchoring Theory, all rhetorical elements in a dialogue or debate that serve any purpose in argumentation are classified and linked.

Our work on these three approaches is laid out in the remaining sections, which are structured as follows: after an introduction to argumentation in Section 2, we describe our corpora in Section 3, followed by our annotation efforts in Section 4. Finally, we introduce some auxiliary resources in Section 5 that we hope will be beneficial to argument mining.

2 Elements of argumentation

Research on argumentation takes many forms, from Plato's search for universal truth to the pragma-dialectical notion of reasonableness introduced by van Eemeren et al. In this section, we establish a brief overview of argumentation research, with a focus on the models and methods used and discussed by computational linguists.

As for argumentation analysis in general, the model first proposed by Stephen Toulmin in 1958 (2003) represented an important milestone and is still relevant for argument mining today (Lytos et al. 2019). This model marks a shift from the strict absolutism of theoretical arguments to a practical approach, favouring justification over inference. According to Toulmin, every practical argument must consist of at least a claim (what the arguer wishes to convince someone about), grounds (evidence supporting the claim), and a warrant (the reasoning by which the grounds constitutes a valid support for the claim). While Toulmin initially focused on legal arguments, revised editions show how it can be applied to other kinds of debates.

In order to better classify types of argumentation, argumentation schemes allow us to describe structures of inference. Perhaps the best known schemes are the ones presented by Walton (Walton, Reed, and Macagno 2008). Walton presents 60 schemes which are meant to represent the type of argumentation found in everyday reasoning but also schemes present in more specialized domains. Schemes are formalized as seen below, with a minor premise, a major premise, and a conclusion. Each scheme also has a set of critical questions by which the scheme can be weakened or defeated, if the questions can't be answered. The questions can also be used to infer missing premises.

Argument from Position to Know
Major premise: Source *a* is in a position to know about things in a certain subject domain S containing proposition A.
Minor premise: *a* asserts that A (in domain S) is true (false).

Conclusion: A is true (false.)
Critical question 1: Is *a* in a position to know whether A is true (false)?
Critical question 2: Is *a* an honest (trustworthy, reliable) source?
Critical question 3: Did *a* assert that A is true (false)?

A strength of the argumentation schemes is that they often represent defeasible arguments, something which is often present in ordinary argumentation but not in traditional logic argumentation. In artificial intelligence research, argumentation has been introduced as a form of reasoning. Argumentation schemes are proposed to be used both for computational reasoning and as a tool for retrieving and analysing argumentation in speech or texts. For example, if a scheme is identified in a text, the critical questions could be used to infer what information is assumed.

Another important contribution to argument theory was the pragma-dialectical approach heralded by Frans van Eemeren and Rob Grootendorst, starting with their systematic analysis of speech act in argumentative discussions (Eemeren and Grootendorst 2010) and culminating in their book A Systematic Theory of Argumentation in 2003 (Eemeren and Grootendorst 2003). Grounded in pragmatics, this model regards argumentation as a complex form of discourse activity, and aims to describe how argumentation is carried out in practice. In the authors' opinion, speech act theory provides the necessary basis for dealing with dialogue that aims to resolve a difference of opinion. While it is far from trivial to incorporate this approach in argument mining, great strides have been made using several applicable methods, such as inference anchoring theory, which we will describe in Section 4.3.

2.1 Argumentation in natural language processing

As shown in the previous section, there are several aspects of argumentation that can be modelled and studied, and several ways in which this can be done. Argumentation annotated datasets for natural language processing (NLP) purposes reflect this and there are datasets annotated with models from various areas in argumentation theory. (There are also datasets without any clear connection to argumentation theory.) These datasets are often created as training sets, to be used by some kind of machine learning algorithm to learn from. The aim is then to automatically identify and analyse argumentation, in what is called *argumentation mining*. The task of identifying argumentation, and thus the task of modelling it, is often presented in these three steps (Stab and Gurevych 2017; Lippi and Torroni 2016):

1. Component identification;
2. Component classification;
3. Structure identification.

Component identification refers to identifying what is argumentative or not, although this step is often skipped (Ajjour et al. 2017). Component classification refers to which roles these parts are playing in argumentation, for example labelling claims and premises. After labelling components, relations such as attack or support are identified, both within individual arguments and between the arguments. Many studies or datasets do not include all these steps, as it is a complicated task. There are also more complex ways to structure the task (see, for example, Lawrence and Reed 2020). When the components themselves have been identified, some studies have explored further aspects of argumentation: for example Hidey et al. (2017) identified ethos, pathos, or logos in argument components; Park and Cardie (2014) classified components as verified or unverified. In Section 4.1, identifying argumentation schemes is explored.

3 Argumentative corpora

One of Språkbanken Text's central research tools is Korp, a corpus search and browsing tool which provides access to a collection of richly annotated corpora spanning more than 13 billion tokens (Borin, Forsberg, and Roxendal 2012). A more detailed description of Korp can be found in Fridlund et al. (2022).

The corpora we have been working on for the purposes of argumentation mining and analysis are *Anföranden*, annotated and augmented debates from the Swedish parliament Rødven-Eide (2020), as well as a collection of social media texts from two popular Swedish internet forums (Lindahl 2020). In addition, we have analysed annotation of argument schemes in a number of newspaper editorials (Lindahl, Borin, and Rouces 2019).

3.1 Parliamentary debates

During the last 15 years, access to parliamentary data has been greatly improved, especially in Europe following the signing of the Council of Europe Convention on Access to Official Documents in 2009.[1] In large part thanks to the ParlaCLARIN

---

[1] https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/205

workshops of 2018[2] and 2020,[3] significant corpora of parliamentary debates have been published and enhanced with metadata for research, such as those from the parliaments of Norway (Lapponi et al. 2018), Slovenia (Pančur, Šorn, and Erjavec 2018) and the UK (Nanni et al. 2018), to name but a few.

The Swedish parliament has published digital versions of its minutes for all parliamentary debates from 1971 onward. 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. Debates from 1993 onwards are, however, also available in a separate dataset, aptly named *anföranden* (meaning parliamentary speeches), where each speech is complemented with appropriate metadata such as speaker, party, topic and speech order. We have processed, enhanced, and augmented this resource in order to improve and simplify research on the debates, through the reduction of noise in the data, the adding of linguistic annotation, and augmenting the resource with a semantic graph, described later in this chapter. Our version of this dataset consists of 325,202 speeches, totalling 122,079,937 tokens.

In Table 1, 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 edited the time stamp in the data, leaving only the dates, which are correct. A more thorough description of the various data can be found in Rødven-Eide (2020).

After processing the documents to fix noisy data, we imported the resulting files into Korp, via the Sparv pipeline. Korp is a tool for searching and exploring corpora (Borin, Forsberg, and Roxendal 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.

The linguistic annotation provided by Sparv is thorough and multifaceted, ranging from part-of-speech and word sense to compound and dependency anal-

---

[2] https://www.clarin.eu/ParlaCLARIN

[3] https://www.clarin.eu/ParlaCLARIN-II

yses. A complete list of the available annotations can be found on the Sparv web page[4] and its user manual.[5] The annotated corpus can be explored with Korp.[6]

Table 1: A typical speech document.

Property	Description
dok_hangar_id	Internal document ID
dok_id	Meeting and speech number
dok_titel	Protocol title
dok_rm	Parliamentary year
dok_nummer	Number of meeting in succession during a year
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

3.2 Social media

Our social media dataset is made up of threads from the two Swedish internet forums Flashback and Familjeliv ('Family life'). These forums are among the most popular in Sweden and are rich in debates and argumentation, of varying levels of sophistication. They are thus suitable for studying informal argumentation. The discussions on Familjeliv are often focused on family and relations while Flashback is known for more political topics, but both forums contain a wide range of topics.

Both forums are split up into a set of main sections (19 on Familjeliv, 16 on Flashback) dedicated to different topics, with many subsections in each section. The discussions on these forums are shown in thread structures, where a user

---

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

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

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

creates a thread by posting a question or topic and other users reply. The answers are shown in chronological order. The users are able to cite each others' posts, but there is no tree-structure similar to that on, for example, Reddit.[7]

For the annotation, nine threads from these forums were chosen at random but only among the threads which had about 30 posts. As threads on these forums can end up with hundreds of posts, this was done to enable us to annotate a wider range of topics. The most recent threads were considered, which at the time were threads created in Spring 2020. The dataset used for our annotation project has a total of 28,000 tokens. The statistics of this dataset are shown in Table 2.

Table 2: Statistics of the social media dataset.

number of threads	number of posts	number of users	number of tokens	number of cite tokens	total number of tokens
9	266	150	21292	7173	28465

Apart from the annotated social media dataset, most available content posted on Flashback and Familjeliv has been collected in Korp. As much of the content is argumentative in its nature, this data could be used for studies of argumentation in these domains. The data also could be used as a supplement to supervised machine learning or unsupervised machine learning, for argumentation mining or other NLP purposes.

4 Annotating argumentation

Argumentation can be modelled and analysed in several different ways and from different aspects, and there are thus many different ways to annotate it, depending on one's goal and interest. When selecting a model for annotation of argumentation, you want to select a model which is complex enough to capture interesting information but also easy to annotate. You also want a model which a machine can learn from, if the goal is to use the data for machine learning. The choice of model might also depend on the domain. A model which is suitable in a monologic domain, such as editorials or news, might not be a good fit for a more dialogic domain, such as online forums.

When annotating different linguistic phenomena, such as argumentation, it is important to reach as high a degree of inter-annotator agreement (between

---

[7] It would be possible to construct a cite tree, but it can't be seen in the user interface.

as many annotators) as possible. This is to be sure that the annotation is reliable and captures what one seeks to study. There exist several measurements of agreement, such as Cohen's or Krippendorff's, with their respective strengths and weaknesses. Depending the task, certain thresholds are deemed acceptable, although no objective scale exists. The Landis and Koch scale (Landis and Koch 1977) is often referred to in argumentation annotation.

Annotating argumentation is challenging and time-consuming. Reaching high inter-annotator agreement is difficult, especially in unstructured domains such as user-generated content. Efforts in annotating argumentation usually do not reach as high a level of inter-annotator agreement as other tasks in NLP. A reason for this is that whether something is argumentative or not can depend on the context. For example, a statement like "I like cats" could be seen as argumentative or not, depending on which of the following statements precedes it.

1. Which animals do you prefer?
2. We should get a cat.
3. Let's get a dog.

- I like cats

If it follows 1, it could be seen as neutral, while in response to 2 or 3 it could be seen as agreement or disagreement.[8] Argumentation also often relies on implicit assumptions and unstated information. This makes it difficult for annotators to agree, because they might interpret a situation differently, and it is not always clear if there is one correct answer. It also makes it time-consuming to annotate, because the annotators often have to interpret intentions or infer missing information. Annotators might also need training in applying the chosen argumentation model, which can take time.

4.1 Annotating argumentation schemes

Our first argumentation annotation was carried out a corpus of editorials, originally described in Lindahl, Borin, and Rouces (2019). The editorials stem from Swedish newspapers originally collected by Hedquist (1978) in order to study emotive language. They were collected in the period May–September 1973 and consist of 30 editorials from 6 newspapers with about 19,000 words (Lindahl, Borin, and Rouces 2019). The newspapers were together deemed to reflect the views of the parties in the Swedish parliament at the time. The editorials from this

---

[8] Example inspired by a tutorial by Budzynska and Reed (2019).

study are annotated for emotive language, but this was not shown when annotating argumentation.

The corpus was annotated with Walton's argumentation schemes (Walton, Reed, and Macagno 2008), described in Section 2. Out of the 60 schemes described by Walton, 30 were used for the annotation. These schemes were originally presented in Walton (1996). The annotation was carried out by two annotators with a background in linguistics. For instructions, they were given Walton's book describing the schemes. The annotation was done in the annotation tool Araucaria (Reed and Rowe 2004), which has support for annotating the schemes. Using this tool, an annotator annotates arguments by first annotating argument components. A component is a span of text, labelled with the role "conclusion" or "premise".[9] These components are then connected to form an argument, which consists of one conclusion and one or more premises. A component can be reused. For example, it is possible for a premise to be connected to two different conclusions, but the premise will then be considered to be two different occurrences. The argument is then labelled with a scheme. An example of an annotated argument from the editorials is seen below.

Premise: It is already showing in the form of increasing oil and gas prices.
Conclusion: But now energy crisis is not far away.
Scheme: Argument from Sign

The annotation was evaluated on component, argument, and scheme level. The annotators annotated a varying number of components and they also varied in how they connected them to form arguments. Annotator 1 (A1) annotated more arguments and thus more conclusions than annotator 2 (A2) (each argument has only one conclusion) but they annotated about the same number of premises. This could be explained by the way they chose to connect components to arguments, as A1 often constructed arguments consisting of only one premise and a conclusion, and then reused the conclusion but chose another premise. A2 chose instead to construct arguments with several premises.

The annotators mostly used the same four or five schemes, and together they used 22 out of the 30 available schemes. The most popular schemes for both annotators were *Argument from Consequences*, *Argument from Sign* and *Argument from Cause to Effect*. A1 uses *Argument from Evidence to a Hypothesis* the most, while this scheme is used only six times by A2.

Because the annotators were free to use any span of text, the agreement measure was based on how much their annotated spans overlap. Given a certain

---

[9] The distinction between major and minor premise was not made in this annotation.

threshold, two spans were considered to be a match if their overlap was much as or over the threshold. Overlap was calculated as the ratio between the longest common span and the longest of the two spans. Thresholds of 0.9 and 0.5 were used. The agreement was then calculated as seen in (1), where *a1* and *a2* are the number of instances of the component and m the number of matches.

$$(1)\ c = 2 * |m|/(|a_1| + |a_2|)$$

Because a conclusion can be supported by different premises and a premise can support different conclusions, they were compared separately and together. The annotators agreed the most when comparing premises. With a threshold of 0.5, c is 0.37 (99 matches) for spans labelled as premises, regardless of whether they are connected to the same conclusion. For conclusions c was 0.34, with 92 matching conclusions. Out of these 92 conclusions, 33 share at least one premise. For these premises c is 0.71. In the 33 cases where a conclusion and at least one premise matched, the schemes were compared. Four schemes out of these matching conclusions and premises were the same. Comparing only matching conclusions (92), nine schemes were the same. It thus seems that even when annotators agree on how an argument was composed, they did not agree on which scheme was appropriate.

The disagreement between the annotators could be due to several reasons, including the setup of the task and the instructions itself. For example, it might have been better to structure the task so that the annotators first annotated arguments and in a later step annotated only schemes.

Some of the disagreement can be explained by differences in how the annotators structured and composed the arguments. When manually inspecting the annotations, it became clear that there is more than one possible interpretation of how to use the components. For example, below is an example of a premise supporting two different conclusions. It is difficult to say that either one of these should be the “correct” annotation.

Premise: A shift of power will result in us not risking any socialistic experiment during the elected term and instead we can further build on the foundations of the welfare society.
Conclusion A1: Voters should vote for the opposition
Conclusion A2: Do not vote away collaboration!
Scheme A1: Argument from Consequences
Scheme A2: Causal Slippery Slope Argument

Another example of this is shown below, where two different premises support the same conclusion. Again, it is difficult to say whether one is right and the other is wrong. The premises could possibly be used together.

Premise A1: It is already showing in the form of increasing oil and gas prices.
Premise A2: We are not especially used to saving anything in this country.
Conclusion A1 & A2: But now the energy crisis is not far away
Scheme A1: Argument from Sign
Scheme A2: Argument from Cause to Effect

It is not surprising that the annotators have chosen different schemes in the above examples, because different components are involved. In the few cases where they agree on components they mostly do not agree on the schemes. However, as with the components, it is possible that more than one scheme could be suitable in the annotated examples. Below is an example where annotators agreeing on conclusion and premise, but not the scheme.

Premise: It is not unlimited.
Conclusion: It is widely considered necessary to economize energy.
Scheme A1: Argument from Consequences
Scheme A2: Argument From Sign

These two schemes, *Argument from Sign* and *Argument from Consequences*, were among the most frequently used by both annotators. They are quite general and could possibly both be applicable in this case. Another example of scheme disagreement is shown below. These two schemes co-occurred 12 times out of the matching 71 conclusions (0.9 overlap threshold). Again, it is possible that two schemes might be suitable at the same time.

Premise: The high unemployment rate in Sweden is not acceptable from any angle, this must be firmly established.
Conclusion: To create new jobs must be the most important task for now.
Scheme A1: Argument from Consequences
Scheme A2: Argument from Popular Practice

Because of the disagreements between the schemes, the scheme annotation was evaluated by sorting the schemes into three groups. These three groups were originally suggested by Walton as a way to classify the schemes. This increased the agreement a little.

This dataset illustrates the difficulties of evaluating argumentation based solely on agreement between annotators, as there can be many possible interpretations of the arguments presented. It also shows the need for explicit instructions, ensuring that the annotators are coherent as possible.

4.2 Annotation of argumentation in social media

The nine threads of the social media corpus, originally described in Lindahl (2020), were annotated with spans of argumentation. Previous annotations of social media or online forums with labelled argumentation components (Habernal and Gurevych 2017; Rosenthal and McKeown 2012; Morante et al. 2020) have not reached very high levels of agreement. Because of this, the aim of this annotation effort was to investigate if it is possible to reliably annotate argumentative spans, and thus distinguish them from the non-argumentative parts of the text. If successful, these spans could be further annotated with, for example, components in an iterative annotation process. Iterative annotation processes have been previously shown to increase agreement (Miller, Sukhareva, and Gurevych 2019).

The guidelines for the annotation included a definition of argumentation, a set of control questions and tests the annotators could use when annotating. Defining what was to be considered argumentation was a bit of a challenge, as there are different definitions that do not all overlap. The definition we decided upon was inspired by van Eemeren's description of argumentation (Eemeren et al. 2014) and modified by what we found when inspecting the domain. Persuasiveness was also added to the definition, as it is often used as a criteria for argumentation (see, for example, Habernal and Gurevych (2017)). This definition was not intended to capture everything which could be considered argumentation, as this can vary, but rather to describe something which we hoped could be distinguished as argumentation. We thus defined argumentation as follows:

1. A standpoint/stance.
2. This standpoint is expressed with claims, backed by reasons.
3. There is a real or imagined difference of opinion concerning this standpoint, which leads to:
4. The intent to persuade a real or imagined other part about the standpoint.

Together with the definition, the annotators were given three questions:

- Does the poster's text signal that he or she is taking a stance / has a standpoint?
- Does the poster motivate why?
- Do you perceive the poster as trying to persuade someone?

Together with the definition and the questions, two tests were given to the annotators. These tests aimed to guide the annotators, not provide definite answers. The first test asked the annotators to insert “I agree/disagree” in the post. The idea behind this test was to capture if the text expressed any difference of opinion which might not be explicitly stated. If adding “I disagree” did not change how they perceived the text, this was probably the case.

The second test asked the annotators to reformulate the argumentative span as “A because of B”. This was to help them clarify what the stance and the motivation for the stance was. Half of the annotators were asked to write this reformulation down in the annotation tool. Examples of the test were included in the guidelines, as exemplified below.

> **I don’t agree.** Of course you shouldn’t put the dog down! It’s a life we are talking about, you can’t just throw the dog away when it doesn’t suit you anymore. Go to a professional. The dog isn’t feeling well. If you can’t help the dog you’ll have to relocate it.
> **Reformulation**: [Do not put the dog down because it has a life which shouldn’t be thrown away.]

For the annotation seven annotators were employed, split into two groups. The first group also included one of the authors, resulting in four annotators in each group. All annotators had linguistic experience through either studies or work. The annotation tool WebAnno (Eckart de Castilho et al. 2016) was used. Both groups received the same guidelines and the same threads to annotate. After the first group had annotated, a meeting was held to discuss their experiences. With the second group, a meeting was held before annotation started, in which the guidelines and the annotators’ interpretation of them were discussed. The second group was also told to write down their reformulations from the tests with the hope that this would increase agreement.

The annotation results were first compared on token level. The annotators annotated between ca 30–60% of the tokens as argumentation, although one annotator only annotated 10%. The annotators most often annotated one or more sentences in their annotations spans, following sentence boundaries. Because of this sentences instead of spans of text were compared. Most of the annotators included 4–5 sentences on average in their spans, but two of them annotated fewer sentences per span. Even though the annotators varied in how many sentences they included in a span, it was most common to only annotate one span per post. Because of this, post-level agreement was examined.

The inter-annotator agreement is shown in Table 3.[10] As there was no clear difference in agreement between the two groups of annotators, IAA is shown for both groups together. Krippendorff’s varied over threads. Unsurprisingly, post-

---

[10] The numbers here are slightly different than previously reported. This is due to a previous error in the calculations, which has been corrected.

level agreement is the highest at 0.51. According to the Landis and Koch scale (Landis and Koch 1977), this is considered moderate agreement. The observed agreement increases if one chooses to look at majority vote (five out of eight annotators agree).

Table 3: IAA for the social media dataset.

Unit	Krippendorff's α	Observed agreement	Observed agr. majority
Token	0.34	31%	74%
Sentence	0.34	31%	75%
Post	0.51	45%	84%

A manual inspection of the disagreements was also made in order to understand why they occurred. Inspection of the reformulations from the second group showed that the annotators had written similar reformulations when they had marked the same spans. Most annotators annotated around 4–5 sentences per argumentation span. In these cases, some of the annotators chose to annotate two spans instead of one, leaving one or more sentences unmarked in between the two spans. This means some annotators has interpreted a particular span of text as parts of the same argumentation, while others have found the same particular span to be to two different distinct argumentation spans, with different standpoints. This difference in argumentation spans has an effect on the sentence and token–level IAA, but not the post–level. This might be the reason why post–level results are the highest out of the three units.

Below is an example of an annotated post, exemplifying the differences in selected spans. Four annotators annotated only the part in bold. One annotator annotated the whole post. Another annotator annotated the first part as one argument, and the second part (the bold part) as another argument. The final annotator also annotated the whole post as two arguments but split the spans at the last sentence.[11]

> I agree. Little children can be bothersome and put a strain on relationships, yes. And to prefer one parent is completely normal, although it is sad, of course. **What has the three year old to be grateful for? That she should be happy and grateful that you 'sacrificed yourself' and moved there to live with them is too complicated and too much to ask of a three-year-old regardless if he/she likes to live with you or not.**

---

[11] One annotator did not annotate the post at all.

These differences highlight the difficulties with annotating argumentation, especially in unstructured domains. All but one annotator agreed this post contained argumentation but not on which parts should be included. In these domains, one standpoint is not always clearly distinguishable from another or they may be implicit. It also not always easy to decide what should be included in the argumentation. These difficulties are probably be the reasons why the annotators chose different spans.

Inspecting what the annotators had marked as annotation, it seemed that when the post authors were very explicit in their standpoints and in their disagreement or agreement, the annotators agreed among themselves. But, when sarcasm or irony was involved, or there was much left unsaid, the annotators disagreed. Thus, when the conditions in the guidelines were explicitly met, the annotators agreed. Examples of this can be seen in the two examples below. They are from the same thread and could be seen having the same message, although the second one is very implicit. In the first post all annotators agreed the post contained argumentation, whereas only three annotators annotated the second example as argumentation.

> So? And how do you think the children are feeling right now? That it’s so hard to live with their with their dad that they’d rather refrain from doing it altogether? It doesn’t matter that you thought it was boring to not to live with your boyfriend. I agree with the others in this thread that you should stop living together. For the sake of the children. You can’t just think of yourself.
>
> A three-year old should be grateful because you split up his parents? Oh my god! Are you for real?

The annotation of this dataset showed that it is possible to annotate argumentation on post-level but distinguishing the boundaries of the argumentation within a post is more difficult. Further annotations of this dataset would need to consider this. For example, can one ensure that the annotators agree on how to interpret standpoints or should one figure out a way to interpret standpoints even if annotators disagree? Stricter instructions on how to select standpoints might help with this.

4.3 Annotation of argumentation in political debates

A similar approach was used for *anföranden*, where some of the same annotators were tasked with identifying argumentation in the transcript of a single debate. The hope was that we through this would be able to create a gold standard, but first we wanted to see whether the difference in domain and structure made a significant difference to inter-annotator agreement. In contrast to the forum discussions, a parliamentary debate has a relatively formalized and predictable structure. On the other hand, any given entry in a parliamentary debate is usually longer, and may touch upon several points raised in several of the previous entries. Although it is performed orally, a parliamentary speech – especially after having been transcribed – bears characteristics of professionally written argumentation, using carefully constructed formulations, whereas forum discussions often try to emulate spoken language, inserting extra vowels into a word such as “loooong” or including interjections like “*sigh*”. Another aspect of parliamentary debates is that their very purpose is to be argumentative. Every speech voices obvious support or opposition to something, and does so in a clearly argumentative way. One could therefore assume that almost everything in a debate is argumentative. From the annotations, we saw that this was, to some extent, a reasonable expectation. A majority of the annotators found 67% of sentences to be argumentation, compared to 30% for the internet forum discussions.

In order to ensure comparability between the annotation efforts on the internet forums and the parliamentary debates, we decided to preserve as much as reasonably possible of the instructions, the main difference being that the examples were changed. However, after noticing that allowing the annotators to mark arbitrary spans as being argumentation somewhat complicated both the argumentation process and the measurement of agreement, we decided to ask annotators to always mark complete sentences in the debates, though spans of more than one sentence were allowed.

Taking all annotators into account, IAA on sentence level was even lower than for the social media dataset, at 0.29 α. Seeing that one of the annotators had marked considerably fewer sentences than the others, we measured IAA among the five other annotators and found it increased to 0.39 α. For the four annotators most in agreement, it rose further to 0.45 α. From this, we can see that the level of agreement was similar to that of the social media annotations.

On the other hand, we saw a major difference with regards to observed agreement among the majority. While we found that all annotators agreed on 25.9% of sentences, again slightly fewer than for the forums, the majority was in agreement of 89%, indicating that it may be easier to agree on argumentation in parliamentary debates, given the right approach. Further analysis of the results of this process is still ongoing, with plans to publish both annotations as well as gold standard evaluation data based on them. An overview of IAA with comparison to the social media dataset is provided in Table 4.

Table 4: IAA comparison on sentence level.

Dataset	α	Observed agreement	Observed agr. majority
Social media	0.34	31%	75% (5 of 8 annotators)
Debates (6 annotators)	0.29	25.9%	89% (4 of 6 annotators)
Debates (5 annotators)	0.39	46.2%	79.5% (4 of 5 annotators)

Another ongoing effort is annotation and analysis of parliamentary debates in accordance with Inference Anchoring Theory (IAT) (Budzynska and Reed 2011). This is a relatively complex method, as it considers all elements of a dialogue or debate that have any purpose in or effect on the argumentation. It is closely related to Rhetorical Structure Theory (Mann and Thompson 1988), but specifically adapted for analysing argumentation. Most importantly, IAT allows for anchoring inference in links between locutions, and not just locutions themselves (Budzynska et al. 2014). As current tools for IAT annotation are designed with the type of dialogue present in radio and TV debates in mind (Janier, Lawrence, and Reed 2014), we found through our initial annotation attempts that the length and complex rhetorical structure of parliamentary debates made them difficult apply in our case. Our project on applying IAT annotation to debates is therefore still ongoing.

5 Auxiliary resources

Due to the complex nature of argumentation, it is not unlikely that various knowledge resources could be helpful for argument mining. We have been working on some resources for this purpose, and as they are general in nature, we hope they will be useful even beyond the task of identifying and classifying arguments.

As a complement to the corpus of parliamentary debates, we published the *Swedish PoliGraph* (Rødven-Eide 2019), a graph of all members of parliament in Sweden. It is, in essence, a semantic database that keeps track of MPs' parliamentary activities, from speeches to responsibilities on commissions and in Governmental roles. One purpose of this graph is to combine it with named entity recognition and resolution, in order to automatically establish the argumentative structure of a given debate. Given the task of mapping a single debate, the procedure would be as follows:

1. Find all speeches with a given *rel_dok_id*.
2. Determine the meeting(s) this was debated in.
3. Establish the chronological order of the speeches during these meetings.

4. Analyse each speech and attempt to determine which previous speech or speeches (if any) was/were addressed or argued against.

For the Swedish PoliGraph, we combined the speech information from *anföranden* with metadata from the MP category, which includes basic biographical information as well as a complete history of their roles in the parliament. Such roles are usually their time working as an MP and commission work, but longer sick leave is also listed here, as well as their substitutes in those cases. In addition to the essential identifiers "name" and "party", links are also created to MPs' Wikidata-IDs and their listed name there, which sometimes provide more detail, as they are stored in the parliament's own database, while simultaneously allowing other data to be pulled from Wikipedia. The structure of the graph is shown in Figure 1.

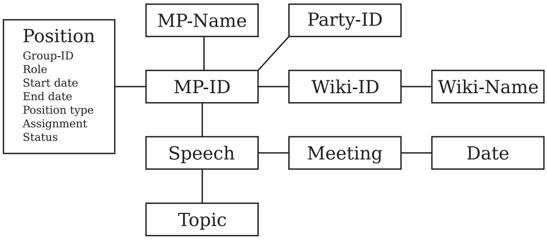

Figure 1: A semantic graph of Swedish MPs and debates.

Roles of MPs are generally described in terms of positions, where each assignment (or leave from that assignment) is stored as a factual predicate with eight arguments:

1. MP-ID
   A unique ID for each MP.
2. Agency code
   An identifying code for the agency. This can be ambiguous, as parties and commissions sometimes use the same identifier.
3. Role
   The MP's role in the agency, e.g., parliamentarian, commission chair, or substitute.
4. From
   Starting date of the position.
5. To
   End date of the position.

6. Type
   The type of position, usually either "kammaruppdrag" for the parliament or "uppdrag" for commission work.
7. Uppdrag
   The info here varies. For commission work and other extraparliamentary duties, it contains the full name of the commission or equivalent. For extended leave, it lists the name of substitutes.
8. Status
   The MP's presence or absence during the given period.

While the Swedish PoliGraph was created for the specific purpose of establishing the structure of parliamentary debates, it was designed to be detailed and flexible enough to be used outside of its planned scope.

Work on named entity recognition has also been initiated, with a number of speeches annotated for six different types of named entities:

1. People, real or fictional
2. Political roles, such as ministerial posts
3. Organizations
4. Locations
5. Works of art and culture, as well as brands
6. Time periods and points in time

These categories, as well as the annotation guidelines were derived from a SWE-CLARIN project that aimed to create a new gold standard for named entity recognition and classification in Swedish (Ahrenberg, Frid, and Olsson 2020). We did, however, choose to remove two of their categories – those pertaining to medical symptoms and treatments – as they were deemed very unlikely to show up in a significant number in the parliamentary debates. On the other hand, we added the category of political roles, in order to capture MPs who were not referred by name. Furthermore, we asked our annotators to designate whether a named person was a member of parliament or not, and whether organizations mentioned were political or not.

We are currently in the process of evaluating the classification methods used by SWE-CLARIN on our data, with the expectation that the Swedish BERT model developed by Kungliga Biblioteket (Malmsten, Börjeson, and Haffenden 2020). We will then proceed to automatically classify the remaining parliamentary debates and release both the manually and the automatically annotated data as a resource.

6 Conclusion

In this chapter we presented our ongoing efforts to create resources for studying argumentation and argumentation mining. As demonstrated here, the annotation of phenomena such as argumentation is complex and challenging. It needs to be carefully thought through, especially the evaluation of such annotations. However, these efforts enable studies from many angles and perspectives. As discussed in Hajičová et al. (2022) in this book, an annotated corpus can both be a resource for linguistic studies and open up new research questions.

The corpora we presented here could be useful for many types of studies aiming to analyse argumentation in the domains covered. Even though the purpose of the annotations have been for use in machine learning, it should be possible to use the annotations for other quantitative studies. For example, are there any specific patterns or words which are more frequent in argumentation than in non-argumentative exchanges? Are there any other patterns to be found, for example between speakers in a debate or users on an online forum?

Much of this chapter has focused on the complexity of argumentation and the disagreement between the annotators. A dataset where annotators disagree might not be the best for machine learning purposes, but it could be used to answer other questions. The disagreements themselves could be studied: are there any patterns to where the annotators agree or disagree? Could one annotator's annotations be easier for a machine learning algorithm to learn compared to the others?

The emergence of the NLP sub-field of argumentation mining has enabled new ways of researching argumentation. This field covers a wide range of possible and envisioned tasks, from argument component identification (Trautmann et al. 2020) to automatic evaluation of arguments or their claims (Sathe et al. 2020). Argumentation mining techniques would also be useful in information retrieval or as teaching aids. But for these tasks to be developed successfully, argumentation annotated corpora from a wide range of domains are essential (Stede and Schneider 2018).

As the annotated parts of the corpora presented here are currently small in size, as is the case for many argumentation corpora due to the challenging nature of the task, their usefulness as machine learning training data is still an open question. In recent years it has been become possible to use smaller amounts of training data due to the introduction of pre-trained language models and the possibility of fine-tuning them, but it still seems that larger amounts of training data is preferred. However, there exists other suggested solutions to the problem of data scarcity in argumentation mining. For example, a small corpus could be suitable for evaluation of unsupervised machine learning methods (Levy et al. 2017) or as a starter for boot-strapping more data (Ein-Dor et al. 2020).

Bibliography

Ahrenberg, Lars, Johan Frid & Leif-Jöran Olsson. 2020. A new resource for Swedish named-entity recognition. SLTC 2020, University of Gothenburg.

Ajjour, Yamen, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth & Benno Stein. 2017. Unit segmentation of argumentative texts. *Proceedings of the 4th Workshop on Argument Mining*, 118–128. Copenhagen: Association for Computational Linguistics.

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer & Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. *Proceedings of SLTC 2016*. Online: Umeå University.

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. *International Conference on Language Resources and Evaluation (LREC) 8*, 474–478. Istanbul: European Language Resources Association.

Budzynska, Katarzyna, Mathilde Janier, Chris Reed, Patrick Saint-Dizier, Manfred Stede & Olena Yaskorska. 2014. A model for processing illocutionary structures and argumentation in debates. *International Conference on Language Resources and Evaluation (LREC) 9*, 917–924. Reykjavik: European Language Resources Association.

Budzynska, Katarzyna & Chris Reed. 2011. Whence inference? Technical Report, University of Dundee.

Budzynska, Katarzyna & Chris Reed. 2019. Advances in argument mining. *Proceedings of the 57th annual meeting of the Association for Computational Linguistics: Tutorial abstracts*, 39–42. Stroudsberg, PA: Association for Computational Linguistics.

Eckart de Castilho, Richard, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank & Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In Erhard Hinrichs, Marie Hinrichs and Thorsten Trippel (eds.), *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 76–84. Osaka: COLING.

Eemeren, Frans H. van, Bart Garssen, Erik C. W. Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij & Jean H. M. Wagemans. 2014. Argumentation theory. *Handbook of argumentation theory*, 1–49. Dordrecht: Springer.

Eemeren, Frans H. van & Rob Grootendorst. 2003. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge: Cambridge University Press.

Eemeren, Frans H. van & Rob Grootendorst. 2010. *Speech acts in argumentative discussions*. Reprint 2010. Berlin: De Gruyter Mouton.

Ein-Dor, Liat, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou et al.. 2020. Corpus wide argument mining – a working solution. *The 34th AAAI Conference on Artificial Intelligence*, 7683–7691. New York: AAAI Press.

Fridlund, Mats, Daniel Brodén, Tommi Jauhiainen, Leena Malkki, Leif-Jöran Olsson & Lars Borin. 2022. Trawling and trolling for terrorists in the digital Gulf of Bothnia: Cross-lingual text mining for the emergence of terrorism in Swedish and Finnish newspapers, 1780-1926. In Darja Fišer & Andreas Witt (eds.), CLARIN. The infrastructure for language resources. Berlin: deGruyter.

Habernal, Ivan & Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics* 43 (1). 125–179.

Hajičová, Eva, Jan Hajič, Barbora Hladká, Jiří Mírovský, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, Pavel Straňák, Barbora Štěpánková & Šárka Zikánová. 2022. Corpus annotation as a feasible and scientifically beneficial task. In Darja Fišer & Andreas Witt (eds.), CLARIN. The infrastructure for language resources. Berlin: deGruyter.

Hedquist, Rolf. 1978. *Emotivt språk: En studie i dagstidningars ledare* [Emotive language: A study in newspaper editorials]. Umeå: Umeå University, Dept. of Nordic Languages.

Hidey, Christopher, Elena Musi, Alyssa Hwang, Smaranda Muresan & Kathleen McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In Ivan Habernal, Iryna Gurevych, Kevin Ashley, Claire Cardie, Nancy Green, Diane Litman, Georgios Petasis, Chris Reed, Noam Slonim and Vern Walker (eds.), *Proceedings of the 4th Workshop on Argument Mining*, 11–21. Copenhagen: Association for Computational Linguistics.

Janier, Mathilde, John Lawrence & Chris Reed. 2014. Ova+: an argument analysis interface. In Simon Parsons, Nir Oren, Chris Reed and Federico Cerutti (eds.), *Computational models of argument*, Frontiers in artificial intelligence and applications, 463–464. Amsterdam: IOS Press.

Landis, J. Richard & Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33 (1). 159–174.

Lapponi, Emanuele, Martin G. Søyland, Erik Velldal & Stephan Oepen. 2018. The talk of Norway: A richly annotated corpus of the Norwegian parliament, 1998–2016. *Language Resources and Evaluation* 52 (3). 873–893.

Lawrence, John & Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics* 45 (4). 765–818.

Levy, Ran, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov & Noam Slonim. 2017. Unsupervised corpus–wide claim detection. *Proceedings of the 4th Workshop on Argument Mining*, 79–84. Stroudsberg, PA: Association for Computational Linguistics.

Lindahl, Anna. 2020. Annotating argumentation in Swedish social media. In Elena Cabrio and Serena Villata (eds.), *Proceedings of the 7th Workshop on Argument Mining*, 100–105. Online: Association for Computational Linguistics.

Lindahl, Anna, Lars Borin & Jacobo Rouces. 2019. Towards assessing argumentation annotation – a first step. In Benno Stein and Henning Wachsmuth (eds.), *Proceedings of the 6th Workshop on Argument Mining*. Stroudsberg, PA: Association for Computational Linguistics.

Lippi, Marco & Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)* 16 (2). 10.1–10.25.

Lytos, Anastasios, Thomas Lagkas, Panagiotis Sarigiannidis & Kalina Bontcheva. 2019. The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management* 56 (6). 102055.

Malmsten, Martin, Love Börjeson & Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT. Preprint: https://arxiv.org/abs/2007.01658 (accessed 23 March 2022).

Mann, William C. & Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text & Talk* 8 (3). 243–281.

Miller, Tristan, Maria Sukhareva & Iryna Gurevych. 2019. A streamlined method for sourcing discourse-level argumentation annotations from the crowd. In Jill Burstein, Christy Doran and Thamar Solorio (eds.), *Proceedings of NAACL 2019*, 1790–1796. Stroudsberg, PA: Association for Computational Linguistics.

Morante, Roser, Chantal Van Son, Isa Maks & Piek Vossen. 2020. Annotating perspectives on vaccination. *International Conference on Language Resources and Evaluation (LREC) 12*, 4964–4973. Marseille: European Language Resources Association.

Nanni, Federico, Mahmoud Osman, Yi-Ru Cheng, Simone Paolo Ponzetto & Laura Dietz. 2018. UKParl: A semantified and topically organized corpus of political speeches. In Darja Fišer, Maria Eskevich and Franciska de Jong (eds.), *Proceedings of the LREC 2018 Workshop on Creating and Using Parliamentary Corpora*, 29–32. Miyazaki: European Language Resources Association.

Pančur, Andrej, Mojca Šorn & Tomaž Erjavec. 2018. SlovParl 2.0: The collection of Slovene parliamentary debates from the period of secession. In Darja Fišer, Maria Eskevich and Franciska de Jong (eds.), *Proceedings of the LREC 2018 Workshop on Creating and Using Parliamentary Corpora*, 8–14. Miyazaki: European Language Resources Association.

Park, Joonsuk & Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In Nancy Green, Kevin Ashley, Diane Litman, Chris Reed and Vern Walker (eds.), *Proceedings of the 1st workshop on argumentation mining*, 29–38. Stroudsberg, PA: Association for Computational Linguistics.

Reed, Chris & Glenn Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools* 13 (04). 961–979.

Rødven-Eide, Stian. 2019. The Swedish PoliGraph: A semantic graph for argument mining of Swedish parliamentary data. In Benno Stein and Henning Wachsmuth (eds.), *Proceedings of the 6th Workshop on Argument Mining*, 52–57. Stroudsberg, PA: Association for Computational Linguistics.

Rødven-Eide, Stian. 2020. Anföranden: Annotated and augmented parliamentary debates from Sweden. In Darja Fišer, Maria Eskevich and Franciska de Jong (eds.), *Proceedings of the 2nd ParlaCLARIN Workshop*, 5–10. Marseille: European Language Resources Association.

Rosenthal, Sara & Kathleen McKeown. 2012. Detecting opinionated claims in online discussions. *2012 IEEE 6th International Conference on Semantic Computing*, 30–37. Palermo: IEEE.

Sathe, Aalok, Salar Ather, Tuan Manh Le, Nathan Perry & Joonsuk Park. 2020. Automated fact-checking of claims from Wikipedia. *International Conference on Language Resources and Evaluation (LREC) 12*, 6874–6882. Marseille: European Language Resources Association.

Stab, Christian & Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics* 43 (3). 619–659.

Stede, Manfred & Jodi Schneider. 2018. Argumentation mining. *Synthesis Lectures on Human Language Technologies* 11 (2). 1–191.

Toulmin, Stephen E.. 2003. *The uses of argument*. 2nd edition. Cambridge: Cambridge University Press.

Trautmann, Dietrich, Johannes Daxenberger, Christian Stab, Hinrich Schütze & Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. *AAAI Conference on Artificial Intelligence* 34 (1). 9048–9056.

Walton, Douglas. 1996. *Argumentation schemes for presumptive reasoning*. Mahwah: Lawrence Erlbaum Associates.

Walton, Douglas, Christopher Reed & Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge: Cambridge University Press.