Towards Assessing Argumentation Annotation – A First Step

Anna Lindahl
Språkbanken Text
University of Gothenburg
Sweden

Lars Borin
Språkbanken Text
University of Gothenburg
Sweden

Jacobo Rouces Gonzalez
Språkbanken Text
University of Gothenburg
Sweden

(anna.lindahl|lars.borin|jacobo.rouces)@svenska.gu.se

Abstract

This paper presents a first attempt at using Walton's argumentation schemes for annotating arguments in Swedish political text and assessing the feasibility of using this particular set of schemes with two linguistically trained annotators. The texts are not pre-annotated with argumentation structure beforehand. The results show that the annotators differ both in number of annotated arguments and selection of the conclusion and premises which make up the arguments. They also differ in their labeling of the schemes, but grouping the schemes increases their agreement. The outcome from this will be used to develop guidelines for future annotations.

1 Introduction

Argumentation mining – the automatic recognition and classification of arguments and their components in text – is a useful technology for a number of practical text-processing applications, both commercial and academic, and in the latter case not least as a component of research tools in the digital humanities and social sciences.

Many different annotation schemes for argument analysis have been proposed in the literature (Lippi and Torroni, 2016; Macagno et al., 2017; Visser et al., 2018; Song et al., 2014), and a central concern in the context of argumentation mining is to arrive at a scheme which is both expressive enough for the intended tasks and explicitly defined in a way which makes it amenable to high-accuracy automatic processing.

Automatic linguistic annotation often requires the use of a ground-truth data set – a gold standard – for evaluating – and often training – different kinds of algorithms and software. Since the gold standard annotations will invariably need to be introduced by humans, we require an annotation scheme which human annotators can learn (in a reasonable amount of time) to apply with high accuracy and high inter-annotator agreement.

One of the most elaborate and extensive efforts to devise a comprehensive set of argumentation schemes is that by Walton et al. (2008), which builds on a long line of works in philosophy and law studies. Walton et al. (2008) further explicitly intend their schemes to be usable "in AI". The 60 schemes (with additional sub-schemes in many cases) presented in the book are given detailed, formalized descriptions, and in the present paper we describe and discuss the initial stage in an effort intended to evaluate the suitability and usefulness of this set of schemes for argumentation mining.

As indicated above, a prerequisite for this is that a sufficient amount of suitable text can be manually annotated with high inter-annotator agreement. Consequently, we have initiated an annotation effort (the first of several), where a small set of Swedish political texts (newspaper editorials) have been annotated using the schemes of Walton et al. (2008). To the best of our knowledge, this is the first annotation study which applies Walton's schemes directly to text, without any pre-annotated structure step beforehand. In the present paper, we present and discuss the results of this exercise, and outline what the next steps of this effort should be, based on these results.

Related Work

In Walton et al. (2008) an argumentation scheme is defined by a set of premises and a conclusion, and a label for the scheme. For most schemes, there is also a set of critical questions which are used for identification and evaluation. Walton's schemes, or modified versions of them, have been used to classify argumentation in many cases (Feng and Hirst, 2011; Green, 2015; Song et al., 2014; Lawrence and Reed, 2016). However, when annotating argumentation schemes, in these cases the annotation has been done on already pre-

Proceedings of the 6th Workshop on Argument Mining, pages 177–186
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

segmented text, identified as containing argumentation.

Visser et al. (2018) use Walton's original schemes for annotating nodes in an argumentation structure. They reach an inter-annotator agreement of $\kappa = 0.723$ (Cohen's Kappa), but note that there are some schemes that are difficult for the annotators to distinguish, despite the use of a decision tree based on Walton et al. (2008). The issue of distinguishing schemes and the need for a taxonomy or classification of the schemes have been also been discussed in Walton (2012), and there have been many suggestions for this (Walton et al., 2008; Walton and Macagno, 2015; Macagno et al., 2017). Because of this, in addition to using the original schemes, we also use groups suggested in the classification system mentioned in Walton et al. (2008).

2 Data Set Creation and Annotation

The Data Set

The data for this study were originally compiled by Hedquist (1978), who investigated emotive language[1] in Swedish newspaper editorials. He selected a total of 30 editorials from 6 newspapers, all published in the period May–September 1973, shortly before the Swedish national parliament elections at the end of September 1973. The newspapers were selected so as to reflect the five political parties then represented in the Swedish parliament, and the editorials were selected on the basis of topic, with two general and three specific topics per newspaper. The total number of words in this data set is about 19,000, for an average word count per editorial of about 640.

For his investigation Hedquist annotated all texts manually for emotive language, using a scheme which he developed specifically for this work. Together with the existing and planned argumentation annotation described in this paper, this data set comprises a small but rich foundation for future work on argumentation mining in Swedish in particular, but also in more general terms on the relationship between argument structures and sentiment.

Annotation Procedure

The editorials were annotated by two annotators with solid training in linguistic analysis, master students of linguistics at Uppsala University.

The instructions given to the annotators were minimal. In preparation for the annotation task, they were initially given three editorials, asked to identify and classify all arguments in them manually according to Walton et al. (2008). After this they met with the project leader, for a discussion of differences and difficulties. Other than that, they were expected to be able to understand the description of the argumentation schemes as given by Walton et al. (2008), as it was believed that somebody with their extensive training in linguistics should be well equipped to understand and apply these descriptions, which are couched in terms quite familiar to somebody who has been exposed to linguistic semantics and pragmatics.

The annotation was done with the Araucaria tool for argument analysis (Reed and Rowe, 2004) which has support for Walton's argument schemes. For the annotation, the 30 most common schemes were used, as originally presented in Walton (2013). In Araucaria, for a given text, the annotator selects any consecutive passage of text, labels it and possibly connects it to any other labeled passage of text. From here on, these passages are referred to as *units*. The available labels are 'premise' and 'conclusion'. A premise can only have one conclusion, but a unit can be annotated multiple times. This is suitable for chained argumentation. After labeling, an argument scheme is connected to a conclusion and one or more premises, and these parts together make up the argument. Araucaria also allows adding so-called 'missing' units if an annotator feels a conclusion or premises are left unstated/implied.

3 Results

The results from the two annotators differ significantly, both regarding what is annotated and how it has been annotated. More specifically, they differ both in numbers of arguments annotated and the distributions of units, and they even differ in how they use the annotation tool, which results in different structure of the file containing the argument

[1] The phenomena investigated and described by Hedquist largely come under the heading of what is now generally referred to as *sentiment analysis*.

	Annotator 1	Annotator 2
No. of arguments	345	195
Avg. no. of premises per arg.	1.26	2.03
Premises, in text	395	380
Premises, missing	42	16
Conclusions, in text	292	194
Conclusions, missing	53	1
Total no. of units	782	591

Table 1: Annotation statistics

information, although the retrieved information is the same.

The number of annotated argument instances and units is shown in Table 1. Annotator 1 (A1) has annotated about 150 more argument instances than annotator 2 (A2), although the latter has annotated more premises on average. By inspection it was observed that A1 often pairs a conclusion with each of its premises into individual arguments with one conclusion and one premise each, but assigns all of these arguments the same scheme. A2 usually includes all premises attached to a conclusion as a single argument. This could be either a difference in interpretation or usage of the tool, but this may be the reason for the difference in the average number of premises.

The annotators have used the option of adding missing units differently, with A1 having added about 100 missing units and A2 17 as shown in Table 1. The identification of implied conclusions or premises is a well-known problem, and might be the reason for this discrepancy. In Table 2 the statistics of multiple occurrences are shown. A1 has both more units repeating as conclusions, and occurring as both premise and conclusion. On the other hand, A2 has 26 repeating premises while A1 has none. Most of A2's occurrences are only repeating once, but A1 has many conclusions which occur many times. This is related to the difference in how the annotators divide the premises between arguments. If a conclusion has 6 premises and A1 turns each conclusion-premise pair into a separate argument, then the conclusion will occur 6 times.

Of the 30 schemes described by Walton (2013), A1 uses 12 and A2 uses 21. Together they use 22 different schemes.[2] Both annotators use 4–5 schemes for the majority of identified arguments, with the rest of the schemes having only a few occurrences each. Even though A1 annotates more argument instances, fewer schemes are used. Table 3 shows the the used schemes and their occurrences for A1 and A2. The schemes ARGUMENT FROM CONSEQUENCES and ARGUMENT FROM SIGN are both heavily used by both annotators. The description of these schemes are seen below.

ARGUMENT FROM SIGN:
Premise: A is true in this situation.
Premise: Event B is generally indicated as true when its sign, A, is true in this kind of situation.
Conclusion: B is true in this situation.

ARGUMENT FROM CONSEQUENCES:
Premise: If A is brought about, then good (bad) consequences will (may plausibly) occur
Conclusion: A should (not) be brought about.

From these descriptions it is seems that these schemes could be applied to a wide range of argumentation, and this is probably why the annotators have used them the most. Compared to some of the descriptions of the other schemes, they are also possibly easier to understand and therefore easier to apply. But they are also very general, and this raises the question in which cases an annotator chooses the more specific scheme in favor of a more general one. Interestingly, the scheme A1 annotated the most (ARGUMENT FROM EVIDENCE TO A HYPOTHESIS) is only used 6 times by A2. Likewise, A2's most annotated scheme (ARGUMENT FROM CORRELATION TO CAUSE) is only used 5 times by A1. The descriptions of these two schemes are seen below. These schemes both describe correlation between events, and one could possibly see the first as a subset of the second. The similarities of the schemes are further explored in the next section.

[2] 23 of the annotated units of A1 are not marked with an argument scheme and are thus not included.

Annotator 1		Annotator 2	
Units as both conclusion and premise	72	Units as both conclusion and premise	12
Units as repeating conclusion	80	Units as repeating conclusion	7
Units as repeating premises	0	Units as repeating premises	26

Table 2: Occurrences of units

Scheme	A1	A2
Argument from Evidence to a Hypothesis	105	6
Argument from Consequences	90	20
Argument From Sign	47	22
Argument from Cause to Effect	30	18
Argument from Falsification of a Hypothesis	30	4
Argument from Commitment	11	3
Argument from Verbal Classification	9	15
Argument from Expert Opinion	8	7
Argument from Popular Opinion	7	12
Argument from Correlation to Cause	5	42
Argument from Analogy	2	1
Ethotic Argument	1	–
Argument from Popular Practice	–	17
Argument from Position to Know	–	8
Argument from Bias	–	5
Causal Slippery Slope Argument	–	4
Argument from Precedent	–	3
Argument from an Established Rule	–	2
Argument from Arbitrariness of a Verbal Classification	–	2
Circumstantial Argument Against the Person	–	2
Argument from Vagueness of a Verbal Classification	–	1
Argument from an Exceptional Case	–	1
Total	195	345

Table 3: Usage of schemes for Annotator 1 and 2

ARGUMENT FROM EVIDENCE TO A HYPOTHESIS:
Premise: If hypothesis A is true, then a proposition B, reporting an event, will be observed to be true.
Premise: B has been observed to be true in a given instance .
Conclusion: A is true.

ARGUMENT FROM CORRELATION TO CAUSE:
Premise: There is a positive correlation between A and B.
Conclusion: A causes B.

4 Evaluation

In order to measure inter-annotator agreement (IA) we use the measure in Equation 1 based on the Sørensen–Dice coefficient, where a_1 and a_2 are the sets of annotations from each annotator, and m is the set of pairs of annotations from a_1 and a_2 that are matching (i.e. they are considered equivalent). Annotations can be either units (spans of text representing premises or conclusions) or arguments (a conclusion with one or more spans).

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

We don't use measures such as Fleiss' kappa or Krippendorff's alpha because these measures calculate agreement over annotation tasks that consist of assigning a discrete label or score to each element in a set, which is different to annotating spans over continuous text. Previous work on argumentation annotation such as as in Stab and Gurevych (2017) uses them because their annotation task is defined as marking whether predefined spans of texts do or do not contain annotations or units, but in our annotation task the annotators themselves create the spans.

To determine if two units are matching, the amount of overlap between the strings representing the units is compared to a given threshold α. The strings are defined as ranges of character indices within the text. The amount of overlap is measured as the ratio between the length of the longest common continuous substring to both strings and the length of the longest of both strings. For example, the units below have an overlap of 0.68.

Unit 1. *7:48 Utgången kan leda till regeringsbyte, men den kommer inte att leda till någon förändring av trygghetspolitiken i det svenska välfärdssamhälllet.*[3]
'The result might lead to a change of government, but it will not lead to any change in the Swedish welfare state.'
Unit 2. *den kommer inte att leda till någon förändring av trygghetspolitiken i det svenska välfärdssamhälllet.*
'it won't lead to any change in the Swedish welfare state.'

Two values of α are used in the experiments. A strict one of 0.9, which can still account for small differences in whitespace, and a more lenient threshold of 0.5. In order to compare how well the annotators agree, the arguments are compared unit by unit. First, the conclusions of the arguments are compared, and if the conclusion matches, the premises are compared. Given both a matching conclusion and premise, the schemes of the two matching arguments are compared. If a unit occurs more than once, it will belong to different arguments. Each occurrence is thus treated as a unique occurrence.

Conclusions

In Table 4 the number of matching conclusions is shown. The IA is calculated as per Equation 1, and is 0.26 for an α of 0.9. The average number of matching conclusions per editorial is 2.37, with two editorials having no matches and one having seven matches.

	α	
Conclusions	0.9	0.5
m	71	92
IA	0.26	0.34

Table 4: IA and m for conclusions.

Premises

Given a matching conclusion between two arguments, the premises of the same arguments are compared. Since the number of premises in an argument can vary between the annotators, both matches with all premises matching and at least one is displayed in Table 5. With the full overlap α, used for both premises and conclusions, the IA is 0.56 for at least one matching premise. With the same α, only 6 of the matching conclusions have all premises matching. Using the 0.5 α, the IA is 0.71 for at least one matching premise, and 0.20 all premises matching. The IA within all arguments is low for both α.

	α	
At least one matching premise	0.9	0.5
m	20	33
IA, within matching conclusions	0.56	0.71
IA, within all arguments	0.07	0.12
All premises match		
m	6	9
IA, within matching conclusions	0.17	0.20
IA, within all arguments	0.02	0.03

Table 5: IA and m for premises, given a matching conclusion.

It is important to note that even if two arguments have a matching conclusion this does not necessarily mean that they should have the same premises, a conclusion can be reached through different premises and argumentation. This could explain why there are 71 matching conclusions, but only 20 of them share at least one premise. An example of this can be seen below:

Premise A1: *den visar sig redan i form av kraftiga höjningar av olje- och bensinpriserna.* 'It is already showing in the form of increasing oil and gas prices.'
Premise A2: *Vi är i det här landet inte särskilt vana att spara på något.* 'We are not especially used to saving anything in this country.'
Conclusion: *Men nu är energikrisen inte långt borta*
'But now the energy crisis is not far away'
Scheme A1: ARGUMENT FROM SIGN
Scheme A2: ARGUMENT FROM CAUSE TO EFFECT

In the same way, a premise can be used for different conclusions. Table 6 shows the matching premises, regardless of whether they have a matching conclusion or not. There are 14 arguments

[3] The number at the beginning of the sentence is sentence numbering present in the source of the texts.

	α	
At least one premise match	0.9	0.5
m	74	99
IA, within all arguments	0.27	0.37
All premises match		
m	14	20
IA, within all arguments	0.05	0.07

Table 6: IA and m for only premises.

where all premises match. Of these 14 matches, three have also a matching argumentation scheme. This means that even if the premises match, there is disagreement about which scheme they participate in. The two following examples show this. The first example is a match in both conclusion and premises, but the schemes differ. The next example has the same premise but different conclusion and scheme. This indicates that a premise can be used for different schemes, and result in different conclusions.

Premise: *Den är inte obegränsad*
'It is not unlimited.'
Conclusion: *Allmänt sett är det nödvändigt att hushålla med energin*
'It is widely considered necessary to economize energy.'
Scheme A1: ARGUMENT FROM CONSEQUENCES
Scheme A2: ARGUMENT FROM SIGN

Premise: *En växling vid makten medför att vi inte riskerar några socialistiska experiment under valperioden utan kan bygga vidare på välfärdssamhällets grund.*
'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: *Väljare bör rösta på oppositionen*
'Voters should vote for the opposition'
Conclusion A2: *Rösta inte bort samverkan!*
'Do not vote away collaboration!'
Scheme A1: ARGUMENT FROM CONSEQUENCES
Scheme A2: CAUSAL SLIPPERY SLOPE ARGUMENT

Argument schemes

After finding which arguments match in conclusion and premise, the argumentation schemes are compared. Using 0.9 as the α, only 2 arguments have a match in scheme, conclusion and premises. The schemes in these two arguments are ARGUMENT FROM SIGN and ARGUMENT FROM CAUSE TO EFFECT. Using 0.5 as α instead, there are 4 matches. Three of them have only 1 premise and they all overlap fully. The last one has half of the premises matching.

Based on the low numbers of matching schemes in the case where both conclusion and premise match, conclusions and premises were compared separately. Of all the matching conclusions, 9 have the same scheme, see Table 7. Figure 1 shows how the schemes co-occur when the conclusion is the same. The schemes that match are the schemes which are most commonly used by both annotators.

	α	
Scheme matches	0.9	0.5
m	9	10
IA, within matching conclusion	0.25	0.22
IA, within all arguments	0.02	0.02

Table 7: IA and m for schemes, given a matching conclusion.

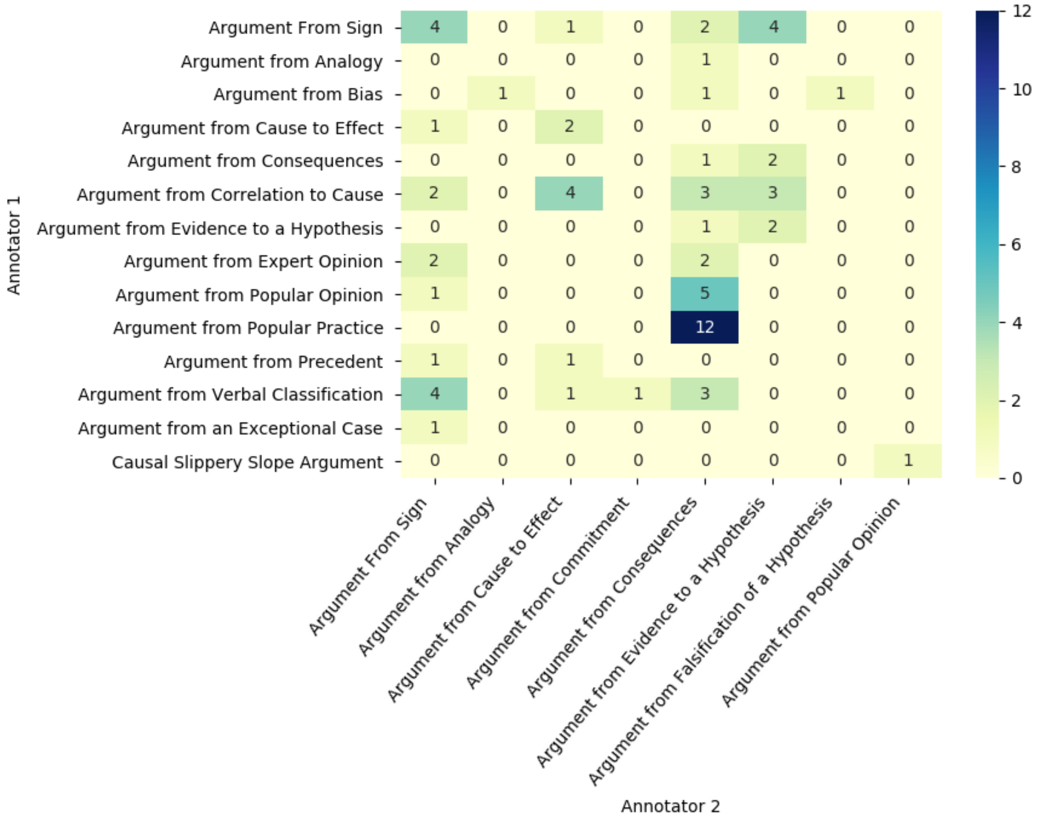
In Figure 1 we can see that ARGUMENT FROM CONSEQUENCES and ARGUMENT FROM POPULAR PRACTICE have a high co-occurrence, compared to the others. This could be because the annotators have have chosen different premises, for the same conclusion and thus chosen different schemes. The descriptions of these schemes are shown below.

ARGUMENT FROM CONSEQUENCES:

Premise: If A is brought about, then good (bad) consequences will (may plausibly) occur.
Conclusion: A should (not) be brought about.

ARGUMENT FROM POPULAR PRACTICE:
Premise: If a large majority (everyone, nearly everyone, etc.) does A, or acts as though A is the right (or an acceptable) thing to do, then A is a prudent course of action.
Premise: A large majority acts as though A is the right thing to do.
Conclusion: A is a prudent course of action.

It seems that the difference between these schemes is dependent on the reason for a proposed action. Should it be done because there is a desired outcome (Consequences) or is the right thing to do because it is a popular practice? An example of this disagreement is seen below. Possibly this example could be argued to be both schemes.



	Argument From Sign	Argument from Analogy	Argument from Cause to Effect	Argument from Commitment	Argument from Consequences	Argument from Evidence to a Hypothesis	Argument from Falsification of a Hypothesis	Argument from Popular Opinion
Argument From Sign	4	0	1	0	2	4	0	0
Argument from Analogy	0	0	0	0	1	0	0	0
Argument from Bias	0	1	0	0	1	0	1	0
Argument from Cause to Effect	1	0	2	0	0	0	0	0
Argument from Consequences	0	0	0	0	1	2	0	0
Argument from Correlation to Cause	2	0	4	0	3	3	0	0
Argument from Evidence to a Hypothesis	0	0	0	0	1	2	0	0
Argument from Expert Opinion	2	0	0	0	2	0	0	0
Argument from Popular Opinion	1	0	0	0	5	0	0	0
Argument from Popular Practice	0	0	0	0	12	0	0	0
Argument from Precedent	1	0	1	0	0	0	0	0
Argument from Verbal Classification	4	0	1	1	3	0	0	0
Argument from an Exceptional Case	1	0	0	0	0	0	0	0
Causal Slippery Slope Argument	0	0	0	0	0	0	0	1

Figure 1: Co-occurrence matrix for the schemes with the same conclusion (α 0.9)

Premise: *Den höga arbetslösheten i Sverige är inte acceptabel ur några synpunkter, detta måste slås fast med skärpa.* 'The high unemployment rate in Sweden is not acceptable from any angle, this must be firmly established.'
Conclusion: *Att skaffa fram nya jobb, måste vara den viktigaste uppgiften för närvarande.* 'To create new jobs must be the most important task for now.'
Scheme A1: ARGUMENT FROM CONSEQUENCES
Scheme A2: ARGUMENT FROM POPULAR PRACTICE

As mentioned above, matching premises were also compared, regardless of conclusions. One could expect this to generate more scheme matches, as similar premises would possibly be used in similar kinds of argumentation. However, as noted in the previous section, of all the 540 arguments, only 14 have all premises matching. Out of these, only 3 have the same scheme, as compared to 9 scheme matches for the conclusions.

Because of the noted difficulty of distinguishing the schemes, both here and in previous research, and the low number of matches, the schemes were divided into groups and these groups were compared instead. This division is suggested by Walton et al. (2008) as a classification system for the schemes.

	α	
Matching schemes	0.9	0.5
m	3	7
IA, within matching	0.08	0.15
IA, within all arguments	0.01	0.03
Abductive reasoning	2	5
Casual reasoning	1	1
Practical reasoning	0	1

Table 8: Matching schemes with the new groups of schemes, given a matching conclusion and at least one premise.

Using the new groups results in more matching schemes, but still the numbers are low for a match of both conclusion and premise, see Table 8. Table 9 shows the same numbers but for only conclusions. The co-occurrence matrix is again showed for an α of 0.9 and only conclusions, see Figure 2. Most noteworthy are the 10 matches in the ABDUCTIVE REASONING group and the 17 co-occurrences between the groups ARGUMENTS FROM POPULAR PRACTICE and PRAC-

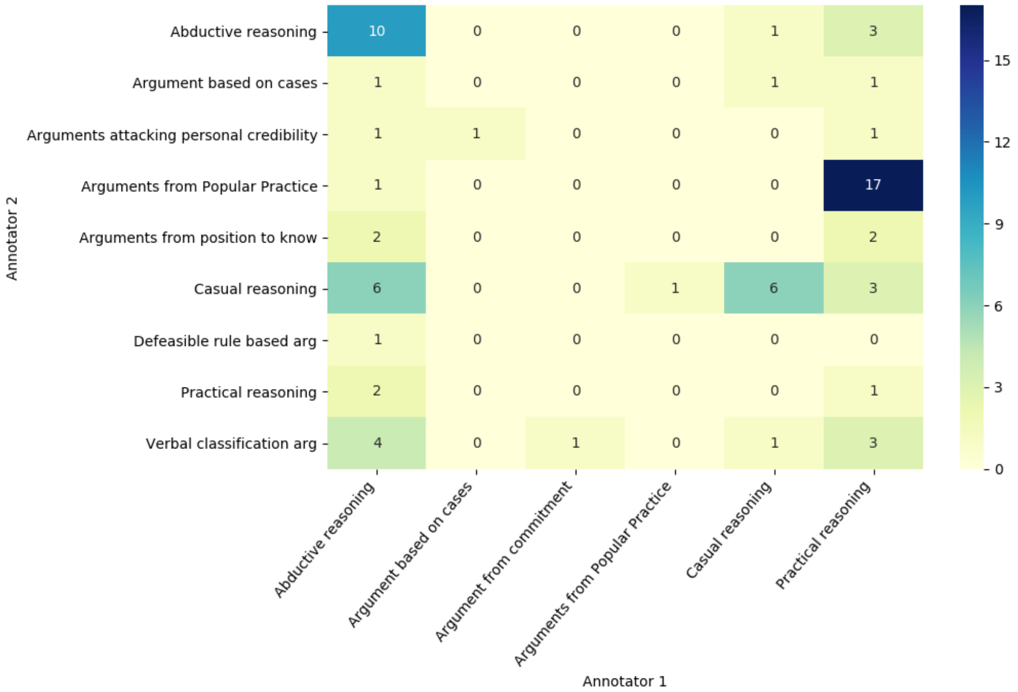


Figure 2: Co-occurrence matrix for the schemes in new groups, with the same conclusion (α 0.9)

	α	
Matching schemes	0.9	0.5
m	17	20
IA, within matching conclusions	0.48	0.43
IA, within all arguments	0.06	0.07
Abductive reasoning	10	12
Casual reasoning	6	6
Practical reasoning	1	2

Table 9: Matching schemes with the new groups of schemes, only conclusions.

TICAL REASONING. This mismatch in groups is due to the previously discussed co-occurrence of the schemes ARGUMENT FROM CONSEQUENCES and ARGUMENT FROM POPULAR PRACTICE. The former scheme is in the group PRACTICAL REASONING and the latter scheme is in the ARGUMENTS FROM POPULAR PRACTICE group, thus transferring the co-occurrences to the new groups.

Again, a comparison of the schemes in the new groups but for only matching premises was done. This however only led to 4 scheme matches and no pattern in the co-occurrences.

5 Conclusions and Future Work

In this first annotation exercise, we wanted to investigate whether annotators with a strong background in linguistics but who were given little explicit instruction for this specific annotation task would be able to recover the argumentation schemes described by Walton et al. (2008). This turned out not to be the case, with the annotators agreeing neither on whole arguments nor on the units and schemes which make them up.[4] This could be for at least three reasons: (1) that the annotators would have needed more detailed and precise instructions; (2) that the argumentation schemes themselves are too difficult to recover from free natural text (despite their seeming formal characterization); or (3) that the annotation task should be structured differently, in a first step where spans representing argument instances are identified followed by a second step where the instances and their components are labeled.[5]

Some of the differences between the annotators would have been avoided if they had more spe-

[4] The use of only two annotators possibly influenced the result, making it difficult to conclude when we are dealing with 'normal' disagreement or not.

[5] It was suggested by the anonymous reviewers that this would make for more effective annotation and higher inter-annotator agreement. We are not aware of any strong arguments in the literature unequivocally supporting this view, nor of any empirical studies comparing the end-to-end efficiency and efficacy of these two annotation workflows while controlling for other potentially relevant variables. We note this as an interesting topic for future research.

cific instructions for the tool. More strict and detailed instructions for the annotation itself could probably improve the inter-annotator agreement, but might come with a loss of information. For example, a rule such as marking sentences instead of spans would result in some loss of information, since an argument might not be restricted to sentences. However, most of the disagreements come from differences in the interpretation of argument components and schemes, as shown in the examples in the previous section. For example, the same premises and conclusions are used in different schemes, and a single premise is used more than one scheme. In order to minimize information loss but achieve high inter-annotator agreement a necessary next step in annotating argumentation needs to be a discussion of what should be marked as premises and conclusions and why the annotators have made the choices they did.

Interpretation seems also to be the reason for the difference in the annotation of the argumentation schemes, although the low inter-annotator agreement in the argument components evaluated before the schemes might influence this. If the annotators were given already annotated units they would possibly agree more. The results of Visser et al. (2018) indicate this, where they use already predefined nodes and reach a high inter-annotator agreement.

As previously shown, and also observed by others (Walton and Macagno, 2015; Macagno et al., 2017), the original schemes can be difficult to distinguish from each other. If they are to be used by annotators, then they need better instructions on when to use which scheme. As the post-annotation grouping of schemes improved agreement, perhaps it would be effective to collapse them already in advance, instructing annotators to use coarser groupings in cases of doubt.

For the immediate future we plan to design two annotation exercises to follow up on the experiment described in this paper and to address some of the questions raised above. Further, the exercises will be carried out using two different annotation workflows. In the first exercise, the annotators will be instructed to use the schemes of Walton et al. (2008), but this time according to an explicit annotation manual. In the second exercise the annotators will be asked to annotate the same texts according to some other proposed scheme, possibly a less fine-grained version of the original schemes as this was shown to have a positive effect on the inter-annotator agreement, but the exact scheme to be used remains to be determined. We will also organize two versions of each exercise, one corresponding to the previous annotation round, where annotatoras are asked to identify argumentation spans and classify them in one operation, and another where argumentation span identification is separated from labeling of schemes and components.

In all cases we plan to employ more than two annotators and there will be a different set of annotators for each of the four annotation setups. The texts to be annotated will include the editorials used for the work described in this paper, but we may also decide to extend the data set. Hopefully, the planned experiments will allow us to gain a better understanding of the advantages and disadvantages of different schemes for argumentation annotation, as well as for alternative organizations of the annotation workflow.

Acknowledgments

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 would also like to thank the anonymous reviewers for their detailed and constructive comments.

References

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 987–996.

Nancy Green. 2015. Identifying argumentation schemes in genetics research articles. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 12–21. Association for Computational Linguistics.

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

John Lawrence and Chris Reed. 2016. Argument mining using argumentation scheme structures. In *COMMA*, pages 379–390.

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

Fabrizio Macagno, Douglas Walton, and Chris Reed. 2017. Argumentation schemes. history, classifications, and computational applications. *History, Classifications, and Computational Applications (December 23, 2017). Macagno, F., Walton, D. & Reed, C*, pages 2493–2556.

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

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78.

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

Jacky Visser, Lawrence John, Wagemans Jean, and Reed Chris. 2018. Revisiting computational models of argument schemes: Classification, annotation, comparison. In *Proceedings of the 7th International Conference on Computational Models of Argument (COMMA 2018)*.

Douglas Walton. 2012. Using argumentation schemes for argument extraction: A bottom-up method. *IJCINI*, 6:33–61.

Douglas Walton. 2013. *Argumentation schemes for presumptive reasoning*. Routledge.

Douglas Walton and Fabrizio Macagno. 2015. A classification system for argumentation schemes. *Argument & Computation*, 6(3):219–245.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press, Cambridge.