Towards assessing depth of argumentation

Manfred Stede
Applied Computational Linguistics
UFS Cognitive Sciences
University of Potsdam / Germany
stede@uni-potsdam.de

Abstract

For analyzing argumentative text, we propose to study the 'depth' of argumentation as one important component, which we distinguish from argument quality. In a pilot study with German newspaper commentary texts, we asked students to rate the degree of argumentativeness, and then looked for correlations with features of the annotated argumentation structure and the rhetorical structure (in terms of RST). The results indicate that the human judgements correlate with our operationalization of depth and with certain structural features of RST trees.

1 Introduction

In recent years, *argumentation mining* has emerged as a discipline that aims to identify portions of argumentation in text and moreover – in some of the work – to relate them to one another, so that the full structure of arguments is being represented. A natural application for this new field is the mining of arguments in customer reviews (e.g., (Villalba and Saint-Dizier, 2012)), where the goal is to move beyond finding positive and negative statements by also finding the reasons that customers provide for their judgements. But the general task can be applied in many other domains as well. For example, the early research of Mochales Palau and Moens (2009) sought to identify premises and conclusions in legal text; another early work on finding theses in student essays (Burstein et al., 2003) has recently been extended to detecting more argumentative structure in such essays (Stab and Gurevych, 2014), (Nguyen and Litman, 2015); other genres that are being tackled include online dialogue (Oraby et al., 2015), multi-party dialogue (Budzynska et al., 2013) and scientific papers (Kirschner et al., 2015).

Despite the manifold recent activities, the field is still young, and a number of basic issues still require attention. This concerns the design of annotation schemes and possibly finding a consensus on them, but also the more fundamental problem of defining the notion of, and identifying instances of, *argumentative text*. While it is tempting to assume that earlier work on text-level subjectivity classification can be used to determine whether some (portion of) text is argumentative, we argue below that this is generally not the case. Going further, we posit that 'argumentativeness' is a matter of degree, and that determining this degree should also be considered a subtask of argumentation mining. We use the genre of newspaper editorials to conduct a small study where human raters are asked to assess how 'argumentative' different texts are, or in other words, what their 'argumentative depth' is. We will defend the introduction of this term by demonstrating that depth is to be distinguished from argument *quality*, which is already an object of study in the community.

Our next task then is to explain how such judgements arise, i.e., to find features that differentiate texts of different argumentative depths. We show that some simple surface features are not sufficient, and then turn to the underlying pragmatics, as it is – to some extent – captured in analyses according to Rhetorical Structure Theory (RST; (Mann and Thompson, 1988)). It was designed as a tool to make the reasons for a text's coherence explicit, and the specific notion of coherence relation used in RST (as opposed to other approaches such as the Penn Discourse Treebank (Prasad et al., 2008) or Segmented Discourse Representation Theory (Asher and Lascarides, 2003)) is decidedly *intentional*. This makes RST a good

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers,
pages 3308–3317, Osaka, Japan, December 11-17 2016.

candidate for checking whether its text representations are able to predict argumentative depth. The corpus we are using in our study has already been annotated with RST trees in earlier work, so we can use this data when looking for correlations with argumentation structure.

The paper is structured as follows. Section 2 introduces the notion of 'argumentativeness' and compares it to that of 'subjectivity', and introduces our approach to characterising the depth of argumentation. Section 3 describes the newspaper corpus we are using, and Section 4 presents the pilot study on annotating and measuring depth. Finally, Section 5 provides a summary and an outlook on future work.

2 Subjectivity versus Argumentation

To define our terminology, we start from the notion of 'text type' (Werlich, 1975) or 'discourse mode' (Smith, 2003), which posits that a passage of text can be of one of the following types: narrative, descriptive, instructive, expository, or argumentative. These classes refer to the central purpose of the passage, and – as especially Smith has shown – they correlate with certain linguistic features (aspect, Aktionsart, tense progression, etc.). The function of argumentative text, in general, is to influence the beliefs or attitudes of the readers by assembling a constellation of propositions that serve to defend a particular standpoint on a controversial issue (cf. (van Eemeren et al., 1996)).

'Subjectivity' in linguistics encompasses several different phenomena, which all serve to distinguish it from the 'objective' statement, but can do so in quite different ways, such as making speculations on future events, reporting on one's feelings, attributing content to third parties or to hearsay, or evaluating some entity in terms of valence (positive/negative). The latter is the notion that is mostly used in Computational Linguistics, and it is the central target of most work on subjectivity classification. This task can be applied on sentence or text level, and it has been done with bag-of-words models, PoS tags, or linguistic features that tend to be more roubst against domain changes (Petrenz and Webber, 2011), (Krüger et al., to appear). Importantly, subjectivity is orthogonal to the text types mentioned above: Texts of all types may be predominantly subjective, and most may not (a likely exception being the argumentative type).

What has – to the best of our knowledge – not been addressed in depth so far is the distinction between the tasks of classifying subjectivity (versus objectivity) on the one hand and argumentativeness (versus non-argumentativeness) on the other. To see why this is necessary, consider the following examples from the Brexit reader-discussion website of the *Irish Times*:[1]

(1) Sir, Born in London of an Irish mother and English father during the second World War and living in Dublin since 1968, I endeavoured to encourage my English relatives not to exit.
It looks like a case for returning my British passport, if Ireland will look favourably on my application for a replacement. Yours, etc.

(2) Sir, A black day for Europe. The age of populism and demagoguery is well and truly upon us, from Donald Trump to Ukip. The centre is under threat and is buckling. We must passionately embrace the centre ground and with it the EU. The EU is undoubtedly flawed but it has guaranteed peace in Europe for 70 years and made the idea of war laughable. It has united countries that for millennia were enemies and all it asks, as per Article 2 of the European Union Treaty, is respect for human dignity, freedom, democracy, equality, among other lofty goals. (...) Yours, etc.

Both are clearly subjective and both express an opinion (incidentally the same) on the issue, but only (2) states a thesis/claim and provides reasons in support (and even grants a possible objection), thus clearly qualifies as presenting an argument. (1), in contrast, gives a very short personal narrative followed by a description of the speaker's attitude.

We find the same difference in product reviews. Consider these camera reviews from an Amazon website:[2]

[1] http://www.irishtimes.com/opinion/letters/brexit-the-readers-have-their-say-1.2698710

[2] https://www.amazon.co.uk/product-reviews/B00IK01PJC/ref=cm_cr_arp_d_paging_btm_51?ie=UTF8

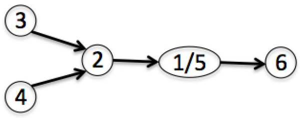

Figure 1: Argumentation structure of example (4)

(3) Camera completely broke whilst on a once in a lifetime holiday and I was nowhere where I could buy a new one. Absolutely gutted as it was first time using camera!

(4) [You'll be lucky if you manage to focus this camera.]$_1$ [Bought to take pictures of jewellery for online sale, the macro function which boasts as having a 5cm macro ability is awful.]$_2$ [For one it is an automatic macro which means that the camera should automatically detect whether the picture requires macro or not - it does not.]$_3$ [For two it should focus on the item in question at even 10cm distance - it does not.]$_4$ [You end up with blurry edges and unfocused pictures.]$_5$ [I was expecting SO much more from an Amazon best seller.]$_6$

Like in the Brexit examples, we sense basically the same opinions (which here resulted in the same low star rating), but the first user gives merely a narration while the second argues (in a fairly complex way with recursive support structure). We added brackets and indices to the sentences so that we can refer to them in our analysis of the argument structure, which was built using the scheme of Peldszus and Stede (2013) and is shown in Figure 1.[3] A brief description of this scheme will be provided in the next section; for now, notice that text segments can be related by SUPPORT and ATTACK relations, and collectively form a tree whose root is the central claim of the text. In our example, all relations are SUPPORT, and we omit the labels here. The node '1/5' indicates that these units in the text convey essentially the same message. Intuitively, to check the representation, a pair of segments in a support relation can be paraphrased with a *why*-link: "I was expecting much more." - "Why?" - "Pictures are unfocused." - "Why?" - "The macro function is awful" - "Why'?" - "Camera doesn't detect macro mode, and camera doesn't focus at close distance."

Using the example, we can begin to make the notion of argumentative depth explicit. For one thing, the measure should contain the proportion of the number of tokens that are taking part in the argumentation analysis: This reflects the amount of non-argumentative information in the text. Furthermore, the measure should reflect the complexity of the tree structure – a flat tree where a claim is supported by a few independent backings is intuitively (and technically) less 'deep' than a structure involving recursion. To measure this, we here simply calculate the average length of the paths from each leaf node to the root. (More elaborate measures could distinguish between the branching factor at the root node and elsewhere, etc.) These two values then constitute our description of argumentative depth – for the time being, we refrain from proposing a way of combining them into a single value. For example (4) and its structure, the pair is: 1.0 / 3.0.

Depth of argumentation, we wish to emphasize, is not the same as quality of argumentation. A writer can build up a fairly complex structure, yet the individual claim/support relations may be weak or even flaky. Here is a (constructed) short camera review to illustrate the point:

(5) I bought this camera yesterday, and I really like it. So even if some people say it's a bad product, you can go ahead and buy it, too. Because it is simply wonderful!

Quality, in this view, is a feature of the reasoning schemes underlying the argumentation, whereas depth is a purely structural measure of the argument as it is presented by the author in the text.

[3] For reasons of space, we cannot discuss all aspects of the analysis here and omit, *inter alia*, a justification of defining the boundaries of segments (argumentative discourse units). See (Peldszus and Stede, 2013).

3 Argumentation in Newspapers: Our Corpus

Our overall goal is to determine whether argumentative depth can be (i) reliably annotated by humans, and (ii) measured automatically. To this end, we wish to correlate it with different layers of linguistic analysis. In order to start this, we favour a corpus of argumentative text that is annotated with argument structure and also with additional layers. The German 'Potsdam Commentary Corpus' (Stede and Neumann, 2014) fulfills this criterion, as it comes with sentence syntax, connectives and their arguments, coreference, and rhetorical structure. We will concentrate here on rhetorical structure; note that our findings do not depend on any specifics of German; a corpus in any other language could be used in the same way.

3.1 The genre of commentary

Most newspapers offer 'opinion' pages with editorials that comment upon current affairs. The specifics of this genre depend to some extent on the different national traditions, but we expect that classifications like that of Schneider and Raue (1996) for German commentary do largely apply to the press in some other countries as well. The authors propose six categories, which include the 'opinion article' (a long piece that slowly builds up a position and encourages readers to reflect), the 'pro and contra' (a crisp and traceable argumentation in favour of a position), the 'on the one/the other hand' (a piece that looks at both sides and remains undecided), and the 'pamphlet' (a strong opinion with little real argumentation). This is, of course, an analytical framework only – actual articles in papers are usually not labelled in this way.

3.2 Data

The Potsdam Commentary Corpus is drawn from two different newspapers: One part is 'pro and contra' (which is in fact their headline) pieces from *Tagesspiegel*; the other is a mix of articles from the opinion page of the local newspaper *Märkische Allgemeine Zeitung* (henceforth: MAZ). All texts are about 10-12 sentences long. The pro/contra articles always come as a pair: One is in favour, one is against the issue under discussion, which may be of local (e.g., Should Berlin build more refugee centers), national (e.g., Should there be mandatory fees for public radio and TV in Germany), or global (e.g., Should we push for a stronger proposal on fighting climate change) relevance. The reader may then decide with whom to side. Given this setting, plus the brevity, we can safely assume a high degree of argumentative depth.

The opinion articles from MAZ, on the other hand, are quite diverse in their argumentative nature: Some merely re-state a piece of news and add some subjective evaluation to it; some try to explain why some event happened; some indulge in a piece of trivia; some take a clear position on a political question and defend it, like a pro/contra text does. Therefore, we can expect to find texts of rather different argumentative depths here, similar to the differences between example texts (1) and (2), and between (3) and (4) above.

For the experiments described below, we selected 15 texts for feature development, and 14 texts for testing; 80% come from MAZ, the rest are 'pro and contra' texts. Both sets were not drawn randomly: They should reflect the spectrum from low to high degree of argumentative depth in roughly equal proportion; these decisions were made by the author (and subject to confirmation by our annotators).

3.3 Annotation layers

For the purposes of this paper, two of the layers provided in the corpus are relevant. The (German) annotation guidelines for these and other layers are freely available.[4]

Rhetorical structure: We are using the RST annotations that were produced for version 2 of the Potsdam Commentary Corpus in 2014. They are largely conforming to the proposals by Mann and Thompson (1988), but the annotation guidelines make some minor modifications. As a preperatory step, the text is first segmented into a sequence of Elementary Discourse Units (EDUs); our guidelines explain which types of clauses constitute such an EDU and which do not. A central concept for our analysis below is

[4] https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/index/index/docId/8276

nuclearity: Most coherence relations in RST distinguish between a central text span (nucleus) and a less important one (satellite); a few treat both (or more) spans as equally important, these are called 'multinuclear'. The result of an RST analysis is a tree structure that joins adjacent EDUs, and then recursively the larger spans.

Using the nuclearity information at each level of that tree, one can read off the central units of the text: those that can be reached from the root of the tree by following only 'nucleus' edges towards the leafs (EDUs). In the second tree of Figure 2 below, EDU 8 is the single central unit; in the the first tree, it is the set of EDUs reached by following all four Joint segments downward.[5] (For an introduction to RST, see also (Taboada and Mann, 2006)).

Argumentation structure: The 'pro/contra' texts in the corpus are already annotated for argumentation, following the scheme proposed by Peldszus and Stede (2013), which is codified in our guidelines. For the MAZ texts, we produced these annotations now, following the same guidelines.

Like with RST, an argumentation structure is also based on a segmentation into elementary units, but these may be larger ('argumentative discourse units' or ADUs). The analysis is a tree whose root represents the 'central claim' unit of the text. The other nodes correspond to ADUs, and the edges to SUPPORT or ATTACK relations, where the latter distinguish between REBUT (challenging the validity of the assertion in the ADU) and UNDERCUT (challenging not the validity of individual ADUs, but the supposition of a SUPPORT relation between two). In contrast to an RST analysis, there is no constraint that only adjacent segments may be conjoined by a relation. Furthermore, the argumentation structure need *not* span the text completely: Parts of it can be deemed irrelevant for the argument made, which is important for our notion of argumentative depth as introduced above.

4 Experimental study

4.1 Features

Simple features. Depth of argumentation is unlikely to be detectable with straightforward surface (or 'near-surface') features. Nonetheless, we tested two simple features that have been used in subjectivity classification (see above) and that might be relevant here: The presence of modal verbs and of 'argumentative' connectives: those that signal a contrastive or causal (in the wide sense) relation.

Argumentation structure: depth measure. The Peldszus/Stede annotation scheme had so far been applied by its authors to pro/contra commentaries and to very short 'microtexts' (Peldszus and Stede, 2016a). Our present annotation is thus its first application to texts which are not as easy to handle: Portions of text might be irrelevant for the argument, and both the central claim and the attachment points of relations can be hard to identify. We will test if our depth measure reflects differences in 'argumentativeness' as perceived by readers.

Rhetorical structure. In the literature there is an ongoing and so far not quite conclusive discussion on the relationship between RST analysis and argumentation analysis (e.g., (Azar, 1999), (Peldszus and Stede, 2013), (Green, 2015), (Peldszus and Stede, 2016b)). Clearly, as mentioned in the beginning, the design of the relations by Mann and Thompson (1988) suggests that the theory be especially amenable to argumentative text. Thus it can be expected that a structural analysis in terms of semantic and pragmatic relations, equipped with nuclearity, might capture what we are characterizing as depth of argumentation here. The first feature that comes to mind is the distribution of semantic versus pragmatic relations (as they have been categorized by the RST and the guideline authors); in particular, we treat Antithesis, Concession, Evaluation, Evidence, Justify, and Reason as 'pragmatic' relations. In the category 'semantic', we include all others except the 'textual' relations Joint, List, Preparation, Conjunction. Here, the hypothesis is "The more pragmatic relations, the more depth" (and conversely for the semantic relations). Then, we use a more elaborate feature to examine some structural properties of the RST tree. The data

[5] The notation in the figure uses straight lines for nucleus links and curved arrows for satellite links. Notice that in the upper tree, EDUs 4-8 have been collapsed in order to fit the tree onto the page. In both trees, segment 1 is missing, because the headline of the text is left out of the analysis.

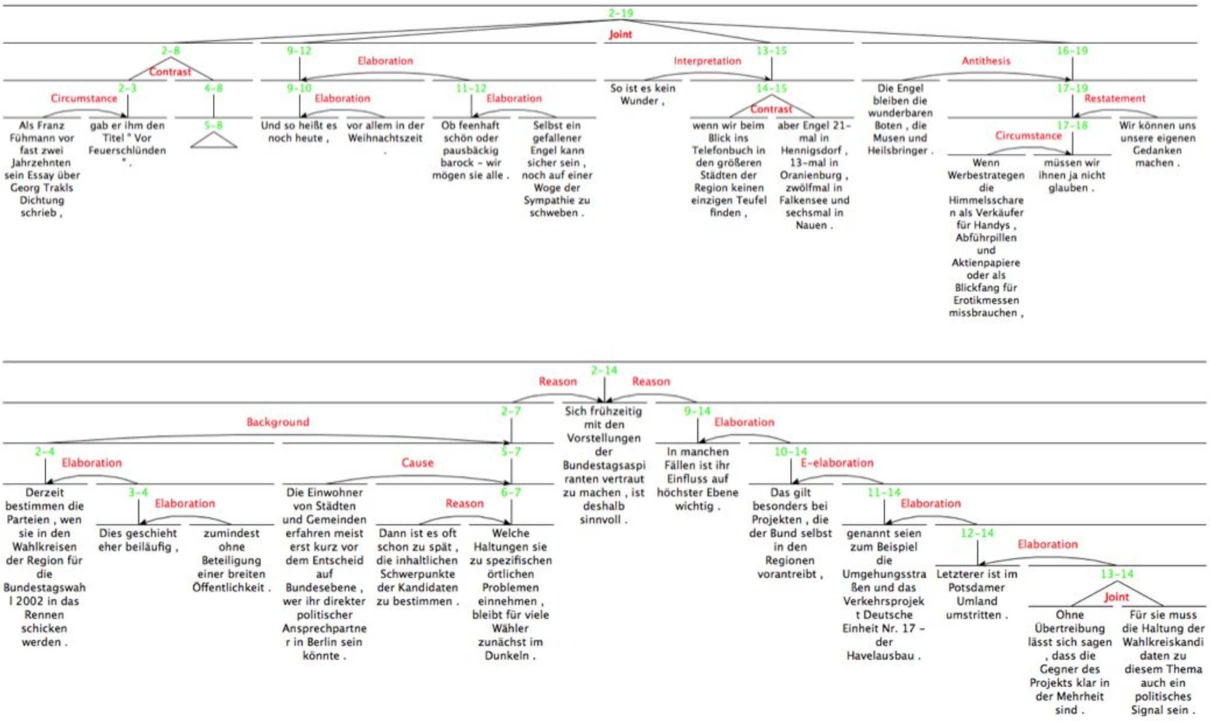
2-19
Joint
2-8
Contrast
2-3
Circumstance
Als Franz Fühmann vor fast zwei Jahrzehnten sein Essay über Georg Trakls Dichtung schrieb ,
gab er ihm den Titel " Vor Feuerschlünden " .
4-8
5-8
9-12
Elaboration
9-10
Elaboration
Und so heißt es noch heute ,
vor allem in der Weihnachtszeit .
11-12
Elaboration
Ob feenhaft schön oder pausbäckig barock - wir mögen sie alle .
Selbst ein gefallener Engel kann sicher sein , noch auf einer Woge der Sympathie zu schweben .
Interpretation
13-15
So ist es kein Wunder ,
14-15
Contrast
wenn wir beim Blick ins Telefonbuch in den größeren Städten der Region keinen einzigen Teufel finden ,
aber Engel 21-mal in Hennigsdorf , 13-mal in Oranienburg , zwölfmal in Falkensee und sechsmal in Nauen .
Antithesis
16-19
Die Engel bleiben die wunderbaren Boten , die Musen und Heilsbringer .
17-19
Restatement
17-18
Circumstance
Wenn Werbestrategen die Himmelsschare n als Verkäufer für Handys , Abführpillen und Aktienpapiere oder als Blickfang für Erotikmessen missbrauchen ,
müssen wir ihnen ja nicht glauben .
Wir können uns unsere eigenen Gedanken machen .
2-14
Reason
Reason
2-7
Background
Sich frühzeitig mit den Vorstellungen der Bundestagsaspi ranten vertraut zu machen , ist deshalb sinnvoll .
9-14
Elaboration
2-4
Elaboration
Derzeit bestimmen die Parteien , wen sie in den Wahlkreisen der Region für die Bundestagswah l 2002 in das Rennen schicken werden .
3-4
Elaboration
Dies geschieht eher beiläufig ,
zumindest ohne Beteiligung einer breiten Öffentlichkeit .
5-7
Cause
Die Einwohner von Städten und Gemeinden erfahren meist erst kurz vor dem Entscheid auf Bundesebene , wer ihr direkter politischer Ansprechpartne r in Berlin sein könnte .
6-7
Reason
Dann ist es oft schon zu spät , die inhaltlichen Schwerpunkte der Kandidaten zu bestimmen .
Welche Haltungen sie zu spezifischen örtlichen Problemen einnehmen , bleibt für viele Wähler zunächst im Dunkeln .
In manchen Fällen ist ihr Einfluss auf höchster Ebene wichtig .
10-14
E-elaboration
Das gilt besonders bei Projekten , die der Bund selbst in den Regionen vorantreibt ,
11-14
Elaboration
genannt seien zum Beispiel die Umgehungsstra ßen und das Verkehrsprojek t Deutsche Einheit Nr. 17 - der Havelausbau .
12-14
Elaboration
Letzterer ist im Potsdamer Umland umstritten .
13-14
Joint
Ohne Übertreibung lässt sich sagen , dass die Gegner des Projekts klar in der Mehrheit sind .
Für sie muss die Haltung der Wahlkreiskandi daten zu diesem Thema auch ein politisches Signal sein .

Figure 2: RST trees with very different nuclearity mass distribution

in the development set suggests a tendency for 'deeply argumentative' texts to have a clearly-identifiable central nucleus, whereas other texts can have multinuclear relations toward the top of the tree, which signifies that several points are perceived as being equally important to the author of the text. For illustration, consider Figure 2, which shows the structure of two texts from our development set. We thus seek to measure the distribution of the 'nuclearity mass' (NM) across the tree. We are not aware of such a measure having been described in the literature, and here suggest two variants. Call the number of satellite links on a path from leaf node to root the 'sat value' of that leaf node.

- NM1: The proportion of leaf nodes with a sat value of 0 or 1 (i.e., 'central' units).
- NM2: Let l_i be the length of the path (irrespective of the nuc/sat distinction) from leaf i to the root. Then, NM2 is the sum of those l_i where i has a sat value of 0 or 1, divided by the sum of l_i for all i of the tree.

While NM1 considers just the number of central nodes, NM2 also takes their distance from the root into consideration. For the upper tree in Figure 2, the two values are 0.93 and 0.88; for the lower tree they are 0.23 and 0.11.

4.2 Collecting judgements on argumentative depth

Having obtained hypotheses from the development set, we proceeded to test them on our second set (14 texts). We had students rate these texts for argumentativeness and then split them into two groups of 'low' and 'high' argumentative depth.

It is not trivial to elicit a judgement on 'argumentative depth', since this is obviously not an established concept. We were interested in intuitive judgements of readers that were not influenced by attempts on (mentally) building an explicit representation of the argumentation. For this reason, we worked with first-year students that had not received any training in argumentation analysis. Our procedure was to

Rank	Value	Source	Expect.	Rank	Value	Source	Expect.
1	4.8	ProCon	high	9	2.8	MAZ	high
2	4.2	MAZ	high	10	2.8	MAZ	medium
3	3.8	MAZ	high	11	2.4	MAZ	low
4	3.8	MAZ	high	12	2.2	MAZ	low
5	3.4	ProCon	high	13	2.0	MAZ	low
6	3.4	MAZ	medium	14	1.8	MAZ	low
7	3.4	MAZ	low	15	1.0	News	low
8	3.2	ProCon	high	16	1.0	News	low

Table 1: Text ranking obtained via students' answers to questionnaire

present them with a questionnaire posing the following questions, where the answer was to be given as a position on a 1..5 Likert scale:[6]

- Is the text a news report or an opinion text?
- If the text is a news report, is the reporting clear or unclear?
- If the text is an opinion text, is the position of the author clearly represented?
- If the text is an opinion text, does the author provide clearly recognizable arguments for his or her position?

The first question was meant as a first proxy for the amount of 'opinion' in the texts and in particular to identify non-commentaries. To make it work properly, we added filler items to the texts: two news reports taken also from the *Tagesspiegel*, which were of roughly the same length as the commentaries. The second question then merely served to balance the question set – the answers were not used in determining the ranking, which was based only on the answers to the last two questions. Notice that the last one does not refer to structural depth but merely to the presence of arguments and their clarity; this means that the students did not rate 'depth' directly. (In a future version of the study we plan to add a question to that effect.)

Every student judged three texts, which were mixed according to our expectations on argumentative depth. Most, but not all, students saw one of the news texts (filler) in their set. 30 students participated, so we obtained 90 judgements in total, which amounts to 5 per text (14 commentaries plus 2 news), over which we calculated averages.

4.3 Results

We translated the ratings into a ranking of the texts, ranging from low to high opinion/argument. Table 1 shows for each text its rank and its accumulated value on the Likert scale, its source, and the depth category we had originally expected for it. The first good news is that the news reports (fillers) were easily identified – they are thus discarded from the subsequent evaluation. Then, using the arithmetic mean, we determined the cutoff to form the two groups of eight 'high' and six 'low' argumentative texts (shown in the two halves of the table). Below we call the groups H and L.

As the table shows, our expectations on ranking are mostly confirmed, with two exceptions at rank 8 and 9, which we had estimated higher up.

Simple features. No difference between the two groups can be found for the frequency of modal verbs and of contrastive/causal connectives (neither on the development nor on the test set).[7]

[6] In addition, we asked whether the text was considered generally 'understandable' so that we could discard judgements where the student apparently had not understood the text well enough. This happened only in two instances.

[7] As pointed out by a reviewer, simple features might be indicative of just the first component of our depth measure (the proportion of argumentative tokens); testing this is left for a follow-up study with a larger corpus.

	pragmatic relations	semantic relations	NM1	NM2
group H	d: 43.17% / t: 38.3%	d: 35.0% / t: 45.8%	d: 37.29% / t: 40.79%	d: 28.1% / t: 33.82%
group L	d: 45.13% / t: 28.03%	d: 32.5% / t: 55.88%	d: 49.75% / t: 55.99%	d: 41.22% / t: 46.63%

Table 2: Results for RST features. d = dev set / t = test set

Argumentation structure. On the test set, the average proportion of tokens participating in the argumentation is 0.81 in group H and 0.31 in group L – a very clear difference. The average path lengths in the argumentation tree also confirm the expectation: In group H it is 1.59, and in group L only 1.09. It seems that our depth measure can capture the difference between the groups.

Rhetorical structure. Table 2 summarizes the results for the various RST features. For one thing, it shows that the results on our (more or less equally small) development and test sets can differ considerably, which suggests that more data is needed to obtain more reliable results. However, other tentative conclusions can be drawn: The figures indicate that the distribution of semantic/pragmatic relations is not a good feature for depth of argumentation, whereas both ways of measuring the distribution of nuclearity mass yield clear differences.

5 Summary

Constraining ourselves to monologue text, we suggested assessing the 'depth' of argumentation as a subtask of argumentation mining, when it aims at reconstructing the full structure of the argument. Our measure combines (i) the proportion of the text that contributes to the argument and (ii) a simple account of the complexity of the tree representing the argument: the average length of the paths in that tree. We hinted at the possibility of using more sophisticated variants of (ii). Importantly, we pointed out that depth is a purely structural measure, which is to be distinguished from argument 'quality' as it is being addressed in related work.

To test our measure, we conducted a pilot study with a small corpus of annotated newspaper commentary texts, supplemented by filler items (news reports). Using a set of questions, we obtained judgements on opinionatedness and argumentativity from human raters, which led to ranking the texts and then splitting them into two groups of high and low argumentative depth, respectively. We found that the depth measure can serve to differentiate the groups.

In addition, we were interested to see whether the human judgements correlate with simple surface features and with aspects of the rhetorical structure of the texts. As for the former, we tested modal verbs and connectives (standardly used in subjectivity classification) but could not find an effect. Regarding rhetorical structure, interestingly, the distribution of 'semantic' versus 'pragmatic' relations (as defined in RST) turns out not to differentiate between the groups. However, two measures of distributing the 'nuclearity mass', which we proposed here, do reflect the distinction. We conjecture that these measures can account for the clarity and focus of the argumentation. One next step is to experiment with variants of the RST measures, trying to combine relational and structural features.

To obtain more evidence for the validity of the results, both on the depth measure and the correlations with rhetorical structure, we plan to supplement the study with another experiment on a larger number of texts from the commentary corpus. Given a broader set of data, proper statistical measures can be computed for annotator agreement and the correlations. Furthermore, as noted earlier, the questionnaire will be extended in such a way that raters make judgements on depth more explicitly; in the pilot, we only asked for an intuitive judgement of the presence of clearly recognizable arguments (irrespective of their potential recursive structure).

Afterwards, it will be interesting to apply the approach to other genres, such as student essays, where we expect to find similar differences of argumentative depth. For the genre we studied here, newspaper commentary, we see our work as a contribution to identifying sub-genres (different types of commentary) and explaing the differences between them.

Acknowledgements

Many thanks to the anonymous reviewers for their constructive comments und suggestions on improving the paper.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

M. Azar. 1999. Argumentative text as rhetorical structure: An application of Rhetorical Structure Theory. *Argumentation*, 13:97–114.

K. Budzynska, M. Janier, C. Reed, and P. Saint-Dizier. 2013. Towards extraction of dialogical arguments. In *Working Notes of the 13th Workshop on Computational Models of Natural Argument (CMNA 2013)*, Rome.

J. Burstein, D. Marcu, and K. Knight. 2003. Finding the WRITE stuff: automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–39.

Nancy Green. 2015. Annotating evidence-based argumentation in biomedical text. In *Proceedings of the IEEE Workshop on Biomedical and Health Informatics*.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2015 NAACL-HLT Conference*. Association for Computational Linguistics, June.

K. Krüger, A. Lukowiak, J. Sonntag, S. Warzecha, and M. Stede. to appear. Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering*.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *TEXT*, 8:243–281.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the ICAIL 2009*, pages 98–109. Barcelona, Spain.

Huy Nguyen and Diane Litman. 2015. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28, Denver, CO, June. Association for Computational Linguistics.

Shereen Oraby, Lena Reed, Ryan Compton, Ellen Riloff, Marilyn Walker, and Steve Whittaker. 2015. And that's a fact: Distinguishing factual and emotional argumentation in online dialogue. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 116–126, Denver, CO, June. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Andreas Peldszus and Manfred Stede. 2016a. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon 2015 / Vol. 2*, pages 801–816, London. College Publications.

Andreas Peldszus and Manfred Stede. 2016b. Rhetorical structure and argumentation structure in monologue text. In *Proceedings of the Third Workshop on Argumentation Mining*, Berlin. Association for Computational Linguistics.

Philipp Petrenz and Bonnie Webber. 2011. Stable classification of text genres. *Computational Linguistics*, 37(2):385–393.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The Penn Discourse Treebank 2.0. In *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

W. Schneider and P. Raue. 1996. *Handbuch des Journalismus*. Rowohlt, Hamburg.

Carlota Smith. 2003. *Modes of discourse. The local structure of texts*. Cambridge University Press, Cambridge.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 1501–1510, Dublin.

Manfred Stede and Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 925–929, Reikjavik.

Maite Taboada and William Mann. 2006. Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies*, 8(4):423–459.

F.H. van Eemeren, R. Grootendorst, and A.F. Snoeck Henkemans, editors. 1996. *Fundamentals of argumentation theory*. Lawrence Erlbaum, Mahwah/NJ.

M.G. Villalba and P. Saint-Dizier. 2012. Some facets of argument mining for opinion analysis. In *Computational Models of Argument. Proceedings of COMMA 2012*, pages 23–34, Amsterdam. IOS Press.

Egon Werlich. 1975. *Typologie der Texte*. Quelle und Meyer, Heidelberg.