Dialogue & Discourse 9(1) (2018) 107–127 doi: 10.5087/2018.104

ChangeMyView Through Concessions: Do Concessions Increase Persuasion?

Elena Musi EM3202@COLUMBIA.EDU
Data Science Institute
Columbia University

Debanjan Ghosh DG513@MIT.EDU
McGovern Institute for Brain Research
Massachusetts Institute of Technology

Smaranda Muresan SMARA@COLUMBIA.EDU
Data Science Institute
Columbia University

Editor: Maite Taboada

Submitted 07/2017; Accepted 07/2018; Published online 08/2018

Abstract

In discourse studies concessions are considered among those argumentative strategies that increase persuasion. We aim to empirically test this hypothesis by calculating the distribution of argumentative concessions in persuasive vs. non-persuasive comments from the ChangeMyView subreddit. This constitutes a challenging task since concessions are not always part of an argument. Drawing from a theoretically-informed typology of concessions, we conduct an annotation task to label a set of polysemous lexical markers as introducing an argumentative concession or not and we observe their distribution in threads that achieved and did not achieve persuasion. For the annotation, we used both expert and novice annotators. With the ultimate goal of conducting the study on large datasets, we present a self-training method to automatically identify argumentative concessions using linguistically motivated features. We achieve a moderate F1 of 57.4% on the development set and 46.0% on the test set via the self-training method. These results are comparable to state of the art results on similar tasks of identifying explicit discourse connective *types* from the Penn Discourse Treebank. Our findings from the manual labeling and the classification experiments indicate that the type of argumentative concessions we investigated is almost equally likely to be used in winning and losing arguments from the ChangeMyView dataset. While this result seems to contradict theoretical assumptions, we provide some reasons for this discrepancy related to the ChangeMyView subreddit.

Keywords: concessions, argumentation, subreddit, discourse relations, classification task

1. Introduction

A major challenge for *Argument Mining*—the automatic identification of argumentative structures within discourse—is the identification of linguistic features which characterize a winning argument. Discourse moves that allow a speaker to achieve persuasion have been investigated since Antiquity, being at the core of *Rhetoric*, and are still a central concern in contemporary discourse studies and *Argumentation Theory*. The automatic identification of persuasive discourse is also receiving more

and more attention in computational linguistics. Concessions have been unanimously deemed as strategies which increase persuasion (Section 2). However, not every concession is part of an argument. We provide a *definition of argumentative concessions* in semantic and pragmatic terms. On this basis, we *empirically test the theoretically-informed hypothesis that argumentative concessions work as persuasive strategies* by calculating their distribution in persuasive vs. non persuasive discourse. An ideal dataset to carry out this empirical investigation is the ChangeMyView subreddit platform, where multiple users negotiate opinions on a certain issue willing to change their point of view through other users' arguments. When their point of view is changed, they award a Δ point and back it up with a reason. Thus, this platform provides us with a clear user intent — persuasion — and a clear signal when a message is perceived as persuasive (Δ point). Even though it would have been convenient to observe the function of concessions in different datasets, the lack of explicit signals of persuasion prevents us to carry out such a comparison. We use the ChangeMyView dataset released by Tan et al. (2016).

Our task faces two challenges. First, in order for concessions to increase persuasion, they need to be part of an argument. However, not every concession is argumentative (Grote et al., 1997): The pragmatic function of the sentence "Although it is December, there is no snow" is not convincing the hearer about the lack of snow — which is an observable fact not subject to doubt — but simply that of expressing surprise for an unusual combination of events. This sentence could work in certain contexts as an argument (e.g., for the standpoint "Global warming is worse and worse") but does not semantically presuppose any controversy and, thus, the presence of argumentation. To address this challenge, we provide a *semantically based methodology to identify argumentative concessions and compare them with other types of concessions* (Section 3). Second, the automatic retrieval of argumentative concessions through discourse markers is difficult due to the polysemy with contrast and other discourse relations (Prasad et al., 2014). State of the art discourse parsers trained on the Penn Discourse Treebank (Prasad et al., 2008) achieve low accuracy in the identification of concessive uses of discourse connectives in general due to the small number of instances in the training data; this result is reflected also on our task of identifying *argumentative concessions*. Therefore, we first observe the distribution of concessions in manually labeled data both through expert annotation and through crowdsourcing (Section 5). We focus our analysis on four discourse connectives: *but*, *though*, *however*, and *while* for two reasons: (1) they constitute 85% of the overall occurrences of potential markers of concessions in our dataset, and (2) they are highly polysemous and thus require disambiguation (Section 4). We used 20% of the overall occurrences of these markers for our crowdsourcing study (i.e., 2,440 instances) as well as a separate dataset of 1,000 instances for the expert annotation task. Second, as a step towards large-scale analysis of persuasive discourse, we use the manually labeled data as training, development and test sets to build computational models to detect argumentative concessions (Section 6). We achieve a moderate F1 of 57.4% on the development set and 46.0% on the test set via a self-training method (Section 7).

Our findings from the manual labeling (using both expert and crowdsourcing annotations) indicate that the type of argumentative concessions we investigate is almost equally likely to be used in winning (Δ-awarded) and losing (non-Δ) arguments. While this result seems to contradict theoretical assumptions, we provide some reasons related to the nature of the ChangeMyView subreddit (Section 8). In addition, we present a preliminary analysis on running our computational models on a different dataset, The Yahoo News Annotated Comments Corpus, (*ERIC*: Engaging, Respectful, and/or Informative Conversations) (Napoles et al., 2017a), where comments are labeled as persuasive or not via crowdsourcing (persuasion is a "binary label indicating whether a comment contains

persuasive language or an intent to persuade"). Differently from *Delta points* in ChangeMyView, the persuasiveness label does not inform us about what discourses achieved persuasion among the participants of an actual interaction, but provides hints as to what type of language is perceived as persuasive by third parties. We observe that argumentative concessions are present more in persuasive comments than in non-persuasive comments in the ERIC dataset. The dataset and code are freely available.[1]

2. Related Work

Concessions have received various non-overlapping definitions. We take as a default definition the one provided by Grote et al. (1997), which focuses on the semantic relation holding between two connected propositions (A, B): "On the one hand, A holds, implying the expectation of C. On the other hand, B holds, which implies not C, contrary to the expectation induced by A":

$A \rightarrow C$ e.g., "This dress is gorgeous" $\rightarrow$ it is worth buying

$B \rightarrow \neg C$ e.g., "but, it is expensive" $\rightarrow$ it is NOT worth buying

The conflict resides in the expectations generated by A and B which are mutually exclusive. Besides this concessive configuration, which is called indirect concession (Azar, 1997; Izutsu, 2008), there are direct concessions where *B* and the following implication rule do not have to be verbalized. In these cases the main clause constitutes the negation of the expectation arising from proposition *A* (e.g., "This dress is gorgeous, but I am not buying it"). The distinction between direct and indirect concessions is particularly relevant from a discourse perspective since direct concessions are more suitable to be used nonargumentatively to describe states of affairs and to increase interest for an unexpected contradiction.

The pragmatic function played by concessions has been widely addressed in linguistically oriented *Discourse Studies*. To cite just a few, drawing from Quintilian *Institutio Oratoria*, Perelman (1971) explains that through concessions,

> "one gives a favorable receipt to one's opponent's real or presumed argument. By restricting his claims, by giving up certain theses or arguments, a speaker can strengthen his position and make it easier to defend, while at the same time he exhibits his sense of fair play and his objectivity".

Mann and Thompson (1988) list concessions among presentational rhetorical relations, aimed at increasing the addressee's positive attitude towards the speaker's beliefs. According to Pragma-dialecticians (Van Eemeren et al., 2007) concessions are attested either in the confrontation stage of an argumentative discussion, where the difference of opinion between two parties is stated, and the opening stage, where the common starting points of the discussion are established. In a corpus study of interview transcripts involving environmental activists, Uzelgun et al. (2015) show that the concessive construction "yes ... but" constitutes a privileged viewpoint to investigate the (dis)agreement space singling out what is accepted and what is criticized. Antaki and Wetherell (1999) identify a series of rhetoric strategies through which concessions strenghten the speaker's positions undermining counterarguments. Similarly, Couper-Kuhlen and Thompson (2000) exemplify how concessive repairs are used by speakers to back down their overstatements and foster credibility. In their investigation of myside bias in written argumentation, Wolfe et al. (2009) show that texts which present and rebut other-side arguments achieve better ratings of agreement, quality

1. https://github.com/debanjanghosh/concessions

and overall impression of the author. However, the presence of concessions preceding rebuttals did not lead to a higher perception of the arguments' quality. It has to be remarked that students that were asked to evaluate the texts were not individually engaged in an interactive conversation; therefore, face-threatening risks were not the same as those characterizing a face to face argumentative exchange. The argumentative value played by concessive discourse relations is recognized by Green (2010) who inserts concessions among the RST relations needed to represent argument presentation in a biomedical corpus.

As far as datasets are concerned, while resources annotated as to agreement and disagreement are provided (Walker et al., 2012), cases of partial (dis)agreement are neglected. Computationally, Somasundaran and Wiebe (2009) develop an unsupervised method for stance detection in online debates, taking into consideration also concessionary opinions. From a pragmatic perspective, a growing interest is devoted to the identification of persuasive discourse strategies in order to implement classification experiments. Young et al. (2011) released a corpus of blog posts annotated as to persuasive tactics according to sociological studies (e.g., promises/threats, mentions to duties/generalizations, appeal to reason). The evaluations of the predictive power of the strategies for the identification of persuasion show that the label *Reason* guarantees accuracy; however the unigram SVM baseline for *Reason* is poor due to the difficulty in identifying rhetorical relations. Drawing from the same set of tactics Young et al. (2011) presented a corpus of 37 transcripts from four sets of hostage negotiation transcriptions annotated as to persuasion features. Using supervised learning algorithms they show that persuasion tactics constitute machine-learnable features. Tan et al. (2016) analyzed shallow linguistic and interactional features which happen to be persuasive in ChangeMyView, a subreddit where users exchange opinions and assign a *Delta point* to the user that managed to change their view: dissimilarity at the lexical level seems to play a major role, together with the order and the number of interactions. Wei et al. (2016) investigated the performance of different sets of features in predicting persuasion: they have found that argumentation-based features perform better than shallow textual features. Habernal and Gurevych (2016) have recently released a corpus of 16k pairs of arguments over 32 topics annotated as to persuasiveness using crowdsourcing. Annotators were also asked to provide reasons behind their choices. Experiments with feature-rich SVM (using LIBSVM tool; (Chang and Lin, 2011)) and Long Short-Term Memory (LSTM) neural networks (Hochreiter and Schmidhuber, 1997) reveal that predicting persuasion is a task that requires analytic skills still hard to attain computationally.

The detection of rhetorical/argumentative relations is bound to the performance of state of the art discourse parsers trained on the Penn Discourse Treebank (PDTB), the largest annotated corpus so far available (Prasad et al., 2008). Pitler et al. (2009) showed that syntactic feautures play a crucial role in disambiguating explicit connectives. Drawing from their work, Lin et al. (2014) built a four-step pipeline for the identification of implicit and explicit discourse relations passing through the identification of text spans functioning as argument. Connectives indicating concessions have been proven by Swanson et al. (2015) to significantly correlate with the presence of argumentation, even though not with argument quality. For the task of detecting the *type* rather than the *class* of discourse connectives, which is more similar in nature to our problem, the performance of state of the art models is still modest (56.91 F-measure even for explicit cases) (Biran and McKeown, 2015). Biran and McKeown (2015) treat PDTB discourse parsing as two separate tagging tasks and show that connective-specific grammatical features promise to improve the results. On these grounds, we combine various lexical patterns, pragmatic as well as semantic features in conjunction

with a self-training approach for our task of identifying whether a discourse connective introduces an argumentative concession or not.

3. Argumentative concessions

In order to answer our research question — whether argumentative concessions increase persuasion — we need to provide a definition of argumentative concessions. Drawing from Musi (2017), we consider concessions as displaying an argumentative function when the proposition introduced by the connective — B —, which denies the expectations brought about by a preceding proposition, expresses the speaker's standpoint. This happens when the following two conditions are met: (i) proposition B is asserted by the speaker who is committed to its truth at the moment of utterance; and (ii) proposition B is non factual — its truth is not self-evident. Concessive relations such as "Although it is already very warm, [there are no buds on the trees]$_B$" and "He states that, despite the difficult situation, [the company will not go bankrupt]$_B$" are, therefore, not argumentative since propositions B are an unassailable fact and a prediction with which the speaker does not necessarily agree, respectively. As far as proposition A is concerned, what is conceded can either be: (i) a point that could *possibly* be made by another speaker (i.e., "[Sure, we could argue about some hypothetical other religion in the region causing similar problems]$_A$, but that hypothetical world is not this world") or that belongs to *common ground knowledge* (e.g., "[He's no Hitler, of course]$_A$, but only because North Korea isn't powerful enough to start annexing its neighbors and has no substantial minority population to send to death camps"), or (ii) a claim previously made by another speaker in the discussion. Concessions of type (i) can be used to prevent a counterargument. However, they do not necessarily display an interactional value, since the potential disagreement imagined by the speaker may have never happened. Concessions of type (ii) are, instead, used in conversations as mitigating strategies to avoid disruptive disagreement. They, are, therefore, inherently argumentative. Here are two examples.

- Speaker 1:[. . .] Basically, I'm glad Jackson opted to make excellent **movies,** instead of attempting a shot for shot visualization of a **book,** and I think the extended cuts subvert his success in that regard. [. . .]
- Speaker 2: As far as Peter Jackson's opinion, the opinion of the author or producers of a content doesn't dictate how it should be interpreted. He is a person who thinks A, that doesn't mean that A is the be all and end all. Like any piece of media once it's released into the public it becomes its own beast that can have its interpretations judged by others. [I agree with you that the quality of the movie matters]$_A$, but [if both versions are good movies and didn't make unnecessary changes then I think they are different rather than one being better]$_B$.

As explained by Couper-Kuhlen and Thompson (2000) this concession type is meant to be persuasive: a speaker, recognizing the validity of a point made by the hearer before expressing disagreement, avoids face-threatening acts and is perceived as reasonable by the hearer. We, thus, consider the description provided by Interactional Linguistics (Couper-Kuhlen and Thompson, 2000, 2005) as our definition of argumentative concessions:

- 1st move: Speaker1 states something or makes some point
- 2nd move: Speaker2 acknowledges the validity of this statement or point (the conceding move)

- 3rd move: Speaker2 goes on to claim the validity of a potentially contrasting statement or point

At a semantic level, proposition A (the conceding move of Speaker2) is always an evaluative proposition since the speaker positively qualifies a standpoint advanced in the preceding post directly through the expression of positive sentiment (e.g., "I like your post, but [. . .]") or agreement (e.g., "You are right, but [. . .]"). At an informational level, the sentiment of the evaluation constitutes new information in the discourse flow, while what is evaluated constitutes old information which coincides with a point that has been previously made by another speaker.

4. Data

To empirically test our theoretically-informed hypothesis we use ChangeMyView (CMV) subreddit: "dedicated to the civil discourse of opinions, and built around one simple idea: in order to resolve our differences, we must first understand them". Users start a discussion thread expressing their opinion (original post, OP) about a certain issue and other users challenge their view in a network of subsequent comments. ChangeMyView constitutes a particularly suitable environment for the study of persuasive argumentation. First, the users, in order to get their posts published, have to respect a series of submission and comment rules, which ensures the presence of argumentation:[2]

- Submission rules:
  - (s1) Try to explain the reasoning behind your view, not just what that view is (500+ characters required).
  - (s2) You must personally hold the view and be open to it changing.
  - (s3) Submission titles must adequately sum up your view and include CMV: at the beginning.
  - (s4) Only post if you are willing to have a conversation with those who reply to you, and are available to start doing so within 3 hours of posting.
- Comment Rules:
  - (c1) Direct responses to a CMV post must challenge at least one aspect of OPs stated view (however minor), or ask a clarifying question.
  - (c2) Don't be rude or hostile to other users.
  - (c3) Refrain from accusing OP or anyone else of being unwilling to change their view.
  - (c4) If you have acknowledged/hinted that your view has changed in some way, please award a delta (Δ).
  - (c5) Comments must contribute meaningfully to the conversation.

This combination of submission and comment rules matches the main rules for conducting an ideal critical discussion (Van Eemeren et al., 2013): rules (s1), (c2) and (c3) constitute an application of the freedom rule ("Parties must not prevent each other from advancing standpoints or from casting doubt on standpoints"), while rule (s4) guarantees that the burden of proof rule ("A party that advances a standpoint is obliged to defend it if asked by the other party to do so") is respected.

2. The following quotations are taken from the CMV wiki (https://www.reddit.com/r/changemyview/wiki/index).

Rule (c1) can be interpreted as a sort of standpoint rule ("A party's attack on a standpoint must relate to the standpoint that has indeed been advanced by the other party"). Finally, rule (c4) presupposes the closure rule according to which "A failed defense of a standpoint must result in the party that put forward the standpoint retracting it". In addition, ChangeMyView constitutes a unique environment for the study of persuasion: although corpora annotated as to the presence of persuasive language exist (e.g., Napoles et al. (2017b)), CMV, as far as we know, is the only available dataset containing first-hand information (through awarded Δ points) as to what arguments are perceived as persuasive by actual language users.

Finally, since the discussed issues cover very different topics and the participants who are anonymous can have different epistemic backgrounds, the results of the analysis in terms of persuasion strategies can be generalized across different text genres and contexts. In our study, we used a dataset collected from the ChangeMyView platform introduced by Tan et al. (2016), where only the replies by the root challenger are considered, defining all the replies by the root challenger in a path as the rooted path-unit. For each rooted path-unit that wins a Δ, they select a rooted path-unit in the same discussion tree that did not win a Δ but was the most "similar" in topic (similarity computed using Jaccard similarity measure). With this setup, the goal is to de-emphasize what is being said, in favor of how it is expressed. We thus end up with a paired dataset that contains Δ-awarded and no-Δ comments, which can be used to test the hypothesis that argumentative concessions are persuasive strategies by computing their distribution in the Δ-awarded and no-Δ comments.

We focus our analysis on four discourse markers: *but*, *though*, *however*, and *while* for two reasons: (1) they constitute 85% of the overall occurrences of potential markers of concessions in our dataset (see Table 1), and (2) they are highly polysemous and thus require disambiguation. For each marker, we collect the sentence in which the marker is occurring as well as the previous and the next sentence (from both Δ and no-Δ comments).

Next, we manually annotated two segments of the datasets using expert annotation and crowdsourcing, respectively. The first segment includes a set of 1,000 examples of the 4 discourse markers (e.g., *but*, *though*, *however*, and *while*) and has been annotated by an expert annotator.[3]The second segment consists of two samples each containing 10% of the overall occurrences of the four discourse markers from Δ and no-Δ comments, which resulted in a total of 1,220 instances each (total of 2,440 instances). The annotators were asked to identify whether a marker introduces an argumentative concession (*arg_c*) or not. One of the samples is used as development set (dev) and one as test set ($test$) in the machine learning experiments. The dev set is used for tuning the parameters of the learning model, while the $test$ set constitutes the blind set used to evaluate the learning model. Instead of expert annotators, we use a *crowdsourcing platform* – Amazon Mechanical Turk (MTurk) to identify the *arg_c*. The task is framed as follows: Given a sentence or pairs of two adjacent sentences in ChangeMyView containing one of the four discourse markers, five Turkers on MTurk are asked to identify and label as *arg_c* those occurrences in which the sentence preceding the connective expresses agreement or positive sentiment towards a point previously made by another speaker. The Turkers are provided with detailed instructions of the task and multiple examples. Before conducting the full crowdsourcing annotations tasks, we ran a pilot annotation to evaluate whether Turkers are able to complete such task. In the pilot annotation round, we realized that providing the Turkers with the mere definition of argumentative concessions was not enough since annotators were selecting as argumentative concessions also those conceding statements be-

3. After annotation we noticed that 20 examples were duplicates, and thus we used the remaining set of 980 examples in the experiments.

Marker	Δ	**no-Δ**
admit	26	17
albeit	9	17
although	78	93
but	4403	5908
concede	8	13
despite	89	114
even if	255	314
even though	101	129
even when	31	55
however	132	213
in spite of	10	8
nevertheless	3	10
notwithstanding	1	4
non the less	-	-
nonetheless	7	18
the fact remains that	3	4
though	426	619
whereas	48	73
while	575	763
total	6205	8372

Table 1: Distribution of candidate concessions markers in the dataset collected from ChangeMyView

longing to common knowledge. Therefore, we refined the guidelines to discard this latter type of concessions pointing to the fact that they tend to contain modal adverbs of certainty (e.g., *(for) sure, of course*) that, differently from markers that imply a dialogical exchange (e.g., *agree*, *understand*), signal that the truth of a general statement is taken for granted since it belongs to the common ground (e.g., "of course, runners use the lower body more than the upper body, but the same is true for swimmers in reverse (they use more upper than lower)"). We, moreover, specified that they contain second person pronouns/adjectives that do not work as deictics, but are used impersonally to portray fictive scenarios (i.e., "yes, a portion of *your* money goes to things you don't support; but likewise, things *you* do support are partially funded by other people who may not support those things."). For both dev and $test$ sets, containing 1,220 sentences (or pairs of sentences) each, we obtain 6,100 labels from the Turkers. To assure a level of quality control, only qualified Turkers were allowed to perform the task (i.e., more than 95% approval rate and at least 5,000 approved Human Intellingence Tasks — HITs). Each HIT contained one sentence (or pairs of sentences) to be labeled and the Turkers were paid 5 cents for each HIT. We compute the inter-annotator agreement (IAA) between the Turkers via Fleiss Kappa measure (Fleiss, 1971). We obtain fair agreement both for dev (κ = 0.31) and for $test$ (κ = 0.22) sets. As underlined by Passonneau and Carpenter (2014), standard measures for inter-annotator reliability are not suitable to account for corpus quality, and

in addition, the κ values obtained on semantic annotation tasks is not high. Therefore, we consider the obtained IAA as a reasonable output which does not undermine the relevance of the corpus.

5. Distribution of argumentative concessions in manually labeled data

The expert annotated dataset contains 229 instances of *arg_c* and 751 instances of other types of concessions or instances expressing contrast or other discourse relations, which we denote as *other*.

For the crowdsourcing experiment, we take majority voting among the five Turkers to choose the label (*arg_c* or *other*). We obtained 201 *arg_c* instances (out of 1,220) for the dev set and 174 *arg_c* instances (out of 1,220) for the $test$ set. To better understand the difficulty of the task, we selected the cases where 3 annotators chose one label, while 2 others choose the other label and we asked an expert annotator to label these cases. We compared the expert labels with the labels obtained by majority voting (i.e., the labeled chosen by 3 Turkers). In the $test$ set, out of 225 such instances, in 22 cases there is a mismatch between the label annotated by three Turkers and the expert annotator; in the $test$ set, out of 280 such instances, the mismatch amounts to 67 cases. Zooming into the mismatched examples, in the dev set 4/22 cases are not recognized as argumentative concessions by the Turkers, while 18 cases are misclassified as argumentative concessions. In the $test$ set, the type of mismatch seems more balanced: 31/67 instances are not recognized as argumentative concessions by Turkers, while 36/67 are misclassified as argumentative concessions. We observe that discarding these most confusing cases where majority is formed by only 3 Turkers, the IAA improves both in the dev set (κ = 0.43) and in the $test$ set (κ = 0.35). Taking into account the annotation provided by the expert for those instances, the number of *arg_c* in dev and $test$ sets is 179 and 168, respectively.

From the qualitative analysis, it seems that argumentative concessions are misclassified when the proposition functioning as standpoint (proposition B) expresses a negative evaluation of the opinion held by another speaker, while the proposition A is used to specify the degree of such an attack, more than to express a partial agreement. In a couple of sentences such as "I don't believe the war was worth the suffering, but to say that nothing positive happened because of it is a little disingenuous", the speaker's intention is not that of partially agreeing with the judgment made by the preceding speaker, but to clarify to what extent it is indeed ingenuous. Cases not recognized as argumentative concessions contain a positive sentiment rather than an explicit agreement with what was said by the preceding speaker (e.g., "You have noble goals, but there are very real downsides").

The distribution of argumentative concessions in threads that have and have not been awarded a Δ point does not appear to be significantly skewed amounting to 99 (awarded) vs 130 (not awarded) occurrences in the set annotated by an expert and 165 vs. 194 in the crowdsourcing experiment. Due to the limited quantity of annotated data available, before discussing the relevance of the attested distribution (section 8), we develop a preliminary computational model to classify and trace back argumentative concessions on a larger scale.

6. Computational models to detect argumentative concessions

We seek to automatically identify argumentative concessions from the CMV corpus. We frame the task as a binary classification task: *arg_c* vs. *other*. Since using expert annotators is an expensive process, we use the expert annotated data as a small labeled training data (total of 980 instances; Section 5). Due to this small annotated data scenario, we develop a self-training method (Clark et al., 2003; Mihalcea, 2004), which uses the remaining unannotated 70% of the CMV corpus (section 6.2)

as unannoted data. From our annotation studies it is clear that all the datasets are highly unbalanced (the size of the *arg_c* class is much smaller than the size of the *other* class).

As stated earlier in the previous section, the data annotated through crowdsourcing (the two samples of 10% of the CMV corpus each) is used as development (dev) and test ($test$) sets, respectively.

In the following section we discuss the features as well as the linguistic patterns associated with the argumentative concessions used in our experiments (section 6.1). Second, we discuss the classification task using self training in section 6.2 and the results in section 7, including comparison with an off-the-shelf state-of-the art discourse parser trained on the Penn Discourse Treebank that aims to classify the *type* of discourse connectives, not their *class* (Biran and McKeown, 2015).

6.1 Feature Description

We use linguistically-motivated features inspired by research on discourse relation identification and persuasive argument identification (Stab and Gurevych, 2014; Ghosh et al., 2016). A brief description of the features is given below.

- *bag-of-words:* We selected the entire CMV dataset used by Tan et al. (2016) to extract the bag-of-words features (e.g., unigrams and bigrams). We consider every sentence that contains the four candidate discourse markers (e.g., "but", "while", "however", and "though") as candidate examples. If the sentence starts with the markers we also consider the previous sentence. Next, based on tf-idf scores (i.e., we treated each sentence as a document), we select the top 1,000 unigrams and bigrams as candidate features.

- *personal pronouns and adjectives:* By definition, argumentative concessions dialogically point to the stance taken by the previous speaker. They, therefore, contain personal pronouns and adjectives (e.g., "I see your point", etc.): we consider as features both a list of first person (e.g., "I","me", "my", "mine") and second person (e.g., "you", "your", "you' re") pronouns and adjectives.

- *modal verbs:* modal verbs (e.g., "could", "should") work as indicators of claims since they indicate that what is expressed in a proposition is not unassailable, but might be otherwise (Palau and Moens, 2009). They, thus, frequently appear in propositions B of argumentative concessions which constitute the speakers' standpoints. We define a Boolean feature which indicates if a candidate example contains a modal verb.

- *hedges:* Hedges are linguistic devices used to mitigate the speaker's commitment to the truth of a proposition (Hyland, 1996a), i.e., "I *tend to* accept". They include possibility modals next to other linguistic items expressing the degree of speaker's certainty. They can be ambiguous depending on constructional features: the propositional attitude indicator *I think* works, for instance, as a hedge in parenthetical constructions (e.g., "It is not worth buying it, *I think*"), while it merely signals subjectivity when used as a main verb (e.g., "I think it is not worth buying it"). The use of hedges is common in argumentative concessions since they contribute to avoid a potentially face-threatening act of abrupt disagreement. Tan et al. (2016) argue that depending on the context, hedges can make an argument weaker or easier to accept by softening its tone. Based on their research and also on Hanauer et al. (2012), we collect a set

of candidate hedge cues and use them as Boolean features (presence or absence of a hedge word).

- *Jaccard Similarity:* In argumentative concessions proposition A always expresses positive sentiment towards a claim expressed by the previous speaker. Thus, we use Jaccard Similarity to measure lexical similarity between the sentence/sentences containing the candidate concession and sentences from the original post OP (we removed stopwords). We use the *maximum* similarity value as a feature.

- *sentiment feature:* By definition, argumentative posts tend to contain opinion on the other posts. A post that is tagged with subjectivity will be a useful feature to identify concessions. We use (a) the MPQA Lexicon (Wilson et al., 2005) of over 8,000 positive, negative, and neutral sentiment words, (b) an opinion lexicon with around 6,800 positive and negative sentiment words (Hu and Liu, 2004) to see whether training instances contain sentiment words.

Apart from the above features, we also retrieved lexical patterns that could be indicators of argumentative concessive uses of the discourse markers. These patterns are used in proposition A to express a positive evaluation about another speaker's claim. They could assist in achieving higher precision in identifying *arg_c* instances. Below is a short description of the semi-automatic retrieval of lexical patterns.

Semi-automatic retrieval of lexical patterns We present a bootstrapping algorithm that automatically learns lexical patterns expressing argumentative concessions. The algorithm begins with only two seed phrases – "I agree" and "you are right"– the most common two patterns expressing argumentative concessions according to annotation results done by an expert on a separate development set of one-hundred utterances (not used in the above annotation studies). We used 80% of our ChangeMyView data (i.e., all except the dev as well as the $test$ dataset) for the bootstrapping algorithm. The algorithm operates on a simple structural assumption: In argumentative concessions, words forming the proposition positioned before the marker (for "but" and "though") or in the marker's scope (for "while") express agreement with a point made in a preceding comment. To identify such patterns, first, we retrieve these words (i.e., all words in the proposition before "but"). Second, we extract all possible trigrams, four-grams, and five-grams from this set of words and search for the two seed phrases in the ngrams. For instance, we identify all instances of "[. . .] I [. . .] agree [. . .]", where [. . .] represent zero or more occurrences of words. We obtain patterns such as "I agree [completely]" or "[I think] you are right [about]". Third, using each new pattern we attempt to identify new lexical constructions. For example, given the pattern "I agree [completely]" we search for the occurrences of "I [. . .] completely" where [. . .] represent zero or more occurrences of any words, excluding negation. With this search, we arrive at new patterns, such as "I [understand] completely" and "[I think] you are [correct]".

Since this method relies on structural syntactic similarity, we embed the following semantic rules: (i) keep only patterns, which contain propositional attitude indicators (i.e., verbs "think", "realize", and the constructions "be right", "be correct") or indicators of sentiment (i.e., verbs "love", "like") and (ii) select the patterns which contain the pronoun "you" or the adjective "your"– to retain just those seeds where the target of agreement is an opinion held by another poster. Finally, in the fourth stage, using the new lexicon we again resume the search for new patterns (i.e, using "[. . .] I [. . .] realize [. . .]") and this process continues until we do not find any new patterns. As a result of the bootstrapping algorithm we obtain 329 lexical patterns, which we call $B\_Lexicon$. We, then,

Manually filtered lexicon
I would agree with you
I fully agree that
I see what you
I see where you
I think you are correct

Table 2: Examples of manually filtered lexicon

manually filter the bootstrapped lexicon to eliminate redundancies, merging patterns which instantiate the same linguistic constructions and adjustments based on the dev set (e.g., removing double negatives "I don't disagree"). As a result of manual filtering we finally obtained 116 lexical patterns indicating argumentative concessions ($B\_Lexicon_{MF}$). Table 2 shows some examples from this manually filtered lexicon.

6.2 Self-training method for identifying concessions

Self-training algorithms (Mihalcea, 2004; Clark et al., 2003) start with a small subset of annotated training data and attempt to increase the amount of training data by using a large set of unannotated data. We adopt the approach of Mihalcea (2004) that used self-training as "a tagger that is retrained on its own labeled cache on each round" (Mihalcea, 2004; Clark et al., 2003). Similar to Mihalcea (2004), we start with a small set of labeled data that is the training data (L), in our case the data labeled by the expert annotator consisting of 980 instances. We build a classifier using the linguistically-motivated features described above, and then apply the learned model on a set of unlabeled data (U). For our experiments, the U is the 70% of the CMV data (i.e., apart from the $training$, $test$, and dev sets; each is 10% of the CMV data). Now, instead of classifying directly on the total set of data U, we split U in P random pools where each pool contains U' unlabeled instances. From each P pool we select only those G_c argumentative concession instances and G_nc non-argumentative concession instances with a labeling confidence exceeding a particular threshold. Similar to Mihalcea (2004), while adding the new instances to the training data L, we maintain the original class distribution of training data between *arg_c* and *other* categories. The threshold could be a preset value of probability; in this experiment, we varied the number of data instances to select while keeping the original class distribution. These instances (i.e, G_c and G_nc) in turn are added to the original training set L. The classifier is retrained with the new set of training data (i.e, L + G_c + G_nc) and this process continues for all the P pools. Note, in each step we also evaluate the classifier on the dev set to assess the quality of the new data that is added to the training set L (Table 3). For details of the self-training procedure, please see Mihalcea (2004); Clark et al. (2003). We use the Support Vector Machines (SVM) classifier with RBF kernel (we use Scikit-learn tool (Pedregosa et al., 2011)). The class weights are inversely proportional to the number of instances in the categories.

As a final classifier we used a system combination: if any dev/$test$ instance contains a lexical pattern obtained via bootstrapping described in the previous section, we classify that instance as *arg_c*, otherwise, use the decision of the self-training classifier. The dev data is used to fine tune all parameters in our experiment and to choose the *best* parameters (see next section for detailed discussion).

Self-training Setting		Optimal size (training)		Performance (Max. F1)		
Pool size	G_c	*arg_c*	*other*	P	R	F1
50	10	430	952	56.9	**57.5**	**57.2**
100	10	441	821	56.7	57.0	56.8
1000	10	289	811	**65.9**	45.8	54.0
2000	10	229	751	**65.3**	44.6	53.0
100	50	442	964	64.5	51.3	**57.4**
500	50	424	946	63.9	48.0	54.8
1000	50	415	937	64.1	47.5	54.5
2000	50	279	801	63.1	46.3	53.4

Table 3: Experimental results of the Self-training method on the dev set (**bold** are best scores)

7. Experiments and Results

Before reporting the results of our self-training classifier and our baselines, we report the results of an off-the-shelf parser that aims to label the *types* of discourse connectives (Biran and McKeown, 2015), noted as Ob_{parser} in Table 4. Biran and McKeown (2015) have trained the parser on the Penn Discourse Treebank (PDTB) corpus (Miltsakaki et al., 2004) using discourse connective features, lexical features, syntactic features, etc. Their model can identify the *types* of discourse relations, such as *Concession*, *Contrast*, *Pragmatic Concession*, and *Pragmatic Contrast*, which are *types* of the *class* of discourse relation *Comparison*. The off-the-shelf parser results in a low F1 measure of only 6 with precision of 13.2 and recall of 4 for the dev data.[4] This low performance is not unexpected. First, the parser is trained on the PDTB corpus, which is based on of Wall Street Journal (WSJ) articles and the language is vastly different from the ChangeMyView subreddit. In addition, PDTB has a small number of concessions in general and it is not clear how many of those are argumentative concessions, if any. It has to be noted that state of the art discourse parsers do not take into account different pragmatic values underlying a discourse relation. The level of granularity required to classify different types of concessions is higher than that necessary to classify what discourse relation is conveyed by a single connective. Nevertheless, the performance achieved by our system is comparable to that achieved by Biran and McKeown (2015) in the classification of explicit discourse relations (56.91 F1).

Baselines. We use two baselines: 1) a rule-based system based on the lexical patterns learned via bootstrapping ($B\_Lexicon_{MF}$ in Table 4), and 2) the system combination that uses the $B\_Lexicon_{MF}$ and a SVM classifier that uses all the features but without the self-training process (SVM_{noST}). The performance of both of these baselines is better on the dev set than on the $test$ set, which is expected as dev set was used to check the final pattern lexicon. The recall of the $B\_Lexicon_{MF}$ patterns is low for the $test$ data, meaning that there are different patterns not covered in our collection of lexical patterns and argumentative concessions can exist in many ways that are not retrieved by simple lexical rules.

Self-training. Table 3 show the results of the system combination (lexical patterns and SVM classifier *using self-training*) on the dev set, used to decide the best parameters (pool size and the

4. Since the performance of the off-shelf-parser on the dev data is very low compared to the other methods we did not conduct any further experiment on the $test$ set.

Computational Model	Training size (*arg_c*;*other*)	*dev*			*test*		
		P	R	F1	P	R	F1
Ob_{parser}	-	13.2	4.0	6.0	-	-	-
SVM_{noST}	(229;751)	65.3	44.6	52.9	35.1	**58.9**	44.0
$B\_Lexicon_{MF}$	-	**65.5**	43.9	52.5	**48.3**	25.0	32.9
Self-training (best)	-	64.5	**51.3**	**57.4**	38.0	58.3	**46.0**

Table 4: Experimental results of classifiers on the dev and $test$ set. The self-training results report the best (in **bold**) performing parameters.

number of instances to add to the training set). In column one (i.e., *pool size*) we report the number of random unlabeled instances (i.e, U') that were tested via the classifier. Column two represents the maximum number of instances (G_c) that are added to the training data. For Recall, we also add G_nc to maintain same class distribution between *arg_c* and *other* categories. The next two columns show the size of training data from our two classes respectively that achieve the highest F1 scores for the dev set. We start the experiments with the labeled training data by expert annotators (also used in the baselines) and gradually add G_c and G_nc that are predicted by the classifier. Finally, the last three columns show the P/R/F1 on the dev set. We report all the columns based on the highest F1 achieved by the models. For instance, in the first row of Table 3, the pool size is 50 and G_c is 10. The size of the unlabeled data is close to 8,400 (after removing duplicates), which means here the number of pools is 168 (8400/50). Now, we classify each pool from the set of P, and depending upon the classification result, we add G_c (here, maximum is 10 instances) and G_nc accordingly from each pool. After the classifier has evaluated a certain number of pools the F1 of 57.2 is achieved. Meanwhile, starting from 229 arg_c, the size of the arg_c is now 430. We also observe a common trend between the various runs of the self-training procedure; the accuracy increases till a certain number of G_c + G_nc is added to the original training set of L, but after that the performance drops probably due to added noise.The best model based on the dev set, has the pool size of 100 and G_c of 50 reaching F1 of 57.4 that is close to a 5% improvement compared to the baseline results without self-training on the dev set (see SVM_{noST} from Table 4).

We used this best setting (i.e., pool size of 100 and G_c of 50) to run our system combination using self-training on the $test$ set. We also used the dev data for feature selection. We employ χ^2 based feature selection and observe that the "top 300" features (based on χ^2 scores) performed best for the dev data. Subsequently this setting (i.e, "top 300" features) is used for the $test$ set. This feature set of "top 300" features contains modal verbs such as "may", "could"; Jaccard similarity value; pronouns such as "I", "your'; hedges such as "almost", "probably", "somewhat"; words such as "greatest", "excitement", "interesting" from the sentiment lexicons; unigrams such as, "agree", "recognize", "argument", "think", and finally, bigrams such as, "while you", "argument is", "i absolutely", to name a few. The results of self-training show improvement over the baselines, but to a lesser degree when compared to SVM_{noST} (2%), and a higher degree when compared to $B\_Lexicon_{MF}$ (13.1%) (Table 4).

To understand the quantitative results better, we have randomly selected and qualitatively analyzed a sample of 135 classified occurrences from the development data. We compared recurrent characteristics both of occurrences that the system classified *arg_c* in accord with gold data and those that it classified as *arg_c*, while the human annotators did not. It turns out that occurrences

classified as *arg_c* both by the system and by the annotators tend to include lexical patterns, which unambiguously express agreement paired with second person pronouns/adjectives which anaphorically point to a previous claim (i.e. "I do *agree* with a lot of points in *your* post, but there is a huge disconnect between the title of your post"). Looking at occurrences that have been classified by the expert annotator as *arg_c* while considered by the system as *other*, they tend to lack an unambiguous expression of agreement and an anaphoric reference to a preceding post. More specifically, agreement is expressed through modal adverbs expressing different degrees of certainty (i.e. "*Possibly*, but not because they're missing out on experiences"; "*Of course* it is but if you're okay with the humane treatment of animals to include population control why are you uncomfortable with the humane treatment of animals in other contexts as the end result is the same?"). However, the state of affairs over which the modal adverbs have scope is elliptical, since it is shared as given information by the participants to the discussions (speakers as well as readers). In classification experiments ellipsis is a hard to detect phenomenon since it is dependent on the pragmatic content of the utterance.

8. Discussion: are concessions persuasive strategies?

To test the theoretically-informed hypothesis that argumentative concessions work as persuasive strategies, we look at their distribution in the Δ awarded vs. no-Δ awarded comments. First, we consider the manually labeled data. The small set of training data annotated by the expert annotator, as well as the dev and $test$ sets annotated in the crowdsourcing experiment. The distribution of argumentative concessions in the Δ awarded vs. no-Δ awarded comments on all these datasets is given in Table 5. Second, we look at the distribution when considering the argumentative concessions *predicted* by our best classifier (system combination with self training). We look at the dev and $test$ set, as well as the rest of the unlabeled CMV corpus (Table 6). As a caveat, we have to point out that the due to the low accuracy of classifier, the observed distribution based on the classifier predictions cannot be considered reliable. For example, the predictions of the computational model on the dev and $test$ set miss to identify more than 1/3 of argumentative concessions compared to the gold annotations.

Marker	**Δ-training**	**no-Δ-training**	**Δ-dev**	**no-Δ-dev**	**Δ-$test$**	**no-Δ-$test$**
but	39	59	68	83	83	82
however	25	3	3	5	-	-
though	22	27	4	8	2	1
while	13	13	2	5	-	-
total	99	130	78	101	85	83

Table 5: Distribution of argumentative concessions in Δ and no-Δ comments in the manual labeled datasets

Overall the numbers suggest a fairly equal distribution of argumentative concessions in the Δ and no-Δ comments both in the manually labeled data and in those predicted by the classifier. We have tested the statistical significance of the distribution of argumentative concessions on the manually labeled sets using χ^2 test: the results are not significant at $p < 0.05$ on $training$ and $test$

Marker	Δ-dev	**no-**Δ-dev	Δ-$test$	**no-**Δ-$test$	Δ**-unlabel**	**no-**Δ**-unlabel**
but	49	57	51	47	1100	793
however	2	4	-	-	-	-
though	4	5	-	1	8	2
while	1	2	-	-	-	-
total	56	68	52	48	1112	795

Table 6: Distribution of *predicted* argumentative concessions in Δ and no-Δ comments

sets, while they are significant on dev set, where argumentative concessions are less frequent in winning arguments.

These results suggest that argumentative concessions do not increase persuasion in ChangeMyView, challenging the assumptions made in the rhetorical literature. However, they do not allow us to draw conclusions scalable to different contexts about the persuasive role played by this type of concessions. They rather provide further confirmation that the persuasive value played by lexical meta-discursive features is context-bound and crucially depends on the rhetorical situation. For example, hedges seem to increase persuasiveness in scientific writing (Hyland, 1996b), while they decrease it in other kinds of messages (Blankenship and Holtgraves, 2005). When it comes to argumentative concessions, they embody what in the philosophical tradition has been called *principle of charity* or *principle of rational accommodation* according to which "we make maximum sense of the words and thoughts of others when we interpret in a way that optimizes agreement" (Davidson, 2001): conceding the claims made by the author of the original post the speaker shows his intention of interpreting them in the best possible way. In doing so he reinforces his ethos, presenting himself as a reasonable discussant. At the same time, he minimizes the risk of a face-threatening opposition which could arise from the apparent incompatibility with the statement expressed in the nucleus (proposition B). The principle of charity has, in fact, to be understood as a methodological presumption that guarantees the understanding of another point of view in its argumentatively strongest form to allow a possibly adequate critique and reach agreement through persuasion. This principle is structural in ChangeMyView, being at the core of the subreddit mission of resolving differences of opinions starting from a deep understanding of them. Users who choose to write on this subreddit have to respect the submission rules and are, thus, by default charitable. Therefore, linguistic strategies such as argumentative concessions that encode the principle of charity constitute persuasive strategies in the speakers' mind, but are plausibly perceived as routinized expressions by the addressees. In other words, they do not shape the argumentative profile of single users, but are conventional strategies belonging to the activity type envisioned by ChangeMyView.

As stated in the introduction, the lack of corpora with explicit signals as to what pieces of discourse achieved persuasion makes it difficult to investigate the persuasive roles played by concessions in datasets belonging to different discourse genres and dialogue activity types. To our knowledge, beside the ChangeMyView subreddit, the only other corpus labeled with persuasive features is the Yahoo News Annotated Comments Corpus (YNACC) (Napoles et al., 2017a). This corpus contains around 140K threads, taken from the comments sections of Yahoo News articles as well threads from the Internet Argument Corpus. Discussion posts are annotated by expert and untrained (i.e., Turkers) annotators with specific labels, both at the thread (e.g., constructiveness) and at the comment level (e.g., topic). Among the latter type, persuasiveness is defined as a "A

binary label indicating whether a comment contains persuasive language or an intent to persuade." Differently from *Delta points* in ChangeMyView, the persuasiveness label does not inform us about what discourses achieved persuasion among the participants of an actual interaction, but provides hints as to what type of language is perceived as persuasive by third parties. In other words, we experiment with this dataset to assess whether argumentative concessions are perceived by language users as persuasive strategies, namely strategies used by speakers as an attempt to persuade, despite the actual pragmatic outcome. We selected a subset of YNACC that are annotated by expert annotators. This subset contains 4,719 posts that are annotated with the label "persuasive" and 17,616 posts that are annotated with the label "not persuasive". From this subset we evaluate only the posts that contain the selected markers for our experiments, "but", "while", "though", and "however".

We use the same training data (described in Section 6) to predict the binary label *arg_c* and *other* from the YNACC posts. We use the same features that are described in Section 6.1 except the *Jaccard Similarity* feature since the YNACC dataset does not indicate which post is a reply to another post. After removing the duplicates from the YNACC posts we observe that out of 649 persuasive posts, 321 (49.4%) are classified as *arg_c* (i.e., concessions) whereas out of 1,263 non persuasive posts, 417 (33%) posts are classified as concessions. According to these results, argumentative concessions are deemed as persuasive strategies, since they correlate with the persuasiveness label. In Table 7 we present the count of the main expressions of concessions across the persuasive and the non persuasive datasets: "?" depicts one or zero occurrence of the particular pattern. For example, "pattern_acknowledge" shows that any occurrence of the expression, "I" and following either of "concede", "acknowledge", and "think" is regarded as a concession of the pattern "pattern_acknowledge". Note in the above case, "also" or "too" can appear between the above two words. Other patterns of concessions also exist (e.g., "I appreciate your") but the frequency is very

Explicit_expressions (patterns)	Description	Count	
		persuasive	not persuasive
pattern_yes	[yes|sure|of course|correct|right|true] [,]	24.81	13.62
pattern_acknowledge	[I] ?[also|too] [concede|acknowledge|think]	12.63	9.58
pattern_see	I ?[adverb|modal] [see|get] [?] [you|your]	8.47	4.51
pattern_agree	I ?[adverb|modal] [agree]	0.77	0.63

Table 7: Distribution (in %) of explicit expressions (patterns) of concessions in persuasive and not persuasive instances in the ERIC corpus

low. Overall, it seems that regardless the type of linguistic construction at work, concessions are more frequent in persuasive threads.

9. Conclusion and future work

We tackled the task of empirically validating the theoretically assumed persuasive role played by specific discourse relations, *concessions*, using the ChangeMyView subreddit platform. Drawing from a linguistically-informed typology of concessions, we singled out one type of concessions that prototypically bears an argumentative value. We focused on four discourse markers that constitute 85% of the overall occurrences of potential markers in our data: *but*, *though*, *however* and *while*. We present a computational model based on self-training using linguistically motivated features. We achieve a moderate F1 of 57.4% via the self-training method on the development set and 46.0% on the $test$ set. Our findings both from the manual labeling (both expert and crowdsourcing annotation) and the system predictions indicate that the type of argumentative concessions we investigate is almost equally likely to be used in winning (Δ-awarded) and losing (no-Δ) arguments. While this result seems to contradict theoretical assumptions, we provided some reasons related to the ChangeMyView subreddit. This behavior shows that text-genre rules have to be taken into account in the interpretation of rhetorical patterns: When perceived as conventional genre-specific rules, persuasive strategies may happen not to be pragmatically effective. In future work, we plan to validate this explanation by the following three steps: 1) improving the performance of our computational models by collecting a larger dataset as training data to be able to run the distribution over a much larger dataset; 2) looking at other types of argumentative concessions; and 3) retrieving argumentative concessions in different text-genres.

Acknowledgement

This paper is based on work supported partially by the Advanced Post Doc SNFS Grant for the project From semantics to Argumentation Mining: a Context-Independent Lexicon of Indicators of Argumentative Discourse Relations and DARPA-DEFT program. The views expressed are those of the authors and do not reflect the official policy or position of the SNFS, Department of Defense or the U.S. Government. We would like to thank the annotators for their work and the anonymous reviewers for their valuable feedback.

References

Charles Antaki and Margaret Wetherell. Show concessions. *Discourse studies*, 1(1):7–27, 1999.

Moshe Azar. Concession relations as argumentation. *Text-Interdisciplinary Journal for the Study of Discourse*, 17(3):301–316, 1997.

Or Biran and Kathleen McKeown. PDTB discourse parsing as a tagging task: The two taggers approach. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 96–104, 2015.

Kevin L Blankenship and Thomas Holtgraves. The role of different markers of linguistic powerlessness in persuasion. *Journal of Language and Social Psychology*, 24(1):3–24, 2005.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

Stephen Clark, James R Curran, and Miles Osborne. Bootstrapping POS taggers using unlabelled data. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4*, pages 49–55. Association for Computational Linguistics, 2003.

Elizabeth Couper-Kuhlen and Sandra A Thompson. Concessive patterns in conversation. *Topics in English Linguistics*, 33:381–410, 2000.

Elizabeth Couper-Kuhlen and Sandra A Thompson. A linguistic practice for retracting. In Hakulinen Auli and Margret Selting, editors, *Syntax and Lexis in Conversation: Studies on the Use of Linguistic Resources in Talk-in-interaction*, volume 17, pages 257–288. 2005.

Donald Davidson. *Inquiries into Truth and Interpretation: Philosophical Essays*, volume 2. Oxford University Press, 2001.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 549–554, 2016.

Nancy L Green. Representation of argumentation in text with rhetorical structure theory. *Argumentation*, 24(2):181–196, 2010.

Brigitte Grote, Nils Lenke, and Manfed Stede. Ma(r)king concessions in English and German. *Discourse Processes*, 24(1):87–117, 1997.

Ivan Habernal and Iryna Gurevych. Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1589–1599, 2016.

David A Hanauer, Yang Liu, Qiaozhu Mei, Frank J Manion, Ulysses J Balis, and Kai Zheng. Hedging their mets: the use of uncertainty terms in clinical documents and its potential implications when sharing the documents with patients. In *AMIA Annual Symposium*, page 321, 2012.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177. ACM, 2004.

Ken Hyland. Talking to the academy: Forms of hedging in science research articles. *Written communication*, 13(2):251–281, 1996a.

Ken Hyland. Writing without conviction? hedging in science research articles. *Applied linguistics*, 17(4):433–454, 1996b.

Mitsuko Narita Izutsu. Contrast, concessive, and corrective: Toward a comprehensive study of opposition relations. *Journal of Pragmatics*, 40(4):646–675, 2008.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(02):151–184, 2014.

William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

Rada Mihalcea. Co-training and self-training for word sense disambiguation. In *CoNLL*, pages 33–40, 2004.

Eleni Miltsakaki, Rashmi Prasad, Aravind K Joshi, and Bonnie L Webber. The Penn Discourse Treebank. In *LREC*, pages 2237–2240, 2004.

Elena Musi. How did you change my view? A corpus-based study of concessions argumentative role. *Discourse Studies*, pages 1–19, 2017.

Courtney Napoles, Aasish Pappu, and Joel R Tetreault. Automatically identifying good conversations online (yes, they do exist!). In *ICWSM*, pages 628–631, 2017a.

Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 13–23, 2017b.

Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 98–107. ACM, 2009.

Rebecca J Passonneau and Bob Carpenter. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326, 2014.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Chaim Perelman. The new rhetoric. In Hillel YehoshuaBar, editor, *Pragmatics of natural languages*, pages 145–149. Springer, 1971.

Emily Pitler, Annie Louis, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, pages 683–691. Association for Computational Linguistics, 2009.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse Treebank 2.0. In *Proceedings of LREC*, 2008.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. Reflections on the penn discourse treebank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950, 2014.

Swapna Somasundaran and Janyce Wiebe. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, pages 226–234. Association for Computational Linguistics, 2009.

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

Reid Swanson, Brian Ecker, and Marilyn A Walker. Argument mining: Extracting arguments from online dialogue. In *SIGDIAL Conference*, pages 217–226, 2015.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624. International World Wide Web Conferences Steering Committee, 2016.

Mehmet Ali Uzelgun, Dima Mohammed, Marcin Lewiński, and Paula Castro. Managing disagreement through yes, but constructions: An argumentative analysis. *Discourse Studies*, 17(4):467–484, 2015.

Frans H Van Eemeren, Peter Houtlosser, and AF Snoeck Henkemans. *Argumentative indicators in discourse: A pragma-dialectical study*, volume 12. Springer Science & Business Media, 2007.

Frans H Van Eemeren, Rob Grootendorst, Ralph H Johnson, Christian Plantin, and Charles A Willard. *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*. Routledge, 2013.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. A corpus for research on deliberation and debate. In *LREC*, pages 812–817, 2012.

Zhongyu Wei, Yang Liu, and Yi Li. Is this post persuasive? Ranking argumentative comments in the online forum. In *The 54th Annual Meeting of the Association for Computational Linguistics*, pages 195–200, 2016.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics, 2005.

Christopher R Wolfe, M Anne Britt, and Jodie A Butler. Argumentation schema and the myside bias in written argumentation. *Written Communication*, 26(2):183–209, 2009.

Joel Young, Craig Martell, Pranav Anand, Pedro Ortiz, and Henry Tucker Gilbert, IV. A microtext corpus for persuasion detection in dialog. In *Proceedings of the 5th AAAI Conference on Analyzing Microtext*, AAAIWS'11-05, pages 80–85. AAAI Press, 2011.