Annotating Claims in the Vaccination Debate

Benedetta Torsi, Roser Morante
Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam
The Netherlands
b.torsi@student.vu.nl,r.morantevallejo@vu.nl

Abstract

In this paper we present annotation experiments with three different annotation schemes for the identification of argument components in texts related to the vaccination debate. Identifying claims about vaccinations made by participants in the debate is of great societal interest, as the decision to vaccinate or not has impact in public health and safety. Since most corpora that have been annotated with argumentation information contain texts that belong to a specific genre and have a well defined argumentation structure, we needed to adjust the annotation schemes to our corpus, which contains heterogeneous texts from the Web. We started with a complex annotation scheme that had to be simplified due to low IAA. In our final experiment, which focused on annotating claims, annotators reached 57.3% IAA.

1 Introduction

Argumentation is an important aspect of human communication. The study of argumentation is an interdisciplinary research field that has been gaining momentum because of its relevance in cognitive sciences and its application in artificial intelligence. Because of the richness of information that it offers, data from the Web such as social media, on-line newspapers, forums, or blogs is often the subject of exploration (Lippi and Torroni, 2016). The availability of such data and the advancements in computational linguistics fostered the rise of a new research field called argumentation mining (AM) (Peldszus and Stede, 2013a), whose goal is to automatically extract argument components from text, generating structured data for computational models of argument.

Thus far, most corpora annotated with argumentation information are composed by a certain type of texts, such as argumentative essays (Stab and Gurevych, 2017) and news editorials (Al Khatib et al., 2016), which usually have a specific structure. However, in order to understand public opinions, it is necessary to process textual data from the Web, which is generated by a diversity of users who do not follow a predefined template, resulting in texts of miscellaneous genres and registers.

Our research focuses on annotating argumentation components in a corpus of texts crawled from the Web. We chose this type of texts because they contain precious information about people's opinions and because existing argumentation schemes[1] are not directly applicable to Web data. Thus, one of our aims was to evaluate whether existing argumentation schemes can be applied to this data. Additionally, we focused on texts about the vaccination debate because of the importance of the topic from a societal and health perspective. A longer term goal of our work is to find out and analyze the claims that people make when they engage in the debate. Potential applications of both the analysis and the annotated data would be, for example, detecting misinformation about vaccinations to populate ontologies (Amith and Tao, 2018) or training agents (Rosenfeld and Kraus, 2016) to persuade hesitant users to vaccinate.

The vaccination debate is an exemplary case for the study of on-line debates and opinion forming processes. In 1998, a scientific paper was published by Andrew Wakefield, who argued that there was a link between the Measles Mumps and Rubella (MMR) vaccine and autism in children. The paper incited skepticism in the public about the safety and the effectiveness of vaccines, to such extent that more an more parents are deciding not to vaccinate their children, causing illnesses such as measles to spread and leading to a measles outbreak in the EU in 2017. Nowadays,

[1] By "argumentation schemes" we mean annotation schemes that have been used to annotate argumentation components.

Proceedings of the 5th Workshop on Argument Mining, pages 47–56

with collaborative media, anyone can join in a discussion and share information and opinions. This makes it difficult to attest the reliability of on-line content (Zummo, 2017). Reports show that statistically significant positive correlations exist between monthly Measles cases and Google queries in most EU28 countries from January 2011 to August 2017 (Mavragani and Ochoa, 2018). Since the debate on immunization directly affects public health and safety, analyzing the way people engage in this discussion is of particular interest.

In this paper we present three annotation experiments with different annotation schemes. Our annotation studies were all performed on a corpus of texts related to the vaccination debate (Vax Corpus), which is composed of 294 documents automatically downloaded from the Internet adding up to 445,574 words. The documents that compose the corpus are heterogeneous in length, genre, and style: blog posts, editorials, news articles, and science articles. Below are some examples of statements found in the corpus:

1. These are child who can't be vaccinated. Children who have cancer. Children who are immunocompromised. Children who are truly allergic to a vaccine orpart of a vaccine (i.e anaphylaxis to egg). These children remain at risk. They cannot be protected ... except by vaccinating people around them.
2. Better believe if I ever have kids I won't be vaccinating.
3. Of course we were already blaming antivaxxers for bringing back measles and spreading them around Disneyland, but a fresh new study confirms that yup, is it definitely scientifically their fault, so let's blame them even harder now.
4. As we have already yelled, in all caps and with many exclamation points, JUST VACCINATE YOUR KIDS, DAMNIT!!!

After experimenting with several annotation schemes, we concluded that in order to model arguments in Web data, a simplified scheme is needed. We cannot rely on fine-grained models based on argumentation theory, as they are not directly applicable to heterogeneous texts. The simplified scheme was adopted for the annotation of our corpus, resulting in the annotated resource *VaxClaim Corpus*, which will be made publicly available. Since the simplified scheme does not contain domain-specific features, it should be applicable to other on-line debates. We will test this hypothesis in future work.

Section 2 presents related work. In Section 3 we introduce the pilot annotation studies and we discuss the results and main sources of disagreement. In Section 4 we describe the final annotation experiment, which focused on *claims*, and we also analyze the results and disagreements. Finally, in Section 5 we provide some conclusions.

2 Related Work

Numerous models have been developed to address and understand the internal (micro) structure of arguments (Lippi and Torroni, 2016). Walton adopted the notion of *argumentation scheme*, which allows to identify patterns in the arguments present in everyday discourse (Bentahar et al., 2010). Since then, several argumentation schemes have been put forward. For a general overview of argumentation schemes, refer to Lippi and Torroni (2016). Here we focus on the schemes used to annotate data for argumentation mining purposes.

Saint-Dizier followed a knowledge driven approach to tackle the task of mining arguments in Web data and concluded that it is possible to relate independent statements by means of lexical data, domain knowledge and reasoning schemes (Saint-Dizier, 2016).

Stab and Gurevych created a corpus of 402 argumentative essays selected from essayforum.com and annotated it with the following argument components: *major claims*, *claims*, and *premises* (Stab and Gurevych, 2017). They model the microstructure of arguments as a connected tree structure where the *major claim* is the root node which represents the author's standpoint. The *major claim* is expected to be contained either in the introduction or the conclusion of the essay. The rest of the essay contains *claims* (the cores of the arguments) and *premises*, which support the *claims*.

Al-Khatib et al. (2016) consider that in the editorial genre, the author generally does not only aim at persuading the audience, but she also wants to spread information about the topic. The author defends a thesis that conveys a stance on a controversial subject providing different kinds of evidence. They constructed a corpus by extracting 100 editorials from each of the following websites: aljazeera.com, foxnews.com and theguardian.com. They introduce an annotation task which consisted

of dividing the corpus into segments. Then, each segment is assigned one of the following labels:

- *Common ground*: the segment contains a self-evident fact, it states common knowledge.
- *Assumption*: the segment contains a conclusion, an opinion or a judgment of the author.
- *Testimony*: the segment contains a statement made by some expert, witness or authority.
- *Statistics*: the segment contains the results of a quantitative study or data analyses.
- *Anecdote*: the segment expresses a personal experience, a specific instance, a concrete example.
- *Other*: the segment is not classifiable with any of the above classes.

These two argumentation schemes were adopted in the first pilot study presented in this work because the documents composing the Vax Corpus present characteristics of both argumentative essays and news editorials.

Habernal and Gurevych (2017) created a corpus of user-generated Web content collecting documents of different registers, such as articles, comments on articles, blog posts, forum posts, etc. Their scheme is based on the Toulmin model (Toulmin, 2003) and it is characterized by the following components argument components:

- *Claim*: the conclusion that the author is trying to establish.
- *Grounds*: the evidence and reasoning that constitute the foundation of the claim.
- *Backing*: the set of information that conveys the trustworthiness of the warrant.
- *Qualifiers*: they express the degree of cogency attached to the claim.
- *Rebuttal*: a statement expressing an instance in which the claim might be defeated.
- *Refutation*: a section of text that attacks the *Rebuttal*.

Their scheme was adopted for the second pilot study because it was designed to fit Web data. The belief was that the resulting annotation task would be feasible for the corpus at hand and that the scheme would allow for the identification of interesting argumentation patterns.

3 Annotating Argument Components: Pilot Studies

One of our research goals is to automatically extract claims related to the vaccination debate. In order to train a system, we first needed to annotate a corpus. We performed three annotation studies to test the feasibility of different annotation schemes. The first two studies served as pilots to discover strengths and weaknesses of previously proposed annotation schemes. For the two pilot studies, annotators were asked to annotate texts for 3 and 10 hours respectively. For the final annotation task, they were asked to annotate 100 randomly-selected documents from the Vax Corpus.

The annotation tasks were carried out by two annotators (A and B) with a background in Linguistics. Annotator A was more experienced with the topic of argumentation. Annotations were performed with the open source annotation tool eHOST, [2] which also provides options to calculate inter annotator agreement (IAA). IAA is calculated in eHOST by dividing the annotation matches by all annotations (matches + non-matches). IAA was calculated with lenient matching in order not to penalize disagreements due to details such as punctuation.

3.1 First Pilot Study

We observed that some documents in the Vax Corpus present characteristics of argumentative essays. In particular, arguments are expressed in a hierarchical structure where there is a main claim, one or several sub-claims and premises providing backing for the claims. This is why for the first pilot study, we adopted the scheme put forward by Stab and Gurevych who found that, in argumentative essays, arguments are often characterized by a tree structure (Stab and Gurevych, 2017). Their markables were deemed appropriate and they were all adopted: *major claim*, *claim*, *premise*. Their approach had to be modified, however, to fit the

[2] http://blulab.chpc.utah.edu/content/ehost-extensible-human-oracle-suite-tools

characteristics of the Vax Corpus: more than one *claim* was allowed per paragraph. Furthermore, we allowed the *attack* relationship between *claim* and *major claim*. That means that the *major claim* does not have to be supported by all the *claims* present in the text and that some *claims* might play the role of *rebuttal*. The *rebuttal* presents opposing views to the author's *claim* and it is often presented with the intent of criticizing it, thus strengthening the claim (Habernal and Gurevych, 2017).

We also observed that the type of backing given by the authors in the Vax Corpus often fits with the descriptions of argument components proposed by Al-Khatib et al. (2016). We adopted their scheme with some modifications: (i) The definition of *assumption* seemed similar to the one of *claim*: they both constitute some sort of conclusion or judgment of the author. We left out this component, as it would overlap too much with the concept of *claim*. (ii) The definition of *common ground* seemed difficult to interpret and apply. It is unclear what should be considered common knowledge and what should not; consequently we left out this unit. (iii) The remaining units, *testimony*, *statistics*, *anecdote* and *other* correspond to the types of premise that were observed in the Vax Corpus. Furthermore, since according to the adopted definition of *claim*, the core statement of the argument can be attributed exclusively to the author of the text, the units *testimony* and *statistics* could only coincide with the *premise*. The *anecdote* corresponds to a recounting of a specific episode, thus it is uncommon that it fits the definition of *claim*. Considering all the factors mentioned above, we selected the four components *testimony*, *statistics*, *anecdote* and *other* as attributes for the component *premise*.

3.1.1 Results

The annotators were able to tackle one file per hour which resulted in 3 annotated documents. Table 1 presents the IAA scores. *Major claim* reached the highest agreement score (66.7%), but the annotators did not agree on the annotation of *premise* in any of the cases. Next we discuss cases of disagreement.

Major claims. An analysis of the disagreements revealed that it is difficult to establish which statement best summarizes the stance of the author, as exemplified by the example below, where each annotator marks a different *major claim* for the same text:

Type	IAA	Matches	Non-matches
All	15.8%	6	77
Claim	6.9%	2	27
Premise	0.0%	0	48
MC	66.7%	4	2

Table 1: Two-way IAA Results (Pilot study 1). MC stands for "Major Claim".

- **Annotator A**: If these diseases seem uncommon - or even unheard of - it's usually because these vaccines are doing their job.
- **Annotator B**: Childhood vaccines protect children from a variety of serious or potentially fatal diseases, including diphtheria, measles, polio and whooping cough (pertussis).

From Annotator A's perspective, the text conveys the effectiveness of vaccines in general. However, for Annotator B there is an emphasis on the importance of vaccinating children against the listed fatal diseases. Both interpretations are acceptable. The schema proposed is based on a hierarchy of claims, so when the ranking is not straightforward the annotation becomes prone to disagreement.

Claims. During a preliminary analysis of the corpus, it was noted that it is difficult to differentiate between *claims* and *premises*. *Claims* are controversial statements that express a certain stance or intention. Identifying controversy and intention in a statement is a process that can generate discrepancies in the annotations. Adopting the more restrictive definition of *claim* by Stab and Gurevych (2017), the goal was to limit the room for interpretation. The definition contains the requirements that the source of the *claim* should always be the author of the text and that *claims* should only be accepted if some backing is found in the text. Despite the restrictions and the example provided in the guidelines, the agreement was very low.

While the restriction about the source of the *claim* helped identifying the *claim*, there were no restrictions for the source of the *premise*: it could be attributable to the author or to another explicit source. Also, the granularity of annotation was approached differently in a few instances. Example 5

was marked as *claim* by Annotator A, while Annotator B marked it as *premise* considering that the source of "worry" is different from the author.

5. Some worry too about a preservative called thimerosal, which contained a very low concentration of a mercury compound.

Premise. The *premise* component achieved no matches. One reason is the difficulty of differentiating between *claims* and *premises*. Another plausible explanation is the fact that the merging of the two annotation guidelines concentrates in the *premise* component; the definition was drawn from Stab and Gurevych's work (2017) and the attributes where taken from Al-Khatib et al. (2016).

The high level of disagreement indicated that the guidelines did not describe and define properly the argument components. The definitions for *claim* and *premise* left too much room for interpretation.

3.2 Second Pilot Study

Habernal and Gurevych (2017)'s scheme was used in the second annotation study. The guidelines[3] had to be adapted because they require that annotators first conduct a round of annotations to identify all texts that are persuasive regarding the discussed topic. This step is not necessary because all texts in the Vax Corpus are about vaccinations and are considered on-topic. Furthermore, although the documents are not traditionally argumentative, it is possible to identify the intent to persuade in most of them. The second step required annotators to identify the following argument components: *claim*, *grounds*, *backing*, *rebuttal*, *refutation*.

This experiment was conducted to observe whether an argumentation scheme based on the Toulmin model and conceived as being applicable to Web data could indeed be applied to the Vax Corpus.

3.2.1 Results

The annotators were able to go through 37 documents in 10 hours. Table 2 presents the scores for IAA. The score for *refutation* was the lowest (10.0%), whereas the agreement for *backing* was the highest (25.8%). Considering the fact that the scheme was put forward to annotate Web data, the agreement reached is not satisfactory for any of the classes. Furthermore, the task was time-consuming.

Type	IAA	Matches	Non-matches
All classes	19.9%	210	843
Claim	19.7%	46	187
Backing	25.8%	92	264
Grounds	15.6%	56	303
Rebuttal	18.5%	12	53
Refutation	10.0%	4	36

Table 2: Two-way IAA Results (Pilot study 2).

Claims. Annotators were asked to annotate claims at sentence level, and only if there were other argument components in the sentence they should switch to the token level (Habernal and Gurevych, 2017). This indication caused several disagreements, as shown in the next example:

- **Annotator A**: [Some people have had concerns that ASD might be linked to the vaccines children receive but studies have shown that there is no link between receiving vaccines and developing ASD.]$_{Claim}$
- **Annotator B**: [Some people have had concerns that ASD might be linked to the vaccines children receive]$_{Rebuttal}$ [but studies have shown that there is no link between receiving vaccines and developing ASD.]$_{Refutation}$

Annotator A followed the suggestion to annotate at the sentence level and considered the whole section as a *claim* because she interpreted the first clause as a fact that the author needs to present in order to provide the context for his or her opinion on the matter. Annotator B followed the suggestion to switch to the token level in the presence of multiple components and identified two argument components within the sentence, *rebuttal* and *refutation*, which she related to the claim "There is no causal link between vaccination and autism." Both interpretations are acceptable, since it is difficult to determine which one is more appropriate on the basis of granularity criteria.

Claim and Grounds. The vagueness of the definitions and the nature of the texts caused annotators to assign different labels to the same fragments of text, suggesting that the task to differentiate between argument components in Web data is not an easy one. In particular, the distinction that annotators A and B struggled the most with is the

[3] https://www.ukp.tu-darmstadt.de/data/argumentation-mining /argument-annotated-user-generated-web-discourse/

one between *claim* and *grounds*. Even in the definition provided by Habernal and Gurevych (2017) the closeness of the two components is expressed: "*Grounds, can in fact, also be claims: subsidiary or secondary claims that are used to build up a larger argument. Grounds have to match your claims.*"

Claim and Refutation The differentiation between *claim* and *refutation* caused disagreement as well. The guidelines define *refutation* as follows (Habernal and Gurevych, 2017): "*Rebuttal attacks the Claim by presenting opposing view (or contra-argument). Refutation attacks Rebuttal.*" Consequently, the *refutation* and the *claim* convey similar ideas. The condition that allows for the identification of the *refutation* is the presence of the *rebuttal*. This means that if the annotators identify different argumentation structures in the text, one including a *rebuttal* and another excluding it, the first will very likely be annotated as *refutation* and the second will as *claim*.

Grounds and Backing The component *backing* reached the highest rate of agreement. This could be due to the fact that a large part of text could be identified as such. In fact, the guidelines simply state that "*Backing is additional evidence to support the argument.*" (Habernal and Gurevych, 2017). Still, the differentiation between *backing* and *grounds* generated a lot of disagreement. Considering that they play a similar role in an argument, which is to support the *claim*, this was expectable. Reading the guidelines, one can understand that *grounds* is necessary evidence presented to provide good reasons for the *claim*, whereas *backing* is "additional", it does not seem to be necessary. In practice, this difference did not help the annotators, since it is difficult to determine what is necessary as support for an argument and what is futile.

From this second pilot study we learned that annotation schemes containing numerous components and guidelines that offer vague directives are not easily applicable to annotating argument components in heterogeneous texts where no clear patterns of argumentation structure are followed. The guidelines should contain precise directions, providing restrictions that can be used as reference in difficult cases like the ones presented above. As a result, we decided to simplify the annotation scheme in order to make the task feasible, given the type of texts that the Vax Corpus contains. The task would also become less time-consuming.

4 Annotating Claims

The third annotation task was simplified as much as possible. Since our final goal is to understand people's attitudes and gain insight in the process of opinion formation, we decided to focus on the core of the argument: the *claim*. This choice implicated that it was not possible to adhere to the traditional definition of argument. Conventionally, an argument is composed at least of two components: a *claim* and a *premise* (Palau and Moens, 2009; Peldszus and Stede, 2013b). Since premises are frequently claim-like statements and express the attitude of the source, they were not excluded from the annotation task and they were subsumed in the *claim* component. Therefore, the focus of this task was to identify all claim-like statements.

The definition of *claim* chosen was the following:

> *The claim is the central component of an argument. Claims are sections of text that express the stance of the author. Sometimes, claims are introduced by an explicit source in the text (different from the author). Since they are opinionated statements with respect to the topic, claims are often introduced by stance expressions, such as "In my opinion", "I think that".*

An important requirement is that the *claim* has to be a refutable statement. It follows that the following do not qualify as claims:

- **Rhetorical question**: "Wouldn't it be better to develop immunity naturally?"
- **Backing**: "I am a nurse."
- **Common ground**: "Measles can spread through airborne transmission."
- **Statistics**: "80% of vaccinated children experience serious side effects."
- **Anecdotes**: "I experienced hearing loss after being given the MMR vaccine."
- **Opinions**: "I am against vaccinations."

Additionally, when the person or entity to whom the *claim* could be attributed was an explicit

source different from the author of the text, then annotators should also mark the *source* and the relation *Has_source* between *claim* and *source*.

The guidelines for the annotation task are provided as supplementary material.

4.1 Results

The annotators were able to annotate 100 files in 33.5 hours. Table 3 shows the IAA scores. Making a quantitative comparison with the IAA achieved by other related studies is not possible because they all use different evaluation measures. It is possible to compare the results of the main annotation study with the previous two that were carried out as part of this exploration.

Type	IAA	Matches	Non-matches
All	54.4%	2542	2130
Claim	57.3%	2224	1658
Source of Claim	40.3%	318	472

Table 3: Two-way IAA Results for the annotation of claims.

The *claim* component achieved 57.3% IAA, which is satisfactory as compared to the first (6.9%) and second experiments (19.7%). The higher IAA rate achieved in this experiment was predictable considering the fact that the annotation task was less restrictive. While in the second experiment *major claim* should be assigned to sections of text that fulfilled strict requirements, in the third experiment, the annotation of *claim* was not subject to such restrictions.

4.2 Error Analysis

Even if the IAA is acceptable for the task, the task remains difficult. In order to understand where its difficulty lies, we performed an error analysis focusing on the component *claim*. The main points of controversy noticed in the results are the following:

- *Debatability*, which refers to the degree of debatability that the *claim* needs to express in order to be considered as such.
- *Attributability and commitment*, which refer to the context the *claim* needs to be presented in, specifically looking at whether it can be attributed to a source and how strongly the source needs to commit to it.
- *Relatedness to topic*, which refers to deciding whether to annotate or not statements that fit the description of *claim*, but that do not have a direct relation to the topic of vaccination.
- *Granularity and sources*, which refer to different interpretations of the task of assigning sources to the *claim*, resulting in annotations with different degrees of granularity.

4.3 Debatability

The two annotators followed a different approach when annotating claim-like statements that could be accepted by both the anti-vaccination and the pro-vaccination audiences. These instances raised some questions about how open to discussion a statement should be in order to deserve the *claim* label. These sentences often contain modal verbs such as “may”, “might” and “could”. An example of this phenomenon can be observed in the following statements:

6. Vaccines, like any medicine, **can** have side effects.
7. The increased use of veterinary vaccines **may** be accompanied by an increase in human exposure to the vaccine strain, new methods of vaccine administration **may** result in an increased likelihood of inadvertent exposure, increased use of aerosol administration **may** result in greater human exposure to animal vaccines.

These examples deal with potentially negative situations that could take place as result of vaccine inoculation. At first glance, one might think that they reveal an anti-vaccination stance. However, they simply express the possibility that vaccines might have negative side effects, which is an idea that is welcomed by the pro-vaccination community as well. The statements express the capability of a certain occurrence to take place. This is difficult to debate, especially in the case of vaccinations. These examples are characterized by a high degree of acceptability and a low degree of debatability.

Annotator A had the tendency not to annotate such statements, while annotator B annotated them as *claims*. The approach of annotator A seemed to deem the debatability of *claim* very important because it directly results in the stance-expressiveness of the *claim*; if a statement could be accepted by both parties of the debate, it does

not reveal the stance of the author. The proposed guidelines do not give information on how to approach these cases. Considering the fact that one of the goals of analyzing the vaccination debate is to identify the stance of the participating authors, future versions of the guidelines should suggest the identification of *claims* that are polarizing.

4.4 Attributability and Commitment

Annotators exhibited different behaviors in approaching instances where it was unclear to whom the claim-like statement could be attributed. Here are some examples:

8. If you run across someone claiming that **their religion is against vaccinations**, you can check here, although, admittedly, the article only covers mainstream religions
9. What if a parent makes a claim that they are Jewish, and one rabbi says that **vaccines are bad**?

Annotator B marked the tokens in bold as *claims*, whereas annotator A did not. The clauses in bold could, indeed, fit the definition of *claims*. Taking into consideration the context in which they are presented, it is difficult to attribute them to an explicit source or the author of the text. Their context expresses hypothetical situations where it could happen that someone makes those claims.

Other instances that raise similar questions are the following:

10. Some parents might worry that **the vaccine causes autism**.
11. Some people have had concerns that **ASD might be linked to the vaccines children receive**.

Annotator B marked the text in bold as *claim*, while annotator A did not. The statements also express hypothetical situations. Furthermore, they both deal with worries and concerns, which reveal a lower level of commitment to the claims in bold. The guidelines do not give directives on how to tackle the two groups of interesting cases. Although the examples contain clauses that fit the requirements of the class *claim*, it is not possible to attribute them to an author who is committed to them. One way to solve this issue would be to think about the end-goal of the project. If the goal is to capture the stances of the users who are participating in the vaccination debate by writing blog posts and comments, then the above-reported examples should be left unannotated. If the aim is to identify all the possible attitudes that people have regarding the debate, then it would make sense to mark those *claims*. Nonetheless, the examples express speculative claims. One could assume that those are real arguments that people brandish. Future versions of the guidelines will require annotators to mark as *claim* those statements that are attributable to a source and that reveal a high level of commitment.

4.5 Relatedness to the Topic

Annotators were asked to approach the texts focusing on finding all the statements that fit the description of *claim* without worrying about how related they were to the topic of vaccination. The choice was made because all documents in the Vax Corpus are considered on-topic. This directive was not always respected by both annotators leading to disagreement. Some examples are the following:

12. The gene is 'silent'.
13. God is going to save you.

These instances raised a critical objection. If one takes into account only the goal to capture the stance of the author, how useful is it to mark statements that are indirectly related to the topic of vaccination and that do not reveal information about the attitude of the source? The examples above do fit the description of *claim*. Since one of our goals is to gain insight in the process of opinion formation, these statements are helpful in forming a profile for the users, uncovering part of their background and some of their beliefs. Future versions of the guidelines will highlight more strongly the importance of such statements in order to avoid disagreement.

4.6 Granularity and Sources

The annotation of *sources* caused some disagreement, affecting the granularity of the *claim* annotations. Some examples exhibiting this kind of disagreement are the following:

14. **Annotator A**: [95% of Americans are brainwashed to believe they are doing this for us.]$_{claim}$
    Annotator B: [95% of Americans]$_{source}$ are brainwashed to believe [they are doing this for us.]$_{claim}$

15. **Annotator A**: [Government authorities also claim the sterilization chemical was an "accidental" contamination.]$_{claim}$
Annotator B: [Government authorities]$_{source}$ also claim [the sterilization chemical was an "accidental" contamination.]$_{claim}$

Because explicit sources are mentioned, one annotator decided to annotate *source* and *claim* whereas the other one annotated everything as *claim*. However, marking the explicit sources separately might exclude some information from the *claim*. In the case of Example 14, the author wants the audience to know that most Americans are brainwashed; identifying as the claim just the fragment "they are doing this for us" detracts details from the message that is being conveyed. In Example 15 Annotator B followed the same approach and annotated "Government authorities" as *source* and "the sterilization chemical was an "accidental" contamination" as *claim*. As a consequence the full *claim* made by the author of the text, which Annotator A annotated, is ignored. The reason why Annotator B exhibits this behavior is that she had been previously been trained to annotate attributions (Pareti, 2015) and sometimes she had the tendency to annotate attributions, which indicates that the guidelines should have been more explicit about how to deal with cases in which attributions are embedded in *claims*.

5 Conclusion

Our research goal was to test whether existing argumentation annotation schemes are applicable to heterogeneous texts from the Web in order to detect statements that are meaningful for the study of beliefs that motivate different stances towards a topic, in this case vaccinations.

Two pilot annotation studies were conducted with argumentation schemes used in previous annotation tasks (Stab and Gurevych, 2017; Habernal and Gurevych, 2017; Al Khatib et al., 2016). A quantitative and qualitative analysis of the results revealed that it was necessary to simplify the task because the annotation categories were not well defined and, consequently, the IAA was too low. This is why we decided to focus on annotating only *claims*. The new task was then tested by conducting a third annotation study, which resulted in 57.3% IAA.

The simplification of the annotation scheme made the annotation task more feasible and less time-consuming. Following the simplified scheme, annotators were better able to agree on fragments of text that are representative of the beliefs that people express when talking about vaccinations. Based on a qualitative error analysis we defined four sources of disagreement: debatability, attributability, relatedness to the topic and granularity of sources. Further research will evaluate how informative the fragments are for the analysis of the vaccination debate.

The simplified scheme had several weaknesses. The patterns of disagreement observed reveal the necessity to modify the guidelines for future experiments: (i) more annotated examples of difficult cases should be included; (ii) instructions should be provided on how to tackle instances where the granularity can be interpreted in different ways; (iii) it should also be stressed that claim-like statements that are not directly related to the topic need to be marked, as they are relevant; and (iv) the guidelines should be more restrictive. For example, the analysis of the errors related to *attributability* lead to the conclusion that *claims* should be attributable to a source and that the source should express a high level of commitment to the claim.

A general conclusion based on the quantitative results and the qualitative error analysis is that, even though the simplification of the scheme relieved some of the complexity of the task, agreeing on what a *claim* is still remains a difficult endeavor for human annotators. As future work we intend to propose a better definition of *claim*. Additionally, since the simplified scheme does not depend on domain dependent features, it should be applicable to any on-line debate. We plan to annotate similar corpora of other domains in order to test whether the same results can be obtained. Finally, we are currently developing a claim detection system as a means to measuring also the difficulty of performing this task automatically.

Acknowledgements

This research is supported by the Netherlands Organization for Scientific Research (NWO) via the Spinoza-prize awarded to Piek Vossen in the project "Understanding Language by Machines" (SPI 30-673, 2014-2019). We are also grateful to the anonymous reviewers for their valuable comments.

References

Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *COLING*, pages 3433–3443.

Muhammad Amith and Cui Tao. 2018. Representing vaccine misinformation using ontologies. *Journal of Biomedical Semantics*, 9:22.

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259.

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

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.

Amaryllis Mavragani and Gabriela Ochoa. 2018. The internet and the anti-vaccine movement: Tracking the 2017 eu measles outbreak. *Big data and Cognitive Computing*, 2:2.

Raquel Mochales Palau and Marie-Francine Moens. 2009. 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.

Silvia Pareti. 2015. *Attribution: A Computational Approach*. Ph.D. thesis, University of Edinburgh.

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

Ariel Rosenfeld and Sarit Kraus. 2016. Providing arguments in discussions on the basis of the prediction of human argumentative behavior. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(4):30.

Patrick Saint-Dizier. 2016. Argument mining: The bottleneck of knowledge and language resources. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages pp–983.

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

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

Marianna Lya Zummo. 2017. A linguistic analysis of the online debate on vaccines and use of fora as information stations and confirmation niche. *International Journal of Society, Culture & Language*, 5(1):44.