# Extracting perspectives from NAF

Soufyan Belkaid

**Abstract**

In this paper a perspective extraction tool for a corpus is presented that relies on the NAF and CoNLL format. The tool combines multiple layers of information to create perspectives. These perspectives are built out of attributions, claims and opinions that have been aggregated into source files. The final goal of the extraction tool is to provide a user with more insight on the perspectives of all the authors on the subject of any particular corpus. The result is a database containing the perspectives that can be queried semantically, providing an in-dept way of querying data and extracting information. The extraction tool comes with an user interface to provide insight into the sources, the content, the frame and its elements in the statement. In this paper the Vaccination Corpus is used, a corpus about the online vaccination debate.

**Index Terms**

Python, NAF, Attribution, Semantic Role Labeling, SQL, Information Extraction

## I. Introduction

It is easy to get lost in the vast amount of opinions and claims that are found in online debates. Some of these debates on the internet are more serious than others as for instance the online vaccination debate. It becomes challenging to form a well educated opinion about the matter if the debate is uncontrolled and not moderated, which often leads to misconceptions. Trying to organise and structure these opinions, claims and beliefs by hand is a daunting task and not suited for everyone. In this paper an extraction tool is presented that parses the output files of an NLP pipeline and generates a user interface that gives more insight in the text than just by reading it. The extraction tool builds upon the output of the pipeline and aggregates multiple layers to form perspectives. A perspective can be interpreted as the relation between the source of a statement and the target in that statement that is characterized by the expressed attitude in that statement [Son16]. The extraction tool uses the entities layer, opinion layer and the semantic role labeling layer to generate the perspectives. The extraction result is stored in an SQL database and can be queried on attribute value pairs.

Any debate on the internet could have an impact on society and the decision making regarding the subject. Being able to quickly analyze the beliefs and opinions participants have in a debate could help with adequately and accurately answering questions, form policy or join the debate. Take for instance the current situation in the Netherlands where policy makers relied on social media and online feedback on whether they should close the elementary schools or not. It is of paramount importance to be up to date and to provide conclusive answers in dynamic situations where the lack of experience is great. Having a tool that extracts these perspectives could help policy makers in such a situation. The interface of the tool shows the text and an overview of the perspectives containing the opinions in the claim, the source, the target, the attitude and the entities in the claim. In section 2., the design of the tool is explained. In section 3., different layers of the tool are evaluated. In section 4., the interface is introduced. Section 5, presents the conclusion.

## II. tool Design

### A. Parser

One part of the tool is the parser that extracts information from the NAF files. The corpus has been supplied in XML format containing multiple layers of information. In the same way the parser also uses multiple elements that depend on each other to construct a perspective, starting with event elements present in the texts. Events are defined as any occurrence in the text of an action, process or event state. The parser uses the semantic role labeling layer in the NAF files to extract predicates that intersect with a predefined list of source identifying predicates. Checking if the predicate is marked by the NLP pipeline as a source identifying predicate is the first step and filters out other statements that do not contain any content ascribed to a source. The span of words of the predicate and its roles are the starting point of the perspective and are used by the parser to combine multiple layers. Every predicate in that list is assumed to be the cue that marks the start of an attribution relation. Attribution is a linguistic phenomenon that ascribes the ownership of an attitude to some linguistic material [Pra06]. The attribution is important for the tool because it allows the system to extract what the statement is and the owner of the statement. For every predicate considered to be the cue a list of source identifying predicates is checked for inclusion. If the predicate occurs in that list the subject of the verb is considered to be the source of the statement and the remainder of the sentence with a relation to the predicate is the content. This is the first part of the perspective extraction process where the source and the content of a statement are identified. The next step is the extraction of the attitude values present in a statement. This is done by extracting all the opinions in the statement. The parser moves on to the opinion layer in the NAF file and extracts the expression of an opinion, the target of that expression and the polarity. This list contains all the opinions in the article and needs to be filtered.

| | cue | | frame | People and Organisations | |
|---|---|---|---|---|---|
| risk | 639 | Choosing | 1274 | FDA | 17 |
| get | 532 | Statement | 1272 | Trump | 13 |
| said | 448 | Evidence | 1130 | Offit | 12 |
| reported | 388 | Communication_response | 978 | Poland | 9 |
| immunity | 370 | Categorization | 951 | Merck | 9 |
| know | 356 | Awareness | 913 | CDC | 9 |
| see | 311 | Becoming_aware | 899 | Kennedy | 9 |
| found | 308 | Getting | 890 | California | 7 |
| response | 283 | Daring | 881 | Majumder | 6 |
| exposed | 256 | Scrutiny | 802 | Ig | 6 |
| reactions | 244 | Familiarity | 704 | Wakefield | 5 |
| information | 240 | Reporting | 682 | ACIP | 5 |
| evidence | 237 | Reveal_secret | 623 | Mize | 5 |
| exposure | 229 | Assessing | 551 | Mr Trump | 5 |
| rates | 225 | Taking_sides | 476 | Adalja | 4 |
| think | 222 | Behind_the_scenes | 443 | God | 4 |
| contact | 217 | Grant_permission | 398 | Texas | 3 |
| reaction | 212 | Telling | 382 | Schaffner | 3 |
| known | 193 | Motion | 379 | WHO | 3 |
| called | 186 | Imposing_obligation | 355 | Schuchat | 3 |

Table I
TOP 20 FREQUENT CUES AND FRAMES

In the previous step the span of words of the attribution relation is also extracted, this is used to filter the opinions that are sub-strings of the statement. All the sub-spans are added to the opinion list of the perspective and are stored in the database.

The next state in the extraction process is the assignment of Named Entities to the source of the attribution relation. The NAF file contains a Named Entity layer that assigns a category to the entities. The categories that are used in this step are the Organisation and the Person Category. For every source element of an attribution relation the entity list is checked. If there is an entry in the list the source_entity column in the database is filled, otherwise it is left to NULL. This entry check is also done based on the word span of the entity extracted by the pipeline and the span of the source extracted in the attribution section.

The result of the steps described above is a collection of data parts from different layers that build a perspective. This is stored in a database. For the parts of the perspective that contain multiple attribute-keys, for example the frame elements, the values are stored in JSON columns. This is done to be able to query the data based on the attribute and values available. Another reason to store the data in JSON columns is because the amount of frame elements per sentence varies. This proves to be an important feature for the extraction tool because it allows for complex queries used for perspective extraction. All the parts of a perspective are now structured and stored in the database. In the next section examples are give on how to build perspectives based on the queries a user writes.

### *B. Extraction result*

The result of the parser needs to be aggregated to build perspectives. The perspective table in the database contains 24171 entries. This is an over generation of entries based on the rules described in the previous subsection. The parser did not account for more than one predicate in a sentence and sub clauses containing these predicates are also considered to be separate entries in the database. This does not have a negative effect on the tool because the tool is meant to help by learning from the data and provide insights. The user can filter on what is required.

In table one an overview of occurrence is given for the cue, frame and people and organisations. When filtering on the source_entity not being null, the top 5 frames are Choosing, Statement, Reporting, Behind_the_scences and Evidence. The amount of perspectives when the source_entity constraint is held is 368 perspectives. These perspectives contain 393 opinions discovered by the pipeline. In the appendix an image is shown of the perspectives table containing all the columns.

### *C. Data Insights*

After extracting the data and storing it in the database a user can query the data using multiple constraints. For instance the ability to query on frame elements used by the semantic role labeling layer in NAF or on the frame itself. The semantic role labeling layer in the pipeline uses The Proposition Bank's (PropBank) [PKG04] argument annotation. By constraining the frame element or the frame itself one can contextualize the perspectives and search for perspectives that are similar. An example where a frame element is queried is the temporal adjunct, which allows for queries on statements relating to time. Using the following SQL statement:

```
SELECT statement, source_entity, frame FROM perspectives WHERE source_entity IS NOT null AND
    json_extract(perspectives.roles_span, '$.AM-TMP') IS NOT null;
```

| statement | source_entity | frame |
|---|---|---|
| ProQuad [PDF -425 KB]: FDA approved this vaccine in 2005 for use in children ages 1 through 12 years of age | FDA | Grant_permission |
| Kinrix [PDF – 165 KB]: FDA approved this vaccine in 2008 | FDA | Grant_permission |
| Pediarix [PDF – 242 KB]: FDA approved this vaccine in 2002 | FDA | Grant_permission |
| hepatitis B. Pentacel [PDF – 325 KB]: FDA approved this vaccine in 2008 | FDA | Grant_permission |
| Quaracel [PDF – 178 KB]: FDA approved this vaccine in 2015 | FDA | Grant_permission |
| Decavac [PDF – 276 KB]: FDA approved this vaccine in 1955 | FDA | Grant_permission |
| FDA approved this vaccine in 2003 | FDA | Grant_permission |
| FDA approved this vaccine in 2005 | FDA | Grant_permission |
| Adacel [PDF – 245 KB]: FDA approved this vaccine in 2005 | FDA | Grant_permission |
| In addition FDA regularly inspects places where vaccines are made | FDA | Inspecting |

Table II
CAPTION

In table 2. the result of the SQL query shows different events over time where vaccines are being approved by the FDA. This provides an insightful way to query the data and extract information, as the frame of the cue is also being shown. Another way to query the data, as mentioned earlier, is on the frame itself. Since the vaccination corpus is a collection of texts on thoughts and beliefs it might be interesting to see in which way an author tries to express beliefs and thoughts. Relaxing the query by not enforcing the constraint on the source of the claim being a Named Entity, shows us some claims being made by, in all probability, the author of the article. The texts resulting from such a query could be opinionated texts that might show less credible information compared to claims made by the FDA or the CDC. As can be found on FrameNet the Awareness frame is described as follows: *A Cognizer has a piece of Content in their model of the world. The Content is not necessarily present due to immediate perception, but usually, rather, due to deduction from perceivables. In some cases, the deduction of the Content is implicitly based on confidence in sources of information* ***(believe)****, in some cases based on logic (****think****), and in other cases the source of the deduction is deprofiled (****know****).* Since perspectives can be interpreted as subjective relations between a source and target, it is interesting to see which claims are the result of the following query:

```
SELECT statement, source_entity, frame, opinion_info, doc_id FROM perspectives WHERE frame IS
    'Awareness';
```

The result of this query is interesting in the way that it contains a lot of claims made by authors that seem to be part of the anti-vaccination movement. As the article does not from an official government website. To have insight in this data could help policy makers address the issues surrounding the thoughts and beliefs as to why the anti-vaccination movement pushes for not vaccinating. Part of the list of the claims and the authors of these claims has been added to the appendix.
The option to search for particular entries as a role entity of a frame is also another interesting way of querying the data. A user of the tool could search for a word in a sentence in a certain frame element under a specific frame. It is possible to search which entity used the frame Attempt_suasion which is the frame for the verb recommend and also put a constraint on the patient of the frame. One could query "*who* recommends *what* for *something*" by filling in the role set and using the SQL LIKE statement to query for key words in those place holders.

```
SELECT statement, source_entity, cue, frame, roles_text, opinion_info, sentiment FROM
    perspectives WHERE json_extract(perspectives.roles_text, '$.A1') LIKE '%MMR%' and
    json_extract(perspectives.roles_text, '$.A0') LIKE '%Wakefield%';
```

In the example above the result are statements about *what* doctor Wakefield recommends or suggests about a certain theme. This is a useful way to query the data because the data can be queried specifically in certain semantic contexts.

## III. EVALUATION

### A. Parser Evaluation

One of the weaknesses in this tool is the reliance on the pipeline output. The source files are labeled automatically and contain errors that could ultimately lead to extracted perspectives that are not correct. An example of such an error is in the SRL layer where a predicate is labeled with the wrong frame. Often leading to frame elements being wrongly selected as source and content. Another significant problem is the fact that the claims extracted can not be evaluated by comparing them to the claims extracted from the CoNLL files. The word spans in the NAF files do not match tokenisation and its associated span in the CoNLL files. if the tokenisation would match from both output files the system would have more resources available, for instance in using a voting technique. One could look at the output by hand and trace were it went wrong and adjust the values extracted. Different methods were examined to match the span from the NAF to the CoNLL source file but the experiments yielded no useful results. For that reason it was chosen to build a database that could be queried so that the user of the tool could build perspective graphs based on the queries made. Additionally, improvement in the parser has to be made

to account for multiple clauses in a sentence. As mentioned earlier the parser over generates because it uses a simple algorithm to find attribution relations using the SRL layer. That layer does not indicate whether the content contains embedded clauses. A possible solution to this problem is matching the attributions that share the same text spans in the article. If they match then the parser should check if the predicate of the shorter text span is included in the content. In this way the lexical verbs are identified and the embedded clauses are accounted for.

### *B. Statement Clustering*

As a way to gain more insight into the set of statements extracted by the tool a KMeans clustering algorithm has been used. The goal was to find semantically similar statements. As a first step features were extracted using a DictVectorizer that can take attribute value pairs as input and transform them to fit the expected parameter of the KMeans model. A feature dictionary contains the frame elements, the frame itself and the cue, see the listing below for an example.

```
{
'A0': 'us',
 'AM-MNR': 'effectively',
 'A1': 'american children',
 'A2': 'as a herd of animals',
 'frame': 'Labeling',
 'cue': 'labeling'
 }
```

The KMeans clustering algortihm expects a predetermined amount of clusters, for these statements the amount of clusters was set to six. This was done purely by looking at the yielded results from the clusters and the amount of categories that were supplied in the stance annotation file. The stance file contains the following categories: AntiStr, ProStr, AntiWeak, ProWeak, Mix and Neutral. For every cluster the top 10 features and the statements are printed. One of the clusters returns the statements about the FDA approving a drug which alligns with the SQL select statement in which the temporal adjunct was set. Additionally, as can be seen in the listing below, the features of the cluster also show the frame elements the cluster consists of. The listing shows the statement, the frame and the cue.

```
statements:
CLUSTER 3
0      ProQuad [PDF -425 KB]: FDA approved this vaccine in 2005 for use in children ages 1
    through 12 years of age GRANT_PERMISSION APPROVED
1      Kinrix [PDF  165 KB]: FDA approved this vaccine in 2008 GRANT_PERMISSION APPROVED
2      Pediarix [PDF  242 KB]: FDA approved this vaccine in 2002 GRANT_PERMISSION APPROVED
3      hepatitis B. Pentacel [PDF  325 KB]: FDA approved this vaccine in 2008 GRANT_PERMISSION
    APPROVED
4      Quaracel [PDF  178 KB]: FDA approved this vaccine in 2015 GRANT_PERMISSION APPROVED
5      Decavac [PDF  276 KB]: FDA approved this vaccine in 1955 GRANT_PERMISSION APPROVED
6      FDA approved this vaccine in 2003 GRANT_PERMISSION APPROVED
7      FDA approved this vaccine in 2005 GRANT_PERMISSION APPROVED
8      Adacel [PDF  245 KB]: FDA approved this vaccine in 2005 GRANT_PERMISSION APPROVED
9      FDA also inspects the sites where vaccines are made to make sure they follow strict
    manufacturing guidelines INSPECTING INSPECTS
10     FDA also would examine the firm\'s total compliance situation and take further action
    according to the severity of the reporting violation. SCRUTINY EXAMINE
11     In addition FDA regularly inspects places where vaccines are made INSPECTING INSPECTS
12     the FDA approved vaccine manufacturers GRANT_PERMISSION APPROVED
13     FDA approvals global financials, market strategies, advertising campaigns, public
    relations, and data tracking and research GRANT_PERMISSION APPROVALS
14     FDA also inspects the sites where vaccines are made to make sure they follow strict
    manufacturing guidelines INSPECTING INSPECTS

features:
Cluster 3:
 A0=fda
 frame=Grant_permission
 cue=approved
 A1=this vaccine
 cue=inspects
 frame=Inspecting
 AM-DIS=also
 AM-TMP=in 2005
 AM-TMP=in 2008
 A1=the sites where vaccines are made to make sure they follow strict manufacturing guidelines
```

## IV. Interface

One of the goals of this tool is to provide an interface that gathers all claims extracted from the articles and present this in an overview per article. This allows for semantically representing the claims made in the article. The interface currently shows all the claims extracted from the CoNLL data and all the claims extracted from NAF based on the methodology described in section 2. The claims contain highlighted parts that depict the frame element in the Frame. Every frame element has its own color. The extraction results from NAF can be compared to the extraction results from the CoNLL files. The interface relies heavily on the extracted information in the database. The appendix contains a couple of screenshots of the interface. For a prototype of the interface please refer to the following website https://atm.isda.xyz.

## V. Conclusion

This paper presented an extraction tool that aggregates multiple layers from the NAF source files to present perspectives. The perspectives extracted depict the relation between a source of a statement and a target in that statement which is characterized by the attitude, for instance an opinion. The result is an automatically extracted structured database containing all the different parts of a perspective, ready to be queried through SQL statements. It allows for information extraction by context since the Frame of a proposition is used to extract the attribution relation. It is therefore possible to see how the statements relate to each other. Leading to a more in-depth analysis of the statement made by authors either expressing factual statements, beliefs or even misinformation.

Another option is to query statements based on frame elements. For instance the possibility to query claims that contain the adjunct temporal frame element, which could express a statement containing events in relation to time. The tool also allows for information extraction based on the entity that made the statement by looking for a specific person as the source entity of the statement. Or the option to search for statements in which the target is a Named Entity. The result of these options is a perspective graph for every query a user makes in the database.

Unfortunately, the opinions expressed in the claims extracted by the opinion layer do not clearly provide the attitude value between source and target. Therefore the VADER sentiment analyser was used as a substitution to provide the sentiment value of the whole statement. To extract the relation between source and target, the extraction tool heavily relies on the semantic role labeling layer (SRL). The correctness of the source and target in the claim depends on how well the SRL layer performs in assigning the correct frame to the predicate and selecting the correct frame elements as source and target. The same goes for the entity layer, where some named entities are classified as the wrong kind of entity.

The resulting database comes with an interface that shows the extracted information on a per article basis. The main reason to add an interface is to provide a quick overview of an article and the perspectives extracted from it, which can help in clustering articles based on the claims expressed in the text. Regardless of the texts in the corpus, as long as the source files are supplied in the NAF format and the entities layer, the opinion layer and semantic role labeling layer are present, the perspectives can be extracted.

One of the future efforts could be a second version of the interface showing the actual clusters of texts based on the perspectives expressed by the authors. Allowing to compare the statements made by the authors. Moreover, the option to search for statements semantically without the need for SQL-like statements would make it less technical to query the database. Therefore, the goal is to integrate a wrapper in the interface so that it is not required to write SQL statements.

## References

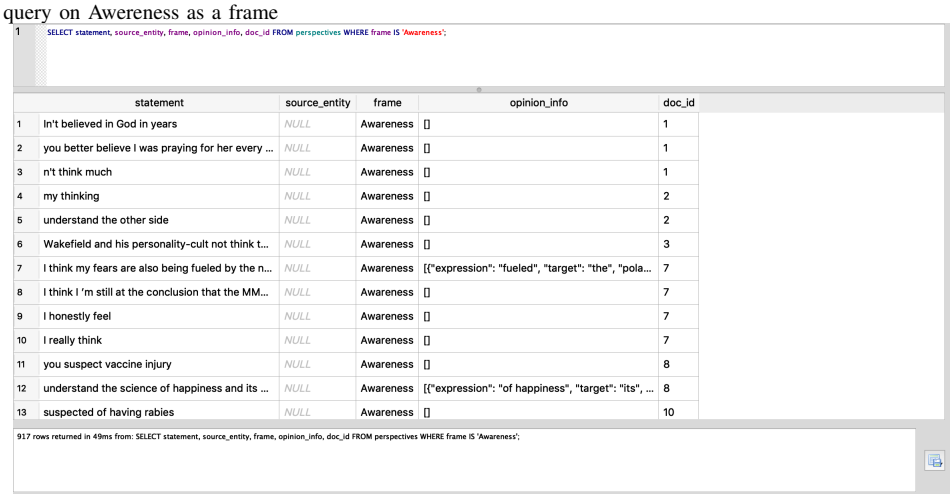
[PKG04] Maria Palmer, Paul Kingsbury, and Daniel Gildea. “The Proposition Bank: An Annotated Corpus of Semantic Roles”. In: (2004).

[Pra06] Rashmi Prasad. “Annotating Attribution in the Penn Discourse TreeBank”. In: (2006).

[Son16] Chantal van Son. “GRaSP: A Multilayered Annotation Scheme for Perspectives”. In: (2016).

# APPENDIX A
# DATABASE

Figure 1. result of the query on Awereness as a frame



```
SELECT statement, source_entity, frame, opinion_info, doc_id FROM perspectives WHERE frame IS 'Awareness';
```

| | statement | source_entity | frame | opinion_info | doc_id |
|---|---|---|---|---|---|
| 1 | In't believed in God in years | NULL | Awareness | [] | 1 |
| 2 | you better believe I was praying for her every ... | NULL | Awareness | [] | 1 |
| 3 | n't think much | NULL | Awareness | [] | 1 |
| 4 | my thinking | NULL | Awareness | [] | 2 |
| 5 | understand the other side | NULL | Awareness | [] | 2 |
| 6 | Wakefield and his personality-cult not think t... | NULL | Awareness | [] | 3 |
| 7 | I think my fears are also being fueled by the n... | NULL | Awareness | [{"expression": "fueled", "target": "the", "pola... | 7 |
| 8 | I think I 'm still at the conclusion that the MM... | NULL | Awareness | [] | 7 |
| 9 | I honestly feel | NULL | Awareness | [] | 7 |
| 10 | I really think | NULL | Awareness | [] | 7 |
| 11 | you suspect vaccine injury | NULL | Awareness | [] | 8 |
| 12 | understand the science of happiness and its ... | NULL | Awareness | [{"expression": "of happiness", "target": "its", ... | 8 |
| 13 | suspected of having rabies | NULL | Awareness | [] | 10 |

917 rows returned in 49ms from: SELECT statement, source_entity, frame, opinion_info, doc_id FROM perspectives WHERE frame IS 'Awareness';

Figure 2. Picture of the Perspectives table extracted from NAF

| id | statement | statement_span | opinion_info | cue | frame | roles_span | order | term_to_word | source_entity | target_entity | sentiment | doc_id |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Filter | Filter | Filter | Filter | Filter | Filter | Filter | Filter | Filter | Filter | Filter | Filter | Filter |
| 1 | 21st Century Wi... | {"t4": "says", "t1"... | [{"expression": "... | says | Choosing | {"A0": ["t1", "t2", "t3"], "A1": ["t5", "t6", "t7", "t8", "t9", "t... | t4 says : A0[t1 2... | {"t4": "says", "t1"... | NULL | NULL | neu | 1 |
| 2 | ... In early 2015... | {"t13": "declared... | [{"expression": "... | declared | Statement | {"AM-TMP": ["t5", "t6", "t7", "t8"], "A0": ["t10", "t11"], "A... | t13 declared : A... | {"t13": "declared... | Big Pharma | NULL | neu | 1 |
| 3 | you n't noticed ... | {"t24": "noticed"... | [] | noticed | Becoming_aware | {"A0": ["t21"], "AM-MNR": ["t23"], "AM-TMP": ["t25", "t26... | t24 noticed : A0... | {"t24": "noticed"... | NULL | NULL | neu | 1 |
| 4 | any American p... | {"t58": "dare", "t... | [] | dare | Daring | {"A1": ["t54", "t55", "t56", "t57", "t58", "t59", "t60", "t61"... | t58 dare : A1[t5... | {"t58": "dare", "t... | NULL | NULL | neu | 1 |
| 5 | that question th... | {"t60": "question... | [] | question | Questioning | {"R-A0": ["t57"], "A1": ["t61", "t62", "t63", "t64", "t65", "t... | t60 question : R... | {"t60": "question... | NULL | NULL | neu | 1 |
| 6 | any mass-produ... | {"t69": "product"... | [] | product | Behind_the_sce... | {"AM-MNR": ["t67"], "A1": ["t66", "t67", "t68", "t69"]} | t69 product : A... | {"t69": "product"... | NULL | NULL | neu | 1 |
| 7 | Back in 2000 th... | {"t77": "announc... | [] | announced | Statement | {"AM-TMP": ["t71", "t72", "t73"], "A0": ["t75", "t76"], "A1"... | t77 announced ... | {"t77": "announc... | NULL | NULL | neu | 1 |
| 8 | 2014 where the... | {"t105": "claims"... | [] | claims | Statement | {"AM-TMP": ["t101", "t102", "t103", "t104", "t105", "t10... | t105 claims : A... | {"t105": "claims"... | NULL | NULL | neu | 1 |
| 9 | the CDC tracke... | {"t108": "tracke... | [] | tracked | Seeking | {"A0": ["t103", "t104"], "A1": ["t109", "t110"]} | t108 tracked : A... | {"t108": "tracke... | NULL | NULL | neu | 1 |
| 10 | they rightly say ... | {"t130": "say", "t... | [] | say | Choosing | {"A0": ["t128"], "AM-ADV": ["t129"], "A1": ["t131", "t132"... | t130 say : A0[t1... | {"t130": "say", "t... | NULL | NULL | neu | 1 |
| 11 | recent cases in ... | {"t138": "constit... | [] | constitute | Communicate_c... | {"A0": ["t132", "t133", "t134", "t135", "t136", "t137"], "A... | t138 constitute ... | {"t138": "constit... | NULL | NULL | neu | 1 |
| 12 | Walt Disney the... | {"t191": "brand",... | [] | brand | Labeling | {"A2": ["t187", "t188"], "A1": ["t189", "t190"]} | t191 brand : A2... | {"t191": "brand",... | NULL | NULL | neu | 1 |
| 13 | on January 7th (... | {"t196": "see", "t... | [] | see | Categorization | {"AM-LOC": ["t192", "t193", "t194", "t195"], "A1": ["t197... | t196 see : AM-L... | {"t196": "see", "t... | NULL | NULL | neu | 1 |
| 14 | successful headl... | {"t222": "declare... | [] | declared | Statement | {"A1": ["t218", "t219", "t220", "t221", "t222", "t223", "t2... | t222 declared : ... | {"t222": "declare... | NULL | NULL | neu | 1 |
| 15 | Even the Preside... | {"t233": "endors... | [] | endorsed | Taking_sides | {"A0": ["t230", "t231", "t232"], "A1": ["t234", "t235", "t23... | t233 endorsed :... | {"t233": "endors... | NULL | NULL | neu | 1 |
| 16 | the campaign st... | {"t236": "stating... | [] | stating | Statement | {"A0": ["t234", "t235", "t236"]} | t236 stating : A... | {"t236": "stating... | NULL | NULL | neu | 1 |
| 17 | Barack Obama t... | {"t257": "told", "... | [] | told | Motion | {"A0": ["t255", "t256"], "A2": ["t258"], "AM-TMP": ["t259"... | t257 told : A0[t... | {"t257": "told", "... | NULL | NULL | neu | 1 |
| 18 | CNN even more ... | {"t270": "stance"... | [] | stance | Assessing | {"A0": ["t264"], "A2": ["t267", "t268", "t269"]} | t270 stance : A... | {"t270": "stance"... | NULL | NULL | neu | 1 |
| 19 | the network no... | {"t278": "touting... | [{"expression": "... | touting | Statement | {"A0": ["t274", "t275"], "AM-TMP": ["t277"], "A2": ["t279"... | t278 touting : A... | {"t278": "touting... | NULL | NULL | neu | 1 |
| 20 | CNN 's Sanjay G... | {"t290": "presse... | [{"expression": "... | pressed | Attempt_suasion | {"A0": ["t286", "t287", "t288", "t289"], "AM-MNR": ["t291... | t290 pressed : ... | {"t290": "presse... | NULL | NULL | neu | 1 |
| 21 | claiming When t... | {"t319": "claimin... | [] | claiming | Statement | {"AM-TMP": ["t322", "t323", "t324", "t325", "t326", "t32... | t319 claiming : ... | {"t319": "claimin... | NULL | NULL | neu | 1 |
| 22 | Tapper later insi... | {"t348": "insiste... | [{"expression": "... | insisted | Statement | {"A0": ["t346"], "AM-TMP": ["t347"], "A1": ["t349", "t350"... | t348 insisted : A... | {"t348": "insiste... | Tapper | NULL | neu | 1 |
| 23 | the federal gove... | {"t356": "convin... | [{"expression": "... | convince | Eventive_cogniz... | {"A0": ["t350", "t351", "t352", "t353"], "A1": ["t357", "t35... | t356 convince : ... | {"t356": "convin... | NULL | NULL | neu | 1 |
| 24 | That could total... | {"t401": "total", ... | [] | total | Adding_up | {"A1": ["t399"], "AM-MOD": ["t399", "t400", "t401", "t40... | t401 total : A1[t... | {"t401": "total", ... | NULL | NULL | neu | 1 |
| 25 | US authorities a... | {"t413": "authori... | [] | authorities | Grant_permission | {"A2": ["t412"], "A0": ["t412", "t413", "t414", "t415"]} | t413 authorities... | {"t413": "authori... | NULL | NULL | neu | 1 |
| 26 | See a full and d... | {"t440": "See", "t... | [] | See | Categorization | {"A1": ["t441", "t442", "t443", "t444", "t445", "t446", "t4... | t440 See : A1[t4... | {"t440": "See", "t... | NULL | NULL | neu | 1 |
| 27 | the family-frien... | {"t459": "brand",... | [] | brand | Labeling | {"A2": ["t456", "t457", "t458", "t459"]} | t459 brand : A2... | {"t459": "brand",... | NULL | NULL | neu | 1 |
| 28 | the media not t... | {"t525": "telling"... | [{"expression": "... | telling | Motion | {"A0": ["t521", "t522"], "AM-NEG": ["t524"], "A2": ["t527"... | t525 telling : A0... | {"t525": "telling"... | NULL | NULL | neu | 1 |
| 29 | the media you (... | {"t528": "surpris... | [] | surprise | Experiencer_obj | {"A1": ["t527"], "A0": ["t526"]} | t528 surprise : ... | {"t528": "surpris... | NULL | NULL | neu | 1 |
| 30 | the media surpr... | {"t530": "surpris... | [] | surprise | Experiencer_obj | {"A1": ["t521", "t522"]} | t530 surprise : ... | {"t530": "surpris... | NULL | NULL | pos | 1 |
| 31 | In fact the CDC '... | {"t600": "experi... | [{"expression": "... | experienced | Feeling | {"AM-DIS": ["t586", "t587"], "A0": ["t598", "t599"], "A1": ... | t600 experience... | {"t600": "experi... | NULL | NULL | neu | 1 |
| 32 | which experienc... | {"t644": "experi... | [] | experienced | Feeling | {"R-A0": ["t643"], "A1": ["t645", "t646", "t647"]} | t644 experience... | {"t644": "experi... | NULL | NULL | neu | 1 |
| 33 | For more infor... | {"t653": "see", "t... | [] | see | Categorization | {"AM-DIS": ["t650", "t651", "t652"], "A1": ["t654", "t655",... | t653 see : AM-... | {"t653": "see", "t... | NULL | NULL | neu | 1 |
| 34 | more informatio... | {"t652": "inform... | [] | information | Telling | {"A2": ["t651"], "A0": ["t654", "t655", "t656", "t657", "t65... | t652 informatio... | {"t652": "inform... | NULL | NULL | neu | 1 |
| 35 | the US media to... | {"t676": "touting... | [] | touting | Statement | {"A0": ["t672", "t673", "t674"], "A1": ["t677", "t678", "t67... | t676 touting : A... | {"t676": "touting... | NULL | NULL | neu | 1 |
| 36 | the US media cit... | {"t685": "citing",... | [] | citing | Adducing | {"A0": ["t672", "t673", "t674"], "A1": ["t686", "t687", "t68... | t685 citing : A0[... | {"t685": "citing",... | NULL | NULL | neu | 1 |
| 37 | California " who... | {"t695": "detecte... | [] | detected | Perception_exp... | {"AM-LOC": ["t690", "t691", "t692", "t693", "t694", "t69... | t695 detected : ... | {"t695": "detecte... | NULL | NULL | neu | 1 |
| 38 | audiences | {"t700": "audien... | [] | audiences | Perception_active | {"A0": ["t700"]} | t700 audiences ... | {"t700": "audien... | NULL | NULL | neu | 1 |
| 39 | believe that the... | {"t702": "believe... | [{"expression": "... | believe | Awareness | {"A1": ["t703", "t704", "t705", "t706", "t707", "t708", "t7... | t702 believe : A... | {"t702": "believe... | NULL | NULL | neu | 1 |
| 40 | US effectively la... | {"t758": "labelin... | [] | labeling | Labeling | {"A0": ["t737"], "AM-MNR": ["t757"], "A1": ["t759", "t760... | t758 labeling : ... | {"t758": "labelin... | US | NULL | neu | 1 |
| 41 | the livestock ter... | {"t754": "immun... | [] | immunity | Taking_sides | {"AM-TMP": ["t748", "t749", "t750"], "A1": ["t753"]} | t754 immunity :... | {"t754": "immun... | NULL | NULL | neu | 1 |
| 42 | Media and CDC ... | {"t775": "insistin... | [] | insisting | Statement | {"A0": ["t767", "t768", "t769", "t770", "t771", "t772", "t7... | t775 insisting : ... | {"t775": "insistin... | NULL | NULL | neu | 1 |
| 43 | herd immunity | {"t780": "immun... | [] | immunity | Taking_sides | {"A1": ["t779"]} | t780 immunity :... | {"t780": "immun... | NULL | NULL | neu | 1 |
| 44 | lifetime immuni... | {"t835": "immun... | [] | immunity | Taking_sides | {"A1": ["t836", "t837", "t838"]} | t835 immunity :... | {"t835": "immun... | NULL | NULL | neu | 1 |
| 45 | vaccine manufa... | {"t851": "prefer"... | [] | prefer | Partiality | {"A0": ["t848", "t849", "t850", "t851", "t852", "t853", "t8... | t851 prefer : A0... | {"t851": "prefer"... | NULL | NULL | neu | 1 |
| 46 | a vaccine which ... | {"t873": "see", "t... | [] | see | Categorization | {"A0": ["t857", "t858", "t859", "t860", "t861", "t862", "t8... | t873 see : A0[t8... | {"t873": "see", "t... | NULL | NULL | neu | 1 |

# APPENDIX B
# INTERFACE

Figure 3. Picture of the Claims table extracted from CoNLL

| id | value | sent_id | token_ids | doc_id |
|---|---|---|---|---|
| Filter | Filter | Filter | Filter | Filter |
| 1 | US media outlets are running a highly coordinated and sensationalized propaganda campaign to hype-up a " national measles outbreak ", while demonizing any American parents that dare to question the efficacy and safety of any mass-produced vaccine product | 2 | "[13, 14, 15, 16... | 1 |
| 2 | There is a visible backlash happening against parents who are challenging government and corporate vaccine orthodoxy | 7 | "[1, 2, 3, 4, 5, 6... | 1 |
| 3 | " There is every reason to get vaccinated — there are n't reasons to not, " | 10 | "[9, 10, 11, 12, ... | 1 |
| 4 | vaccines " do not cause autism " | 12 | "[17, 18, 19, 20... | 1 |
| 5 | " When they [parents] opt-out [of vaccines] they put not only their children but other children at risk too " | 13 | "[12, 13, 14, 15... | 1 |
| 6 | " needs to convince " parents that all vaccines " are safe " | 14 | "[8, 9, 10, 11, 1... | 1 |
| 7 | " measles are currently ravaging the US and so we need all children to receive ALL vaccines mandated by the state " | 15 | "[10, 11, 12, 13... | 1 |
| 8 | What US authorities and media are hiding from the public is a disturbing phenomenon where the measles infection is actually spread through MMR (live measles-mumps-rubella vaccine ) | 17 | "[1, 2, 3, 4, 5, 6... | 1 |
| 9 | The media used the family-friendly Disney brand to hype-up a non-existent " national outbreak " | 19 | "[4, 5, 6, 7, 8, 9... | 1 |
| 10 | The reality, just like the US-based Ebola Hoax, is that there is no ' national outbreak ' | 20 | "[1, 2, 3, 4, 5, 6... | 1 |
| 11 | What the media is not telling you is that the majority of those 644 plus cases originated from two specific sources: an Amish community and a group of travelers who entered the US from the Philippines | 22 | "[1, 2, 3, 4, 5, 6... | 1 |
| 12 | US media outlets, including CNN, have been aggressively promoting the livestock term, " herd immunity ", effectively labeling American children as a herd of animals | 28 | "[9, 10, 11, 12, ... | 1 |
| 13 | In reality, what the media are promoting is a " herd mentality ", ignoring how parents used to bring their young children together to catch chicken pox and measles etc, to guarantee lifetime immunity for their children | 30 | "[5, 6, 7, 8, 9, 1... | 1 |
| 14 | The psychological operation being deployed here is fairly obvious - first to berate and shame, and then to scare American parents into compliance with central government - a government that appears to be in collusion with Big Pharma on this issue | 32 | "[1, 2, 3, 4, 5, 6... | 1 |
| 15 | any parents who might have genuine personal and family health concerns regarding risky vaccines, are " putting the public safety at risk and therefore must comply " | 33 | "[6, 7, 8, 9, 10, ... | 1 |
| 16 | the media appear to be subtly working to characterize measles as a fatal disease, when in reality, the vast majority of measles infections will run their course in just 7-10 days | 37 | "[3, 4, 5, 6, 7, 8... | 1 |
| 17 | " non-vaxxers brought back an eradicated disease! " | 52 | "[30, 31, 32, 33... | 1 |
| 18 | if it 's vaccine-strain measles, then that means it is the vaccinated who are contagious and spreading measles resulting in what the media likes to label " outbreaks " to create panic (a panic more appropriately triggered by our 25 year history of epidemic autism ) | 60 | "[8, 9, 10, 11, 1... | 1 |
| 19 | It would be what one might call vaccine fallout | 61 | "[1, 2, 3, 4, 5, 6... | 1 |
| 20 | the MMR vaccine can lead to measles infection and transmission | 65 | "[27, 28, 29, 30... | 1 |
| 21 | many children are being put at risk of serious, preventable diseases | 7 | "[8, 9, 10, 11, 1... | 2 |
| 22 | he was absolutely right | 15 | "[2, 3, 4, 5]" | 2 |
| 23 | When you start seeing a fraying of herd immunity, you begin to see outbreaks, starting with the most contagious diseases | 16 | "[1, 2, 3, 4, 5, 6... | 2 |
| 24 | There was no biological explanation and virtually no data | 35 | "[1, 2, 3, 4, 5, 6... | 2 |
| 25 | The paper should never have been published | 36 | "[1, 2, 3, 4, 5, 6... | 2 |
| 26 | A lot of misinformation is based on anecdotal evidence | 40 | "[1, 2, 3, 4, 5, 6... | 2 |
| 27 | Anecdotal evidence is a contradiction in terms | 42 | "[1, 2, 3, 4, 5, 6... | 2 |
| 28 | It 's an emotional, personal experience, but it is n't evidence | 43 | "[1, 2, 3, 4, 5, 6... | 2 |
| 29 | The vaccine did it | 46 | "[1, 2, 3, 4]" | 2 |
| 30 | That 's not evidence | 47 | "[1, 2, 3, 4]" | 2 |
| 31 | You could present reams of scientific data showing that that particular vaccine does n't increase the risk of seizures, but it 's hard to convince a mother who says, " I saw what I saw . " | 53 | "[1, 2, 3, 4, 5, 6... | 2 |
| 32 | It 's reasonable enough to ask if the vaccine was responsible | 54 | "[1, 2, 3, 4, 5, 6... | 2 |
| 33 | when we 've done those studies, there 's no association between vaccination and autism | 58 | "[2, 3, 4, 5, 6, 7... | 2 |
| 34 | That 's real scientific evidence | 59 | "[1, 2, 3, 4, 5]" | 2 |
| 35 | The idea of a religious exemption not to vaccinate your child is the equivalent of a religious exemption to practice child abuse | 67 | "[1, 2, 3, 4, 5, 6... | 2 |
| 36 | It seems to me quite the opposite | 69 | "[1, 2, 3, 4, 5, 6... | 2 |
| 37 | Choosing not to get your child vaccinated is not a risk-free decision | 70 | "[1, 2, 3, 4, 5, 6... | 2 |
| 38 | It 's not surprising that parents are afraid | 75 | "[1, 2, 3, 4, 5, 6... | 2 |
| 39 | if you want to do research on the chickenpox vaccine, you have to read the roughly 300 papers that have been published on the chickenpox vaccine | 83 | "[4, 5, 6, 7, 8, 9... | 2 |
| 40 | For that, you would need some expertise in virology, immunology, statistics, and epidemiology | 84 | "[1, 2, 3, 4, 5, 6... | 2 |
| 41 | if you want the best medical advice, turn to the experts | 89 | "[5, 6, 7, 8, 9, 1... | 2 |
| 42 | We should do what California just did, and eliminate all non-medical exemptions | 90 | "[1, 2, 3, 4, 5, 6... | 2 |
| 43 | You ca n't get a religious or philosophical exemption and not strap your kid in | 101 | "[1, 2, 3, 4, 5, 6... | 2 |
| 44 | We compel people to do it because we know it will save lives | 102 | "[1, 2, 3, 4, 5, 6... | 2 |
| 45 | we know that vaccinations will save many, many lives. | 103 | "[2, 3, 4, 5, 6, 7... | 2 |
| 46 | pockets of unvaccinated or undervaccinated people may still present a problem | 5 | "[19, 20, 21, 22... | 3 |

Figure 4. Empty overview

Figure 5. Picture of an article overview