Vaccine Attitude Surveillance Using Semantic Analysis: Constructing a Semantically Annotated Corpus

Stephanie Brien, Nona Naderi, Arash Shaban-Nejad, Luke Mondor, Doerthe Kroemker, David L. Buckeridge
McGill Clinical & Health Informatics, Department of Epidemiology and Biostatistics,
Faculty of Medicine, McGill University, 1140 Pine Ave. W., Montreal, Canada, H3A 1A3, 514-934-1934 ext. 32999
stephanie.brien@mcgill.ca, nona.naderi@mcgill.ca, arash.shaban-nejad@mcgill.ca,
luke.mondor@mail.mcgill.ca, doerthe.kroemker@mcgill.ca, david.buckeridge@mcgill.ca

ABSTRACT
This paper reports work in progress to semantically annotate blog posts about vaccines to use in the Vaccine Attitude Surveillance using Semantic Analysis (VASSA) framework. The VASSA framework combines semantic web and natural language processing (NLP) tools and techniques to provide a coherent semantic layer across online social media for assessment and analysis of vaccination attitudes and beliefs. We describe how the blog posts were sampled and selected, our schema to semantically annotate concepts defined in our ontology, details of the annotation process, and inter-annotator agreement on a sample of blog posts.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**] Natural Language Processing – *Discourse*, *Text analysis*. I.2.4 Knowledge Representation Formalisms and Methods – *Semantic Network*

General Terms
Management, Measurement, Documentation, Design, Languages, Standardization

Keywords
Semantic analysis, Vaccine sentiment, Ontologies, Social network.

1. INTRODUCTION
Online sources such as blogs and news feeds provide timely information about public attitudes and beliefs towards vaccination and are a potentially valuable source for surveillance to guide public health programming. Current approaches to extract information from these online sources, however tend to identify only the general (e.g., positive or negative) or dominant sentiment (e.g., fear or anxiety) or independent mentions of terms, such as the name of a vaccine, disease, or adverse event. This information identifies what is being mentioned, but it does not help public health personnel to understand specific beliefs about vaccines. For example, knowing that a blog post refers to MMR is noteworthy, but knowing that the post asserts MMR vaccination causes autism is considerably more useful. (In this paper, we use the terms "attitude" and "sentiment" interchangeably).

Formal representations of knowledge contained in text, using the Vaccine Sentiment Ontology (VASON) (Figure 1) can help to identify not only instances of concepts, but also important relationships between the instances of concepts expressed in text. Once identified, the concepts and relationships between them can be used to infer vaccination attitudes and beliefs, which public health agencies can use to understand prevalent concerns regarding specific vaccines and to develop effective interventions. Furthermore, understanding the different relationships, properties and axioms that exist in this domain can provide a rich body of knowledge to facilitate semantic analysis.

We aim to develop and evaluate an automated method for Vaccine Attitude Surveillance using Semantic Analysis (VASSA), which we will apply on a large scale to online sources such as blogs and news feeds. The VASSA framework, which aims to support the automated extraction and analysis of text in online sources related to vaccination, consists of a Natural Language Processing (NLP) module, which is used for semantic analysis and classification, and the VASON, which models existing knowledge about vaccine attitudes. VASON aims to capture domain knowledge regarding vaccine beliefs and attitudes, and it can be used to facilitate concept extraction and analyze the concepts and relationships extracted using the NLP module. The development of the ontology is currently in progress and we are now performing several text extraction experiments using the VASON sub-taxonomies adapted and imported from the Vaccine Ontology (VO) [1] and the Disease Ontology (DO) (http://disease-ontology.org) properties and axioms. VASON is intended to provide conceptual structure to organize the vast amount of unstructured data scattered over blog posts, to facilitate blog content analysis, and to enable discovery of patterns of words or phrases in blog text (e.g. specifying topics, dates, themes, sentiments, beliefs and so on). It also assists in revealing opinionated claims and assertions in blogs and relating authors, forms, functions, geographical locations, audiences of blogs, as well as bloggers' motives for assertions about vaccination. The data for creating the VASON conceptual model comes from the literature, databases and some of the existing vocabularies and ontologies including:

1) Sentiment lexicons such as SentiWordNet [2] and WordNet-Affect [3]; 2) The VO [1], which classifies various licensed vaccines, as well as vaccine candidates in research and trail; 3) The Ontology of Adverse Events [4]; and 4) The DO and the MEDIC vocabularies [5].

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

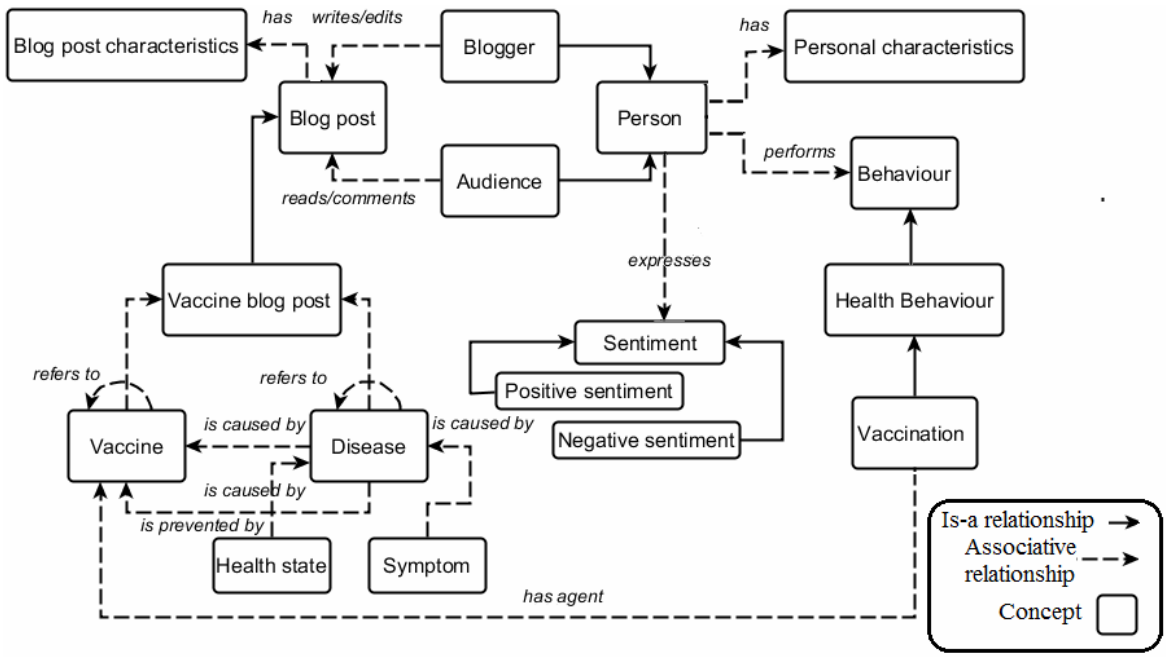

Figure 1. A partial view of the VASON conceptual model and the annotation schema representing the existing knowledge in the domain of vaccine sentiment.

In this paper, we report on our work in progress to semantically annotate blog posts on vaccines for future use as a gold standard corpus to train and evaluate automated text extraction.

2. METHODS

2.1 Sampling Vaccine Blog Posts

Blog posts were sampled using the Google.ca blog search engine with the following search terms: "immunize", "immunise", "immunization", "immunisation", "vaccine", "vaccinate", "vaccination", and "vax". Terms were combined using the "OR" operator and were also prefixed with "pro" and "anti".

We performed two searches on May 11th and June 11th, 2012. We selected the first 200 blog posts from each Google search and screened the results by reviewing the title and scanning the text for keywords and content. We included blog posts with content on human vaccines, in which bloggers were either pro- or anti-vaccine or had no opinion. We excluded duplicate blogs posts (the same entry by the same blogger), posts with content on animal vaccines, posts lacking vaccine search terms in the text, posts comprised of only two sentences or less, non-English language posts, links and posts with content that could not be viewed in our manual annotation software. We initially limited our search to blogspot.com and wordpress.com blog service providers to select blogs with structured profile information for future analysis of the bloggers. However upon the review of our first search results (May 11th), we limited the following search to the blogspot.com blog service provider, as it contained the most structured profile information.

We obtained a sample of 182 (45.5% of all retrieved posts) blog posts after applying our exclusion criteria. From this sample, we randomly selected 10 blog posts to test our annotation schema and guidelines and to measure inter-annotator agreement.

2.2 Annotation Schema and Guidelines

Our preliminary annotation schema, represented within the conceptual model (Figure 1), captures common assertions found on anti-vaccine websites, such as that vaccines cause illness [6]. Given that blogs are an informal source, in which bloggers use lay terms and express themselves in many different ways, we iteratively developed an annotation schema and guidelines.

We annotated the following concepts in the blog posts: vaccines, diseases, health states and symptoms. Table 1 provides examples of these concepts. We were not interested in annotating concepts within proper nouns (e.g., Center for Disease Control), URLs, links or references. Occasionally, bloggers use the terms vaccine or disease in the following context: "vaccine effectiveness" or "disease prevention program", and unless a specific vaccine or disease was mentioned (e.g., HPV vaccine effectiveness), we did not annotate these terms.

Table 1. Examples of annotated concepts.

Concept	Examples
Vaccine	Vaccine, Vaccination, Gardasil, Shot, Inoculation, Immunization
Disease	Measles, HPV, MMR, Chickenpox, Infectious disease, Autism, Swine flu
Health State	Dead, Death, Killed, Harm, Sick
Symptom	Rash, Pain, Redness, Local reaction, Adverse event, Side effect, Anaphylaxis

We also annotated asserted relationships between vaccines and diseases, health states or symptoms. Specifically we annotated

"disease *is prevented by* vaccine" and "disease, health state, or symptom *is caused by* vaccine". We were not interested in health states and symptoms of a disease. Therefore we only annotated health states and symptoms when the text implied a direct relationship between a health state or symptom (*is caused by*) and a vaccine. We also annotated co-references (*refers to)* to disease and vaccine annotations to distinguish between general and specific disease and vaccine mentions in the text. We annotated co-references and relations throughout the blog post and did not restrict our annotation to sentence-level relationships.

Additionally, we annotated the blogger's personal characteristics, such as blogger name and blog post characteristics, such as blog title and blog date. Blog posts were manually classified by sentiment (pro- or anti-vaccine). The time to annotate a blog post was also recorded. These attributes will be used in future analyses (once more blogs are annotated) to explore trends in the concepts and relationships expressed in blog posts and to identify and exclude duplicate blog posts.

Our annotations were restricted to the title and body of the blog post. At this phase, we were mainly interested in the blogger's view of vaccines, therefore we did not annotate the comments of the blog posts, however we intend to in our future work. In general, the concepts and attributes in the text were annotated first, followed by the relations. Given that bloggers can often be sarcastic, we paid special attention to annotate in context.

2.3 Annotation Examples

Below are a few excerpts from our sample to illustrate how concepts and relations were annotated. The words in bold are the terms that were annotated and the words in parentheses denote the concepts and relations.

1. "Despite widespread childhood **vaccination** (vaccine; prevents: Bordetella pertussis) against **Bordetella pertussis** (disease), **disease** (disease; refers to: Bordetella pertussis) remains prevalent." (Pro-vaccine blog post)
2. "The **influenza** (disease) **vaccine** (vaccine; prevents: influenza; causes: anaphylaxis) is "convincingly" linked to causing **anaphylaxis** (symptom), which is why **influenza** (disease) **vaccines** (vaccine; prevents: influenza; causes: killed) have **killed** (health state) so many children." (Anti-vaccine blog post)
3. "**MMR** (diseases) **vaccines** (vaccine; prevents: MMR; causes: measles, seizures, anaphylaxis, health problems) cause **measles** (disease), **seizures** (symptom), **anaphylaxis** (symptom) and other **health problems** (disease)." (Anti-vaccine blog post)
4. "That's why every winter, the vast majority of people who catch the **flu** (disease) are the very same people who were **vaccinated** (vaccine; prevents: flu, causes: flu) against the **flu** (disease)." (Anti-vaccine blog post)

2.4 Annotation and Schema Evaluation

To improve the consistency of manual coding, three annotators, two with backgrounds in epidemiology and a computer scientist participated in the preparation of the annotated blog posts. Following a short training session, the epidemiologists independently annotated ten blog posts (C-10). The computer scientist examined the annotations and identified errors and inconsistencies. After reviewing discrepancies, differences were resolved and two epidemiologists re-annotated 5 of the 10 blog posts (C-5) independently. Together, the two epidemiologists re-annotated all 10 blog posts to create the final sample. Manual annotation and analyses were performed using the General Architecture for Text Engineering (GATE) Developer environment [7].

Inter-annotator agreement was computed using the F-measure metric. For overlapping annotations, we considered two criteria, Strict and Lenient. In Strict, the partially matching annotations are considered as disagreement and in Lenient, the partially matching annotations are considered as agreement. An attribute or a concept was considered a match if the same term(s) were annotated and labeled with the correct attribute or concept. A relation (*is prevented by* and *is caused by*) was considered a match if the same vaccine and disease, health state or symptom terms were annotated and the same disease, health state or symptom term was specified for the relation. A co-reference was considered a match if the same vaccine or disease was annotated and specified in the annotation.

3. RESULTS

Among our sample (N=10) of annotated blog posts, 5 were pro-vaccine posts. On average, a blog post took approximately 24 minutes to annotate. Table 2 shows the number of attributes, concepts and relations, annotated by annotator, for the two sets of posts.

Table 2. Number of annotations, by attributes, concepts and relations, per set and annotator.

Item	Number of Annotations			
	First Set (C-10)		Second Set (C-5)	
	A	B	A	B
Attributes:				
Blog Title	19	18	9	9
Blog Date	10	10	5	5
Blogger Name	16	15	5	5
Concepts:				
Vaccine	246	248	36	36
Disease	289	338	39	39
Relations:				
is prevented by	76	111	20	21
is caused by	112	157	4	6
refers to	34	26	4	5

Table 3 presents the descriptive statistics of the consensus annotations. On average, two blog titles and one blog date and blogger name were annotated in each blog post. Blog posts with a negative sentiment had on average more annotations (mean: 90 vs. 62), *is caused by* relations (mean: 29.4 vs. 0.6) and symptoms and health states compared to posts with a positive sentiment.

Table 3. Number and descriptive statistics of consensus annotations, by concepts and relations, per blog post (N=10).

Item	N	Min	Mean	Max
Overall	621	10	62.1	213
Concepts:				
Vaccine	248	3	24.8	87
Disease	334	3	33.4	122
Relations:				
is prevented by	107	0	10.7	30
is caused by	150	0	15	86
refers to	12	0	1.2	7

Inter-annotator agreement for the two sets of blog posts are presented in Table 4. Agreement was satisfactory for the first set (C-10), however after reviewing the discrepancies, we noticed that most of the differences were due to not following the guidelines (Ex: health states were annotated in the absence of a link to a vaccine and annotations were missed)

Among the re-annotated set, inter-annotator agreement increased for all attributes, concepts and relations (Table 4), with the exception of blog title annotations, which decreased to 55.5%, due to differences in annotating punctuation. For example, one annotator considered quotation marks as part of the annotation, whereas the other annotator ignored them. There were also a few differences in annotating health states and symptoms. One annotator considered "Sudden Vaccine Death Syndrome" as a symptom, while the other annotator considered "Death" as a health state. We also observed differences in annotating co-references. In the following excerpt: "It was about a Hib meningitis outbreak here in Minnesota in 2009, which killed an infant and sickened four others. Hib is short for Haemophilus influenza type B. It's one of the basic childhood vaccinations. It's a terrible disease," disease refers to "Hib meningitis", "Hib" and "Haemophilus influenza type B". However annotators annotated different disease mentions ("Haemophilus influenza type B" and "Hib meningitis") as the co-reference. These differences can be easily resolved by considering the closest mention.

Table 4. Inter-annotator agreement, by attributes, concepts and relations, per set.

Item	Agreement (%)			
	First Set (C-10)		Second Set (C-5)	
	Strict	Lenient	Strict	Lenient
Attributes:				
Blog Title	64.8	97.2	55.5	100
Blog Date	100	100	100	100
Blogger	64.5	77.4	80.0	100
Concepts:				
Vaccine	89.4	93.5	100	100
Disease	73.3	88.7	84.6	97.4
Relations:				
is prevented by	76.5	77.9	91.4	91.4
is caused by	69.2	70.1	66.6	80.0
refers to	47.7	47.7	57.1	57.1

4. DISCUSSION

In this paper, we report on our first step towards an automated method for vaccine attitude surveillance using blog posts. We demonstrate that attributes, concepts and relations can be manually annotated in a reliable manner for a small sample of blog posts on vaccines. There are limitations to our work. We tested our annotation schema and guidelines on a small sample of blog posts. Although this is not sufficient to evaluate our automated extraction approach, blog posts had on average 62 annotated items, allowing us to adequately test our schema and guidelines. Second, the annotated blog posts may not be representative of the blog posts on vaccines present on the web. In future work, we will expand our search to all blog service providers to ensure representativeness. Finally, we have not examined the effect of conflicting statements in blog posts, however we will investigate this in our future work. We intend to continue to annotate blog posts to build a semantically annotated corpus and extend our annotation to correspond with more concepts and relationships in the VASON ontology. We will then use this corpus to evaluate the accuracy with which our semantic framework extracts and classifies text, and identifies relationships and co-references. Additionally, we intend to develop methods for the automatic selection of the relevant blog posts from the web and define more representative semantic queries on the relations between vaccine attitudes, vaccine adverse events, and the risk factors. Finally we will apply our method of extraction and classification to a larger sample of blog posts and news feeds to explore trends in the concepts and relationships expressed in blog posts and news feeds over time, overall and stratified by sentiment, and blogger characteristics.

5. ACKNOWLEDGMENTS

The project is funded by the Public Health Agency/Canadian Institutes of Health Research Influenza Research Network (PCIRN). We would like to thank Jessica Vineberg for screening blog posts.

6. REFERENCES

[1] He, Y., Cowell, L., Diehl, A.D., Mobley, H.L., Peters, B., Ruttenberg, A., Scheuermann, R.H., Brinkman, R.R., Courtot, M., Mungall,C., et al. 2009. VO: Vaccine Ontology. *The 1st International Conference on Biomedical Ontology (ICBO 2009) Nature Precedings.* Buffalo, NY.

[2] Esuli, A., Sebastiani, F. 2006. SentiWordNet: a high-coverage lexical resource for opinion mining. Technical report ISTI-PP-002/2007, Kluwer Academic Publishers, Netherlands.

[3] Strapparava, C. and Valitutti, A. 2004. WordNet-Affect: an affective extension of WordNet. In *Proc. of LREC*, 4: 1083–1086. http://wndomains.fbk.eu/wnaffect.html

[4] Sarntivijai, S., Xiang, Z., Shedden, K.A., Markel, H., Omenn, G.S., Athey, B.D., and He, Y. 2012. Ontology-based combinatorial comparative analysis of adverse events associated with killed and live influenza vaccines. *PLoS ONE*. 7(11): e49941. doi:10.1371/journal.pone.0049941.

[5] Davis, A.P., Wiegers, T.C., Rosenstein, M.C., Mattingly, C.J. 2012. MEDIC: a practical disease vocabulary used at the comparative toxicogenomics database. *Database.* (Mar. 2012), bar057.

[6] Kata, A. 2000. A postmodern Pandora's box: Anti-vaccination misinformation on the Internet. *Vaccine.* 28 (Feb 2010), 1709–1716.

[7] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic. D., Heitz, T., Greenwood, M.A., Saggion, H., Petrak, J., Li, Y., Peters, W. 2011. Text Processing with GATE (Ver. 6). University of Sheffield, CS Department.