GRaSP: Grounded Representation and Source Perspective

Antske Fokkens♣, Piek Vossen♣, Marco Rospocher♢, Rinke Hoekstra♡♣ and Willem R. van Hage♠
♣ CLTL and Computer Science, Vrije Universiteit, Amsterdam, The Netherlands
♢ Fondazione Bruno Kessler, Trento, Italy
♡ Elsevier BV, Amsterdam, The Netherlands
♠ Netherlands eScience Center, Amsterdam, The Netherlands
{antske.fokkens,piek.vossen}@vu.nl, rospocher@fbk.eu
r.hoekstra@elsevier.com, w.vanhage@esciencecenter.nl

Abstract

When people or organizations provide information, they make choices regarding **what** they include and **how** they represent it. These two aspects combined (the content and the stance) represent a **perspective**. Investigating perspectives can provide useful insights into the reliability of information, changes in viewpoints over time, shared beliefs among social or political groups and contrasts with other groups, etc. This paper introduces GRaSP, a generic framework for modeling perspectives and their sources.

1 Introduction

Structured data and knowledge resources typically provide what is seen as factual information. They contain definitions of concepts, ontologies, information about origins, dates, locations, etc. Methods have been developed to automatically extract such information from text (Hearst, 1992; Buitelaar et al., 2004; Wu and Weld, 2010, among others). However, knowledge consists of much more than ontological classifications and basic verifiable properties of objects and people. It involves information about various entities, events and concepts, connecting this information and judging its validity. For social science and humanities, these aspects of knowledge are particularly interesting, i.e. how information is connected, how people judge validity, how knowledge changes, what uncertainty and sentiment that accompanies it.

When people or organizations provide information, they make choices regarding what they include and how they present information. These two aspects together (the content and stance provided by the source) represent a *perspective*, an element of interest for many disciplines. Communication scientists and social psychologists study (e.g.) how common opinions or existing stereotypes are displayed in the media. Political scientists can investigate how various sources present hot topics. Historians may look into how perspectives on historic events change over time. Outside of academia, perspectives can be of interest to information professionals, decision makers, advertisers, journalists and any citizen interested in critical thinking and finding balanced information.

Natural language processing (NLP) can offer support in identifying the topic of text, classifying stances, identifying sentiment and opinions, determining factuality values of events, etc. To our knowledge, these technologies are generally investigated in isolation and have, up to date, not been connected in order to obtain a more full-fledged representation of perspectives. In this paper, we take the first step towards such a representation by introducing a framework that formally represents perspectives: the Grounded Representation and Source Perspective framework (GRaSP). GRaSP is a unique and generic flexible framework that combines the formal representation of the content and of the source perspective in one single model. It is compatible with existing models, but can also model subtleties that can be expressed in natural language but remain challenging for RDF representations.

The rest of this paper is structured as follows. We provide background on GRaSP in Section 2. We then introduce the framework itself in Section 3. We describe an automatically generated dataset represented in GRaSP in Section 4. After discussing related work, we conclude.

2 Background

The origins of GRaSP lie in the projects NewsReader (Vossen et al., 2016) and BiographyNet

Proceedings of Knowledge Resources for the Socio-Economic Sciences and Humanities associated with RANLP-17, pages 19–25,
Varna, Bulgaria, 7 Sep 2017.
https://doi.org/10.26615/978-954-452-040-3_003

(Fokkens et al., 2014). NewsReader aimed at extracting what happened to whom, when and where from large amounts of (financial) news, creating structured data to support decision making. In BiographyNet, we aimed to extract information about individuals in biographical dictionaries for historians. We investigated in connections between people and how the same person or event was depicted in different biographical dictionaries. An essential step for addressing these challenges is to indicate which documents talk about the same entity or event. In addition, the provenance of information is essential in both projects, i.e. end-users need insight into the source of specific information. NewsReader and BiographyNet also shared the vision of comparing differences in information from various sources.

More recent projects dive deeper in perspectives. *Understanding Language By Machines* investigates the relations between events, uncertainty, sentiment and opinons and how this information results into storylines and world views. In *Reading between the lines*, we look at more subtle cues of perspectives addressing questions such as "which background information given when talking about people from different ethnic groups?" or "when do we chose to generalize (e.g. by calling someone a thief rather than a suspect of having stolen something)?". QuPiD2 addresses (among others) what evidence is discussed and how sources build their argumentation around it.

With GRaSP, we aim to design a framework that can support the research questions central to these projects following six requirements. First, we want to represent various perspectives on the same entity, proposition or topic next to each other. Second, it should represent the source of each perspective, so that users can e.g. select all perspectives of a specific source; group sources according to shared or conflicting views on a given content; find all sources that have a perspective on the same content or share a perspective; and, find available background information about the source. Third, we want to provide the means to semantically compare the (propositional) content across statements and represent whether sources mention the same, similar or related content (e.g. more or less specific), or a different framing of content (e.g. *murdered*, which is intentional, or *killed* which may be accidental). Fourth, it should be possible to represent a wide range of perspective-related phenomena, including: sentiment, emotion, judgment, negation, certainty, speculation, reporting, framing and salience. Fifth, we want to make alternative *interpretations* of the same statement explicit, since statements might be (deliberately) ambiguous, not well formulated or difficult to process with Natural Language Processing (NLP) technology. Finally, users should be able to gain insight in the full provenance of any information provided by GRaSP. Next to the source, it should provide information about how this perspective was analyzed (e.g. expert analysis of a text, crowd annotations, text mining).

The first three requirements allow users to place various perspectives next to each other allowing them to compare, among others, which sources agree or disagree on what, which sources change their mind, which sources speculate and whether their predictions were accurate. In addition, they would allow identifying all content and stances given on a specific topic by a source and, for example, display this on a timeline. Researchers can thus investigate what information is important to sources who hold a specific opinion. The fourth and fifth requirement ensure that the model is flexible enough to support various needs of end-users as well as to accommodate the variation of information provided by different systems or datasets. Tools used to gather and interpret information can introduce biases end-users should be aware of (Lin, 2012; Rieder and Röhle, 2012). Providing clear provenance of information (including involved processes) is a necessary component for creating such awareness (sixth requirement).

There are several ontologies that can be used to model perspective-related information. We will outline the most influential ones and explain which part of the requirements they fulfill in Section 5.

3 The GRaSP Framework

Perspectives are expressed by **statements** (which can be spoken or written language, images, signals, etc.) from a specific source. A perspective can be conveyed in many ways, some more explicit than others. Explicit opinions or highly subjective terms are easily identified, but perspectives can be expressed more subtly. The selection and implicit framing of information plays a role (e.g. does an article report on someone's ethnicity, do they report an expert's political preference when citing them on a societal matter) as well as choices

in how information is presented (e.g. using neutral or marked words, certainty, confirming or denying something). We therefore see a perspective as the combination of the **content** of one or more statements (which information is included) and the **stance** sources take on this content.

GRaSP makes the link between the content and stance of a statement as well as to their source explicit. The framework achieves this through a triple layered representation consisting of a **mention layer**, an **instance layer** and an **attribution layer**. The **mention layer** is the central layer of the model. Mentions are physical objects, such as a (piece of) text, (part of) an image or a sound, that signal information and can be embedded in a larger physical object. Mentions can be combined and form a statement that displays a perspective on some propositional content by some source. Propositions are abstract meaning representations that make reference to events and participating entities. Both events and entities are represented as instances in some (presumed) world in the **instance layer**. Finally, the stance expressed by the statement is represented in the **attribution layer**. This layer models attitudinal information such as beliefs, judgments, certainty and sentiment of the source towards the propositional content. This section introduces these layers and illustrate how they are used to model perspectives.[1]

3.1 Grounding

An essential part of representing perspectives is making explicit what the perspective is about, i.e. representing the described (real-world) situation. This is captured by the two top layers of our framework; the instance layer and the mention layer. These two layers, as well as their connecting relation are based on the architecture proposed in the Grounded Annotation Framework (Fokkens et al., 2013, GAF), which is incorporated in GRaSP. Consider the following examples:

1. During 2000-2014, measles vaccination prevent an estimated 17.1 million deaths
2. The search result contained 108 deaths over this period, resulting from four different measles vaccines
3. There have been no measles death reported in the U.S. since 2003

[1]The ontology and examples can be found at: https://github.com/cltl/GRaSP

These sentences above make statements about whether measles or vaccinations cause death. Figure 1 illustrates how this is represented in the top two layers of GRaSP.

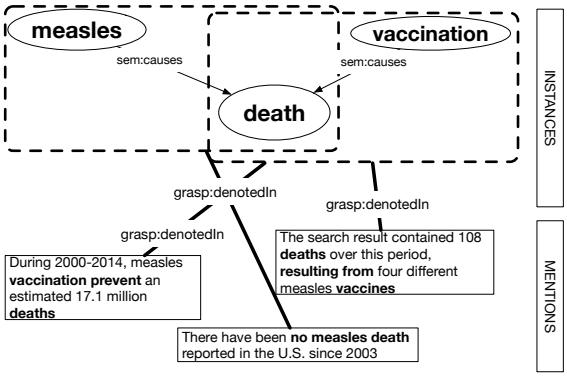

Figure 1: Instance and mention layers

The content of statements is represented in the instance layer. This layer can represent information about events, their participants, their locations or their time, but also information about (generic) concepts or ideas. Typically, propositions are expressed in terms of the Simple Event Model (SEM, (van Hage et al., 2011)), but information in this layer can be represented using other vocabularies as well. SEM is a generic RDF vocabulary for event-participant relations that allows for reasoning over the propositional content of statements. Event-event relations can be represented as well in the instance layer: the example in Figure 1 includes a causal relation between measles and death, and one between vaccination and death.

The second layer represents mentions. Mentions are (pieces of) resources that denote entities or propositions from the instance layer: they can be expressions in text, spoken words, numbers or signals on a display, images, videos, etc. The mention layer allows us to trace all resources where a specific event, a person or idea is mentioned. It also records each specific way in which an instance of interest is presented in a resource. Following Semantic Web practice, GRaSP identifies mentions by IRIs (Internationalized Resource Identifiers). This allows us to link them to additional information, including their surface string (the literal text) and lemma and their exact position within a text or image. This feature is particular relevant for scholars working with automatically analyzed text, since it allows them to easily identify where specific information is mentioned in the original source and hence verify it.

Entities, events and statements in the instance layer can be linked to expressions in the mention layer by the relations `grasp:denotedBy` pointing out the exact words or linguistic structure that expresses the event or statement, or by `grasp:denotedIn` which indicates that the statement is made somewhere in a sentence, paragraph or document. In our example the causal relation between 'vaccination' and 'death' is linked to Sentences 1. and 2. The relation between 'measles' and 'death' is linked to Sentence 3. Through these links, researchers interested in how various sources talk about the risks of measles or vaccinations can find snippets of text that talk about these issues. However, the source and the stance taken are not made explicit yet. The next subsection introduces the attribution layer, which allows us to add this information.

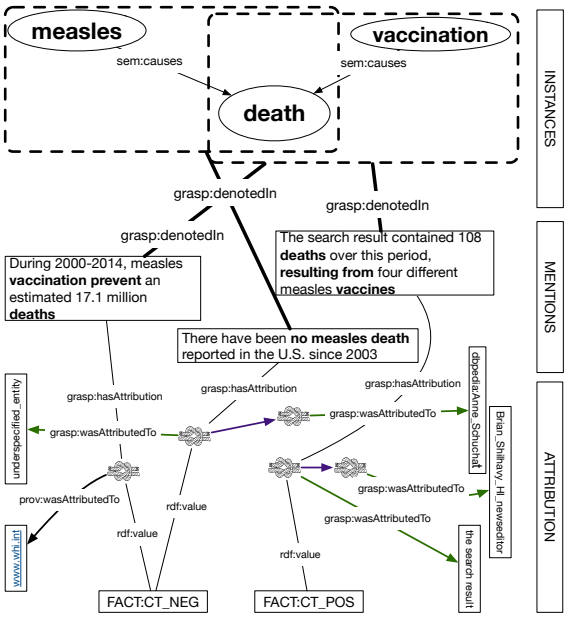

Figure 2: Instance, mention and attribution layers

3.2 Source Perspective

Grounding (modeled by the link between instances and mentions) establishes what a specific message is about. Two components need to be added to complete a framework that can capture perspectives from various sources. First, mentions should be linked to the source that expresses the perspective. Second, we want to represent the stance the source takes on the content of the message. The stance typically includes information on factuality (e.g. does the source confirm or deny, is it certain or hesitant, is it talking about the future?), judgment, sentiment and emotion (e.g. does the source consider the content ethical, is the source scared by the content?).

The third layer, the **attribution** layer, adds these components to GRaSP. Figure 2 adds the attribution layer to our example. Each of the mentions is linked to an attribution node. These nodes are in turn linked to the source that published them and, if applicable, the source reported in the text. We use the PROV-DM (Moreau et al., 2013) to model the source of publication and a specified variant of the `wasAttributedTo` relation introduced by GRaSP for quoted sources. Attribution nodes also receive values that make the stance taken by the source explicit. In this case, the statements expressing opposing views to vaccines leading to deaths or measles leading to death are connected to a factuality value indicating that the source denied this relation without expressing doubt.

Through the addition of this layer, end-users can explore opposing views on the same topic or statement. This allows users to compare sources reported on both viewpoints, identifying what other opinions these sources express and investigating the overall argumentation (the statement that there were no measles death can both been used by people opposing to vaccination, because "measles are not deadly", or by people supporting it, stating that "measles deaths are avoided thanks to vaccination"). Additional information about sources can be gathered leading to investigations on, among others, how reports on specific topics evaluate over time, the difference in certainty expressed by politicians or scientists or how different countries report on the same event.

The examples we have shown here are simplified for the purpose of illustration. Sentence 1. does not express the opposing view from vaccinations causes deaths by saying they do not, but exactly states vaccinations actively prevent death. This relation of prevention can also be modeled in GRaSP. Under this representation, we would model the statement that vaccination prevents death and that this forms an opposing view to them causing death. We can even model that both these statements can be considered to be true by the same source (e.g. believing that vaccinations avoid many deaths and sometimes cause them). The details of how to represent more complex relations and how viewpoints connect are beyond the scope of this paper.

4 GRaSP illustrated

One of the main challenges involved in representing perspectives in GRaSP is the question of how to obtain this information accurately. In principle, GRaSP can be used in combination with close-reading manual methods, where researchers use it to meticulously record the information they base their conclusions on. It becomes more interesting when we can represent massive amounts of data and help researchers find information automatically. Sentiment analysis, factuality classification, opinion mining, event extraction and argumentation mining are challenging tasks. Automatically creating highly accurate representations of perspectives in GRaSP is a challenge for the future. Nevertheless, current methods can provide output that we believe to be useful for researchers interested in perspectives. In this section, we illustrate what information can currently be generated by NLP tools through a dataset that the GRaSP framework for representation made available through an interface providing an open source visualization (van der Zwaan et al., 2016; van Meersbergen et al., 2017).

The GRaSP dataset consists of WikiNews texts[2] by the Open Source pipeline of NewsReader (Vossen et al., 2016). The pipeline includes software for identifying events, relations between events, factuality of events and opinions. The interpretation program turning the linguistic representations of the NLP tools into RDF representation in GRaSP specifically targets Source Introducing Predicates (e.g. *say*, *believe*), identifying who said what according to the text. All content not in the scope of these predicates is attributed to the author of the text.

The interactive visualization showing perspectives on immigration and external EU borders in WikiNews.[3] is available on github and can be explored for better understanding of the following passage.[4] Figure 3 provides a partial screenshot. On the left hand side, the sources are provided. There are two lists of sources, the bottom list provides the authors or publishers of news articles. The top list provides sources quoted in the article. The events mentioned by the sources are displayed in the central image, with actual text on the right. Statistics on sentiment and factuality are provided by the diagrams at the bottom of the visualization. The visualization is interactive: sources and events can be selected leading to updates of perspective information and text.

5 Related Work

GRaSP offers ways to connect statements (in texts, video, images, etc.) to their source, the entities and events they mention and the stance they display. Arguably, this connection can be seen as a form of *annotation*. The Web Annotation Data Model (OA)[5] of the W3C represents annotations as the *related* combination of a body (the annotation) and a target (the annotated source). The relation is *directed*, the body says something about the target, but not vice versa. Directionality of OA, and the annotation view in general, is not compatible with the goals of GRaSP. A traditional annotation would just say that the link between an instance and its mention is a form of semantic enrichment of the text containing mention. The real question is: does the semantic representation of an instance determine how mentions should be understood, or do the combined mentions of an instance collectively determine its semantics? This nuance is of central importance when e.g. studying concept drift across historical sources, and it is the reason that GRaSP commits to the neutral *denotation* relation between instances and mentions. Secondly, the OA specification forces annotation targets to be *dereferencable*, which is problematic for sources that are not owned by the agent producing the annotations. License and other constraints may prohibit republication, and on a technical level dereferenceability cannot be guaranteed for sources hosted at an external location.

Marl[6] provides a model to represent subjective opinions in text. Marl is used by the Onyx ontology[7] for representing emotions expressed in text. It has also been combined with lexical information on sentiment from Lemon (Buitelaar et al., 2013).[8] GRaSP shares this flexibility of being compatible with various models that express aspects of perspectives. Unlike GRaSP, Marl is restricted to text. It furthermore confounds the layers that GRaSP carefully separates: the opinion (attribution in GRaSP) is a central node, that refers to

[2] https://en.wikinews.org/wiki/Main_Page
[3] http://wikinews.org/
[4] http://nlesc.github.io/UncertaintyVisualization/
[5] http://www.w3.org/TR/annotation-model/
[6] http://gsi.dit.upm.es/ontologies/marl/
[7] http://www.gsi.dit.upm.es/ontologies/onyx/
[8] http://lemon-model.net

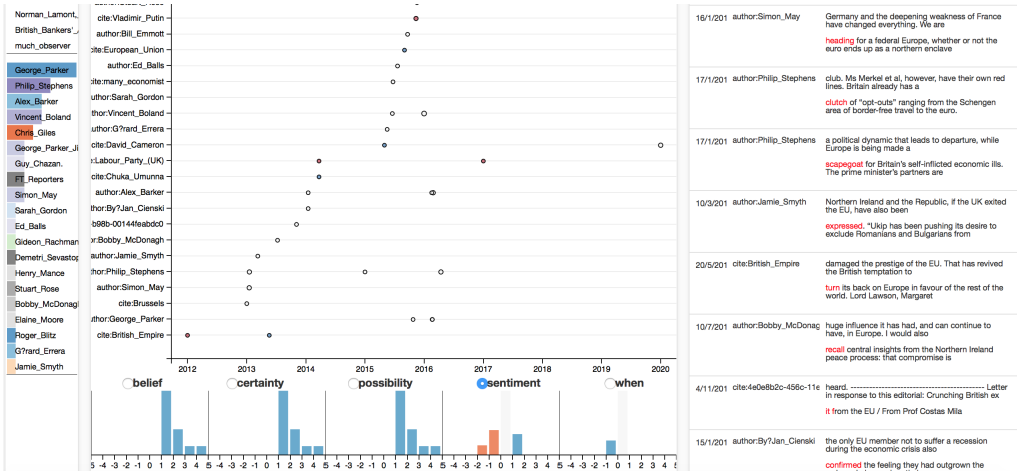

Figure 3: Screenshot of visualized perspective information

an object/feature (instance in GRaSP) and the literal text that reflects the opinion (mentions). This has two consequences. First, Marl only relates the opinion to the source (text or url) in which it was found without making the opinion holder explicit. GRaSP links mentions to their provenance and attributions of stances to the source that expressed the opinion. Marl thus does not seem to provide the means to collect all perspectives from a specific source. Second, GRaSP's separation of these layers makes it more flexible in dealing with alternative interpretations of mentions, both at the attribution and instance layer. Finally, GRaSP is not limited to explicitly subjective opinions, but can connect all stances taken by a source (including factual statements).

GRaSP can be combined with various existing models. We use PROV (Moreau et al., 2013) to model the provenance of mentions and interpretations made on them (i.e. to model the NLP process following Ockeloen et al. (2013)). The NLP Interchange Format (NIF, Hellmann et al. 2013) is an RDF/OWL vocabulary for representing NLP annotations in a common way, to foster interoperability between NLP tools, language resources and annotations. The core of NIF consists of a vocabulary and a URI design that permit describing strings and substrings, to which arbitrary annotations can be attached using vocabularies external to NIF. NIF itself does not specifically address the representation of source or attribution information, but can be combined with GRaSP. GRaSP bases the format of IRIs of mentions on NIF and uses it to represent some mention layer attributes (e.g. char offset in the text). Finally, GRaSP uses the grounding relations provided by GAF, as mentioned above. GRaSP's main contribution compared to GAF is that GRaSP adds an attribution layer tying sources and their stances to mentions.

6 Conclusion and Discussion

This paper introduces GRaSP, a formal framework to represent perspectives on content. The GRaSP framework was designed out of need from various NLP projects that deal with automatically identifying perspectives. We explained how GRaSP provides the structure to study perspectives from various view points (starting with a topic, source, sentiment, or stance). We provide a dataset actively using GRaSP that allows users to study the perspective various sources express on events in WikiNews.

The way perspectives are expressed in natural language is highly complex. Space limitations prevented us to illustrate how phenomena such as scope, alternative interpretations and framing can be represented in GRaSP. The wide range of possibilities for applying this and how researchers can deal with (lack of) accuracy of NLP tools also requires more space than available in a short paper. We plan to address these issues in future work.

Acknowledgement

The work presented in this paper was funded by the European Union through the FP7 project NewsReader and by the Netherlands Organization for Scientific Research (NWO) via the Spinoza grant, awarded to Piek Vossen and via VENI grant 275-89-029 awarded to Antske Fokkens.

References

Paul Buitelaar, Mihael Arcan, Carlos A Iglesias, J Fernando Sánchez-Rada, and Carlo Strapparava. 2013. Linguistic linked data for sentiment analysis. In *2nd Workshop on Linked Data in Linguistics*. page 1.

Paul Buitelaar, Daniel Olejnik, and Michael Sintek. 2004. A protégé plug-in for ontology extraction from text based on linguistic analysis. In *European Semantic Web Symposium*. Springer, pages 31–44.

Antske Fokkens, Serge Ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, and Guus Schreiber. 2014. Biographynet: Methodological issues when nlp supports historical research. In *LREC*. pages 3728–3735.

Antske Fokkens, Marieke van Erp, Piek Vossen, Sara Tonelli, Willem Robert van Hage, Luciano Serafini, Rachele Sprugnoli, and Jesper Hoeksema. 2013. GAF: A grounded annotation framework for events. In *The 1st Workshop on Events*. Atlanta, USA.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, pages 539–545.

Sebastian Hellmann, Jens Lehmann, Sren Auer, and Martin Brmmer. 2013. Integrating NLP using Linked Data. In *Proc. of ISWC*. pages 98–113. See also http://persistence.uni-leipzig.org/nlp2rdf/.

Yu-wei Lin. 2012. Transdisciplinarity and digital humanities: Lessons learned from developing text-mining tools for textual analysis. In *Understanding digital humanities*, Springer, pages 295–314.

Luc Moreau, Paolo Missier, Khalid Belhajjame, Reza B'Far, James Cheney, Sam Coppens, Stephen Cresswell, Yolanda Gil, Paul Groth, Graham Klyne, et al. 2013. Prov-dm: The prov data model. *Retrieved July* 30:2013.

Niels Ockeloen, Antske Fokkens, Serge Ter Braake, Piek Vossen, Victor De Boer, Guus Schreiber, and Susan Legêne. 2013. Biographynet: Managing provenance at multiple levels and from different perspectives. In *LiSc-Volume 1116*. CEUR-WS. org, pages 59–71.

Bernhard Rieder and Theo Röhle. 2012. Digital methods: Five challenges. In *Understanding digital humanities*, Springer, pages 67–84.

Janneke van der Zwaan, Maarten van Meersbergen, Antske Fokkens, Serge ter Braake, Inger Leemans, Erika Kuijpers, Piek Vossen, and Isa Maks. 2016. Storyteller: Visualizing perspectives in digital humanities projects. In *Computational History and Data-Driven Humanities: Second IFIP WG 12.7 International Workshop, CHDDH 2016, Dublin, Ireland, May 25, 2016, Revised Selected Papers 2*. Springer, pages 78–90.

Willem Robert van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 2011. Design and use of the Simple Event Model (SEM). *Journal of Web Semantics* 9(2):128–136.

Maarten van Meersbergen, Piek Vossen, Janneke van der Zwaan, Antske Fokkens, Willem van Hage, Inger Leemans, and Isa Maks. 2017. Storyteller: Visual analytics of perspectives on rich text interpretations. In *Proceedings of Natural Language Processing meets Journalism*. Copenhagen, Denmark.

Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, Marco Rospocher, and Roxane Segers. 2016. Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems* 110:60–85.

Fei Wu and Daniel S Weld. 2010. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 118–127.