Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3162–3171
Marseille, 11–16 May 2020

Large-scale Cross-lingual Language Resources for Referencing and Framing

Piek Vossen[a], Filip Ilievski[a], Marten Postma[a], Antske Fokkens[a], Gosse Minnema[b], Levi Remijnse[a]
[a]Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
{piek.vossen,f.ilievski, m.c.postma, antske.fokkens, l.remijnse}@vu.nl
[b]Rijksuniversiteit Groningen
Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, The Netherlands
g.f.minnema@rug.nl

Abstract

In this article, we lay out the basic ideas and principles of the project *Framing Situations in the Dutch Language*. We provide our first results of data acquisition, together with the first data release. We introduce the notion of cross-lingual referential corpora. These corpora consist of texts that make reference to exactly the same incidents. The referential grounding allows us to analyze the framing of these incidents in different languages and across different texts. During the project, we will use the automatically generated data to study linguistic framing as a phenomenon, build framing resources such as lexicons and corpora. We expect to capture larger variation in framing compared to traditional approaches for building such resources. Our first data release, which contains structured data about a large number of incidents and reference texts, can be found at `http://dutchframenet.nl/data-releases/`.

Keywords: framing, FrameNet, reference. situation semantics, events, cross-lingual text corpora

1. Introduction

We use language to tell stories and reflect on situations in the world. We describe these situations in many different ways, which often reflects different perspectives. Although there are many corpora capturing language, hardly any of these also represent the actual situations that texts refer to, let alone provide indications of which texts refer to the same situation. Event coreference corpora could serve this purpose as they are annotated for mentions of the same event. However, available event coreference corpora are very small and they exhibit hardly any *ambiguity*, i.e. there typically is one referent for each expression, nor *variation*, i.e. there are only one or few expressions for each referent (Ilievski et al., 2016; Postma et al., 2016).

Not having sufficient texts that refer to the same or similar situations, or not knowing which texts do, makes it difficult to investigate the different ways in which people make reference. It also hampers the development of systems to automatically resolve (cross-document) coreference and to understand and develop technology that detects how events are framed.

Imagine you want to create a text corpus that represents the language used to describe murders. How to proceed? You can use public corpora such as the Gigaword corpus (Napoles et al., 2012) and search for texts using keywords. How many murders will you find and will you find all murders? Referring to events as *murder* is actually already subjective and may miss situations that some people describe differently. Even if you get a substantial amount of texts about murders, we still do not know which texts make reference to the same murder. Such referential grounding is however a prerequisite to study differences in framing these events.

The project *Framing Situations in the Dutch Language*[1] tries to tackle this problem using the data-to-text method described in Vossen et al. (2018), which compiles massive text data (so-called *reference texts*) in different languages that is referentially grounded to specific event instances represented as so-called *microworlds*. We not only ground these texts but also automatically disambiguate mentions of these events in texts following a *one-sense-per-event-type* principle. Furthermore, we automatically derive the typical vocabulary and FrameNet frames (Baker et al., 2003) for different event types.

We believe that inferring typical expressions and frames is an efficient and comprehensive way to enrich text collections with framing interpretations. From the texts and referential data collected in this way, we eventually derive FrameNet lexicons, and automatic frame labelers.

In this paper, we describe our theoretical assumptions and hypotheses that form the basis of our approach to learn framing from referentially grounded texts. We further describe the Multilingual Wiki Extraction Platform (MWEP), which is the first publicly available implementation of the data-to-text method. We describe the first result of applying MWEP to some event types and languages to obtain typical expressions and frames.

This paper is structured as follows. We first describe our theoretical assumptions in Section 2. We then introduce our approach in Section 3. and our formal data model in Section 4. Section 5. provides the details of the MWEP platform. We validate our first results in Section 6. We conclude in Section 7.

2. Theoretical assumptions

In order to learn the typical ways of framing events, we need to obtain massive amounts of texts for the same event types. We assume that events of a single type exhibit similar coherence relations that form roughly similar stories. The FrameNet frames evoked by these events should therefore also reflect similar coherence relations and stand-out as prototypical frames.

[1] http://dutchframenet.nl

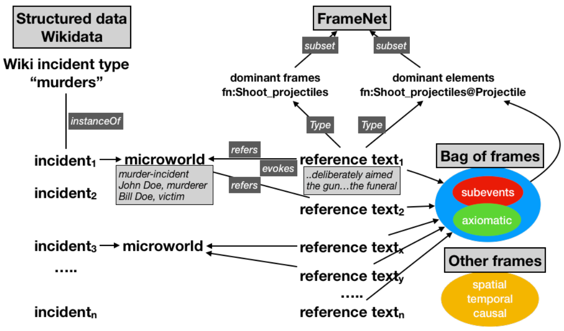


Figure 1: Overall schema for acquiring coherent text collections.

By contrasting evoked frames across different event types, we expect to find certain *typical frames* to occur more dominantly for specific event types, while others occur across many different types, such as abstract frames that express causal, temporal and spatial properties of specific situations. Frames that stand out with respect to one or a few types of events may point to typical subevents, e.g., many *murders* involve *shooting* as a subevent, but not all murders involve shooting, or they may be axiomatic, e.g., in the case of a *murder*, there must be an aspect of intentionality. We call the set of frames evoked with statistical significance for a type of event a *bag-of-frames*. Knowing the typical *bag-of-frames* for an event type supports framing research in two directions. First, we can create a strong expectation for the frames to be found in the reference texts of specific incidents of the same type. Secondly, we can compare specific reference texts grounded to the same incident for framing differences, i.e. which frames are used from the potential bag-of-frames and which are not. The former will help us to disambiguate expressions for the frames that they evoke, while the latter will learn the vocabulary for framing and show how much variation there is to frame the same incident across sources. For example, occurrences of the word *fire* in texts that refer to *murder* incidents are more likely to evoke the *fn:Shoot_projectiles* frame than the *fn:Firing* frame for employment relations. Similarly as observed by Cybulska and Vossen (2010), an incident such as the fall of Srebrenica can be described as a violent conflict with *shootings* and *transport* of women and children, focusing on the reporting, or as *deportation* and *genocide*, focusing on the intention and the responsibility.

There are two crucial questions to address for this to work: 1) at what level of *abstraction* do we need to aggregate text such that we maximize the volume of text that still exhibits coherent typical frames? 2) at what level of *granularity* do we need to aggregate text such that we maximize the volume of text and still obtain coherent temporal sequences of subevents.

The first question addresses the level of abstraction of different events at which they still share sufficient coherence relations. If we aggregate texts that report on very different situations, we may have a lot of data, but it will lack coherence. On the other hand, if texts are too specific, we will have little data per event type, similar data is unnecessarily dispersed, and we do not exploit the maximal generalization that is possible within the coherence constraint.

Following Rosch (1978), Rifkin (1985), and Morris and Murphy (1990), we assume that there is a natural *basic level* to categorize events, similar to the way we categorize objects. Above this basic level, we find superordinate levels of events that share only a few properties. Below this basic level, we find subordinate levels, at which events do not differ significantly in terms of properties. For example, *race* is expected to be a superordinate concept because there are many types of races each having very different properties, while *horse race*, *dog race*, and *marathon* may be at the basic level just below *race*. More specific races, such as *Kentucky Derby*, *Epsom Derby*, and *Grand National* are at a subordinate level of *horse race* and share many properties among each other.

In analogy with the findings for the basic level of concrete objects, we expect most event instances to be labeled with basic level categories (Hypothesis 1), and most subevents to be shared between events at the basic level compared to events at the superordinate level (Hypothesis 2). For our approach, it thus makes sense to aggregate event instances and reference texts at this basic level. This should maximize the acquisition of event instances that are still coherent in terms of frame relations, and at the same time maximize the prediction of frames and frame relations for the events described.

The second question addresses the granularity of the events to cover. We can describe events at very fine-grained levels, e.g., *(sub)atomic* and *chemical* events, but also as global or universal processes, e.g., *crime*, *climate change*, *evolution*, the *expanding universe*, or anything in between. Event granularity has a temporal dimension, in the sense that fine-grained events tend to have short durations, whereas global and universal processes have extreme large durations or are unbound. This meronymic-temporal dimension crosscuts the hyponymic basic-level dimension, comparable to Vossen (1995). We expect that people will group series of events at a granularity that fits their daily life and interest, for which Rosch (1978) already provided evidence in a pilot study. Hence, we expect people to register those events as *incidents* that consist of sequences of more fine-grained events. For instance, a *murder* event have subevents such as *pulling a gun*, *pointing*, *firing*, *hitting*, and *dying*. We do not expect people to register and describe the fine-grained subevents as such nor the fact that events such as *murders* or *races* could be embedded in more global phenomena, e.g., *life*, *crime*, *sport*. The *encyclopedic* information that people typically tend to add to Wikipedia and Wikidata includes certain events that are valued as noteworthy *incidents*. We therefore further distinguish *incidents* as specific social-cultural constructs that are a subtype of *event instances*. Whereas any change or relation can be an event, we see *incidents* as those events that are culturally and cognitively considered as explanatory containers at a typical granularity (Hypothesis 3). Incidents thus contain prototypical subevents that reflect these causal relations which explain why things happen and are considered more important, see Caselli and Vossen (2017) and Vossen et al.

(2015).

Our hypotheses can be summarized as follows: selecting the correct basic level of abstraction and granularity of *incidents* will yield a coherent set of subevents and frames that are significantly different from other incidents (H2 and H3), while most incidents are typically categorized at this level (H1). We want to exploit these hypotheses to structure the extracted data and predict the relevant framing for each dataset on the basis of the correct level of incident classification. If correct, typical frames as in Figure 1 can be obtained automatically using automatic frame labelers that are available for some languages, such as Open-SESAME for English (Swayamdipta et al., 2017), and by contrastive quantitative analysis of expressions in our reference texts. The frame distributions obtained in such a way can be projected to reference texts in other languages, e.g., using vocabulary mappings or cross-lingual embeddings. In the rest of this paper, we describe the implementation of our ideas in the MWEP platform and our first trials to approximate a basic level incident acquisition and frame extraction.

3. Overall approach

FrameNet's hierarchical structure of frame and subevent relations are still limited. Furthermore, it contains a wide variety of frames at different levels of granularity and abstraction, while the relations between these frames are dispersed. FrameNet thus does not lend itself directly for selecting a basic level of events. We therefore rely on the taxonomy of Wikidata and derive typical frames a posteriori.

In Figure 1, we show an overview of our approach. Following the data-to-text approach (Vossen et al., 2018), we query Wikidata for an assumed basic-level type of events to get the registered incidents, e.g., *murder* incidents. The Wikidata API will return the records with some structured data, from which we derive a so-called microworld. A microworld is defined as the minimal referential data to identify incidents in the world. It typically consists of the incident *type*, *date*, *location*, and *entities* that participate. In many cases, Wikidata also provides links to Wikipedia text pages in various languages that support the data. We consider these pages as secondary reference texts that report on the incident. In addition, the Wikipedia pages may point to primary reference texts that in turn support each Wikipedia page and also refer to the same incident. Likewise, we can rapidly aggregate several reference texts (possibly in different languages) that are referentially grounded to the same incident.

Pre-structuring the referentially-grounded texts for types of incidents has a number of advantages: 1) we can learn which frames are potentially relevant for a type of incident without relying on the FrameNet relations or having to consider all FrameNet frames, 2) texts can be pre-annotated automatically with these frames as ambiguity is reduced or even resolved, 3) referentially-grounded mentions of events and participants can be annotated in a more consistent way. Furthermore, the notion of a causal-temporal container that represents the incidents can be used to limit the annotation to events that fit in this container, in analogy to the notion of a temporal container used in the annotation of the Richer Event Description corpus (O'Gorman et al., 2016). Finally, the grounding will make it possible to analyze the reference texts for the different ways in which the same event is framed.

4. Model

In this section, we describe how we formalize the concepts defined in the previous section and the required data elements. Let R be a registry of real-world event instances. Let R_i be a real-world event instance and let $R_i \in R$. Each R_i contains structured data about the real-world event instance, e.g., the period or time when the event happened, its location, and information about which participants played a role and in which capacity. We model the structured data on the events according to the Simple Event Model (SEM, Van Hage et al. (2011)), which is an RDF model that formally distinguishes instances from their types and relates time, location and participant instances to event instances. The SEM representation of an event instance forms a *microworld*, which is a tuple consisting of [R_i, T_t, L_l, P_p] where T_t is a date instance in T, L_l is a location instance in L and P_p is a participant instance in P. These event instances or microworlds have an *rdf:instanceOf* relation with an event type. Let E_t be an event type, which is a categorization of a real-world event instance. The most abstract event type is *sem:Event*. More specific event types will have *rdf:subclassOf* relations with *sem:Event*, eventually forming a hierarchy of event types. Typically in our framework, the event hierarchy comes from Wikidata, and the most abstract event type is *event* (Q1656682).

In addition to the structured data, we collect reference texts. Let $PrimS$ be a primary source describing a real-world event (R_i), e.g., a news article as a reference text describing what happened. Let Sec be a secondary source describing the real-world event instance, which contains interpretations of primary sources. Typically, a secondary source makes use of several primary sources, which in some event registries are directly linked to one another. Each R_i can have multiple primary and secondary sources, all containing information about the same real-world event instance in many different languages. These sources are not necessarily parallel since their sentences and tokens may not be aligned but can be comparable since they provide information about the same real-world event instance.

Reference texts are sequential language structures whose expressions have meanings in contexts. In the reference text, there will be expressions e that can make reference to instances $R_i \in R$. Regardless of the reference, expressions have meanings. Let m be the meaning of an expression in the reference text. A meaning m can coincide or be equivalent to a type E_t of an event instance. Typically in our framework, event expressions are mapped to FrameNet frames. FrameNet frames may have some relation to Wikidata types, which is what we want to learn.

The reference relation of expressions to instances is formally captured by the Grounded Annotation Format (GAF, (Fokkens et al., 2013)). GAF[2] models reference relations as

[2] GAF is superseded by the more elaborate model GRaSP, the Grounded Representation and Source Perspective model (Fokkens et al., 2017). However for the current framework, GAF suffices to represent the basic referential relations.

gaf:denotedBy and *gaf:denotes* relations between expressions and instances. Expressions that make reference are mentions of the instance they refer to. Following a Fregean approach to reference, expressions with different meanings m can have a *gaf:denotes* relation to the same instance R_i. This formally models different ways of framing the same event instance.

In Figure 2, we provide a simplified example of the formal modeling of the data that is generated by our platform (see the next section). At the heart of the graph, we find an event instance derived from Wikidata with the identifier *Q618463*. It is related to other Wikidata instances through *sem:hasActor, sem:hasPlace* and *sem:hasTimeStamp* relations. Furthermore, it has labels in various languages coming from Wikidata and *rdf:subClassOf* relations to *sem:Event* and a FrameNet frame *fn:Change_of_Leadership*.

On the left side of the graph, we see *gaf:denotedIn* relations that indicate which primary and secondary reference texts mention the incident *Q618463*. We use the Dublin Core Terms (DCT) ontology[3] to represent their properties: title, description, source, type, and language. The *gaf:denotedBy* relations between specific expressions, and the instances are not shown here for readability. Within the reference text, specific expressions likely refer to specific instances modeled here or to properties or subevents of these instances.

This model allows us to study framing in a number of ways. The primary and secondary sources of the same real-world event instance provide us with insights about how sources describe the properties of the real-world event instance. In the case of a soccer match, some sources may describe it as one team winning a match. In contrast, other sources will focus on another team losing.

Primary sources about the same real-world event instance may differ in the information they cover. Some may focus on only some parts of structured data about the real-word event instance, e.g., only mention who won the election but not talk about specific candidates, whereas others may also contain a lot of background information.

The mapping of real-word event instances to event types is a valuable source of information. By grouping together the sources that describe the same event types, we are able to analyze how sources talk about the same event type.

5. Multilingual Wiki Extraction Pipeline (MWEP)

In this section, we describe the platform and resources for our incident extraction pipeline, and our data model. All code is freely available on GitHub: `https://github.com/cltl/multilingual-wiki-event-pipeline/`.

5.1. Resources

Wikipedia and Wikidata are two projects led by the Wikimedia community. These are simultaneously developed, which means that their information and guidelines are mutually consistent.

Wikipedia is a free online encyclopedia whose content has been collaboratively created and continuously updated by volunteers worldwide.[4] Wikipedia contains information in 307 languages, 297 of which are in active development.[5] The Wikipedia pages which describe the same topic across languages are explicitly connected by 'langlinks' (language links).

Each Wikipedia page consists of an initial 'abstract' description of an item, followed by a number of sections where specific aspects of that item are detailed further. Finally, the Wikipedia pages contain a list of external links (news documents, reports, books, ...) on which the Wikipedia page content is based.

Wikidata (Vrandečić and Krötzsch, 2014) is a free and open knowledge base that can be read and edited by both humans and machines. Wikidata is one of the largest knowledge bases in the linked data cloud today: at the moment of writing this paper, it described 62,557,696 items. Unlike Wikipedia and many other structured knowledge bases like DBpedia, Wikidata has a single, language-agnostic description of an item. This description contains various semantic information about an entity or event, such as a person's nationality and date of birth, or an event time and location. The items in Wikidata have labels in all entered languages, which mostly correspond to titles of language-specific Wikipedia pages that describe that item. In addition, there are explicit links between the Wikidata identifiers and the Wikipedia pages in various languages, which can be retrieved from the Wikimedia API.[6]

Wikidata organizes items through an ontology. The two most dominant relationships are *subclassOf* (Property 279) and *instance of* (Property 31). The *subclassOf* relationship is expressed between types, e.g., *presidential election* is a type of event that is a *subclassOf* the type *election*, whereas *instanceOf* relationships are expressed between an instance, e.g., *2012 French presidential election*, and a type *presidential election*.

We represent all *subclassOf* relations as a directed graph

$$G = (V, A)$$

where V is the set of nodes, i.e., the Wikidata items, and A is the set of directed edges, i.e., the *subclassOf* relations. In total, the directed graph of Wikidata contains over 2.3 million nodes and over 2.9 million edges. The average in- and out-degree is 1.3, and the root node is *entity* (Identifier Q35120).

We focus on a subgraph of the entire directed graph, i.e., we only make use of all nodes under the *event* node (Q1656682). For each event type, we query Wikidata to obtain the number of Wikidata items that are linked to the respective event type via an *instanceOf* relationship, which we call the instance frequency (Inst Freq). For example, the Wikidata item *2017 French presidential election* (Identifier

[3] `https://www.dublincore.org/specifications/dublin-core/dcmi-terms/2012-06-14/?v=terms`

[4] `https://en.wikipedia.org/wiki/Main_Page`

[5] `https://meta.wikimedia.org/wiki/List_of_Wikipedias`, visited on November 20th 2019

[6] `https://www.mediawiki.org/wiki/REST_API`

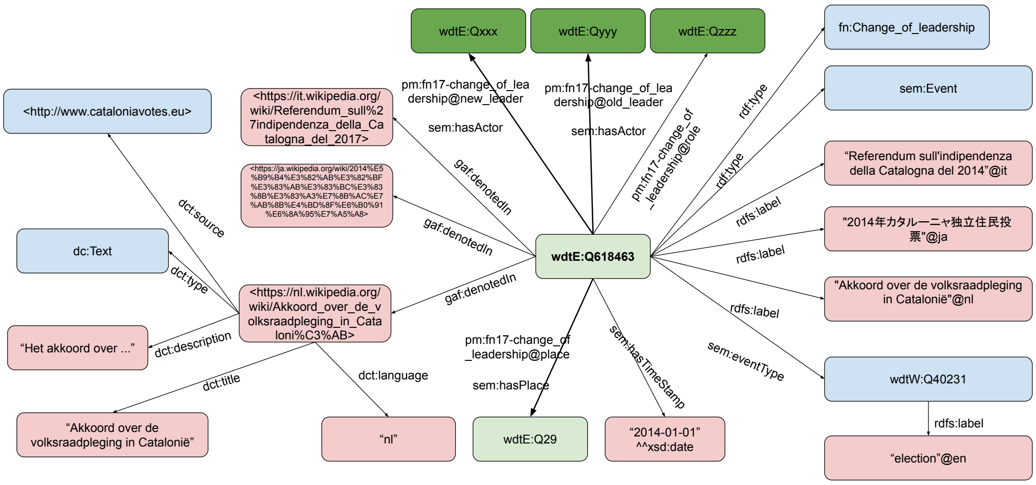

wdtE:Qxxx
wdtE:Qyyy
wdtE:Qzzz
fn:Change_of_leadership
sem:Event
<http://www.cataloniavotes.eu>
<https://it.wikipedia.org/wiki/Referendum_sull%27indipendenza_della_Catalogna_del_2017>
<https://ja.wikipedia.org/wiki/2014%E5%B9%B4%E3%82%AB%E3%82%BF%E3%83%AB%E3%83%BC%E3%83%8B%E3%83%A3%E7%8B%AC%E7%AB%8B%E4%BD%8F%E6%B0%91%E6%8A%95%E7%A5%A8>
pm:fn17-change_of_leadership@new_leader
sem:hasActor
pm:fn17-change_of_leadership@old_leader
sem:hasActor
pm:fn17-change_of_leadership@role
rdf:type
rdf:type
"Referendum sull'indipendenza della Catalogna del 2014"@it
rdfs:label
"2014年カタルーニャ独立住民投票"@ja
rdfs:label
"Akkoord over de volksraadpleging in Cataloniё"@nl
rdfs:label
dct:source
gaf:denotedIn
gaf:denotedIn
gaf:denotedIn
dc:Text
dct:type
wdtE:Q618463
<https://nl.wikipedia.org/wiki/Akkoord_over_de_volksraadpleging_in_Catalonі%C3%AB>
"Het akkoord over ..."
dct:description
dct:title
dct:language
pm:fn17-change_of_leadership@place
sem:hasPlace
sem:hasTimeStamp
sem:eventType
wdtW:Q40231
rdfs:label
"Akkoord over de volksraadpleging in Cataloniё"
"nl"
wdtE:Q29
"2014-01-01" ^^xsd:date
"election"@en

Figure 2: RDF model

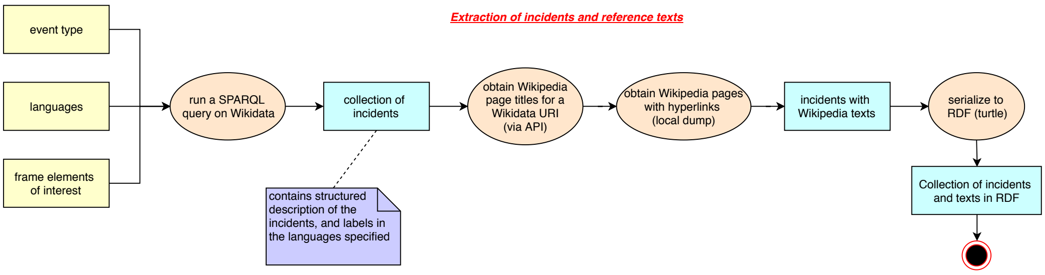

Extraction of incidents and reference texts
event type
languages
frame elements of interest
run a SPARQL query on Wikidata
collection of incidents
contains structured description of the incidents, and labels in the languages specified
obtain Wikipedia page titles for a Wikidata URI (via API)
obtain Wikipedia pages with hyperlinks (local dump)
incidents with Wikipedia texts
serialize to RDF (turtle)
Collection of incidents and texts in RDF

Figure 3: Extraction of incidents and reference texts.

Q7020999) is an instance of *presidential election in France*. Finally, we use the term *subsume* when an event type is a descendant of another event type via the *subclassOf* relationship. We remove all leaf nodes that have less than three Wikidata items linked to it via the *instanceOf* relationship. In total, there are 3,374 event types, i.e., items that are subsumed by the *event* item via the *subclassOf* relationship, of which 3,016 are leaf nodes. The total Instance frequency for all considered event types is 416,627.

5.2. Extraction

Figure 3 presents a schema of our extraction process. Through this process, we leverage the wealth of information in Wikidata and in Wikipedia, as well as their connections, in order to extract rich information about many incidents belonging to various types.

The input to the extraction process contains three parameters: languages, incident types, and optionally, frame elements of interest mapped to Wikidata properties. This means that the extraction process is designed and implemented in a generic way, and hence, is able to produce an incident collection for any specified list of incident types and languages. For instance, suppose that the languages specified are English, Dutch, and Italian; whereas the incident types are murder (`https://www.wikidata.org/wiki/Q132821`) and conflagration (`https://www.wikidata.org/wiki/Q168983`).

As a first step, the script fires a SPARQL query to the Wikidata endpoint. The results of this query are all found incidents of the requested types, together with their labels in the three requested languages (if found) and some structured information (time and country, if no further frame mappings were set through the script inputs).

Next, we query the Wikimedia API to obtain Wikipedia page titles in all three languages for each of the incident URIs in Wikidata. Then, we search for these potential page titles and the incident labels from Wikidata in our collection of Wikipedia pages.[7] When found, the Wikipedia page text and its hyperlinks are stored as a reference text for the corresponding incident. The set of Wikipedia pages provides parallel descriptions of the same incident in multiple languages: their content is not exactly the same, but referentially we can safely assume that they describe the same main incident. After this process, we have an updated incident collection that contains both structured and unstructured descriptions of each incident. The collection is also serialized as RDF, following the model in Figure 2.

5.3. Selection of highest quality data and further processing

Figure 4 shows the next steps of data selection and semantic processing. Based on a set of quality/completeness criteria, we make a selection of incidents from the general set. These, for example, might be incidents with most-complete metadata or multilingual descriptions. The aggregated collection is firstly enriched with primary reference

[7] Currently, we use a local dump of Wikipedia from July 20th 2019, loaded through a customized reader: `https://github.com/cltl/Wikipedia_Reader`.

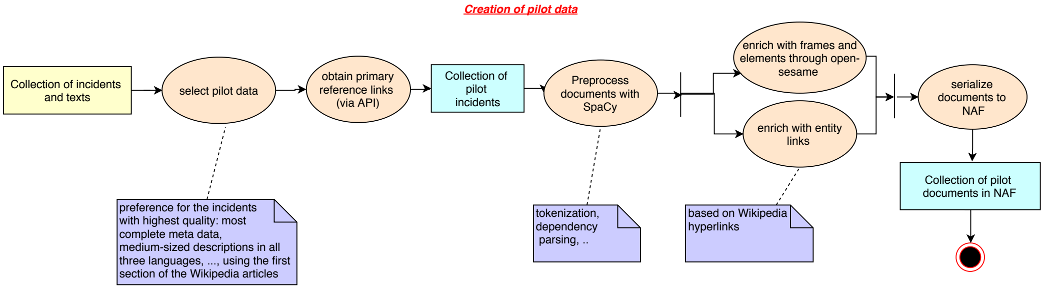

Creation of pilot data
Collection of incidents and texts
select pilot data
obtain primary reference links (via API)
Collection of pilot incidents
Preprocess documents with SpaCy
enrich with frames and elements through open-sesame
enrich with entity links
serialize documents to NAF
Collection of pilot documents in NAF
preference for the incidents with highest quality: most complete meta data, medium-sized descriptions in all three languages, ..., using the first section of the Wikipedia articles
tokenization, dependency parsing, ..
based on Wikipedia hyperlinks

Figure 4: Creation of pilot data.

links, which are external references listed in Wikipedia pages. We obtain these via the Wikimedia API.

Next, we enrich the text with semantic annotations, both based on given information and automatically induced. The annotated Wikipedia page is stored as NAF (NLP Annotation Format, (Fokkens et al., 2014)), a stand-off, multilayered annotation schema for representing linguistic annotations.[8] We process each document with spaCy[9] and store the tokenization results in the NAF token layer. We attach the previously preserved hyperlinks as entity links in NAF. We run Open-SESAME (Swayamdipta et al., 2017) to produce potential FrameNet frames and store these in the Semantic Role Layer of NAF.[10]

The final result of our extraction pipeline is an incident collection that contains structured descriptions for each incident, but also a NAF file per Wikipedia page with information on its tokens, entities, and semantic roles and partial referential relations for entities.

6. Results

In this section, we report on the first results of extracting data and frame statistics. We first describe the MWEP pilot corpus and explain how we can set the basic level for incidents to derive a comprehensive set of event types and show what the contrastive frame analysis may yield. Finally, we gain more insights into basic level events through manual annotation.

6.1. MWEP corpora

Table 1 shows statistics for ten diverse event types and three languages (English, Italian, and Dutch). We typically obtain (tens of) thousands of incidents with at least one Wikipedia (secondary) reference text. Many of these incidents have a description in more than one language. A subset of these incidents contains information for all fields of our structured data, ranging between 93 for exhibition events and 1,260 for award ceremonies.

As elaborated in Figure 4, the seed collection of incidents is used as a basis for selecting high-quality and well-described incidents by setting different criteria. In Table 1, we show statistics per criterion: having an English description and “full info” (data for each of the structured data properties). As expected, we obtain fewer incidents with a higher density of information (presumably, incidents that are well known and for which there is more data). For the pilot incidents, we also obtained a list of primary reference text URIs. The average number of primary reference texts per incident type is generally high but varies a lot (between 4.18 and 57.4).

The time needed to extract incidents for all ten event types is slightly over an hour on a simple machine (Ubuntu 18.04 operating system, 2 CPUs, and 8 GBs of RAM memory), allowing us to scale up extraction to many incidents for various event types and languages. This process can be optimized by parallelization of the MWEP pipeline.

6.2. Basic level event detection

Following our coherence principle, we want to extract as much data as possible per event type, which still provides us coherent sets of frames and sufficient variation in framing. We reimplemented the procedure from Izquierdo et al. (2007) to detect basic levels in the Wikidata hierarchy.

The following parameters are used in the procedure: 1. **node weight property**: the property used to weigh the node in the graph. 2. **subsumer threshold (ST)**: the number of event types that an event must subsume to be considered a basic level. 3. **root weight zero**: whether the weight of the root node is set to zero. Following Izquierdo et al. (2007), we use the notion of *local maximum* to select candidates for basic level types. A local maximum occurs when the node weight of the child and the parent of a node are lower than that of the node itself. Given a path from a leaf node to a root node, there can be multiple local maxima. Different from Izquierdo et al. (2007), we apply this to *instance of* relations in Wikidata rather than just *subclass of* relations.

The procedure to detect basic levels is as follows. For each leaf node in the graph, we query all paths to the root node and determine their local maxima. We select the path with the local maximum with the highest node weight. The result is a set of basic level types shared by multiple leaf nodes. A local maximum is discarded if the number of event types it subsumes is lower than the subsumer threshold. Furthermore, if a basic level is superseded by another type, we prefer the more specific type.

We apply the basic level detection procedure with the following settings: 1. **node weight property**: we use the instance frequency (Inst Freq) of an event type as its node weight. This setting is inspired by Morris and Murphy (1990), who provided evidence that humans label incidents

[8] https://github.com/newsreader/NAF

[9] https://spacy.io

[10] Our wrapper is available at: https://github.com/cltl/run_open-sesame.

		total			*pilot*		
event type	Q ID	# inc	# SRTs	# full info	# inc	# SRTs	mean PRTs
conflagration	Q168983	693	769	155	146	189	28.6
murder	Q132821	1,667	2,141	209	200	326	57.4
exhibition	Q464980	1,682	2,122	93	80	160	20.45
festival	Q132241	7,389	9,012	644	521	804	16.83
award ceremony	Q4504495	4,983	6,691	1,260	1,057	1,387	5.63
horse race	Q3001412	2,604	2,633	274	272	275	4.18
film festival	Q220505	2,296	2,802	468	425	627	11.96
marathon	Q40244	1,297	1,452	555	65	126	4.96
military operation	Q645883	2,763	3,924	193	172	273	22.04
beauty pageant edition	Q62391930	1,595	1,826	1,197	1,104	1,149	7.59

Table 1: Statistics on extracting incidents for the languages EN, IT, and NL, for 10 event types. *Columns:* event type, Q ID from Wikidata, initial number of incidents (total # inc), initial number of secondary reference texts (total # SRTs), initial number of incidents with full information (total # full info), number of incidents in the pilot collection (pilot # inc), number of secondary reference texts in the pilot collection (pilot # SRTs), average number of primary reference text URIs in the pilot data (pilot mean PRTs), total time in seconds.

ST	# of BL	# of unique BL	Avg Depth	Avg Desc	Avg Inst Freq
0	1153	102	4.0	14.6	3151.7
10	1615	39	3.8	61.0	8258.1
20	1536	25	3.6	88.2	9913.8
30	2269	16	3.2	187.2	11119.3
40	2209	12	3.5	238.2	12809.2

Table 2: For each subsumer threshold (ST), the number of basic levels (# of BL), the unique number of basic levels (# of unique BL), the average node depth of the basic levels (Avg Depth), and the average number of descendants (Avg Desc). Finally, we show the average instance frequency, i.e. how many Wikidata items have been tagged with a certain event type.

cut-off point	5	10	20	40
precision	0.27	0.28	0.28	0.26

Table 3: Precision of validated typical frames per cut-off point.

#	**Presidential election**	**FFICF**	**Tennis tournament**	**FFICF**
1	Leadership	0.65	Judicial_body	0.53
2	Change_of_leadership	0.55	Calendric_unit	0.39
3	Apellations	0.36	Performers_and_roles	0.36
4	Political_locales	0.15	Spatial_contact	0.26
5	Calendric_unit	0.13	Part_whole	0.24

Table 4: Example top-5 frames with **FFICF** scores for two event types.

predominantly at a basic level (Hypothesis 1). 2. **subsumer threshold (ST)**: the number of event types that an event must minimally subsume to be considered a basic level. We experiment with the following settings: 0, 10, 20, 30, and 40. 3. **root weight zero**: we set the weight of the root to zero. Table 2 shows the number of basic level nodes we obtain given different thresholds.

We see that setting no threshold will give us 102 basic level types, and setting a threshold of 40 gives us 12 unique event types. We can also see that the average depth of the types ranges from 4.0 to 3.2, suggesting that the smallest sets, i.e., thresholds 30 and 40, are at higher levels of abstraction and dominate most descendant nodes: 238.2 descendant types. By way of illustration, very specific event types are identified as a basic level when not applying a threshold, e.g., *gubernatorial election* and *Esperanto meeting*, whereas very general ones are detected at a threshold of 40, e.g., *recurring sporting event* and *legal transaction*. We expect the optimal frame coherence between these extremes. The table shows how we can spread the basic level types. What threshold yields the most coherent and largest data set is to be determined empirically. We can already see in the instance frequency column (Avg Inst Freq) that within the two extremes, we will obtain a large number of incidents per event type.

Based on manual inspection, we selected the set of event types resulting from running the Basic Level Event Detection system with a threshold of ten for our first data release. From the set of 39 basic level events, we selected the event types for which, according to our Wikidata representation, between 500 and 10,000 incidents are tagged with those event types. 25 event types met these requirements. After running MWEP for the selected set of event types, we obtained a total of 26,778 reference texts (average: 1,071) and 557,616 tagged frames by Open-SESAME (average: 22,304). Our first data release, which contains structured data about the incidents and reference texts of the described event types, can be found at `http://dutchframenet.nl/data-releases/`.

6.3. Typical frame detection

To evaluate which frames are most typical for a collection of texts for a particular event type, we applied a quantitative analysis based on the Open-SESAME frame annotations as part of our first data release. We derived a typicality score for each frame in each collection by using an adapted version of the **TFIDF** metric. We call this the **FFICF** metric, where *CF* is the number of collections in which a frame occurs, and *FF* is the frame frequency in that collection. For every event type, we ranked all of the frames occurring in the corresponding text collection by this metric and selected the top-40 frames for further analysis. First, we manually judged the typicality of each frame-event type pair. We then determined different cutoff points to find out how well high-ranked frames correspond to the manually annotated set. Table 3 shows precision scores for these cutoffs, averaged over event types, and Table 4 shows the top-5 frames for two event types, *presidential election* and *tennis tournament*.

We observe that the overall precision score is not only low, but remains stable across cut-off points, which means that the typical frames are not necessarily the most highly ranked ones, but are spread across the top-40. Yet, a closer examination of the data reveals that variables were possibly affecting the overall precision score. First of all, there is much variation across event types. For some event types, such as *presidential election*, the top-ranked frames closely correspond to manual judgments, with a precision score of 0.35 for the top-40 and 0.80 for the top-5, which includes the frames *fn:Leadership* and *fn:Change_of_leadership*. On the other end, *tennis tournament* shows an overall low precision score of about 0.28, decreasing to 0.20 for the top-5, which includes frames such as *fn:Calendric_unit* and *fn:Spatial_Contact*. We observe that overall, event types related to sports tournaments and races show low precision scores with typical frames occurring at a low cut-off point.

Another potential factor influencing the quality of the frame rankings is the difference in size between the text collections for different event types. One might expect that a larger volume of reference texts per event type leads to higher precision. However, this is not supported by the data: *tennis tournament* displays the largest number of texts (4,988), while it gets an overall low precision score. Meanwhile, *Contract* displays the lowest number of texts (7), while it gets an overall higher precision score, with typical frames such as *fn:Make_Agreement_On_Action* and *fn:Be_In_Agreement_On_Action* in the top-5 ranking. *Contract* might benefit from its conceptually distinct character in comparison to sports, races, and elections. Even though the FF for the frames in this event type is low given the small number of reference texts, the ICF is also low since the frames are not likely to occur in the contrasting event types.

Additionally, the low precision scores can also be attributed to frame identification errors by Open-SESAME. For example, the highest-ranking frame for *tennis tournament* is the irrelevant frame *fn:Judicial_Body*. Our lexical unit analysis shows that this is due to the frequently occurring lexical unit "court", which is wrongly classified as referring to a judicial court rather than to a tennis field. This is consistent with findings from the FrameNet semantic role labeling literature that show that frame identification is a major challenge (Hartmann et al., 2017), especially on datasets outside of the domains covered by the FrameNet corpus.

We conclude that our first results and the contrastive analysis did not yet confirm our hypotheses, i.e., we could not derive high-quality typical frames for all types of events. This can be due to a number of factors: 1) FrameNet frames are not specific enough to distinguish these types of events, 2) Open-SESAME generates too much noise due to frame identification errors, 3) our basic level assignment is too specific (too much similarity across event types) or too general (too much lumping of similar events). To investigate 2), we will experiment with other disambiguation strategies in the future. We tested factor 3) by lumping together all sports events and recalculating the FFICF scores. This did, however, not result in a more consistent ranking. To obtain a more precise division in types of events, we, therefore, carried out a manual annotation of events that stand in a *subclassOf* relation, which is described in the next section.

6.4. Basic Level Event Annotation

We carried out a pilot study to determine basic level events through manual annotation. We base the experimental setup on two experiments in which evidence was provided for basic level events as performed by Morris and Murphy (1990). In Experiment 1 of Morris and Murphy (1990), subjects were asked to list actions of an event. The outcome indicated that as event names increase in generality from a subordinate level to a basic level, there was a limited loss of information. Also, the results from Experiment 4b, in which subjects were asked to provide an event label to a story, showed that the most concrete information is provided at the basic level, and subordinate events may only be used when this is required for communication.

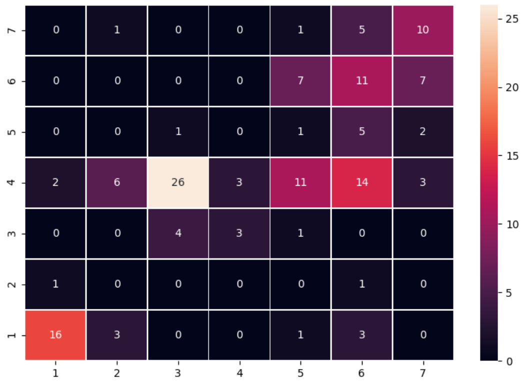
Along a similar line, we presented two annotators with two event labels connected through a *subclass of* Wikidata relationship. They had to answer two questions about a pair of event labels generalising from the child to its parent label: 1. *participants* What proportion of the main type of participants are alike? The main type of participants can be *bike* and *riders* in the case of a *cycling race*. 2. *subevents* What proportion of the main subevents are alike? Subevents of a *cycling race* may include *start* and *finish*.

The annotators had to indicate their judgment on a seven-point Likert scale (Likert, 1932). Additionally, there was a *do not know* option in case the annotator is unsure and a *non-sensible* option in case irrelevant.

We selected a subgraph of the Wikdata representation from Subsection 5.1. for annotation. We removed all leaf nodes from the graph as they are too specific to be basic level events. From the resulting trimmed graph of the previous step, we only retained the leaf nodes that had 25 or more incidents linked to it. Also, given a parent with more than two children that are leaf nodes, we only keep the two children with the highest incident frequency. We selected the subgraphs from the top ten children from the event node (Q1656682) with the highest cumulative incident frequency.

Two trained linguists annotated a total of 168 Wikidata edges. For the subevents subtask, both annotators provided a numeric answer for 149 pairs, i.e. *do not know* or *non-sensible* for 17 pairs. There was a full agreement for 20% of the cases and a maximum difference of one for 60%. For the participants subtask, the agreement was higher, for which a detailed comparison is shown in Figure 5. There was a full agreement for 30% of the cases and 76% for a maximum difference of one. We will further analyze the results of the subtask with the highest agreement, which was the participants subtask.

We focus the analysis on nodes for which there are annotations for both the edges to the subordinates and the superordinates. We call these nodes candidate basic level events. For the participants subtask, we formulate the *basic level eventness* (see Figure 6) of a candidate basic level event by computing the difference between: 1. Similarity$_{\text{subordinates}}$: the average Likert values of the subclass of edges between the subordinates nodes and the candidate basic level node.



	1	2	3	4	5	6	7
7	0	1	0	0	1	5	10
6	0	0	0	0	7	11	7
5	0	0	1	0	1	5	2
4	2	6	26	3	11	14	3
3	0	0	4	3	1	0	0
2	1	0	0	0	0	1	0
1	16	3	0	0	1	3	0

Figure 5: The annotation comparison between the two annotators for the participants task.

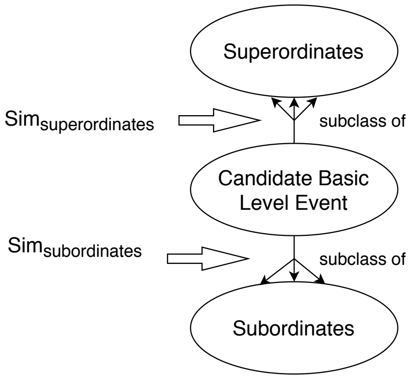

Superordinates
Sim$_{\text{superordinates}}$
subclass of
Candidate Basic Level Event
Sim$_{\text{subordinates}}$
subclass of
Subordinates

Figure 6: Determining Basic Level Eventness score.

2. Similarity$_{\text{superordinates}}$: the average Likert values of the subclass of edges between the candidate basic level node and the superordinate nodes. The *basic level eventness* score is determined by subtracting the Similarity superordinates by the Similarity subordinates.

The participants subtask contains 53 candidate basic level events. The basic level eventness scores range from -2 to 4.2, with an average score of 1.23. A positive score indicates that the similarity from the subordinates to the candidate basic level was higher than the similarity to the superordinates. The highest scoring nodes were *festival*, *ceremony*, and *holiday*, whereas the lowest scoring nodes are *recurring tournament*, *Olympic sport*, and *tournament*. Most of the high-scoring nodes, i.e., a score of 3 or higher, are situated relatively high in the graph with an average node depth of 2.8 and a standard deviation of 0.83. The lower-ranking nodes, i.e., with a score lower than 3, are situated lower in the graph, i.e., with an average depth of 3.46, and also have a higher standard deviation of 1.22, i.e., they are more dispersed. The depth property appears to be important in detecting basic level events. Still, other properties such as the incident frequency and the number of subordinates are also likely to play an essential role in determining basic level events. Figure 5, shows the heat-map for the participant annotation comparing the two Likert scores. We clearly see a high degree of overlap in the diagonal areas and most zero cases of in the extreme disagreement areas. This shows a high overall agreements in judgements with respect to the basic level criteria. In our future research, we will use the manually annotated data to calibrate our automatic techniques to derive the basic level and to improve the extraction of typical frames.

7. Conclusions

In this paper, we introduced the project *Framing Situations in the Dutch Language*, which started in April 2019. We described our theoretical assumptions and hypotheses, and we described the first implementation of the data-to-text approach as the Multilingual Wiki Extraction Pipeline (MWEP) platform. The data resulting from this platform contains structured data about incidents and a large number of reference texts that all make reference to the same incidents. We reasoned over the Wikidata event ontology to determine a set of event types used for a first data release, for which we also gained more insight using manual annotation. Also, we predicated typical frames for each event type in the first data release by contrasting collections of event types. We performed a first validation of our hypotheses on frame coherence, granularity of events, and dominance in relation to an assumed basic level. Our current data does not yet provide strong evidence for these hypotheses, i.e., we could not derive high-quality typical frames for all types of events. We described a number of possible explanations for these results. In future experiments, we will experiment with better frame disambiguation approaches than Open-SESAME and alternative approaches for contrastive analysis. We also carried out a manual annotation to establish a basic level. This will allow us to further calibrate the automatic detection of the optimal level in a future data release. The validated typical frames will then be used to calibrate the levels for the collections and to guide the annotation of the reference texts. Our first data release contains incidents and reference texts for 25 event types and can be found at `http://dutchframenet.nl/data-releases/`. For the future, we plan the following work in the project: 1. annotations and annotation comparisons 2. projection of frames across languages 3. automatic frame labelers for English, Dutch, and Italian 4. FrameNet lexicons and lexicon extensions.

8. Acknowledgements

The research reported in this article was funded by the Dutch National Science organisation (NWO) through the project *Framing situations in the Dutch language, VC.GW17.083/6215.*

9. Bibliographical References

Caselli, T. and Vossen, P. (2017). The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction. In *Computing news storylines and events workshop, ACL-2017.*

Cybulska, A. and Vossen, P. (2010). Event models for historical perspectives: Determining relations between high and low level events in text, based on the classification of time, location and participants. In *Proceedings of* *the Seventh International Conference on Language Resources and Evaluation (LREC'10).*

Fokkens, A., van Erp, M., Vossen, P., Tonelli, S., van Hage, W. R., Serafini, L., Sprugnoli, R., and Hoeksema, J. (2013). GAF: A Grounded Annotation Framework for Events. In Eduard Hovy, et al., editors, *Proceedings of the 1st workshop on Events: Definition, Detection, Coreference, and Representation at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL2013)*, Atlanta, GA, USA, Jun 9-15. Association for Computational Linguistics.

Fokkens, A., Soroa, A., Beloki, Z., Ockeloen, N., Rigau, G., Van Hage, W. R., and Vossen, P. (2014). NAF and GAF: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 9–16.

Fokkens, A., Vossen, P., Rospocher, M., Hoekstra, R., and van Hage, W. (2017). GRaSP: Grounded Representation and Source Perspective. In *Proceedings of KnowRSH*, Varna, Bulgaria.

Hartmann, S., Kuznetsov, I., Martin, T., and Gurevych, I. (2017). Out-of-domain FrameNet semantic role labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 471–482, Valencia, Spain, April. Association for Computational Linguistics.

Ilievski, F., Postma, M., and Vossen, P. (2016). Semantic overfitting: what world do we consider when evaluating disambiguation of text? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1180–1191.

Izquierdo, R., Suárez, A., and Rigau, G. (2007). Exploring the Automatic Selection of Basic Level Concepts. In *Proceedings of RANLP*, volume 7.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*, 22(140).

Morris, M. W. and Murphy, G. L. (1990). Converging operations on a basic level in event taxonomies. *Memory & Cognition*, 18(4):407–418.

Postma, M., Ilievski, F., Vossen, P., and van Erp, M. (2016). Moving away from semantic overfitting in disambiguation datasets. In *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*, pages 17–21.

Rifkin, A. (1985). Evidence for a basic level in event taxonomies. *Memory & Cognition*, 13(6):538–556.

Rosch, E. (1978). Principles of categorization. *Cognition*.

Swayamdipta, S., Thomson, S., Dyer, C., and Smith, N. A. (2017). Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold. *arXiv preprint arXiv:1706.09528*.

Van Hage, W. R., Malaisé, V., Segers, R., Hollink, L., and Schreiber, G. (2011). Design and use of the Simple Event Model (SEM). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):128–136.

Vossen, P., Caselli, T., and Kontzopoulou, P. (2015). Storylines for structuring massive streams of news. In *Proceedings of the First Workshop on Computing News Storylines*, pages 40–49, Beijing, China, July 26-31.

Vossen, P., Ilievski, F., Postma, M., and Segers, R. (2018). Do not Annotate, but Validate: a Data-to-Text Method for Capturing Event Data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).*

Vossen, P. (1995). *Grammatical and Conceptual Individuation in the Lexicon*. Studies in language and language use. IFOTT.

10. Language Resource References

Baker, C. F., Fillmore, C. J., and Cronin, B. (2003). The Structure of the FrameNet Database. *International Journal of Lexicography*, 16(3):281–296.

Napoles, C., Gormley, M., and Van Durme, B. (2012). Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100. Association for Computational Linguistics.

O'Gorman, T., Wright-Bettner, K., and Palmer, M. (2016). Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.