From Incidents to Framing: a Dutch and English Frame-Semantic Corpus and Lexicon

Piek Vossen, Pia Sommerauer, Levi Remijnse
Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081HV, Amsterdam, The Netherlands
piek.vossen@vu.nl, pia.sommerauer@vu.nl, levi_remijnse@hotmail.com

Abstract

This paper reports on the final results of the Dutch FrameNet project. The project followed a new approach to aggregate an event corpus starting from registered events in Wikidata and collecting text in different languages that refer to these events. The resulting corpus is not only referentially grounded, but it is also grouped by the type of event, e.g. mass shootings, elections, sports events. A subset of the texts has been annotated with FrameNets frames for all references to the registered events and participants. The result is a unique corpus with comparable texts across languages that make reference to the same and similar events. From the annotations, we derived Dutch and English FrameNet lexicons, as well as reference lexicons. These lexicons allow us to infer abstractions from the annotations that also reflect sociocultural differences in framing the same entities and events.

Keywords: FrameNet corpus, data to text, FrameNet lexicon

1. Introduction

Linguistic resources on framing are hard to obtain, especially to enable comparing framing across languages and different sources. The Berkeley FrameNet project (Baker et al., 1998) created a lexical database and an English text corpus annotated with the frames and frame elements defined in the database. This initiative has also been followed by several groups for other languages, now organized in the Global FrameNet Initiative (Torrent et al., 2020).

However, although the FrameNet resources for the non-English languages follow the same approach and database of frames and frame elements, the selection and annotation of texts is done independently in each project. This makes it difficult to compare differences in framing similar events and situations across languages and cultures.

This paper reports on the final results of the Dutch FrameNet project. The project followed a new approach to aggregate an event corpus starting from registered events in Wikidata and collecting text in different languages that refer to these events, the so-called data-to-text method (Vossen et al., 2018a). In this way, the corpus is not only referentially grounded, but it is also grouped by the type of event, e.g. shoot downs, elections, sports events. Grouping texts by events and event types helps annotating mentions in the text both with frames and with reference relations. Event mentions belonging to the same type will tend to share frames and frame elements. Texts grounded to the same event instance, will share reference relations.

We used the data-to-text method to collect Dutch and English texts and annotated these with frames and frame elements, resulting in a unique corpus with comparable texts across languages that make reference to the same and similar events. From the annotations, we derived Dutch and English FrameNet lexicons, as well as reference lexicons. These lexicons allow us to infer abstractions from the annotations that also reflect sociocultural differences in framing the same entities and events.

Our contributions are:

- We applied the data-to-text method to create a freely available corpus of English and Dutch texts that are referentially grounded to events and entities in Wikidata.
- We provide the structured data for the events and entities that are referenced in the texts.
- We provide annotations of the corpus in terms of FrameNet frames, frame elements and Wikidata references.
- We derived English and Dutch FrameNet lexicons from the annotations.
- We derived English and Dutch reference lexicons from the annotations.

All our code, corpora and lexicons are available on Github under an open source license.[1]

2. Related Work

There are many ways in which people can tell the same story. FrameNet (Baker et al., 1998) is the first attempt to address this diversity of perspectives by developing a database of conceptual schemes for events, called *frames*, which can be used to

[1] https://cltl.github.io/resources/dutchframenet/.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3684–3696
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

describe how events and their participants can be framed in the English language. Today, there are FrameNets in various other languages following a similar approach (Torrent et al., 2020) but not yet for Dutch. The creation of FrameNet lexicons roughly follows a bottom-up approach starting from text corpora that first have been parsed syntactically. Next, the main predicate of a clause is annotated with a frame, and the complementation with the associated frame elements.

The downsides of the bottom-up approach are that 1) all frames and related frame elements need to be considered for each sentence in a diverse set of texts by the annotators and 2) it is difficult to study variation in framing because different texts most likely talk about different events. For these reasons, Vossen et al. (2018a) developed a so-called data-to-text method that collects texts for events that are registered on Wikidata. This makes it possible to gather many texts that are referentially grounded and talk about the same event or describe the same type of event. Annotating these texts is more efficient and coherent as groups of texts are considered that are either grounded to the same event or describe similar events, hence relate to a small subset of all frames and frame elements. This method has been followed by others (Eirew et al., 2021, 2022) and has also been shown to allow annotating frames and elements of frames at the discourse level rather than just the clause level ((Remijnse et al., 2022, 2021). The method has been shown to increase the agreement between annotations and the speed of the annotation (Vossen et al., 2018a).

In this paper, we describe the result of applying this method to specific types of events to extract referentially grounded corpora for English and Dutch and infer from these both FrameNet lexicons and ReferenceNet lexicon (Vossen et al., 2018b) for English and Dutch. Comparing the frames and frame elements across different texts describing similar events at scale provides insight into the different ways that these events are framed.

3. From Data to Text

The data-to-text method was applied in various cycles to create referentially grounded corpora for various types of events. A subset of these events was also annotated with frames, frame elements, and references to events and their participants.

The approach is centered around the concepts **event types**, **incidents**, and **reference texts**. Event types organize events at a conceptual level; for instance, different murder incidents all share certain conceptual properties. Conceptually, we organize them under a single event type. The specific incidents are registered[2] instances of events of a certain type, which results in referential grounding; they can be defined in terms of actual participants, place, and time. The texts reporting on these incidents are what we call **reference texts**.

3.1. Data Collection

The data collection was driven by the goal of creating referentially grounded corpora, that is, corpora for which it is known what events, participants, time periods, and places the texts refer to. This approach has the following key advantages: (1) it allows for collecting many different texts that describe the same incident (or same type of event). This is likely to lead to a high degree of variation in how the same event can be referred to. (2) it supports the annotation process as it reduces ambiguity. Annotators can start from a point where they know the key facts about an event and can use this knowledge of context when approaching a text. A referentially grounded corpus allows to study (variation in) framing and to develop (cross-)document entity and event coreference at scale.

The data collection was conducted in several steps leading to different versions of the corpus. In this paper, we present the developments leading up to the complete corpus.

Event type corpus The first release of the current corpus was presented by Vossen et al. (2020). The corpus was collected using the Multilingual Wiki Extraction Pipeline, a tool designed to extract reference texts in multiple languages starting from a Wikidata identifier (i.e. a Q-number referring to a specific incident). The extraction process works as follows: via Wikidata, the links to Wikipedia pages in the target languages are retrieved. The texts describing the incidents on the Wikipedia pages serve as **secondary reference texts**. We call these texts secondary reference texts, as they are based on primary sources listed in the *Reference* section of a Wikipedia article. In addition to extracting these secondary reference texts, we can use references with a hyperlink to a primary source to also extract the primary source texts (called **primary reference texts**). In the case of incidents, the primary source texts tend to be news texts. By extracting reference texts of the same incident in multiple languages, it is possible to collect texts grounded in the same situation. Such texts are not truly parallel (i.e. not direct translations of one another), but share the same referential context.

The initial collection focused on extracting data representative of conceptually meaningful event

[2]Note that not all events are registered. We tend to only register socially and culturally important events with some impact.

types that could easily be linked or would contain data representative of specific FrameNet frames. This rationale resulted in an approach that aimed to identify basic-level event types in the Wikidata hierarchy (for details, refer to Vossen et al. (2018a)). Initially, 25 basic-level event types were selected after manual inspection based on a number of criteria, each containing between 500 and 10,000 incidents and a total of 26,778 reference texts.

Incidents with Impact Collecting incidents and reference texts via Wikidata and Wikipedia tends to result in a bias where English-language texts and incidents related to the Anglophone world are emphasized. To counter-balance this tendency, several specific incidents with a high impact on the Netherlands were chosen to increase the coverage of Dutch texts. The incidents that were added were the 2019 Utrecht Shooting, the MH 17 incident, the Dutch curfew riots during the Covid 19 pandemic and the Eurovision Song Contest (planned for 2020 and held in 2021 in Rotterdam). Due to this approach, the distribution of incidents and reference texts over event types is different in English and in Dutch (further discussed in Section 4).

3.2. Frame Semantic Annotation

The goal of the annotation process was to study the extent of variation of framing in referentially grounded corpora. A subset of the referentially grounded texts was annotated in terms of frame semantic information based on FrameNet frames (where possible). We used an in-house annotation environment that was specifically developed to enable this process and support the data-to-text method (first introduced in (Remijnse et al., 2022)).

The annotation process and tool were designed to take advantage of the data-to-text approach by using the contextual information available through the referential grounding to help the disambiguation process and increase annotation speed. Texts were annotated by event type and, given a set of incidents of that type, by incident and the specific reference texts per language. In addition to the reference text itself, the annotators were also presented with the referential grounding of a text visible in a box next to the text. This box contained information about an incident retrieved from Wikidata, such as place, participants involved in the event, the incident identifier itself, and the event type the incident is subsumed under.

The goal of the annotation process was to link referential information with conceptual information and anchor it in linguistic realizations in the reference texts. In a first step, annotators were asked to link linguistic realizations of the referential grounding to the respective referential information. For example, mentions of a specific event were linked to the event type information displayed in the referential data, mentions of the participants to the participants and other known facts (see Figure 2 for an example). Annotators could edit and add to the referentially grounded data when necessary.

Next, frame semantic annotations were conducted on top of the referential annotations. Rather than focusing on all predicates mentioned in a text (as would be customary in traditional FrameNet-style annotation efforts), the annotators focused on the referential annotations from the previous step. In addition to all mentions of the main incident, the annotators also frame-annotated all subevents of the main incident.

Frame-elements of all event-mentions were annotated within as well as across sentence boundaries. In traditional frame-annotation, frame elements can only be annotated within the same sentence as a predicate. In our approach, we extended the annotation of frame elements to include discourse-level annotations. If a core frame element was not mentioned in the same sentence, but was mentioned in the text, the closest mention of it was annotated (see Figure 1 for an example).

The English part of the corpus was pre-annotated with Open-SESAME (Swayamdipta et al., 2017), a FrameNet-style semantic parser. In later stages of the annotation process, the annotators could make use of these automatic pre-annotations. We filtered them using the notion of *typical frames* introduced in Vossen et al. (2020). Typical frames result from comparing machine-annotated frames of an event-type corpus against frames in other event-type corpora using the principles of TF-IDF comparison (in our implementation called FF-ICF, where CF is the number of corpora in which a frame occurs and FF is the frequency of a frame in a specific corpus).

In Figure 1, we show a screen dump of the annotation environment with some text loaded showing the mentions with the annotation marked. Figure 2 shows the visualization in the environment of the referential and frame information that is associated with a selected mention.

4. Corpus

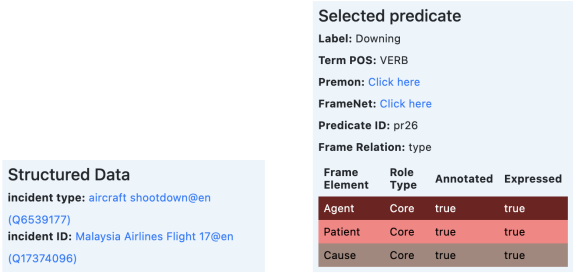
In this section, we present the referentially grounded corpus.

4.1. Overview

Table 1 presents an overview of the data. In total, we collected texts from 38 event types and 22,058 incidents. This collection resulted in 28,642 texts of which 1,312 have been annotated manually. Not all event types and incidents are covered in both languages. We can also observe that the English

en zijn verantwoordelijkheid[pr37] te aanvaarden . Een woordvoerder van de Duitse regering zegt dat Rusland " zijn verantwoordelijkheid[pr38] moet nemen , zodat deze tragedie[pr12] volledig kan worden opgehelderd[pr13] en de daders[pr14] verantwoordelijk gehouden[pr15] kunnen worden " . ' Stop met liegen ' Het Amerikaanse ministerie van Buitenlandse Zaken zegt in een verklaring het volledige vertrouwen te hebben in de resultaten van het Joint Investigation Team van gisteren . Volgens het Amerikaanse ministerie van Buitenlandse Zaken heeft het JIT aangetoond[pr16] dat de Buk - raket vanuit Rusland naar Oekraïne is gebracht[pr17] , vervolgens vanuit een door Rusland gecontroleerd[pr20] gebied is afgevuurd[pr18] en daarna weer is teruggebracht[pr19] naar Russisch territorium . " Het is tijd dat Rusland stopt met liegen en zijn verantwoordelijkheid neemt voor de rol in het neerhalen[pr21] van de vlucht[pr22] " , schrijft het ministerie in een verklaring . De VS beschuldigt Rusland van een " hardvochtige desinformatiecampagne " en roept het land op die te staken . NAVO - baas Jens Stoltenberg zegt ook dat Rusland zijn verantwoordelijkheid[pr39] moet nemen . " Het neerhalen[pr23] van vlucht[pr24] MH17 was een wereldwijde tragedie[pr25] . Degenen die daarvoor verantwoordelijk zijn moeten verantwoordelijk worden gehouden . Ik roep Rusland op om die verantwoordelijkheid[pr40] te nemen . " Rusland ontkent Eerder vandaag stelden Australië en Nederland de Russen aansprakelijk[pr41] voor het neerhalen[pr26] van vlucht[pr27] MH17 op 17 juli 2014 , na de bevindingen[pr28] van het internationale onderzoeksteam van gisteren . Dat concludeerde dat de Buk - raket waarmee de

Figure 1: Annotations in the text of the file "33f08208-e72e-46b6-b074-b7d600868ae4.naf" with the annotations of the frame and frame-elemements associated with the word *neerhalen* (to take down).



(a) Referential information about the incident and event type.

(b) Overview of the frame annotation of a selected predicate (*neerhalen*.

Figure 2: Information displayed to the annotators next to the text shown in Figure 1 about the shoot-down of the flight MH17.

part of the corpus generally has a higher coverage. Even though the corpus has not been designed to be a 'true' parallel corpus, we do cover a subset of event types and incidents in both languages (27 event types and 2,277 incidents). The texts are not translations of each other, but rather reflect the different perspectives on incidents that have been taken in different cultural contexts. The collection process involved retrieving primary and secondary reference texts. We can see that the majority of texts in our corpus are secondary reference texts. The manual annotation effort has been focused on annotating primary reference texts.

4.2. English

In this section, we zoom in on the English subcorpus. Table 2 shows the distribution of incidents and texts over event types.

We can see that there are two extremes in terms of distribution of incidents over event types **event types with a high number of incidents** (e.g. *local election* with 4,066 incidents) and **event types with a low number of incidents** (e.g. *aircraft shoot-down* referring to the MH17 incident, and *natural disaster* referring to the 2021 European Floods). Event types with a high number of incidents are likely to reflect a higher degree of diversity in framing.

Table 2 also shows the distribution of texts over incidents and event types. Again, we can see two extremes of the distribution: **event types with a high number of texts per incidents** as reflected by the ratio (e.g. *aircraft shootdown* with 182 texts covering a single incident) and **event types with a low number of texts per incident** (e.g. several sports events such as *marathon*, *single-day road race*, *golf tournament* and others having, on average, one text per incident).

4.3. Dutch

Table 3 shows the distribution of incidents and texts over event types in the Dutch subcorpus. We can observe similar tendencies as in the English subcorpus; several event types have a high number of different incidents (e.g. *legislative election*, several sports events) and several event types with a low number of different incidents (e.g. *aircraft shoot-down* and *natural disaster*, which refer to the same incidents as in the English corpus). The event types with a high number of texts per incident are similar to the English corpus (*aircraft shootdown*, *disease outbreak*, *natural disaster*).

Overall, we can observe that the data-to-text approach resulted in fewer texts per incident (and event type), which is a reflection of the English-language bias of Wikipedia (and Wikidata). Incidents that were highly relevant in the Netherlands did lead to comparatively many texts, such as the MH17 incident (the flight MH17 had departed from Amsterdam, and the Netherlands played a central role in the subsequent investigation and trial).

4.4. Annotation coverage

In this section, we present the coverage of the manual and automatic frame-semantic annotations. Most of the English texts were annotated automatically with Open-SESAME before manual annota-

	total	en	nl	both
types	38	37	28	27
incidents	22,058	22,002	2,333	2,277
texts	28,642	25,648	2,993	n.a.
human-annotated	1,312	975	337	n.a.
system-annotated	24,670	24,670	0	n.a.
primary reference texts	7,903	6,729	1,173	n.a.
human-annotated	1,312	975	337	n.a.
system-annotated	5,751	5,751	0	n.a.
secondary reference texts	20,739	18,919	1,820	n.a.
human-annotated	0	0	0	n.a.
system-annotated	18,919	18,919	0	n.a.
token-count of annotated texts	4,839,844	4,431,000	408,251	n.a.
human-annotated	699,876	583,683	116,193	n.a.
system-annotated	3,844,778	3,844,778	0	n.a.

Table 1: Overview of annotated texts.

tion (see Section 3.2). Most automatic annotations have not been evaluated but were merely used for inferring the Typical Frames for the event types. A selection of English texts were manually validated, where annotators could confirm the automatic annotation or add another Frame. We kept all automatic and manual annotations/validations with a timestamp and source trace. For Dutch, we did not conduct such a pre-annotation step, as there is yet no reliable frame annotator for Dutch.

As reported in (Remijnse et al., 2022), the Inter Annotator Agreement for the frame annotations (English and Dutch) is very high (from 73.7% up to 97.6% depending on the type of annotation) due to the approach to annotate references to events of the same type or subevents of these events.[3]

Table 4 shows the coverage of the manual annotation in the two subcorpora. The annotation efforts were focused on producing high-quality frame semantic annotations for specific event types (one incident and event type at a time) and on annotating the same event types and incidents in Dutch and English. This approach led to high coverage in Dutch and English for event types such as *music festival*, *mass shooting*, and *aircraft shootdown*. At the same time, this approach meant that there are several event types for which no manual annotations are available.

5. FrameNet Lexicons

From the corpus, we derive all the unique lexical forms that have been annotated. These forms are combinations of lemmatized words and their part-of-speech, extended with the list of distinct frames, frame elements, and references that have been associated with it. These frames and elements represent different concepts ("Sinn" in the Fregian tradition) associated with the form, while the references ("Bedeutung", following Frege) represent the world entities and events that they refer to (Frege et al., 1892). The resulting lexicons form a raw registration of the usage of these words and expressions to frame and refer, which could be the basis for a more theoretically grounded definition of the meaning by lexicographers.

5.1. Corpus derived FrameNet lexicons

Table 5 shows the statistics on the lexicons derived from the Dutch and English texts, where we distinguish between:

1. F-lex: a lexicon of forms that were annotated with a frame as denoting or evoking the frame.
2. FE-lex: a lexicon of forms that were annotated as frame elements of frames.

We can see some analogies and differences between these lexicons and between languages, where it should be noted that the English data have been mainly annotated automatically (94%) and only for a small part manually (6%). The Dutch data were fully annotated manually. This shows in the larger number of annotations in English. The polysemy (average frames per entry) in English is substantially higher than for Dutch as well. More coverage of texts also results in more annotations for a word and hence more frames. Furthermore, Open-SESAME annotations are agnostic for specific types of events, whereas the corpora for the manual annotations have been selected for specific types of events. In all lexicons, nouns dominate. However, in the frame lexicons, verbs are equally

[3]The annotators were well-trained master and PhD students in computational linguistics and linguistics which are native Dutch and English speakers or fluent non-native English speakers

	inc.	texts	ratio
aircraft shootdown	1	182	182.0
disease outbreak	2	189	94.5
natural disaster	1	48	48.0
music festival	15	554	36.93
economic crisis	4	119	29.75
royal wedding	17	320	18.82
mass shooting	88	997	11.33
legal case	40	383	9.57
riot	74	448	6.05
storm	61	260	4.26
auto race	138	181	1.31
unknown	1,180	1,296	1.1
festival	35	38	1.09
presidential election	3,008	3,235	1.08
multi-sport event	317	322	1.02
association football competition	41	42	1.02
local election	4,066	4,093	1.01
legislative election	2,395	2,408	1.01
tennis tournament	3,378	3,410	1.01
championship	550	556	1.01
stage race	86	87	1.01
marathon	39	39	1.0
holiday	2	2	1.0
ceremony	1,139	1,140	1.0
tournament	25	25	1.0
single-day road race	1,574	1,574	1.0
round	52	52	1.0
military operation	2,203	2,210	1.0
contract	7	7	1.0
1.1 (cycling race)	100	100	1.0
motorcycle race	633	634	1.0
rally	161	161	1.0
regatta	51	51	1.0
annual event	21	21	1.0
1.2 (cycling race)	57	57	1.0
golf tournament	201	201	1.0
wrestling event	205	206	1.0

Table 2: Overview of coverage of event types in terms of incidents and texts in English. Several incidents are not subsumed under an event type (*unknown*).

dominant. Remarkable is the large proportion of proper nouns in the English frame element lexicon and, to a lesser extent, in the Dutch lexicon. Obviously, the names of the event participants are often used to refer to them.

The Total Frames column shows the unique frames and frame elements covered in the F-lex and FE-lex resources. Berkeley FrameNet has over 1,200 frames, of which 943 are covered in our English lexicon (F-lex-en) and 643 in our Dutch lexicon (F-lex-nl). Frame elements are more dispersed as there are more than 10K in the FrameNet database of which about 18 to 21% are covered in both lexicons.

	inc.	texts	ratio
aircraft shootdown	1	161	161.0
disease outbreak	2	291	145.5
natural disaster	1	18	18.0
riot	5	58	11.6
music festival	5	54	10.8
mass shooting	10	76	7.6
legal case	1	4	4.0
wrestling event	28	42	1.5
unknown	118	135	1.14
association football competition	5	5	1.0
local election	4	4	1.0
auto race	1	1	1.0
regatta	3	3	1.0
presidential election	5	5	1.0
tournament	6	6	1.0
rally	146	146	1.0
golf tournament	26	26	1.0
multi-sport event	98	98	1.0
championship	72	72	1.0
festival	14	14	1.0
marathon	37	37	1.0
ceremony	96	96	1.0
tennis tournament	898	899	1.0
motorcycle race	557	557	1.0
legislative election	175	175	1.0
contract	5	5	1.0
annual event	6	6	1.0

Table 3: Overview of coverage of event types in terms of incidents and texts in Dutch. Several incidents are not subsumed under an event type (*unknown*).

In Figure 3, we show a fragment of the entry "neerhalen" (downing) with a single frame and a single annotation that corresponds to the mention in the text shown earlier in the annotation environment (see Figure 1). The annotation provides provenance information about the annotation process (project, time, annotation identifier), the way of making reference (*type* indicates the event is an instance of the frame, whereas *evoke* would indicate the frame is associated with a participant), and, finally, information about which part of a specific text received the annotation (a so-called mention) in terms of the text file, the term identifiers, the tokens and the actual sentence.

5.2. Aggregated Dutch FrameNet lexicon

In previous works (Vossen et al., 2018), part of the Dutch SoNaR corpus (Oostdijk et al., 2013) was annotated with FrameNet frames and frame elements. For this, a classical approach was followed that starts from ProbBank annotations (Kingsbury

	en man	sys	nl man
mass shooting	292	705	68
music festival	220	334	45
aircraft shootdown	172	7	148
royal wedding	151	169	0
legal case	122	261	1
storm	12	248	0
presidential election	4	3,231	0
riot	2	446	50
1.2 (cycling race)	0	57	0
annual event	0	21	0
association football competition	0	42	0
regatta	0	51	0
economic crisis 9	0	119	0
contract	0	7	0
holiday	0	2	0
1.1 (cycling race)	0	100	0
disease outbreak	0	189	17
natural disaster	0	48	8
festival	0	38	0
legislative election	0	2,408	0
local election	0	4,093	0
single-day road race	0	1,574	0
tennis tournament	0	3,410	0
championship	0	556	0
ceremony	0	1,140	0
tournament	0	25	0
unknown	0	1,296	0
auto race	0	181	0
round	0	52	0
marathon	0	39	0
military operation	0	2,210	0
motorcycle race	0	634	0
multi-sport event	0	322	0
rally	0	161	0
stage race	0	87	0
golf tournament	0	201	0
wrestling event	0	206	0

Table 4: Overview of the number of annotated texts per event type in English and Dutch.

and Palmer, 2002), assigning the frame to the main predicate and the frame elements to the arguments. From the annotations in the SoNaR corpus, we extracted 1,246 lexical entries (mostly verbs).

In addition, we extracted monosemous lexical units from the Referentie Bestand Nederlands (Van der Vliet, 2007) that were included as synonyms in synsets of the Open Dutch WordNet (Postma et al., 2016). If these synsets had a mapping to synsets from the Princeton WordNet (Fellbaum, 1998), we used SemLink (Stowe et al., 2021) to infer the corresponding frame from FrameNet if any. This resulted in a lexicon with 951 entries (462 nouns, 316 verbs and 173 adjectives).

We merged these two lexicons with the frame and frame element lexicons derived from the


```
"neerhalen:VERB": {
 "lemma": "neerhalen",
 "pos": "VERB",
 "frames":
 {"fn17-downing":
 {"annotations": [
  {"project": "DutchFrameNet",
   "status": "manual",
   "annotator": "fYmWjPYXm-a94lVDkG-Vrl2dU_2LTh8f",
   "timestamp": "2021-10-24T19:55:37UTC",
   "reftype": "type",
   "mention": [
   {"doc": "33f08208-e72e-46b6-b074-b7d600868ae4.naf",
   "term": "t328",
   "tokens": [ {"token_id": "w328","sent": "18",
               "offset": "1975","length": "9"}],
   "text": "Rusland ontkent Eerder vandaag stelden Australië
    en Nederland de Russen aansprakelijk voor het neerhalen
    van vlucht MH17 op 17 juli 2014 , na de bevindingen
    van het internationale onderzoeksteam van gisteren ."}]
  }]}}}
```


Figure 3: Fragment of the lexical entry for "neerhalen" (downing) as a verb with a single annotation for the frame fn17-downing.

Wiki-based corpus to create a combined Dutch FrameNet lexicon. When lexical entries are shared across the lexicon, we check if they also share the associated frames. If so, we add the annotations to the frame. If not, we add a new frame to the entry with the annotations. If entries are not shared, they are added with their frames and annotations. Table 6 shows an overview of the result. In total, almost 7K entries were included, most of which are nouns and verbs. The polysemy (frames per entry) is 3.01 and on average there are 12.41 annotations per entry. In total, 2,311 frames and frame elements are covered. Most annotations originate from the data-to-text corpus described in this paper and were created manually. Some annotations (lexical baseline) come from monosemous words in the Open Dutch Wordnet linked to FrameNet.

5.3. Grounded framing

In addition to FrameNet annotations, the corpus also has reference annotations to entities and events in Wikidata. From these annotations, we can derive a so-called reference lexicon, which is a lexicon that captures the expressions that are used to refer to specific entities and events in the world (Vossen et al., 2018b).

Likewise, we extracted a Dutch and English reference lexicon from the data, see Table 7. References to entities and events are made not only through names but also through various phrases among which nouns, verbs, adjectives, and pronouns (5 different pronouns in the Dutch and 11 pronouns in the English lexicon). However, many phrases are also used: in total 2,708 English and 910 Dutch unique sequences of parts-of-speech are included in the reference lexicons.[4]

Because the corpus is referentially grounded, we can get FrameNet annotations for mentions referring to the same entities. This allows us to compare

[4]The longest sequence being a noun phrase with a conjunction of two verb phrases

Lexicon	Total entries	Nouns	Names	Verbs	Adj	Frames per entry	Annotations per entry	Total Annotations	Total Frames
F-lex-nl	2,662	1,003	48	1,019	212	4.14	18.64	11,986	643
FE-lex-nl	3,588	2,159	195	532	257	2.00	14.19	25,494	1,796
F-lex-en	7,194	2,778	591	2,157	1,170	7.63	973.57	918,079	943
FE-lex-en	6,121	2,490	1,879	856	401	2.85	35.03	75,146	2,145

Table 5: Statistics on the FrameNet lexicons for Dutch (nl) and English (en) derived from the manual and automatic (Open-SESAME) annotations. Separate lexicons are derived from the frame (F) and frame element (FE) annotations. Total entries are given, as well as the main part-of-speech division. Polysemy is represented by the average frames per entry. Further columns give the average number of annotations per entry, the total annotations and the number of distinct frames and frame elements covered.

Dutch FrameNet v1.0	Counts
Total entries	6,964
Nouns	2,883
Names	218
Verbs	2,621
Adjectives	530
Frames per entry	3.01
Annotation per entry	12.41
Total annotations	28,670
Frame coverage	2,311
Manual annotations	27,841
Lexical baseline	829
Data-to-text annotations	22,435
RBN-Wordnet-FrameNet mappings	954
Sonar annotations	5,281

Table 6: Overview of the Dutch FrameNet lexicons combining three data sets: Data2Text corpus, the RBN-WordNet-FrameNet mapping and the SoNaR annotations

the framings across the two languages in texts that are not translations (comparable texts).

We inspected three of the most frequent entities in the data: *Russia*, *Gökmen Tanis* and a *rioter*, framed in both English and Dutch texts. These entities have a high number of overlapping frames: 65, 41, and 31 respectively. For *Russia*, we see that the Dutch mentions tend to frame it more as a suspect of a criminal investigation (in relation to the MH17 shoot down), whereas the English mentions focus on communication, exchange and supply. In the case of *Gökmen Tanis*, we see that Dutch mentions frame him as a terrorist and a perpetrator in a crime, whereas the English mentions focus on the killing and shooting itself (e.g. the bearing_arms is only used in English texts). Finally, the unidentified entity *rioter*, referred to in texts about the curfew riots during COVID, is framed in relation to chaos in both languages, but again the Dutch mentions frame rioters as aggressors, perpetrators, attackers far more often than the English mentions.

These examples show that our data can reveal socio-cultural differences in framing the same entities in the context of the same events. Clearly, the events that were annotated had a big impact in the Dutch context and less in the Anglo-Saxon world. This explains the stronger moral judgments expressed in the framing of the Dutch texts.

6. Discussion and conclusions

The paper reports on the final results of the Dutch FrameNet project. We described the corpora and lexicons that have been created following the data-to-text method and that have been annotated with frames and frame elements at the discourse level. The fact that the data are referentially grounded provides unique possibilities to compare framing across texts and across languages/cultures. It also allows us to train and test frame labeling and coreference resolution systems at the discourse level and across documents in the future.

Several aspects distinguish our lexicon from traditional lexicons and frame-annotated corpora. We start our annotation process from referentially grounded data. Firstly, this means that annotators have much more knowledge of the context in which a text was produced than in traditional annotation scenarios. They can disambiguate potentially minimal, vague, or ambiguous references to the target incident and participants because they know that the texts are written about the same event. As they annotate multiple documents grounded to the same incident, or otherwise, to the same type of incident, the annotators will be more consistent in making similar choices for frames and frame elements, without having to consider all possible options. They will also become more aware of differences in framing across documents.

Secondly, this annotation approach has consequences for the lexicons derived from our annotations. Deriving the lexicons from the annotations means that the evidence for associating a form with frames is included. This provides a powerful starting point for reflective and analytic editing by lexicographers to generalize these annotations. For example, the verb "neerhalen" is annotated with three

Lexicon	Total entries	Nouns	Names	Verbs	Adj.	Referents per entry	Annotations per entry	Total annotations	Total referents
R-lex-nl	3,170	472	88	39	73	11.01	34.21	9,852	288
R-lex-en	8,005	282	743	46	61	12.65	62.68	39,676	633

Table 7: Statistics on the Reference lexicons for Dutch (nl) and English (en) derived from the manual annotations with Wikidata identifiers. Total number of entries are broken down by the main part-of-speech. Referents per entry shows the average polysemy in making reference, whereas annotations per entry shows the average number of references annotated. Total referents are the unique Wikidata identifiers referred to.

different frames: *downing*, *vehicle_landing*, and *domination*. The latter annotation is a metaphorical use in the sense of taking someone down in a power play. The other annotations are closely related and differ in subtle ways. The *downing* frame implies an agent outside the vehicle, whereas *vehicle_landing* does not and can be considered as a more general frame. This would grant a general basic meaning *vehicle_landing* with two derived specializations, one with a specific implicature for the agent and the other metaphorical.

The fact that we take knowledge about the referential grounding into account and that we annotate references to specific incidents and entities (reference annotation) allows us to also derive a reference lexicon next to a frame lexicon. The reference lexicon contains entries that would traditionally not be part of a lexicon that follows a conceptual organization. Nevertheless, such as lexicon can still inform us about the language use in relation to certain types of entities and events.

Third, by inspecting mentions of the same entity in both lexicons, we can investigate which frames are used to express different perspectives on the same entity and event instances. This will inform the developers of FrameNet to be aware of relations between frames and frame elements across different frames.

The combination of referential and conceptual data raises many questions about how to construct and organize a lexicon. In future work, we plan to investigate how to exploit the referential and conceptual information in our corpus and lexicon in such a way that we do justice to the richness of the data while still making the lexicon informative. Obviously, we need to stretch this approach to increase the coverage across event types and different genres of texts.

7. Acknowledgements

The research reported in this article was funded by the Dutch National Science organisation (NWO) through the projects *Framing situations in the Dutch language, VC.GW17.083/6215.* and *Understanding-languages-by-Machines, Vossen.Spinoza*. We thank the reviewers for their feedback and the students for annotating the texts.

8. Limitations

The current data are limited to events that are registered in Wikipedia and for which at least secondary reference texts were found. To find lesser-known events, the code can also be used to find entities (people, places) in Wikidata that participate in certain event types, such as diseases, conflicts, creative works, marriages, etc. By searching the Wikipedia pages of these entities for the mentions of these events and possible primary reference texts, we can reconstruct the event data in which the entity participated. In future work, we will extend our data with such less-common events as well.

The current data was built for English and Dutch texts only. However, the method can also be used for other languages. For annotation, the manual environment can be used or cross-lingual FrameNet labellers need to be developed. Some of the data has been annotated manually but many English texts grounded to the events were automatically annotated. The automatic annotations by OpenSESAME were used to learn the typical frames associated with events of certain types. We have not assessed the quality of these annotations but we did spot errors due to biases, e.g. "court" for tennis events receiving legal frames. Future work could investigate the impact of these errors and biases.

9. Ethics

Obviously, framing in data reflects personal and socio-cultural biases towards entities and events. These biases can be harmful to people. Such harm, as a reflection of how people framed things, should be considered when consulting the data (corpus and lexicon).

10. Bibliographical References

Alon Eirew, Avi Caciularu, and Ido Dagan. 2022. Cross-document event coreference search: Task, dataset and modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 900–913.

Alon Eirew, Arie Cattan, and Ido Dagan. 2021. Wec: Deriving a large-scale cross-document event coreference dataset from wikipedia. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2498–2510.

Gottlob Frege et al. 1892. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100(1):25–50.

Levi Remijnse, Marten Postma, and Piek Vossen. 2021. Variation in framing as a function of temporal reporting distance. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 228–238.

Levi Remijnse, Piek Vossen, Antske Fokkens, and Sam Titarsolej. 2022. Introducing frege to fillmore: A framenet dataset that captures both sense and reference. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 39–50.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *arXiv preprint arXiv:1706.09528*.

Tiago Timponi Torrent, Collin F Baker, Oliver Czulo, Kyoko Ohara, and Miriam RL Petruck. 2020. Proceedings of the international framenet workshop 2020: Towards a global, multilingual framenet. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*.

Piek Vossen, Filip Ilievski, Marten Postma, Antske Fokkens, Gosse Minnema, and Levi Remijnse. 2020. Large-scale cross-lingual language resources for referencing and framing. In *12th Language Resources and Evaluation Conference: LREC 2020*, pages 3162–3171. European Language Resources Association (ELRA).

Piek Vossen, Filip Ilievski, Marten Postma, and Roxane Segers. 2018a. Don't 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)*.

Piek Vossen, Filip Ilievski, and Marten Postrma. 2018b. Referencenet: a semantic-pragmatic network for capturing reference relations. In *Proceedings of the 9th Global Wordnet Conference*, pages 219–228.

11. Language Resource References

Baker, Collin F and Fillmore, Charles J and Lowe, John B. 1998. *The berkeley framenet project*.

Fellbaum, Christiane. 1998. *WordNet: An electronic lexical database*. MIT press.

Kingsbury, Paul R and Palmer, Martha. 2002. *From TreeBank to PropBank.*

Oostdijk, Nelleke and Reynaert, Martin and Hoste, Véronique and Schuurman, Ineke. 2013. *SoNaR user documentation*.

Postma, Marten and Van Miltenburg, Emiel and Segers, Roxane and Schoen, Anneleen and Vossen, Piek. 2016. *Open dutch wordnet*.

Stowe, Kevin and Preciado, Jenette and Conger, Kathryn and Brown, Susan Windisch and Kazeminejad, Ghazaleh and Gung, James and Palmer, Martha. 2021. *SemLink 2.0: Chasing lexical resources*.

Van der Vliet, Hennie. 2007. *The Referentiebestand Nederlands as a multi-purpose lexical database*. Oxford University Press.

Vossen, Piek and Fokkens, Antske and Maks, Isa and van Son, Chantal. 2018. *Towards an open Dutch Framenet lexicon and corpus*.

12. Appendix

12.1. Lexical entries

Below we show some fragments for the frame lexicon (Figure 4) with a few associated frames and an annotation and another fragment for the reference lexicon (Figure 5) with associated Wikidata identifiers and an annotation in the DutchFrameNet corpus.


```
{
    "ramp:NOUN": {
        "lemma": "ramp",
        "pos": "NOUN",
        "frames": {
            "fn17-catastrophe": { }
        }
    },
    "betrokkenheid:NOUN": {
        "lemma": "betrokkenheid",
        "pos": "NOUN",
        "frames": {
            "fn17-participation": { }
        }
    },
    "onderzoek:NOUN": {
        "lemma": "onderzoek",
        "pos": "NOUN",
        "frames": {
            "fn17-criminal_investigation": { },
            "fn17-inspecting": { },
            "fn17-scrutiny": { },
            "fn17-research": { },
            "fn17-controller_object": { },
            "fn17-abandonment": { }
        }
    },
    "denken:VERB": { },
    "handelen:VERB": { },
    "meewerken:VERB": { },
    "neerhalen:VERB": {
        "lemma": "neerhalen",
        "pos": "VERB",
        "frames": {
            "fn17-downing": {
            "annotations": [
                    {
                        "project": "DutchFrameNet",
                        "status": "manual",
                        "annotator": "fYmWjPYXm-a941VDkG-Vr12dU_2LTh8f",
                        "timestamp": "2021-10-24T19:34:28UTC",
                        "reftype": "type",
                        "mention": [
                            {
                                "doc": "33f08208-e72e-46b6-b074-b7d600868ae4.naf",
                                "term": "t74",
                                "tokens": [
                                    {
                                        "token_id": "w74",
                                        "sent": "4",
                                        "offset": "400",
                                        "length": "9"
                                    }
                                ],
                                "text": "Het neerhalen van vlucht"
                            }
                        ]
                    },]
            },
            "fn17-domination": { },
            "fn17-vehicle_landing": { }
        }
    }
}
```


Figure 4: Fragment of the Dutch frame lexicon derived from the FrameNet corpus, showing several entries with associated FrameNet associations and for "neerhalen" (downing) an annotation in a text.


```
{
    "België:NOUN": {
        "lemma": "België",
        "pos": "NOUN",
        "references": {
            "http://www.wikidata.org/entity/Q31": { },
            "http://www.wikidata.org/entity/Q46": { },
            "http://www.wikidata.org/entity/Q16493497836999": { }
        }
    },
    "de_ramp:DET_NOUN": { },
    "vlucht_mh17:NOUN_NOUN": {
        "lemma": "vlucht_mh17",
        "pos": "NOUN_NOUN",
        "references": {
            "http://www.wikidata.org/entity/Q16318642716l2": {
                "annotations": [
                    {
                        "project": "DutchFrameNet",
                        "status": "manual",
                        "annotator": "undefined",
                        "timestamp": "2021-10-24T19:20:40UTC",
                        "reftype": "entity",
                        "mention": [
                            {
                                "doc": "33f08208-e72e-46b6-b074-b7d600868ae4.naf",
                                "term": "t15t16",
                                "tokens": [
                                    {
                                        "token_id": "w15",
                                        "sent": "1",
                                        "offset": "102",
                                        "length": "6"
                                    },
                                    {
                                        "token_id": "w16",
                                        "sent": "1",
                                        "offset": "109",
                                        "length": "4"
                                    }
                                ],
                                "text": "Verschillende landen willen dat Rusland verantwoording aflegt
                                over zijn betrokkenheid bij de ramp met vlucht MH17 ."
                            }
                        ]
                    },
                    { },
                    { },
```


Figure 5: Fragment of the Dutch reference lexicon derived from the FrameNet corpus, showing the entries for "België" and "vlucht_mh17" with Wikidata references and the annotation of a mention of the latter in a text.

12.2. Referred entities

Table 8 shows the most-frequently referred entities in the Dutch and English data, represented by Q-numbers. Note that new Q-numbers were created when the annotators could not find a proper referent in Wikidata (Lebel=None). The table is sorted by the number of overlapping frames across the Dutch and English mentions. In addition we see the most frequent mention (Dutch or English) and the total Frames and Frame elements used in each language.

Referent	Label	Most Frequent Mention	NL Frames	EN Frames	Overlap
Q159	Russia	Russia	105	202	65
Q17374096	Malaysia Airlines Flight 17	crash:NOUN	77	120	53
Q62116513	Gökmen Tanis	t. (Tanis)	113	88	41
Q1636580448374	None	Ukraine	60	124	31
Q1631864085791	None	rebel	38	201	31
Q1651075371225	None	organiser	47	116	25
Q1631025548534	None	slachtoffer (victim)	69	104	23
Q1631864271612	None	MH17	34	78	22
Q1646835057489	None	audience	35	135	21
Q29999	Kingdom of the Netherlands	nederland:NOUN	59	74	20
Q30973589	Eurovision Song Contest 2020	contest	32	94	19
Q15821620	Joint investigation team	jit	51	43	19
Q1673010005817	None	nabestaande (surviving relative)	59	54	17
Q1638962685452	None	artist	26	135	14
Q1689167507986	None	politie (police)	46	47	12
Q67197819	Volodymyr Tsemakh	tsemach	49	22	11
Q1632409367599	None	slachtoffer (victim)	33	30	10
Q1646212736373	None	rioter	74	20	10

Table 8: Most frequently referenced entities with their Wikidata label, if present, and the most frequent referring expression (English or Dutch). Columns show the number of unique Frames associated with the Dutch and English mentions and the number of overlapping Frames. If the Label is None, the entity is not in Wikidata but created by the annotator on the spot with a non-existing random Q-number

Gökmen Tanis	NL & EN	NL	EN
fn17-offenses@perpetrator	39	19	20
fn17-suspicion@suspect	33	11	22
fn17-hit_target@agent	29	17	12
fn17-attack@assailant	25	13	12
fn17-killing@killer	22	8	14
fn17-terrorism@terrorist	22	15	7
fn17-suspicion	21	7	13
fn17-trial@defendant	19	14	5
fn17-statement@speaker	11	9	2
fn17-committing_crime@perpetrator	12	8	4
fn17-use_firearm@agent	15	3	12
fn17-arrest@suspect	13	2	11
fn17-detaining@suspect	15	4	11
fn17-bearing_arms	0	0	11

Table 9: Most frequent frames for the entity "Gökmen Tani" (not in Wikidata). First columns gives the combined frequency English-Dutch, followed by the separate counts

rioter	NL & EN	NL	EN
fn17-chaos@entity	19	11	8
fn17-chaos@state	5	1	4
fn17-chaos	5	1	4
fn17-committing_crime@perpetrator	4	3	1
fn17-people	4	3	1
fn17-robbery@perpetrator	2	1	1
fn17-emotion_directed@experiencer	2	1	1
fn17-verdict@defendant	2	1	1
fn17-setting_fire@kindler	2	1	1
fn17-hostile_encounter@side_2	2	1	1
fn17-protest@protester	0	9	0
fn17-attack@assailant	0	8	0
fn17-causation@actor	0	6	0
fn17-arrest@suspect	0	5	0
fn17-cause_harm@agent	0	5	0
fn17-damaging@agent	0	4	0
fn17-destroying@destroyer	0	4	0
fn17-evading@evader	2	0	2
fn17-resolve_problem@problem	2	0	2

Table 10: Most frequent frames for the entity "rioter" (not in Wikidata). First columns gives the combined frequency English-Dutch, followed by the separate counts

Q159 (Russia)	NL&EN	NL	EN
fn17-statement@speaker:	55	13	42
fn17-responsibility@agent:	48	36	12
fn17-affirm_or_deny@speaker:	47	12	35
fn17-criminal_investigation@suspect:	37	23	14
fn17-supply@supplier:	34	3	31
fn17-exchange@exchanger_1:	22	5	17

Table 11: Most frequent frames for the entity "Russia". First columns gives the combined frequency English-Dutch, followed by the separate counts.