Towards an Open Dutch FrameNet lexicon and corpus

Piek Vossen, Antske Fokkens, Isa Maks, Chantal van Son
Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam
The Netherlands
{piek.vossen, antske.fokkens, isa.maks, c.m.van.son}@vu.nl

Abstract

This paper reports on the progress of the development of an Open Dutch FrameNet lexicon and annotated corpus. We started the project in 2017 with the annotation of a Dutch corpus of written Dutch that was previously annotated with PropBank predicates and roles. The corpus represents a diverse set of written Dutch texts. We discuss the annotation results and process. From this corpus, we have derived an initial Dutch lexicon with FrameNet frames. In the meanwhile, we designed a method to collect texts that exhibit a large degree of variation in framing similar events. We will apply this method in the future to extend the representative corpus vertically for certain types of events to obtain more insight into variation of framing.

Keywords: Dutch, frame semantics, corpus annotation

1. Introduction

Languages are rich instruments for framing situations or events in various ways. A report on a football game, for instance, can be written from the perspective of the winner, the loser, or a neutral observer; a financial transaction can be reported from the buyer or the seller; a medical case can be framed from the perspective of the patient or the doctor. We use different words and expressions in language to frame similar situations differently depending on our interest, our motivation, and audience. The perspective on a situation that is associated with the choice of words is what we call linguistic framing. It reflects what we see as important and what as background, it expresses emotions and judgments, and it suggests motivations and expectations. A concrete case in point is work by Cybulska and Vossen (2010), who demonstrate how the *Fall of Srebrenica* is framed differently depending on the time passed between the event taking place and the moment of reporting. As historic distance increases, less detail (e.g. abstracting from the precise time, location and participants) but more explanations, motivations and judgments (deportation, genocide) were given. Fokkens et al. (2018) investigate how stereotypes and created images are reflected in textual micro-portraits (framings of individuals in stories) and show, for instance, that Dutch newspapers mostly specifically label people as "Dutch" when they win in sports.

Clearly, language is a powerful instrument to shape our view of the world, and it is therefore important to get a good understanding of how framing works. Yet, little is known about framing in Dutch. What are the Dutch words and expressions used to frame the same situations or events in different ways? How does Dutch framing differ from other languages? How much variation exists and what are the underlying semantic and pragmatic factors for using these variants in contexts?

This paper reports on the initial development of the Open Dutch FrameNet similar to multilingual FrameNets described in (Baker, 2008). We started the development of a Dutch FrameNet in 2017 with the annotation of a corpus of written Dutch that was previously annotated with PropBank predicates and roles (Kingsbury and Palmer, 2002); see Sections 2. and 3.. From this corpus, we derived an initial FrameNet lexicon (Section 4.). For future work (Section 5.), we will use a method to collect texts that exhibit a large degree of variation in framing similar events.

2. Overall Approach

Our first objective is to capture the usage of FrameNet frames and elements in a representative Dutch corpus and to derive a Dutch FrameNet lexicon from this corpus. We therefore took the following design decisions:

- We use a balanced corpus with diverse genres;
- We apply an all-sentences-approach, which means:
  - we take the sentences of a document as given
  - we do not apply any preselection of lexical units nor a preselection of example sentences;
  - we also do not preselect frames or frame elements;
  - but for each sentence a preselection of the main predicate and the arguments is already given;
- Frame identification should fit the usage of the predicate in the sentence;
- Roles are assigned after the sentence-frame is selected with the corresponding roles.

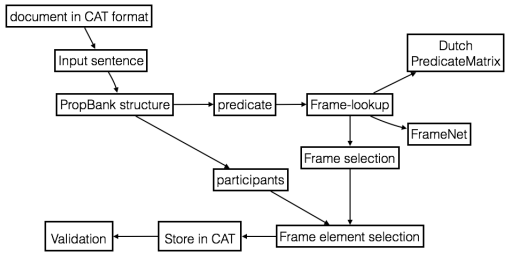

Figure 1: Overview of the annotation process of the SoNaR documents with PropBank annotation in the CAT format.

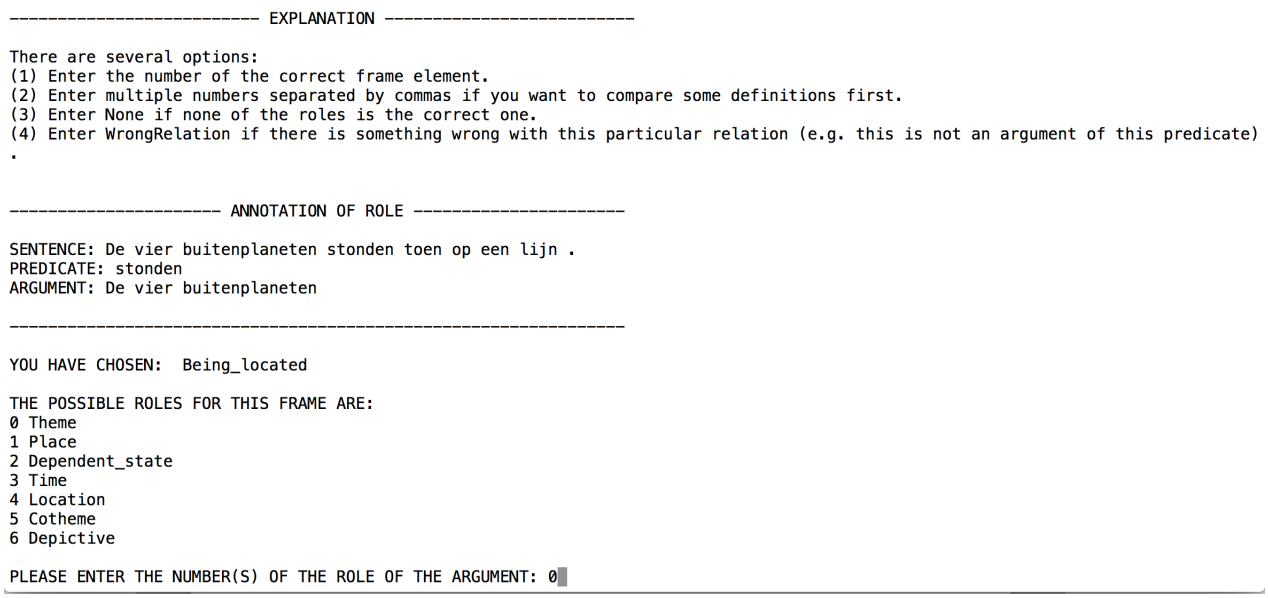

Figure 2: Screenshot of the annotation interface showing instructions, the target sentence and the target predicate and an argument according to the PropBank structure for which a frame element needs to be selected, given the frame Being_located that was assigned to the predicate *stonden* (stood).

We used SoNaR as a corpus, which is a corpus of written Dutch (Oostdijk et al., 2008). Part of this corpus was already annotated with PropBank relations (De Clercq et al., 2012). Figure 1 shows the further process starting with documents from SoNaR in the format of the CAT annotation tool (Lenzi et al., 2012). Our annotators first add FrameNet annotations to these previously annotated PropBank predicates (verbs) and their arguments. Because the annotators proceed sentence-by-sentence through a highly varied set of texts, they have to consider all frames from the English FrameNet version 1.7. We therefore developed a specific annotation tool[1] to support the annotators, which loads the annotated PropBank relations one by one and presents the annotator with the sentence, the predicate and the arguments. The annotation task consists of two steps: (1) frame annotation, and (2) frame element annotation. For the first step, the tool supports searching for frames in FrameNet by entering the predicate and/or equivalents in both Dutch and English. Equivalents are generated using the PredicateMatrix (derived from SemLink (De Lacalle et al., 2014)), which provides mappings between English and Dutch lexical units through the Open Dutch WordNet (Postma et al., 2016). After entering the predicate and/or equivalents, the annotator is then presented with the definitions of all associated frames and selects the most fitting one (if any). More experienced annotators can also directly enter the name of the frame. Once a frame is selected for the predicate, the tool iterates over the arguments to select the frame elements. Figure 2 shows a screenshot of the frame element annotation after the frame Being_located has been selected for the sentence in Example 1 from the Dutch Wikipedia article on the solar system.

(1) *De vier buitenplaneten* ***stonden*** *toen op een lijn*
The four outer planets **stood** then in one line.
"The four outer planets were aligned in those days."

Texts annotated by two annotators are processed to mark mismatches and disagreement. We distinguish between mismatches between frames that stand in a super-subtype relation in FrameNet and other mismatches. Texts with marked agreement and disagreement are visualised for analysis and adjudication using the CAT tool.

3. Frame Corpus

Four students worked for four months, eight hours a week. All texts have been double annotated. In total, 3,898 verb tokens have been annotated with 679 frames. Table 1 shows the statistics for the annotated corpus, showing the distribution of texts and the number of annotated predicates for each genre. The most represented genres are financial, periodicals and wikipedia.

theme/genre	nr_of_files	nr_of_annotated_verbs
background-news	3	110
financial	17	1756
medical	1	88
news	5	499
newsletter	3	111
periodicals	37	821
policy	12	352
teletext	3	169
websites	1	49
wiki	34	1295
totals	116	5250

Table 1: Corpus statistics on the different genres and the number of files in the SoNaR corpus that have PropBank annotations with the total number of annotated predicates in each genre.

[1] https://github.com/cltl/FrameNet-annotation-tool

We measure the inter-annotator agreement (see Table 2) counting exact matches (47%, Kappa 0.46) and lenient matches. In the case of lenient matches, we consider frames to be matches if they are closely related by one of FrameNet's frame-to-frame relations such as Inheritance (lenient agreement-I) or any relation (lenient agreement-II). Inter-annotator agreement increase with 3% and 7% respectively when lenient matching is applied. Agreement in annotating frame elements given agreement on the frame was much higher (79%). Frame agreement is lower than agreement scores reported by, for example, Søgaard et al. (2015) and Benešová et al. (2008), who respectively report scores of 85% (frames) and 78% (frame elements) on English Twitter data, and 69% and 85% on Czech lexical units for communication verbs. However, in these studies, the annotation tasks were much more restricted in the types and/or number of frames to be considered. Following the procedure explained in the previous section, our annotators need to proceed sentence-by-sentence, considering very different predicates and all types of frames and all possible relations.

Type of agreement	Percentage
strict agreement	0.47
lenient agreement -I: only inheritance relations	0.51
lenient agreement -II: all relations	0.54
agreement on frame elements (with matching frames)	0.79

Table 2: Inter-annotator agreement statistics on frames and frame elements.

The annotators struggle both with consistently selecting frames from the large set available in FrameNet and with coverage problems of FrameNet (in which case the frame "None" is assigned). In Table 4, we show the most frequently confused frames. As was also found by Padó (2007, p. 63), some of these disagreements are due to subtle or difficult distinctions between frames in meaning that may not be clear from the context. Therefore, we further analyzed the disagreements by determining the distance between the confused frames in the frame hierarchy (taking all relations into account) and the type of relations between them. We found that in 20% (552 instances) of all disagreements, the frames were directly related through one of the ten frame relation types in FrameNet (frame-frame distance of 1). The distribution of the relation types in these cases is shown in Table 3. For example, there is an Inheritance relation between many of the most frequent frame confusion pairs, e.g. {Activity_start, Process_start}, {Creating, Intentionally_create}. Other frequent cases include those frames standing in a Using relation; for example, the frame Communication is used in many other frames, such as Statement and Expressing_publicly. The ReFraming_Mapping relation between two frames indicates that lexical units were moved into a new frame (Petruck et al., 2004), as is the case for the pair {Attempt_suasion, Request}. In many of these cases, one frame may be more specific than the other, but both are likely to fit the lexical unit found in the text. For example, both Creating and Intentionally_create are technically correct for the lexical unit *maken* in Sentence 2, even though Intentionally_create would be more specific.

(2) *maar wij moeten het beter doen en minder van*
but we must it better do and less of
die regels ***maken***
those rules **make**
"but we have to do better and make less of those rules."

Frame relation type	Percentage
Inheritance	0.40
Using	0.21
ReFraming_Mapping	0.14
Causative_of	0.12
See_also	0.09
Inchoative_of	0.01
Perspective_on	0.01
Precedes	0.01
Subframe	0.01
Metaphor	0.0

Table 3: Distribution of types of relations between confused frames with a frame-frame distance of 1.

Other confusions, however, seem to involve frames with different core elements and restrictions on these core elements (such as +CONTROL or -CONTROL) which are not likely to be both correct for one context, as with the pair {Operate_vehicle, Self_motion}. However, even these distinctions are not always clear. For example, the correct frame in Sentence 3 for *gereden* seems to be Operate_vehicle, whereas Self_motion seems less correct. However, the definition of Self_motion does mention that "many of the lexical units in this frame can also describe the motion of vehicles (e.g., as external arguments) [and are treated] as belonging in this frame."

(3) *Doorgaans wordt vanwege de risico's in konvooi* ***gereden***
Usually being because the risks in convoy **driven**
'Usually, vehicles are **driven** in convoy because of the risks."

The other frame confusion pairs had a frame-frame distance of two (15%), three (17%), more (42%), or were not related at all (7%). Even though frame confusions were never counted as correct in our agreement scores if their frame-frame distance is larger than one, some of them are still understandable. For example, the frames Daring and Attempt are not directly related to each other, but both inherit from Intentionally_act, which makes them sister frames (distance=2). We also encountered 'grandparent' relations, such as {Finish_competition, Activity_finish} linked through Finish_game (distance=2). Frame pairs with larger distances are more likely to exhibit significant semantic differences, as with {Path_shape, Sign_agreement} (distance=5), but not necessarily, as with {Opinion, Regard} (distance=5).

In Table 5, we show agreement and disagreement for the most frequent frames. We can see that the (dis)agreement varies considerably across frames: e.g. Desiring (69), Attempt_suasion (65) and Statement (64) as highest scoring and Circumscribed_existence (6), Intentionally_create (7) as lowest scoring. High agreements could be due to frequency of certain predicates with clear meaning and little ambiguity. Low agreements seem idiosyncratic.

Our annotations are open source and freely downloadable as well as some of the original texts.[2] Part of the original

[2] https://github.com/cltl/Open-Dutch-Framenet

19	Activity_start	Process_start
14	Creating	Intentionally_create
14	Cause_change_of_position_on_a_scale	Change_position_on_a_scale
12	Using	Using_resource
12	Opinion	Regard
10	Cooking_creation	Manufacturing
8	Getting	Receiving
8	Expressing_publicly	Statement
8	Existence	Presence
8	Awareness	Grasp
7	Operate_vehicle	Self_motion
7	Finish_competition	Finish_game
7	Causation	Evidence
7	Being_named	Name_conferral
6	Perception_active	Perception_experience
6	Intentionally_create	Text_creation
6	Giving	Grant_permission
6	Cure	Medical_intervention
6	Cause_to_perceive	Expressing_publicly
6	Beat_opponent	Finish_competition
6	Awareness	Certainty
6	Accomplishment	Getting
5	Reference_text	WrongRelation
5	Preventing	Thwarting
5	Perception_active	Reference_text
5	Have_associated	Possession
5	Finish_competition	Success_or_failure
5	Competition	Finish_competition
5	Communication	Statement
5	Communication	Expressing_publicly

Table 4: Frame confusion pairs across annotators sorted by frequency.

frame	agreements	disagreements	percentage agreement
Desiring	25	11	0.69
Attempt_suasion	33	18	0.65
Statement	108	60	0.64
Request	19	19	0.5
Removing	16	18	0.47
Causation	38	45	0.46
Receiving	22	31	0.42
Self_motion	17	23	0.42
Change_position_on_a_scale	27	39	0.41
Perception_active	17	24	0.41
Activity_start	23	35	0.4
Event	30	46	0.39
Coming_to_be	30	46	0.39
Being_located	22	35	0.39
Using	17	28	0.38
Cause_change	13	21	0.38
Possession	38	64	0.37
Intentionally_act	24	49	0.33
Opinion	18	38	0.32
Participation	12	27	0.31
Existence	26	62	0.3
Evidence	14	34	0.29
Becoming_aware	12	33	0.27
Inclusion	12	39	0.24
Process_start	8	26	0.24
Arriving	8	26	0.24
Cause_to_perceive	11	36	0.23
Awareness	14	50	0.22
Accomplishment	6	27	0.18
Giving	7	35	0.17
Communication	6	30	0.17
Finish_competition	5	35	0.12
Intentionally_create	4	52	0.07
Circumscribed_existence	2	32	0.06
None	2	40	0.05

Table 5: Most frequently assigned frames with the agreements and disagreements.

texts must however be obtained through a license (freely available for research): "SoNaR-klein-commercieel" enriched with PropBank annotations.[3]

4. Initial frame lexicon

We can derive an initial FrameNet lexicon for Dutch from the annotations made so far. In total more than 1,336 predicate types or lexical entries have been annotated. We list all the different frames that have been assigned to these predicates with their frequency. If we consider each lemma-frame pair as a lexical unit, we would get 4,755 different lexical units distributed across 671 frames. Figure 3 shows a few examples of this derived lexicon. We see that the annotator assigned six different frames to the polysemous Dutch word *afsluiten* (close, settle, end). Some of these frames are closely related to each other representing three of the main meanings of the word: the meaning *close a building or door* is represented by the frames Locale_closure and Change_activity, the meaning *settle an agreement* is represented by Make_agreement_on_action and Sign_agreement and the meaning *finish a process* by Activity_finish and Process_end. The example shows that in this way not only coarse-grained senses, but also more fine-grained nuances of word senses are captured.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<fnLexicon lang="nl">

<ENTRY lemma="inschakelen" pos="v">  /* switch on */
 <frameAnnotation frame="Installing" annotations="1"/>
 <frameAnnotation frame="Process_start" annotations="1"/>
</ENTRY>
<ENTRY lemma="mankeren" pos="v">  /* be inadequate */
 <frameAnnotation frame="Medical_conditions" annotations="2"/>
 <frameAnnotation frame="Undergoing" annotations="2"/>
</ENTRY>
<ENTRY lemma="baseren" pos="v">  /* base on */
 <frameAnnotation frame="Evidence" annotations="5"/>
 <frameAnnotation frame="Justifying" annotations="1"/>
 <frameAnnotation frame="None" annotations="1"/>
 <frameAnnotation frame="Reliance" annotations="3"/>
</ENTRY>
<ENTRY lemma="afsluiten" pos="v">  /* close, settle, end */
 <frameAnnotation frame="Make_agreement_on_action" annotations="1"/>
 <frameAnnotation frame="Sign_agreement" annotations="1"/>
 <frameAnnotation frame="Locale_closure" annotations="2"/>
 <frameAnnotation frame="Change_accessibility" annotations="1"/>
 <frameAnnotation frame="Activity_finish="11"/>
 <frameAnnotation frame="Process_end" annotations="3"/>
</ENTRY>
```

Figure 3: Example of a lexical entry in the Dutch FrameNet lexicon derived from the corpus annotations.

Using the annotations as input for creating a lexicon has an additional advantage. We will explore whether we can group certain annotations and frames and eliminate errors. By addressing the annotations from a lexical point of view, we can critically assess the annotations.

5. Future Plans

The annotation carried out so far follows a traditional **text-to-data** method, where linguists first collect texts and then annotate it with interpretations, e.g. frames. The process is labor-intensive and the IAA is low as explained above. The annotators have to consider a highly diverse set of texts on very different topics. Since they have to annotate every predicate from the PropBank annotation, sentence-by-sentence, they also have to consider all the FrameNet frames and elements continuously.

In future work, we therefore continue with a **data-to-text** approach, described in more detail in Vossen et al. (2018b). This approach starts from a-priori registrations of events in structured data and provides so-called reference texts that report on these specific events. Starting from structured data that defines what the event is, but also who is involved, when and where it took place, the data-to-text approach guarantees a large variety of texts on similar situations and events from various perspectives. Annotators will consider sets of documents that involve more or less the same frames and elements simultaneously in relation to the same or very similar events.

This data-to-text method has several advantages over a classical text-to-data annotation method: 1) we already have predefined a formal representations of events or incidents,

[3] http://tst-centrale.org/nl/tst-materialen/corpora/sonar-klein-corpus-commercieel-detail

often with information on the time, location and participants without having to rely on error-prone automatic processing of text or labor-intensive manual annotation, 2) we obtain a large variety of texts from different sources, genres and languages that make reference to the same events, likely in very different ways, 3) we do not need to interpret everything that is written in the text but can focus on the text parts that relate to the structured data, 4) we can compare many different pairings of structured data and reporting texts for the same type of events and therefore generalize our observations to the level of frame types, 5) annotators can focus on similar events that share frames and frame elements for many texts, 6) annotators can focus more on the variation in framing of similar events.

As explained in (Vossen et al., 2018a; Vossen et al., 2018c), we used this method to annotate 510 documents for event coreference for the SemEval2018-Task5 *Counting Events and Participants in the Long Tail* (Postma et al., 2018). All the documents report on manually registered gun violence incidents and have been annotated given the structured data on the incident a priori.[4] Annotators mark in the text any reference to the incident as a whole and specific subevents. Table 6 lists the most used expressions for the different event types represented by frames. The table shows a wide range of closely related predicates. Note that some of the references to frames can be very indirect, e.g. *surgery* implies *Experience_bodily_harm* and *funeral* implies *Death*. By starting from similar incidents, we not only expect to cover a wider range of predicate and frames but also provide input for possible frame relations that can be added to FrameNet.

Frame	Most common expressions
Death	dead (305) died (285) killed (283)
Use_Firearm	shooting (680) gunshot (247) went off (72)
Hit_Or_miss	shot (801) shooting (83) struck (46) missed (1)
Incident	accident (57) shooting (260) incident (164) tragedy (11) it (88)
Experience_bodily_harm	wound (175) injured (75) injuries (68) surgery (1)

Table 6: Most common expressions used for frames in the Gun Violence corpus

By complementing the current balanced corpus through this vertical extensions by the data-to-text method, we hope to obtain a good mixture of a corpus that on the one hand strives for representing the diversity of language genres and topics and on the other hand for variation in framing similar events across texts. The data-to-text method is different from FrameNet annotation approaches that start from a specific frame and try to find sentences with related lexical units. The event registries do not come with a selection of frames or lexical units and we expect that the annotation of the related texts may involve a substantial variety of related frames. Obviously, only a restricted range of events are covered by the event registries. As such, we consider this method as complementary to other approaches and hope to learn from the differences in variation across these annotations.

[4]https://github.com/cltl/GunViolenceCorpus

6. Conclusion

In this paper, we described the first steps towards an Open Dutch FrameNet lexicon and annotated corpus. The first contribution of this paper is the description of the current status of the annotation process and lexicon. These annotations consisted of adding FrameNet frames and element annotation to a component of the Dutch SoNaR corpus that was already annotated with PropBank predicates and roles. The corpus in question contains a diverse set of written Dutch texts.

A total of 3,898 verbs covering 1,336 predicate types have been annotated with frames and their arguments with frame elements. Due to the high variety of data and lexical types that had to be considered, inter-annotator agreement was lower than in other studies where annotators focused on more selective data. Agreement was 47% for exact match, 51% when counting frames standing in a heritage relation as correct and 54% when accepting frames standing in any relation. Problems were mainly found in the lack of coverage of FrameNet and in mismatches between frames whose distinction is subtle as also observed by Padó (2007). Overall, a lexicon based 4,755 pairings between lexical units and frames could be derived from our data, covering 671 frames.

The second contribution of the paper is that it proposes to use a new method, the data-to-text method (Vossen et al., 2018b) for creating annotated data with a high variation in framing similar events. We plan to apply this method in future work.

7. Acknowledgements

The work presented in this paper was funded by the Netherlands Organisation for Scientific Research (NWO) via the Spinoza grant, awarded to Piek Vossen in the project "Understanding Language by Machines" and VENI grant 275-89-029 awarded to Antske Fokkens.

8. Bibliographical References

Benešová, V., Lopatková, M., and Hrstková, K. (2008). Enhancing czech valency lexicon with semantic information from framenet: The case of communication verbs. *Proceedings of the 1st International Conference on Global Interoperability for Language Resources*, pages 18–25.

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 *LREC*.

Fokkens, A., Ruigrok, N., Gagenstein, S., van Atteveldt, W., and Beukeboom, C. (2018). Microportrait detection for identifying stereotypes in dutch news. In *LREC-2018, Myazaki, Japan*.

Lenzi, V. B., Moretti, G., and Sprugnoli, R. (2012). Cat: the celct annotation tool. In *LREC*, pages 333–338.

Padó, S. (2007). *Cross-lingual annotation projection models for role-semantic information*. Ph.D. thesis, Saarland University.

Petruck, M. R., Fillmore, C. J., Baker, C. F., Ellsworth, M., and Ruppenhofer, J. (2004). Reframing framenet

data. In *Proceedings of The 11th EURALEX International Congress*, pages 405–416.

Postma, M., Ilievski, F., and Vossen, P. (2018). Semeval-2018 task 5: Counting events and participants in the long tail.

Søgaard, A., Plank, B., and Martinez Alonso, H. (2015). Using frame semantics for knowledge extraction from twitter. *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2447–2452.

Vossen, P., Postma, M., and Ilievski, F. (2018a). From data to text: Capturing long tail events through microworlds and reference texts. In *LREC-2018, Myazaki, Japan*.

Vossen, P., Ilievski, F., Postma, M., and Roxane, S. (2018b). Don't annotate, but validate: a data-to-text method for capturing event data. In *LREC2018, Myazaki*.

Vossen, P., Postma, M., and Ilievski, F. (2018c). Referencenet: a semantic-pragmatic network for capturing reference relations. In *Global Wordnet Conference 2018, Singapore*.

9. Language Resource References

Baker, Collin. (2008). *FrameNet, present and future*.

De Clercq, Orphée and Hoste, Veronique and Monachesi, Paola. (2012). *Evaluating automatic cross-domain Dutch semantic role annotation*.

De Lacalle, Maddalen Lopez and Laparra, Egoitz and Rigau, German. (2014). *Predicate Matrix: extending SemLink through WordNet mappings.*

Kingsbury, P. and Palmer, M. (2002). From treebank to propbank. pages 1989–1993.

Oostdijk, N. and Reynaert, M. and Monachesi, P. and Noord, G. van and Ordelman, R. and Schuurman, I. and Ghinste, V. van. (2008). *From D-Coi to SoNaR: A reference corpus for Dutch*. Paris, France: ELRA.

Postma, Marten and van Miltenburg, Emiel and Segers, Roxane and Schoen, Anneleen and Vossen, Piek. (2016). *Open Dutch WordNet*.