Variation in framing as a function of temporal reporting distance

Levi Remijnse, Marten Postma, Piek Vossen
Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
l.remijnse,m.c.postma,piek.vossen{@vu.nl}

Abstract

In this paper, we measure variation in framing as a function of foregrounding and backgrounding in a co-referential corpus with a range of temporal distance. In one type of experiment, frame-annotated corpora grouped under event types were contrasted, resulting in a ranking of frames with typicality rates. In contrasting between publication dates, a different ranking of frames emerged for documents that are close to or far from the event instance. In the second type of analysis, we trained a diagnostic classifier with frame occurrences in order to let it differentiate documents based on their temporal distance class (close to or far from the event instance). The classifier performs above chance and outperforms models with words.

1 Introduction

To understand streams of news and blogs in terms of the ways in which events can be framed, we need to model how these streams develop over time in relation to the common ground that is created. The common ground between interlocutors plays an essential role in how they refer to real-world event instances.[1] Following pragmatic theory (Grice, 1975; Horn, 1998; Clark et al., 1977), when this common ground is low, the speaker, in an attempt to be cooperative, needs to be as informative as possible, using detailed and marked descriptions of the main event instance. When the common ground is high, the speaker can optimally use less marked expressions and hence background the main event instance in order to foreground related events with a higher informative value (see also the grounding principles of Grimes (2015). The less marked expression then implicates prior knowledge of the event instance, which has become unnecessary to explicate. This is shown in the next two examples that report on instances of the same event type at different points in time (the-day-before versus a-week-ago). In example (1a), reference to the event instance is marked by using multiple indefinite expressions of different syntactic categories in reference to subevents: a *shooting* in which a man *died*. In example (1b), reference to the event instance is restricted to one definite expression *last week's murder*, which presupposes the event instance as shared knowledge and implicates its details. The rest of the text in the example focuses on other events.

(1) a. One man died in a shooting early Thursday morning in southwest Houston. [2]

b. One of the four suspects wanted in last week's murder of Keith Thompson was arrested Wednesday morning at a home in Springfield, according to the Jacksonville Sheriff's Office. [3]

Given this theory about variation in referential expressions, we can expect that, from the onset of an unexpected real-world event instance (e.g., a shootout), the constantly developing narrative of related events (e.g. pursuits, arrests, trials) will enforce these mechanics of foregrounding and backgrounding based on growing mutual knowledge. In other words, the common ground determines the extent to which the speaker is able to background (i.e., use minimal expressions or implicatures) the main event in order to foreground related subjects.

[1] In this paper we use the term *event instance* for event instances of a specific event type, e.g., an instance of shooting

[2] https://www.chron.com/houston/article/Shooting-levaes-man-dead-in-SW-Houston-6688587.php, published on the same day as the event instance.

[3] http://www.news4jax.com/news/crime/1-arrest-in-westside-murder, published a week after the event instance.

Proceedings of the 14th International Conference on Computational Semantics, pages 228–238
June 17–18, 2021. ©2021 Association for Computational Linguistics

Suppose we want to test these principles empirically by examining references in a large dataset, e.g., a referentially grounded corpus. This requires a large collection of documents all referencing single event instances, with a large spread of temporal distance between the publication dates and the event instance date. However, most of the available co-referential corpora hardly contain multiple reference texts for the same event instance, let alone with a strong range of publication dates (Ilievski et al., 2016; Postma et al., 2016).

In this paper, we propose to overcome the data sparsity by merging data of event instances of the same event type to study foregrounding and backgrounding phenomena. We assume that it takes approximately the same amount of time for information, on for instance shooting events, to become common ground between members of a society. Furthermore, such a specific event type activates a coherent set of conceptual properties typically used in reference (Vossen et al., 2020; Morris and Murphy, 1990). Yet, this use of reference might depend on mutual knowledge. Based on the discussed pragmatic principles, we claim that both referential expressions and their meanings vary across documents with different temporal distances to event instances: over time, relevant information about the event instance is left implicit as a means to background reference to the event instance and foreground reference to novel information.

In order to find evidence for our claim, we use FrameNet (Fillmore et al., 2003) as a proxy to characterize event semantics. Our prediction is that those frames typically associated with an event type, called *typical frames*, will also show a different foregrounding and backgrounding distribution as a function of the increased common ground. We expect that subevents of the event instance are foregrounded in texts with little temporal distance, whereas related disjoint events are expected to be foregrounded in texts with large temporal distance. This difference should be reflected by their frames.

To test this hypothesis, we applied a method based on Grootendorst (2020) to learn frame typicality rates for event types from a large collection of news reports that were processed with an automatic frame-labeler (Swayamdipta et al., 2017). Furthermore, we trained a Linear Support Vector Machine classifier to distinguish between referential texts with close temporal distance and further temporal distance on the basis of the typical frames evoked by the texts. We contrast this classifier against models trained on words. We provide evidence that frame distributions are learned by the classifier to perform the task, whereas this is lesser the case for word based models. Our analysis of the results shows that the typical frames evoked in texts with a short temporal distance are backgrounded in texts of larger temporal distance by means of implicature.

The main contributions of our work are:

- We present *HDD* (*Historical Distance Data*), an extensive corpus of reference texts for event instances grouped under event types, with a large spread of temporal distance to the event instance;
- We derive a ranking of typical frames cross-event types;
- We show that frames are more informative than their predicates in training a Linear Support Vector to predict the temporal distance class given a document;
- We show that when contrasting frames for an event type between temporal distance classes, the top ranked frames reflect foregrounded topics.

Our results will help future systems in detecting events in texts and their framing but also help the computational modeling of pragmatics and implicatures.

This paper is structured as follows. We first describe relevant past work in Section 2. We then introduce our methodology in Section 3. Section 4 provides the results of our experiments, which we discuss in Section 5. We conclude in Section 6.

2 Background

In this section, we discuss previous work that has been done with respect to event foregrounding, (2.1 FrameNet (2.2), temporal distance (2.3) and event corpora (2.4).

2.1 Event foregrounding

Different studies have focused on the recognition and characterization of foregrounded events. On the sentence level, foregrounded events show high probability of appearing in main clauses, being actively voiced and having a high transitivity (Kay and Aylett, 1996; Decker, 1985). These observations are applied by Upadhyay et al. (2016) to identify the most significant event in a news article.

On the discourse level, it has been observed that normalized frequencies of co-referential event mentions play an important role in detecting the central event of a document (Filatova and Hatzivassiloglou, 2004a,b). According to Choubey et al. (2018), another crucial factor is the scope of the chain of co-referential mentions throughout the document. These mentions foreground subevents in reference to the central event. The discussed examples in Choubey et al. (2018) show that backgrounded events scarcely occur throughout the document, supporting the reader in grounding the foregrounded central event in a commonly known prior event. In line with their proposal, both *died* and *shooting* in (1a) form a chain of foregrounded subevents that make reference to the central event of the document. In (1b), *arrested* is the foregrounded central event, but *murder* is a backgrounded event.

In this paper, we propose that the mentions that foreground the central event instance also activate a coherent set of FrameNet frames typically used in reference to the event type. In analyzing HDD, we find that this set of typical frames is different for documents written long after the event instance, as an effect of backgrounding that event instance and foregrounding related disjoint events.

2.2 Frames as implicatures

We use FrameNet as a proxy to characterize event semantics in this paper.[4] FrameNet is a lexicographic project anchored in the paradigm of frame semantics (Fillmore et al., 2003; Fillmore and Baker, 2010; Baker et al., 2003). Its lexical database consists of over 1200 semantic frames. Each frame is considered a schematic representation of a situation involving semantic roles, and is assumed to be *evoked* by a *lexical unit*, i.e., a lemma in one of its senses. Each frame exhibits an inventory of lexical units. Below, (1) is extended with FrameNet annotations.

(2) a. One man DEATH$_{\odot}$*died* in a KILLING$_{\odot}$*shooting* [...]

b. One of the four SUSPICION$_{\odot}$*suspects* wanted in last week's KILLING$_{\odot}$*murder* of Keith Thompson was ARREST$_{\odot}$*arrested* [...]

With respect to inferential relations between frames, literature largely focuses on different types of *frame-to-frame relations*, i.e., asymmetric relations between two frames. The FrameNet database registers frame-to-frame relations between the frames to form a network. For example, the Precedes relation specifies a sequential order between two frames, e.g., ARREST shows a Precedes relation to ARRAIGNMENT (Ruppenhofer et al., 2010). Thus, when ARRAIGNMENT is evoked in a document, we can infer ARREST as an implicature. Frames connected through Precedes relations form a coherent set in which any frame implicates the "preceding" frames. The output of our experiment can be used as input for FrameNet to form more of these cohesive sets of frames with temporal relations.

2.3 Temporal Distance

The effect of the temporal distance between a reference text's publication date and the event date on variation in reference has been explored in a few studies.[5] Staliūnaitė et al. (2018) focus on co-reference to entities in the *New York Times Annotated Corpus* (Sandhaus, 2008), which contains articles spanning 20 years. They show that as a function of common knowledge, references to the same entity become definite, of shorter length, i.e., less marked, and with less use of appositives.

Cybulska and Vossen (2010) carried out a statistical analysis on a corpus of reference texts concerning the Srebrenica Massacre. The corpus consisted of 52 news articles (evenly distributed over two news journals) published within a time range of 10 days after the event, and 26 "historical" texts published years later. They created a word-based frequency distribution of references. They showed a strong discrepancy in type-token ratio between the two conditions of temporal distance: the sub-corpus written close to the event shows a higher number of word types than the sub-corpus written years later. The authors conclude that difference in temporal distance correlates with variation in language use. Short temporal distance leads to more variation in descriptions, due to focus on sub-

[4] https://framenet.icsi.berkeley.edu/fndrupal/

[5] On discourse level, referential variation as an effect of common ground has been studied more intensively. See Yoshida (2011); Markert et al. (2012); Del Tredici and Fernández (2018).

events, while longer distance leads to less variation in descriptions due to focus on the main event.

Our research aims to contribute to Cybulska and Vossen (2010) in the following ways. Our HDD is restricted to news articles under the assumption that variation in reference can also be observed within genres. Hence, the potential confounding variable of variation between text genres in their study is eliminated. Second, HDD covers reference texts of multiple event instances of a single event type. Third, we use FrameNet to measure variation in typically evoked frames on top of expressions. Finally, the dimension of temporal distance in our experiments ranges to 30 days after the event instances, instead of years.

2.4 Event corpora

In event co-reference research of the last decade, the corpus datasets show a small number of documents referencing events. Vossen et al. (2018) provide an overview of the nine governing text corpora (e.g., OntoNotes (Pradhan et al., 2007), ECB (Bejan and Harabagiu, 2010), ACE2005 (Peng et al., 2016)) and observed that their sum consists of less than four thousand documents. The number of mentions of events is small within documents (10 mentions per document on average) and only a subset of the corpora contains cross-document event co-reference. Also more recent attempts to manually create annotations for all sentences in articles did not cover a high number of documents (Cybulska and Vossen, 2014; Song et al., 2015; O'Gorman et al., 2016).

Since we need a substantial amount of event reports of the same event type for our experiment, we used the Multilingual Wiki Extraction Platform (MWEP) (Vossen et al., 2020) to obtain a large corpus of referentially grounded news texts. MWEP follows the data-to-text method and takes event types as input to query Wikidata (Vrandečić and Krötzsch, 2014) for event instances. For the obtained event instances, MWEP crawls the corresponding Wikipedia pages and their primary reference texts. These pages are processed by NLP systems, resulting in a corpus of multilayered linguistic annotation files.

3 Methodology

In this section, we describe the methodology used for both the between-event type and within-event type experiments.[6] This includes the resources, (3.1), data processing (3.2), contrastive analysis (3.3, hypotheses (3.4) and evaluation (3.5).

3.1 Resources

The model used to describe our data relies on three main concepts: event type, incident, and reference text. Let E be a set of event types, let I be a set of real-world event instances, and let R denote a registry of reference texts. Each real-world instance $L_i \in I$ is an instance of one or more event types. Also, there can be reference texts that refer to a particular real-world instance L_i. For example, the reference text *Significance of Orlando gunman calling 911 during standoff*[7] refers to the real-world event instance *Orlando nightclub shooting*[8], which is an instance of several event types according to Wikidata, including *mass shooting*[9] and *mass murder*.[10] Based on Wikidata, the incident date can be obtained.

Commonly, our pointer to a reference text is an URL. We apply the following steps to locate, retrieve, and process the reference text. First, we make use of the Internet archive Wayback Machine[11]. Please note that this step is not successful for all URLs. Second, we apply news-please (Hamborg et al., 2017) to crawl the reference text as well as the publication date. Third, we process the text using spaCy (Honnibal et al., 2020) for sentence splitting, tokenization, lemmatization, and dependency parsing. Finally, we apply Open-SESAME (Swayamdipta et al., 2017), which was retrained in order to be used. The collection process results in a document with annotations for various NLP tasks, including frame identification, and the publishing date of the document is typically known.

We make use of two routes to obtain data for HDD according to our model. We apply MWEP on three Wikidata event types: *presidential election* (Q858439), *storm* (Q81054), and *music festival* (Q868557). The second source is a Kaggle dataset called *Gun Violence Data* (Ko, 2018), which con-

[6] the code is available at https://github.com/cltl/HDD$_a$nalysis.

[7] https://www.cbsnews.com/news/orlando-shooting-investigation-gunman-omar-mateen-911-call/

[8] https://www.wikidata.org/wiki/Q24561572

[9] https://www.wikidata.org/wiki/Q21480300

[10] https://www.wikidata.org/wiki/Q750215

[11] https://web.archive.org/

tains approximately 260,000 real-world instances regarding the event type *gun violence*, containing links between reference text URLs and the real-world instances. The four event types are selected due to their differentiation of conceptual properties, which makes them suitable for a contrastive analysis. The descriptive statistics of applying our retrieval and processing software are shown in Table 1.

For each of the four selected event types, Table 1 presents the descriptive statistics. MWEP is capable of generating data for various different event types. However, the number of incidents and reference texts are limited, while the number of reference texts per incident is relatively high. The gun violence dataset, on the other hand, provides a high number of incidents for one specific event type, i.e., gun violence, but the number of texts per incident is relatively low.

Finally, we compute the *temporal distance*, which we define as the number of days between the incident date and the publishing date of a reference texts that makes reference to it. We visualize the distribution of temporal distance for the event type gun violence (Q5618454) for those reference texts for which we were able to obtain a publishing date, see Figure 1.

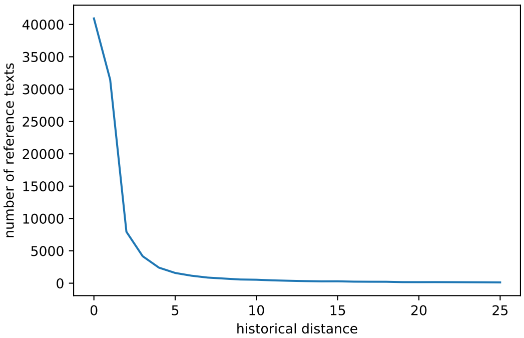

Figure 1: The distribution of the temporal distance is shown for the reference texts that are published within 25 days of the incident, which holds for approximately 90% of the reference texts for the event type gun violence (Q5618454)

Figure 1 visualizes the distribution of temporal distance for the event type gun violence (Q5618454). Most texts are published at the day of the incident. As time passes, the number of documents written about an incident decreases. Still, more than 10,000 are written after 25 days have passed.

3.2 Processing the corpus

We chose to train our diagnostic classifier on the gun violence data, since this subcorpus of HDD shares the largest volume of texts. The following steps were taken to preprocess the data for training. First, a subset of 6,290 documents containing less than 10 annotated frames were removed. These are most likely documents whose URLs were not successfully retrieved by the Wayback Machine, resulting in raw text of error messages, cookies etc. We also removed a subset of 16,237 documents for which news-please was not able to retrieve the publication date.

Next, we specified two temporal distance classes: "day 0" and "day 8-30". The remainder of documents were categorized into those classes according their publication date. After this step, day 0 covers 38,930 documents and day 8-30 covers 6,291 documents. We chose to train a Linear Support Vector model with both this unbalanced variant and a balanced variant in which the documents of day 0 are reduced to a randomized set of equal size as day 8-30.

Per document, both the frequencies of the frames and of their predicates were extracted and separately implemented as features in a data frame. A column was added with the temporal distance classes as labels. Each data frame was split into a training set (80%), a development set (10%) and a test set (10%). We ended up with data frames for both predicates and frames in a balanced and unbalanced corpus condition (4 experiments).

For each experiment, LinearSVC from Scikit Learn (Pedregosa et al., 2011) was used to train a Linear Support Vector with both the features of the experiment and the temporal distance classes as labels. This diagnostic classifier was applied to the test set and evaluated as a multi-class task per experiment.

3.3 Typical frame detection

The HDD corpus was first used for a contrastive analysis between event types and between temporal distance classes of gun violence. The aim was to derive typical frames, i.e., frames that are typically evoked in reference to a certain event type. The following steps were taken to process the data. We selected the data for the event types presidential election, storm and music festival, from which a total set of 57 documents containing less than 10 annotated frames were removed. From the event type

event type	# Li	# of Ri	Avg # of Ri per Li
presidential election (Q858439)	111	408	3.7
storm (Q81054)	60	256	4.3
music festival (Q868557)	13	205	15.8
gun violence (Q5618454)	103,090	123,659	1.2

Table 1: Descriptive statistics regarding the key data concepts of the data forming HDD, used for the experiments. The first three rows originate from using MWEP to obtain data, whereas Gun Violence Data (Ko, 2018) is used for the data of the last row. The first column indicates the event types and the Wikidata identifier of the event type. The second column, L_i, indicates the number of real-world incidents that belong to the event type. The third column, R_i, presents the total number of reference texts, each referring to one of the real-world incidents. Finally, the average number of reference texts per real-world event instance are shown.

gun violence, we used the documents for which the publication date could not be retrieved. Next, the corpus was randomly sampled by equalizing the volume of texts to the smallest collection (N=191), resulting in an equal amount of reference texts per event type.

For the analysis between event types, all frame annotations were extracted from the documents and compiled per event type. Next, we apply an **FFICF** metric (a derivative of C-TFIDF), where *FF* stands for the frame frequency in a subcorpus, and *ICF* is the inverse collection frequency (Vossen et al., 2020). This results in an FFICF score (henceforth *typicality score*) per frame per event type.

C-TFIDF was designed by Grootendorst (2020) with the purpose of determining the topic of a word cluster based on the set of highest scoring words. We have the advantage that, due to the data-to-text approach, the documents in HDD are already clustered based on predefined topics, i.e., event types. It follows that if we apply C-TFIDF to our corpus, we merely have to validate the highest scoring frames. Adapted to collections of frames, the mathematical model reads as follows:

$$FF - ICF_i = \frac{t_i}{f_i} \times log\frac{m}{\sum_j^n t_j} \tag{1}$$

where the frequency of each frame *t* is extracted for each event type *i* and divided by the total number of frames of that event type. Then, the total number of documents *m* across event types is divided by the total frequency of the frame *t* across event types *n*.

We applied this metric to our sampled subset of HDD and ranked the typicality scores per event type. Furthermore, we performed a similar FFICF procedure between the temporal distance classes of gun violence.

3.4 Hypotheses

1. FFICF between event types
We expect the frames with high typicality scores to differ between event types. The frames with the lowest typicality scores may be similar across event types, being a-typical.

2. FFICF between temporal distance classes
We expect the frames with high typicality scores to differ between texts from the same event type gun violence but from different temporal distance classes due to foregrounding and backgrounding.

3. Training and testing the Linear SVM
We expect the diagnostic classifier to perform above chance in predicting the temporal distance class given a document, when the texts are represented by the typically evoked frames. With frame frequencies as features, the model will outperform word based models.

3.5 Evaluation

In order to validate the outcome of the contrastive analysis between temporal distance classes, we presented two annotators with frames from the subcorpus of gun violence for which the publication dates could not be retrieved. Frames with three or less occurrences across this subcorpus were filtered out. For each of the remaining 282 frames, the annotators were asked to provide a binary judgment about whether it is typical in reference to an incident of gun violence at day 0. We utilized the notion *narrative container* (NC) from Pustejovsky and Stubbs (2011), i.e., the scope between the onset of the event instance and the document creation time, to estimate the possible subevents that have a high chance of being referred to in a document on day 0. The annotators had to judge whether each frame is part of the NC. We used Cohen's kappa (Cohen, 1960) to obtain a measure of interannotator-agreement. We expect that the frames annotated as part of the NC of day 0, also occur in the top rank of FFICF scores for this class, whereas frames annotated as falling outside of the NC occur in the top rank of FFICF scores for day 8-30.

We evaluated the output of the diagnostic classifier in a multi-class classification report with precision, recall and F1-score in addition to accuracy, macro average and weighted average.

4 Results

For the contrastive analysis between event types, Table 2, shows the top and bottom ranked FFICF scores for the event types gun violence and music festival. The top ranked frames differentiate between event types and appear to reflect their typical properties. In contrast, the bottom ranked frames are the same for both types and reflect generic event properties.

For the contrastive analysis between temporal distance classes, Table 3 shows the top and bottom ranked FFICF scores between the two temporal distance classes of the event type gun violence. LAW_ENFORCEMENT_AGENCY, KILLING and CATASTROPHE, which were in the top ranking in Table 2, ended in the bottom ranking here. Furthermore, except for two frames, the top ranking of both classes in Table 3 is occupied by different frames.

The annotators show a Cohen's kappa of .48, which is moderate. However, their judgments on the frames in the top and bottom ranking of FFICF ratings in Table 3 show a rather high agreement (20 out of 26 frames, 77%). Half of the top ranked frames in day 0 are annotated as part of the NC of day 0, and almost all top ranked frames at day 8-30 are annotated as not belonging to that same NC. Note that the three frames that are both in the top ranking of scores between event types and at the bottom ranking of scores between temporal distance classes, are also annotated as part of the NC.

Table 4 displays the evaluation report of the experiments with the Linear SVM classifier. In the unbalanced conditions, the accuracy is above 0.85, but biased towards the performance for day 0. The model performed below chance in predicting day 8-30. For frames, the performance in this class is higher than for predicates. In the balanced conditions, the performance decreases for day 0, but increases for day 8-30. For predicates, the model performs around and above chance, with higher recall for day 0 and lower recall for day 8-30. For frames, the F1 is above 0.75, with consistent precision and recall.

5 Discussion

In Table 3, we find that SHOOT_PROJECTILES and JUDGMENT_COMMUNICATION remain in the top ranking, each in a different class. All other frames in the top ranking are typically used in reference to the events of their respective class. Many of those frames can be considered typical for gun violence (e.g., EXPERIENCE_BODILY_HARM, JUDICIAL_BODY), but their evocation is subjected to the temporal distance class. The frames on day 0 refer to subevents of the central event instance, while the frames on day 8-30 refer to related disjoint events, as is generally validated by the annotators. We interpret this variation as an effect of foregrounding and backgrounding. Most typical frames on day 0 are backgrounded in day 8-30 due to the high common ground. They are pragmatically implicated in order to foreground the frames of day 8-30, which carry the highest informative value, but are not typically used in reference to the central event instance of day 0.

Recall that in order to implicate shared knowledge, one uses minimal or less marked expressions. Thus, if the typical frames of day 0 have become shared knowledge in day 8-30, then the writer optimally uses short and definite expressions to implicate them. Such expressions then evoke a strong typical frame, an *anchor* frame, that is sufficient to both refer to the event type and implicate the typical frames as shared knowledge. Such an anchor frame should show a high typicality score for the event type, but a low score across temporal distance classes, due to its frequent usage. KILLING and CATASTROPHE in Table 2 and Table 3 meet both requirements and refer to the main event instance. These might behave as anchor frames on day 8-30, backgrounding the main event instance and implicating the typical frames of day 0 as shared knowledge. This is demonstrated in (1b), where KILLING is evoked in the backgrounded noun phrase.

Finally, the results of the diagnostic classifier in Table 4 show that frame occurrences are more informative for the model than predicate occurrences. The above-chance performance of the model in the balanced/frames condition shows that it is capable to learn temporal patterns, just by paying attention

rank	music festival	gun violence
1	PERFORMING_ARTS (1)	ARREST (1)
2	SOCIAL_EVENT (.990)	LAW_ENFORCEMENT_AGENCY (.991)
3	CREATE_PHYSICAL_ARTWORK (.984)	WEAPON (.980)
4	PARTICIPATION (.975)	HIT_TARGET (.977)
5	ORIGIN (.967)	SHOOT_PROJECTILES (.952)
6	COMMERCE_SELL (.965)	KILLING (.946)
7	LOCALE_BY_EVENT (.965)	JUDGMENT_COMMUNICATION (.929)
8	EXPERTISE (.964)	SCRUTINY (.926)
9	COMPETITION (.964)	LOCATING (.919)
10	MANUFACTURING (.960)	CATASTROPHE (.919)
...	...	...
714	PEOPLE (.862)	CARDINAL_NUMBERS (.777)
715	LOCATIVE_RELATION (.840)	POLITICAL_LOCALES (.765)
716	CARDINAL_NUMBERS (.804)	LOCATIVE_RELATION (.763)
717	LEADERSHIP (.757)	LEADERSHIP (.629)
718	POLITICAL_LOCALES (.631)	PEOPLE (.601)
719	STATEMENT (.568)	STATEMENT (.040)
720	CALENDRIC_UNIT (0)	CALENDRIC_UNIT (0)

Table 2: The top 10 highest ranked frames (FFICF score) and the 7 bottom ranked frames for the event types music festival (Q868557) and gun violence (Q5618454). The scores were remodeled from (-1,1) to (0,1)

rank	day 0	day 8-30
1	STATE_OF_ENTITY (.007566) [D]	JUDICIAL_BODY (.007431) [N]
2	EXPERIENCE_BODILY_HARM (.006752) [Y]	DOCUMENTS (.007431) [N]
3	CAUSE_HARM (.006729) [Y]	JUDGMENT_COMMUNICATION (.006781) [N]
4	EVENT (.006607) [Y]	THEFT (.006538) [D]
5	MEDICAL_CONDITIONS (.006393) [Y]	INTOXICANTS (.006307) [N]
6	TAKING_TIME (.006317) [N]	BAIL_DECISION (.00623) [N]
7	SHOOT_PROJECTILES (.006266) [Y]	ORDINAL_NUMBERS (.006139) [N]
8	DIRECTION (.006037) [D]	CATEGORIZATION (.005915) [N]
9	RESPONSE (.006009) [N]	EVIDENCE (.005842) [N]
10	INFORMATION (.006006) [D]	UNATTRIBUTED_INFORMATION (.005827) [N]
...	...	...
710	KILLING (-.00196) [Y]	KILLING (-.00229) [Y]
711	VEHICLE (-.00299) [D]	VEHICLE (-.00302) [D]
712	LEADERSHIP (-.00422) [D]	CATASTROPHE (-.00421) [Y]
713	ROADWAYS (-.00552) [N]	LEADERSHIP (-.00421) [D]
714	CATASTROPHE (-.005763) [Y]	ROADWAYS (-.00457) [N]
715	AWARENESS (-.00763) [N]	AWARENESS (-.00656) [N]
716	BUILDINGS (-.0093) [Y]	BUILDINGS (-.0072) [Y]
717	LAW_ENFORCEMENT_AGENCY (-.01465) [Y]	LAW_ENFORCEMENT_AGENCY (-.01063) [Y]
718	PEOPLE (-.0379) [Y]	PEOPLE (-.03166)[Y]
719	CALENDRIC_UNIT (-.06463) [Y]	CALENDRIC_UNIT (-.05501) [Y]
720	STATEMENT (-.09892) [N]	STATEMENT (-.08964) [N]

Table 3: The top 10 highest ranked frames (FFICF score)[annotators' score: Y = yes, N = no, D = disagreement] and the 11 bottom ranked frames for the classes “day 0” and “day 8-30” within the event type gun violence. The scores range between -1 and 1.

	precision	recall	F1	support
1. predicates/unbalanced				
day_0	0.861	0.998	0.925	3896
day_8-30	0.357	0.008	0.016	630
Accuracy			0.860	4526
macro avg	0.609	0.503	0.470	4526
weighted avg	0.791	0.860	0.798	4526
2. frames/unbalanced				
day_0	0.891	0.974	0.931	3896
day_8-30	0.627	0.267	0.374	630
Accuracy			0.880	4526
macro avg	0.759	0.620	0.653	4526
weighted avg	0.855	0.876	0.854	4526
3. predicates/balanced				
day_0	0.562	0.676	0.614	630
day_8-30	0.594	0.473	0.527	630
Accuracy			0.575	1260
macro avg	0.578	0.575	0.570	1260
weighted avg	0.578	0.575	0.570	1260
4. frames/balanced				
day_0	0.746	0.789	0.767	630
day_8-30	0.776	0.732	0.753	630
Accuracy			0.760	1260
macro avg	0.761	0.760	0.760	1260
weighted avg	0.761	0.760	0.760	1260

Table 4: Classification reports providing, precision, recall, F1 and support for the performance of the Linear SVM on the test sets of four different experiments: 1. predicate frequencies/unbalanced corpus; 2. predicate frequencies/balanced corpus; 3. frame frequencies/unbalanced corpus; 4. frame frequencies/balanced corpus. Accuracy, macro average and weighted average are also provided per condition.

to frame occurrences.

We performed a model analysis to derive a ranking of the most important frames that the model used as margins to derive the hyperplane. The top 5 reads: TEMPORAL_SUBREGION, BECOMING_SILENT, SELF_MOTION, STORE and ENFORCING. None of these frames get a high typicality score in Table 3. Although the typical frames in the FFICF analysis show strong effects of foregrounding and backgrounding, idiosyncratic generic frames in the data seem more informative for the model in finding the most optimal separating hyperplane. TEMPORAL_SUBREGION might be a strong generic contender across event types due to its inherent temporal properties.[12] BECOMING_SILENT[13], SELF_MOTION[14] and ENFORCING[15] might show a significant frequency in a specific class in reference to the main event instance or subevents.

We assume that the results of our analysis can be generalized over unpredicted event types. From the onset, the common ground increases over time, affecting the pragmatic principles of foregrounding and backgrounding. Thus, if we would be able to obtain enough texts for the event type storm, we would expect the variation in framing between temporal classes to only occur with this event type as well. Since presidential election and music festival are rather anticipated events, the common ground at day 0 is at maximum height and build up from texts in preceding days. Thus, for these event types, temporal distance classes should be determined from preceding days up until the event itself.

6 Conclusion

In this paper, we measured variation in framing as a function of pragmatic foregrounding and backgrounding. We hypothesized that difference in common ground determine the extent to which the writer is able to background frames typically used in reference to the main event instance. We presented HDD, a corpus consisting of reference texts grouped under event types and enriched with publication dates. HDD was used to both perform FFICF between event types and between temporal distance classes, and train a diagnostic classifier. The former resulted in a ranking of typical frames per event type and between classes. The Linear SVM to a large extent was able to differentiate documents of different temporal distance classes. Frames turned out to be more informative than their predicates in training the model. Yet, The diagnostic classifier prefers idiosynchratic frames for learning the hyperplane.

In future work, we extend our experiments to more event types and we want to learn the specific frame-to-frame relations from the typical frames for event types. We expect to learn subevent relations from texts with short temporal distance and (causal) sequence relations from typical frames in texts with larger temporal distance.

Acknowledgments

This research was funded by the Dutch National Science Organisation (NWO) under the project number VC.GW17.083, Framing Situations in the Dutch language.

References

Collin F. Baker, Charles J. Fillmore, and Beau Cronin. 2003. The Structure of the FrameNet Database.

[12] Examples of lexical units: *later.a*, *earlier.a*, *early.a*.

[13] Examples of lexical units: *quiet.v*, *silence.v*

[14] Examples of lexical units: *walk.v*, *run.v*, *rush.v*

[15] Examples of lexical units: *enforcement.n*, *enforce.v*

International Journal of Lexicography, 16(3):281–296.

Cosmin Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422, Uppsala, Sweden. Association for Computational Linguistics.

Prafulla Kumar Choubey, Kaushik Raju, and Ruihong Huang. 2018. Identifying the most dominant event in a news article by mining event coreference relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 340–345, New Orleans, Louisiana. Association for Computational Linguistics.

Herbert H Clark, S Haviland, and Roy O Freedle. 1977. Discourse production and comprehension.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Agata Cybulska and Piek Vossen. 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)*, Valletta, Malta. European Language Resources Association (ELRA).

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland. European Language Resources Association (ELRA).

Nan Decker. 1985. The use of syntactic clues in discourse processing. In *23rd Annual Meeting of the Association for Computational Linguistics*, pages 315–323, Chicago, Illinois, USA. Association for Computational Linguistics.

Marco Del Tredici and Raquel Fernández. 2018. The road to success: Assessing the fate of linguistic innovations in online communities. pages 1591–1603.

Elena Filatova and Vasileios Hatzivassiloglou. 2004a. Event-based extractive summarization. In *Text Summarization Branches Out*, pages 104–111, Barcelona, Spain. Association for Computational Linguistics.

Elena Filatova and Vasileios Hatzivassiloglou. 2004b. A formal model for information selection in multi-sentence text extraction. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 397–403, Geneva, Switzerland. COLING.

Charles J Fillmore and Collin Baker. 2010. A Frames Approach to Semantic Analysis. In *The Oxford handbook of linguistic analysis*. Oxford University Press.

Charles J Fillmore, Christopher R Johnson, and Miriam RL Petruck. 2003. Background to framenet. *International journal of lexicography*, 16(3):235–250.

H Paul Grice. 1975. Logic and conversation. *Cole, P., and Morgan, J.(Eds.)*, 3.

Joseph E Grimes. 2015. *The thread of discourse*, volume 207. Walter de Gruyter GmbH & Co KG.

Maarten Grootendorst. 2020. Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A generic news crawler and extractor. In *15th International Symposium of Information Science (ISI 2017)*, pages 218–223.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Lawrence R Horn. 1998. Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature. *Pragmatics: Critical Concepts*, 4:383–417.

Filip Ilievski, Marten Postma, and Piek Vossen. 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, Osaka, Japan. The COLING 2016 Organizing Committee.

Roderick Kay and Ruth Aylett. 1996. Transitivity and foregrounding in news articles: Experiments in information retrieval and automatic summarising. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 369–371, Santa Cruz, California, USA. Association for Computational Linguistics.

James Ko. 2018. Gun Violence Data. `https://www.kaggle.com/jameslko/gun-violence-data`.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 795–804, Jeju Island, Korea. Association for Computational Linguistics.

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

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 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, Austin, Texas. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Haoruo Peng, Yangqiu Song, and Dan Roth. 2016. Event detection and co-reference with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 392–402, Austin, Texas. Association for Computational Linguistics.

Marten Postma, Filip Ilievski, Piek Vossen, and Marieke van Erp. 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, Austin, TX. Association for Computational Linguistics.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160, Portland, Oregon, USA. Association for Computational Linguistics.

Josef Ruppenhofer, Michael Ellsworth, Miriam R L Petruck, Christopher R Johnson, and Jan Schefczyk. 2010. FrameNet II: Extended theory and practice.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Ieva Staliūnaitė, Hannah Rohde, Bonnie Webber, and Annie Louis. 2018. Getting to "hearer-old": Charting referring expressions across time. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4350–4359, Brussels, Belgium. Association for Computational Linguistics.

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*.

Shyam Upadhyay, Christos Christodoulopoulos, and Dan Roth. 2016. "making the news": Identifying noteworthy events in news articles. In *Proceedings of the Fourth Workshop on Events*, pages 1–7, San Diego, California. Association for Computational Linguistics.

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 *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3162–3171, Marseille, France. European Language Resources Association.

Piek Vossen, Filip Ilievski, Marten Postma, and Roxane Segers. 2018. 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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Etsuko Yoshida. 2011. *Referring expressions in English and Japanese: patterns of use in dialogue processing*, volume 208. John Benjamins Publishing.