Resource Interoperability for Sustainable Benchmarking: The Case of Events

van Son, C.M.; Inel, O.A.; Morante Vallejo, R.; Aroyo, L.M.; Vossen, P.T.J.M.

published in
Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)
2018

document version
Publisher's PDF, also known as Version of record

Resource Interoperability for Sustainable Benchmarking: The Case of Events

Chantal van Son, Oana Inel, Roser Morante, Lora Aroyo, Piek Vossen
Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
{c.m.van.son, oana.inel, r.morantevallejo, lora.aroyo, piek.vossen}@vu.nl

Abstract

With the continuous growth of benchmark corpora, which often annotate the same documents, there is a range of opportunities to compare and combine similar and complementary annotations. However, these opportunities are hampered by a wide range of problems that are related to the lack of resource interoperability. In this paper, we illustrate these problems by assessing aspects of interoperability at the document-level across a set of 20 corpora annotated with (aspects of) events. The issues range from applying different document naming conventions, to mismatches in textual content and structural/conceptual differences among annotation schemes. We provide insight into the exact document intersections between the corpora by mapping their document identifiers and perform an empirical analysis of event annotations showing their compatibility and consistency in and across the corpora. This way, we aim to make the community more aware of the challenges and opportunities and to inspire working collaboratively towards interoperable resources.

Keywords: resource interoperability, annotation consistency, events

1 Introduction

Language resources are at the core of most research in computational linguistics and natural language processing (NLP) for system development and benchmarking. There is already a wealth of resources available and the collection is continuously expanding. With this growth, monitoring their creation and making them interoperable is becoming increasingly important for finding, accessing and reusing existing corpora. For example, different annotation layers applied to the same document provide valuable opportunities for studying and leveraging interdependencies between different types of annotation, but these opportunities are hampered when too laborious conversion steps are required to resolve structural or conceptual differences in their representations (Chiarcos, 2012a). Despite the efforts of many initiatives working towards solutions (Section 2), interoperability issues still persist today and become more and more problematic. On the one hand, this is because reaching consensus on standards and best practices is not a straightforward task and naturally takes time. On the other hand, we hypothesize that the problems that arise when interoperability is lacking have not been illustrated enough for the necessity of solutions to be widely recognized.

Although there are various types of language resources, we limit our discussion to annotated text corpora for which we define interoperability at the levels presented in Figure 1. At the corpus-level, interoperability involves documenting and representing metadata of the data collection as a whole such as its name, language, type, genre, source, creator, year, size, etc., to enable resolving the identity of corpora in a uniform way. At the document-level, we can further distinguish between the metadata of the document (e.g. filename, language, size) and its body, where the latter consists of a textual string and the linguistic annotations of its substrings (e.g. sentences, phrases, tokens). With respect to the annotations, we adopt the distinction between structural interoperability (annotations of different origin are represented using the same formalism) and conceptual interoperability (annotations of different origin are linked to a common vocabulary) as defined by Chiarcos (2012a).

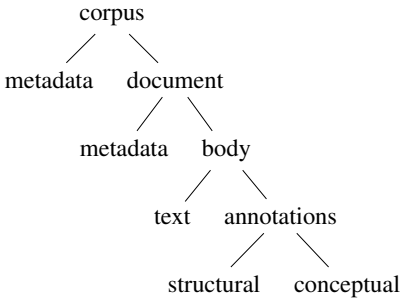

Figure 1: Levels of interoperability

We present an analysis on the document-level interoperability across a set of 20 corpora that have been annotated with events, predicates or propositions. The NLP community defines these terms in various ways, often using each other in their definitions, as in: a *proposition* is formed by a *predicate* with its arguments; *events* are expressed by *predicates* describing situations that happen/occur (Saurí et al., 2006); *predicates* can be of "*propositional*" type (representing an *event*, state, etc.) (Meyers, 2007). There is, however, little agreement on the degree of meaning overlap and relatedness. For the sake of clarity, we will use in this article the term *event* to refer to all three of the overlapping and interrelated notions. Many existing event corpora contain annotations of different aspects of events that are often applied to the same documents (Pustejovsky et al., 2005), providing an interesting use case for analysing interoperability. Furthermore, event annotations involve a wide range of properties and phenomena which makes it ultimately rewarding to achieve interoperability and combine these annotations.

The contributions of this paper are the following:

- **a comprehensive overview of interoperability issues**

across event corpora that result from differences in document naming conventions, textual content and structural/conceptual representations of annotations;
- **a method for aligning diverse language resource corpora** to identify divergent and overlapping corpora;
- **an overview of document intersections across event corpora** revealing opportunities for comparing and combining different annotation layers;
- **an empirical analysis of event annotation compatibility and consistency** in and across corpora.

The remainder of this paper is structured as follows. Section 2 discusses existing initiatives for improving resource interoperability and related studies comparing annotation schemes across corpora. Section 3 provides an overview of the event corpora included in our study. Section 4 reviews metadata-level interoperability across these corpora and shows their overlap in documents, after which we focus on a subset: PropBank/NomBank (PB/NB), FactBank (FB) and TempEval-3 (TE3). We analyse their interoperability at the level of text in Section 5 and at the level of annotations in Section 6. Finally, Section 7 concludes with our lessons learned and proposes some best-practice guidelines.

2 Related Work

There is a range of initiatives collecting and indexing metadata of language resources at the corpus-level to support researchers in finding the right one for their task or application. These include OLAC (Simons and Bird, 2003), Language Grid (Ishida, 2006), the LRE Map (Calzolari et al., 2012), the ELRA Universal Catalog,[1] the LDC Catalog,[2] META-SHARE (Piperidis, 2012), CLARIN (Krauwer and Hinrichs, 2014) and Linghub (McCrae and Cimiano, 2015). The last decades have also seen various metamodel proposals for representing annotations that facilitate structural interoperability, most of which are also translatable to each other. These include GATE (Cunningham, 2002), UIMA (Ferrucci and Lally, 2004), LAF/GrAF (Ide and Romary, 2004; Ide and Suderman, 2007), NIF (Hellmann et al., 2013), NAF/GAF (Fokkens et al., 2014) and PAULA/POWLA (Chiarcos and Erjavec, 2011; Chiarcos, 2012b). Repositories of linguistic annotation terminology, such as GOLD (Farrar and Langendoen, 2003), ISOcat (Windhouwer and Wright, 2012) and its successor CCR,[3] make it possible to overcome the heterogeneity of annotation schemes by acting as an interlingua that allows mapping annotations from one scheme to another, thus addressing conceptual interoperability (Chiarcos, 2012a).

However, far from all corpora that we use today follow the principles mentioned above. This may be because they were created in a time where these standards simply did not yet exist. For more recently created corpora, however, there is presumably a plethora of reasons. We hypothesize that one of them is that whereas working groups such as the Open Linguistics Working Group (OWLG)[4] actively promote resource interoperability, there seem to be few examples (that we know of) that actually clearly illustrate the extent of the problems. Some studies indirectly discuss conceptual interoperability by comparing annotation schemes. For example, Aguilar et al. (2014) compare the Events, Entities and Relations represented in ACE, ERE, TAC-KBP Slot-filling, and FrameNet. Werner et al. (2015) compare the factuality/committed belief annotations in FactBank and the Language Understanding (LU) corpus. The differences between the representations of semantic propositions in PropBank, VerbNet and FactBank have been extensively described and even leveraged to build SemLink (Palmer et al., 2014). Close to our work is that of Pustejovsky et al. (2005), who discuss the issues involved in creating a Unified Linguistic Annotation (ULA) by merging the annotation schemes of PropBank, NomBank, TimeBank, the Discourse Treebank and Coreference Annotation. However, their work remains on theoretical ground by limiting their discussion to overlapping and conflicting annotations in example sentences. Our approach is unique in the sense that we provide empirical evidence by discussing the overlap of the actual annotations for the complete resources when aligning them on the same texts, as well as more general distributional similarities and differences with respect to their linguistic types (part-of-speech (POS), lemma).

3 Overview of Event Corpora

For this study, we focus on 20 corpora that are connected to each other in terms of annotations and/or document bodies. More specifically, the corpora have been annotated with events or predicates, or they share pieces of texts and, therefore, contain at least some documents annotated with events.[5] These corpora provide a range of opportunities to compare or merge annotations of events and their aspects. On the one hand, the definition of what constitutes an event varies across the corpora given the initial goal of the corpus. On the other hand, additional annotations capture various aspects of events, such as predicate-argument relations, semantic roles, event types, event coreference, event factuality, event time-stamping and event-event relations.

In the following overview we present a general description of the corpora, which we group according to their annotation standards (these groupings, however, are not clear-cut).

3.1 PropBank, OntoNotes & Abstract Meaning Representation (AMR)

PropBank (Palmer et al., 2004) has been one of the most influential corpora in NLP research. It provides semantic role annotations for all verbs in the 1M word Wall Street Journal (WSJ) section of the **Penn Treebank** (Marcus et al., 1999). Its creation led to that of **NomBank** (Meyers et al., 2008), which marks the sets of arguments that co-occur with nouns in the same set of documents. Most of the annotations of Penn Treebank and PropBank are now, slightly adapted, available from their successor **OntoNotes 5.0** (Weischedel et al., 2013), which contains other annotation layers such as coreference and

[1] http://universal.elra.info
[2] https://catalog.ldc.upenn.edu
[3] https://www.clarin.eu/ccr
[4] https://linguistics.okfn.org

[5] This is by no means a comprehensive list of all corpora meeting these criteria, but we hope this selection provides a good starting point.

named entities and also covers the Chinese and Spanish languages. Furthermore, it includes an additional 200K of broadcast news, 200K of broadcast conversation, 145K of P2.5 data and 200K of Web data taken from other sources. PropBank's representation of semantic roles is also used in the **Abstract Meaning Representation (AMR) corpus** (Knight et al., 2014), which represents the semantics of English sentences as single rooted, directed graphs with the aim of abstracting away from syntactic idiosyncrasies. It uses a variety of sources for its data, including WSJ news.

3.2 Automatic Content Extraction (ACE) & Entities, Relations and Events (ERE)

The key content extraction tasks of the Automatic Content Extraction (ACE) program (Doddington et al., 2004; Strassel et al., 2008), which ran between 1999 and 2008, were defined as the automatic detection and characterization of real-world Entities, Relations, and Events. However, the program mostly focused on entities and relations between them. Event annotations are available only in the **ACE 2005 Multilingual Training Corpus** (Walker et al., 2006), where annotators tagged the extent, trigger, polarity, tense, genericity, modality, participants and attributes for a constrained set of event (sub)types. This data has been reused in several other corpora, including OntoNotes and **Datasets for Generic Relation Extraction (reACE)** (Hachey et al., 2011) We also included **ACE-2 version 1.0** (Mitchell et al., 2003), which originally does not contain event annotations, but a selection of its documents was used in TimeBank, which does (Section 3.3).
ACE was followed by **Light ERE**, which was designed as a lighter-weight version of ACE with the goal of making annotation easier and more consistent. Modifications to ACE for Light ERE included a reduced inventory of entity and relation types, a slightly modified and reduced event ontology, and the addition of event coreference. In turn, Light ERE has transitioned to the more complex **Rich ERE**, with the latter enabling a more comprehensive treatment of phenomena such as event coreference (Song et al., 2015).[6]

3.3 TimeML

TimeML (Pustejovsky et al., 2003a) is a specification language for events and temporal expressions, designed to capture their attributes, to link them (event time-stamping) and to determine the temporal order between events. It has been applied in several corpora, including the **AQUAINT TimeML Corpus** (Brandeis University, 2008) and **TimeBank 1.2** (Pustejovsky et al., 2006). The documents in TimeBank come from PropBank and the ACE-2 corpus. In turn, data from TimeBank and AQUAINT TimeML was used to build **FactBank 1.0** (Saurí and Pustejovsky, 2009), adding a representation of factuality interpretation to the event annotations, and the evaluation datasets of the TempEval shared tasks, from which we selected the **TempEval-3 dataset** (UzZaman et al., 2013).
The TimeML specifications were followed to annotate events in **EventCorefBank (ECB)** (Bejan and Harabagiu, 2010) which was built to encode event structures with relations like SUBEVENT or REASON, and intra- and cross-document event coreference. ECB 1.0 consists of 482 documents from Google News clustered into 43 topics. A first extension to ECB was released by Lee et al. (2012), who revised and completed the original annotations and added entity coreference relations following the OntoNotes annotation guidelines for coreference (Pradhan et al., 2007). We included a second extension called **ECB+** (Cybulska and Vossen, 2014), which contains another corpus consisting of 502 documents, completely (re)annotated according to new guidelines. The corpus is annotated with event classes (based on TimeML), locations and times (based on ACE and TimeML), and intra- and cross-document coreference.

3.4 Other Annotation Standards

The full-text annotations in **FrameNet** (ICSI Berkeley, 2017) capture the frame semantic structures as defined in its lexical database (Fillmore et al., 2003). The documents come from different sources, including PropBank, the AQUAINT Program[7] and the **Lexical Understanding (LU) Annotation Corpus** (Diab et al., 2009). The latter contains annotations of dialog acts, event coreference, event relations and entity relations, but is best known for its annotations of committed belief, i.e. the strength of the author's beliefs and the degree of commitment to their utterance (similar to FactBank).
The **EventStatus Corpus** (Huang et al., 2017) annotated approximately 3,000 English and 1,500 Spanish news articles with temporal and aspectual properties of major societal events, that is, whether an event has already happened, is currently happening or may happen in the future. Its English documents were sourced from English Gigaword Fifth Edition (Parker et al., 2011), which also served as a source for other resources such as the ACE corpora.
Finally, in the **Richer Event Description (RED) Corpus** (O'Gorman and Palmer, 2016) a number of event-related annotation layers are integrated into a single representation of events and participants. It consists of 95 discussion fora and newswire documents annotated with entities, events, times, their coreference and partial coreference relations, and the temporal, causal and subevent relationships between the events. Its documentation mentions that this includes 55 documents annotated by a range of DEFT annotation formats, but it does not specify which ones.

From this overview, it is clear that the relations across the different corpora and their annotations are complex and not trivial, making it difficult to combine them. In the next sections, we will discuss their interoperability in more detail.

4 Document Interoperability: Metadata

The first step towards analysing similar or merging complementary annotation layers of different origin on the same pieces of text is to determine which documents are shared across the corpora of interest. Many corpora select all or a subset of their documents from existing corpora and possibly add additional ones for annotation, which is usually

[6] At the moment, the ERE data is not yet publicly available, but the LDC kindly provided us with a list of training data filenames.

[7] http://www-nlpir.nist.gov/projects/aquaint/

described in their documentation. This way we learn, for example, that the Wall Street Journal articles first released as part of Penn Treebank have also been used in PropBank, NomBank, OntoNotes, FrameNet, the AMR corpus, TimeBank, and more. We know that TimeBank also sourced documents from the ACE program and in turn was together with the AQUAINT TimeML Corpus the basis for creating FactBank and the TempEval-3 evaluation dataset. However, when documents are taken from multiple sources, it is not always specified in the documentation exactly which documents were sourced from where. Additionally, documents often go through several stages of corpus selection and extension. Lack of precise documentation makes it often difficult to trace back the complete history of corpus development and document overlap.

To obtain an overview of document intersections for the corpora mentioned in Section 3 we decided to use document identifiers. However, we found that a direct comparison of the identifiers did not reveal all intersections because documents are often renamed to match different naming conventions. Therefore, we first mapped the filenames for each of the corpora to uniform identifiers. These uniform identifiers are all fully capitalized and do not contain file extensions. Directory names are excluded, unless they are needed for disambiguating files. For example, `WSJ/00/WSJ_0006.MRG` (Treebank-3) and `wsj_0006.txt` (FactBank) are both mapped to the uniform filename `WSJ_0006`. Some corpora required some more work than capitalization and stripping extensions and prefixes. For example, all annotations of the AMR corpus are collected in one file per data collection containing the identifiers of the source documents as meta-information for each annotation, and OntoNotes has renamed its files to simpler filenames (e.g., `NBC20001003.1830.0755` was renamed to `nbc_0001`), but we could use the mapping files specifying the original filenames that are provided in its release for each of its data collections.

Figure 2[8] visualizes the document intersections that we found, revealing the complex network resulting from more than a decade of data selection and extension.[9] To begin with, we observe the expected intersections as discussed above: 480 documents are shared between ECB 1.0 and its extension ECB+, 1,728 documents between PropBank and its successor OntoNotes, and 132 WSJ documents between Treebank, PropBank, NomBank, TimeBank, TempEval-3 and FactBank (one of which also occurs in FrameNet). It also reveals some less obvious intersections, e.g. 20 documents between the RED and AMR corpora. There is only one corpus that completely stands on its own: EventStatus. Finally, it also reveals unexpected lack of overlap between some corpora. For example, OntoNotes and TimeBank/FactBank are partially built on top of PropBank, but they do not share any documents with each other. It appeared that 25% of PropBank was not carried over into OntoNotes, i.e. a set of documents that were considered too

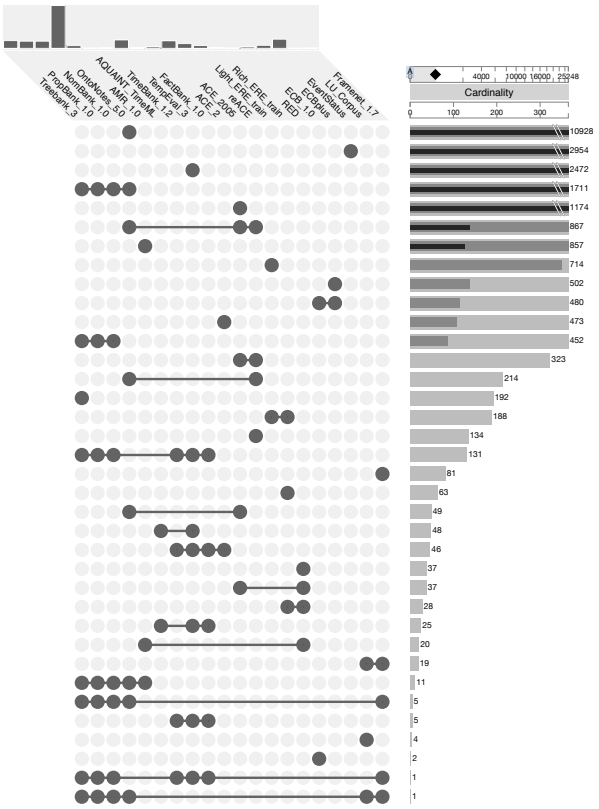

Figure 2: Document intersections of all event corpora

domain-specific because of their strictly financial content. The documents in TimeBank/FactBank sourced from PropBank were all part of this subset. The document intersections reveal several opportunities for merging and comparing annotations across corpora.

5 Document Interoperability: Text

To emphasize the interoperability challenges identified at the text and annotations levels, we performed a small-scale analysis on a subset of corpora, namely PropBank/NomBank (PB/NB), FactBank 1.0 (FB) and TempEval-3 (TE3). For TE3, we only consider the TBAQ-cleaned subcorpus[10] as described in (UzZaman et al., 2013). TBAQ-cleaned corresponds to the complete TimeBank and the complete AQUAINT TimeML Corpus (A-TimeML) with revised event annotations, FB contains the complete TimeBank and part of A-TimeML with their original event annotations, and PB/NB corresponds to the verbal/nominal predicates annotated in the WSJ section of Treebank-3. Therefore, we indirectly analyse these corpora as well. Figure 3 summarizes the document intersections of these three corpora and related ones.

Table 1 summarizes the content of the three datasets. PB/NB is the largest corpus in terms of number of documents (2,312), number of sentences (49,208), number of tokens (1,174,165) and also number of events (227,491). FB consists of 208 documents which are split into 3,839 sentences and 77,231 tokens and contains 9,492 events. TE3

[8] Figures 2 and 3 were created using UpSet (Lex et al., 2014), see: http://caleydo.org/tools/upset

[9] An interactive and more detailed version of the UpSet visualization can be viewed by following the instructions at https://github.com/cltl/CorpusComparison.

[10] https://www.cs.york.ac.uk/semeval-2013/task1/

Event Dataset	#Documents	Total # Sentences	Total # Tokens	# Sentences with Events	# Sentences without Events	Total # Events	Avg. #Events / Sentence
PB/NB	2,312	49,208	1,174,165	47,394	1,814	227,491	4.79
FB	208	3,839	77,236	2,807	1,032	9,492	3.38
TE3	256	3,955	99,384	3,604	351	11,129	3.08

Table 1: Content overview of selected event corpora

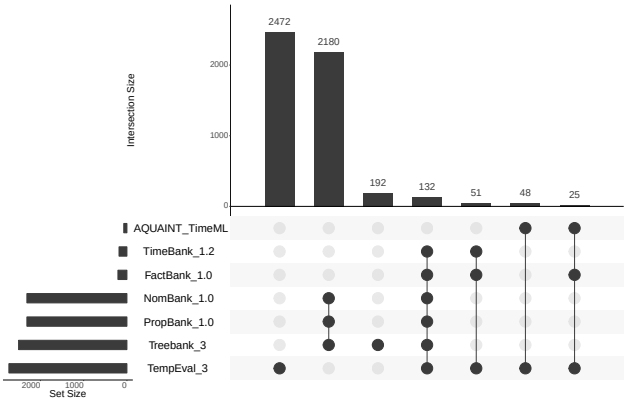

Figure 3: Document intersections of selected corpora

consists of 256 documents, 3,955 sentences[11] and about 100k tokens of which 11,129 have been annotated as events. With respect to the text-level interoperability among these three datasets, we can mention that while PB/NB and TE3 contain only sentences from the article body, FB includes metadata such as the document id, its creation date and its title as the first sentences of the text. These sentences account for the major part of the large number of sentences without annotated events in FB - 1,032 (26.88%), as opposed to 1,814 (3.69%) in PB/NB and 351 (8.87%) in TE3. TE3 has the lowest number of events annotated per sentence, i.e. 3.08 compared to 3.38 in FB and 4.79 in PB/NB.

6 Document Interoperability: Annotations

Chiarcos (2012a) and Chiarcos et al. (2013) define structural interoperability as annotations of different origin being represented using the same formalism, such that different resources can be processed in a uniform way and that their information can be easily merged. Following Ide and Pustejovsky (2010), they define conceptual interoperability as "the ability to automatically interpret exchanged information meaningfully and accurately in order to produce useful results" which can be achieved by linking annotations of different origin to a common vocabulary. In Section 6.1 we summarize the structural interoperability issues that we encountered for the three datasets. In Section 6.2 we review their conceptual interoperability by empirically comparing their annotations of events. This involves a direct comparison between the aligned annotations in FB and TE3 on the basis of their event identifiers, and a type-based analysis, where we abstract away from the annotations in context and provide a more general overview of the annotated types (i.e. POS, lemmas) in all three datasets.

[11] We used Stanford CoreNLP (Manning et al., 2014) in order to split the documents into sentences.

6.1 Structural Interoperability

The data in PB/NB has been released in different formats. Originally, the semantic role annotations were represented with PropBank pointers, i.e. stand-off annotations pointing to locations in the parse tree in Treebank, which in turn was represented by simple labelled brackets in a text file. PropBank pointers are only useful in combination with the corresponding tree structures, since they include empty elements such as traces in the token count. We loaded the data using NLTK.[12] As part of OntoNotes, the data was also released in CoNLL-format, with each line representing a single word with a series of tab-separated fields.[13] However, as mentioned before, OntoNotes does not contain the complete PB/NB. FB also uses a stand-off annotation format and represents the data through a set of 20 tables. For example, one table contains all linguistic information relative to each token (e.g. token id, POS tag), one contains all information relative to each event (e.g. event identifier, text), and one contains the factuality degree values assigned to each event. Finally, TE3 uses the TimeML XSD schema,[14] with XML elements for representing metadata (e.g. <DOCID>, <EXTRAINFO>), the main body of the annotated text (<TEXT> with in-line <EVENT> annotations) and event instances (<MAKEINSTANCE> with tense/aspect of events as XML attributes).

Merging and comparing the annotations of the three datasets is not straightforward due to these structural differences and the textual differences mentioned in Section 5. Both PB/NB and FB use {DOC_ID, SENT_ID, TOKEN_ID} to identify the position of an event in text, but since FB includes the document id, creation date and title as part of the text and PB/NB includes empty elements in the token count, there are mismatches in the sentence and token identifiers. Although the in-line annotations of TE3 do not allow for direct comparison with other datasets, the event identifiers can be used for comparison in this matter.

To analyse the conceptual interoperability of the annotations, we converted all three corpora to CoNLL-format with each line representing a token and information about its document id, sentence id, token id, token text, lemma and POS. If possible, we used the gold sentence splitting, tokenization, lemmatization and POS tagging. If that was not available (e.g. TE3 only has POS tags for most but not all events, PB/NB only has lemmas for events), we used the Stanford CoreNLP pipeline to retrieve the POS tags and the lemma of all the tokens in the datasets.

[12] http://www.nltk.org/howto/propbank.html

[13] https://github.com/propbank/propbank-release

[14] http://timeml.org/timeMLdocs/TimeML1.2.1.xsd

	Verb	Noun	Adjective	Preposition	Number	Adverb	Particle	Determiner	Oth/Unknown
PB/NB	114,574	109,793	-	-	-	-	-	-	-
FB	6,377	2,498	250	45	46	21	6	2	1
TE-3	5,835	2,451	202	10	-	-	-	-	2,632

Table 2: Distribution of POS tags across annotated single-token events

6.2 Conceptual Interoperability

The event annotations in both FB and TE3 are based on the TimeML 1.2.1 Annotation Guidelines (Saurí et al., 2006), which define an event as a situation that happens or occurs. The guidelines further specify in which cases an event should or should not be annotated as a (separate) event. For example, generic events should never be tagged and causative predicates only in specific cases. Thus, TimeML defines events primarily from a semantic point of view, and allows the annotation of all linguistic realizations, including verbs, nominalizations, adjectives, predicative clauses, or prepositional phrases. Whereas the events in FB are the original ones from TimeBank and A-TimeML, those in TE3 are a result of multiple revisions. UzZaman et al. (2013) mention that TE3 borrowed the events from TempEval-2 and added missing events, but no example is given. Verhagen et al. (2010) mention that all event annotations for TempEval-2 were reviewed to make sure that it complied with the latest guidelines, i.e. a simplified version of the TimeML guidelines (Saurí et al., 2009).

In contrast, PB/NB does not take semantics as a starting point, but the syntactic categories of verbs and nouns. PropBank annotates the arguments and adjuncts of each verb with their semantic roles. NomBank does the same for nouns, but defines semantic restrictions with respect to which noun phrases and which constituents of noun phrases are *markable*. For instance, the head noun must be of a "propositional" type (representing an event, state, etc.) and the noun phrase must contain at least one argument and one "proposition-modifying" adjunct (Meyers, 2007).

The semantic and syntactic requirements for annotating events are thus slightly different for TimeML and PB/NB. Another difference is the extents (i.e. span of tokens) of events. In PB/NB, the single noun or verb is annotated as the event, but for phrasal verbs the particle is concatenated with the verb to form a single predicate lemma (Bonial et al., 2010). TimeML implements the notion of minimal chunk, i.e. only the head of the constituent should be annotated and not the whole phrase. As opposed to PB/NB, only the verbal part (and not its particle) of a phrasal verb is marked as event. In the early TimeML guidelines, an exception was made for exocentric elements (i.e. if it has no single head), in which case the entire expression was to be marked (e.g. *on board*). For TempEval-2, however, the annotators always had to annotate only the head.

6.2.1 Overlap of Annotations

As mentioned in Section 6.1, alignment of the annotations of PB/NB with the other datasets is not straightforward due to textual and structural differences. We did, however, perform an analysis on the overlap of event annotations in FB and TE3 for their 208 shared documents by aligning them on the basis of their unique identifiers, i.e. the combination of {`DOC_ID, EVENT_ID`}. The overlap was 8,227 events out of a total of 9,492 in FB and 8,248 in TE3. The revised annotations of TE3 included 24 new events, while 1,265 events from FB were removed. First of all, we observe that all events annotated in TE3 indeed consist of a single token. There were 4 multi-token events that were changed into single-token events: *coup d'etat* $>$ *coup*, *March for Life* $>$ *March* and *World War II* $>$ *War* (2 occurrences). All other multi-token events in FB, 241 in total, were removed. Most of them (238) contained a cardinal number (e.g. *$4.375 a share* and *about 12%*). From the single-token events that were removed (1,265), 46 also consisted of a cardinal number. The rest of the single-token events consisted mainly of nouns (713), verbs (138) and adjectives (82). The 24 new events consisted of 11 nouns, 12 verbs and 1 adjective. Surprisingly, many of both the removed and the newly added events seemed to correspond to the 'source-introducing predicates' as defined in (Saurí and Pustejovsky, 2009), e.g. *say*, *think*, *statement*, *plan*, *confident*, which were however also represented in the shared set of events. This may simply be a result of inconsistency (see also Section 6.3). In sum, the revisions of event annotations mainly seemed to concern the removal of 'quantitative statements' (Saurí et al., 2006) and multi-token events, but we were not able to find other clear patterns in the many events that were lost in the transition from FB to TE3.

6.2.2 Part-of-Speech Distribution of Events

In Table 2 we show the POS tag distribution for every single-token event in the three datasets. We observe that for all three datasets the majority of the single-token event mentions are either verbs or nouns. While in PB/NB only verbs and nouns are annotated as events, in TE3 and FB there are also adjectives and prepositions that stand as events, but in a smaller proportion. Furthermore, FB contains also event mentions of type number, adverb, particle and determiner. In the TE3 dataset, a considerable amount of event mentions do not have a POS tag assigned but are marked as 'Other' or 'Unknown' instead.

Next, we looked at the multi-token events in PB/NB and FB. We observe that there is no consistency between the two datasets. While PB/NB contains multi-token events composed of verbs in combination with particles, adverbs, prepositions, pronouns, adjectives and nouns, FB has no such event mentions. As mentioned before, the majority of the multi-token events in FB are combinations of numerals and symbols, under various patterns. Furthermore, in PB/NB all the multi-token events have exactly two tokens, but in FB 23 out of 246 events have more than 2 tokens.

6.2.3 Overview of Event Tokens and Lemmas

In Table 4 we show the number of event tokens, the number of distinct event tokens and the number of distinct event

		Verb	Noun	Adjective	Preposition	Adverb	Number	Particle	Determiner
PB/NB	Total Tokens	153,721	355,099	83,341	122,957	39,713	45,046	2,990	100,727
	Events	110,291	115,051	2,107	496	302	2	2,320	2
	(%)	71.74%	32.34%	2.52%	0.40%	0.76%	0.004%	77.59%	0.001%
FB	Total Tokens	10,180	24,208	5,032	8,131	2,257	3,854	194	6,711
	Events	6,386	2,546	240	50	29	291	3	9
	(%)	62.73%	10.51%	4.76%	0.61%	1.28%	7.55%	1.54%	0.13%
TE3	Total Tokens	13,643	30,290	6,246	10,706	3,469	3,652	242	8,843
	Events	8,536	2,391	173	17	12	-	-	-
	(%)	62.56%	7.89%	2.76%	0.15%	0.34%	-%	-%	-%

Table 3: Overview of tokens annotated as event per POS

lemmas. For every event token and event lemma we also consider their POS tag. Because FB and TE3 do not contain information regarding the lemma of the tokens nor events and PB/NB only for the events (their rolesets), we compute these statistics based on the output of Stanford CoreNLP. While PB/NB has the lowest ratio of distinct event tokens (7.2%) and event lemmas (4.43%), FB has the highest ratios, around 32% for tokens and around 22% for lemmas. This high number of distinct event instances is due to the fact that many event mentions contain numerals which do not repeat throughout the dataset.

Dataset	Total #Events	Distinct #Event Tokens	Distinct #Event Lemmas
PB/NB	227,491	16,398	10,089
FB	9,492	3,041	2,171
TE3	11,129	2,883	1,871

Table 4: Overview of distinct event tokens and lemmas

6.3 Consistency of Annotations

In this section we analyse the consistency of event annotations. We analyse the consistency at the level of the POS tags (Section 6.3.1) and at the level of event token and event lemma (Section 6.3.2). This analysis is a work-in-progress performed in the context of the CrowdTruth[15] project (Inel et al., 2014), which shows that crowdsourcing is a feasible method to identify and correct inconsistent annotations (Inel and Aroyo, 2017; Aroyo and Welty, 2012).

6.3.1 Evaluation of Event POS Tags

Table 3 shows the total amount of tokens for each of the following POS tags: verb, noun, adjective, adverb, preposition, number, particle, determiner and symbol, as well as the total amount and percentage of tokens annotated as events for each POS tag. Overall, the tokens categorized as verbs have the highest coverage as events, as more than 62% of those have been annotated as events across the three datasets. For all three datasets, the verbs that were not marked as events were mostly the verbs *be*, *have* and *do*, which we assume to be those occurrences where they act as auxiliaries (91%, 65% and 68% of non-annotated verbs in PB/NB, FB and TE3 respectively). However, we also found some surprising cases. For example, *televise* or *bless* are not annotated as events in TE3 and FB, but they are annotated as events in PB/NB, and occurrences of *say* are often not annotated in all three corpora.

The nouns annotated as events have a lower coverage (between 7.89% and 32.34%) and this coverage varies a lot for each dataset. The coverage of adjectives, prepositions and adverbs is quite similar on the three datasets, as shown in Table 3. However, we know that for PB/NB these are all part of phrasal verbs (*cut loose, dig up*), which also accounts for the high coverage of particles in PB/NB. In contrast, adjectives, prepositions and adverbs can act as independent events in FB and TE3 if they have a propositional meaning (*optimistic*, *down*, *in place*). As we have already discussed, only the events in FB cover numbers, but their coverage is still quite low, around 7.5%.

6.3.2 Evaluation of Event Tokens and Lemmas

In Table 5 we present the overview of inconsistencies encountered at the level of annotated event token and event lemma. For each event token and each event lemma we count how many times it appears in the dataset and how many times it was annotated as event. Based on these occurrences, we compute how many times a single event token or event lemma was not annotated, and for how many distinct single event tokens and event lemmas there are instances in the corpus which are not annotated.

Dataset	Event Tokens		Event Lemmas	
	Total (%)	Distinct (%)	Total (%)	Distinct (%)
PB/NB	146,268 39.46%	6,462 42.97%	178,253 44.27%	4,914 53.3%
FB	11,711 55.88%	1,016 36.04%	12,737 57.94%	921 47.25%
TE3	12,580 53.06%	973 33.74%	13,473 54.76%	837 44.73%

Table 5: Inconsistencies at the level of single-token events: event tokens and lemmas

We first observe that there are inconsistencies at the token level since not all instances of an event are always annotated as such. For example, in TE3 *decision* (noun) is annotated as event in 45 out of 53 cases, *embargo* (noun) in 7 out of 13 cases, and *said* (verb) in 993 out of 1,006 cases. Further-

[15] www.crowdtruth.org

more, there are also inconsistencies at the lemma level since not all lemma instances of an event are always annotated as events. For example, in TE3 *disaster* (noun) is annotated as event in 3 out of 7 lemma-based occurrences, *war* (noun) in 32 out of 52 cases, and *export* (noun) in 1 out of 5. Across all three datasets, the total amount of inconsistencies at the level of event lemma is higher than the total amount of inconsistencies at the level of event token which means that only particular forms of a lemma are usually annotated as events. Further, we acknowledge the fact that TE3 shows the least amount of inconsistencies for distinct event tokens and lemmas, although they are still substantial. We believe this is due to the fact that the annotations in the dataset have been revised multiple times.

Many multi-token events in FB are composed of numbers in combination with symbols (*#*), currencies (*$*, *us$*, *c$*) and percentages (%). It is interesting to observe, however, that in cases where the sign is replaced by the word, i.e. *5 percent* instead of *5%*, only the number is annotated as event (we noticed 3 such cases). We also observe that only cardinal numbers are annotated as event, while numbers such as *million* or *billion* are never annotated. We further noticed the following annotation inconsistencies: *# CD* was annotated only 3/23 times as events, *$CD* was annotated only 216/819 times and *CD%* was annotated only 21/304 times. Furthermore, 6 times only the cardinal was annotated as event in constructions of type *c$CD*. While the TimeML guidelines do specify that quantitative statements should only be annotated in case "their validity is relative to the point in time they have been asserted" (Saurí et al., 2006), we hypothesize that inconsistency may have been one of the reasons that they do not occur in TE3 any more.

Lastly, we checked the multi-token events in PB/NB. At the token level, 527 times a phrase that was previously annotated as a multi-token event was not annotated as such. In 475 out of these 527 cases, only the head of the phrase was annotated as a single-token event; the remainder was not annotated at all. At the level of lemmas, 834 instances of multi-token events were missed, including 784 cases where only the head was annotated. It is not easy to determine which of these cases indicate actual syntactic or semantic differences, and which indicate inconsistencies. For example, the combination *back off* is annotated twice as belonging to the roleset back.11,[16] meaning "to retreat from" (e.g. in WSJ_1000-S4: *big securities firms backed off from program trading*). However, we found 5 other occurrences of *back off* that were not annotated as such, 3 of which were incorrectly assigned to another roleset. Consider the two sentences below; in both cases, the verb *back* was annotated as belonging to roleset back.02, meaning "move backwards". In Sentence 1 this is the correct interpretation, but in Sentence 2 back.11 would have been the correct roleset.

1. [...] creditors committee *backed* off a move to come up with its own alternative proposals [...] (WSJ_0475-S0)
2. Previously, he noted, gold producers tended to *back* off from a rising gold market [...] (WSJ_2045-S22)

[16]https://verbs.colorado.edu/propbank/framesets-english

7 Conclusion

Interoperability of resources has been discussed extensively in many standardisation initiatives and meta-model proposals. However, the practical reality is very far from the ideal solutions that have been proposed. In this paper, we provided a detailed and comprehensive analysis of the incompatibilities that are de facto for a selection of 20 text corpora with event data. Our analysis emphasizes the need for resource interoperability in order to facilitate merging similar or complementary event annotations in corpora that overlap in content and hence, to exploit research opportunities.

We provided a comprehensive overview of the document intersections among a representative set of event-centric textual corpora, which is the stepping stone in advancing the resource interoperability process. Furthermore, we have illustrated through empirical analysis various type-level annotation inconsistencies in and across a subset of these event corpora. Based on this study, we conclude that corpora should provide the means to link and align them to other corpora at different levels, for which we propose the following best-practice guidelines:

Metadata: Documents should be named with unified standards (persistent identifiers) and be provided with mappings to documents in other corpora.

Textual content: If corpora share files, their content should be aligned to facilitate the merging of annotations.

Structural: The community should agree on a data structure to represent annotations that capture their provenance, especially when they are performed on the same content.

Conceptual: Preferably, the community should agree on (the interpretation of) their labels. Alternatively, or additionally, basic statistics on the total number of tokens, words, sentences, number of annotated units per type of annotation and coverage of annotation per lemma and POS, released together with the annotated corpus, would help to properly evaluate and understand the consistency of the annotations within the corpus and their compatibility with similar annotations in other corpora. Any revisions should be sufficiently documented.

Ultimately, the community is responsible for reaching consensus on how to publish and distribute resources in the future. We hope that the availability of tools for corpus and annotation analysis and aggregation will stimulate releasing the resources in a more consistent and transparent way. In the future, researchers should spend less time on conversion and mapping of data and more time on conceptually understanding the content of the annotations.

8 Acknowledgement

We would like to thank the anonymous reviewers for their useful feedback, everyone who kindly answered our questions and provided us with the information necessary for our analysis, and all researchers for the hard work that was put into creating the valuable resources cited in this paper. The work presented in this paper was funded by the Amsterdam Data Alliance in the QuPiD project, by the Netherlands Organization for Scientific Research (NWO) via the Spinoza grant, awarded to Piek Vossen in the project "Understanding Language by Machines", and by CLARIAH-CORE project financed by NWO (www.clariah.nl).

9 Bibliographical References

Aguilar, J., Beller, C., McNamee, P., Van Durme, B., Strassel, S., Song, Z., and Ellis, J. (2014). A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53.

Aroyo, L. and Welty, C. (2012). Harnessing disagreement for event semantics. *Detection, Representation, and Exploitation of Events in the Semantic Web*, 31.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Palmer, M., and Schneider, N. (2013). Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (LAW VII & ID)*, pages 178–186.

Bejan, C. A. and Harabagiu, S. M. (2008). A linguistic resource for discovering event structures and resolving event coreference. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2881–2887.

Bonial, C., Babko-Malaya, O., Choi, J. D., Hwang, J., and Palmer, M., (2010). *PropBank Annotation Guidelines, version 3.0*. Center for Computational Language and Education, Research Institute of Cognitive Science, University of Colorad at Boulder, December.

Calzolari, N., Del Gratta, R., Francopoulo, G., Mariani, J., Rubino, F., Russo, I., and Soria, C. (2012). The LRE Map. Harmonising Community Descriptions of Resources. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1084–1089.

Chiarcos, C. and Erjavec, T. (2011). OWL/DL formalization of the MULTEXT-East morphosyntactic specifications. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 11–20.

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013). Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*, pages 7–25.

Chiarcos, C. (2012a). Interoperability of corpora and annotations. In *Linked Data in Linguistics*, pages 161–179.

Chiarcos, C. (2012b). POWLA: Modeling linguistic corpora in OWL/DL. In *Extended Semantic Web Conference*, pages 225–239.

Cunningham, H. (2002). GATE, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254.

Cybulska, A. and Vossen, P. (2014). Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 4545–4552.

Diab, M. T., Levin, L., Mitamura, T., Rambow, O., Prabhakaran, V., and Guo, W. (2009). Committed belief annotation and tagging. In *Proceedings of the 3rd Linguistic Annotation Workshop (ACL-IJCNLP 2009)*, pages 68–73.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., and Weischedel, R. M. (2004). The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*.

Farrar, S. and Langendoen, D. T. (2003). A linguistic ontology for the semantic web. *GLOT International*, 7(3):97–100.

Ferrucci, D. and Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Fillmore, C. J., Johnson, C. R., and Petruck, M. R. (2003). Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

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

Hachey, B., Grover, C., and Tobin, R. (2012). Datasets for generic relation extraction. *Natural Language Engineering*, 18(1):21–59.

Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. (2013). Integrating NLP using linked data. In *International Semantic Web Conference*, pages 98–113.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL (Short Papers)*, pages 57–60.

Huang, R., Cases, I., Jurafsky, D., Condoravdi, C., and Riloff, E. (2016). Distinguishing Past, On-going, and Future Events: The EventStatus Corpus. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 44–54.

Ide, N. and Pustejovsky, J. (2010). What does interoperability mean, anyway? Toward an operational definition of interoperability for language technology. In *Proceedings of the 2nd International Conference on Global Interoperability for Language Resources*.

Ide, N. and Romary, L. (2004). International standard for a linguistic annotation framework. *Natural language engineering*, 10(3-4):211–225.

Ide, N. and Suderman, K. (2007). GrAF: A graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop*, pages 1–8.

Inel, O. and Aroyo, L. (2017). Harnessing diversity in crowds and machines for better NER performance. In *European Semantic Web Conference*, pages 289–304.

Inel, O., Khamkham, K., Cristea, T., et al. (2014). Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In *ISWC*, pages 486–504.

Ishida, T. (2006). Language Grid: An infrastructure for intercultural collaboration. In *Proceedings of the 2005 Symposium on Applications and the Internet (SAINT'06)*.

Krauwer, S. and Hinrichs, E. (2014). The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. In *Proceedings of the 9th Interna-*

tional Conference on Language Resources and Evaluation (LREC 2014), pages 1525–1531.

Lee, H., Recasens, M., Chang, A., Surdeanu, M., and Jurafsky, D. (2012). Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '14)*, 20(12):1983–1992.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

McCrae, J. P. and Cimiano, P. (2015). Linghub: a Linked Data Based Portal Supporting the Discovery of Language Resources. In *Proceedings of the 11th International Conference on Semantic Systems*, volume 1481, pages 88–91.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. (2004). The NomBank project: An interim report. In *HLT-NAACL 2004 workshop: Frontiers in corpus annotation*, volume 24, page 31.

Meyers, A. (2007). *Annotation Guidelines for NomBank – Noun Argument Structure for PropBank*. New York University.

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

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Palmer, M., Bonial, C., and McCarthy, D. (2014). SemLink+: FrameNet, VerbNet and Event Ontologies. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 13–17.

Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 36–42.

Pradhan, S. S., Ramshaw, L., Weischedel, R., MacBride, J., and Micciulla, L. (2007). Unrestricted coreference: Identifying entities and events in OntoNotes. In *International Conference on Semantic Computing, 2007 (ICSC 2007)*, pages 446–453.

Pustejovsky, J., Castano, J. M., Ingria, R., Saurí, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. (2003a). TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of the 5th International Workshop on Computational Semantics (IWCS)*.

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., and Lazo, M. (2003b). The TIMEBANK corpus. In *Corpus linguistics*, volume 2003, pages 647–656.

Pustejovsky, J., Meyers, A., Palmer, M., and Poesio, M. (2005). Merging PropBank, NomBank, TimeBank, Penn Discourse Treebank and Coreference. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 5–12.

Saurí, R. and Pustejovsky, J. (2009). FactBank: A corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227–268.

Saurí, R., Littman, J., Gaizauskas, R., Setzer, A., and Pustejovsky, J. (2006). TimeML annotation guidelines, version 1.2.1.

Saurí, R., Goldberg, L., Verhagen, M., and Pustejovsky, J. (2009). *Annotating Events in English: TimeML Annotation Guidelines (Version TempEval-2010)*. Brandeis University.

Simons, G. and Bird, S. (2003). The Open Language Archives Community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing*, 18(2):117–128.

Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., and Ma, X. (2015). From Light to Rich ERE: Annotation of Entities, Relations, and Events. In *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*, pages 89–98.

Strassel, S., Przybocki, M. A., Peterson, K., Song, Z., and Maeda, K. (2008). Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

UzZaman, N., Llorens, H., Derczynski, L., Verhagen, M., Allen, J., and Pustejovsky, J. (2013). SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9.

Verhagen, M., Saurí, R., Caselli, T., and Pustejovsky, J. (2010). SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, pages 57–62.

Werner, G. J., Prabhakaran, V., Diab, M., and Rambow, O. (2015). Committed belief tagging on the FactBank and LU corpora: A comparative study. In *Proceedings of the 2nd Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 32–40.

Windhouwer, M. and Wright, S. E. (2012). Linking to linguistic data categories in ISOcat. In *Linked Data in Linguistics*, pages 99–107.

10 Language Resource References

Bejan, C. and Harabagiu, S. (2010). *EventCorefBank (ECB)*. Human Language Technology Research Institute.

Brandeis University. (2008). *AQUAINT TimeML Corpus*. TimeML corpora, ELRA-U-W0394.

Cybulska, A. and Vossen, P. (2014). *The ECB+ Corpus*. The NewsReader Project.

Diab, M., Dorr, B., Levin, L., Mitamura, T., Passonneau, R., Rambow, O., and Ramshaw, L. (2009). *Language Understanding Annotation Corpus*. Linguistic Data Consortium, LDC2009T10, ISRLN 775-964-514-342-7.

Hachey, B., Grover, C., and Tobin, R. (2011). *Datasets for Generic Relation Extraction (reACE)*. The University of Edinburgh, distributed via Linguistic Data Consortium, LDC2011T08, ISLRN 494-554-511-556-5.

Huang, R., Jurafsky, D., and Riloff, E. (2017). *The EventStatus Corpus*. Linguistic Data Consortium, LDC2017T09, ISLRN 173-931-115-382-5.

ICSI Berkeley. (2017). *FrameNet 1.7*. FrameNet Project.

Knight, K., Baranescu, L., Bonial, C., Georgescu, M., Griffitt, K., Hermjakob, U., Marcu, D., Palmer, M., and Schneider, N. (2014). *Abstract Meaning Representation (AMR) Annotation 1.0*. Linguistic Data Consortium, LDC2014T12, ISLRN 637-196-362-554-6.

Marcus, M. P., Santorini, B., Marcinkiewicz, M. A., and Taylor, A. (1999). *Treebank-3*. The Penn Treebank (PTB) project, distributed via Linguistic Data Consortium, LDC99T42, ISLRN 141-282-691-413-2.

Meyers, A., Reeves, R., and Macleod, C. (2008). *NomBank v 1.0*. The NomBank Project, distributed via Linguistic Data Consortium, LDC2008T23, ISLRN 652-357-402-514-3.

Mitchell, A., Strassel, S., Przybocki, M., Davis, J., Doddington, G., Grishman, R., Meyers, A., Brunstein, A., Ferro, L., and Sundheim, B. (2003). *ACE-2 Version 1.0*. The ACE Program, distributed via Linguistic Data Consortium, LDC2003T11, ACE corpora, ISLRN 498-363-793-174-9.

O'Gorman, T. and Palmer, M. (2016). *Richer Event Description 1.0*. Linguistic Data Consortium, LDC2016T23, ISLRN 722-524-976-246-2.

Palmer, M., Kingsbury, P., Babko-Malaya, O., Cotton, S., and Snyder, B. (2004). *Proposition Bank I*. The PropBank Project, distributed via Linguistic Data Consortium, LDC2004T14, ISLRN 874-058-423-080-1.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011). *English Gigaword Fifth Edition*. Linguistic Data Consortium, LDC2011T07, ISLRN 911-942-430-413-0.

Pustejovsky, J., Verhagen, M., Saurí, R., Littman, J., Gaizauskas, R., Katz, G., Mani, I., Knippen, R., and Setzer, A. (2006). *TimeBank 1.2*. Brandeis University, distributed via Linguistic Data Consortium, LDC2006T08, TimeML corpora, ISLRN 717-712-373-266-4.

Saurí, R. and Pustejovsky, J. (2009). *FactBank 1.0*. Linguistic Data Consortium, LDC2009T23, ISLRN 293-001-569-603-0.

Walker, C., Strassel, S., Medero, J., and Maeda, K. (2006). *ACE 2005 Multilingual Training Corpus*. The ACE Program, distributed via Linguistic Data Consortium, LDC2006T06, ACE corpora, ISLRN 458-031-085-383-4.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., El-Bachouti, M., Belvin, R., and Houston, A. (2013). *OntoNotes Release 5.0*. OntoNotes Project, distributed via Linguistic Data Consortium, LDC2013T19, ISLRN 151-738-649-048-2.