Investigating interoperable event corpora

Caselli, Tommaso; Bos, Johan

Published in:
Language Resources and Evaluation

DOI:
10.1007/s10579-023-09643-6

Language Resources and Evaluation (2023) 57:1107–1137
https://doi.org/10.1007/s10579-023-09643-6

ORIGINAL PAPER

Investigating interoperable event corpora: limitations of reusability of resources and portability of models

Tommaso Caselli[1] · **Johan Bos**[1]

Accepted: 30 January 2023 / Published online: 26 February 2023

Abstract
Studies on the applicability of heterogeneous semantically interoperable corpora are rare. We investigate to what extent reusability (both of systems and of annotations) is entailed by corpora whose interoperability is based on compliance to standards. In particular, we look at event detection in English texts, supported by the ISO-TimeML annotation scheme. We run two sets of experiments using a common neural network architecture and extensively evaluate our results on both in-distribution and out-of-distribution settings. In all experimental settings, systems obtain state-of-the-art results on the in-distribution data and underperform out-of-distribution ones, setting limits to the benefits of semantically interoperable corpora. By means of a detailed error analysis, we show that while being compliant to a standard guarantees semantic interoperability, this becomes only a necessary condition for reusability, with factors such as differences in the quality of the annotations having a much stronger impact.

Keywords Event detection · Semantic interoperability · Reusability of data · Portability of systems · Standards

1 Introduction

Supervised machine learning is currently the main technique to develop applications in Natural Language Processing (NLP). To work properly, this approach requires annotated data, and since they are usually generated by humans, this is an expensive endeavour. An interesting alternative is to make labelled data *interoperable*, enabling the reuse of costly resources. Interoperability of language resources is a complex notion with multiple dimensions. This work is concerned with one specific dimension,

✉ Tommaso Caselli
t.caselli@rug.nl

Johan Bos
johan.bos@rug.nl

[1] University of Groningen, Oude Kijk in't Jatstraat, 26, 9712 EK Groningen, The Netherlands

namely that of *semantic interoperability*. In recent years, most efforts have focused on implementing semantic interoperability for language resource infrastructures with dedicated initiatives and EU funded projects (de Jong et al., 2020; Rehm et al., 2020, among others). When it comes to (annotated) data, the debate on what constitutes semantic interoperability and how to best implement it is still open (Witt et al., 2009; Ide & Pustejovsky, 2010; Chiarcos, 2012; Hajičová 2014). Besides specific differences, one of the promises of semantic interoperability is to make annotated data for a specific language phenomenon *reusable*. Reusability of annotated data has different facets: on the one hand, it can be interpreted as possibility of searching across different corpora for the same language phenomenon. On the other hand, it can be interpreted as combination of multiple corpora to increase the diversity of training material for stochastic NLP systems. The aim of this article is to investigate this latter interpretation of semantic interoperability.

Interoperability of (annotated) data is closely related to portability of supervised learning systems. Following Ettinger et al. (2017), good portability of systems would indicate the development of more robust NLP technologies. This means that a system trained for a specific task is expected to perform well (or with very limited losses) across datasets. Moreover, portability requires that predictions made by the system are consistent, even in presence of small perturbations to the input (Wang et al., 2020). To achieve portability one can apply transfer learning and domain adaptation (Blitzer et al., 2006; Daumé III, 2007; Ma et al., 2014; Ruder et al., 2017; Ruder & Plank, 2017; Wu & Huang, 2016).

Another strategy for portability, one that is the focus of this article, is training systems with larger and more varied training data (Tu et al., 2020). Previous work (Alex et al., 2006; van Erp et al., 2016) has shown that the lack of interoperability for the categories used to annotate the data is the perfect recipe for failure in reusability of the annotated resources and the portability of systems. On the other hand, the use of *interoperable annotated resources* has proven to be successful. A case in point are the interoperable treebanks for syntactic parsing (Niu et al., 2009; Stymne et al., 2018), whose experimental data supports the connection between interoperability of annotated resources and portability of systems. This suggests that interoperability can be a faster and more reliable solution to access diversified supervised training material rather than simply concatenating datasets regardless of their annotation specifications (Witt et al., 2009; Poch & Bel Rafecas, 2011).

But one swallow doesn't make a summer. In this contribution we conduct a thorough empirical investigation on the benefits of *semantically annotated* interoperable corpora. The specific case that will be subject to our scrutiny are corpora annotated with information about events. Events have been at the heart of Linguistics, Philosophy, and Artificial Intelligence. Much work has been conducted in this area reaching a level of maturity and consensus that has boosted the development of dedicated semantically interoperable corpora. We will empirically assess, for the first time as far as we are aware, the extent to which the relationship between interoperability and reusability (of systems and data) holds for event-annotated corpora. In particular, we address a non-trivial dimension of interoperability, namely the use of a shared vocabulary and markables across annotation schemes. Our main contributions are:

– an in-depth investigation of the interoperability and compatibility of event-annotated corpora for English compliant to the ISO-TimeML standard;
– a detailed analysis of the factors that impact the interoperability of language resources;
– state-of-the-art models for event trigger detection.[1]

This article is organised as follows. First, in Sect. 2, we give a thorough background on how interoperability is conceived in the NLP community and the presumed advantages of having semantically interoperable language resources. We then provide an overview of the selected task, i.e., event detection (Sect. 3) and discuss current state-of-the-art methods that have been used to address it. Section 2 discusses the ISO-TimeML annotation scheme (Pustejovsky et al., 2010), a standard for event and temporal annotation, and introduces three ISO-TimeML compliant corpora. The corpora will be used to investigate semantic interoperability, with particular attention to two aspects: portability of systems (Sect. 5) and reusability of annotated data (Sect. 6). Section 7 presents a detailed analysis to identify differences in the application of the annotation guidelines that may have had an impact on the interoperability experiments. We conclude with directions for future work in Sect. 8.

2 Syntactic and semantic interoperability

The increasing popularity of data-driven methods in NLP results in the development of a varied and large number of linguistic corpora with an even larger variation of annotations. For instance, the LRE Map[2] documents more than 2,608 written corpora for different languages. Similarly, there has been a parallel development of tools to represent, annotate, and visualise such varied data. This proliferation of tools is accompanied by a bottleneck: the representation formalisms could not be shared or combined, thus limiting the possibilities of investigating different annotation layers to the same piece of text or integrating tools in more complex text processing pipelines. As Chiarcos (2012) points out, the desire to address this bottleneck was the driving motivation for investigating interoperability.

Interoperability is a composite notion that takes into account different dimensions and levels of analysis including metadata, data, and tools. Interoperability is now a key goal of standardisation efforts (e.g., ISO TC/37 SC4) and of language resource infrastructures (e.g., META-SHARE,[3] CLARIN-ERIC,[4] European Language Grid[5]). Two macro areas of application of interoperability must be distinguished: interoperability of NLP systems, and interoperability of corpora. These two areas of applications are

[1] To facilitate replicability, the scripts used for training and evaluating the systems, the data—within the limits of licenses for redistribution—and the best models are freely available GitHub: https://github.com/tommasoc80/event_interoperability.

[2] https://lremap.elra.info/? last access 2022-07-14.

[3] http://www.meta-share.org.

[4] https://www.clarin.eu.

[5] https://www.european-language-grid.eu.

strictly connected, although in this contribution we will investigate only the latter: the interoperability of annotated corpora.

There is a consensus in distinguishing two types of interoperability when it comes to corpora: *syntactic* (sometimes referred to as structural) and *semantic* interoperability. Syntactic interoperability is defined as convergence towards a common, or pivot, formalism of annotations of different origins to allow uniform processing of different resources (Chiarcos, 2012). Syntactic interoperability aims at representing all annotations of a corpus in a way that allows their storage and querying regardless of their original annotation layer. Essentially, syntactic interoperability is seen as "the ability of systems to process exchanged data either directly or via conversion." (Calzolari et al., 2011, p. 45). A generic example of a syntactic interoperable format is XML or OWL/RDF. Standardisation efforts such as TMF (ISO 16642), SynAF (Declerck, 2006), LAF/GrAF (Ide & Suderman, 2007), and the NLP Interchange Format (NIF) (Hellmann et al., 2013) all qualify as examples of syntactic interoperability that propose and define data models that allow uniform representations of different annotation layers.

Semantic interoperability, in contrast, is more challenging and addresses the heterogeneity of linguistic annotations. While representing richness of analyses, heterogeneity of annotations is a major hurdle for reusability and, consequently, for the portability of systems. Following Ide and Pustejovsky (2010), semantic interoperability of corpora is "the ability to automatically interpret exchanged information meaningfully and accurately in order to produce useful results". The key idea behind semantic interoperability is that sharing a common vocabulary, or a *repository of linguistic annotation terminology*, across language resources is a way to enrich knowledge and exchange of information. The use of a common annotation terminology enables the resolution of the linguistic information from one corpus against the information from another corpus: "[r]eference definitions [i.e., the common terminology repository (*au.*)] provide an interlingua that allows mapping linguistic annotations from annotation scheme A to annotations in accordance with scheme B" (Chiarcos, 2012, p. 163).

This sounds great in theory, but what does it mean in practice? The creation of shared linguistic repositories, such as ISOCat (Kemps-Snijders et al., 2009; Windhouwer & Wright, 2012) among others, is not inherently free from problems since different communities develop and maintain them independently. In some cases, linguistic repositories may fail to be compatible with the definitions they provide. The rise of the Semantic Web and the adoption of the Linked Open Data (LOD) principles have contributed to establishing innovative practices and solutions for attaining semantic interoperability. This has led to the development of new community standards, recommendations, and shared vocabularies, e.g., Ontologies of Linguistic Annotation (Chiarcos & Sukhareva, 2015) for linguistic data categories, or OntoLex-Lemon (McCrae et al., 2017) for lexical resources. Using LOD principles to make resources available allows them to be always uniquely identifiable, linked to one another in a uniform way, and immediately retrieved and processed through standard Web protocols. One of the most innovative aspect of this approach is that the same formalism (OWL/RDF) is used to address simultaneously syntactic and semantic interoperability (Chiarcos, 2012; Chiarcos et al., 2013).

A complementary approach to LOD to achieve semantic interoperability is the development and use of common annotation schemes. Hajičová (2014) distinguishes two ways to instantiate this approach: (i) the application of a common scheme to different languages; or (ii) the convergence towards a single representation format, in other words, an interlingua (Witt et al., 2009), for a common phenomenon encoded by different annotation schemes. Clear examples of the first method are initiatives such as the Universal Dependency Treebank (Nivre et al., 2016, 2020), SemEval 2010 Task 13: TempEval-2 (Verhagen et al., 2010), SemEval 2015 Task 5: Aspect-based Sentiment Analysis (Pontiki et al., 2016), or SemEval 2020 Task 12: OffensEval (Zampieri et al., 2020). The second method, i.e., the convergence towards a common scheme, is a more specific definition of semantic interoperability when compared to those by Ide and Pustejovsky (2010) and Chiarcos (2012), stressing concrete ways in which the "meaningful and accurate" interpretation of information can be achieved. Examples of this latter method are usually exemplified by conversion between different representation formats (de Marneffe et al., 2006; Niu et al., 2009; O'Gorman et al., 2018). Both methods call for a vision of interoperability that directly promotes the reuse of annotated data rather than creating them from scratch. At the same time, the adoption of a common representation scheme goes a step further by directly promoting also the reuse of systems.

Directly reusing corpora and their annotations has important consequences for the development of NLP systems. Stochastic NLP approaches are data-hungry, thus having access to multiple corpora for the same phenomenon annotated in an interoperable format is a way to address this issue (Comeau et al., 2013). The promoted vision here is that more (varied) data will lead to more robust systems, and consequently, systems that are more portable across heterogeneous datasets (Tu et al., 2020). This does not necessarily guarantee that these systems have also learned to "generalise well" or, in other words, that they freed themselves from the chains of their datasets and "learned" a language phenomenon. Nevertheless, increasing the robustness of models is a necessary step to attain generalisability (Ettinger et al., 2017).

This contribution investigates interoperability of event-annotated corpora. In particular, we consider the case of event annotation promoted by ISO-TimeML (Pustejovsky et al., 2010) for English texts. ISO-TimeML represents a peculiar case for the study of semantic interoperability. First, *ISO-TimeML is a common representation format* for annotating events (an interlingua). Second, available annotation schemes and *corpora are all described as ISO-TimeML compliant*, meaning that ISO-TimeML is used as a common reference format to develop new annotation schemes and guidelines. This is the opposite of what is considered interoperability by Hajičová (2014). Third, *ISO-TimeML compliant corpora match the definition of semantic interoperablity following* (Chiarcos, 2012; Ide & Pustejovsky, 2010): corpora share a common vocabulary for the definition of the events allowing to "automatically interpret exchanged information meaningfully and accurately" (Ide & Pustejovsky, 2010). At the same time, they instantiate a case of inverted interlingua approach. In particular, the annotation schemes are derived from the interlingua (ISO-TimeML), introducing further issues concerning the compatibility and complementarity of the specific annotations.

3 Event detection: a short overview

In the previous section we presented background information on the notion of semantic interoperability and its advantages for the development of NLP systems. In this section we introduce the specific NLP task we selected to study interoperability: event detection. In particular, we explain what constitutes event detection in the context of NLP, give an overview of the most important annotated corpora, and present the state-of-the-art approaches developed for this task.

3.1 Task definition

Event detection is a complex task with a strong tradition in NLP (Ahn, 2006; Bethard, 2013; Ji & Grishman, 2008; Nguyen & Grishman, 2015; Ritter et al., 2012). The problem has been framed as follows: given a document D, identify all pairs of linguistic expressions $(w_i, w_j) \in D$, where w_i is an instance of an event trigger and w_j is an instance of an event participant. Event triggers correspond to those linguistic expressions in a document that denote the happening of something, or a state being valid (i.e., *what*) (Bach, 1986). Participants refer to the tokens (or phrases) that expresses the actors (i.e., *who* and *whom*), the location (i.e., *where*), and time of occurrence (i.e., *when*). Given this formulation, it is easy to notice that events are hubs of information that explicate complex relationships between people, places, objects, actions and states.

The identification of event triggers is a particularly challenging task: the "eventiveness" of a linguistic expression w_i is highly dependent on its context because there exists a continuum between eventive and non-eventive interpretation in the space of event semantics (Araki et al., 2018). Event participants, on the other hand, are defined as "participant roles that can be filled" (Linguistic Data Consortium, 2005) and their identification is not guided by the syntactic structure of the predicates but by semantic schemes that represent event-related scenarios.[6]

3.2 Computational approaches

Previous work on event detection has mainly adopted supervised approaches.[7]. Two major waves of systems can be identified, namely: (i) feature-based ones; and (ii) neural network architectures. Feature-based models (e.g., Support Vector Machines—SVM, Conditional Random Fields—CRF) make use of hand-crafted symbolic features combining linguistic and domain specific knowledge (Ji & Grishman, 2008; Jung & Stent, 2013; Chen & Ji, 2009; Caselli & Morante, 2018; Venugopal et al., 2014). On the other hand, neural network architectures, either based on Convolutional Neural Networks (CNN) or Recurrent Neural Network (RNN) and its extensions (e.g., Gated Recurrent Unit—GRU, or Long Short-Term Memory—LSTM), have shown their effectiveness in

[6] Indeed, event participant annotation can be compared to frame annotation (Fillmore, 2006; Rupenhofer et al., 2010) However, these two annotation frameworks are embedded in different theoretical and applied frameworks and theories.

[7] A notable exception is the Liberal Event Extraction approach (Huang et al., 2016)

reducing the dependencies of the systems on the use of toolkits and external resources for feature extraction by automatically learning features from the data[8] (Araki, 2018; Huang et al., 2018; Nguyen & Grishman, 2015; Nguyen et al., 2016).

A difference that cuts across the specific algorithm is how the task is modeled. Early systems followed a two-stage approach: first, event triggers are identified (and classified), and subsequently this information is used to predict the participants in the event. More recent approaches propose predicting event triggers and participants jointly (Li et al., 2013). The advantages of joint modeling are mainly in a reduction of error propagation across the NLP pipeline. Furthermore, since event triggers of the same type tend to co-occur with the same set of participants, joint approaches benefit from this additional information and obtain better results. For instance, on the ACE corpus, the joint model by Nguyen et al. (2016) obtains an F1 score of .693 for event trigger detection and classification and .554 for argument roles on ACE, improving against the pipeline model by Nguyen and Grishman (2015).

More recently, pre-trained language models have been successfully applied to this task (Caselli & Üstün, 2019; Yang et al., 2019) reaching new state-of-the-art results. Next to this, a new wave of systems has been proposed based on the development of transferable neural network learning techniques with a common semantic space of shared embedding representations (Huang et al., 2018). The approach, also labeled as *share-and-transfer*, first learns the extraction models over this common space, and subsequently applies it to the target data. The advantage of the method is that the learned event knowledge can be transferred, i.e., becomes available, for recognizing unseen content in low-resource settings.

3.3 Corpora

Most previous work on event detection has focused on contemporary texts covering different domains, including news articles (Pustejovsky et al., 2003; Linguistic Data Consortium, 2005; Song et al., 2015; Minard et al., 2016), (bio-)medical documents (Bethard et al., 2016, 2017), and social media (Ritter et al., 2012). Recently, event extraction has been applied also to historical texts (Sprugnoli & Tonelli, 2019). Evaluation campaigns and dedicated workshops[9] have played a big role in boosting research by promoting the availability of numerous benchmark corpora and opening discussions for annotation proposals and refinements of the definition of events. For instance, in the Message Understanding Conference (MUC)'s tasks (Sundheim, 1992; Chinchor, 1998), event detection is restricted to predetermined event instances (e.g., joint venture announcements or rocket launching) based on a scenario filling task of template elements, with specific fields roughly corresponding to the event participants. More fine-grained annotations have been proposed in the Automatic Content Extraction (ACE) campaign, TempEval (Verhagen et al., 2007, 2010; UzZaman et al., 2013)

[8] Normally, the network is initialized by making use of word embedding representations, i.e., continuous vector representations obtained from large amounts of data and that provide a semantic representation of the tokens in a sentence.

[9] *Workshop on Temporal and Spatial Information Processing* (2001); *EVENTS: Definition, Detection, Coreference, and Representation* (four editions; 2013–2016); *Computing News Storylines/Events and Storylines in the News* (four editions; 2015–2018).

and Clinical TempEval (Bethard et al., 2016, 2017), the Knowledge Base Population track at the Text Analysis Conference Know (TAC KBP),[10] and i2b2 challenge.[11]

Variations in the annotation of events affect different dimensions, the most relevant being: (i) the definition of what constitutes an event and how to annotate its linguistic realization(s); and (ii) the assignment of events to specific classes. Such differences make most of the event-labeled corpora incompatible with each other preventing the direct reuse of their data and their automatic conversion across formats. For instance, ACE restricts the annotation of events to the occurrence of eight semantic classes (e.g., *Life*, *Movement*, *Conflict*, *Business*, *Contact*, *Personnel*, *Justice*, *Transaction*) in news articles. The TAC KBP campaigns adopt an approach similar to ACE by limiting the annotation to specific semantic classes. But they introduce a new annotation scheme, Entities Relations Events (Mitamura et al., 2015; Song et al., 2016), where events are annotated on the basis *event nuggets*, i.e., "a semantically meaningful unit that expresses an event [being] either a single word (verb, noun, or adjective) or a phrase which is continuous or discontinuous" (Araki, 2018, p. 116).

The TempEval campaigns have promoted the use of TimeML (Pustejovsky et al., 2003a) which rejects any restriction on semantic classes and linguistic realizations of events. The TimeML annotation philosophy is surface-oriented. The event classes do not correspond to any semantic category but rather are based on both the lexical aspect (Vendler, 1967) and their contextual syntactic structure. The consolidation of TimeML into an ISO standard had a beneficial effect in promoting the development of annotation initiatives that share some (minimal) common elements, such as (i) a common definition of the target phenomenon (i.e., what an event is); (ii) a similar annotation philosophy (i.e., how to annotate events); and (iii) neither semantic nor morpho-syntactic restrictions,, i.e., what can realize an event (Caselli & Sprugnoli, 2017). This makes ISO-TimeML compliant corpora suitable for investigating interoperability. In particular, in this work we investigate the interoperability of three such corpora in English, namely TempEval-3 (UzZaman et al., 2013), RED (O'Gorman et al., 2016), and MEANTIME (Minard et al., 2016).

In Table 1 we report an overview of the major English corpora available for event detection illustrating their domain, definition of event, and event classes. In Table 2, we illustrate the annotation layers available in each corpus.

In the following section, we will describe in details the properties and the annotation characteristics of ISO-TimeML, and of the three compliant corpora we have selected for our study on interoperability.

4 The ISO-TimeML standard for event markup

ISO-TimeML is the International Organization for Standardization ISO/TC37 standard for event annotation (Pustejovsky et al., 2010). As it happens, the annotation of event participants is not addressed by ISO-TimeML nor by ISO-TimeML compliant corpora (Table 2). Hence, in this article, we investigate interoperability only for the event trigger

[10] https://tac.nist.gov.

[11] https://www.i2b2.org/NLP/DataSets/Main.php.

Table 1 Event annotated corpora in English—overview of text domain, event definition, and event classes

Corpus	Domain	Event definition	Event classes
MUC-7	News	Relationship between a predicate and participants	Aircrash scenarios; missile launches scenarios
ACE 2005	Newswires	Something that happens a change of state	Life, Movement, Conflict, Business, Contact, Personnel, Justice, Transaction
TempEval 1–3	News	Something that happens a state that holds as true	Occurrence, I_Action, I_State, State, Aspectual, Reporting Perception
i2b2	Clinical notes	Clinical concepts clinical departments evidentials occurrences	N/A
Clinical TempEval	Clinical notes	Anything relevant on the clinical timeline of a patient	NA, Aspectual, Evidential
TAC-KBP 2014–2017	News, Discussion Forum	Explicit occurrence of an event with or without participants of any of the selected classes	Business, Contact Manufacture, Life, Transaction, Movement Personnel, Justice
RED	News, Discussion Forum	Any occurrence, action, process or event state	N/A
MEANTIME	News	Something that can be said to obtain or hold true to happen or to occur	Grammatical, Speech_Cognitive, Other

Table 2 Event annotated corpora in English—overview of available annotation layers other than event

Corpus	Event participants	Entities	Temporal expressions	Event coreference	Temporal relations	Causal relations
MUC-7	✓	✓				
ACE 2005		✓	✓			
TempEval 1–3			✓		✓	
i2b2			✓		✓	
Clinical TempEval			✓		✓	
TAC-KBP 2014–2017	✓	✓		✓		
RED		✓	✓	✓	✓	
MEANTIME		✓	✓	✓	✓	✓

detection task. We will first give an overview of the main characteristics for event annotation following ISO-TimeML standard (Sect. 4.1). Then we show how each of the three corpora implement this standard and provide a detailed analysis of these corpora (Sect. 4.2). This serves to illustrate potential and unexpected mismatches between the annotation schemes and the actual guidelines that may affect interoperability. Finally, we provide an empirical analysis of the similarities and differences of the data distribution composing these corpora to better assess their compatibility and impact for portability and reusability (Sect. 4.3).

4.1 Event definition and annotation in ISO-TimeML

As Pustejovsky et al. (2010) highlight, ISO-TimeML distinguishes between abstract syntax and concrete syntax. The abstract syntax "specifies the elements making up the information in annotations, and how these elements may be combined to form complex annotation structures" (Pustejovsky et al., 2010, p. 394). The combinations of annotation structures are independent of any specific representation format. On the other hand, the specification of how to represent the annotation structures is delegated to the concrete syntax. While XML is used to represent concrete ISO-TimeML annotations, any representation format that is faithful to the ISO-TimeML abstract syntax can be readily converted into a corresponding ISO-TimeML concrete syntax representation, i.e., XML. The abstract syntax is the key to the interoperability of annotations across different concrete specification formats: "[t]he fact that this semantics is associated with the abstract syntax, rather than with a particular concrete syntax, explains why all concrete representations of ISO-TimeML annotations are semantically equivalent"(Pustejovsky et al., 2010, p. 394).

ISO-TimeML defines an event as anything that happens or occurs, or a state in which something obtains or holds true. This definition describes what is commonly referred to as *eventuality* (Bach, 1986). Event mentions are annotated using the tag `EVENT`. From a morpho-syntactic perspective, ISO-TimeML considers every possible realisation of an eventuality as valid, including verbs, nouns, adjectives, and (some) prepositional phrases. Every annotation of an event is also enriched by various attributes specifying the class, the tense, the grammatical aspect, the polarity (negative or positive), the presence of modal operators, and the cardinality.

ISO-TimeML also presents core normative instructions on how to annotate events. In particular, the identification of the textual extent of an event is mainly syntax driven and based on the notion of *minimal chunk*, i.e., the most meaningful component of a phrase mentioning an event. Higher constituents (e.g., a verb phrase) are discarded to avoid nesting of multiple events. In practical terms, the minimal chunk approach means that only a single token of an event mention is annotated, as the following examples demonstrate[12]:

[12] All examples are taken from the ISO-TimeML Annotation Guidelines.

(1) There is no reason why we would not be fully <EVENT eid="e1" eiid="ei1" class="OCCURRENCE" pos="ADJECTIVE" tense="NONE" aspect="NONE" polarity="NEG" modality="would"/>prepared</EVENT>.

(2) Israel has been <EVENT eid="e1" eiid="ei1" class="IACTION" pos="VERB" tense="PAST" aspect="PERFECTIVE_PROGRESSIVE" polarity="POS"/>scrambling </EVENT> to buy more masks abroad.

(3) A fresh <EVENT eid="e1" eiid="ei1" class="OCCURRENCE" pos="NOUN" tense="NONE" aspect="NONE" polarity="POS"/>flow </EVENT> of lava, gas and debris erupted there Saturday.

4.2 Annotating events in TempEval-3, RED, and MEANTIME

There are three ISO-TimeML compliant corpora for English, to wit:

- TempEval-3 (TE3) (UzZaman et al., 2013);[13]
- Richer Event Description (RED) (O'Gorman et al., 2016); and
- Meantime (MNT) (Minard et al., 2016).

In this section we review the schemes and the associated guidelines for event annotation in these three corpora. We will first have a close look at their definition of event. Then we discuss how events are annotated, and conclude by illustrating the composition of the corpora.

Event definition As already stated, ISO-TimeML adopts a broad definition of event corresponding to the notion of eventuality. As illustrated by Table 1, the corpora follow the same definition, whereby all eventualities are eligible for annotation. Furthermore, the corpora do not filter events using a predefined ontology or set of classes, like in ACE. From this perspective, the corpora are perfectly interoperable, given they have a shared vocabulary for defining an event. Sharing of vocabulary goes even further: TE3 and RED adopt the same tag (EVENT) to mark events mentions, while MNT uses a slightly different naming (EVENT_MENTION). The different naming in MNT is needed to distinguish between event mentions and co-referential event instances (see Tonelli et al. (2014) for details).

Event annotation The annotation guidelines of the corpora follow the same philosophy of adherence to the surface structure of a document.

Unexpected differences emerge on restrictions to what should not be marked as an event. TE3 excludes generic events from the annotation (see example 4).[14] RED, on the other hand, introduces a restriction concerning the "place upon a timeline"[15] of events. As a result, verbs and nouns expressing grammatical encoding to relationships (see example 5) or epistemic status (see example 6) are not annotated as events.[16] RED

[13] We used only the gold annotated portion.

[14] The example is taken from Saurı et al. (2006).

[15] https://github.com/timjogorman/RicherEventDescription/blob/master/guidelines.md#what-is-an-event.

[16] Examples 5 and 6 are taken from the RED annotation guidelines.

and MNT also allow the annotation of pronouns as events when they are co-referential with an event antecedent (see example 7[17]).

(4) Use of corporate jets for political travel is legal. [GENERIC]

(5) The earthquake struck during the parade. [GRAMMATICAL]

(6) John seems to like chocolate. [EPISTEMIC]

(7) The economic *crisis* began in 2007: **it** started with a banking crisis

When it comes to the textual extent of the event tags, all guidelines apply the notion of minimal chunk. However, additional differences emerge at this level. MNT has a more flexible application of this rule allowing the identification of multi-token events. This is restricted to very limited cases corresponding to phrasal verbs, idioms, and prepositional phrases if attested as a single entry in a reference dictionary. On the other hand, RED allows multi-token events only when they correspond to named events (e.g., *World War II*).

These differences introduce an issue concerning the compatibility of the corpora. While interoperability as *shared vocabulary* and *meaningful and accurate exchange of information* is preserved, the specific implementations of the annotation schemes (i.e., the guidelines) have introduced changes that may prevent full portability and reusability.

Table 3 illustrates the distribution of annotated event tokens across the three corpora. Here we show the number of annotated events in the train, development and test splits for each corpus, and give the distribution for the parts of speech that occur in the corpora: verb, noun, adjective, pronoun, and multi-token types. We have kept the official splits into train, dev(elopment), and test for RED only. For TE3, we have created a development section by using all test documents from the TempEval-2 evaluation campaign. No changes have been done to the TE3 official test data. The MNT corpus does not have an official split and we created one for this study. Development splits are used to optimize the training of the models. TE3 and RED annotated full documents, while MNT only considered the first five sentences of each document.

Unsurprisingly, most of the annotated events correspond to verbs. This is in line with studies in lexical semantics that identify verbs as the prototypical part of speech for the realization of events (Lyons, 1977). Nouns form the second most frequent part of speech for annotated events. The number of event nouns is higher in RED than in TE3 and MNT. In particular, around 16% of all nouns in this corpus are marked as events, while this corresponds to approximately 8% in TE3 and 12% in MNT. A similar difference affects adjectives. These differences are neither expected nor foreseen, considering the definition of event, the annotation scheme, and the annotation guidelines. They appear to be idiosyncrasies of each corpus which may result from a combination of factors such as the topics and the impact of different annotators. Although they do not affect the interoperability of the corpora, these variations in the distribution of the annotations may impact the portability of the systems and the reusability of the data. In particular, it could be the case that the use of TE3 data results in the weakest portable models on RED and MNT because of the annotation

[17] Example 7 has been taken from Tonelli et al. (2014)

Table 3 Distribution of events across part of speech and split (train, dev(elopment), and test) for three corpora of this study

Part of Speech	TE3			RED			MNT		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
Verb	8141	393	542	3467	365	415	2128	115	204
Noun	2268	124	175	2772	346	343	1111	72	110
Adjective	165	8	21	383	59	61	95	7	9
Pronoun	0	0	1	118	10	13	32	5	3
Other POS	29	1	7	274	83	14	96	8	6
Multi-token	0	0	0	36	4	0	111	7	7
Total	10,603	526	746	7050	867	846	3543	214	339

differences and event POS distributions (see Table 3), regardless of the fact that TE3 is the largest annotated corpus.

Composition of the corpora Table 4 presents a summary of the raw data of the three corpora. TE3 is the largest manually annotated corpus for event detection for its size both at document and at token levels. It is also the corpus that covers the longest time period (i.e., 24 years). This contrasts with RED and MNT whose sizes are smaller, corresponding, respectively, to around 74% and around 12% of the tokens of TE3. They also span over a shorter time period of about five years, largely overlapping. A further notable difference concerns event density with respect to the total amount of tokens. MNT is the most densely annotated corpus, with almost 30% event tokens. On the other hand, TE3 and RED have similar proportions around 11%. Although MNT limits the annotation to the first five sentences of each document (including the title), the higher density of events appears to be a peculiarity of the documents composing the corpora.

Looking at the sources of the documents, TE3 contains only news articles from established news outlets (e.g., Wall Street Journal, BBC, New York Times, Associated Press, among others). MNT has news articles written by volunteers on an online platform, *Wikinews*. Finally, RED is the only corpus that offers different text types by combining news articles and posts from online forums. Overall, there is a large overlap across the corpora in terms of text types: 93% of all documents are news articles. Nevertheless, the varied distribution of documents in time raises additional warnings concerning the compatibility of the corpora and their impact on the reusability of the data.

Table 4 Basic statistics of the three ISO-TimeML corpora in this study

Corpus	Docs	Tokens	Events	Period	Source
TE3	276	106,982	11,875	1989–2013	News
RED	95	79,561	8763	2002–2014	News; Discussion Forum
MNT	119	13,874	4096	2009–2013	(Wiki)News

Table 5 Interoperability of corpora: Jensen–Shannon divergence of train and test distributions of the three corpora)

Train	Test		
	TE3	RED	MNT
TE3	**.361**	.430	.467
RED	.402	**.357**	.503
MNT	.422	.480	**.273**

Table 6 Interoperability of corpora: normalized Out-Of-Vocabulary (OOV) rate (percentages) across train and test distributions of the three corpora

Train	Test		
	TE3 (%)	RED (%)	MNT (%)
TE3%	**32.52**	36.81	26.15
RED%	40.80	**27.19**	35.90
MNT%	64.13	60.11	**2.16**

4.3 Compatibility of data distributions

The surface level analysis presented in Sect. 4.2 indicates that we have three corpora that broadly speaking belong to the same general domain. However, additional differences due to a variety of factors may affect the data distributions and consequently could impact the portability of systems. For instance, it may be the case that one of the corpora is skewed to a series of topics limiting the portability of a trained model on another corpus without violating interoperability. To better assess the impact of such factors we analyse the corpora for their similarity and diversity (Plank & Van Noord, 2011; Liakata et al., 2012; Ruder & Plank, 2017).

Similarity and diversity Following previous work (Ruder & Plank, 2017), we investigate these aspects by means of two metrics: the Jensen–Shannon (JS) divergence and Out-of-Vocabulary rate (OOV).

The JS divergence assesses the similarity between two probability distributions, q and r. We followed the JS implementation in Ruder and Plank (2017):

$$JS(q,r) = \frac{1}{2}\left[D\left(q||\frac{1}{2}(q+r)\right) + \frac{1}{2}D\left(r||\frac{1}{2}(q+r)\right)\right] \quad (1)$$

where first Kullback–Leibler (KL) divergence is computed and then averaged between the two probability distributions. The JS divergence has been computed using the count of each token normalized over the entire vocabulary.[18] On the other hand, the OOV rate helps in assessing the differences between the corpora as it highlights the percentage of unknown tokens. Table 5 illustrates the JS divergence between train and test splits, while Table 6 shows the normalized distributions of OOV rates.

When looking at the JS divergence, MNT_{test} is the least similar to TE3_{train} and RED_{train} (Table 5). The figures for the OOV rates complement the analysis. In this case, MNT is the corpus that is more homogeneous between train and test splits (only

[18] We exclude: tokens less than 2 characters and English stopwords—as provided by NLTK.

2.16% of OOV tokens) and a potential challenging test set for TE3 and RED. Besides MNT_{test} limited size, the OOV rates with TE3_{train} and RED_{train} distributions are 26.15% and 35.90%, respectively. At the same time, we observe that more than 60% of the tokens in TE3_{test} and RED_{test} are not present in MNT_{train}.

These figures suggest that, regardless of the differences in size between TE3_{train} and RED_{train} with respect to MNT, both TE3_{train} and RED_{train} may struggle to achieve comparable performances on the MNT_{test} against a system trained on MNT_{train}. Furthermore, the OOV rate and JS score suggest that system trained on TE3_{train} would obtain better results than a system that uses RED_{train} when tested on MNT_{test}.

The comparison between TE3 and RED, on the other hand, tells a different story. In general, both corpora seem to occupy a relatively homogeneous space given their JS scores (see Table 5). In particular, the differences between the in-distribution splits and the out-of-distribution ones are not very large. The OOV rate, on the other hand, suggests that a system trained on TE3_{train} should be more portable than one trained on RED_{train}.

The main takeaways from this overview can be summarized as follows:

- semantic interoperability can be preserved at an abstract level but disregarded in the actual realization of a corpus;
- corpora may be interoperable but not necessarily compatible—even if covering the same broad domain;
- TE3, RED, and MNT present differences in the distribution of the annotations for events that were unexpected on the basis of formal checks of their respective annotation schemes and guidelines.

The fact that the TE3, RED, and MNT adopt the same definition of event and types of texts offers an optimal setting to investigate the portability of systems (Sect. 5) and the reusability of their annotations to create more robust systems (Sect. 6). The differences we have observed will help us to formulate expectations on the performance of systems, their portability, and the reuse of the annotated data.

5 Portability of systems

Previous work has investigated portability and robustness in terms of system performance under a distribution shift (Novielli et al., 2018; Hendrycks et al., 2020; Wang et al., 2022). Under this perspective, portability is closer to the notion of domain abstraction or generalization to unforeseen distribution shifts (Muandet et al., 2013; Hendrycks et al., 2020). One of the assumptions is that performance should not degrade due to (minor) differences in the data (e.g., new domain, grammatical errors, speakers' dialect) (Wang et al., 2022). The focus is mainly on the portability of a *system's architecture*, backgrounding the differences in the data. However, assessing the compatibility of the data composing the corpora is a necessary requirement to study portability (Alex et al., 2006; van Erp et al., 2016; van Son et al., 2018). In Sect. 4.2, we have identified critical issues concerning the compatibility of the three corpora mainly in terms of their annotation. In Sect. 4.3 we have highlighted similarities and differences in the data distributions and formulated expectations on the

behavior of systems. As a matter of fact, data from the same domain (e.g., news) may differ due to a variety of factors, some of which are openly identifiable (e.g., time of publication, topics, writing styles) and others yet to be enumerated (Plank, 2016; Ramponi & Plank, 2020).

To properly investigate the portability of systems, it is important to first assess the loss due to data distributions i.e., the *expected loss*. To do so, we apply methods developed to predict the performance drop of NLP systems in presence of domain-shift (Elsahar & Gallé, 2019) to the raw data of our corpora (Sect. 5.1). This will return an expected loss of the system performances across the corpora due to intrinsic differences of the data rather than due to human annotations. Only after this threshold has been identified, the portability of the systems can be assessed (Sect. 5.2). In the light of the previous analysis, we may expect that the differences in annotation and event part-of-speech distributions may negatively affect portability, besides the corpora being semantically interoperable. This means that we may register losses in performance higher than the expected loss.

For the purpose of our experiments on event trigger detection, portability, and reusability, we deemed it sufficient to use a state-of-the-art method based on a Bi-LSTM network with a CRF classifier as the last layer (Reimers & Gurevych, 2017b).[19] rather than more complex architectures such as pre-trained Language Models (PLMs). We do not fine-tune the hyper-parameters, but follow the suggestions in Reimers and Gurevych (2017a), Reimers and Gurevych (2017b) for sequence labeling tasks. The network is composed by two LSTM layers of 100 units each, trained with the *Nadam* optimizer. Variational dropout is applied,[20] with gradient normalization ($\tau = 1$), and batch size of 8. We train the models for a maximum of 30 epochs, with early stopping after 5 consecutive epochs with no improvements. Komninos and Manandhar (2016) pre-trained word-embeddings are used to initialize the network[21] and concatenated with character-level features Ma and Hovy (2016). Notice that in all experiments the development sets used during the training to optimize the losses are always from the same distributions as the train. Using an out-of-distribution development set is already a form of domain adaptation, an aspect which is out of the scope of this contribution.

5.1 Assessing the impact of data distributions

Differences in data distributions are known causes of performance drops. To assess their impact and identify an expected performance losses imputable to differences in the data rather than differences in the annotations, we run a series of experiments inspired by the idea of reverse classification accuracy (RCA) (Fan & Davidson, 2006; Zhong et al., 2010). RCA provides us with a quantitative measure of the expected loss directly comparable against the performance of a systems. RCA uses a classifier trained on a source dataset, or domain, C_s, to label data of a different dataset, or domain, B_t. The newly annotated dataset, B_t, is then used to train a new model, C'_t,

[19] https://github.com/UKPLab/ emnlp2017-bilstm-cnn-crf.

[20] Variational dropout value is (0.5, 0.5)

[21] Reimers and Gurevych (2017b) showed that the best results for this task are obtained with these word embeddings.

and evaluated against the same held-out portion of the source dataset, A_s. Originally, RCA has been proposed as a strategy to select the best performing model or dataset for a given target domain.

Following Elsahar and Gallé (2019), we frame RCA as follows: we train one classifier using a source corpus, C_s (e.g., TE3$_{train}$). We then use this classifier to annotate the train splits of the other two target corpora, B_t and D_t (e.g., RED$_{train}$ and MNT$_{train}$). We further train the same system's architecture with the newly annotated train portions, obtaining two classifiers, C'_b and C''_d. Finally, we test the performance of systems against the test split of corpus A_s (e.g., TE3$_{test}$). The difference in performance that we observe reflects the potential performance loss due to the differences in their intrinsic language varieties. The results of these experiments are illustrated in Table 7.

In all settings, we observe large drops between the in-distribution and the out-of-distribution splits—as it appears by comparing the results on the same test set along every row of Table 7. MNT_{train} is the split that suffers most when used to make predictions on $TE3_{test}$ (F1.805, −.015 points) and RED$_{test}$ (F1.750, -.137 points). In line with the JS score, MNT_{train} is quite dissimilar from TE3$_{test}$ and RED$_{test}$ and it presents the largest OOV rates. We observe that using TE3$_{train}$ as a re-annotated corpus for training new systems produces the lower losses. We confirm that RED has issues when applied to MNT$_{test}$, while being more homogeneous to TE3 on the basis of the JS score and OOV rate. Interestingly, the expected loss has a strong negative correlation (Spearman's $\rho - .885$, $p < .05$) with the OOV rates. On the other hand, the correlation with JS scores is moderate negative (Spearman's ρ -.552) but not significant ($p > .05$). Although these results cannot be considered definitive, the correlation outcome indicates that OOV rate appears as a potential stronger prediction of expected loss than JS divergence.

While the corpora are compatible for the task (i.e., event detection) and text types, we have identified an expected loss due to different linguistic properties of the corpora and mainly due to differences in their vocabulary, as supported by the correlation between OOV rates and cross-corpora results. These differences in performance can be used as lower thresholds, or expected minimal losses, when testing for the portability of systems.

5.2 Experiments and results

Our initial working hypothesis was that systems trained on interoperable resources (on the same domain) should minimise their loss when applied across interoperable (test) distributions. We have identified intrinsic differences across the corpora due to their data distribution. In addition to this, we have observed that while preserving interoperability, there are compatibility issues related to the actual annotations of the data. This finding has a non-negligible impact, since it may be the cause of extra loss in performance. The combination of all these elements forces us to reformulate our expectations in more homogeneous and uniform way, prioritizing the annotation compatibility issues. This means that for portability, we may expect the following types of behavior:

Table 7 Compatibility of data distribution

Scheme and Test	Re-annotated train corpus								
	TE3			RED			MNT		
	R	P	F1	R	P	F1	R	P	F1
TE3 → TE3	$.867_{.017}$	$.778_{.012}$	$\mathbf{.820}_{.003}$	$.864_{.008}$	$.770_{.008}$	$.81.002$	$.857_{.017}$	$.758_{.014}$	$.805_{.006}$
RED → RED	$.854_{.009}$	$.765_{.003}$	$.807_{.003}$	$.905_{.012}$	$.871_{.006}$	$\mathbf{.887}_{.003}$	$.800_{.026}$	$.707_{.014}$	$.750_{.007}$
MNT → MNT	$.947_{.002}$	$.934_{.006}$	$.940_{.003}$	$.932_{009}$	$.908_{.009}$	$.919_{007}$	$.995_{.002}$	$.991_{.004}$	$\mathbf{.993}_{.003}$

Columns represent the raw text corpus used to (re)train the systems using the three different annotation schemes. Each row reports the results on the same test split (e.g., $\texttt{TE3}_{test}$) of three different systems, each trained with the same annotation scheme (e.g., **TE3**) but different raw text data (columns TE3 RED, and MNT). All scores are averaged over 5 different runs, standard deviations are reported in subscripts

– annotation differences will trigger higher losses than the expected ones;
– systems trained with TE3_{train} will be the weakest portable ones because their annotations are the least compatible with RED and MNT;
– besides the differences in JS score and OVV, systems trained using RED_{train} will be more portable on MNT_{test} because their distribution of event part-of-speech is closer than that of TE3_{train}.

The figures in Table 8 are in line with our expectations, offering new insights on interoperability and portability of event corpora: all losses across the test distributions are bigger than the expected ones. While the losses in Table 7 range between 0.7% ($\text{RED}_{train} - \text{TE3}_{test}$) up to a maximum of 15.44% ($\text{MNT}_{train} - \text{RED}_{test}$), here we face losses ranging between 27.84% ($\text{MNT}_{train} - \text{RED}_{test}$) and 14.02% ($\text{RED}_{train} - \text{TE3}_{test}$). A more detailed analysis shows that RED_{test} is challenging for systems trained on TE3 and MNT, while TE3_{test} seems to be the "easiest". A further peculiar behavior concerns the differences across out-of-distribution data: systems trained using RED_{train} and MNT_{train} consistently tend to maximize recall while those trained with TE3_{train} maximize precision.

These results appears to be in line with our revised expectations: differences in annotations have a stronger impact on portability than data distribution. TE3 systems suffer, as expected, when applied both to RED_{test} and MNT_{test}. The lower recall scores clearly point out issues in the ability of the systems to correctly identify events. In spite of having a larger amount of annotated data, the systems perform poorly. Applying such systems on new data distributions would result in a weak event annotation (many positive cases will be missed). RED systems appear to overgeneralize the identification of events on out-of-distribution data, as indicated by the high recall. On closer inspection, we observe that on MNT_{test} these systems behave as expected, being more portable than those trained on TE3_{train}. As for MNT, systems are expected to suffer when applied on out-of-distribution data because of the limited size of the training and OOV rates. At the same time, the distribution of the annotated data indicates a behavior in line with RED systems, as shown by the results on TE3_{test} and RED_{test}.

The outcome of this first round of experiments clearly points out that *interoperability does not necessarily entail portability* of systems. The raw data indicates that each corpus is a specific variety of English and that differences in data distribution appears to have a minimal impact on system performances when the annotation scheme is not changing (see Table 7). On the other hand, while interoperability can be preserved,

Table 8 Portability of systems

Test →	TE3				RED				MNT			
Train ↓	R	P	F1	E-F1	R	P	F1	E-F1	R	P	F1	E-F1
TE3	.867	.778	**.820**	–	.575	.749	.650	.807	.671	.923	.777	.940
RED	.891	.583	.705	.814	.905	.871	**.887**	–	.816	.799	.807	.919
MNT	.898	.566	.693	.805	.673	.611	.640	.750	.995	.991	**.993**	–

Best F1 scores are in bold. All scores are averaged over 5 different runs, standard deviations have been omitted. Rows correspond to the training corpus, columns to the test corpus. Differences for the same test split can be observed per column. **E-F1** indicates the expected F1 from Table 7

unexpected differences in the distribution of the annotations plays a major role on the portability of current neural models.

6 Reusability of annotations

This section will investigate to what extent semantically interoperable corpora of the kind we are dealing with can be reused. The focus here is on direct reuse of the annotated data. By using an increased amount of training data, systems would gain access to a larger and diversified set of examples of contexts where different lexical items can realise events. This should result in better and more portable systems (Wang et al., 2022). Notice that, in comparison to previous work, interoperability does not require domain adaptation (Daumé III, 2007; Kim et al., 2016; Ruder & Plank, 2017). As shown in Sect. 4, the corpora share common characteristics that qualify them as belonging to the same domain. Additionally, we are not merging data originally annotated for different phenomena into a unified representation (Karan & Šnajder, 2018). This setting directly tests the assumption of reusability of annotated data promoted by semantic interoperability. The experiments in this section also aim at providing a different analysis for portability of models.

Following solutions proposed in previous work (Daumé III, 2007; Niu et al., 2009) the simplest and most immediate method to check the reusability of annotations is to concatenate the data and train a system. Table 9 reports the results of such an experimental setting where different concatenations of the train sets of each of the three corpora has been used and the resulting systems tested on every test distribution. For all experiments we used the same system architecture, Bi-LSTM with a CRF, as described in Sect. 5.1. To facilitate an easy comparison of the results, we repeat the scores reported in Table 8 at the bottom of Table 9.

The outcome of this set of experiments is different from our expectations on the use of semantically interoperable corpora. At the same time, these results are in line with respect to our insights on the data distributions and differences in annotation. Focusing on reusability of data only, the picture is better than that coming from Table 8. In general, the combination of data annotated for the same language phenomenon has a positive impact when applied to out-of-distribution tests. Not surprisingly, the largest improvements concern the combination of training materials that are subsequently applied to an in-distribution test. For instance, when TE3+RED is applied to TE3_{test}, the results are better than systems trained solely on RED_{train} or MNT_{train}.

This second battery of experiments, however, confirms some of the limits of interoperability as realized by these corpora. While concatenating data has a positive effect (also for the portability of systems), the differences in the annotations prevent systems from improving when compared to the in-distribution test splits. Hence, the results of our experiments question the reusability of these corpora as a strategy to improve on in-distribution results.

From a more general perspective, our experiments consistently indicate that the promise of interoperability has limits and requires a more careful formulation, at least according to the dimensions we are investigating with these corpora. As already stated, the three corpora are semantically interoperable: they share a common vocabulary and

Table 9 Reusability of annotations—concatenated training data

Test →	TE3			RED			MNT		
Train ↓	R	P	F1	R	P	F1	R	P	F1
TE3+RED	$.876_{.009}$	$.736_{.016}$	$.799_{.008}$	$.856_{.018}$	$.861_{.005}$	$.858_{.010}$	$.708_{.014}$	$.893_{.009}$	$.789_{.008}$
TE3+MNT	$.871_{.006}$	$.761_{.011}$	*.812*$_{.007}$	$.603_{.018}$	$.717_{.016}$	$.655_{.009}$	$.939_{.023}$	$.963_{.006}$	$.951_{.014}$
RED+MNT	$.915_{.008}$	$.581_{.011}$	$.710_{.006}$	$.890_{.018}$	$.870_{.016}$	*.879*$_{.006}$	$.972_{.009}$	$.963_{.002}$	*968*$_{.005}$
TE3+RED +MNT	$.887_{.010}$	$.711_{.011}$	$.789_{.003}$	$.839_{.024}$	$.857_{.010}$	$.847_{.008}$	$.901_{.011}$	$.952_{.005}$	$.926_{.007}$
TE3	$.867_{.017}$	$.778_{.012}$	**.820**$_{.003}$	$.575_{.016}$	$.749_{.017}$	$.650_{.016}$	$.671_{.020}$	$.923_{.010}$	$.777_{.011}$
RED	$.891_{.016}$	$.583_{.011}$	$.705_{.006}$	$.905_{.012}$	$.871_{.006}$	**.887**$_{.003}$	$.816_{.021}$	$.799_{.025}$	$.807_{.019}$
MNT	$.898_{.016}$	$.566_{.021}$	$.693_{.014}$	$.673_{.013}$	$.611_{.007}$	$.640_{.007}$	$.995_{.002}$	$.991_{.004}$	**.993**$_{.003}$

All scores are averaged over 5 different runs, standard deviations are reported in subscripts. Best scores per test set are in bold. Second best scores are in italics

definitions of events and they have a similar annotation philosophy. Nevertheless, it appears that **interoperability, in its dimension of independent schemes compliant to a standard, does not entail reusability**. As far as the combination of these corpora can enrich each other, as suggested in Table 9, and can provide a large collection of data for the linguistic investigation of events in English, they marginally contribute to the creation of more robust models.

7 Exploring the impact of annotation compatibility

Our analyses have shown that corpora can be semantically interoperable and yet they may fail to fulfill expectations concerning portability and reusability. We have identified that one important factor is represented by differences in the actual annotations of the data. In this section we will perform an error analysis on the different systems as a proxy to shed light on the actual annotations of the data and their impact for portability and reusability.

For each corpus, we use the best system and apply it to the development splits for both in- and out-of-distribution settings. We exclude the concatenated models, e.g., TE3+RED, because they would prevent the understanding of the specific features of each composing corpus (TE3, RED, and MNT) and their impact on reusability.

We focus the analysis on the errors per part of speech, as they are more relevant to understand the behaviors of the different systems in light of the data presented in Table 3 in Sect. 4.2. In the case of the false positives (FPs), the TE3 system tends to label more often verbs rather than nouns as events. This results in 62.22% of all FPs on TE3_{dev} to be verbs, followed by nouns (32.59%), adjectives (2.96%), and other parts of speech (2.22%). The contrary can be observed for the system trained on RED, where 61.90% of the FPs are nouns, while verbs are only 32.53%. Adjectives and other POS, on the other hand, are comparable with TE3. The MNT system, finally, distributes its FPs on adjectives (66.66%) and other parts of speech (33.33%) only. However, the limited amount of training data and the high similarity between the train and test distribution makes this quite a special corpus for the analysis of in-distribution errors. The analysis of the false negatives (FNs) is more in line with previous error analyses of systems for event detection (Mirza & Tonelli, 2014; Caselli & Morante, 2018). All systems struggle to correctly identify event nouns although with different proportions: 59.52% for TE3, 46% for RED, and 50% for MNT.

The analysis of the errors in the out-of-distribution data is complementary and helps to better understand the differences in precision and recall. Regardless of the target corpus, the RED system tends to predict nouns as events: 66.66% of FPs on TE3_{dev} and 78.12% of FPs on MNT_{dev}. The same system fails more often to detect verbal events: 73.33% of FNs on TE3_{dev} and 44.89% MNT_{dev}. When systems trained on TE3 and MNT are applied on RED_{dev}, the distributions of the errors is swapped: the majority of FPs are verbs (73.48% for TE3 and 56.15% for MNT, respectively) while the majority of FNs are nouns (63.48% for TE3 and 62.63 for MNT).

The behavior of the systems is a direct consequence of the distribution of the annotations in the respective train corpora. Although smaller in size, the differences in the amount of event nouns and verbs in RED_{train} are statistically significant (chi-

square test, $p < .05$) when compared to TE3_{train} and MNT_{train}. We thus explore the performance of the best trained models per corpus only on events realised by verbs and nouns. The results of this extra evaluation setting will help to better understand whether errors can be restricted to differences in the annotations of events. Figures are reported in Table 10.

In absolute terms, identifying nominal events appears as a more challenging task than identifying verbal events. This difference is best understood when framed within the notion of continuum of eventiveness of lexical items (Araki, 2018) and adopting recent theoretical frameworks of word classes as continua (Simone, 2000; Sasse, 2001; Simone, 2003). Nevertheless, performance issues are still at play in the out-of-distribution setting, with RED_{test} being the most challenging. When looking at the performance drops of the systems, none of them can actually be considered minimal nor in line with expected loss due to language variety. However, it appears that **portability of systems (and potentially of annotations) is subject to specific parts of speech**.

On the basis of the annotation scheme and guidelines, the impact of event part of speech on the systems' performance is not expected. In no part of the annotation guidelines of each corpus is stated that event nouns are to be treated differently. This opens questions about the annotation process and the quality of the corpora. Previous work (Derczynski & Gaizauskas, 2010) identified inconsistencies in the annotation of the TimeBank corpus (Pustejovsky et al., 2003) that were supposed to be addressed for the release of the TE3 corpus (UzZaman et al., 2013). Another contribution (Inel & Aroyo, 2019) investigated the consistency and completeness of TE3 for event annotation and time expressions. Their analysis show that the TE3 corpus still contains annotation inconsistencies such as not annotated sentences and missed event tokens.

We assess the differences in event annotations using the average of the ratio between the frequency of each type annotated as event an event ($freq(ev_{wi})$, and its overall frequency (regardless of the event annotation) ($freq(w_i)$) in the corpus distribution (C). We call this measure *average event ambiguity* (AEA) and it can be expressed as follows:

Table 10 Verbal and nominal event evaluation—in- and cross-distribution data for best model only

Train ↓	Test →								
	TE3			RED			MNT		
	R	P	F1	R	P	F1	R	P	F1
Verbs									
TE3	.943	.809	*.873*	.862	.739	.796	.811	.942	.872
RED	.902	.808	.852	.949	.889	*.918*	.856	.945	.898
MNT	.977	.687	.807	.903	.616	.733	.965	.955	*.960*
Nouns									
TE3	.657	.680	*.668*	.271	.744	.397	.490	.898	.634
RED	.908	.347	.503	.863	.836	*.894*	.888	.644	.747
MNT	.805	.413	.546	.495	.656	.564	.990	.981	*.986*

Best score per POS and trained model are in italics

Table 11 Average event ambiguity (AEA) for TE_{train}, RED_{train}, and MNT_{train}

Corpus	AEA-overall	AEA-verbs	AEA-nouns
TE	.813	.919	.668
RED	.833	.910	.837
MNT	.902	.966	.884

$$AEA(C) = avg\left(\sum_{i=1}^{n} \frac{freq(ev_{wi})}{freq(w_i)}\right). \quad (2)$$

AEA can have multiple functions. First, it can be used to assess the diversity of "eventiveness" of different tokens. Consider cases of deverbal nouns like "building" or complex types (Pustejovsky, 1995) like "eulogy": for these nouns not all of their occurrences give rise to eventive readings.[22] Second, it gives an estimate of the density of events, i.e., how many times a type is marked as an event with respect to all of its occurrences. Third, it can facilitate the identification of potential inconsistencies in the manual annotation. In general, the nearer to 1.0 AEA is, the less the variety of eventiveness, the higher the density, and the less the inconsistencies are. We report AEA scores in Table 11 for the train splits of the corpora and for verbs and nouns.

The outcome of AEA further confirms that there are remarkable differences in the annotations of event nouns across the three corpora. This helps to explain the unexpected results for precision and recall for the trained systems when applied out-of-distribution. The boost in recall we observe for the RED and MNT systems is definitely due to a more consistent annotation of nominal events. Furthermore, we can safely claim that the lower performance of MNT are mainly due to size rather than the annotation of the data.

As a final check, we manually annotated a random portion of 20% of the FP nouns predicted by RED on $TE3_{test}$ using the TimeML v1.2 annotation guidelines. It turned out that 54.40% of them could reliably be annotated as instances of events.

Besides their being semantically interoperable, **the differences in the actual annotations** of the data in the three corpora *limit the potentially positive effects of using interoperable corpora*.

8 Conclusion and future directions

In this paper we have tested for the first time to what extent the promise of reusability promoted by interoperability of language resources holds when applied to semantically interoperable corpora for events. In particular, we have investigated a dimension of interoperability where resources share a common vocabulary based on a standard but implement independent schemes and guidelines.

We controlled for factors that may negatively influence the outcome of such analysis, namely differences in text types and language variety of the corpora involved. We have

[22] Consider the following examples were an event reading is not plausible: (a) "The building is beautiful"; (b) "The lunch was delicious".

shown that the three corpora are composed by homogeneous text types (Sect. 4) and that differences in language variety can be used as expected, lower bound thresholds that negatively impacts the performance of systems (Sect. 5.1).

We have conducted an extensive set of analyses and experiments showing the limits of interoperability based on a shared common vocabulary for portability and reusability. Differences in annotations (Sect. 7) and the quality of the annotations—see Derczynski and Gaizauskas (2010) and Inel and Aroyo (2019)—appear as the major factors that impacts the reuse of interoperable resources. Annotation quality cannot be estimated only on the basis of annotation guidelines or reported IAA scores. The TimeBank Corpus, on which TE3 is derived, reports high IAA scores for event annotation. And so do RED and MNT. We propose to use measures like AEA to explore at a deeper level of analysis potential differences in annotations that do not emerge by an analysis of the annotation schemes and guidelines.

Even complex system architectures such as neural networks are very sensitive to the composition of the training data, as illustrated by unsatisfying performances when reusing systems trained on TE3 in out-of-distribution splits. Interoperability, and consequently reusability, seems best achieved by applying the same annotation scheme (e.g., Universal Dependencies) rather than by creating standard compliant schemes, as in our case.

A further aspect that cannot be ignored concerns the peculiarity of the event detection task. Events are complex entities that are at the heart of the syntax-semantic interface. The semantics of an event in some cases requires access to knowledge that may not be explicitly expressed in texts. This is a further factor that may have played a role in the outcome of our experiments. A natural follow-up question for future work would require the creation of new test sets on multiple and different data distributions against which all systems that have been developed in this work can be further tested. This may help to better understand the potential advantages of semantic interoperable resources.

In our closing remarks we want to clarify our stance on resource interoperability. We believe that standardisation initiatives such ISO-TimeML and UD represent valuable contributions for computational linguistics and natural language processing. They promote discussion and advancements for the analysis and documentation of language phenomena, because they create an environment that can be used by multiple communities potentially triggering new contributions and cross-fertilization across disciplines. Being able to query, find, and observe the same information in different datasets (and potentially different languages) is useful to expand knowledge of language phenomena, refine theoretical frameworks, and help develop more robust systems. In the end, to make substantial progress in the field we need more semantically interoperable resources of higher-quality.

References

Ahn, D. (2006). The stages of event extraction. In: *Proceedings of the workshop on annotating and reasoning about time and events, association for computational linguistics*, (pp 1–8).

Alex, B., Nissim, M., & Grover, C. (2006). The impact of annotation on the performance of protein tagging in biomedical text. In: *Proceedings of the Fifth international conference on language resources and evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy*, http://www.lrec-conf.org/proceedings/lrec2006/pdf/398_pdf.pdf

Araki, J. (2018). Extraction of event structures from text. PhD thesis

Araki, J., Mulaffer, L., Pandian, A., Yamakawa, Y., Oflazer, K., & Mitamura, T. (2018). Interoperable annotation of events and event relations across domains. In: *Proceedings 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation, Association for Computational Linguistics*, (pp 10–20). http://aclweb.org/anthology/W18-4702

Bach, E. (1986). The algebra of events. *Linguistics and Philosophy, 9*(1), 5–16.

Bethard, S. (2013). Cleartk-timeml: a minimalist approach to tempeval 2013. In: *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, (vol 2, pp 10–14).

Bethard, S., Savova, G., Chen, WT., Derczynski, L., Pustejovsky, J., & Verhagen, M. (2016). Semeval-2016 task 12: Clinical tempeval. In: *Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), association for computational linguistics, San Diego, California*.

Bethard, S., Savova, G., Palmer, M., & Pustejovsky, J. (2017). Semeval-2017 task 12: Clinical tempeval. In: *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017), association for computational linguistics, Vancouver, Canada*, (pp 565–572) http://www.aclweb.org/anthology/S17-2093

Blitzer, J., McDonald, R., & Pereira, F. (2006). Domain adaptation with structural correspondence learning. In: *Proceedings of the 2006 conference on empirical methods in natural language processing, Association for Computational Linguistics*, (pp 120–128).

Calzolari, N., Monachini, M., & Quochi, V. (2011). Interoperability framework: The FLaReNet action plan proposal. In: *Proceedings of the workshop on language resources, technology and services in the sharing paradigm, Asian Federation of Natural Language Processing, Chiang Mai, Thailand*, (pp 41–49) https://www.aclweb.org/anthology/W11-3306

Caselli, T., & Morante, R. (2018). Systems' agreements and disagreements in temporal processing: An extensive error analysis of the tempeval-3 task. In: chair) NCC, Choukri K, Cieri C, Declerck T, Goggi S, Hasida K, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, Tokunaga T (eds) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Paris, France

Caselli, T., & Sprugnoli, R. (2017). It-timeml and the ita-timebank: Language specific adaptations for temporal annotation. In: Pustejovsky J, Ide N (eds) Handbook of Linguistic Annotation - Vol. II, Springer, pp 969–988

Caselli, T., & Üstün, A. (2019). There and back again: Cross-lingual transfer learning for event detection. In: Proceedings of the Sixth Italian Conference on Computational Linguistics, CEUR Workshop Proceedings (CEUR-WS.org), vol 2481

Chen, Z., & Ji, H. (2009). Language specific issue and feature exploration in Chinese event extraction. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, Association for Computational Linguistics, Boulder, Colorado, pp 209–212, https://www.aclweb.org/anthology/N09-2053

Chiarcos, C. (2012). Interoperability of corpora and annotations. In: Linked Data in Linguistics, Springer, pp 161–179

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

Chiarcos, C., & Sukhareva, M. (2015). OLiA-Ontologies of linguistic annotation. *Semantic Web, 6*(4), 379–386.

Chinchor, NA. (1998). Overview of MUC-7. In: Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29–May 1, 1998, https://www.aclweb.org/anthology/M98-1001

Comeau, DC., Islamaj Doğan, R., Ciccarese, P., Cohen, KB., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F., Torii, M., & et al. (2013). Bioc: a minimalist approach to interoperability for biomedical text processing. Database 2013

Daumé, H., III. (2007). Frustratingly easy domain adaptation. *ACL, 2007*, 256.

de Jong, F., Maegaard, B., Fišer, D., Van Uytvanck, D., Witt, A., & et al. (2020). Interoperability in an infrastructure enabling multidisciplinary research: The case of clarin. In: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association (ELRA), pp 3406–3413

de Marneffe, MC., MacCartney, B., & Manning, CD. (2006). Generating typed dependency parses from phrase structure parses. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy, http://www.lrec-conf.org/proceedings/lrec2006/pdf/440_pdf.pdf

Declerck, T. (2006). SynAF: Towards a standard for syntactic annotation. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy, http://www.lrec-conf.org/proceedings/lrec2006/pdf/652_pdf.pdf

Derczynski, L., & Gaizauskas, R. (2010). Analysing temporally annotated corpora with CAVaT. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, http://www.lrec-conf.org/proceedings/lrec2010/pdf/546_Paper.pdf

Elsahar, H., & Gallé, M. (2019). To annotate or not? predicting performance drop under domain shift. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, pp 2163–2173, https://doi.org/10.18653/v1/D19-1222,

Ettinger, A., Rao, S., Daumé III, H., & Bender, EM. (2017). Towards linguistically generalizable nlp systems: A workshop and shared task. In: Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems, pp 1–10

Fan, W., & Davidson, I. (2006). Reverse testing: an efficient framework to select amongst classifiers under sample selection bias. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 147–156

Fillmore, C. (2006). Frame semantics. In: Geeraerts D (ed) Cognitive Linguistics: Basic Readings, De Gruyter Mouton, Berlin, Boston, pp 373–400, originally published in 1982.

Hajičová, E. (2014). "three dimensions of the so-called ""interoperability"" of annotation schemes". In: Chair) NCC, Choukri K, Declerck T, Loftsson H, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland

Hellmann, S., Lehmann, J., Auer, S., & Brümmer, M. (2013). Integrating nlp using linked data. In: International semantic web conference, Springer, pp 98–113

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., & Song, D. (2020). Pretrained transformers improve out-of-distribution robustness. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, pp 2744–2751, https://doi.org/10.18653/v1/2020.acl-main.244

Huang, L., Cassidy, T., Feng, X., Ji, H., Voss, CR., Han, J., & Sil, A. (2016). Liberal event extraction and event schema induction. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol 1, pp 258–268

Huang, L., Ji, H., Cho, K., Dagan, I., Riedel, S., & Voss, C. (2018). Zero-shot transfer learning for event extraction. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, pp 2160–2170, https://doi.org/10.18653/v1/P18-1201,

Ide, N., & Pustejovsky, J. (2010). What does interoperability mean, anyway? toward an operational definition of interoperability for language technology. In: Proceedings of the Second International Conference on Global Interoperability for Language Resources. Hong Kong, China

Ide, N., & Suderman, K. (2007). Graf: A graph-based format for linguistic annotations. In: proceedings of the Linguistic Annotation Workshop, pp 1–8

Inel, O., & Aroyo, L. (2019). Validation Methodology for Expert-Annotated Datasets: Event Annotation Case Study. In: Eskevich M, de Melo G, Fäth C, McCrae JP, Buitelaar P, Chiarcos C, Klimek B, Dojchinovski M (eds) 2nd Conference on Language, Data and Knowledge (LDK 2019), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, OpenAccess Series in Informatics (OASIcs), vol 70, pp 12:1–12:15, https://doi.org/10.4230/OASIcs.LDK.2019.12,

Ji, H., & Grishman, R. (2008). Refining event extraction through cross-document inference. Proceedings of ACL-08: HLT pp 254–262

Jung, H., & Stent, A. (2013). Att1: Temporal annotation using big windows and rich syntactic and semantic features. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Association for Computational Linguistics, Atlanta, Georgia, USA, pp 20–24, http://www.aclweb.org/anthology/S13-2004

Karan, M., & Šnajder, J. (2018). Cross-domain detection of abusive language online. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Association for Computational Linguistics, Brussels, Belgium, pp 132–137, https://doi.org/10.18653/v1/W18-5117,

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., & Wright, S. E. (2009). Isocat: remodelling metadata for language resources. *International Journal of Metadata, Semantics and Ontologies, 4*(4), 261–276.

Kim, YB., Stratos, K., & Sarikaya, R. (2016). Frustratingly easy neural domain adaptation. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, pp 387–396, https://www.aclweb.org/anthology/C16-1038

Komninos, A., & Manandhar, S. (2016). Dependency based embeddings for sentence classification tasks. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 1490–1500

Li, Q., Ji, H., & Huang, L. (2013). Joint event extraction via structured prediction with global features. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 73–82

Liakata, M., Saha, S., Dobnik, S., Batchelor, C., & Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. Bioinformatics 28(7):991–1000, https://doi.org/10.1093/bioinformatics/bts071,

Linguistic Data Consortium. (2005). ACE (Automatic Content Extraction) English Annotation Guidelines for Events. 5th edn

Lyons, J. (1977). Semantics: Volume 2, vol 2. Cambridge University Press

Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint arXiv:1603.01354

Ma, J., Zhang, Y., & Zhu, J. (2014). Tagging the web: Building a robust web tagger with neural network. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol 1, pp 144–154

McCrae, JP., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). The Ontolex-Lemon model: development and applications. In: Proceedings of eLex 2017 conference, pp 19–21

Minard, AL., Speranza, M., Urizar, R., Altuna, B., van Erp, M., Schoen, A., & van Son, C. (2016). Meantime, the newsreader multilingual event and time corpus. In: Chair) NCC, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Paris, France

Mirza, P., & Tonelli, S. (2014). Classifying temporal relations with simple features. *EACL, 14*, 308–317.

Mitamura, T., Yamakawa, Y., Holm, S., Song, Z., Bies, A., Kulick, S., & Strassel, S. (2015). Event nugget annotation: Processes and issues. In: Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, Association for Computational Linguistics, Denver, Colorado, pp 66–76, https://doi.org/10.3115/v1/W15-0809,

Muandet, K., Balduzzi, D., & Schölkopf, B. (2013). Domain generalization via invariant feature representation. In: International Conference on Machine Learning, PMLR, pp 10–18

Nguyen, TH., & Grishman, R. (2015). Event detection and domain adaptation with convolutional neural networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Lin-

guistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), vol 2, pp 365–371

Nguyen, TH., Cho, K., & Grishman, R. (2016), Joint event extraction via recurrent neural networks. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, pp 300–309, https://doi.org/10.18653/v1/N16-1034,

Niu, ZY., Wang, H., & Wu, H. (2009). Exploiting heterogeneous treebanks for parsing. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Association for Computational Linguistics, Suntec, Singapore, pp 46–54, https://www.aclweb.org/anthology/P09-1006

Nivre, J., de Marneffe, MC., Ginter, F., Goldberg, Y., Hajič, J., Manning, CD., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, pp 1659–1666, https://www.aclweb.org/anthology/L16-1262

Nivre, J., de Marneffe, MC., Ginter, F., Hajič, J., Manning, CD., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, pp 4034–4043, https://www.aclweb.org/anthology/2020.lrec-1.497

Novielli, N., Girardi, D., & Lanubile, F. (2018). A benchmark study on sentiment analysis for software engineering research. In: 2018 IEEE/ACM 15th International Conference on Mining Software Repositories (MSR), IEEE, pp 364–375

O'Gorman, T., Pradhan, S., Palmer, M., Bonn, J., Conger, K., & Gung, J. (2018). The new Propbank: Aligning Propbank with AMR through POS unification. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, https://www.aclweb.org/anthology/L18-1231

O'Gorman, T., Wright-Bettner, K., & Palmer, M. (2016). Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In: Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016), Association for Computational Linguistics, Austin, Texas, pp 47–56, http://aclweb.org/anthology/W16-5706

Plank, B. (2016). What to do about non-standard (or non-canonical) language in nlp. arXiv preprint arXiv:1608.07836

Plank, B., & Van Noord, G. (2011). Effective measures of domain similarity for parsing. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics, pp 1566–1576

Poch, M., & Bel Rafecas, N. (2011). Interoperability and technology for a language resources factory. In: Proceedings of Workshop on Language Resources, Technology and Services in the Sharing Paradigm; 2011 Nov 12; Chiang Mai, Thailand. Chiang Mai (Thailand): ACL; 2011. p. 32-40., ACL (Association for Computational Linguistics)

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, SM., & Eryiğit, G. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, San Diego, California, pp 19–30, https://doi.org/10.18653/v1/S16-1002,

Pustejovsky, J., Castao, J., Ingria, R., Saurì, R., Gaizauskas, R., Setzer, A., & Katz, G. (2003a). TimeML: Robust Specification of Event and Temporal Expressions in Text. In: Fifth International Workshop on Computational Semantics (IWCS-5)

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., & et al. (2003). The TimeBank corpus. In: Corpus linguistics, vol 2003, p 40

Pustejovsky, J., Lee, K., Bunt, H., & Romary, L. (2010). ISO-TimeML: An international standard for semantic annotation. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, http://www.lrec-conf.org/proceedings/lrec2010/pdf/55_Paper.pdf

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press.

Ramponi, A., & Plank, B. (2020). Neural unsupervised domain adaptation in NLP—A survey. In: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), pp 6838–6855, https://doi.org/10.18653/v1/2020.coling-main.603, https://www.aclweb.org/anthology/2020.coling-main.603

Rehm, G., Galanis, D., Labropoulou, P., Piperidis, S., Welß, M., Usbeck, R., Köhler, J., Deligiannis, M., Gkirtzou, K., Fischer, J., & et al. (2020) Towards an interoperable ecosystem of ai and lt platforms: A roadmap for the implementation of different levels of interoperability. In: Proceedings of the 1st International Workshop on Language Technology Platforms, pp 96–107

Reimers, N., & Gurevych, I. (2017a). Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. CoRR abs/1707.06799, arxiv:1707.06799,

Reimers, N., & Gurevych, I. (2017b). Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, pp 338–348, https://www.aclweb.org/anthology/D17-1035

Ritter, A., Etzioni, O., Clark, S., & et al. (2012). Open domain event extraction from twitter. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 1104–1112

Ruder, S., & Plank, B. (2017). Learning to select data for transfer learning with bayesian optimization. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, pp 372–382, https://www.aclweb.org/anthology/D17-1038

Ruder, S., Ghaffari, P., & Breslin, JG. (2017). Data selection strategies for multi-domain sentiment analysis. arXiv preprint arXiv:1702.02426

Rupenhofer, J., Ellsworth, M., Petruck, MRL., Johnson, CR., & Schefczyk, J. (2010). FrameNet II: Extended theory and practice. https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf

Sasse, H. J. (2001). Scales between nouniness and verbiness. *Language typology and language universals, 1*, 495–509.

Saurı, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., & Pustejovsky, J. (2006). Timeml annotation guidelines version 1.2. 1

Simone, R. (2003). Masdar,'ismu al-marrati et la frontière verbe/nom. In: Estudios ofrecidos al Profesor José Jesús de Bustos Tovar, Servicio de Publicaciones, pp 901–918

Simone, R. (2000). Cycles lexicaux. *Studi italiani di linguistica teorica ed applicata, 29*(2), 259–287.

Song, Z., Bies, A., Strassel, S., Ellis, J., Mitamura, T., Dang, HT., Yamakawa, Y., & Holm, S. (2016). Event nugget and event coreference annotation. In: Proceedings of the Fourth Workshop on Events, Association for Computational Linguistics, San Diego, California, pp 37–45, https://doi.org/10.18653/v1/W16-1005,

Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., & Ma, X. (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, Association for Computational Linguistics, Denver, Colorado, pp 89–98, https://doi.org/10.3115/v1/W15-0812,

Sprugnoli, R., & Tonelli, S. (2019). Novel event detection and classification for historical texts. Computational Linguistics 45(2):229–265, https://doi.org/10.1162/coli_a_00347, https://www.aclweb.org/anthology/J19-2002

Stymne, S., de Lhoneux, M., Smith, A., & Nivre, J. (2018). Parser training with heterogeneous treebanks. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, pp 619–625, https://doi.org/10.18653/v1/P18-2098,

Sundheim, BM. (1992). Overview of the fourth Message Understanding Evaluation and Conference. In: Fourth Message Uunderstanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992, https://www.aclweb.org/anthology/M92-1001

Tonelli, S., Sprugnoli, R., Speranza, M., & Minard, AL. (2014). Newsreader guidelines for annotation at document level. Technical Re- port NWR2014-2-2, Fondazione Bruno Kessler

Tu, L., Lalwani, G., Gella, S., & He, H. (2020). An empirical study on robustness to spurious correlations using pre-trained language models. Transactions of the Association for Computational Linguistics 8:621–633, https://doi.org/10.1162/tacl_a_00335, https://aclanthology.org/2020.tacl-1.40

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., & Pustejovsky, J. (2013). Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In: Second Joint

Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Association for Computational Linguistics, Atlanta, Georgia, USA, pp 1–9, http://www.aclweb.org/anthology/S13-2001

van Erp, M., Mendes, P., Paulheim, H., Ilievski, F., Plu, J., Rizzo, G., & Waitelonis, J. (2016). Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In: Chair) NCC, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Paris, France

van Son, C., Inel, O., Morante, R., Aroyo, L., & Vossen, P. (2018). Resource interoperability for sustainable benchmarking: The case of events. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.

Vendler, Z. (1967). *Causal relations. The Journal of philosophy, 64*(21), 704–713.

Venugopal D, Chen C, Gogate V, & Ng V (2014) Relieving the computational bottleneck: Joint inference for event extraction with high-dimensional features. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, pp 831–843, https://doi.org/10.3115/v1/D14-1090,

Verhagen M, Gaizauskas R, Schilder F, Hepple M, Katz G, & Pustejovsky J (2007) Semeval-2007 task 15: Tempeval temporal relation identification. In: Proceedings of the 4th International Workshop on Semantic Evaluations, Association for Computational Linguistics, pp 75–80

Verhagen M, Sauri R, Caselli T, & Pustejovsky J (2010) Semeval-2010 task 13: Tempeval-2. In: Proceedings of the 5th international workshop on semantic evaluation, Association for Computational Linguistics, pp 57–62

Wang, X., Wang, H., & Yang, D. (2022). Measure and improve robustness in NLP models: A survey. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, pp 4569–4586, https://aclanthology.org/2022.naacl-main.339

Wang, T., Wang, X., Qin, Y., Packer, B., Li, K., Chen, J., Beutel, A., & Chi, E. (2020). CAT-gen: Improving robustness in NLP models via controlled adversarial text generation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, pp 5141–5146, https://doi.org/10.18653/v1/2020.emnlp-main.417,

Windhouwer, M., & Wright, SE. (2012). Linking to linguistic data categories in isocat. In: Linked Data in Linguistics, Springer, pp 99–107

Witt, A., Heid, U., Sasaki, F., & Sérasset, G. (2009). Multilingual language resources and interoperability. *Language Resources and Evaluation, 43*(1), 1–14.

Wu, F., & Huang, Y. (2016). Sentiment domain adaptation with multiple sources. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol 1, pp 301–310

Yang, S., Feng, D., Qiao, L., Kan, Z., & Li, D. (2019). Exploring pre-trained language models for event extraction and generation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, pp 5284–5294, https://doi.org/10.18653/v1/P19-1522,

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, Ç. (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), pp 1425–1447, https://www.aclweb.org/anthology/2020.semeval-1.188

Zhong, E., Fan, W., Yang, Q., Verscheure, O., & Ren, J. (2010). Cross validation framework to choose amongst models and datasets for transfer learning. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, pp 547–562