

# Sunken Ships Shan't Sail: Ontology Design for Reconstructing Events in the Dutch East India Company Archives

Stella Verkijk<sup>1,2,\*</sup>, Piek Vossen<sup>1</sup>

<sup>1</sup>*Department of Language, Literature and Communication, Vrije Universiteit Amsterdam, The Netherlands*

<sup>2</sup>*Huygens Institute, The Netherlands*

## Abstract

This short paper describes ongoing work on the design of an event ontology that supports state-of-the-art event extraction in the archives of the Dutch East India Company (VOC). The ontology models Dynamic Events (actions or processes) and Static Events (states). By modelling the transition of a given to a new state as a logical implication that can be inferred automatically from the occurrence of a Dynamic Event, the ontology supports implied information extraction. It also considers implied sub-event detection and models event arguments as coreferential between event classes where possible. By doing so, it enables the extraction of much more information than is only explicitly stated in the archival texts with minimal annotation effort. We define this complete event extraction task that adopts both Natural Language Processing techniques as well as reasoning components as Event Reconstruction. The Event Reconstruction module will be embedded in a search interface that facilitates historical research in the VOC archives.

## Keywords

Natural Language Processing, event modelling, computational history, ontology design, reasoning

## 1. Introduction

The Dutch East India Company (VOC) played a major and in many ways controversial role in Asian history. Researching the practices of the VOC can deepen our understanding of the processes of early globalization and colonialism in the seventeenth and eighteenth centuries. Fortunately, the VOC Archives contain a wealth of detailed descriptions of the company's day-to-day activities in the Indian Ocean World. However, it has been extremely challenging to conduct focused research with this corpus, as it consists of over twenty-five million pages of handwritten Early Modern Dutch, with little metadata available.

We define Event Reconstruction (ER) as a task in Natural Language Processing (NLP) consisting of automatically extracting events described explicitly or implicitly in free text, using reasoning to provide a more complete representation of the events described. The first step in ER is extracting explicit information: detecting explicit event mentions, classifying event types

---

*CHR 2023: Computational Humanities Research Conference, December 6 – 8, 2023, Paris, France*


\*Corresponding author.

✉ s.verkijk@vu.nl (S. Verkijk); p.t.j.m.vossen@vu.nl (P. Vossen)

🆔 0009-0000-9263-2272 (S. Verkijk)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

and detecting and classifying event arguments (who, what, where, etc.). The second step involves the classification of latent events denoted by verbs, nominal expressions or nouns (e.g., *friendship* evokes a *relationship* event): a type of implicit information extraction. Finally, the last step of ER deals with a deeper level of implied information extraction, where we extract implied events that cannot be anchored directly to a mention in the text, but are the logical consequence of a mentioned event. In our case this entails logical reasoning using an ontology to extract implied *states*. Obvious consequences or assumptions are often not reported explicitly but need to be inferred to reconstruct a complete story (a mention of one geopolitical entity losing a battle from another implies they were at war). Other information that is often implied in text that we also aim to make explicit is how multiple smaller events (*subevents*) sometimes implicitly describe a relevant supra-event (e.g., when the same ships *leave* a harbour and *arrive* at a different harbour at the same time these subevents describe a *voyage* event).

Figure 1 illustrates these three levels of information extraction applied on a sample sentence of our corpus.<sup>1</sup> It shows how the custom ontology enables the modeling of ship voyages and their locations over large time frames. The aim of the ontology presented in this paper is thus to allow inferencing of states or *Static Events* (such as *being at a location* or *being in conflict*) from extracted *Dynamic Events* stated explicitly or implicitly in the text (such as *leaving a location* or *attacking*). This type of common sense reasoning can be classified as causal reasoning: inferring logical implications of events.

We provide a taxonomy of event classes defined by specialist historians, covering a wide range of researchable topics in the VOC archives. It serves as the basis for a more elaborate event ontology that enables the detection of both explicit and implicit information needed for event reconstruction (currently under development).<sup>2</sup> To our knowledge, this paper describes the first event ontology designed for a source dating as far back as the Early Modern period. We describe a project that applies state-of-the-art automatic inferencing for the humanities, presenting the first ontology made for the historical field that supports reasoning with implicit information from unstructured text for comprehensive event reconstruction.

## 2. Related Work

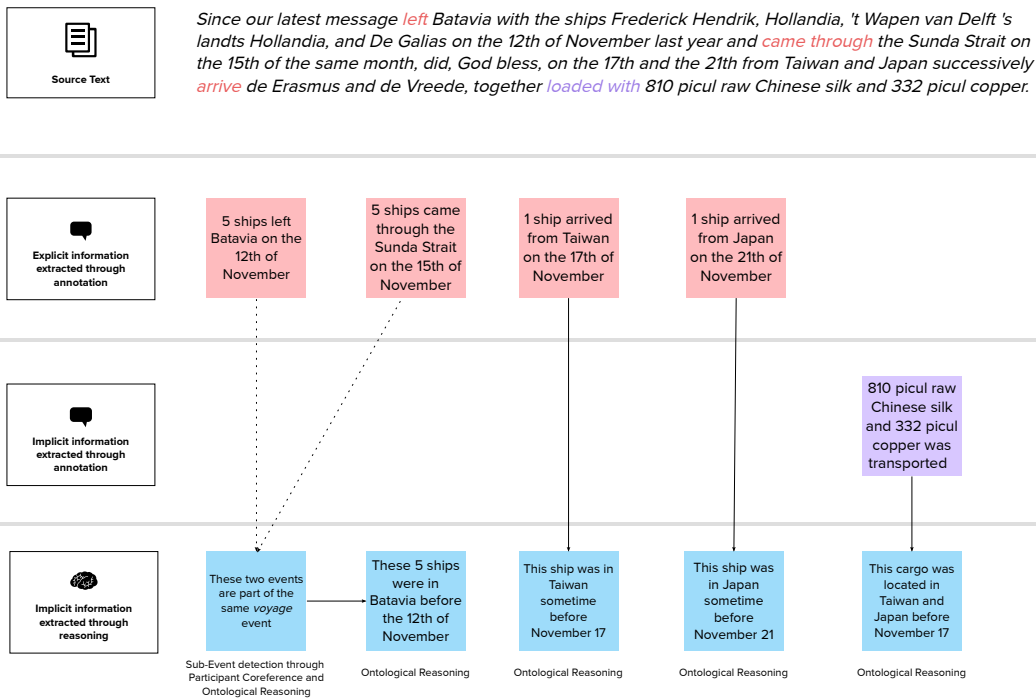
A vast amount of research has been conducted on knowledge representation for NLP, resulting in many ontologies and resources. However, none were made specifically for our domain, and little leverage commonsense reasoning over event causality. In this section, we discuss the ontologies, lexical resources and theories most relevant to the current project. When assessing re-usability, we consider what information is needed to perform event reconstruction on (pre-)colonial sources.

Most historical event detection research has been conducted on datasets from the 20th century [20, 21, 6, 3]. These frameworks are often not directly applicable to themes that are highly

---

<sup>1</sup>The source text in this figure is a comprehensive translation to English; for the original text and a literal translation to English, see 8.1

<sup>2</sup>For a representation of the taxonomy in OWL see [https://github.com/globalise-huygens/nlp-event-detection/tree/main/built\\_resources](https://github.com/globalise-huygens/nlp-event-detection/tree/main/built_resources), for a Wiki describing each event in detail see <https://github.com/globalise-huygens/nlp-event-detection/wiki>, and for our annotation guidelines see <https://docs.google.com/document/d/1ZL9BbEmGky1tJeTR-yvnU-bvqVfA14mVKOWlGgEJT0g/edit?usp=sharing>



**Figure 1:** Example of how event information is extracted from source text at different levels. For clarity reasons, the individual ship names are not stated in the squares representing extracted information.

relevant for our source (ship movement; trade; geopolitical/social relationships; power imbalances). More importantly, these historical event models do not model implied states (event implications), which is necessary for detecting information not explicitly mentioned.

ER deals with storyline plausibility. For a system to be able to understand stories (chains of events) it should have knowledge of causal reasoning [10, 11]. In order to make correct judgments on storyline plausibility a system must consider pre- and postconditions of a given event (e.g., a ship that sinks at sea will never arrive at its destination). With the results that state-of-the-art language models (LMs) are showing at several NLP tasks, it could be questioned whether the integration of an extra module for common sense knowledge integration that makes certain event causalities explicit is needed for implied event detection. However, Qasemi et al. (2021) show that several state-of-the-art language models fail to reason with preconditions of events [14]. Their results show a 10-30% gap between LMs' and human performance on three tasks evaluating the ability to understand situational preconditions. Storks et al. (2021) conclude that LMs struggle to support predictions on storyline plausibility with valid supporting evidence [22]. Many more studies show that LMs struggle to manage logical and causal reasoning [7, 8, 25]. For our project, we expect these gaps to be even bigger due to the challenging nature of the historical text (see Section 3). Therefore, we want to rely on an ontology to fill these gaps. A hybrid approach to event implication detection, combining the

strong associations captured in an LM (to bridge a possible recall gap) and an ontology as a checking and addition module (to bridge a precision gap), is also a possible line of research.

Since the ontology supports an ongoing project, we aim to create a flexible and comprehensive ontology to which event classes can be added in an iterative manner. Comprehensibility is also important to guard annotation quality and modeling possibilities. General-purpose ontologies like SUMO [13] and FrameNet [15] aim to model the whole world at a very high level of granularity, making them less flexible and comprehensible. Also, SUMO distinguishes between objects and processes but not between static and dynamic events. Therefore, although SUMO allows for reasoning, it does not formalize logical implications of events or processes as new states. FrameNet does not model causality or implications of events at all.

The Rich Event Ontology aims to integrate lexical and ontological resources [1] and also models causal relations between events [2]. Another aspect of the REO that is highly relevant to ER is the differentiation between *hasSubclass* and *hasSubevent*, the latter capturing how subevents are temporally contained within their supra-class. We can draw on many principles of the REO, however, they do not formally define how dynamic events logically imply new states.

Modeling causal relationships between events ontologically has been investigated extensively by Segers [19, 18, 16, 17]. Segers offers different versions of a detailed ontology (the Circumstantial Event Ontology (CEO)) that models pre-, during- and post-conditions of calamity events and the roles of the entities affected by the event. By doing so, the ontology represents implied causal relations between event classes. This framework served as a basis of our ontology.

A key aspect of event reconstruction is modelling the participants of an event, often referred to as semantic roles or event arguments. Li et al. (2022) show how event prediction can be informed by implied information extraction [9], taking semantic roles into account. They utilize graph schema induction as a means of predicting new nodes that represent future events. They employ a copy mechanism to generate coreferential arguments (for example, the *Detainee* argument is the *Attacker* of a previous *Attack* event).

We adopt this idea and rely on PropBank [12] for the linguistic argumentation behind this type of participant modelling. PropBank, different from the semantics-central approach of FrameNet, is based on syntax. The roles in PropBank are recyclable in the sense that they would fit in any sentence with any mentioned event. For almost any event, a Patient can be defined (similar to ARG1 in PropBank) and in most events an Agent can be defined (similar to ARG0). The downsides of PropBank for our purposes are threefold: i) it is a sentence-based approach, which is not desirable for our project because of how information is presented in the archival material (Section 3), ii) it is lexical and does not generalize over variants or synonyms that express the same event and, iii) the PropBank roles are not consistent enough to recycle among event classes.<sup>3</sup> While PropBank arguments rely on the syntactic structure of a sentence, which is linked to the transitivity of a specific verb, we want to rely on the semantic implications for an argument, the role it plays in an action and how that role can be translated to the role it plays in an inferred situation.

---

<sup>3</sup>For example, the location of an event sometimes takes up the ARG1 slot, sometimes ARG2, etcetera, depending on the syntactic structure a predicate (event) dictates in a sentence

### 3. Data

The project in which our ER module is embedded focuses on a subset of the VOC archives, namely the *Overgekomen Brieven en Papieren* ('Received Letters and Papers', OBP) and within that the *Generale Missiven* ('General Missives'). These are documents written by the governor general and council of Batavia summarizing the status of, for example, trade and conflict in Asia. They are written in Early Modern Dutch, which differs from contemporary Dutch with respect to lexical variation, spelling, style, and grammar. Most noteworthy is the frequent occurrence of very long-range dependencies, with pronouns sometimes referring to an entity several sentences or even pages back. Imperfect automatic transcriptions of handwritten material result in even more variation. Given the nature of the historical data and the lack of training data, an ontology can have a strong added value to guide the interpretation of any models that need to process these texts and reconstruct events. See 8.2 for some corpus characteristics.

## 4. Ontology Description

### 4.1. Ontology Requirements

The design of the ontology is based on six requirements following from the interdisciplinary collaboration between NLP scientists and historians and the needs for an ER system.

**R1 Domain-specific event classes**

The ontology should contain event classes that historians have indicated to be relevant in this specific corpus, and the classes should be defined by historians.

**R2 Scalar and binary change**

The ontology must accommodate for scalar change as well as binary change. This enables, for example, the modeling of price changes or the conditions of ships over time through ER.

**R3 Implied event detection and reasoning**

The ontology must accommodate automatic reasoning over events in a way that enables it to extract implied states and events.

**R4 Flexibility**

The ontology must have a (taxonomic) structure that allows for the addition of events in the future through an iterative process.

**R5 Comprehensibility**

The ontology must remain as comprehensible and transparent as possible in order to guard both annotation quality and modeling possibilities, taking into account the complexity and noisiness of the source data. The ontology should serve to reduce annotation effort where possible.

### 4.2. Event classes and ontology structure

After considering several linguistic frameworks and ontologies, we opted to take building blocks from different resources to create an ontology that is as robust as possible for our dataset

and research purposes. The CEO served as a basis for the definition of event classes because it focuses less on how events are semantically manifested in text (like FrameNet), and more on the practical notion that dynamic events often imply a static event. However, many event classes had to be adapted, deleted or added, since the CEO was built for processing modern news.<sup>4</sup> We copied the notion that tokens can be annotated with semantic roles cross-sententially from FrameNet (as opposed to PropBank where the whole syntactic structure of one sentence should be represented in the annotation). We took the notion that event arguments do not have to be re-defined for each event class specifically from PropBank. We defined our own limited set of argument types in order to guard annotation quality and modeling purposes.

The present ontology contains events that can be categorised under three themes deemed relevant by historians: ship movement, trade, and (geo)political/social relations. The event classes are defined so that they mainly represent *observable* events, steering clear of concepts that have an inherently subjective character. For example, whether an action is legal or illegal depends on the context, which should be studied firstly by a historian. Which events can be classified as *observable* and which as *subjective* is an open and ongoing interdisciplinary discussion. Since there also exist causal relations that are debatable or subjective [24], we only aim to model logical implications of events (which means only modelling necessary and not possible consequences of dynamic events).

For an overview of the current taxonomy of dynamic events, see Figure 2. Most dynamic events automatically imply the transition from a given to a new static event. The static events are modeled as event classes with their own arguments. This allows us to annotate and extract them at the same complexity level as dynamic events. For a sample of the graph indicating how event classes relate, see Figure 3. Our Wiki describes all event classes.<sup>5</sup>

Previous research [23] shows that events are often described as relative changes on a scale rather than in absolute values. A scalar model allows for capturing and reasoning over such relative and under-specified reports. We solve the inclusion of scalar changes without having to annotate and model the absolute value of change by including events like *Increasing*, *Decreasing*, *Repairing* and *Damaging* and linking them to states like *HavingInternalState+* and *HavingInternalState-*.<sup>6</sup>

### 4.3. Event participant modelling

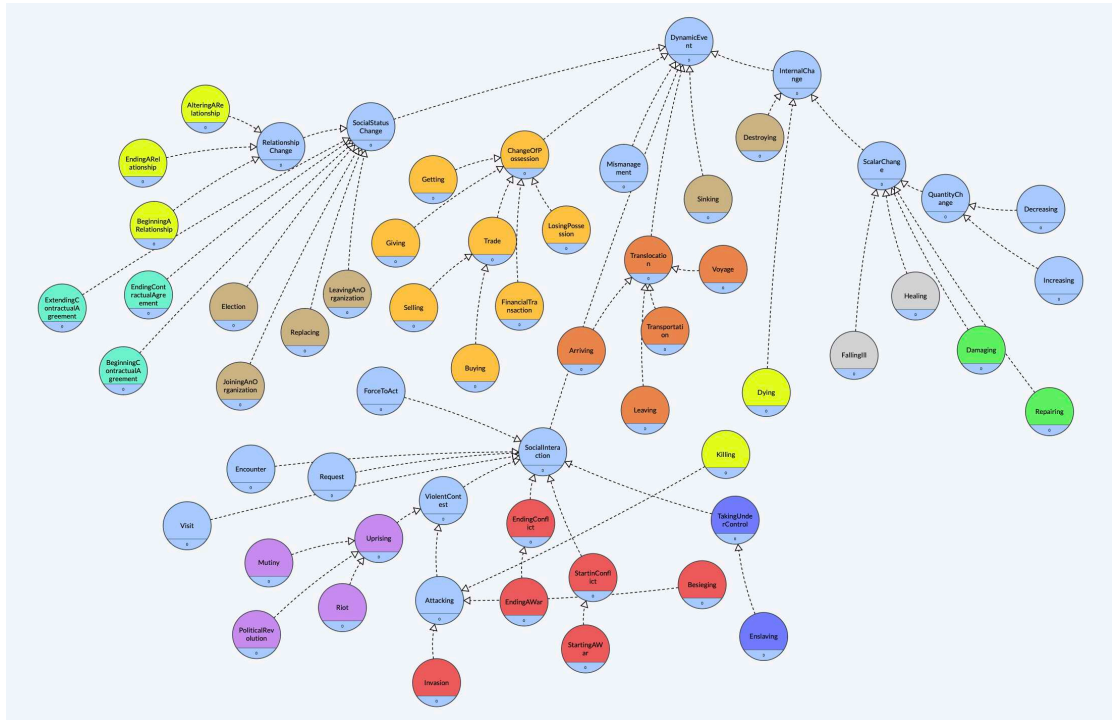
FrameNet and the CEO define participants specific to each event class. In order to be able to model actors of implied changes of state, semantic roles have to be recyclable from one event class to another (if necessary with minimal conversion rules). That is to say, we want to infer that the Agent in an *Attacking* event is a Patient in a *BeingInConflict* state. FrameNet shifts the semantics from the event class to the role, which limits the generalization across roles. The CEO offers a solution to this with the incorporation of assertion rules. The goal for our ontology is to have even more intuitive and general roles than in the CEO because the textual data we are working with is extremely noisy and information about one event is stretched out over large portions of text (see Section 3). By adopting more intuitive and general roles,

---

<sup>4</sup>The final ontology will link classes to the CEO, SUMO, FrameNet and other relevant ontologies with SKOS relations

<sup>5</sup><https://github.com/globalise-huygens/nlp-event-detection/wiki>

<sup>6</sup>A ship being repaired implies *HavingInternalState+* and a deteriorated city implies *HavingInternalState-*



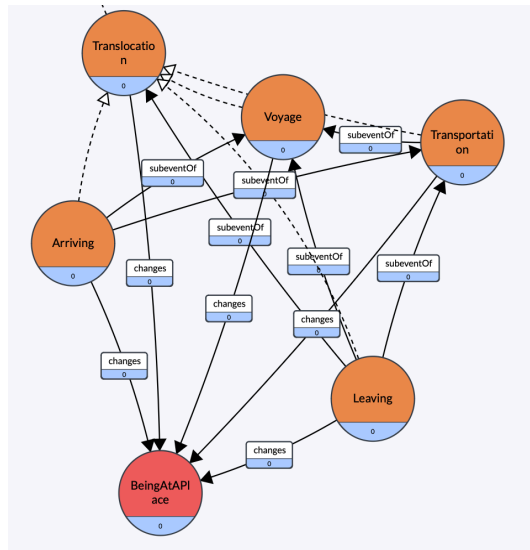
**Figure 2:** Simplified visualization of the current version of the taxonomy of Dynamic Events. Classes that imply a transition in the same static event are grouped by position and/or colour. The default blue colour for a class indicates they do not automatically imply a transition to a new state.

we simplify the annotation process as much as possible, which will lead to more consistent annotation. Also, we suspect that the definition of more general semantic roles might be more informative for a machine learning system that needs to generalize on limited training data.

Table 1 gives an overview of our arguments. Participants can be animate and non-animate. The AgentPatient role is specifically defined for events in which a causer is simultaneously undergoing the event *SocialInteraction*, *BeginningARelationship*, *EndingAContractualAgreement*. For now, Time can be anything from a date to an hour to a day in the week. Source and Target are defined for *Translocation* events that may indicate both the starting and ending point of a movement. Roles are defined per event in our Wiki.<sup>7</sup> In our ontology, event participants are represented as properties: i.e. event classes have properties like *hasAgent*, *hasPatient*.

The roles can sometimes be recycled directly from a dynamic event class to their supra-class as well as the static class they change, and sometimes they need conversion rules. See Table 2 for an example of how roles are translated between classes. It shows how the Beneficiary of a *Selling* event (the person to whom is sold) is an Agent in the static event *HavingInPossession*, and it can be deduced that this Agent has the Patient of the *Selling* event (the thing sold) in possession from the moment the event happened onward.

<sup>7</sup><https://github.com/globalise-huygens/nlp-event-detection/wiki>



**Figure 3:** Detail of graph with subclass (dotted arrow), subevent and implication relations. Orange classes are Dynamic Events; the red class is a Static Event. The *subEventOf* relation indicates that an instance of the subject class is a possible sub-event of an instance of the object class.

**Table 1**  
Semantic roles / event arguments

Participants	Spatial	Temporal
Agent	Location	Time
Patient	Source	
AgentPatient	Target	
Beneficiary	Path	
Cargo		
Instrument		

## 5. Future steps: Operationalization

An annotation pilot has been conducted and the ontology and event guidelines have been updated accordingly. We plan to enhance the ontology in this iterative manner over a few annotation rounds. When the ontology and guidelines are stabilized, we will run consecutive annotations through the pipeline presented in Figure 4. The resulting Knowledge Graph (KG) is to be integrated in the complete KG supporting the online infrastructure for historical research.

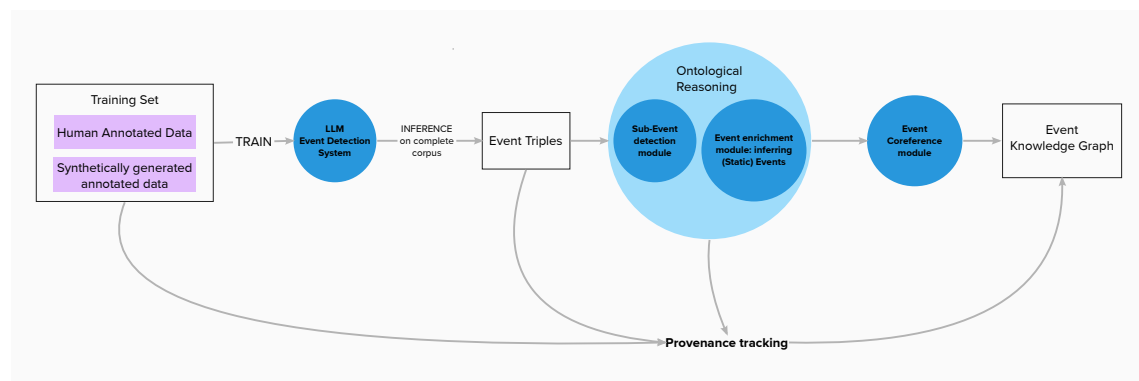
When representing information extracted from text it is important to be transparent about provenance [4]. For each event triple represented in the KG, we will show whether this triple was retrieved through human annotation or through the NLP pipeline. The GRASP model [5] allows for provenance tracing so that every triple is part of a claim by a source.



**Table 2**

Example of role recycling from dynamic event superclasses to dynamic event subclasses to static event  
 AP = AgentPatient; A = Agent; P = Patient; B = Beneficiary

DynamicEvent	Role	DynamicEvent (subclass)	Role	StaticEvent	Role	Situation
StartingAConflict	AP AP	StartingAWar	AP AP	BeingInConflict	P1 P2	P1 inConflict with P2 P2 inConflict with P1
Attacking	A P	Besieging	A P	BeingInConflict	P1 P2	P1 inConflict with P2 P2 inConflict with P1
Uprising	A P	Mutiny	A P	Unrest	P1 P2	Unrest between P1 and P2 Unrest between P1 and P2
	A P	PoliticalRevolution	A P	Unrest	P1 P2	Unrest between P1 and P2 Unrest between P1 and P2
ChangeOf Possession	A1 A2 P	Giving	A B P	HavingInPossession	A1 A2 P	A1 notHasInPossession P A2 hasInPossession P
	A1 A2 P	Selling	A B P	HavingInPossession	A1 A2 P	A1 notHasInPossession P A2 hasInPossession P
	A2 A1 P	Buying	A B P	HavingInPossession	A2 A1 P	A2 hasInPossession P A1 notHasInPossession P



**Figure 4:** Visualisation of NLP pipeline containing the ontology as a reasoning step.

## 6. Conclusion

This paper has offered an overview of considerations that arise when designing an event ontology for a specific domain and a specific purpose within an interdisciplinary effort, relying, where possible, on existing resources. We make clear that event classes should be defined

according to the characteristics of the dataset and needs of the final product: in our case defining events specific enough to the corpus but general enough to leave space for historical and contextual interpretation. We distinguish between event extraction and event reconstruction, defining the latter as a combination of several NLP-tasks and reasoning. We propose an ontological method of extracting implicit information and incorporating it in an event reconstruction pipeline.

By focusing on this case-study of event reconstruction in Early Modern Dutch documents we expect to come to new insights about how to push the state-of-the-art in event extraction (using ontologies) forward. Our data and ontology can serve as a robust use-case for investigating to what extent LMs need external common sense knowledge integration to successfully perform certain NLP tasks. Finally, we aim to facilitate a new way of doing historical research with the software supported by the presented ontology.

## 7. Acknowledgements

This research is part of the GLOBALISE project, funded by the Dutch Research Council (NWO) under project number 175.2019.003. We thank Kay Pepping, dr. Lodewijk Petram, dr. Pia Sommerauer and dr. Micky Cornelissen for their help revising the paper, as well dr. Manjusha Kuruppath, Maartje Hids, Henrike Vellinga and Brecht Nijman for their involvement in the annotation pilot.

## References

- [1] C. Bonial, S. Brown, M. Palmer, and G. Kazeminejad. “The Rich Event Ontology: Ontological Hub for Event Representations”. In: *Computational Analysis of Storylines: Making Sense of Events* (2021), pp. 47–66.
- [2] S. W. Brown, C. Bonial, L. Obrst, and M. Palmer. “The rich event ontology”. In: *Proceedings of the Events and Stories in the News Workshop*. Vancouver, Canada, 2017, pp. 87–97.
- [3] A. Cybulska and P. Vossen. “Historical event extraction from text”. In: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Portland, OR, USA, 2011, pp. 39–43.
- [4] A. Fokkens, S. Ter Braake, N. Ockeloen, P. Vossen, S. Legêne, G. Schreiber, et al. “BiographyNet: Methodological Issues when NLP supports historical research”. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC’14)*. Reykjavik, Iceland, 2014, pp. 3728–3735. DOI: [https://doi.org/10.1162/coli\\_a\\_00347](https://doi.org/10.1162/coli_a_00347).
- [5] A. Fokkens, P. Vossen, M. Rospocher, R. Hoekstra, W. R. van Hage, and F. B. Kessler. “Grasp: Grounded representation and source perspective”. In: *Proceedings of the Workshop Knowledge Resources for the Socio-Economic Sciences and Humanities associated with RANLP*. Varna, Bulgaria, 2017, pp. 19–25.
- [6] N. Ide and D. Woolner. “Historical ontologies”. In: *Words and Intelligence II: Essays in Honor of Yorick Wilks* (2007), pp. 137–152.

- [7] O. Ignat, Z. Jin, A. Abzaliev, L. Biester, S. Castro, N. Deng, X. Gao, A. Gunal, J. He, A. Kazemi, M. Khalifa, N. Koh, A. Lee, S. Liu, D. June Min, S. Mori, J. Nwatu, V. Perez-Rosas, S. Shen, Z. Wang, W. Wu, and R. Mihalcea. “A PhD Student’s Perspective on Research in NLP in the Era of Very Large Language Models”. In: *arXiv preprint arXiv:2305.12544* (2023).
- [8] Z. Jin, A. Lalwani, T. Vaidhya, X. Shen, Y. Ding, Z. Lyu, M. Sachan, R. Mihalcea, and B. Schölkopf. “Logical fallacy detection”. In: *arXiv preprint arXiv:2202.13758* (2022).
- [9] M. Li, S. Li, Z. Wang, L. Huang, K. Cho, H. Ji, J. Han, and C. Voss. “The future is not one-dimensional: Complex event schema induction by graph modeling for event prediction”. In: *arXiv preprint arXiv:2104.06344* (2021).
- [10] K. Ma, F. Ilievski, J. Francis, E. Nyberg, and A. Oltramari. “Coalescing Global and Local Information for Procedural Text Understanding”. In: *arXiv preprint arXiv:2208.12848* (2022).
- [11] P. Mirza. “Event Causality”. In: *Computational Analysis of Storylines: Making Sense of Events* 106 (2021), pp. 106–124.
- [12] M. Palmer, D. Gildea, and P. Kingsbury. “The proposition bank: An annotated corpus of semantic roles”. In: *Computational linguistics* 31.1 (2005), pp. 71–106.
- [13] A. Pease, I. Niles, and J. Li. “The suggested upper merged ontology: A large ontology for the semantic web and its applications”. In: *Working notes of the AAAI-2002 workshop on ontologies and the semantic web*. Vol. 28. 2002, pp. 7–10.
- [14] E. Qasemi, F. Ilievski, M. Chen, and P. Szekely. “Paco: Preconditions attributed to commonsense knowledge”. In: *arXiv preprint arXiv:2104.08712* (2021).
- [15] J. Ruppenhofer, M. Ellsworth, M. Schwarzer-Petruck, C. R. Johnson, and J. Scheffczyk. *FrameNet II: Extended theory and practice*. Tech. rep. International Computer Science Institute, 2016.
- [16] R. Segers, T. Caselli, and P. Vossen. “The circumstantial event ontology (CEO)”. In: *Proceedings of the Events and Stories in the News Workshop*. Vancouver, Canada, 2017, pp. 37–41.
- [17] R. Segers, T. Caselli, and P. Vossen. “The Circumstantial Event Ontology (CEO) and ECB+/CEO; an Ontology and Corpus for Implicit Causal Relations between Events”. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC’18)*. Miyazaki, Japan, 2018.
- [18] R. Segers, M. Rospocher, P. Vossen, E. Laparra, G. Rigau, and A.-L. Minard. “The event and implied situation ontology (eso): Application and evaluation”. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC’16)*. 2016, pp. 1463–1470.
- [19] R. Segers, P. Vossen, E. Laparra, G. Rigau, M. Rospocher, and A.-L. Minard. “The Event and Implied Situation Ontology (ESO)”. In: *Clin26*. 2015.
- [20] R. Sprugnoli and S. Tonelli. “Novel event detection and classification for historical texts”. In: *Computational Linguistics* 45.2 (2019), pp. 229–265.

- [21] R. Sprugnoli and S. Tonelli. “One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective”. In: *Natural language engineering* 23.4 (2017), pp. 485–506. doi: <https://doi.org/10.1017/S1351324916000292>.
- [22] S. Storcks, Q. Gao, Y. Zhang, and J. Chai. “Tiered reasoning for intuitive physics: Toward verifiable commonsense language understanding”. In: *arXiv preprint arXiv:2109.04947* (2021).
- [23] P. Vossen, R. Agerri, I. Aldabe, A. Cybulska, M. van Erp, A. Fokkens, E. Laparra, A.-L. Minard, A. P. Apro시오, and G. Riga. “NewsReader: using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news”. In: *Knowledge-Based Systems* 110 (2016), pp. 60–85. doi: <https://doi.org/10.1016/j.knsys.2016.07.013>.
- [24] P. Vossen, T. Caselli, and R. Segers. “A Narratology-Based Framework for Storyline Extraction”. In: *Computational Analysis of Storylines: Making Sense of Events* 125 (2021), pp. 125–140.
- [25] M. Willig, M. ZEČEVIĆ, D. S. Dhimi, and K. Kersting. “Causal Parrots: Large Language Models May Talk Causality But Are Not Causal”. In: *preprint* (2023).

## 8. Appendices

### 8.1. Additional examples

- **Original source**

‘Tsedert onsen Jongsten (hiernevens in copie gaande) met de schepen Frederick Hendrik, Hollandia, ’t Wapen van Delff, ’s landts Hollandia, ende de Galias den 12en November passato in compagnie van Batavia *gescheyden* ende den 15en ditto door de Strate Sunda *geraeckt*, sijn hier, Godtloff, den 17en ende den 21en van Teyouhan ende Jappan successive wel *aengecomen* ’t yacht Erasmus ende ’t schip de Vreede, t’samen geladen 810 picol rouwe Chineesche syde ende 332 picol coper.’<sup>8</sup>

- **Literal translation**

Since our youngest (herewith in copy) with the ships Frederick Hendrik Hollandia, ’t Wapen van Delff, ’s landts Hollandia, and de Galias on the 12th of November passato in company from Batavia *seperated* and the 15th of the same *have made its way through* the Sunda Strait, did here, God bless, on the 17th and the 21st from Teyouhan and Jappan successively *arrive* ’t yacht Erasmus and the ship de Vreede, together loaded 810 picul raw Chinese silk and 332 picul copper.

### 8.2. Corpus characteristics

---

<sup>8</sup>General Missive of 6-1-1628’, National Archive, The Hague, The Netherlands, 1.04.02 (Archive of the VOC), inventory no. 1092, folio 1, r.

Name	Period	Description	Number & size	Avg doc length
General Missives	1618-1792	Narrative reports from Council of India (Batavia) to Gentlemen Seventeen (Dutch Republic). Often ordered by region; small summaries in margins	923 documents; 191.725 handwritten pages	207 pages
OBP	1610-1796	Collection of General Missives and varied documents on which these missives are based.	c. 250,000 documents; c. 7 million handwritten pages	c. 28 pages (rough estimate)