Language Resources and Evaluation (2023) 57:819–848
https://doi.org/10.1007/s10579-022-09597-1

ORIGINAL PAPER

Constructing a cross-document event coreference corpus for Dutch

Loic De Langhe[1] · Orphée De Clercq[1] · Veronique Hoste[1]

Accepted: 16 May 2022 / Published online: 4 June 2022

Abstract

Event coreference resolution is a task in which different text fragments that refer to the same real-world event are automatically linked together. This task can be performed not only within a single document but also across different documents and can serve as a basis for many useful Natural Language Processing applications. Resources for this type of research, however, are extremely limited. We compiled the first large-scale dataset for cross-document event coreference resolution in Dutch, comparable in size to the most widely used English event coreference corpora. As data for event coreference is notoriously sparse, we took additional steps to maximize the number of coreference links in our corpus. Due to the complex nature of event coreference resolution, many algorithms consist of pipeline architectures which rely on a series of upstream tasks such as event detection, event argument identification and argument coreference. We tackle the task of event argument coreference to both illustrate the potential of our compiled corpus and to lay the groundwork for a Dutch event coreference resolution system in the future. Results show that existing NLP algorithms can be easily retrofitted to contribute to the subtasks of an event coreference resolution pipeline system.

Keywords Event coreference resolution · Event annotation · Entity coreference resolution

1 Introduction

Researching the links between individual entities and events in texts is paramount to a good understanding of natural language. Knowing which textual events refer to one another allows us to weave a narrative within a given text or across different texts. In the past, within-document event coreference resolution

✉ Loic De Langhe
loic.delanghe@ugent.be

[1] LT3, Language and Translation Technology Team, Ghent University, Groot-Brittanniëlaan 45, 9000 Ghent, Belgium

has successfully been applied in the areas of template filling (Humphreys et al., 1997), automated population of knowledge bases (Ji & Grishman, 2011), question answering (Narayanan & Harabagiu, 2004) and contradiction detection (De Marneffe et al., 2008). In recent years, research interest in event extraction and event coreference resolution has been steadily growing in popularity for the English language domain (Lu & Ng, 2018) with the expressed goal of working towards end-to-end event coreference resolution in a cross-document setting. This specific objective is particularly interesting because despite many efforts, current NLP techniques usually still rely on word-level lexical semantics. Researching an application such as ECR in which discourse-level relations are important and which simultaneously breaks down topic -and document barriers (Bugert et al., 2020) can provide us with many insights on language at a different structural level. In addition to this, cross-document event coreference resolution can potentially be of great benefit to practical multi-document applications such as summarization recommendation (Liu & Lapata, 2019), content-based news recommendation (Vermeulen, 2018) and reading comprehension recommendation (Yan et al., 2019). For content-based news recommendation in particular, it has been shown that the development of a new generation of news recommendation systems relies on the identification, extraction and analysis of key news events in texts followed by the linking of said news events, both in within- and cross-document settings (Colruyt et al., 2019a). The latter being critical, as it would provide the reader with access to different points of view discussing the same event.

Despite the aforementioned efforts, event extraction and coreference resolution remain challenging tasks within the field of Natural Language Processing (NLP), both conceptually and in practice. First and foremost, finding a satisfying definition for news events (and indeed, events in general) is difficult, as concepts such as "relevancy" and "importance" are hard to translate to a practical setting. Defining a clear cut-off point of what constitutes an event is thus somewhat problematic. As a direct result, annotation of events becomes a complex task in which many fundamental decisions rely entirely on the judgement of the annotators, sometimes resulting in inconsistent annotations (Vossen, 2018). Another problem that routinely plagues event extraction and coreference studies is scope. Keeping in mind practical applications such as content-based news recommenders, ideally, one would design a system equipped to deal with unrestricted events, i.e events which do not belong to a set of predefined topics or themes. However, due to practical limitations, research tends to often focus on events belonging to a certain topic, within a self-defined taxonomy (Aone & Ramos-Santacruz, 2000). Moreover, almost all prior research was conducted on high-resourced languages (Lu & Ng, 2018), while studies exploring event extraction and coreference resolution for other languages remain scarce. Even for languages such as Dutch, which are typically well-resourced for many NLP tasks, to date, no large-scale event coreference corpora exist.

This paper presents the initial efforts towards an event coreference resolution algorithm for the Dutch language. We introduce the ENCORE corpus, a collection of 1,115 Dutch news texts in which coreference between news events is annotated, both at a within and cross-document level. The articles in the corpus belong to a large variety of topics, ranging from geopolitical events to local news. This work represents, to the best of our knowledge, the first effort to create a large-scale resource for event coreference in (untranslated) Dutch news texts.

We hope that the introduction of this corpus will boost discussion and research in event extraction and coreference studies and its possible practical applications for low-resourced and less-studied languages. Our annotation effort have resulted in a document collection which is comparable to the largest and most popular English language corpora for event coreference resolution. Besides introducing the ENCORE corpus, a set of preliminary experiments have also been conducted, which can give us an initial indication on the viability of using certain well-established methods and techniques for event coreference resolution. Earlier research on coreference resolution has shown that pipeline-based architectures, in which the problem is resolved gradually, tend to be quite effective for English (Lu & Ng, 2016a). One aspect of such a pipeline is identifying shared participants between different candidate event mentions. Logically, coreferring events will have the same real-life participants engaging in them. Knowing which textual entities refer to the same real-world entities is thus quite valuable information when trying to determine whether or not two candidate event mentions refer to one another. We employ two well-established methods for entity coreference in Dutch and adapt one of them for participant coreference, as event argument participants do not always correspond to traditional textual entities. We demonstrate that with some minor modifications, existing Dutch NLP systems can be successfully adapted to certain tasks within a pipeline-based event coreference resolution system. We managed to improve the baseline results of rule-based coreference systems and machine learning mention-pair models with an averaged F1 score of 9.3% and 3.4% respectively for coreference resolution between event participants.

In this paper we first give a comprehensive overview of existing event coreference datasets. We discuss the scope, size, annotation process and potential benefits or drawbacks for the most popular ECR corpora (Sect. 2). Next, we discuss the creation of our own corpus. This includes the collection, processing and annotation steps of the corpus, as well as a summarized overview of our annotation guidelines. In addition to this, we also provide a broad examination of the finished corpus and perform a set of inter-annotator agreement (IAA) experiments to measure the quality of the annotations (Sect. 3). Finally, we discuss the methodology, results and error analysis of the argument coreference resolution experiments (Sect. 4).

2 Related work

Coreference links are the glue that hold language together. For human language understanding and interpretation, knowing which text fragments refer to one another is crucial. Naturally, computer-assisted language applications can also greatly benefit from the integration of this type of information (Elango, 2005). Knowing this, it is no surprise that coreference resolution has been one of the core tasks in NLP for many years. Typically, coreference resolution can be performed on an entity or event level. When trying to resolve coreference between entities, one attempts to automatically link textual entities, which are often noun phrases, to one another when they refer to the same extra-linguistic entity (Ng, 2017). Event coreference resolution (ECR), on the other hand, is a task that aims to resolve coreference between text fragments referring to the same real-world events, see below for an example.

1. SP.A brengt winterjassen bijeen voor kansarmen *EN: SP.A gathers winter coats for the underprivileged.*
2. Op de Werelddag tegen Armoede hield de SP.A van Dendermonde op de binnenkoer van het ABVV in de Dijkstraat een inzameling van winterjassen *EN: On the day against poverty the SP.A faction of Dendermonde organised a collection of Winter coats on the ABVV courtyard in the Dijkstraat.*

In an entity coreference task, we would only be interested in drawing a coreferential link between the textual entities *SP.A* and *de SP.A van Dendermonde*, as the 'broader' entity *SP.A* refers to the subgroup *de SP.A van Dendermonde* in this context i.e the specific faction of the SP.A that was present in Dendermonde to collect winter coats. In an event coreference scenario, however, we aim to draw a link between these two sentences in their entirety. Capturing all the complexities of textual events in order to be able to perform event coreference resolution can be quite a daunting task. When comparing both sentences, we observe that they have different verbal triggers (*brengt bijeen (EN: gathers)* vs. *hielden een inzameling (EN: organised a collection)*), that information on the time and location of the event may or may not be present (*Op werelddag tegen Armoede*, *op de binnenkoer van het ABVV in de Dijkstraat*) and that different surface forms of people and objects can participate in the event (*SP.A* vs. *SP.A van Dendermonde*). Moreover, one also has to consider the event type; most applications tend to work with information relating to a specific theme such as economy, politics or technology (Minard et al., 2016). As a consequence, posing no restriction on specific themes will only further complicate the ECR task. As stated before, most studies regarding event coreference that work within predefined themes opt for a fixed event typology. This setup works well in closed domain settings where data is typically restricted to newspaper articles falling within one particular subject i.e economy or politics. However, it has been shown in Dutch event extraction studies that finding a taxonomy which covers all possible news event types is not straightforward and that the exceptions arising from such a typology invoke a slew of conceptual problems on their own, both at the annotation and extraction level (Colruyt et al., 2019b). Last and not least, ERC can be performed both in a within and cross-document setting.

When compared to well-studied areas in coreference resolution such as entity coreference (Sukthanker et al., 2020), ECR thus tends to be much more complex and multi-faceted. Nevertheless, in recent years more work has emerged on event-based tasks (Lu & Ng, 2018). A large number of models and approaches have been proposed for resolving coreference between events ranging from mention-pair (Cybulska & Vossen, 2015) and mention ranking models Lu and Ng (2017) that use pipeline-based architectures in order to detect, extract and

Table 1 Overview of the most popular corpora annotated for event coreference, both within-document (WD) and cross-document (CD)

Corpus	#Documents	Genre	Coref	Languages
ACE 2005	600, 500	News articles, conversations	WD	EN, CH
OntoNotes	600	Financial news articles	CD	EN
TAC KBP	1000, 800, 400	News articles, forum discussions	WD	EN, SP, CH
ECB	480	News articles	CD	EN
ECB+	982	News articles	CD	EN
Newsreader Meantime	120	News articles	CD	EN, DU, IT, SP

resolve coreference in text to joint inference methods (Lu et al., 2016) and even rule-based sieve approaches (Lu & Ng, 2016b). Of the aforementioned methods, supervised learning algorithms are the most popular. Advancements in event coreference resolution have generally followed trends within the broader field of NLP. In recent years, attention in coreference research has shifted from traditional machine learning approaches such as maximum entropy models (Ahn, 2006) and support vector machines (Chen and Ng, 2014) to deep learning architectures (Nguyen et al., 2016) and span-based embedding models (Lee et al., (2018; Joshi et al., 2020) . In addition to this, several semi-supervised (Chen & Ng, 2016) and unsupervised learning methods have also been proposed. The latter typically include non-parametric probabilistic models (Bejan & Harabagiu, 2010). Note that all of the aforementioned research was exclusively done for English or Chinese. To our knowledge, there have been no attempts yet to create coreference resolution systems for unrestricted events, i.e events not belonging to a certain predefined theme, for languages such as Dutch. While the creation of a Dutch reference corpus for ECR remains our primary objective, working with events from a large number of topics and themes is also of paramount important because posing no restrictions on the type of events involved is an important step towards the integration of event coreference resolution systems into practical applications.

A first prerequisite for creating an ECR system is data. However, data for event coreference resolution is notoriously sparse, especially when compared to other tasks involving coreference (Choubey et al., 2018). Unlike entities, reference to the same real-world event is made only sparingly throughout texts. ECR corpora are often compiled using news articles from a series of different sources discussing the same event. Note that while this method of collecting data greatly increases the chance of having much-valued event coreference relations in the dataset, it does not completely solve the aforementioned sparsity problem. Several large-scale corpora for event coreference resolution exist for English and other high-resourced languages, such as Chinese and Spanish (Lu & Ng, 2018).

The following section gives an overview of the most used event coreference corpora. Additionally, Table 1 provides a side-by-side comparison of the sizes of each corpus, the genres they comprise, the way in which coreference was annotated, either within (WD) or cross-document (CD) and the languages included.

Among the most popular of event coreference corpora is the EventCorefBank+ (ECB+) dataset (Cybulska & Vossen, 2014b). The corpus is an extension of the EventCorefBank (ECB) (Bejan & Harabagiu, 2010) corpus and includes both within and cross-document event coreference annotations, as well as an extensive annotation scheme that covers many aspects of textual events (e.g Time, Location, Participants etc.). Another large-scale resource for ECR is the OntoNotes corpus (Pradhan et al., 2007). Much like ECB+ and its predecessors, the OntoNotes corpus covers events of all types. In this corpus both entity and event coreference has been annotated in a cross-document fashion. However, because no distinction is made between entities and events in the annotation scheme, this dataset tends to be more suited for tasks other than coreference resolution, e.g. automatic ontology population (Su et al., 2019).

Next, there are the ACE corpora which are published by the Linguistic Data Consortium (LDC) as an ongoing effort to provide resources for tasks related to automatic content extraction. Of the corpora provided to the public by the LDC, ACE 2005 (ACE English Annotation Guidelines, 2008) is most suited for the task of event coreference resolution. This corpus also provides a (more limited) set of Chinese documents for ECR, making it one of the few corpora which provides resources for languages other than English. However, a notable drawback of the ACE 2005 corpus is that its coreference annotations are limited to within-document event links. In addition to this, only events belonging to specific event types are annotated, rendering the corpus less effective for extraction and resolution tasks of unrestricted events. Following an approach similar to the one of ACE 2005, the TAC KBP corpora (Mitamura et al., 2015) were created, in these corpora only within-document event coreference is annotated and only if these events belong to a specific type. In addition, the corpus includes a more limited set of Chinese and Spanish documents for event coreference resolution. Notable differences between ACE 2005 and TAC KBP include a more complex annotation scheme and a more expanded set of event types.

A final ECR corpus that should be mentioned is the Newsreader Meantime dataset (Minard et al., 2016). While this corpus is limited in size, 120 news articles, it has extensive event annotations and includes both within and cross-document coreference. Moreover, it includes documents in English, Italian, Dutch and Spanish. This makes the Meantime corpus, to our knowledge, the only linguistic resource for ECR in Dutch. Besides its limited size, the articles in Dutch, Spanish and Italian were translated from the original English source news articles which is arguably a non-optimal way of collecting data. A final note regarding the MeanTime corpus is that the Italian and Spanish data was not annotated directly, but rather through cross-lingual projection.

Table 2 Three randomly drawn clusters with their overarching topic and number of documents included

Cluster id	Topic	# of documents
47	Tim Burton exposition in Genk	11
75	Royal Wedding Prince Harry	24
87	Election of Cuban president	12

3 Corpus creation

3.1 Data collection

All data was sourced from a large collection of Dutch (Flemish) newspaper articles, amounting to no less than 631,559 news articles[1]. These articles were collected during the calendar year 2018 from a variety of different news sources and cover a broad and diverse number of topics ranging from geopolitical issues to local news. This collection comprises articles from the online versions of a number of national (*De Morgen*, *Het Nieuwsblad*, *Het Laatste Nieuws*, *De Standaard*) and regional (*Het Belang van Limburg*) newspapers, as well as articles published on the news website of the Flemish public broadcasting agency (*VRT News*). Given that event coreference resolution is a task that typically suffers from data sparsity (Bugert et al., 2020) additional steps were taken to ensure a large number of event coreference links within our corpus. To this purpose, articles with the same overarching topics were grouped into event coreference clusters of around ten to twenty news articles. This was done by first randomly drawing an article from the entire document collection, after which all named entities within this article were retrieved using the LeTs preprocessing toolkit (Van de Kauter et al., 2013a). Next, a pass was made through the document collection and articles containing five or more overlapping named entities with the core article were added to the cluster. This entity-based method resulted in a cluster in which not only entities, but also events were likely to overlap. We decided that, in order for a named entity/topic cluster to be included in the corpus, it should contain at least 10 and no more than 50 news articles. These cut-off points were established to combat data sparsity on the one hand, while also ensuring the clusters would not grow to disproportionate sizes which could complicate manual annotation. This process led to 122 clusters. Finally, all clusters were manually pruned in order to remove irrelevant articles and duplicates, leading to a final set of 91 clusters. Table 2 displays three randomly drawn clusters from the collection following the aforementioned selection process.

[1] These articles were collected as a part of the NewsDNA project (https://www.ugent.be/mict/en/research/newsdna)

3.2 Data annotation

In order to create a complete and comprehensible dataset for Dutch ECR it was crucial to annotate: all possible events (i), together with all relevant information regarding these events (ii), coreference relations between entities that function as arguments to the aforementioned events (iii) and finally, event coreference, both within and across documents (iv). Note that annotation of cross-document event coreference was only conducted within a given event cluster, as manual annotation over the entire corpus would be an overwhelming task. A possible solution for this scaling problem could be found in semi-supervised annotation methods, which in recent years have been gaining popularity for large-scale text annotation tasks (Caicedo et al., 2022). However, due to the complexity and intricacies involved in this type of annotation we do not yet esteem these methods as viable for event coreference annotation specifically.

We use the guidelines developed for the widely popular ECB+ corpus as a building block for our own data, as we believe its annotation scheme (Cybulska & Vossen, 2014a) is straightforward, logical and universal. The ECB+ style of annotation is quite extensive, especially when compared to the more succinct styles of the ACE 2005, OntoNotes and KBP corpora (see Sect. 2 for more details). In ECB+, events as a whole can be represented by syntactic clauses, infinitival constructions or noun phrases. Typically, each event is composed of a series of *event arguments*. These arguments provide additional information regarding the real world event and correspond well to the wh-questions: what, who, where, when, why and how. The ECB+ guidelines specify four types of event arguments: EVENT-ACTION, EVENT-TIME, EVENT-LOCATION and EVENT-PARTICIPANTS. With the EVENT-PARTICIPANTS arguments containing two major subtypes: HUMAN-PARTICIPANTS and NON-HUMAN-PARTICIPANTS that partake in the event. The example below illustrates the typical form of an event annotation:

3. [[Het vliegtuig van vlucht MH17]$^{Non-humanParticipant}$ werd [op 17 juli 2014]^{Time} boven [Oost-Oekraïne]^{Location} uit de lucht [geschoten]^{Action} door [een Buk-raket, een wapen van Russische makelij]$^{Non-humanParticipant}$]^{Event} *EN: The airplane of flight MH17 was shot down on july 17th 2014 above eastern Ukraine by a Russian-made BUK-missile.*

3.2.1 Event annotation

Before going into detail about the annotation scheme itself, it is useful to consider what actually constitutes an event. One of the most commonly used definitions of textual events can be found in the TimeML specifications (Pustejovsky et al., 2003), where events are defined as "situations that happen or occur". However, we propose to modify this definition to the following: "Any real, hypothetical or fictional situation that occurs, occupying a space-time and involving a number of participants". This interpretation draws from earlier work (Quine, 1985; NIST, 2005) and more importantly allows for the inclusion of fictional and hypothetical events.

We made a series of modifications to the existing ECB+ annotation scheme, taking into account certain limitations we were faced with, as well as our ambition to develop an event coreference resolution system which can be used in practical applications such as news recommendation algorithms. First, we make changes regarding the interpretation of verbs signalling an action of reporting. This type of action is very prevalent in news texts, but annotation can be inconvenient in some cases, especially when trying to create a lexically rich qualitative corpus. Consider the example below:

4. Fouad Belkacem zegt dat hij zich zal verzetten tegen de uitspraak *EN: Fouad Belkacem says he will resist the verdict.*

Arguably, one might discern two separate events: *zeggen*/*say*) and *verzetten*/ *resist*). While the action of zeggen does satisfy the aforementioned conditions, i.e we can trace this action to a specific point in time when the expression was made and the action is well defined, we do not consider verbs of this type as events. Compared to main events that constitute the articles, actions that signal a report of said events hold very little informational value. As verbs of this type are extremely common in news texts, annotating all of them would take up valuable time and effort for only a meager reward. However, events of this type are annotated when they are at the foreground of the article in question (i.e. it is the event being reported upon). In this manner important events (such as courtroom verdicts) can still be annotated, while quotes and insertions by the reporters can be safely disregarded. Consider the examples below.

5. [Het contact met correspondent Rudy Vranckx werd verbroken terwijl [hij rapporteerde in Jemen]] *EN: Contact with correspondent Rudy Vranckx was broken while he was reporting in Yemen.*
6. [De rechter zal het vonnis op maandagmorgen uitspreken]] *EN: The judge will provide the final verdict on monday morning.*
7. Het Laatste Nieuws rapporteerde op maandag dat de laatste resultaten er positief uitzagen.] *EN: Het Laatste Nieuws reported on monday that the latest results seemed positive.*

For this set of examples we distinguish a clear difference in the informational value between 1 and 2 on the one hand and 3 on the other hand. While the first two verbs of reporting signal an important event within the context of the article, the third example does not.

A second modification to the existing annotation scheme is a reduction of the number of argument subclasses. While a more nuanced definition of event arguments can ultimately benefit the search for events that corefer, it also results in more complexity being added to an already complicated task. For instance, the ECB+

corpus guidelines distinguish four subsidiary classes for the EVENT-TIME argument alone(*Date*, *Time of Day*, *Duration* and *repetition*), based on the TimeML annotation scheme (Pustejovsky et al., 2003). Attempting the extraction and classification of event arguments to such detail might still be somewhat premature for a low-resource language such as Dutch (Colruyt et al., 2019b). We therefore remove the subsidiary classes for the ACTION, TIME and LOCATION components in the annotation scheme. In addition to this, we distinguish only two subcategories for the HUMAN-PARTICIPANT and NON-HUMAN-PARTICIPANT tags: *Named Entities* and *Non-named Entities*.

The third modification we apply is the annotation of a set of *event properties* (Colruyt, 2020). For each event that is annotated in the text, annotators are asked to decide on three distinct characteristics of said event. Firstly, a distinction is made between *main* and *background* events. This property reflects the event's role within a given article. The main event forms the backbone of the article in question and is the reason why the article was written whereas the background event is only present within the document to provide some context or supplemental information to the main event. Secondly, the *realis* property is annotated. A *certain* realis denotes that the event has or has not occurred in the past or will certainly or not occur in the future, while an *uncertain* realis signifies that there is serious doubt on whether or not the event has occurred or will occur in the future. The realis property allows us to distinguish between the hypothetical and the real, which is particularly useful as our definition of events explicitly allows the inclusion of hypothetical events. Consider the examples below. Examples 3 and 5 would be marked as *certain*, whereas example 4 is to be marked as *uncertain*.

8. [Duitse president weigert wet te ondertekenen]^{Certain} *EN: German president refuses to sign law*
9. [Maradona komt volgende maand misschien naar Limburg en België]^{Uncertain}. *EN: Maradona might come to Limburg and Belgium next month*
10. [België gaat door naar de halve finale van het WK]^{Certain}. *EN: Belgium advances to the semi-finals of the World cup*

Finally, annotators are asked to associate a sentiment from the following list to each event: *positive*, *negative*, *neutral* or *conflict*. Sentiment is annotated from the reader's perspective i.e the emotion evoked after perceiving the event. Any implicit opinions that may be present in the way the event is framed in the text are not considered. Furthermore, annotators are asked to judge the real-world events in the text from a European/Western point of view. Logically, the positive tag is used when an event is considered to be positive, while the negative tag is used when the opposite is true. The conflict tag is reserved for politically sensitive events where the annotator's own political stance might influence judgement[2]. Note that the sentiment annotated within the documents is mainly implicit, as news articles generally lack the

[2] More details on the annotation of implicit sentiment for events can be found here https://github.com/Cyvhee/ImplicitSentimentAnnotations.

explicit sentiment cues that more subjective texts have. Sentiment was annotated not only because it could help in the task of ECR, but also because if we want to create more fine-grained content-based news recommenders it is crucial to also be able to detect implicit polarity or the level of controversy or sentiment.

3.2.2 Coreference annotation

In addition to the changes regarding event and argument annotation discussed in the previous section we also adapted the annotation scheme for the annotation of coreference relations, both at the entity/argument and event level.

Coreferential relations between entities and arguments can be quite nuanced. Taking into account this more fine-grained information can be useful from a linguistic point of view and might aid the establishment of coreference links between arguments in the future. We distinguish three possible links between event arguments: identity, part-whole and type/token (Ng, 2017). First, identity relations are very straightforward. Two entities are in an identity relation when they refer to exactly the same real world entity.

11. De laatste e-mails van **de leraar** aan de schooldirectie voor **hij** werd onthoofd. *EN: **The teacher's** final e-mails to the school board before **he** was decapitated*

Second, part/whole coreference links are established when one of the arguments is connected to another argument, but only to a part of it.

12. **De auto** raakte van de weg of omstreeks half 9. Getuigen zeiden dat **de lichten** niet werkten. *EN: **The car** slipped of the road at around eight thirty. Witnesses said that **the lights** weren't operational*

Third, two arguments can also be linked through a type/token relationship. In this case, two arguments refer to the same object type but have a different token. In other words, the arguments do not refer to the same real world object, but to one of a similar description.

13. Premier Michel koos op het fotomoment voor **de blauwe vlag**, terwijl Tom Van Grieken naar **de gele** greep. *EN:Prime minister Michel chose a **blue flag** for the photo op, while Tom Van Grieken went for **the yellow one***

The final modification is a more nuanced annotation of event coreference relations. When deciding on whether or not two event mentions refer to the same real-world event, three criteria are typically set: Events should occur at the same time (i), in the same place (ii) and the same participants should be involved (iii). When these three criteria have been fulfilled, an event coreference link is made as we can be assured that both mentions fully refer to the same real-world event and to each other. This is typically called an *identity* link. However, some cases, which at first glance seem to satisfy the conditions for the establishment of a coreference identity link, do require some further examination. Consider the examples below:

14. (a) [Politieke aardbeving in Israël][1] *EN: Political earthquake in Israel.*
    (b) [De Israëlische premier Ariel Sharon heeft zijn lidmaatschap van de Likoedpartij opgezegd][2] en [het ontslag van zijn regering aangeboden][3]. *EN: The Israeli premier Ariel Sharon has terminated his membership of the Likud party and proposed the resignation of his government.*

There are three event mentions in this example. We can draw coreference links between both mention 1 and 2 and mention 1 and 3, respectively. However, because both *Het lidmaatschap opgezegd/terminated his membership* and *Het ontslag van zijn regering/resignation of his government* contribute to the event *politieke aardbeving/political earthquake* it is hard to assign an identity relation to these links. It is thus better to say that both mention 2 and 3 relate to mention 1 in a *part-whole* structure, as mention 1 constitutes their combined individual event actions. We therefore distinguish two types of event coreference relations: the *identity* relation and the *part-whole* relation.

The sections above described how the widely popular ECB+ annotation scheme was adapted to our own needs. For a more detailed and complete explanation of our annotation process please refer to the final version of the annotation guidelines (De Langhe et al., 2021).

3.3 Annotation process

Six annotators (all graduate students in Applied Linguistics) were hired over a two-month period. Each annotator worked part-time and was assigned 20 event clusters for annotation, which corresponds to an average of around 200 news articles of varying length. Annotators were given the following step-by-step guidelines:

1. Read through the document and highlight all full event mention spans
2. For each annotated event span, determine the *event properties*
3. Annotate all *event arguments* for each of the highlighted events
4. Annotate argument coreference links and subtypes
5. Once all events in a given cluster are annotated, establish event coreference links

The annotators worked at their own pace and individually, with occasional calls for advice from an expert supervisor. The event identification and entity coreference tasks were completed with the WebAnno annotation tool (Sarwar et al., 2001), whereas the cross-document annotation of events was performed using the knowledge base structures in the Inception language annotation tool (Rubin et al., 2015). After an initial training period to get familiar with the task, all annotators were presented with the same set of articles in order to determine inter-rater reliability.

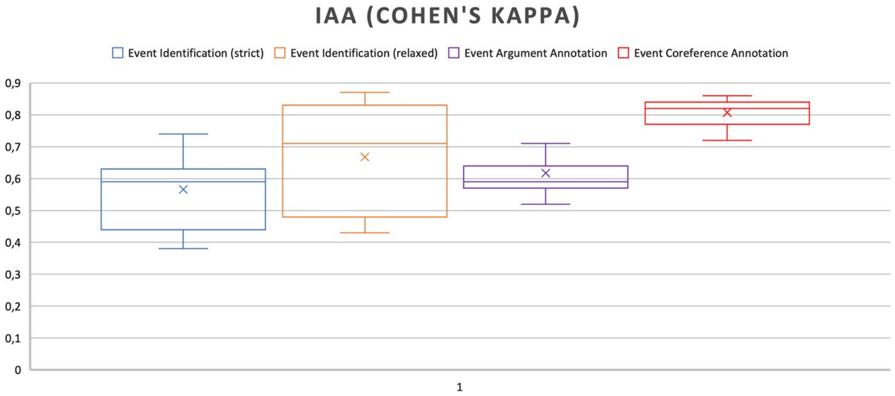

Fig. 1 Graphical representation of computed agreement for each of the annotation tasks

3.4 Inter-Annotator agreement

In order to determine the inter-annotator agreement (IAA) scores on the event annotation task, two clusters were set aside, which comprise a total of 25 articles. For the coreference task, IAA was studied by setting aside two other clusters, containing 21 articles in which events and arguments had already been annotated.

Three main components of the corpus annotation were evaluated. The first component entails the event identification. Annotators were asked to select the full mention span of any event they encounter that satisfies the restrictions presented in the annotation guidelines. Because span annotations are sometimes problematic, as two annotators might highlight the same event with a slightly different span, we chose to evaluate the annotations using both a strict and relaxed matching mechanism. The strict matching mechanism considers the span annotation of both annotators and considers them equal if and only if the full strings of both spans match. Conversely, relaxed matching allows a buffer zone of two tokens between strings for the calculation of the IAA event mention scores. The example below illustrates this and represents how an event which was annotated differently by two annotators (A and B) is still considered as a match.

15. [[Mensen worden opgepakt zonder degelijk onderzoek]^{B} of reden]^{A} *EN: People are getting arrested without any thorough investigation or reason*

Annotator A selected *Mensen worden opgepakt zonder degelijk onderzoek of reden* as the event span in this case while annotator B highlighted *Mensen worden opgepakt zonder degelijk onderzoek*. In this case, despite not having a full string match, the two mention spans are considered to be the same in regards to the IAA calculation.

The second component relates to the annotation of event arguments and monitors the agreement between annotators on the token selection of the following overarching event argument classes: ACTION, TIME, LOCATION,

Table 3 Comparison of various event coreference corpora

Corpus	Documents	Topics	Event mentions
ECB (English)	482	43	1744
ECB+ (English)	982	43	14884
MeanTime (translated Dutch)	120	4	1510
ENCORE (Dutch)	1115	91	15407

HUMAN-PARTICIPANT and NON-HUMAN-PARTICIPANT. In contrast to the event identification task, no relaxed matching mechanism is applied here.

Finally, annotation of cross-document event coreference was evaluated by presenting the annotators with a separate set of documents in which gold-standard event mentions and arguments were already annotated. The annotators were then asked to establish coreference links between the gold-standard event mentions.

Because our annotation process contains many elements and we employ a relatively high number of annotators, calculating an interpretable IAA score is not straightforward. Figure 1 presents the Cohen's Kappa statistic Cohen (1960), a measurement which considers chance agreement, for each of the annotation tasks. Note that for the event mention span task, one annotator was always considered to be the gold standard annotation. The kappa scores that are presented are an average of the scores for each possible annotator pair (15 pairs total). The average scores of 0.67 (relaxed)/ 0.57 (strict) and 0.62 indicate a substantial agreement for the event mention annotation and event argument annotation tasks, respectively, while the average score of 0.80 for the event coreference task signifies very strong agreement. The boxplots reveal somewhat more variance among the annotations for the first component, i.e. event identification. A table with Cohen's Kappa scores for each of the tasks for all annotator pairings can be found in the appendix.

3.5 Corpus statistics

In total, 1,115 documents were selected for annotation, of which 1,087 contained at least one news event. A total of 15,546 news events and 35,387 arguments were annotated in the corpus. The number of event coreference chains i.e the amount of event clusters that contain two or more events totals to 2,504. This corpus is thus larger than the corpora presented in Table 1, both in terms of actual size (number of documents) and in terms of the total number of event clusters. The ENCORE corpus was the result of a significant annotation effort that will hopefully provide a boost for ECR research in Dutch and low-resourced languages in general. Table 3 provides additional information regarding the size of the corpus presented in this paper when compared to the major English event coreference corpora and the translated Dutch MeanTime corpus.

The following sections provide a more detailed analysis of the corpus' composition, based on its different layers and aspects. From this table we can derive that the ENCORE corpus is on par with the popular ECB+ corpus.

Table 4 Distribution of the various event properties in the ENCORE corpus

Feature	Value	#
Prominence	Main	1866
	Background	13541
Realis	Certain	15026
	Uncertain	373
Sentiment	Positive	217
	Negative	869
	Neutral	14183
	Conflict	165

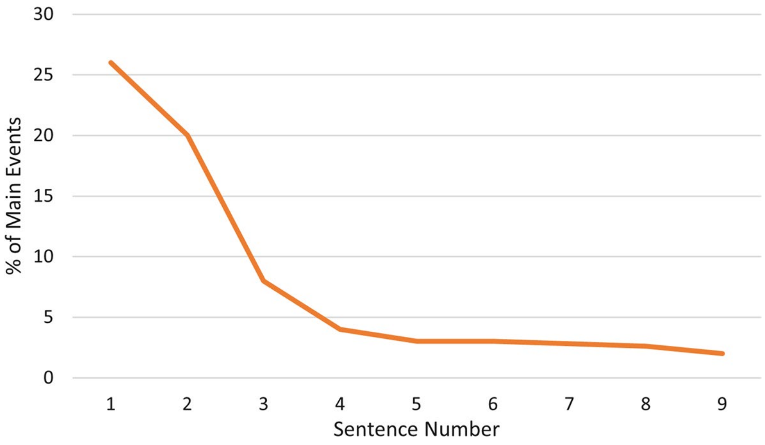

Fig. 2 Distribution of main events by document position

3.5.1 Event annotation

As stated before in Sect. 3.2, a real-life event can generally be represented by syntactic clauses, infinitival constructions or noun phrases. Of the 15,407 events in the corpus, the overwhelming majority has a verbal component. Syntactic clauses make up 63% of all events, while infinitival constructions account for 13% of the annotated mentions. The final 24% of events is represented by a noun phrase.

Table 4 represents the prevalence of the various properties that were annotated for each event. Most mentions are marked as *background* events. This is unsurprising, as most news articles are structured in a way where the *main* event is only briefly referred to in the title paragraph or lead. The rest of the article is then composed of a series of happenings or opinions that provide context to that key event. Figure 2 provides a visualization of the way in which news articles are typically structured. Interestingly, up to 48 percent of the time, the article's main event is found in the first two sentences, while only around 28 percent of main event mentions are found from sentence 10 onward. The average length of a newspaper article in the corpus

Table 5 Distribution of arguments and argument types

Argument	Type	Number	% of total
Time/location	Time	5236	61
	Location	3300	39
Human participants	NE PER	3011	22
	NE ORG	1376	10
	NE LOC	502	3
	Nominal	9340	65
Non-human participants	NE	533	4
	Nominal	12017	96

is around 45 sentences. At first glance these figures might lead one to assume that annotation of coreference in these documents will be extremely sparse because articles are mostly composed of ‘background information’. However, it should be noted that neither Table 4 nor Fig. 2 take into account that background events in some articles might correspond to main events in other articles and vice versa.

When examining the annotation of the *realis* property in the corpus we find that the large majority of the event mentions is marked as *certain*. This was to be expected as news articles typically wish to inform readers on events that have happened in the past. Upon closer inspection we noticed that almost all event mentions that were assigned the *uncertain* tag belonged to two specific types of newspaper articles: either forecasts of election results or of sporting events.

It is perhaps somewhat surprising that up to 92 percent of all event mentions are marked with the *Neutral* sentiment tag, especially compared to other corpora with a similar setup. For instance, one study focuses on the annotation of implicit polarity in the events annotated in the EventDNA corpus (Van Hee et al., 2021), which was composed out of the same larger pool of documents. When annotated, these events show a lot more inherent polarity, as up to 11% of events are judged to be positive, 48% are seen as negative, 36% are considered neutral and the remaining events are ascribed to the *conflict* tag. Two possible explanations can be found for this observation. First, in the EventDNA corpus only the article title and lead paragraph were annotated. Typically, titles and leads in newspaper articles are framed in a more provocative way, as to draw in the reader and incite some type of emotional response (Horne and Adali, 2017). Second, the majority of these earlier studies tend to focus on what is known as geopolitical ’hard news’. As stated before, however, our ambition is to extend event coreference resolution systems to more local, trivial news articles for day to day use. Many of the events in our dataset thus lack the more outspoken sentiment that is present in these ’hard news’ articles, which are often polarizing by their very nature. Consider the examples below.

16. Bij de vulkaanuitbarsting in het zuidwesten van Guatemala zijn al zeker 70 mensen omgekomen *EN: At least 70 people have died after a volcanic eruption in the south-west of Guatemala*
17. Marnix Peeters brengt nieuw boek uit *EN: Marnix Peeters releases new book*

Table 6 Coreference relations in the ENCORE corpus

Type	Number	% of total
(a) Distribution of entity coreference relations		
Identity	4258	82
Part-whole	791	15
Type/Token	108	3
	5157	100
(b) Distribution of event coreference relations		
Identity	9799	90
Part-whole	1073	10
	10,872	100

Naturally, in the first example most newspaper readers will immediately perceive the devastating volcanic eruption, and the loss of life that followed it to be negative. This is contrasted by example 2, which originates from an article describing the award show of a Dutch literary prize. Arguably, fans of the Flemish writer Marnix peeters will consider this event to be positive, whereas others might perceive this as negative. However, it is safe to assume that most newspaper readers will have no strong feelings about this event either way and thus consider it neutral (Table 5).

Table 6 displays the spread of the various event arguments that are annotated in the corpus. As stated in Sect. 3.2, time and location are important markers in the identification of event coreference. Moreover, further analysis reveals that a total of 6,699 event mentions, or around 43% of the entire corpus, have at least one time or location marker. When further examining the *Human participant* and *Non-Human participant* arguments two observations can be made. First, of all arguments marked with the *Human Participant* tag, around 35% are named entities, while the remaining cases are nominal constituents. Second, somewhat unsurprisingly, only 4% of non-human participants are named entities. This can be easily explained by the fact that most participants in this class are either lifeless objects or animals. Note that the Named entity location tag (NE LOC) that falls under the human participants category might be somewhat confusing at first glance. This type of annotation deals with metonymic usage where geographical locations are used to refer to the people that live in said locations (Desmet and Hoste, 2014), as illustrated by the example below.

18. Polen maakt zich zorgen over Wit-Russische druk aan de grens *EN: Poland worried about Belarusian pressure at the border*

3.5.2 Coreference annotation

As previously explained coreference was annotated at two levels. First, entity coreference between all participant arguments was annotated within each document. Second, event coreference was annotated both within and across documents. The former is particularly useful when keeping into account the eventual objective of creating a full event coreference resolution system, as events which corefer logically also have arguments which corefer. Performing coreference resolution between arguments can thus be a useful first step in this respect. However, coreference resolution between event arguments should be distinguished from regular entity coreference resolution. In regular entity coreference all markables that refer to real-world entities are considered for resolution, while in event argument coreference resolution only markables present in events are considered.

After annotation, our corpus comprises 10,546 argument coreference chains that consist of two or more arguments and 2,605 event coreference chains that contain two or more events in the corpus. Of those event coreference chains, 1018 were unique intra-document chains i.e chains only found within a single document and 1587 were cross-document chains i.e chains spanning multiple documents. On average, a cross-document chain contained events of 4 different documents. Table 6 displays the distribution of argument and event coreference link types for every argument (a) or event (b) contained in the respective coreference chains. One might argue that the total number of coreference links at the entity level is (relatively) low. However, entity coreference was only annotated at the intra-document level into entity chains and additionally, only entities that served as arguments for the annotated events were considered. This results in a lower number of entity links when compared to more general entity coreference corpora.

4 Preliminary experiments

As stated before, performing event coreference resolution reliably is a daunting and complicated task, even in high-resourced languages (Lu & Ng, 2018). One element that can aid to establish coreference links between two mentions is an analysis of the accompanying arguments. Intuitively, two events that corefer will have similar entities participating in them. While one could argue that most traditional entity coreference algorithms should be able to resolve coreferential arguments at first sight, two notable problems arise. Firstly, syntactic structures that function as arguments within an event are not always what one would traditionally label as an entity, or are composed of multiple entities. Two such examples are presented below.

19. [[Het contact met correspondent [Rudy Vranx]^{Entity}]^{Argument} werd [verbroken]^{Action}]^{Event}. *EN: The contact with correspondent Rudy Vranx was broken off*
20. [[[Haar]^{Entity} [boek]^{Entity}]^{Argument} wordt volgende maand gepubliceerd] *EN: Her book will be published next month*

Secondly, in our annotated corpus, like many other corpora focusing on event-centric tasks, arguments and entities are only labeled when they occur within an event. Traditionally, entity coreference resolution systems employ proximity-based features in order to resolve most entities based on a set of possible antecedent candidates. In our dataset, however, the aforementioned set of antecedents is virtually always incomplete rendering this type of approach less effective. In order to deal with this specific problem, we modify an entity coreference system to be able to resolve argument resolution more effectively and compare it to two traditional entity coreference approaches.

4.1 Methods

We take a rule-based multi-pass sieve approach for entity coreference resolution (Raghunathan et al., 2010) and adapt it to our purposes. The reason for choosing such a rule-based system is two-fold. Firstly, machine learning and neural approaches may be hampered by data sparseness. As previously mentioned, argument mentions are much less frequent compared to textual entities. Secondly, while subtasks like argument coreference can benefit future work and give us a better understanding of event coreference, they are not the central goal of this project and thus it might be better to prioritize straightforward and easy-to-deploy rule-based methods, as opposed to learning-based methods which often require lengthy training processes and, in some cases, extensive feature engineering. A final argument in favour of rule-based approaches is their observed performance. Despite the emergence of machine learning systems and neural models, rule-based systems still attain state-of-the-art performance in some coreference tasks. This includes systems that are entirely based on rules and hybrid approaches (Lee et al., 2017).

For the experiments we rely on gold-standard arguments. First, named entities are extracted from the documents using the LeTs preprocessing toolkit (Van de Kauter et al., 2013b). Second, all named entities that are not part of an event are filtered out, as we are only interested in event arguments for this coreference resolution task. We then compare the resolution of identity coreference relations with two other Dutch entity coreference resolution systems. The first is a Dutch version of the aforementioned multi-pass sieve approach (van Cranenburgh, 2019)[3], which will serve as a baseline model for entity coreference. The second method is a gradient-boosting machine learning algorithm trained as mention-pair model. Both are described in closer detail below.

4.1.1 Multi-pass sieve approach

Most multi-pass sieve algorithms, including the baseline model (van Cranenburgh, 2019) and our own model attempt to create coreference chains (clusters) in the following manner. At first, each argument is considered to be a singleton cluster. For each of these clusters an antecedent list is generated. This list consists of the heads of each cluster preceding it. Subsequently, each mention is processed by a given sieve and tested against each of the candidate antecedents. If, through the set of predefined rules in the sieve, coreference can be resolved for an antecedent-mention pair, the cluster that this mention belongs to is added to the antecedent cluster and

[3] https://github.com/andreasvc/dutchcoref, v0.1, 22/03/21.

Table 7 Multi-pass sieve architecture

Sieves
1. Exact string match
2. Partial string match
3. Alias detection
4. Precise construct
5. Head matching
6. Head synonymy
7. Pronoun resolution

the cluster list is updated. This is done procedurally, one sieve at a time, until all data has passed through each individual sieve. Finally, the cluster list containing the coreference chains is returned. Note that only the heads of the clusters are processed, which greatly improves the efficiency of the algorithm. As an additional step, before the sieves are passed through, each cluster is parsed using the Dutch Alpino parser (van Noord, 2006), and assigned a set of cluster properties such as number, gender, and entity type. No two clusters can be merged when there is a critical mismatch of these properties.

The paragraphs below explain each of the individual sieves of our modified multi-pass sieve model in more detail, see Table 7 for an overview, and provide an intuition as to why certain modifications were made from the original entity coreference algorithm. Concretely, two new sieves were added to the original model: the *Alias sieve* and the *Head synonymy sieve*, while the *pronoun resolution sieve* was extensively reworked.

Exact String Match Stop words are first removed from both the candidate and antecedent mentions. Two mentions are assigned to the same coreference chain when there is a complete overlap of their surface form. Note that pronominal mentions are excluded from this sieve and are not considered until sieve 6.

Partial string Match Two mentions are assigned to the same coreference chain when the surface form of one of the mentions can be fully found within the other mention.

Alias detection In cases where the mention is a proper name a lexical lookup is conducted on Wikipedia. If the other mention corresponds to one of the alternative names in Wikipedia's alias property, a coreferential chain is formed.

Precise construct This sieve attempts to match precise sentence constructions based on the sentence's parse tree. Two notable examples are appositive constructions and acronyms. The example below demonstrates a typical appositive structure.

21. Het hof van beroep in Antwerpen heeft de Belgische nationaliteit afgenomen van [Fouad Belkacem]^{HumanParticipant}, [de gewezen leider van terreurbeweging Shariah4Belgium]^{HumanParticipant} *EN: The Antwerp Court of Appeal has revoked the Belgian nationality of Fouad Belckacem, former leader of terrorist group Sharia4Belgium*

Note that precise construct matching in earlier studies typically included patterns such as predicate nominals and reflexive and reciprocal pronouns (Ng, 2017). However, due to the manner in which events and arguments are annotated in this corpus predicate nominal constructions are very rare and reflexive pronouns almost never figure as arguments on their own. Therefore, including these additional construct patterns would have no noticeable effect.

Head Matching Mentions are parsed using the Alpino dependency parser (van Noord, 2006) and their syntactic head is determined. Two mentions with the same syntactic head are assigned to the same coreferential chain when there is no conflict in their respective modifiers.

Head synonymy This sieve serves mostly the same purpose as the previous one. Syntactic heads are determined for both antecedent and candidate mentions. Then, the synonym sets (synsets) are determined for both heads using the OpenDutchWordnet lexical database (Postma et al., 2016). If the synsets of both heads overlap and there is no conflict in their respective modifiers, a coreferential link is formed.

Pronoun resolution Pronouns present an interesting problem in resolution applications, as usually some knowledge of the real world and context is needed to be able to link a pronoun with its corresponding antecedent. However, from analysis of our annotated corpus we know that the pool of arguments that correspond to real-world entities in a single article is rather limited. We can use this knowledge to create a greedy heuristic with which we can resolve the coreference of pronouns in a relatively high number of cases. When an entity is recognized as a pronoun, all antecedent mentions are ranked based on their proximity and the candidate mention is assigned to the first cluster with which it has both a gender and number agreement. If no suitable candidate is found within a distance of 5 sentences, the pronoun remains a singleton cluster. The latter sounds very counter-intuitive as almost all research into entity coreference presupposes that every pronoun has at least one non-pronominal antecedent (or a postcedent in the case of a cathaphoric relation). However, as the eventual goal here is to aid the resolution of events and because not all entities in the text correspond to event arguments it is possible that some pronouns completely lack a suitable antecedent. Consider the example below as an illustration:

22. Barack Obama is het het niet eens met het huidige beleid van President Trump. *EN: Barack Obama does not agree with president Trump's current policy.*
23. [[In het nieuwe boek dat hij volgende maand voorstelt] zal die onenigheid ook uitgebreid aan bod komen]] *EN: Those disagreements will prominently feature in the new book that he will present next month*

While the pronoun *hij* in sentence 2 clearly refers to the entity *Barack Obama* in sentence 1, no coreference link is established here. No events are annotated in sentence 1 on the ground that its content refers to a *state of being* rather than a real-world event.

Table 8 Results for the argument coreference experiments using the three approaches: a baseline multi-pass sieve (I), a mention-pair (II) and an adapted multi-pass sieve (III) approach

Approach	MUC	B	CEAFe	LEA
	F1	F1	F1	F1
(a) *Evaluation including predicted singleton clusters*				
I	0.62	0.59	0.61	0.58
II	0.66	0.65	0.68	0.62
III	0.69	0.71	0.70	0.64
(b) *Evaluation excluding predicted singleton clusters*				
I	0.62	0.55	0.60	0.54
II	0.66	0.59	0.63	0.60
III	0.69	0.69	0.70	0.61

4.1.2 Gradient boosted mention-pair approach

In a mention-pair model all entities (or arguments in this case) are paired and the model is tasked with the binary decision on whether or not the mentions refer to the same real-world person, group or object. We use the popular XGBoost algorithm (Chen et al., 2015) trained using 10-fold cross-validation and extensive hyperparameter tuning. The model was trained on a subset of the Dutch SoNaR corpus (Oostdijk et al., 2013). The subset used for training (SoNaR-1) contains 1 million Dutch words, rigorously annotated with named entity and coreference information, coming from a diverse number of sources. The final model configuration can be found in the appendix. Each mention pair is represented by a set of well-performing entity coreference features (Hoste, 2005) for Dutch. The paragraphs below broadly describe the type of features used for this task, as well as the general intuition behind them. A table detailing each individual feature that was used can be found in the appendix. Syntactic, semantic and morphological features were extracted using the Dutch dependency parser Alpino (van Noord, 2006).

Distance features are a set of positional characteristics that detail the sentence distance and number of noun phrases between each candidate anaphor and antecedent pair. Additionally, a binary feature that indicates whether or not the mentions occur within 3 sentences of one another.

String Matching features are a set of binary characteristics that indicate whether or not the surface forms of the antecedent and candidate completely match, partially match or share the same syntactic head. In addition to this, when the candidate anaphor and antecedent are both proper names it is determined whether or not they are an alias of one another.

Morphological Features indicate whether or not the candidate anaphor and antecedent belong to a certain part-of-speech class. Binary features are added for pronouns and proper names as well as reflexive and demonstrative pronouns specifically. In addition, a binary feature for numerical agreement is also included. This agreement feature takes no value when numerical agreement or disagreement cannot be determined.

Table 9 The accuracy score of each individual sieve

Sieve	Sieve accuracy (%)
Exact string match	97.3
Partial String match	83.2
Alias Detection	88.0
Precise construct	96.4
Head Match	86.4
Head synonymy	93.7
Pronoun Resolution	77.2

Syntactic features detail to which syntactic class (subject/object) the candidate antecedent and anaphor belong respectively. These features also specify whether or not an anaphor and antecedent are found within precise syntactic constructs, such as appositives.

Semantic Features are a set of binary characteristics that specify the synonym/hypernym relations between the candidate anaphor and antecedent. In addition to this, this set of features looks at the named entity type of the mention pairs.

4.2 Results

We evaluate the modified multi-pass sieve and mention-pair approaches against the baseline entity coreference model. As stated before, the mention-pair model was first trained on the Dutch SoNaR-1 corpus, which is composed of more than one million words and in which entity coreference relations are annotated. The model was trained using 10-fold cross validation and then evaluated on our own corpus data. The multi-pass sieve models were simply evaluated on the corpus data, as they do not require any training. Table 8 displays the results for the argument coreference resolution task on gold standard data for the baseline multi-pass sieve algorithm (approach I), mention-pair model (approach II) and the modified multi-pass sieve algorithm (approach III). We use four scoring mechanisms that are typically used for coreference evaluation: MUC Vilain et al. (1995), B Bagga and Baldwin (1998), CEAFe Luo (2005) and the more balanced LEA Moosavi and Strube (2016). We would like to repeat that argument coreference was only annotated for within-document settings.

An important decision to make when evaluating coreference resolution systems is whether or not to include singleton clusters i.e entities that are predicted to be in coreference chains with a size of one. Including these free-standing entities is known to somewhat inflate some of the metrics commonly used for evaluation. Concretely, this means that the MUC score might be the best performance indicator for this task when taking singletons into account, as it is more robust against correctly predicted singleton clusters (Cai & Strube, 2010), Whereas B and CEAF metrics might be less suitable in this case. Even the LEA metric, which has long been thought of as providing the most fair assessment of coreference resolution has been shown to be

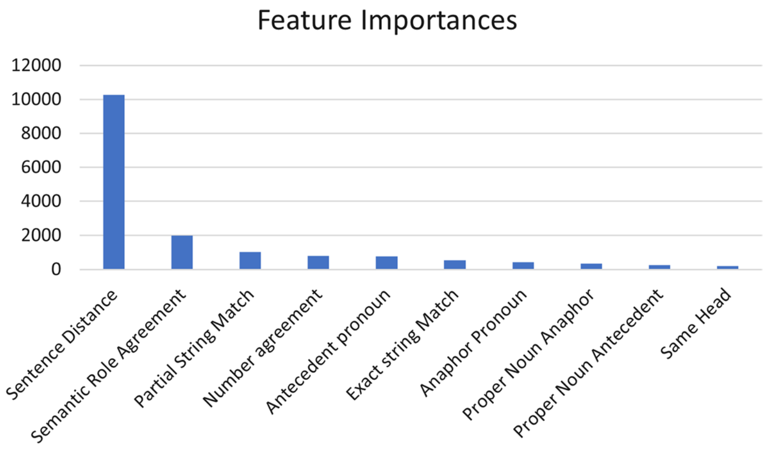

Fig. 3 Importance scores for each of the 13 most important features

prone to some distortion when singletons are present (Poot & Cranenburgh, 2020). For the sake of completeness, we opt to present an evaluation of our coreference system both with and without singleton clusters. In the latter case, singletons are predicted, but are simply removed before calculating the evaluation metrics. For the evaluation including singletons it is demonstrated the task benefits from the minor modifications made to the multi-pass sieve approach (MUC of 0.69 compared to one of 0.62 for the baseline). This could mean that existing NLP methods can be easily adapted to be used as upstream components for event coreference resolution. Lea scores are slightly lower compared to the other evaluation metrics, this is consistent with earlier findings (Moosavi & Strube, 2016). The evaluation scores when singletons are excluded are understandably slightly lower, as earlier explained, but are generally consistent with our earlier observations.

4.3 Error analysis

As results of coreference tasks can be hard to interpret at times, a more detailed analysis of the models' performance might shed some light on the conclusions we can draw from Table 8.

First, Table 9 details the accuracy score of the individual sieves in the modified multi-pass sieve model. As is expected, the string matching, alias and precise construct sieves perform very well. Most of the arguments in the corpus consist of nominal groups. The head synonymy sieve, which was implemented to deal with these arguments specifically also performs remarkably well, with 93.7 percent of antecedent-anaphor pairs classified correctly. The pronoun resolution sieve also works, especially given its simple and naive setup. It should be noted here that the choice to use gold standard argument mentions is a likely explanation for this. If this coreference model were to be embedded in a pipeline architecture with an argument detection aspect, accuracy for this sieve would surely suffer from the amount of inaccurate and undetected argument mentions. Therefore, should this method be used as a downstream task in larger event coreference resolution we should first ensure that we can detect event and argument mentions with great accuracy. In addition to this, the performance of this model compared to the established entity coreference algorithms can also be partly explained by our prior knowledge of the data, which allowed us to tailor the system to this very specific task. While this is a logical step to take when keeping in mind our eventual goal of a fully tailored event coreference resolution systems, it also means that the performance of this model would drop significantly when used for a general entity coreference task.

It is also useful to examine in more detail the results of the machine learning mention-pair model. Traditionally, mention-pair models based on well-established classification algorithms perform well in entity coreference tasks. Many studies have focused on the importance and effectiveness of hand-crafted features and distance features in particular are known to be very informative for entity coreference (Hoste, 2005). Figure 3 displays the informativeness of each feature for the trained mention-pair model. Feature importance was determined by simply calculating the amount of times a given feature was split on by the gradient boosted tree algoritm. From the figure we can infer that, much like in regular entity coreference tasks, distance features are among the most important features for classification. However, it should be noted here that distance feature primarily play a role in case of a negative decision by the classifier i.e entities with a large distance between them are much more likely not to corefer. Further error analysis reveals that cases where there is a relatively large distance between two coreferring entities are almost always wrongly classified. Among other things, We hope to improve classifier performance for these specific cases in future research.

5 Conclusion and future work

In this paper we have presented efforts to create the first large-scale cross-document event coreference corpus for unrestricted events in Dutch. We give an overview of state-of-the-art methods in event coreference resolution and the most widely used event corpora. For our own corpus, we drew inspiration from the popular English ECB+ dataset and made a series of modifications to its annotation scheme. Most notable are the annotation of implicit sentiment for events, a more thorough annotation of event coreference in which a distinction is made between *identity* an *part-whole* relations, annotation of the realis property and a more strict definition of which mentions are considered to be newsworthy events.

The annotation process was discussed and IAA experiments performed for three aspects of the annotation task: event identification, argument annotation and event coreference annotation. Despite the complexity of the task we managed to achieve strong to very strong agreement scores on each of the tasks. During the annotation process, 15,407 events were annotated in a total of 1,115 documents, making the final corpus comparable in size to the biggest English event coreference corpora. In addition, the corpus contains 91 distinct topic clusters, making it one of the most diverse corpora available for the resolution of unrestricted events. Following the publication of this paper, the dataset will be made freely available and we hope that this will stimulate research in event coreference for Dutch, as well as low-resourced languages in general (Table 10).

Existing studies on English event coreference often make use of pipeline architectures in which a series of upstream tasks (ranging from event detection to argument classification) are used to facilitate coreference resolution. As implementing such a full pipeline system for our corpus would be premature we have focused on a minor upstream task that demonstrates how research into event coreference resolution might develop. We performed event argument resolution using two rule-based multi-pass sieve approaches, of which one was specifically equipped to deal with this task, and a machine learning XGBoost algorithm with classical entity coreference features. While event argument resolution is closely related to entity coreference resolution and semantic role labeling, notable exceptions arise which might suggest that relying on well-established entity resolution systems is not optimal for this task. The results reveal that existing well-performing rule-based algorithms can be easily adapted for the specific tasks required in an event coreference resolution pipeline.

In the future, we hope to direct our research efforts toward full event coreference resolution for Dutch. Initially, focusing on pipeline-based architectures and later working towards an end-to-end system.

Appendix: IAA scores

See Table 10.

Table 10 Averaged Cohen's Kappa statistic for all annotator pairings

	Event mention span	Event arguments	Event coreference
A-B	0.71	0.58	0.72
A-C	0.87	0.75	0.82
A-D	0.52	0.62	0.80
A-E	0.83	0.64	0.76
A-F	0.77	0.60	0.83
B-C	0.70	0.57	0.76
B-D	0.43	0.52	0.86
B-E	0.72	0.60	0.84
B-F	0.83	0.83	0.79
C-D	0.48	0.58	0.86
C-E	0.84	0.71	0.81
C-F	0.72	0.59	0.83
D-E	0.48	0.56	0.77
D-F	0.44	0.52	0.82
E-F	0.68	0.59	0.84
Average	0.67	0.62	0.80

References

ACE English Annotation Guidelines for Events (v5.4.3). (2008). Linguistics Data Consortium.

Ahn, D. (2006). The stages of event extraction. Proceedings of the Work-shop on Annotating and Reasoning about Time and Events - ARTE'06 (July), 1–8. Retrieved from https://doi.org/10.3115/1629235.1629236

Aone, C., & Ramos-Santacruz, M. (2000). Rees: A large-scale relation and event extraction system. In *Sixth applied natural language processing conference* (pp. 76–83).

Bagga, A., & Baldwin, B. (1998). Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference* (Vol. 1, pp. 563–566).

Bejan, C., & Harabagiu, S. (2010). Unsupervised Event Coreference Resolution with Rich Linguistic Features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (July), pp. 1412–1422. Retrieved from http://www.aclweb.org/anthology/P10-1143https://doi.org/10.1162/COLI_a_00174

Bugert, M., Reimers, N., Barhom, S., Dagan, I., & Gurevych, I. (2020). Breaking the subtopic barrier in cross-document event coreference resolution. In *Text2story@ ecir* (pp. 23–29).

Cai, J., & Strube, M. (2010). Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the sigdial 2010 conference* (pp. 28–36).

Caicedo, R. W. A., Soriano, J. M. G., & , Sasieta, H. A. M. (2022). Bootstrapping semi-supervised annotation method for potential suicidal messages. Internet Interventions.

Chen, C., & Ng, V. (2016). Joint inference over a lightly supervised information extraction pipeline: Towards event coreference resolution for resource-scarce languages. In *Proceedings of the thirtieth AAAI conference on artificial intelligence*, February 12–17, 2016, Phoenix, Arizona, USA. (pp. 2913–2920). Retrieved from http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12413

Chen, C., & Ng, V. S. (2014). An end-to-end Chinese event coreference resolver [c].

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. (2015). Xgboost: Extreme gradient boosting. R package version 0.4-2, 1 (4).

Choubey, P. K., Raju, K., & Huang, R. (2018). Identifying the most dominant event in a news article by mining event coreference relations, 6.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Colruyt, C. (2020). Eventdna annotation guidelines.

Colruyt, C., De Clercq, O., & Hoste, V. (2019a). EventDNA: Annotation guidelines for entities and events in Dutch News Texts (v1.0) (Technical report).

Colruyt, C., De Clercq, O., & Hoste, V. (2019b). Eventdna: Guidelines for entities and events in Dutch news texts (v1. 0). LT3 Technical Report-LT3 19-01.

Cybulska, A., & Vossen, P. (2014a). Guidelines for ecb+ annotation of events and their coreference. In Technical report. Technical Report NWR-2014-1, VU University Amsterdam.

Cybulska, A., & Vossen, P. (2014b). Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of the ninth international conference on language resources and evaluation* (LREC'14), 4545–4552.

Cybulska, A., & Vossen, P. (2015). Translating Granularity of Event Slots into Features for Event Coreference Resolution. In *Proceedings of the 3rd workshop on EVENTS: Definition, detection, coreference, and representation* (pp. 1–10). Denver, Colorado: Association for Computational Linguistics. Retrieved 2019-02-20, from https://doi.org/10.3115/v1/W15-0801

De Langhe, L., De Clercq, O., & Hoste, V. (2021). Guidelines for annotating events and event coreference in Dutch News Articles (Technical Report).

De Marneffe, M.-C., Rafferty, A. N., & Manning, C. D. (2008). Finding contradictions in text. In *Proceedings of acl-08: Hlt* (pp. 1039–1047).

Desmet, B., & Hoste, V. (2014). Fine-grained Dutch named entity recognition. *Language Resources and Evaluation, 48* (2), 307–343. Retrieved 2019, Dec 09, from http://hdl.handle.net/1854/LU-4246431https://doi.org/10.1007/s10579-013-9255-y

Elango, P. (2005). *Coreference resolution: A survey*. Madison, WI: University of Wisconsin.

Horne, B., & Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the international aaai conference on web and social media* (Vol. 11).

Hoste, V. (2005). Optimization issues in machine learning of coreference resolution (PhD Thesis). Universiteit Antwerpen. Faculteit Letteren en Wijsbegeerte.

Humphreys, K., Gaizauskas, R., & Azzam, S. (1997). Event coreference for information extraction. In *Proceedings of the ACL/EACL Workshop on Operational Factors in Practical, Robus Anaphora Resolution for Unrestricted Texts* (pp. 75–81).

Ji, H., & Grishman, R. (2011). Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 1148–1158).

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). Spanbert: Improving pretraining by representing and predicting spans. *Transactions of the Association for Computational Linguistics, 8*, 64–77.

Lee, H., Surdeanu, M., & Jurafsky, D. (2017). A scaffolding approach to coreference resolution integrating statistical and rule-based models.

Lee, K., He, L., & Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. arXiv preprint arXiv:1804.05392.

Liu, Y., & Lapata, M. (2019). Hierarchical transformers for multi-document summarization. arXiv preprint arXiv:1905.13164.

Lu, J., & Ng, V. (2016a). Event Coreference Resolution with Multi-Pass Sieves, 8.

Lu, J., & Ng, V. (2016b). Event coreference resolution with multi-pass sieves. In *Proceedings of the tenth international conference on language resources and evaluation* (lrec'16) (pp. 3996–4003).

Lu, J., & Ng, V. (2017). Joint Learning for Event Coreference Resolution (pp. 90–101). Association for Computational Linguistics. Retrieved 2018 Jul 02, from https://doi.org/10.18653/v1/P17-1009

Lu, J., & Ng, V. (2018, July). Event Coreference Resolution: A Survey of Two Decades of Research. In *Proceedings of the twenty-seventh international joint conference on artificial intelligence* (pp. 5479–5486). Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization. Retrieved 2019 Jan 14, from https://doi.org/10.24963/ijcai.2018/773

Lu, J., Venugopal, D., Gogate, V., & Ng, V. (2016). In *Joint inference for event coreference resolution*. COLING, 12.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing* (pp. 25–32).

Minard, A.-L., Speranza, M., Urizar, R., van Erp, M., Schoen, A., & van Son, C. (2016). MEANTIME, the NewsReader Multilingual Event and Time Corpus. In *Proceedings of the 10th language resources and evaluation conference* (LREC 2016) (p. 6). Portorož, Slovenia: European Language Resources Association (ELRA).

Mitamura, T., Liu, Z., & Hovy, E. (2015). Overview of TAC KBP 2015 Event Nugget Track. *Kbp Tac, 2015*, 1–31.

Moosavi, N. S., & Strube, M. (2016). Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th annual meeting of the association for computational linguistics* (volume 1: Long papers) (pp. 632–642).

Narayanan, S., & Harabagiu, S. (2004). Question answering based on semantic structures (Technical Report). International Computer Science Inst, Berkeley, CA.

Ng, V. (2017). Machine learning for entity coreference resolution: A retrospective look at two decades of research. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 31).

Nguyen, T. H., Meyers, A., & Grishman, R. (2016). New York University 2016 System for KBP Event Nugget: A Deep Learning Approach. Text Analysis Conference, 7.

NIST. (2005). The ACE 2005 (ACE 05) Evaluation Plan.

Oostdijk, N., Reynaert, M., Hoste, V., & Schuurman, I. (2013, November). The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch (pp. 219–247). https://doi.org/10.1007/978-3-642-30910-6_13

Poot, C., & van Cranenburgh, A. (2020). A benchmark of rule-based and neural coreference resolution in Dutch novels and news. In *Proceedings of the third workshop on computational models of reference, Anaphora and coreference*.

Postma, M., van Miltenburg, E., Segers, R., Schoen, A., & Vossen, P. (2016). Open Dutch WordNet. In *Proceedings of the eight global wordnet conference*, 300–307. Retrieved from http://wordpress.let.vupr.nl/odwn/

Pradhan, S. S., Ramshaw, L., Weischedel, R., MacBride, J., & Micciulla, L. (2007). Unrestricted coreference: Identifying entities and events in ontonotes. ICSC 2007 International Conference on Semantic Computing, 446–453. https://doi.org/10.1109/ICSC.2007.93

Pustejovsky, J., Castano, J., Ingria, R., Saurı, R., Gaizauskas, R., Setzer, A., & Katz, G. (2003). TimeML: Robust Specication of Event and Temporal Expressions in Text. *New Directions in Question Answering, 3*, 28–34.

Quine, W. V. O. (1985). Events and reification. Actions and events: Perspectives on the philosophy of Donald Davidson, pp. 162–171.

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., & Manning, C. (2010, October). A Multi-Pass Sieve for Coreference Resolution. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 492–501). Cambridge, MA: Association for Computational Linguistics. Retrieved 2019 Feb 27, from http://www.aclweb.org/anthology/D10-1048

Rubin, V. L., Chen, Y., & Conroy, N. J. (2015). Deception detection for news: Three types of fakes: Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology, 52*(1), 1–4. https://doi.org/10.1002/pra2.2015.145052010083.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001, April). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web* (pp. 285–295). Hong Kong, Hong Kong: Association for Computing Machinery. Retrieved 2020 Feb 26, from https://doi.org/10.1145/371920.372071

Su, M.-H., Wu, C.-H., & Shih, P.-C. (2019). Automatic ontology population using deep learning for triple extraction. In 2019 *Asia-Pacific signal and information processing association annual summit and conference* (apsipa asc) (pp. 262–267).

Sukthanker, R., Poria, S., Cambria, E., & Thirunavukarasu, R. (2020). Anaphora and coreference resolution: A review. *Information Fusion, 59,* 139–162.

van Cranenburgh, A. (2019). A Dutch coreference resolution system with an evaluation on literary fiction. *Computational Linguistics in the Netherlands Journal, 9,* 27–54.

Van de Kauter, M., Coorman, G., Lefever, E., Desmet, B., Macken, L., & Hoste, V. (2013). Lets preprocess: The multilingual lt3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands journal, 3*, 103–120.

Van de Kauter, M., Coorman, G., Lefever, E., Desmet, B., Macken, L., & Hoste, V. (2013). LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal, 3,* 103–120.

Van Hee, C., De Clercq, O., & Hoste, V. (2021). Exploring implicit sentiment evoked by fine-grained news events. In *Workshop on computational approaches to subjectivity and sentiment analysis (wassa), held in conjunction with eacl* 2021 (pp. 138–148).

van Noord, G. J. (2006). At last parsing is now operational.

Vermeulen, J. (2018). newsdna: Promoting news diversity: An interdisciplinary investigation into algorithmic design, personalization and the public interest (2018–2022).

Vilain, M., Burger, J. D., Aberdeen, J., Connolly, D., & Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Sixth message understanding conference (muc-6): Proceedings of a conference held in Columbia*, Maryland, November 6–8, 1995.

Vossen, P. (2018). NewsReader at SemEval-2018 Task 5: Counting events by reasoning over event-centric-knowledge-graphs, 7.

Yan, M., Xia, J., Wu, C., Bi, B., Zhao, Z., Zhang, J., & Chen, H. (2019). A deep cascade model for multidocument reading comprehension. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 7354–7361).