Guidelines for Annotating Events and Event Coreference in Dutch News Articles

version 1.0

LT3 Technical Report – LT3 21-01

Loic De Langhe Orphée De Clercq Véronique Hoste
LT3 – Language and Translation Technology Team
Department of Translation, Interpreting and Communication
Ghent University
URL: http://www.lt3.ugent.be[1]

February 9, 2021

[1]The reports of the LT3 Technical Report Series (ISSN 2032-9717) are available from http://www.lt3.ugent.be/en/publications/

Contents

Chapter 1

Introduction

Reading online news has become immensely popular over the past years and the number of news sources that provide services online has exploded. As a result of this, the popularity of research in computational technology that can help users navigate the enormous amount of content they are presented with on a daily basis has also been on the rise. News recommendation systems aim to provide readers with new content based on previous reading behaviour and simultaneously allow them to discover new points of view by presenting them similar articles of different news sources. An important aspect of such content-based recommendation systems is the ability to recognize key news events within an article and link those events to other articles. The ENCORE project aims to lay the groundwork for a Dutch content-based recommendation algorithm by employing advances in Natural language processing (NLP) to perform event coreference resolution. The goal of event coreference resolution is to determine which mentions of events in texts refer to the same real-world event.

In order to start developing an algorithm that can perform event coreference resolution we need to compile a large-scale Dutch corpus of news texts in which events and coreference links between them are annotated. Due to the ambition of the ENCORE project to implement event coreference resolution in a news recommendation system two additional difficulties are posed on the dataset. Firstly, in order for the news recommendation algorithm to inform readers as broadly as possible all types of events should be taken into account. Secondly, as the recommendation algorithm should provide alternate points of view from a variety of sources the coreference links in the corpus should span across different documents. The annotation task of the ENCORE data is organized in the following way. Our data consists of a large selection of Dutch newspaper articles from a variety of Dutch news sources discussing a broad number of topics. The first step of the annotation process is to identify all events in a given article. Next, coreference links between the events in said article are established. Finally, coreference links between events in different articles are also annotated. The annotation guidelines of the ENCORE project are largely based on the existing annotation schemes of the ECB+ (Cybulska and Vossen, 2014) and ACE corpora (noa, 2008), but slight adaptations have been made to accommodate this style of annotation for the Dutch language. In addition to this, we annotate certain specific event properties that we esteem will facilitate the eventual goal of engineering a content-based recommender system.

In the following sections we first detail how events should be identified and annotated (sections 2.1 and 2.2). We then specify how additional information regarding the events should be annotated (sections 2.3 and 2.4). Finally, we explain how to establish coreference links between different news events, both within a document (section 3.1) and across all documents (section 3.2).

Chapter 2

Event Annotation

2.1 Event Mention Extent

Events in news articles can be represented by syntactic clauses, noun phrases or infinitival constructions within a sentence. We call these structures event mentions. A first step towards linking events in texts is to identify all possible event mentions. Typically, a single event mention will not cross sentence boundaries i.e one event cannot belong to two different sentences. However, it is entirely possible that one sentence is composed of multiple event mentions. In addition to this, two or more event mentions in the same sentence can overlap, which can sometimes lead to a complicated web of events within a single sentence. As per illustration, the examples below show how an event can be represented by a syntactic clause, a noun phrase and an infinitival construction, respectively. Note that for the examples only one event mention is annotated per sentence even though some of the examples contain multiple event mentions.

1. [Vrouw ontwaakt in camera obscura].
2. [De Eerste wereldoorlog] werd vandaag herdacht in Ieper
3. Een getuige zag [het vliegtuig landen].

Event mentions themselves are often composed of an *event action* i.e a verb or noun phrase that denotes activity and a number 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, how. More detailed guides to identifying the action and arguments of an event can be found in sections 2.2 and 2.3. In the ENCORE project, event mentions are annotated in their entirety. Concretely, that means we select the full event action and all possible arguments of said action. Punctuation that signals the end of a sentence is not considered to be part of the event mention. It is to be noted here that relative clauses part of event arguments are also kept within the event mention.

When annotating events in a document an important factor to consider is scalability. As one can infer from examples 1-3, events occur very frequently and can take many forms. It is useful to pose some restrictions on what constitutes an event and when to annotate an event as annotating all of them would be a near impossible task. While the first question will be mostly answered in the following section, we will first discuss when one should annotate an event. As important and newsworthy events are the focal point of the ENCORE project, large events that have many arguments and hold a lot of information are most valuable to us. Therefore, events that span

(almost) an entire sentence should always be annotated. In addition to this, any events that are embedded in a larger sentence-spanning event should also be annotated. Any event that does not fall into these categories should not be annotated. The latter relates particularly to events in the form of noun phrases. Consider the examples below.

4. Eerder vandaag was er onduidelijkheid of het proces geldig was.
5. [We vallen af door [de economische crisis]]

In the first example, *het proces* is technically an event. However, the sentence's principal verbal actions *onduidelijkheid zijn* and *geldig zijn* are not events but states of being. Therefore, no annotation is required here. In contrast to 4, the nominal event *de economische crisis* in example 5 is annotated as it is both an event in itself and an argument of another sentence-spanning event. The realisation that a group of words can be both an argument and an event at the same time is quite valuable, and passing this information into the learning algorithm might improve its performance. Hence, annotation of these cases is required.

Sometimes, complicated situations arise when coordinate and subordinate clauses enter into the equation. Determining the scope of an event can be tricky when a sentence consists of multiple events that are linked through conjunctions as is the case in example 6.

6. [De zwaartste gevangenisstraf kreeg hij in januari 2016], [toen hij tot twaalf jaar cel werd veroordeeld].

In these cases, we choose to annotate two separate events. Note that the connector *toen* is included in the event mention as it relates to *januari 2016* and thus holds temporal information regarding the event. Conversely, conjunction words like *en* are rarely included in the event mention span due to the fact that they do not give any relevant information regarding the event.

7. [De lichte straf voor de beklaagde zette kwaad bloed] en [de familie van het slachtoffer verliet daarop de zaal].

However, certain cases of conjunction and subordination such as example 7 warrant a different approach.

8. [In oktober 2016 werd de procedure gestart [om hem zijn Belgische nationaliteit af te nemen]].

In this case, we annotate an overarching event that encompasses the subordinate clause. This is done due to the fact that the clause *om hem zijn nationaliteit af te nemen* can be seen as a part of the event in the main clause, as opposed to examples 6 and 7 in which the sentence contains two separate events.

2.2 Event Action

Once the full scope of the event mention is determined it is essential to highlight the event action. The event action can be seen as the "core" of the event. In syntactic clauses and infinitival constructions, the event action is usually a verb while in noun phrases the event action is denoted by

the head phrase. Naturally, one event mention can only contain a single event action. In the first part of this section we will discuss how one can reliable identify the core action in most events, including some exceptions where highlighting a single action might not be as straightforward. The second subsection will detail groups of verbs that are not to be annotated, despite a seemingly present event action.

2.2.1 Identifying event actions

When trying to identify the event action, look for the token(s) that provides the reader with the most lexical information regarding the occurrence. For the three examples that were given in section 2.1 we would annotate the event action in the mention in the following way:

9. [Vrouw [ontwaakt]^{Action} in camera obscura].
10. [[De Eerste wereldoorlog]^{Action}] werd vandaag herdacht in Ieper
11. Een getuige zag [het vliegtuig [landen]^{Action}].

Most event actions will be able to be identified with relative ease. However, there are three situations that one might encounter which warrant some additional explanation. Firstly, for event actions that are composed of a modal verb and a main verb we only tag the main verb with the *Action* tag, as it holds all the lexical information on the event action. Secondly, in Dutch we are often confronted with discontinuous verbs where the event action is composed of a core and a separate particle. In those cases, we annotate both the (lexical) core and particle(s) as demonstrated by the example below.

12. [President Trump [kwam]^{Action} gisteren [aan]^{Action} in de getroffen staat Californië].

Finally, one might encounter syntactic clauses or infinitival constructions where the main verb gives very little information on the event itself. While it is always preferable to have only one verb token as the core action of an event, an exception is made for these cases. Consider the example below:

13. [Fouad Belkacem, het gewezen kopstuk van de terroristische groepering Sharia4Belgium, [gaat in cassatieberoep]^{Action} tegen de intrekking van zijn Belgische nationaliteit].

In example 13 the verb *gaan* alone does not define the event action well enough. Therefore, some additional tokens are included in the *Action* tag. Note that these situations are few and far in between and one's primary instinct when annotating syntactic/infinitival events should always be to limit the event action to a single verb token.

2.2.2 Verb types that do not constitute events

The fact that we work with unrestricted events means we don't limit ourselves to specific types of events that have a predefined structure. This can complicate the identification of certain "borderline" cases, in which it is hard to determine whether or not a concrete event action is present. The sections below discuss certain situations that might be hard to classify at first glance.

Copulas and volition Candidate mentions that contain copular verbs (*zijn*, *schijnen*, *lijken*, *voelen* etc.) often pose problems. When trying to identify event mentions one should always ask themselves the following questions as a rule of thumb: "Is the action of this event clearly defined?", "Can you pinpoint the exact start/end of the event on a temporal scale?". If the answer to both of these questions is no, the candidate mention is most likely not an event. Consider the following example:

14. Fouad Belkacem voelt zich Belg en heeft geen band met Marroko.

Now consider this example from a different article regarding the same topic:

15. [Fouad Belkacem verzette zich hevig tegen het vonnis].

In the first case, the acts of *voelen* or *hebben* can hardly be classified as real actions, nor can we really indicate when they began. We do not identify any event in that sentence as they represent a state of being. The second example however, does clearly contain an event. In this case, the subject performed the action *verzetten*, which can be traced back to an exact moment in time.

Related to this are expressions of volition. As these are quite common in news texts it is useful to include them in this section.

16. Staatssecretaris voor Asiel en Migratie Theo Francken (N-VA) wil Fouad Belkacem zo snel mogelijk het land uitzetten als hij vrijkomt.

The example above should make clear that similarly to copulas, it is hard to pinpoint such expressions on a temporal scale as they present a state of being rather than an actual action. Therefore, they are not annotated as events. While *Fouad Belkacem zo snel mogelijk het land uitzetten* constitutes an event, it is not annotated. This is because we only consider an event when the main verb's action spans the entirety of the sentence. Since the main verb of this sentence is *willen*, no annotation is required.(cfr. 2.1).

Direct and indirect speech Besides copulas and expressions of volition, text segments that denote direct and indirect speech are also not annotated. Direct speech can be easily omitted from the annotation by disregarding any text between quotation marks. However, indirect speech is sometimes harder to identify. In most cases, indirect speech will be signaled by a verb that denotes some type of reporting. Consider the example below.

17. Fouad Belkacem zegt dat hij zich zal verzetten tegen de uitspraak.

When examining the example above, one might discern two separate events: *zeggen* and *verzetten*. 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. Therefore, no annotation is required. Much like example 15, we do not annotate the rest of the sentence either due to the fact that its main verb is not classed as an event action. Note that in some cases verbs of reporting does not signal indirect speech and is to be annotated after all. This happens when the action of reporting is at the focal point of the article such as in the example below:

18. [Het contact met correspondent Rudy Vranx werd verbroken terwijl [hij rapporteerde in syrië]].

2.3 Event Properties

Other than the event mentions and their arguments, we also annotate some additional information regarding the events in the articles.

2.3.1 Main and Background Events

It is not uncommon for news articles to refer to multiple events. Often times, the authors try to link certain events described in an article to other events that occurred in the past or are ongoing at the time of writing in order to provide the reader with some additional context and information. When annotating the articles, we determine one central or main event that forms the backbone of the article in question. When trying to find the main event always keep in mind the following questions: *Why was this article written?* or *What is this article exactly trying to inform me of?*. In most cases, looking at the title of the article will provide you with a good intuition regarding the main event. Once the main event of an article is established we mark it as such and then mark all event mentions referring to other events with the *background* type. When a main event is mentioned multiple times, all mentions of this event are naturally marked with the *Main* type. Consider the following examples that each consist of the title and a segment of the associated article:

19. (a) Neckermann vraagt bescherming tegen schuldeisers aan
    (b) [Neckermann vraagt de rechtbank om een procedure van gerechtelijke reorganisatie (WCO) op te starten]^{Main}. [De reisorganisatie, die [zwaar geraakt is door de coronacrisis]^{Background}, hoopt zo tijd te kopen om zijn schulden te heronderhandelen]^{Background}. 'We hebben een adempauze nodig.'

20. (a) Chileense burgers mogen nieuwe grondwet schrijven
    (b) [De Chilenen hebben zondag met een luide 'ja' laten weten dat ze een nieuwe grondwet willen]^{Main}. [Daarmee wordt de erfenis van Pinochet eindelijk begraven]^{Background}.

In some cases, it is not straightforward to determine the seminal event of an article. This can occur when the title of an article is so broad that it can encompass multiple events or when the author gives the same level of importance to several events. When this occurs, try to determine the event(s) that are absolutely crucial to the article. If none of the events can be singled out as the most important one, multiple main events may be included for this given article.

21. (a) Politieke aardbeving in Israël
    (b) [De Israëlische premier Ariel Sharon heeft zijn lidmaatschap van de Likoedpartij opgezegd]^{Main} en [het ontslag van zijn regering aangeboden]^{Main}.

2.3.2 Event Sentiment annotation

Implicit polarity of each event is also taken into account. For this task the annotator should determine whether an event is positive, negative or neutral. We largely base ourselves on the

sentiment annotation scheme that was proposed for the EventDNA project (Colruyt, De Clercq, and Hoste, 2019) It is important here to not only focus on lexical elements within the sentence but also to reflect on the event based the annotator's own intuition and world knowledge. However, certain events are politically charged and their annotation may differ based on the annotator's own preference or opinions. These events should be annotated with the *conflict* type instead, as to avoid inconsistencies in the annotation process.

22. Minstens 16 doden bij [bomaanslag in Pakistan]^{Negative}
23. [Belgische vrouw (26) gered na vijf uur in zee bij Koh Samui]^{Positive}
24. [Stephen King komt in maart met een nieuw boek]^{Neutral}
25. [De brexit]^{Conflict} kaapte donderdag de meeste aandacht weg op de eerste dag van de Europese top in Brussel.

2.3.3 Realis

Event mentions refer to real-world events, sometimes even regardless of whether or not the event actually happened. To distinguish between events that have occurred and events that might occur, we associate a positive/negative tag to each event. If the event did not yet take place and we only have a low degree of certainty that the event will take place the tag is negative. Naturally a positive tag means that the event occurred or will occur with a high degree of certainty. Note that the tag description *negative* can be somewhat misleading, as the 'not happening' of a certain event is the event itself e.g in example 9, the refusal of the new law is the event.

26. [Duitse president weigert wet te ondertekenen]^{Pos}
27. [Maradona komt volgende maand misschien naar Limburg en België]^{Neg}.
28. [België gaat door naar de halve finale van het WK]^{Pos}.

2.4 Event Arguments

Event arguments constitute a crucial component of an event mention. By identifying the arguments associated with an event action, we learn information regarding the who, what, how, when and where of the event. In the following sections, we discuss the identification and annotation of possible arguments. Note that similar to corpora like EBC+, the ENCORE annotation style is event-centric, meaning that we only annotate arguments if they are linked to the event action, as opposed to expressions of time and location or entities that occur in the text with no relation to the given action. Following the annotation of events proposed in (Cybulska and Vossen, 2014) and (noa, 2008), We distinguish between four main argument types: **event time**, **event location**, **human participants** involved in the event and **non-human participants** involved in the event. Before discussing the details regarding argument subtypes it is useful to examine two situations that can make argument annotation significantly harder.

Embedded Arguments

When annotating event arguments one should consider an argument in its entirety. Consider the following example:

29. Koning Filip (60) heeft woensdag de zes winnaars van de hoopdoetleven-challenge ontvangen op [het Koninklijk Paleis Van Laken]^{LOC}.

In this case, *het Koninklijk Paleis Van Laken* is tagged as the location of the event. However, this argument is actually composed of two location entities: *het Koninklijk Paleis* and *Laken*. Yet, we annotate it as a single location argument due to the fact that it refers to a specific entity, namely *Het koninklijk Paleis* which is located in *Laken*, as opposed to the Belgian royal palaces in other locations such as *Hertoginnedal* or *Belvédère*. We do not tag these instances separately, nor do we embed the separate entities within the location argument. This hold true not only for locations, but for human and non-human participants as well.

Subordinate clauses in event arguments

Subordinate clauses are often found in texts as a way to give more information regarding one or more objects in the main clause. Consider the following example:

30. [Boze Trump weigert advocaat, die hij met toeters en bellen binnenhaalde, te betalen.]

When examining the main clause of this sentence, we distinguish 2 arguments: *Boze Trump* and *advocaat*. However, due to the fact that a clause is subordinate to one of the arguments the second argument is actually *advocaat, die hij met toeters en bellen binnenhaalde* We make the decision to **not** annotate subordinate clauses separated by a comma as part the argument, as including them would result in long-winded and (from a computer's perspective) hard to interpret event components. Note that events found within these subordinate clauses should be annotated if the main clause also includes an event.

2.4.1 Event Time

The Event Time argument consists of explicit time expressions. This argument refers to when an event took place. The time conveyed can be situated in the past, present, future or can denote an ongoing event. When annotating event time within an event mention we select the entire time expression including any prepositions that may or not be present. The examples below illustrate various ways the time argument can be expressed within an event action. Note that we do not distinguish subtypes in the annotation itself.

Calendar dates are a very straightforward way of expressing time, especially within news articles. They are not necessarily limited to arguments that explicitly state a day, month or year, but can also include more vague statements that situate the event in time with respect to the present/time of writing.

31. Het vliegtuig van vlucht MH17 werd [op 17 juli 2014]^{TIME} boven Oost-Oekraïne uit de lucht geschoten door een Buk-raket, een wapen van Russische makelij.
32. [Vanaf maandag]^{TIME} zullen aan daklozen in de Brusselse straten 'attesten van niet-huisvesting' verdeeld worden door veldwerkers, mobiele teams en spoeddiensten van de ziekenhuizen Sint-Jan en Sint-Pieters.
33. Het was de derde officiële koudegolf van [deze winter]^{TIME}.

Another method of explicitely denoting time is to refer to a specific time of the day. Note that compound time expressions such as *At 9 a.m. Friday, October 1, 1999* are tagged as a single time expression..

33. [Donderdagmorgen om half tien]^{TIME} is in het Koningin Geraldine Kraamziekenhuis in Tirana prinses Geraldine geboren, zo heeft het Albaanse hof
34. Wolvestraat [donderdagmiddag]^{TIME} zonder water door dringende huishoudelijke aftakking
35. [Om half vijf 's morgens]^{TIME} was de gijzeling afgelopen.

Time expressions are also tagged when they denote ongoing events or events that go on for a specific time window.

36. [Twee weken lang]^{TIME} moeilijk parkeren rond gemeentehuis door afbraak
37. Heracles - Feyenoord is één van de duels [die avond]^{TIME}
38. Delhaize geeft vanaf [dinsdagochtend tussen 8 en 9 uur]^{TIME} expliciete voorrang aan 65-plussers.

Certain time expressions describe sets of time for repeating events. These expressions often do not contain specific time statements but rely on adverbs that signal a repetition.

39. We zien vooral single mannen en vrouwen, [vaak]^{TIME} 60-plussers, ten prooi vallen aan vriendschapsfraude.
40. Vanaf dit weekend zullen in 'VertelHotel' vrijwilligers [elke eerste zaterdag van de maand]^{TIME} voorlezen voor luisterende oren in de bibliotheek van Mortsel
41. [Om de twee weken]^{TIME} is er in België een Tine Nys: Wim Distelmans schetst breder plaatje op euthanasieproces.

2.4.2 Event location

The Event Location argument corresponds to physical locations where an event takes place. As with arguments of time, we do not limit ourselves to the head of the expression. However, prepositions are **not** included in the annotation here. Three major subtypes are distinguished based on the specifications included in (noa, 2008). When annotating event location it is important to keep in mind that we only want to annotate the location where the event has taken place. This is often signalled by a preposition in front of the location entity.

Geographical regions are a straightforward choice for the event location argument. It is important to note that this tag is only given when the entity in question refers to the territory itself. For cases in which the geographical entity metonymically refers to the people inhabiting said territory a different annotation should be used (cfr. 2.4.3). In addition to socially defined geographical regions we also include locations that are identified based on astronomical features and landmarks (e.g. landmasses, bodies of water and geological formations).

42. Op het moment dat in [Antwerpen]^{LOC} de avondklok werd ingevoerd, kwam er in die provincie elke dag 203 nieuwe bevestigde besmettingen bij
43. De jongen was om een onbekende reden in [een zijtak van de Dijle, vlakbij het Sportkot]^{LOC}, gesukkeld.

Buildings, real-estate improvements and other man made locations where an action has taken place are also tagged as event locations. Similarly to the Geographical locations, instances where facility entities are used to refer to the people that inhabit them/are present in them we use a different annotation tag (2.4.3).

44. De ene week op [school]^{LOC}, de andere thuis les: dat zou ideaal zijn.
45. Inferno in [Filipijnse fabriek]^{LOC}: zeker 30 doden
46. In de Brusselse gemeente Sint-Jans-Molenbeek is deze namiddag een vrouw neergestoken op [straat]^{LOC}.

Locations that do not fall under the geographical location or facility tag can also be annotated using the *Event Location* tag.

47. Man doet dutje in [eigen bed]^{LOC} tijdens het paragliden.
48. Er wordt ook huishoudelijk afval gedumpt in of naast [de vuilnisbak]^{LOC}.

2.4.3 Human participants

The human participant argument includes entities that are expressed as syntactic subjects or objects within the event mention. The word entity is an umbrella term for people, objects, organizations, locations... that figure in a given text. An entity can be any string of words, continuously referring to the same (real world) object. We distinguish 2 categories for entities: named entities and nominal entities. While a named entity refers to a real world participant with a proper name, nominal entities do so through noun phrases or pronouns. As with the previous argument types, we annotate the entire phrase and do not limit ourselves to the lexical head. The name 'Human participant' can be somewhat misleading, as we do not limit ourselves to direct mentions of human entities such as names of specific people. A human participant can be described as any entity crucial to the event that is either human, controlled by humans or used to metonymically refer to humans. As mentioned before, within the human participant tag we distinguish between 2 subtypes.

Named Entities

As mentioned before, named entities refer to agents in the text that are represented by a proper name. Note that this includes abreveations, acronyms and aliases too. We distinguish 3 subtypes for human participants in the named entity category.

Person *Part.Ent.PER*

The *Part.Ent.PER* tag is used when referring to specific individuals in the real world by name.

49. Ook [Janine Bischops]$^{Part.Ent.PER}$ valt het overlijden van [Johny Voners]$^{Part.Ent.PER}$ zwaar
50. [Lachaert]$^{Part.Ent.PER}$ en [De Wever]$^{Part.Ent.PER}$ willen conflict uitpraten na bijna vier maanden radiostilte.

Organizations *Part.Ent.ORG*

Groups and agencies that have a defined organisational structure are also considered human.

51. [De Staatsveiligheid]$^{Part.Ent.ORG}$ brengt een brochure uit om reizende Belgen te waarschuwen voor spionage en beïnvloeding vanuit het buitenland.
52. [De Universiteit Gent]$^{Part.Ent.ORG}$ opende het academiejaar op een valse noot: het digitaal studentenportaal Ufora lag er maandagochtend al meteen een uur uit.

Note that the *Part.Ent.ORG* tag is also used for buidlings and structures in the text that refer to humans the that manage or work in them.

53. [Volvo Cars]$^{Part.Ent.ORG}$ verzekert toekomst van de fabriek.
54. [Home Boudewijn]$^{Part.Ent.ORG}$ verbroedert met buren op Vosrock.

Location *Part.Ent.LOC*

As mentioned before in 2.4.2 Geopolitical entities denoted by geographical regions in the text can refer to the government or inhabitants of said regions. Those entities are annotated with the *Part.Ent.LOC* tag.

55. [Polen]$^{Part.Ent.LOC}$ wil actief meewerken aan euroredding.
56. [Brussel]$^{Part.Ent.LOC}$ wil passagiersterminal voor grote cruiseschepen.

(pro)Nominal entities

In some cases, events take arguments that are not named entities but still refer to humans. These arguments are often entire noun phrases that either refer to humans without mentioning a specific person or refer metonymically to humans in a way that is hard to fit in any of the categories mentioned in section 2.4.3.1. In addition to this, any pronoun that refers to a human participant in the action is also tagged as a nominal entity.

Firstly, the *Part.Nom* tag is given to noun phrases that constitute an event argument in their entirety. Note that these noun phrases might contain a named entity. However, due to the fact that we consider the entire phrase to be the argument they are not tagged as such.

57. [De winnaars van De Standaard Solidariteitsprijs]$^{PART.NOM}$ zeilen scherp aan de maatschappelijke wind.
58. [Twee op de drie Vlamingen]$^{PART.NOM}$ zijn niet akkoord met de plannen van de regering.
59. [Staatssecretaris voor Asiel en Migratie Sammy Mahdi]$^{PART.NOM}$ stelde afgelopen week zijn beleidsverklaring voor in het parlement. zijn niet akkoord met de plannen van de regering.

Note that following the rules we established regarding the embedding of arguments in the introduction of section 2.4, *De Standaard* and *Sammy Mahdi* are not separately tagged as an entity in example 56 and 58 respectively. However, as always some exceptions arise. Compare the following sentence that includes a relative clause below to example 59 and note the difference.

60. [Sammy Mahdi]$^{PART.NAM.PER}$, [staatssecretaris voor Asiel en Migratie]$^{PART.NOM}$ verweerde zich gisteren in de media.

A human participant can also be seen as any type of vehicle that is controlled by a human agent, as these inanimate objects cannot perform actions on their own. Note that when referring to the vehicle as a static object we use the *Non-Human Participant* tag instead.

61. [Belgische F-16's]$^{PART.NOM}$ hebben twee Russische bommenwerpers onderschept die onderweg waren naar het Verenigd Koninkrijk.
62. [Reddingsschepen van twee Duitse ngo's]$^{PART.NOM}$, die in totaal 49 migranten hebben gered op de Middellandse Zee, hebben van Malta de toelating gekregen om aan te meren.

We can also refer to humans in more vague statements. In these cases, we also use the *Human participant* tag.

63. Uit een studie van energieregulator CREG over het energieverbruik in 2018 blijkt dat [veel Belgische gezinnen]$^{PART.NOM}$ in energiearmoede leven.
64. In tirade van een uur valt geërgerde Trump opnieuw ['oneerlijke' pers]$^{PART.NOM}$ aan.

2.4.4 Non-Human Participants

The non-human participant tag is ascribed to any remaining entities that do not fall under human participant tag, yet still contribute to the meaning of the main event. In most cases, non-human participants will either be expressed by lifeless objects, concepts and ideas, prepositional phrases or syntactic clauses within the sentence. As in the previous section, we make a distinction between named non-Human participants and nominal non-human participants.

Before discussing these subtypes it should be noted that certain non-Human participants can also be marked as events themselves. As is the case in the following example:

65. Afgelopen vrijdag was het exact 75 jaar geleden dat [de Tweede Wereldoorlog]$^{NHPART.Nam}$ eindigde.

The non-human participant (object) that was tagged in this sentence also constitutes an event of itself. Keep in mind that in situations like this both events of the sentence should be fully annotated.

Named entities

Certain objects or concepts can be referred to by name. These cases are annotated with the *NH-part.Nam* tag.

66. [Het Lam Gods]$^{NHPART.Nam}$ is weer compleet: panelen worden onder politiebewaking gezet.
67. [Bouchez (MR): "[Vivaldi]$^{NHPART.Nam}$ niet coalitie van onze dromen, maar wel één die stabiliteit kan geven".

(pro)Nominal entities

Arguments that are neither human nor a proper name are placed in the *NHpart.nom* tag. Note that this category is not limited to objects specifically, but include certain prepositional phrases or entire syntactic clauses.

68. Mogelijk stemmen we in oktober met [potlood]$^{NHPART.nom}$ en [papier]$^{NHPART.nom}$.
69. [Koe op wandel]$^{NHPART.nom}$ houdt zich aan de snelheid

Chapter 3

Coreference Annotation

As mentioned before, coreference resolution in natural language processing is the task of linking language components that refer to the same real world phenomena. For the annotation of the ENCORE dataset, we annotate both entity coreference and event coreference.

3.1 Entity coreference

As stated before, the word entity is an umbrella term for people, objects, organizations, locations... that figure in a given text. An entity can be any string of words, continuously referring to the same (real world) object. Going back to section 2.4, entities within texts will fill important event argument slots. Naturally, determining whether entities that are part of different events refer to the same real world objects is a crucial step towards linking the events themselves. For this purpose, we annotate links between all entities that serve as arguments to an event in a given news article. For the ENCORE corpus the annotation of entity coreference is restricted to entities that occur in the same document, as linking all entities throughout the corpus would be a gargantuan and often frustrating task. However, entity coreference links that cross sentence boundaries are annotated. Other than direct mentions, pronouns are also taken into account and should be annotated as well. We distinguish three different types of entity coreference. Note that the type of coreference for entities is not annotated and that the examples below serve mostly as an illustration as to how entity coreference might be encountered in a document.

NP Coreference When two noun phrases refer to the same real world entity we can draw a coreference link between them. In order for two entities to be considered coreferent they must have the same entity type (PER/ORG/LOC). In the examples below all entities that are coreferent are indicated.

70. **Elio Di rupo** was daar niet mee opgezet. **De Waals minister-president** sprak zich tegen de beslissing uit in Le Soir.

Anaphora Coreference links can also be established between pronouns and noun phrases. In the case of anaphora an entity that in itself has no meaning (the pronoun) takes the meaning of an aforementioned noun phrase (the antecedent).

72. Volgens **Belkacem** een totaal onrechtvaardige daad en bovendien 'geen toeval': "Staatsveiligheid vroeg **me** om informant te worden.

Cataphora Although less common, pronouns can also take the meaning of a noun phrase that succeeds them. These are called cataphora and are also annotated as they still refer to the same extralinguistic entity.

73. De vrouw zelf zei dat **hij**, **Yves Leterme** haar geen aandacht meer schonk.

3.1.1 Entity coreference relations

Determining wether two entity mentions corefer is not always as straightforward as in the examples above. Certain coreference links that are drawn require additional information to paint a full picture. Therefore, it is useful to specify what type of entity coreference relation is used to connect the entities. We distinguish 3 important types of possible links.

Identity coreference Identity relations are very straightforward. Two entities are in an identity coreference relation when they refer exactly to the same real world entity.

74. De laatste e-mails van **de leraar** aan de schooldirectie voor **hij** werd onthoofd.

Part/whole Coreference Part/whole coreference links are established when one of the entities is connected to the other entity, but only in part of it.

75. **De auto** raakte van de weg of omstreeks half 9. Getuigen zeiden dat **de lichten** niet werkten.

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

76. Terwijl Van Grieken feestelijk **de resultaten van zijn partij** meedeelde was De Wever **er** op het partijhoofdkwartier minder blij om.

3.2 Event Coreference annotation

Once the event mentions in a given article are annotated, we can start drawing relations between them. One possible relation between two event mentions is coreference. Event coreference occurs when two event mentions refer to the same real-world event. One of the explicit aims of the ENCORE project is to create a corpus that shows how different sources present the same news facts. Naturally, annotating event coreference is a crucial step towards realizing this goal. We make a distinction between two types of event coreference: within document coreference and cross-document coreference.

3.2.1 Within-Document coreference

As the name implies, when annotating within-document coreference we link event mentions that are found in the same document. Establishing a coreference link between two event mentions is not a straightforward task. Deciding whether or not mentions truly refer can be quite subjective in itself and in addition to that, certain events refer only partly to the same real-world event which further complicates matters. In order to aid the annotator in this task we present three crucial criteria that should be fulfilled when drawing a coreference link between two event mentions:

- Events should occur at the same time
- Events should occur in the same place
- The same participants should be involved

The following examples are annotated with these criteria in mind:

77. Two events mentions that corefer
    (a) [Een uitgekookt Frankrijk heeft de Rode Duivels verslagen met 1-0 in de halve finale op het WK voetbal].
    (b) [De Rode Duivels hebben hun halve finale op het WK in Rusland tegen Frankrijk verloren met 1-0].
78. Two events mentions that don't corefer
    (a) [Gisteren is een 25-jarige vrouw uit Tielt om het leven gekomen bij een verkeersongeval in Baarle, een deelgemeente van Drongen].
    (b) [Een vrouw op een elektrische step is vanochtend bij een verkeersongeval om het leven gekomen op de Boomsesteenweg in Wilrijk].

In many cases event mentions will not include the all the information needed to verify the aforementioned conditions directly e.g: the time/place of the event might not be included in one of the mentions. Should this issue arise, annotators are asked to intuitively judge the events in question. If there is a reasonable suspicion that the mentions refer to the same real-world event e.g. through additional context presented in the article coreference may be annotated.

79. Two events that intuitively corefer
    (a) [SP.A zamelt winterjassen in voor kansarmen]
    (b) [Op de Werelddag tegen Armoede hield de SP.A van Dendermonde op de binnenkoer van het ABVV in de Dijkstraat een inzameling van winterjassen].

Coreference type As mentioned before, certain events refer only in part to one another. Although rare, these cases merit specific attention as their semi-ambiguous nature might pose problems for the learning algorithm later on. For this purpose, we introduce a final parameter that can help us identify coreference links between events: the coreference type. We distinguish two subtypes, namely the **identity** link and the **part-whole** link. The identity link subtype is straightforward to understand, it occurs when two event mentions refer exactly to the same real world event. However, the part-whole coreference link does warrant some additional explanation. In this subtype two event mentions satisfy the three criteria mentioned in 5.1, but one of the events constitutes only one component of the other. Reconsider this example that was given in section 2.

80. (a) [Politieke aardbeving in Israël][1]
    (b) [De Israëlische premier Ariel Sharon heeft zijn lidmaatschap van de Likoedpartij opgezegd][2] en [het ontslag van zijn regering aangeboden][3].

There are three event mentions in the example. We can draw coreference links both between mention 1 and 2 and mention 1 and 3 respectively. However, due to the fact that both *Het lidmaatschap opgezegd* and *Het ontslag van zijn regering* contribute to the event *politieke aardbeving* 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.

3.2.2 Cross-document Coreference

As stated before, the ENCORE project's goal is to be able to link news articles from different sources that discuss the same event. In order to achieve this we need not only annotate coreference within documents but also across different documents. Because cross-document annotation over the entire news corpus would be an insurmountable task, we introduce event clusters.

Event Clusters In theory, there is no real difference between performing within-document coreference annotation and cross-document coreference annotation. The same criteria introduced in section 3.1 are used to determine whether or not two event mentions refer to the same real-world event. However, due to the much larger pool of candidate mentions throughout the corpus, performing cross-document coreference would take an immense amount of time to complete. Event clusters can be a practical solution to this problem. Each cluster contains around 10 news articles on average. These articles are pre-selected from the larger corpus based on their topics and entities as follows. Firstly, one article is randomly drawn from the entire document collection which will serve as the centre of the cluster. Secondly, all named entities are retrieved from this article. Thirdly, a pass is made through the document collection and articles that contain a high number of overlapping entities with the core article are added to the cluster. This entity-based method results in a cluster in which events are likely to overlap. Finally, the clusters are manually pruned in order to remove irrelevant articles. In order to retrieve a satisfactory amount of coreference links from the coprus and keep the task of annotating the corpus feasible cross-document coreference is only annotated within these clusters. The example below shows the titles of news articles from a part of a sample cluster in the corpus:

81. Zaak over nationaliteitsafname van Fouad Belkacem morgen verder behandeld
82. Belkacem geraakt niet in rechtbank door cipierstaking
83. Fouad Belkacem is Belgische nationaliteit kwijt
84. Belkacem weet volgende maand of hij Belg mag blijven
85. Theo Francken (N-VA) wil Fouad Belkacem zo snel mogelijk land uit

Chapter 4

Annotation Example for Cross-Document Event Coreference

We use the WebAnno tool (Yimam et al., 2013) for annotating events and entity coreference in the document collection. The following section is an illustrated guide on how to setup WebAnno, load event clusters and annotate the documents within those cluster. Once the all events are annotated, we use the Knowledge base (KB) supported search functionality of the INCEpTION annotation tool to annotate cross document event coreference links. .

4.1 Annotating Events in WebAnno

When a document is opened in WebAnno you will see the sentences displayed in the "Annotation" tab. Each row displays one sentence of the document. The default setting will let you view 5 sentences at a time. However, when annotating corference it is useful to have an overview of the entire text in front of you. You can adjust the number of rows that are displayed by clicking the **Settings** button.

Figure 4.1: The WebAnno interface

The first thing to do when starting the annotation of a new document is to indicate all events in the article. To do this, make sure the **Layer** parameter on the right side of the screen is set to "Event" and simply highlight all events in their entirety. In the case where you make a mistake you can double click the event and press the delete button in order to remove the annotation layer.

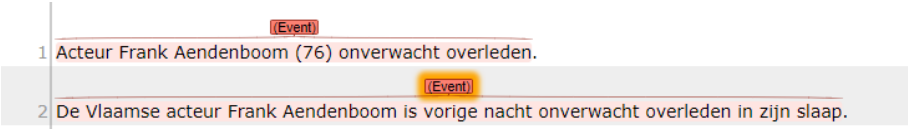

Figure 4.2: Annotation of events

Once all events in the document are highlighted go through each event individually and fill in their event properties and action. You can select the event properties for a given event on the right hand section of the screen. Simply designate a subtype for each of the listed properties through the dropdown menu.

In order to annotate the event action you must click on the **add** button in the "Event Action" subsection on the right hand side of the screen. You can now select the action by highlighting it with your cursor.

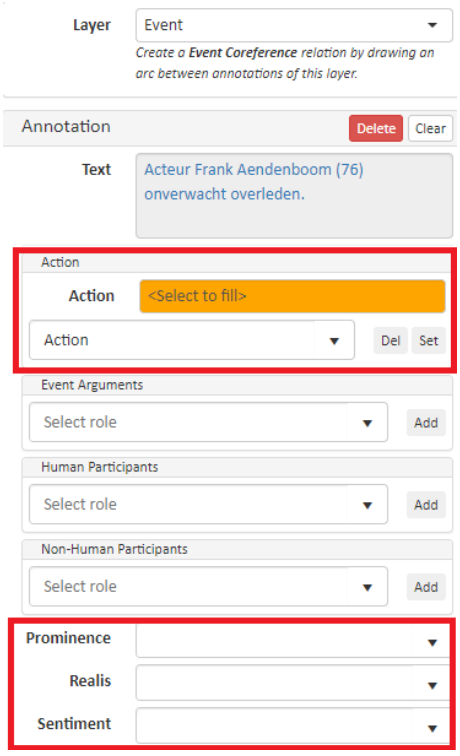

Figure 4.3: Annotation of the event action and properties

In a third pass through the document select all event arguments. The time and location argument types are found under the ”Event argument” subsection, while the Human Participant and Non-human Participant argument types have their own subsection.

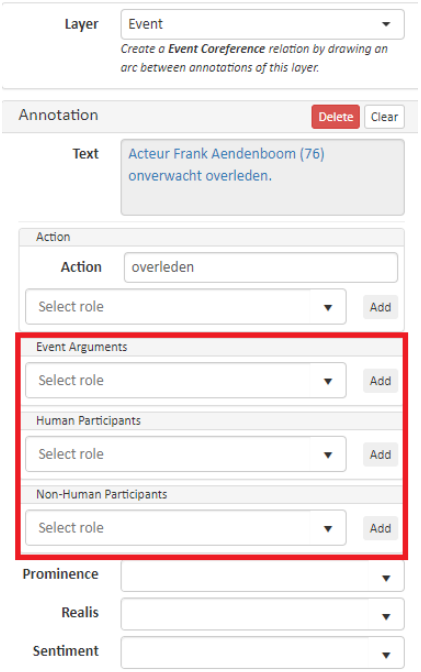

Figure 4.4: Annotation of the event arguments

In order to add an argument to the event you must first select the role of the argument in question. Then click the add button and highlight the relevant portion of text similarly to how you annotated the event action.

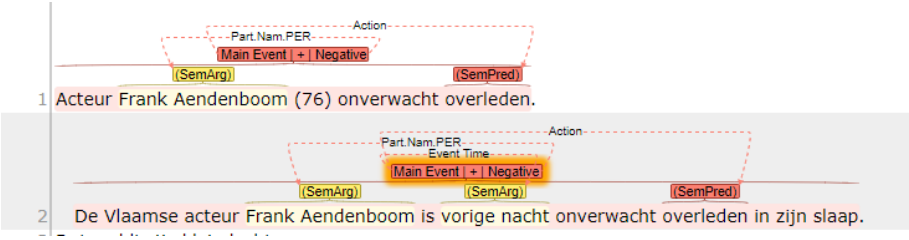

Figure 4.5: An example of a fully annotated event with its arguments

The next step of the annotation is to draw coreference links between all entities. You do not have to change the layer parameter for this task. Simply select an argument in the annotation and draw a link to another argument in the text that it corefers with. A coreference link is created automatically. Then, select the type of coreference that links the two entities.

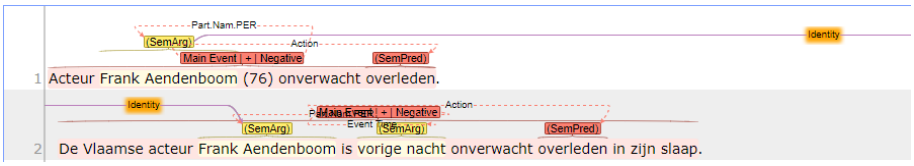

Figure 4.6: Annotation of entity coreference

4.2 Establishing cross-document coreference links in INCEpTION

Once events and entity coreference links are annotated for the entire dataset we import the document collection in the INCEpTION annotation tool (Klie et al., 2018). The knowledge base functionality (Boullosa et al., 2018) allows us to link the annotated events in the documents to concepts within a KB. We create a KB for each event cluster (cfr. 3.2.2). We then traverse all documents within that cluster and add a new concept to the KB's taxonomy when a new real-world event is encountered. When an event mention is encountered that refers to a previously established concept within the KB, we simply link that event mention to that concept, creating a web of cross-document event coreference links. The image below shows the KB taxonomy for the example, as there is only one real-life event mentioned the number of KB concepts is limited to one.

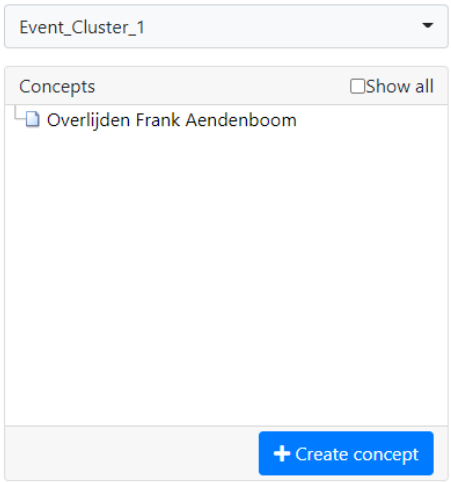

Figure 4.7: Knowledge Base taxonomy for the annotated example

4.3 Summary of the annotation process

1. Make a first pass through the article and annotate all events.
2. Now highlight the event actions and determine the event properties for each of the events.
3. Make a third pass through the article and annotate all arguments for each of the events
4. In a fourth pass establish coreference links between the entities that were annotated as arguments.
5. Once all articles in a cluster are annotated, make a final pass through all the articles and annotate cross-document links between the events through the knowledge base functionality.

References

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

Boullosa, Beto, Richard Eckart de Castilho, Naveen Kumar Laskari, Jan-Christoph Klie, and Iryna Gurevych. 2018. Integrating knowledge-supported search into the inception annotation platform. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 127–132.

Colruyt, Camiel, Orphée De Clercq, and Véronique Hoste. 2019. EventDNA: Annotation Guidelines for Entities and Events in Dutch News Texts (v1.0). Technical report.

Cybulska, Agata and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *LREC*, pages 4545–4552.

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9.

Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6.