Event coreference in the news

Cybulska, Agata Katarzyna

2021

VRIJE UNIVERSITEIT

Event coreference in the news

Who, what, where and when?

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor
aan de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Geesteswetenschappen
op donderdag 15 april 2021 om 11.45 uur
in de online bijeenkomst van de universiteit,
De Boelelaan 1105

door

Agata Katarzyna Cybulska

geboren te Warschau, Polen

promotor: prof.dr. P.T.J.M. Vossen

Abstract

This dissertation is about cross-document coreference between events in the news. The first two parts focus on data used to study event coreference and the last two parts contribute to modelling the event coreference phenomenon. Firstly, I investigate the available data sets to determine their representativeness with regard to the referential and lexical diversity of event coreference in the news. Thereafter, I explore how one can make a data set more representative of event coreference in the news by creating the ECB+ corpus. Next, I research the best ways to model the phenomenon of gradable coreference and to consider partial coreference in event coreference resolution. Finally, I deliberate about the role of event times and entities in event coreference resolution. The last part of the work results in developing the Bag of Events approach to event coreference resolution which makes use of partial coreference between mentions of event components from a unit of discourse. This dissertation provides a rigorous account of how the diversity of event coreference in the news can be sampled and modelled to perform event coreference resolution. The outcome of this research lays the foundation for highly accurate coreference resolvers.

Acknowledgements

This research has been carried out at the VU University Amsterdam as part of the Semantics of History research project funded by the interfaculty research institute CAMeRA (Center for Advanced Media Research) and within the NewsReader project supported by the European Commission within the 7th Framework Program under grant agreement FP7-ICT-316404.

Writing a dissertation is at once a solitary and collective endeavor. Solitary because the researcher is solely responsible for the work. Collective because science is never conducted in isolation. This is so not only because of the collaborative nature of university work, but also because output is always delivered, as the saying goes, on the shoulders of giants.

Beginning with the giant on whose shoulders I stood, I wish to thank first and foremost my advisor, promotor, and mentor all rolled into one: Prof. Piek Vossen. It was a pleasure working with Piek from the early days as an intern at Irion and eventually a PhD student. Piek's tireless support, steadfast dedication through the (very) long path leading to my defense, and friendship helped transform a rough PhD proposal into the dissertation in your hands.

I am also appreciative of the people with whom I worked at the university. The VU Amsterdam provided a vibrant community of research scientists, scholars, and academics whose company and camaraderie was remarkable. Specifically, my work benefited from coffees, chats, and friendly interactions with my colleagues from the CLTL: Isa Maks, Hennie van der Vliet, Rubén Izquierdo Beviá, Marten Postma, Paul Huygen, and Attila Görög, as well as my office-mates Bertie Kaal and Ruth Wester. I am also grateful to the annotators of the ECB+ corpus, Elisa Wubs and Melissa Dabbs.

I wish to thank Eduard Hovy for hosting me during my exchange semester at the Information Sciences Institute (ISI) of the University of Southern California. The colleagues there all made me feel welcome and made my visit into the impactful event that it turned out to be.

I would like also to express my gratitude to the Reading Committee, the Defense Committee and the anonymous reviewers whose keen insights, advice and input helped sharpen my arguments.

Thank you one and all!

Finally, the foundation for the delivery of this work, owes to my friends and family who directly or indirectly helped me deliver this dissertation. My parents, Elżbieta and Antoni Cybulscy and my sister Monika Cybulska helped me to become the person I am today. They gave me the strength to leave familiar Warsaw and chart new territory in The Netherlands. Friends like Francesca Righetti, Marielle De Silva De Freitas and Silke Hömstreit were great company and distractions from the challenges of PhD research. I wish also to thank my husband, Matt Coler. True though it may be that it is mainly his fault that this dissertation took me over a decade to complete; after all, our wedding (which I planned) and two children (who I birthed) slowed progress considerably – nonetheless, he is my rock. The love of my family sustained me throughout this process.

Dla Zosi i Julci.

Contents

“Neither the category of substance nor the category of change is conceivable apart from the other.” Davidson [1969]

Chapter 1

Introduction

1.1 Background

This dissertation is about events in the news. While every news story is comprised of events, the same event can be described in many different ways, from different individual perspectives to different ideological perspectives. Consider the downing of flight MH17 over Ukraine. While a Russian and Dutch writer may agree that an airplane crashed, their accounts of that event may be immensely different depending on the political persuasion of the writer.

The claim that the same event may be described in different words may seem extremely simple, even obvious. However, it has consequences which go beyond effective search and even beyond research into computational linguistics. This topic encompasses the very fabric of democracy itself, particularly in an era of fake news.

The number of news stories on the internet is incalculably large, and is growing still larger every second. The sheer quantity of text means that accurate search is crucial for scientists, professionals and the general public to make sense of current events. To be able to search news systematically, first we need to identify events described in text and then we need to find all texts that describe the same event. This is a very complex task that involves different research topics: event theory, event extraction, resolution of event coreference and identification of other event relations than event identity, for example when events are part of a larger structure like in Schank's scripts (Schank [1990]). At the crux of this issue is **event coreference**. If there are more than one event mentions describing the same event, we say that the mentions are coreferent. Event coreference resolution is the task of determining whether two event mentions refer to the same event. Resolution of coreference between event descriptions in the news is the first and crucial step for many NLP applications such as text search, information extraction, question answering or text summarization. To solve event coreference, a coreference resolver must consider that the same event could be described with diverse formulations and that different events can be described with the same formulation. Some of the reasons why event coreference resolution on news data poses a big challenge for natural language processing applications are:

- there is an unknown number of world events
- only a subset of those makes the news
- some events develop over time

- there are event descriptions in different languages
- the same event can be described with different formulations in one language
- event descriptions can reflect different ideological perspectives
- entity references by function or title can be temporary
- event descriptions often spread over multiple sentences.

Let us examine these confounding factors in more detail. There are millions of news articles on the web that describe an **unknown number of events**. For example according to the USGS (earthquake.usgs.gov) during 7 days between 25 February 2020 and 03 March 2020 there were 389 earthquakes registered worldwide of at least 2.5 magnitude. During 30 days from 02 February 2020 to 03 March 2020 there were 1988 earthquakes worldwide. At the same time when you search the word "earthquake" in the news section of the Google search engine, restricting the time frame to "past month" (search on 03 March 2020) there were 18 pages of search results returned. Not all earthquakes make the news. For the ones that do, some are reported immediately, others are picked up by the news agencies only after some time and some are not reported at all. For people who investigate a certain topic based on a news archive, it is a challenge to get a clear picture of reality with so much news data. At the same time, news articles describe only a **subset of events**. Event coreference resolution is crucial to get an idea about what events are described in the news.

Consider also how some news **events develop over time**. A case in point is the COVID-19 pandemic, which at the time of writing, receives varying amounts of press as the virus spreads, recedes, and returns across different areas, as told from different historical, political, and regional perspectives in various languages in different ways. This is an example of a developing story with new events happening and making the news slowly over time. The news data in this case is extremely complex, with a relatively expansive time frame and in **different languages**.

Even if we limit the discussion to a simple case of a short bounded event which is described in one language, the **ways of description can strongly vary**. To resolve event coreference an NLP application must know *what* happened as well as *when* and *where* it happened and *who* was involved with the event, even if the ways of description are different. Compare *a car bombing in Madrid in 2001* and *an explosion in Spain in 2001*. These two descriptions could, but need not, refer to the same event.

Consider also the following example: When an army enters the territory of another country, one could describe that event as *a military presence, intervention, liberation* or *invasion*. In natural language we can select the appropriate formulation that can be relatively neutral or that could reflect some **subjective marking**. Compare how military activity in a foreign country can be described as *an invasion, an occupation* or simply as *a presence*.

An additional challenge is the fact that entity references can describe a person by their **function or title which can be temporary**. Consider how the referent of US President is different today than it was in 2015. Thus, understanding what event is meant in descriptions like *Sri Lanka's President visited the US President* depends entirely on the time in question.

Event coreference resolution is made even more difficult by the fact that descriptions of events are often **spread over multiple sentences**. Following the Gricean Maxim of quantity (1975), writers do not repeat pieces of information that were already communicated within discourse borders.

These are some of the reasons why event coreference resolution on news data poses a big challenge for natural language processing applications.

In this dissertation, I explore the topic of cross-document resolution of coreference between events in news articles. The main research questions investigated in this work are the following: **How can we model the phenomenon of event coreference across news and how can we design a corpus used to study the phenomenon in the context of the quickly changing world and growing data?** To get insights into the main research questions, we will look at four research sub-questions in the four parts of the manuscript. Parts I and II of the dissertation focus on data used to study event coreference and parts III and IV contribute to modelling the event coreference phenomenon. We will begin in part I by investigating how representative of the event coreference problem are the available data sets. Next, in part II we will explore how we can make a data set more representative of event coreference in the news. In part III we will research the best ways to model the phenomenon of gradable event coreference. Finally, in part IV we will deliberate about the role of event times and entities for event coreference resolution. Taken together, these four parts will provide a detailed understanding of how the diversity of event coreference in the news can be sampled and modelled for the purpose of event coreference resolution, which in turn will pave the way for creation of accurate coreference resolvers.

1.2 Method and roadmap

In order to understand event coreference in the news, first we analyse corpora annotated with event coreference in chapter 3. Most of our interest are data sets annotated with cross-document event coreference because they most closely resemble event coreference in news articles. As the corpora turn out not to be optimal for our research question, we extend and re-annotate one of the existing data sets in chapter 4.

To get a better understanding of event coreference, in this dissertation we research the phenomenon in two kinds of data. (1) In chapter 3 we look at multiple events described in a number of texts from one text type. We consider here the ECB corpus (Bejan and Harabagiu [2010]) with descriptions of 43 events in on average 11 news texts per event. (2) We zoom in on different descriptions of one selected event in many diverse text types to ensure as high as possible diversity of formulations used to describe the selected event. We research descriptions of the Srebrenica massacre in 78 texts in chapter 5.

Following a deeper analysis of event descriptions in different kinds of data, this dissertation makes a contribution toward determining event granularity in chapter 5 and modeling event coreference in chapter 6 based on semantic relations between mentions of events. Finally, after running some preliminary experiments with events and entities in chapter 7, in chapter 8, we propose a new approach to event coreference resolution that makes use of entities and discourse structure to solve coreference between events.

1.3 Contribution

The main scientific contributions of this dissertation are:

- a review of corpora annotated with event coreference - chapter 3

- creation of an extended corpus targeting cross-document event coreference resolution - chapter 4
- a taxonomy for determining event granularity - chapter 5
- a model of event coreference based on semantic relations - chapter 6
- a new approach to event coreference resolution that employs entities and discourse structure to solve coreference - chapter 8.

1.3.1 Data and publications

This dissertation resulted in the release of the ECB+ data set (Cybulska and Vossen [2014b]) for the creation of which the ECB+ annotation guideline was used (Cybulska and Vossen [2014a]). The ECB+ data set and the ECB+ annotation guideline can be downloaded from `http://www.newsreader-project.eu/results/data/the-ecb-corpus/` or `https://github.com/cltl/ecbPlus`. The paper accompanying the release of the ECB+ corpus:

> Cybulska, Agata, and Piek Vossen. "Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution." (2014b)

as of December 2020 has been cited 100 times according to Google Scholar, and the ECB+ annotation guideline:

> Cybulska, Agata, and Piek Vossen. "Guidelines for ECB+ annotation of events and their coreference." (2014a)

has been cited 30 times.

Besides the 2014 publications on the ECB+ corpus, the following papers have been published as a result of this dissertation:

> Cybulska, Agata, and Piek Vossen. "Event Models for Historical Perspectives: Determining Relations between High and Low Level Events in Text, Based on the Classification of Time, Location and Participants." (2010)
>
> Cybulska, Agata, and Piek Vossen. "Historical event extraction from text." (2011)
>
> Cybulska, Agata, and Piek Vossen. "Using semantic relations to solve event coreference in text." (2012)
>
> Cybulska, Agata, and Piek Vossen. "Semantic relations between events and their time, locations and participants for event coreference resolution." (2013)
>
> Cybulska, Agata, and Piek Vossen. "Translating Granularity of Event Slots into Features for Event Coreference Resolution." (2015b)
>
> Cybulska, Agata, and Piek Vossen. " "Bag of Events" Approach to Event Coreference Resolution. Supervised Classification of Event Templates." (2015a).

Furthermore Cybulska and Vossen [2015b] presented a newly created granularity taxonomy, see also Appendix 9.

Chapter 2

Relevant terminology and concepts

This chapter provides an overview of relevant terminology and concepts. To better understand the challenges relating to event coreference in texts, a few necessary definitions are in order. To those ends, we start with the definition of a corpus, which lays at the foundation of linguistic studies. Then we focus on events, the notion of event granularity and annotation of events in text. Thereafter we look at event coreference, annotation of coreference in text, and metrics used to evaluate coreference resolution.

2.1 Corpus

According to Sinclair (2004), "a **corpus** is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research". Following Sinclair, the validity of corpus-based studies depends on the notion of **representativeness** of a corpus. If a corpus is not representative of the sampled language population, one cannot be sure that the results of experiments obtained on it can be generalized onto the intended language population. The issue of representativeness is crucial for parts of this research.

In part I chapter 3 of the dissertation we look at corpora annotated with events and their coreference and we investigate whether they are representative of event coreference in the news. We set requirements on which a corpus used for experiments on unrestricted cross-document event coreference resolution must fulfil so that the results of the experiments can be generalised onto the language population of news articles. We test whether the existing event coreference corpora fulfil on these requirements.

2.2 Event

There are a number of functional definitions of **events** used in the computational linguistics community. In the annotation guidelines of the Automatic Content Extraction program (ACE), an event is defined as a specific occurrence of something that happens, often a change of state, involving participants (LDC [2005b]). In the TimeML specification, events are described as "situations that happen or occur" that can be punctual

1. action		*transported, crash*
2. time		*on Wednesday*
3. location		*to a nearby hospital, in Miramar*
4. participant	human	*driver*
	non-human	*car*

Table 2.1: Event Components

or durational, as well as stative predicates describing "states or circumstances in which something obtains or holds true" (Pustejovsky et al. [2003]).

According to the Quinean view of events (1985), events occupy a space-time. Without time and place information event actions are just denotations of abstract classes of concepts. With the exception of proper name events (such as *World War II*), events need to be anchored in time and space to become instantiated.[1]

Based on the above definitions, and as a consequence of the Quinean view, our definition of events makes reference to the following four components:

1. an event **action** component describing what happens or holds true

2. an event **time** component anchoring an action in time describing when something happens or holds true

3. an event **location** component specifying where something happens or holds true

4. a **participant** component that gives the answer to the question: who or what is involved with, undergoes change as a result of or facilitates an event or a state; we divide event participants into human participants and non-human participants.[2]

In EXAMPLE 2.2.1, *driver* is a human participant involved with the event, *car* is a non-human participant, *on Wednesday* tells us when the event happened, *to a nearby hospital* and *in Miramar* describe the locations where the two events happened while *transported* and *crash* constitute actions (see TABLE 2.1). The four components of our event model are in line with the four core classes of the formal SEM model: sem:Event (what happens), sem:Actor (who or what participated), sem:Place (where), sem:Time (when) (Hage et al. [2011]).

Example 2.2.1 *On Wednesday a driver has been transported to a nearby hospital after a car crash in Miramar.*

The four component event model introduced in this section lays at the foundation of this work and is employed throughout this dissertation. In part II chapter 4, the ECB+ corpus is developed. The corpus is annotated in line with the four component event model. In part III we model events and their coreference following the four component model. In chapter 5 of part III we focus on granularity of events in the context of event coreference, whereas event granularity is determined along the four event components. In chapter 6 a model of gradable event coreference based on semantic relations between mentions is developed; whereas semantic relations are established for an event

[1] As events occupy a space-time, the time and place of events will be crucial for solving event coreference. Coreference will only make sense for events within the same time and place. Time and place in which an event happened will form the starting point for solving event coreference (compare: *genocide in Srebrenica* with *genocide in Rwanda*).

[2] Non-human participants contribute to the meaning of the event action and will often be expressed as direct objects of a sentence or PP phrases not in object position e.g. instrument phrases.

component. Finally, in part IV we evaluate the role of the four components for event coreference resolution.

2.3 Event granularity

When analyzing event coreference, event **granularity** plays a role. In this section we introduce the notion of granularity.

The notion of granularity was described by Keet [2008] as the ability to represent and operate on different levels of detail in data, information, and knowledge. "Granularity deals with organizing data, information, and knowledge in greater or lesser detail that resides in a granular level or level of granularity and which is granulated according to certain criteria, which thereby give a [granular] perspective (...) on the subject domain". Mulkar-Mehta et al. [2011a] define granularity as the level of detail of description of an event or object. "Granularity is the concept of breaking down an event into smaller parts or granules such that each individual granule plays a part in the higher level event."

People view the world at different granularities. Humans are able to switch among different granularities of world conceptualizations (Hobbs [1985]). In a reasoning process, a granularity level is distinguished depending on what is relevant for a particular situation.

Fine granularities, that is, a lower granularity level, provides a more detailed data representation than a more abstract higher level, that is, **coarse granularities** which leave out some details.

In our work in part III we look at granularity of events. We consider coarse vs. fine granularities across the four event components. We formalize a model of event coreference that considers granularity based on semantic relations between mentions of the four event components.

2.4 Event mention annotation

In this work we distinguish between an event **mention** and an event **instance**. The former are descriptions of events in text, whereas the latter are "references" of mentions, that is, their "meaning" or "reference", as understood by Frege (1892). For example, mentions, *World War I, WWI, First World War* and *the Great War* all refer to the same instance; that is, the military conflict which occurred between 1914 and 1918.

In addition to mentions and instances, other relevant lexico-semantic units in a corpus annotated with coreference include: tokens, lemmas and concept types. Mentions are comprised of tokens, which can be grouped into lemmas. A **token** is an individual word occurrence in language use (Peirce [1906]). A **lemma** is "a base word and its inflections" (Nation and Waring [1997]). In the sentence from EXAMPLE 2.4.1, there are nine word tokens and seven lemmas. Note that "lemmas" are distinct from "types". We define "types" after Peirce (Peirce [1931–58]) as unique tokens.

Example 2.4.1 *My cat is a cat that hates other cats.*

Concept types, such as event or entity types, group instances of the same semantic type. For example *World War II, World War I, the Vietnam War* and *the Iraq War* all belong to the concept type (here event type) "war". These mentions refer to different instances but have the same “denotation” as understood by de Saussure or the same “sense” as explained by Frege (1892).

Some prominent event annotation efforts and terminology defined for the purpose of these annotation tasks are described in sections 3.1 and 3.2. In section 4.2 an event coreference annotation guideline was developed (Cybulska and Vossen [2014a]) that builds on LDC [2005b] and Pustejovsky et al. [2003]. The ECB+ guideline determines how to annotate mentions of event actions, event times, event locations and event participants in text as well as how to mark the coreference relation between mentions of an event component.

When working with annotated data, mentions which are annotated in a data set are seen as **true mentions** or the gold standard annotations. When our system extracts mentions, the mentions generated by our pipeline are called **system mentions**.

2.5 The interplay of the lexico-semantic units in a coreference corpus

In this section we will look at how the different lexio-semantic units found in a corpus annotated with coreference interact with each other, see “tokens”, “mentions”, “lemmas”, “instances” and “concept types” in FIGURE 2.1. The pyramid shape indicates

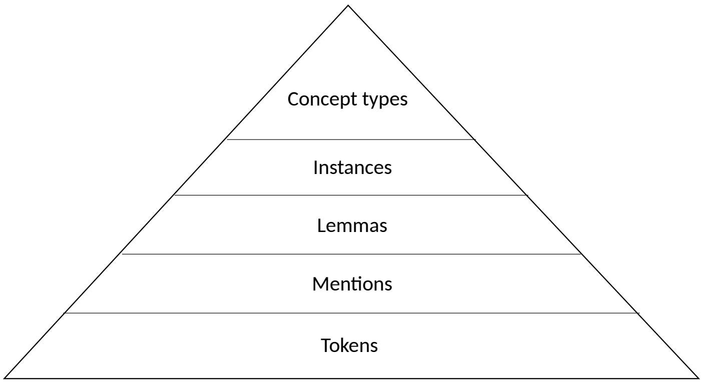


Figure 2.1: Lexico-semantic units in a coreference resource.

that the number of lemmas or lemma combinations (in case of multi-word unit mentions) will never exceed the number of mentions. There must be at least one mention for a lemma to be represented in a corpus so there is a possibility of a coreference corpus with only one mention per lemma or one lemma could cover multiple mentions. The same relationship applies to concept types and instances. The number of concept types cannot be higher than the number of instances. There can be the same amount of annotated event types and instances or there can be multiple instances annotated per event type.

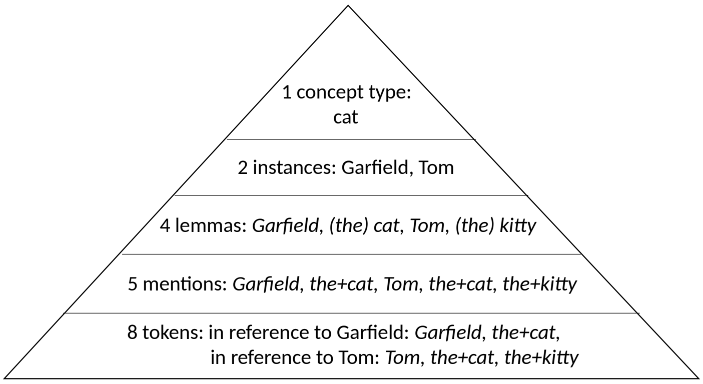

1 concept type:
cat
2 instances: Garfield, Tom
4 lemmas: *Garfield*, *(the) cat*, *Tom*, *(the) kitty*
5 mentions: *Garfield*, *the+cat*, *Tom*, *the+cat*, *the+kitty*
8 tokens: in reference to Garfield: *Garfield*, *the+cat*,
in reference to Tom: *Tom*, *the+cat*, *the+kitty*

Figure 2.2: Synonymy example.

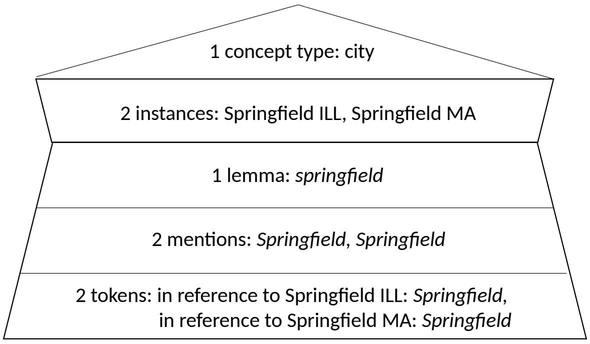

1 concept type: city
2 instances: Springfield ILL, Springfield MA
1 lemma: *springfield*
2 mentions: *Springfield*, *Springfield*
2 tokens: in reference to Springfield ILL: *Springfield*,
in reference to Springfield MA: *Springfield*

Figure 2.3: Referential polysemy example.

The pyramid accounts for synonyms that refer to one or multiple instances from the same concept type. Consider examples from FIGURE 2.2 where two instances of the concept *cat* – *Tom* and *Garfield* – are described with five mentions that cover four lemmas, two of which are synonymous: *cat* and *kitty*.

However the relation between the different "levels" in the pyramid can be distorted by polysemy, whereas one lemma refers to two different instances whether from one or from multiple concept types e.g. *Hollywood* can refer to multiple places in the US, Ireland and the UK, the American film industry in general, the movie stars themselves, or it can be a proper name, the name of a particular tv series, a programming language, or a type of tree which grows in Australia. Note that one can distinguish between referential polysemy vs. lexical polysemy. In case of the lexical polysemy e.g. a word *mouse* can refer to an animal or to an electronic device (at least two instances and two concepts). In case of the referential polysemy e.g. *Springfield* can refer to two different cities, that is, two instances that both represent one concept, see FIGURE 2.3.

Terms and concepts related to annotation of event mentions introduced in this and in the previous section are especially relevant when looking at corpora annotated with event coreference in part I chapter 3 as well as in part II chapter 4 where the ECB+ corpus is developed. Furthermore, event annotation terminology is employed in the context of experiments performed in all four parts of the dissertation: experiments on lexical and referential diversity in parts I and II, corpus analysis experiments in part III and experiments with the role of the different event components in part IV.

2.6 Event coreference

Coreference happens when more than one mention has the same reference. **Event coreference** occurs if there are more than one event mentions with the same reference, that is, if at least two event mentions refer to the same event instance. If coreferent mentions occur in the same text, we call this **within-document** coreference. If coreferent mentions occur in different texts, it is **cross-document** coreference.

The two sentences below may refer to the same event, although as the same president signs different bills and different presidents sign different bills, the example sentences may have different references so they may refer to two different instances of events. If one can determine based on the context that two event mentions have the same reference, they can be considered as coreferent.

Example 2.6.1 *The president signed the bill on Tuesday.*

Example 2.6.2 *President Barack Obama on Tuesday signed into law a landmark health care reform bill.*

In the annotation guideline developed in section 4.2, we determine that two event descriptions corefer if they refer to the same instance of an event, i.e. when the same actions happen or hold true: (1) in the same time (2) in the same place (3) with the same participants involved.

Hovy et al. [2013] have identified three levels of event identity:

- fully identical: identical in all respects as far as one can tell from the context (agents, location and time are identical or compatible)
- partial-/quasi-identity: mostly the same but there is some additional information on the side of one mention or the other that is not shared; membership or subevent relationship between events

- not identical: full independence of events, mentions do not corefer.

To solve coreference we compare different mentions. **Active mentions** are two mentions which are being compared and analysed to determine whether they are coreferent.

A **coreference chain** is a set of mentions from one instance with cardinality bigger than 1. So if a mention is linked through coreference relation to at least one other mention, we can identify these two mentions as a coreference chain.

If a mention is not linked as coreferent with any other mention in the corpus, we can identify it as a **singleton mention**. In this case the instance is constituted by only one single mention; it is a set of mentions with cardinality equal to 1.

When preparing a corpus that is meant to be a resource for studies of event coreference, it is common to first select a number of event instances, descriptions of which we want to capture in the corpus so that event coreference chains are captured in the data set. We will call these events **seminal events**. A seminal event in a document determines the topic of the document. It is connected to most events from its surrounding context (Allan [2002]).

The relevance of event coreference within the context of this dissertation is detailed in the final paragraph of the following section.

2.7 Coreference evaluation metrics

Corpora annotated with coreference can be used to develop and test coreference resolvers. Evaluation of coreference resolution is not straightforward. Until recently, there was no consensus in the field with regard to evaluation measures used to test approaches to coreference resolution. Coreference resolution results are analysed in terms of recall (R), precision (P) and F-score (F) by employing a few coreference resolution evaluation metrics. The most commonly used measurements follow: MUC (Vilain et al. [1995]), B3 (Bagga and Baldwin [1998]), CEAF (Luo [2005]), BLANC (Recasens and Hovy [2011]) and CoNLL F1 (Pradhan et al. [2011]).

The **MUC** measure (Vilain et al. [1995]) determines the minimum number of pairwise links for the gold standard and system output by subtracting the number of instances from the total number of mentions. The number of shared coreference pairwise links between the gold standard and system output is counted and divided by the minimal number of pairwise links needed to represent the gold standard (recall) and the system output (precision). This approach promotes system outputs that produce overmerged chains.

To overcome the shortcomings of MUC the **B3** measurement (Bagga and Baldwin [1998]) was introduced which computes the recall and precision scores for each mention. By doing so, in contrast to MUC, it considers the singleton mentions more than the MUC does. However, in case of data that contains many singleton mentions, the B3 becomes too inflated and approaches 100% accuracy.

The **CEAF** (Luo [2005]) measure looks for the best one-to-one mapping between instances from the gold standard and system output. Luo developed a mention-based and an entity-based CEAF depending on the similarity function used. The mention-based CEAF was used more widely. CEAF calculates the number of shared mentions between every two aligned instances and divides the number by the total number of mentions. The drawback of the measurement is that just as the B3 score it promotes outputs that contain a higher number of singletons.

The **BLANC** measure (BiLateral Assessment of Noun-phrase Coreference, Recasens and Hovy [2011]), considers both coreference and singleton links by averaging their F-scores. By doing so it does not ignore singletons (like the MUC does) but it also does not allow singleton mentions have too strong influence on the evaluation scores (like the B3 and CEAF measures).

Finally, the evaluation measure used during the CoNLL-2011 Shared Task on coreference (Pradhan et al. [2011]) is the **MELA** (Denis and Baldridge [2009]) which is a weighted average of three metrics: MUC, B-CUBED and CEAF that all focus on different aspects of coreference evaluation. For the CoNLL-2011 task the unweighted mean of the three metrics was used to compare coreference resolution systems and determine the winning one. The entity-based CEAF was used instead of the mention-based CEAF.

To conclude, when discussing event coreference evaluation measures it must be noted that some of the commonly used evaluation metrics focus on particular aspects of coreference. They promote and penalize some types of coreference resolution output. This makes them dependent on the evaluation data set, with scores going up or down with the number of singleton items in the data (Recasens and Hovy [2011]). For example, the MUC measure promotes longer chains. B3, on the other hand, seems to give additional points to responses with more singletons. CEAF and BLANC as well as the CoNLL measures (the latter being an average of MUC, B3 and entity CEAF) give more realistic results.

Terms and concepts related to event coreference introduced in this and in the previous section are crucial when analyzing corpora annotated with event coreference in part I chapter 3 and when creating the ECB+ corpus annotated with event coreference in part II chapter 4. Furthermore event coreference forms the foundation of the work in part III where we model event granularity in the context of event coreference resolution (chapter 5) and where a model of gradable event coreference is developed (chapter 6) as well as in part IV where we experiment with event coreference resolution to evaluate the role of the four event components.

Part I

Are the existing corpora representative of the event coreference problem?

In part I we will look at corpora annotated with event coreference that are used to study event coreference resolution. As introduced in section 2.1, a corpus is a sample of a language population (Sinclair [2004]). It is used for the purpose of linguistic research. For the results of a study based on a corpus to be reliable, the sample corpus must as far as possible represent the sampled language population with special attention to representing the diversity of the phenomenon of interest. In chapter 3 we will evaluate the existing corpora with regards to their representativeness of the event coreference in the news by looking at two kinds of their diversity, the referential and the lexical diversity. Part I addresses the following research questions.

Research questions

- Do the data sets annotated with event coreference reflect the diversity of news articles?
- What are the requirements for a data set for experiments on unrestricted cross-document event coreference?
- Is there an English language data set that fulfils the requirements?

Chapter 3

Analysis of event coreference resources[1]

To gain insight into event coreference, we will analyze descriptions of events and coreference between them in the news. The first step is to research corpora annotated with event coreference. We are most interested here in data sets annotated with cross-document event coreference because they resemble event coreference in news articles the most.

In this chapter we will take a closer look at English language corpora annotated with event coreference. We will survey available resources and examine them from the perspective of the task of cross-document event coreference resolution. This part of the research, performed in 2014, led to the creation of a new resource called ECB+ (Cybulska and Vossen [2014b]), described in chapter 4.

In section 3.1 we will describe corpora annotated with within-document event coreference. In section 3.2 we will introduce the only data set (at the time when this research was performed) annotated with cross-document event coreference: the ECB corpus (Bejan and Harabagiu [2010]). Next, in section 3.3 we will look at requirements that a corpus used for coreference experiments should fulfil so that the corpus is representative of the coreference problem. In section 3.4 we will evaluate the ECB corpus as a resource used to develop and test approaches to event coreference resolution by quantifying the average referential and the average lexical diversity of the corpus. We will conclude in section 3.5.

There are four preeminent English language corpora annotated with event coreference, which are available for studies of event coreference resolution. Three of them were only annotated with within-document event coreference. These are described in section 3.1:

1. ACE 2005 data set (LDC [2005b])

2. OntoNotes corpus (Pradhan et al. [2007])

3. Intelligent Community - IC corpus (Hovy et al. [2013]).

At the time of writing, there was only one corpus available that was annotated with cross-document event coreference: the EventCorefBank (ECB). The corpus was cre-

[1] The contents of this chapter have been published in Cybulska and Vossen [2014b].

ated by Bejan and Harabagiu in 2010. Two years later it was reannotated by Lee et al. [2012] (ECB 0.1). This is described in section 3.2.

3.1 Within-document event coreference corpora

In 2014 there were three preeminent English corpora available that were annotated with within-document coreference relations: the ACE 2005 corpus, the OntoNotes corpus and the IC corpus. In the following sections we will briefly describe each of the three data sets. We will discuss four aspects with regards to the event annotation effort per corpus:

1. mention extent
2. mention part of speech (POS)
3. topic coverage
4. event relations.

3.1.1 ACE 2005

The ACE 2005 data set (LDC [2005b]) was used in the 2005 Automatic Content Extraction evaluation. The English part of the data is annotated for entities, events and relations. The corpus contains 535 documents marked with within-document coreference of events.

(1) Mention extent
In the ACE data set, event triggers are annotated together with event arguments, that is event participants and attributes. A sentence that describes a taggable event constitutes "the event extent". "The event trigger" is the word that most clearly expresses the event. In ACE the so called "light verbs" are not annotated, while the nouns that they occur with are marked. Neither the "grammatical verbs" nor "aspectuals" are annotated. For phrasal verbs, particles are only annotated if they are not discontinued in a sentence. "Event participants" annotated in ACE are all entities involved with an event mentioned within the extent of an event ("event scope"). Event time and place are seen as "event attributes" together with any other event arguments which are not entities.

(2) Mention POS
Event triggers in the ACE data can be expressed by verbs, nouns, pronouns and adjectives; and more precisely by verbs, adjectives or past participles in the function of sentence predicates, participles or adjectives in modifier position and nouns or pronouns.

(3) Topic coverage
A restricted set of subtypes from the following eight event types is annotated in the ACE data set: LIFE, MOVEMENT, TRANSACTION, BUSINESS, CONFLICT, CONTACT, PERSONNEL and JUSTICE.

(4) Event relations
Additionally, four properties of events are annotated in the ACE corpus: event polarity, tense, genericity and modality. Event coreference restricted to event identity is marked within the scope of a document.

3.1.2 OntoNotes

The English part of the OntoNotes corpus (Pradhan et al. [2007]) consists of 597 texts annotated with within-document identical (anaphoric) and appositive NP coreference (pronominal, nominal and named entity coreference). Events are selectively annotated with coreference mainly if expressed by a noun phrase (NP). Verbal event coreference is marked only if there is a link present to an NP event.

(1) Mention extent

In the OntoNotes corpus (entire) noun phrases and pronouns are annotated as mentions. Verbal mentions are marked as single-word spans, expressed by heads of verb phrases.

(2) Mention POS

Event mentions in OntoNotes can be expressed by (heads of) verb phrases, noun phrases, named entity mentions and pronouns.

(3) Topic coverage

Neither event annotation nor annotation of event coreference were the focus of the OntoNotes annotation effort, which mainly targeted entity mention and entity coreference annotation. Noun phrase annotation in OntoNotes is not restricted to a number of selected semantic types as it is the case with the ACE data set. Verbs (heads of verb phrases to be specific) are annotated in the context of entity coreference annotation.

(4) Event relations

The OntoNotes corpus distinguishes between the identical, that is anaphoric, and appositive coreference (appositives functioning as attributions). Identical coreference is marked with IDENT tag. Appositive coreference is marked as APPOS.

3.1.3 IC corpus

In 2014 the Intelligence Community (IC) domain corpus contains 65 gold-standard documents annotated with a rich set of within-document coreference links (Hovy et al. [2013]) between violent events belonging to an event ontology of ca. 50 terms.

(1) Mention extent

In the IC corpus, single-word spans are annotated as event mentions.

(2) Mention POS

Event mentions can be constituted by verbal, nominal and pronominal phrases.

(3) Topic coverage

The annotation tagset distinguishes between domain and communication events or reportings. Domain events are violent events like *bombings, killings, wars* from an event ontology of ca. 50 terms. There are two types of reportings: locutionary verbs such as *say, report, announce* and speech acts like *condemn, promise, support, blame*. Violent events were annotated with links to any reporting events that introduce a domain event.

(4) Event relations

In the IC corpus not only event identity is considered but several event relations are annotated: full coreference, subevent relation and membership.

3.1.4 Conclusion

TABLE 3.1 summarizes information about corpora described in section 3.1. All three resources have major limitations which made them unfit for our research on unrestricted cross-document event coreference.

First of all, all three corpora discussed here are annotated with within-document event coreference relations. Cross-document relations are not captured. Secondly, in all

		ACE	**OntoNotes**	**IC**
Number of texts		535	597	65 (as of 2014)
Topic coverage		8 event types	unrestricted	violent & reporting events
Within-doc. event coreference	verbal	+	-	+
	nominal	+	+	+
Cross-doc. event coreference		-	-	-

Table 3.1: Corpora annotated with within-document event coreference.

three data sets event coreference annotation is restricted. In ACE only eight event types are annotated with low agreement. The IC corpus considers domain events limited to violent events from an ontology of ca. 50 terms. In the OntoNotes corpus on the other hand, coreference annotation is focused on noun phrases. Verbal event coreference is only captured if there is a link to a noun phrase event.

3.2 Cross-document event coreference in the ECB

		ECB	**ECB 0.1**
Number of topics		43	
Number of texts		482	
Number of action mentions		1744	2533
Number of entity mentions	locations times human participants non human participants	None	5447 entity mentions, no subtypes marked
Number of event coreference chains	within-document	1302	
	cross-document	208	total of 774

Table 3.2: ECB statistics.

The EventCorefBank (ECB, Bejan and Harabagiu [2010]) was, at the time when this research was performed, the only freely available data set annotated with cross-document event coreference.[2] The corpus was specifically designed for the purpose of studies on cross-document event coreference. The data set is organized around corefering events, as opposed to annotating event coreference in a collection of news articles from a topic, from a time period or selected with any other criterion in mind than event coreference. Organizing a data set with the tested phenomenon in mind ensures a more comprehensive coverage of the researched phenomenon. The ECB corpus consists of 43 topics, each corresponding to a seminal event, which in total contain 482 texts from the GoogleNews archive (`http://news.google.com`).

TABLE 3.3 lists the 43 topics / seminal events covered by the ECB (note that topics 15 and 17 are missing). On average there are ca. 11 texts describing a seminal event in

[2] Note that since the time that this work was performed in 2014 new data sets containing cross-document event coreference annotations were created. For instance in 2016 the RED corpus was created (the Richer Event Descriptions corpus, O'Gorman et al. [2016]) and the EER corpus was built (the Event-Event Relation Corpus, Hong et al. [2016]). Corpora released after the release of the ECB+, were not considered in this work.

ECB topic	**Seminal event description**
1	T. Reid checks into rehab in 2008
2	H. Jackman announced as next Oscar host 2010
3	Courthouse escape Brian Nicols Atlanta 2008
4	B. Page dies in LA 2008
5	Philadelphia 76ers fires M. Cheeks 2008
6	”Hunger Games” sequel negotiations C.Weitz 2008
7	W. Klitchko defended IBF, IBO, WBO titles from H. Rahman 2008
8	Bank explosion Oregon 2008
9	Bush changes ESA 2008
10	Angels made an eight year offer to M. Teixeira 2008
11	Parliamentary election in Turkmenistan 2008
12	Indian Navy prevents a pirate attack on an Ethiopian vessel Gulf of Aden 2008
13	Wassila Bible Church fire in Alaska 2008
14	Waitrose supermarket fire in Banstead, Surrey 2008
16	Avenues Gang assassination of J.A. Escalante Cypress Park 2008
18	Deadly office shooting Vancouver 2008
19	Riots in Greece over teenagers death 2008
20	Qeshm island earthquake 2008
21	Bloomington hit and run 2008
22	S.D. Crawford Smith accused of killing co-workers Staunton 2008
23	M. Vinar dies in a climbing accident on Mount Cook 2008
24	4 robbers in drag steal jewelry in Paris 2008
25	The Saints put R. Bush on injured reserve 2008
26	Mafia member G. L. Presti dies in prison Sicily 2008
27	Microsoft releases an IE patch 2008
28	Mark Felt dies in CA 2008
29	Colts beat Jaguars, secure no. 5 seed in the playoffs Fla. 2008
30	France Telecom cable disruption in the Mediterranean 2008
31	T. Hansbrough becomes all-time leading scorer N.C. 2008
32	Gary Gomes double murder New Bedford 2009
33	J. Timmons on trial for stray bullet killing of a 10 year old girl Albany, N.Y. 2008
34	Sanjay Gupta nominated for U.S. Surgeon General 2009
35	V. Jackson arrested under DUI in San Diego 2009
36	W. Blackmore, J. Oler polygamy trial Canada 2009
37	6.1 earthquake Indonesia 2009
38	Small earthquake in Sonoma County 2009
39	Matt Smith role take over “Doctor Who” 2009
40	Apple announces new MacBook Pro CA 2009
41	Israel bombs Jabaliya camp 2009
42	T-Mobile USA adds new BlackBerry model to portfolio 2009
43	AMD acquires ATI 2006
44	Hewlett-Packard acquires EDS 2008
45	S. Peterson found guilty of killing pregnant wife L. Peterson CA 2004

Table 3.3: Overview of the 43 seminal events in the ECB.

the corpus. In ECB texts, a selection of sentences was annotated with within- and cross-document event coreference (amongst other relations). Event mentions were annotated in accordance with the TimeML specification (Pustejovsky et al. [2003]).

The annotation of the ECB, was extended by Lee et al. [2012] into ECB 0.1, following the OntoNotes annotation guidelines (Pradhan et al. [2007], see section 3.1.2). The re-annotation process resulted in more complete sentence annotation and annotation of entity mentions and NP coreference relations, however no specific annotation of entity types was performed. Almost 800 EVENT mentions were annotated and 5447 entity mentions were marked with a cumulative ENTITY tag. TABLE 3.2 compares statistical information about the first version of the ECB corpus and its extended version from 2012.

The EventCorefBank is an important resource, that has been frequently used in studies of event coreference resolution, including those of Bejan and Harabagiu [2008], Bejan et al. [2009], Lee et al. [2012]. Considering ECB's popularity as a data set in event coreference experiments, it is crucial to analyze and be aware of its limitations and how these limitations influence the results of experiments performed on the ECB.

In section 3.3 we will define some requirements that a corpus used for experiments on unrestricted cross-document event coreference should fulfil. Then, in section 3.4, we will examine the representativeness of the ECB corpus as a data set for training and testing of event coreference resolution systems.

3.3 What are the requirements for a data set for experiments on unrestricted cross-document event coreference?

Corpus studies are only valid if the corpus used in them is representative of the sampled language population (Sinclair [2004]). Sampling the diversity of a population is crucial for a sample to be representative of the population. In this section we will take a closer look at the representativeness of a corpus in the context of the task of event coreference resolution. We define some diversity requirements for a corpus that is used as a sample of event coreference in news articles. There are two kinds of diversity that are crucial for an event coreference resource: (1) the diversity of event instances from an event type e.g. multiple presidential elections described and (2) the diversity of event descriptions, that is, coverage of different formulations describing an event instance e.g. *presidential election* or *presidential vote*. We will call the former "referential diversity" and the latter "lexical diversity". In the next section 3.3.1 we will define the referential diversity of a coreference resource. Then we will discuss lexical diversity in section 3.3.2.

3.3.1 Referential diversity

The term "referential diversity" indicates the distribution of event instances that are described in a corpus in relation to the observed concept types (see sections 2.2, 2.4 and 2.5 for definitions of key terms such as event, instance or concept type). The referential diversity of a coreference corpus can be defined as the number of instances covered in a corpus per concept type. Let ARD be the average referential diversity, let ET represent all event types (concept types) annotated in a corpus and let I stand for all instances annotated in the corpus, then:

$$ARD = |1 - (ET/I)| \tag{3.1}$$

Rehab, death, fire, earthquake are examples of event types covered by the ECB corpus. For example, in the ECB there is only one instance of a rehab check-in described, namely the 2008 rehab check-in of Tara Reid so the referential diversity for this event type equals 0. On the other hand, there are 3 earthquakes represented in the ECB corpus: topics 20, 37 and 38. For comparison, a search was done in the earthquake catalog of the USGS (United States Geological Survey, `https://earthquake.usgs.gov`). The search query was intended to find earthquakes of magnitude higher than 2.5 which happened between 1 and 31 July 2014. 2427 earthquakes were found! This indicates the task cut out for a coreference resolver which would be used on daily news streams with thousands of news articles published every working day. A coreference resolver working with online news would have to be able to distinguish between many earthquakes. To make a coreference corpus more representative of event coreference in the news on the web, multiple earthquake descriptions and similarly multiple rehabs and multiple instances of every event type represented in the coreference corpus should be included.

In comparison to the referential diversity of events that can be observed in articles online, there could be a huge discrepancy with event instance diversity in the ECB. In principle, the lower the referential diversity per event type represented in a corpus, the lower the representativeness of the whole event coreference resource. We define the ARD to get an estimate of the referential diversity of a coreference resource.

Let us calculate the ARD for an imaginary corpus in which the following three event types are represented by single instances:

1. event type: "earthquake"
   - instance A: 2011 earthquake in Japan
2. event type: "company acquisition"
   - instance B: Oracle bought NetSuite in 2016
3. event type: "war"
   - instance C: World War II.

The ARD for the imaginary corpus equals |1-(3/3)| which gives us an ARD of 0 which correctly indicates that there is no referential diversity in a corpus in which all event types are only represented by single event instances.

Let us consider another example of a corpus with one more instance covered for one of the three event types:

1. event type: "earthquake"
   - instance A: 2011 earthquake in Japan
2. event type: "company acquisition"
   - instance B: Oracle bought NetSuite in 2016
   - instance C: Sun acquisition by Oracle in 2010
3. event type: "war"
   - instance D: World War II

The ARD for this imaginary corpus equals |1-(3/4)| which gives us an ARD of 25%.

Finally let us consider an imaginary corpus with the most referential diversity compared with the two other example corpora.

1. event type: "earthquake"
    - instance A: 2011 earthquake in Japan
    - instance B: 2008 Sichuan, China earthquake
    - instance C: New Zealand earthquake from 2011
2. event type: "company acquisition"
    - instance D: Oracle bought NetSuite in 2016
    - instance E: Sun acquisition by Oracle in 2010
3. event type: "war"
    - instance F: World War II
    - instance G: Iraq War
    - instance H: World War I
    - instance I: the Vietnam War

The ARD for this imaginary corpus can be calculated as |1-(3/9)| which gives us an ARD of ca. 66%.

The number of event types is not likely to be equal to the number of instances in a coreference corpus, in which case the ARD would equal zero. Moreover, the number of event types from the ARD formula by definition cannot exceed the number of instances, so the ARD will never reach 100%.

3.3.2 Lexical diversity

In this section we will discuss the "lexical diversity" which captures the lexical variation in a corpus. This kind of diversity operates at the level of mentions. Lexical diversity can be defined as the variation in lemmas or lemma combinations (in case of multi-word-unit mentions) among mentions referring to the same instance (see chapter 2 for key term definitions such as lemma, mention or instance). Let ALD stand for the average lexical diversity. Let L represent all lemma combinations annotated as mentions of the same instance, let M stand for all mentions of an instance and let I be the number of all instances from a corpus then we can calculate ALD as follows:

$$ALD = (\sum(L/M))/I \tag{3.2}$$

The calculation of the ALD makes most sense for corpora with coreference instances, that consist of more than one mention. Note that singleton chains by definition have an ALD of 1. The ALD calculation considers the size of coreference chains, meaning that the higher the number of mentions in a chain, the higher number of distinct lemmas is necessary to achieve a higher ALD score. If there are any coreference chains present in a data set, the ALD will never equal zero.

Let us calculate the ALD for an imaginary corpus with the following three instances annotated:

- instance A: *died, (was) killed, death*

- instance B: *acquired, acquisition, purchase, acquired*
- instance C: *World War II, WWII,the Second World War*.

The average lexical diversity for the resource equals ((3/3)+(3/4)+(3/3)) / 3 and will amount to 2.75 / 3 which gives us an ALD of 0.916. The high ALD score captures the relatively high number of distinct lemmas per instance given the number of mentions. The score does not reach 100% as one lemma (*acquire*) is used twice. Next, let us calculate the ALD for an imaginary corpus with two instances both of each described with two mentions expressed with different lemmas:

- instance A: *died, killed*
- instance B: *acquired, purchased*.

In this case, the ALD will equal 1 which indicates an average lexical diversity of 100%. As this example shows the ALD reaches 100% if all annotated mentions from all instances are expressed with different lemmas or lemma combinations.

The fact that coreference chains from a corpus can reach an ALD of 100% does not imply that the degree of ALD in the sampled population is 100%. The score is meant as an indication of the lexical diversity of a sample data set which can be used for comparison of the lexical diversity of coreference resources.

If we consider an imaginary corpus with two instances, both of which described with two mentions expressed with the same lemmas:

- instance A: *died, died*
- instance B: *acquired, acquired*

then the ALD will equal 0.5. The ALD score of 50% reflects the fact that there are four mentions in the imaginary corpus and that those mentions are expressed with a total of two distinct lemmas. Note that the ALD is intended as a relative score for comparison of two coreference resources of comparable size. Our examples illustrate that the bigger the data set, the more distinct lemmas are necessary to achieve a high ALD score. This is a limitation of the measurement but if one takes this deficiency into account, they can successfully use the metric to compare coreference resources.

A coreference resource with low ARD and / or with low ALD might not be representative of diversity that one can come across in articles available online where different formulations might be used to describe multiple instances of events and entities from the same concept type. Low lexical and referential diversity in a coreference corpus will have implications for coreference resolvers trained on such data. In the next section we will evaluate the diversity of the ECB corpus.

3.4 Is the ECB corpus representative of the event coreference problem?

In this section we examine the representativeness of the ECB 0.1 corpus (Bejan and Harabagiu [2010], Lee et al. [2012]) as an evaluation data set for experiments with unrestricted cross-document event coreference resolution. We evaluate the extent to which the ECB 0.1 can be used as a sample of event coreference in news articles. First we quantify the diversity of the corpus, by calculating the average referential diversity in section 3.4.1 and the average lexical diversity in section 3.4.2. Then we illustrate how low corpus diversity influences the task of coreference resolution on the ECB 0.1 in section 3.4.3.

3.4.1 Measuring the average referential diversity of the ECB 0.1

In this section we will take a closer look at the average referential diversity of the ECB 0.1 corpus as the only resource available in 2014 marked with unrestricted cross-document event coreference. We will calculate the average referential diversity (ARD) based on a number of event types represented in the corpus, which we will call ET, together with the number of event instances annotated in the corpus referred to with I. We will use formula 3.1 introduced in section 3.3.1 to determine the average referential diversity of the ECB 0.1.

To be able to calculate the ARD of the ECB 0.1 corpus we need to determine the number of event types (ET) annotated in the corpus and the number of event instances (I) marked in the data set. For the purpose of the calculation, we will assume that Wordnet synsets correspond to event types. We determine the number of ET in the corpus by assigning WordNet synsets to all event mentions manually annotated in the data set.

We used Beautiful Soup [3] - an HTML/XML parser for Python to read in ECB 0.1 corpus texts. EXAMPLE 3.4.1 provides a fragment of file number 1 from topic 38 of the ECB 0.1 corpus.

Example 3.4.1 *<ENTITY COREFID=“4”>An earthquake with a preliminary magnitude of 4.4 </ENTITY><EVENT COREFID=“1”>struck </EVENT>in <ENTITY COREFID=“9”>Sonoma County </ENTITY><ENTITY COREFID=“7”>this morning </ENTITY>near <ENTITY COREFID=“6”>The Geysers </ENTITY>, according to <ENTITY COREFID=“3”>the U.S. Geological Survey </ENTITY>. <ENTITY COREFID=“4”>The earthquake </ENTITY><EVENT COREFID=“1”>struck </EVENT>at <ENTITY COREFID=“7”>about 9:30 a.m. </ENTITY>and <EVENT COREFID=“106”>had </EVENT><ENTITY COREFID=“28”>a depth of 2.7 miles </ENTITY>, according to <ENTITY COREFID=“3”>the USGS </ENTITY>. The quake was centered about two miles from The Geysers and 13 miles east of Cloverdale. <ENTITY COREFID=“29”>Earlier this morning </ENTITY>, <ENTITY COREFID=“5”>an earthquake with a preliminary magnitude of 2.0 </ENTITY><EVENT COREFID=“2”>struck </EVENT>near <ENTITY COREFID=“6” >The Geysers </ENTITY>, according to <ENTITY COREFID=“3”>the USGS </ENTITY>. <ENTITY COREFID=“5”>The earthquake </ENTITY><EVENT COREFID=“2”>struck </EVENT>at <ENTITY COREFID=“29”>about 7:30 a.m. </ENTITY>and <EVENT COREFID=“107”>had </EVENT><ENTITY COREFID=“30” >a depth of 1.4 miles </ENTITY>, according to <ENTITY COREFID=“3”>the USGS </ENTITY>.*

The following six events were extracted from the text.

<event corefid=“1”>struck</event>
<event corefid=“1”>struck</event>
<event corefid=“106”>had</event>
<event corefid=“2”>struck</event>

[3] https://www.crummy.com/software/BeautifulSoup/bs3/documentation.html

```
<event corefid="2">struck</event>
<event corefid="107">had</event>
```

The following lemmas were assigned by the NLTK's WordNet Lemmatizer to the six events. Note that this is the output of the lemmatizer; we will consider these to be lemmas even if they do not always comply with our definition of lemmas as specified in chapter 2.

```
u'struck'
u'struck'
u'had'
u'struck'
u'struck'
u'had'
```

Then the following synsets were assigned to the lemmas on the first position.

```
Synset('strike.v.01')
Synset('strike.v.01')
Synset('have.v.01')
Synset('strike.v.01')
Synset('strike.v.01')
Synset('have.v.01')
```

After removal of duplicates we had two synsets left for file 1 from topic 38:
Synset('strike.v.01') and Synset('have.v.01'). We assumed that these synsets represent event types as defined in chapter 2.

To calculate the number of event types in the corpus for our ARD calculation, first we counted unique synsets per topic and then we extended the set to the entire corpus. This procedure has some implications for the calculation of the total number of event synsets from a data set. In ECB 0.1 one can find the so called "light verbs" annotated without their sentence object which is necessary to express the meaning of an event. In EXAMPLE 3.4.1 we see "have" annotated twice with different event coreference IDs. Both events refer to an earthquake's depth so in this case we will correctly assume that the event type corresponding to both synsets "have" is the same. This assumption is, however, not always correct. On the other hand, if we considered synsets unique per topic, we would end up with an inflated number of event types for all "content-rich" action mentions that carry the same meaning across topics of the corpus such as *war, earthquake, say, murder, arrest, acquire* etc. We preferred to avoid this by considering synsets unique per corpus.

Considering synsets unique for the whole corpus, we ended up with 866 synsets. Note that the actual number of event synsets is in reality somewhat higher because of the issue with "content-poor" actions but this should still give us a good indication of the ARD for comparison with other corpora; assuming that we use the same methodology to count synsets and to calculate the ARD.

To calculate the ARD we also need to know the number of event instances covered by a corpus. According to Lee et al. [2012] there are 774 events annotated in the corpus. However it is not entirely clear how events are defined and whether they correspond to instances as defined in chapter 2. We determined the number of event instances covered in the ECB 0.1 by counting unique event coreference IDs per topic of the corpus. We consider unique coreference IDs per topic, because the IDs do not hold across topics. We calculated that there are 1043 unique coreference IDs in the ECB 0.1. We will use this number for calculation of the ARD.

Following our ARD formula, we calculate the average referential diversity of the ECB 0.1 as |1 - (866 / 1043)|. This gives us a low ARD of 17% for synsets unique in the whole corpus.

3.4.2 Measuring the average lexical diversity of the ECB 0.1

In this section, we measure the average lexical diversity of the ECB 0.1 corpus.

We use formula 3.2 introduced in section 3.3.2 to determine the average lexical diversity of the ECB 0.1. We calculate the average lexical diversity (ALD) per event instance annotated in the corpus, based on a number of unique lemmas L from mentions M describing an event instance. The formula considers the total number of event instances I annotated in the corpus.

Let us look again at EXAMPLE 3.4.1 from the previous section that provides a fragment of the text number 1 from topic 38 of the ECB corpus. In the example text there are six event mentions annotated that describe four event instances - coreference IDs: 1, 2, 106 and 107. By means of the NLTK's WordNet lemmatizer we obtain two unique lemmas assigned to the six event mentions: u'struck' and u'had'. Let us fill out our ALD formula for this example text. In the text there are four event instances mentioned. Instance 1 is referred to with two mentions: *struck*. Instance 2 is also referred to with two mentions: *struck*. Then we have instances IDs 106 and 107, each described by one mention: *had*. If one adds up the calculation of L/M per event instance one would get the following: 1/2 (instance ID 1) plus 1/2 (instance ID 2). Then (1/2 + 1/2) / 2 gives us 1/2 which is an ALD of 50%. Of course this calculation was only done for a single text to exemplify the calculation of the ALD. For a complete score of the ALD for ECB 0.1 we also need to consider any other mentions from other texts of the instances 1, 2 and also 106 and 107 if they are not singleton chains. We do this for all corpus texts and all event instances with all their mentions and lemmas.

As in section 3.4.1, we consider unique coreference IDs per topic. There are 1043 unique coreference IDs in the ECB 0.1. The sum of unique lemmas from mentions of each instance amounts to 187.3. The ALD can be calculated as 187.3 / 1043 = 0.179.

The ALD of the ECB 0.1 corpus is similar to the ARD. It is also very low, only 18%. For this calculation singleton instances were not considered as explained in section 3.3.2.

3.4.3 Influence of low corpus diversity on cross- and within-topic coreference resolution on ECB 0.1

Sections 3.4.1 and 3.4.2 show that the average referential and the average lexical diversity of the ECB 0.1 corpus are very low. We hypothesise that the low average lexical and referential diversity of a coreference resource have a strong influence on results of coreference experiments. To determine the interplay between the low diversity of the ECB 0.1 corpus and the task of event coreference resolution, we perform two experiments in which we consider coreference chains generated based on lemma matches of event actions across topics and within topics of the corpus.

Method

For the purpose of the experiments, we created chains of corefering events based on lemma matches of mentions of event actions. We used tools from the Natural Language Toolkit (Bird et al., 2009, NLTK version 2.0.4): the NLTK's default word tokenizer and POS tagger (POS tagger for the purpose of proper verb lemmatization) and WordNet lemmatizer[4].

We ran two experiments. We will call them E1 and E2. In E1 we made the assumption that a lemma represents an event instance for the whole corpus. In E2 we assumed that a lemma represents an event instance per topic of the corpus.

	MUC			**B3**			**CEAF**	**BLANC**			**CoNLL**
	R	P	F	R	P	F	R/P/F	R	P	F	F
E1	54.29	53.80	54.04	60.04	59.05	59.54	40.00	61.56	54.98	56.91	52.52
E2	51.83	83.16	62.19	59.40	92.75	71.23	61.03	63.10	84.09	65.53	64.76

Table 3.4: Event coreference resolution based on lemma matches of actions in experiment E1 cross-topic and in experiment E2 within-topic matches, evaluated on the ECB 0.1 in MUC, B3, mention-based CEAF, BLANC and CoNLL F.

Results

TABLE 3.4 shows the results of E1 and E2 achieved by means of lemma matches of event action mentions in terms of recall (R), precision (P) and F-score (F) by employing five coreference resolution evaluation metrics: MUC (Vilain, 1995), B3 (Bagga and Baldwin, 1998), mention-based CEAF (Luo, 2005), BLANC (Recasens and Hovy [2011]), and CoNLL F1 (Pradhan et al. [2011]).[5] Both lemma approaches E1 and E2 achieved surprisingly good results. In the cross-topic lemma experiment E1 coreference between events was solved with an F-score of 54.04% MUC, 59.54% B3, 40.00% CEAFm, 56.91% BLANC F and 52.52% CoNLL F1. Restricting lemma matches to action mentions from a topic in the experiment E2 bought us a 20-30 point increase of precision across the evaluation metrics; and an 8-20 point improvement of the F-scores.

The following results were achieved in studies of event coreference resolution in related work, for easier comparison see TABLE 3.5.

- Bejan and Harabagiu [2010]: 83.8% B3 F, 76.7% CEAF F on the ACE (2005) data set and on the ECB corpus 90% B3 F, 86.5% CEAF F-score
- Chen et al. [2011]: 46.91% B3 F on the OntoNotes 2.0 corpus
- Lee et al. [2012]: 62.7% MUC, 67.7% B3 F, 33.9% (entity based) CEAF, 71.7% BLANC F and 54.8% CoNLL F-score on the ECB 0.1 corpus.

Note that the different approaches from TABLE 3.5 are not directly comparable. In our experiments we used true mentions instead of system mentions (for definitions of true and system mentions see section 2.2), as the point of our tests was to expose the division of work in a multi-step approach to event coreference resolution. Furthermore, a comparison with studies performed on different data sets is difficult. Bejan

[4] www.nltk.org/_modules/nltk/stem/wordnet.html

[5] See section 2.7 for a discussion of coreference resolution evaluation metrics.

	MUC-F	B3-F	CEAF-F	BLANC-F	CoNLL-F
B&H on ACE	-	83.8	76.7	-	-
B&H on ECB	-	90	86.5	-	-
Chen on OntoNotes	-	46.91	-	-	-
Lee on ECB 0.1	62.7	67.7	33.9	71.7	54.8
E1-lemma baseline on ECB 0.1	54.04	59.54	40	56.91	52.52
E2-lemma baseline on ECB 0.1	62.19	71.23	61.03	65.53	64.76

Table 3.5: Event coreference resolution based on lemma matches of actions in experiment E1 cross-topic, in experiment E2 within-topic and in related work, evaluated in MUC, B3, mention-based CEAF, BLANC and CoNLL F.

and Harabagiu [2010] worked on the ACE data set and on the ECB which are not appropriate for and simplify the task of cross-document event coreference resolution (see section 3.1 and 3.2 for details). The same applies to the OntoNotes corpus, which Chen et al. [2011] worked on. A comparison with Lee et al. [2012], who like us, worked with the ECB 0.1 corpus is the most informative. In their study the same data set was used that is evaluated in the lemma experiments. Furthermore, Lee et al. [2012] report the most reliable BLANC-F and CoNLL-F scores for their experiment (see section 2.7 for a discussion of evaluation measures used in coreference resolution).

Even though the lemma approaches neither perform anaphora resolution nor employ entities or any syntactic features, the CoNLL-F score reached in E1, is only ca. 2 points lower than the CoNLL-F score achieved by the sophisticated approach by Lee et al. [2012]. The CoNLL-F score reached in E2 is even 10% higher that the CoNLL-F score reached by the approach by Lee et al. [2012], however it must be noted that E1 and E2 solve coreference on true mentions which makes the task easier than working with system mentions. Looking at the BLANC-F measure, the within-topic lemma match heuristic from E2 reached 65.53 BLANC F-score which is only ca. 6 points less than the state of the art approach on the ECB 0.1 corpus by Lee et al. [2012]. It is remarkable to see that with the simple lemma match heuristic we obtained results comparable to those achieved by means of a sophisticated machine learning approach trained on the same corpus in related work.

Discussion: low referential and lexical diversity simplify the task of event coreference resolution

The high scores achieved by both lemma approaches in E1 and in E2 give rise to at least two crucial conclusions. First, coreference chains based on lemma matches reach very high evaluation scores. There is relatively little lexical diversity in descriptions of event actions from ECB 0.1 coreference chains. Second, event entities, that is participants, times or locations of events, do not seem to play a crucial role in event coreference resolution, at least not if one evaluates on this data set.

Event coreference chains in the news are much more lexically and referentially diverse than in the ECB 0.1. Compare the following two pairs of event descriptions as:

- *car bombing in Madrid in 1995*

- *bombing in Spain in 2009*,

or

- *massacre in Srebrenica*
- *genocide in Rwanda.*

The ECB corpus, while containing multiple documents describing particular seminal events, in most cases captures only one seminal event per topic. For instance, texts from ECB topic one, describing Tara Reid's check-in into rehab in 2008, constitute the only rehab-related event coreference chain in the corpus; and so the only instance of a rehab check-in event captured by the corpus. It is understandable that if testing event coreference resolution on such data set, event entities will not seem to play a big role in resolution of coreference between events. The number of seminal events described per topic is limited. As in most cases there is only one seminal event per topic, with exception of a few topics like *earthquake, acquisition, death* and *fire*, event descriptions from a particular topic tend to share their entities (see section 4.1 for a complete overview of seminal events in the ECB). The referential diversity is low. And with a low referential diversity of a corpus, the event coreference task becomes simplified to topic classification.

It is a common practice to perform text categorization before solving event coreference for example Lee et al. [2012] use a variant of the Expectation Maximization (EM) algorithm to perform document clustering. When performing topic classification or topic modelling texts are grouped into topical classes based on their vocabulary. If topics in a coreference corpus correspond to seminal events, once documents are classified with topics, an event coreference resolver does not need to do any reasoning to distinguish between event instances from the same event type within a topic as there is only one event instance described per event type. Lee et al. [2012] observed that document clustering by means of the EM model on the ECB corpus performed well on the training data and only two topics describing different instances of earthquake events were merged into one document cluster. This is expected when working with a corpus with a limited number of seminal events per topic. However, this is not what one will come across in the news.

In experiment E2 again we used lemma matches of action mentions to generate event coreference chains. However, this time we considered lemma matches within each topic of the ECB corpus. E2 illustrates the interaction between event coreference and the referential diversity per topic. The preliminary topic classification strongly influences the coreference resolution on the ECB corpus. The results expose the diversity of event coreference chains within a topic, resembling the task of solving event coreference after the first step of topic classification, as performed in most recent approaches to event coreference resolution. E2 exposes the division of work in a multi-step machine learning approach to coreference resolution. We see that much of the work on a data set like the ECB is done within the topic classification step. We make the assumption that the situation looks different if one considers a large corpus of news articles from a longer period of time, where different topics are represented by multiple event instances of the same type (for instance multiple celebrities going into rehab, or the same celebrity reentering a rehab facility). Our expectations are that when solving event coreference on a corpus with multiple instances representing an event type, topic classification will still make the task easier. The task difficulty, however, will significantly increase, as on top of matching lexically compatible action mentions (which in the second experiment gave us an CoNLL F score of ca. 65%), a system will also have to make a distinction between mentions of different instances of the same event type.

3.5 Conclusion

In this chapter we analyzed corpora annotated with event coreference relations. Most event coreference data sets available for research in 2014 were only annotated with within-document event coreference. Only the ECB corpus was annotated with unrestricted cross-document event coreference.

In the context of the task of event coreference resolution, we analyzed the diversity of the ECB as a sample of the language population of news articles. We introduced a methodology to quantify the referential and the lexical diversity of a coreference resource. We calculated the average referential diversity and the average lexical diversity of the ECB 0.1 data. Both the ARD and ALD scores for ECB 0.1 are very low. The ARD of the ECB 0.1 corpus amounts to 17% while the ALD is 18%. These low diversity scores quantify the low complexity of the data set from the point of view of event coreference.

The analysis of the corpus and some tests with lemma heuristics show that low referential and lexical diversity of the ECB 0.1 strongly influences event coreference evaluation results achieved on the corpus. The ECB in most cases covers one seminal event per domain, which considerably simplifies event instance and also language diversity that one would come across in the news. The results obtained on the ECB corpus cannot be generalized onto the sampled language population which is expected to be much more diverse referentially and lexically. One cannot assume that approaches to event coreference resolution tested on the ECB would perform with comparable accuracy when solving event coreference between event mentions extracted from news online which cover multiple instances of events from an event type.

To increase the referential diversity of the ECB corpus, we extended the corpus with a new corpus component of 502 texts covering new instances from event types covered by the ECB. In chapter 4 we present the new resource called ECB+. After describing the new corpus, we will evaluate its contribution in section 4.8 by measuring the referential and lexical diversity and comparing the ARD and the ALD scores with those achieved on the ECB in chapter 3.

Part II

How can we make a corpus more representative of the event coreference problem?

In the first part of the dissertation we researched corpora annotated with event coreference with regard to their representativeness of news articles. The ECB corpus was at the time the only data set with focus on capturing cross-document event coreference. However, my analysis showed that the corpus has a low referential and lexical diversity which strongly influence event coreference experiments. In the second part of the dissertation we will investigate how a corpus can be made more representative of the event coreference problem. We will primarily focus on representing the referential diversity of news articles in a corpus. We will reuse the ECB 0.1 data set and we will extend it to increase its diversity and to make the ECB sample more appropriate to study event coreference resolution. This part of the dissertation addresses the research questions below.

Research questions

- How can one obtain an empirically valid data set on event coreference in the news that is representative of the language population of news articles?
- How should a data set that is meant for research on event coreference be organized?
- How should one reflect the lexical and the referential diversity of news articles in a corpus?
- What is the importance of entities for event coreference resolution and how should an event coreference corpus be annotated to facilitate research on the topic?

Chapter 4

ECB+ data set[1]

This chapter presents the ECB+ corpus (Cybulska and Vossen [2014b]), which is an extension to the ECB corpus (Bejan and Harabagiu [2010]). The new data set was created to increase the referential diversity of the ECB corpus so that the corpus is more representative of event coreference in news articles (see chapter 3). Section 4.1 describes how we augmented the ECB corpus to increase its referential diversity. Section 4.2 discusses the annotation guideline used in the creation of the new resource to mark event information in text. Section 4.3 elaborates on the event-centric annotation style of the ECB+ and 4.4 on the annotation of the coreference relation. In section 4.5 the setup of the annotation task is described and in 4.7 we discuss the inter-annotator agreement. Section 4.9 offers a conclusion.

4.1 Increasing the referential diversity of an event coreference corpus

This work makes a distinction between "mentions" and "instances" that mentions refer to (see terminology definitions in chapter 2). All mentions of events that refer to the same event instance are related through coreference relation. The term "referential diversity" introduced in section 3.3.1 indicates the distribution of entity or event instances that are mentioned in a corpus.

This section explains how the referential diversity with regards to event instances has been increased from the ECB to ECB+. The ECB corpus consists of 482 texts[2] from 43 topics. Each topic is a collection of texts describing one seminal event. The vast majority of seminal event types are represented in the corpus by a single event instance. In section 3.4.1 we measured the average referential event diversity of the ECB 0.1. The ARD of the corpus is only 17%.

With the objective to make the ECB corpus more representative of news articles, we increased the referential diversity of the dataset. We augmented the 43 topics of the ECB with 502 texts reporting different instances of the same event types provided in the ECB. Thus, we were targeting events that happened at a different time, place or with different participants. For example, the first ECB topic consists of texts outlining

[1] The contents of this chapter have been published as Cybulska and Vossen [2014b] and as Cybulska and Vossen [2014a]

[2] Note that two texts (text 4 from topic 7 and text 13 from topic 19) were missing from the version of the ECB 0.1 data (Lee et al. [2012]) which we found on the web.

T	Seminal event ECB	Seminal event new component ECB+
1	T. Reid checks into rehab in 2008	L. Lohan checks into rehab in 2013
2	H. Jackman announced as next Oscar host 2010	E. Degeneres announced as next Oscar host 2014
3	Courthouse escape Brian Nicols Atlanta 2008	Prison escape A.J. Corneaux Jr. Texas 2009
4	B. Page dies in LA 2008	E. Williams dies in LA 2013
5	Philadelphia 76ers fires M. Cheeks 2008	Philadelphia 76ers fires J. O’Brien 2005
6	”Hunger Games” sequel negotiations C.Weitz 2008	”Hunger Games” sequel negotiations G. Ross 2012
7	W. Klitchko defended IBF, IBO, WBO titles from H. Rahman 2008	W. Klitchko defended IBF, IBO, WBO titles from T. Thompson 2012
8	Bank explosion Oregon 2008	Bank explosion Athens 2012
9	Bush changes ESA 2008	Obama changes ESA 2009
10	Angels made an eight year offer to M. Teixeira 2008	Red Socks made an eight year offer to M. Teixeira 2008
11	Parliamentary election in Turkmenistan 2008	Parliamentary election in Turkmenistan 2013
12	Indian Navy prevents a pirate attack on an Ethiopian vessel Gulf of Aden 2008	Indian Navy prevents a pirate attack on merchant vessels Gulf of Aden 2011
13	Wassila Bible Church fire in Alaska 2008	Mat-Maid Dairy fire in Alaska 2012
14	Waitrose supermarket fire in Banstead, Surrey 2008	Waitrose supermarket fire in Wellington 2013
16	Avenues Gang assassination of J.A. Escalante Cypress Park 2008	Hawaiian Gardens assassination of sheriffs deputy J. Ortiz Hawaiian Gardens 2005
18	Deadly office shooting Vancouver 2008	deadly office shooting Michigan 2007
19	Riots in Greece over teenagers death 2008	riots in Brooklyn over teenagers death 2013
20	Qeshm island earthquake 2008	Qeshm island earthquake 2005
21	Bloomington hit and run 2008	Queens hit and run 2013
22	S.D. Crawford Smith accused of killing co-workers Staunton 2008	Y. Hiller accused of killing co-workers Philly 2010
23	M. Vinar dies in a climbing accident on Mount Cook 2008	R. Buckley, D. Rait die in climbing accidents on Mount Cook 2013
24	4 robbers in drag steal jewelry in Paris 2008	4 robbers steal jewelry in Paris 2013
25	The Saints put R. Bush on injured reserve 2008	The Saints put P. Thomas on injured reserve 2011
26	Mafia member G. L. Presti dies in prison Sicily 2008	Mafia member V. Gigante dies in prison Montana 2005
27	Microsoft releases an IE patch 2008	Microsoft releases an IE patch 2013
28	Mark Felt dies in CA 2008	Fred LaRue dies in Miss. 2004
29	Colts beat Jaguars, secure no. 5 seed in the playoffs Fla. 2008	Colts beat Chiefs, secure no. 5 seed in the playoffs Missouri 2012
30	France Telecom cable disruption in the Mediterranean 2008	Seacom cable disruption Egypt 2011
31	T. Hansbrough becomes all-time leading scorer N.C. 2008	D. McDermott becomes all-time leading scorer Missouri 2013
32	Gary Gomes double murder New Bedford 2009	John Jenkin double murder Cumbria 2013
33	J. Timmons on trial for stray bullet killing of a 10 year old girl Albany, N.Y. 2008	A. Lopez on trial for stray bullet killing of Z. Horton Brooklyn 2011
34	Sanjay Gupta nominated for U.S. Surgeon General 2009	Regina Benjamin nominated for U.S. Surgeon General 2013
35	V. Jackson arrested under DUI in San Diego 2009	J. Williams arrested under DUI in San Diego 2009
36	W. Blackmore, J. Oler polygamy trial Canada 2009	Jeff Warren polygamy trial Texas 2011
37	6.1 earthquake Indonesia 2009	6.1 earthquake Indonesia 2013
38	Small earthquake in Sonoma County 2009	Small earthquake in Sonoma County 2013
39	Matt Smith role take over “Doctor Who” 2009	Peter Capaldi role take over “Doctor Who” 2013
40	Apple announces new MacBook Pro CA 2009	Apple announces new MacBook Pro CA 2012
41	Israel bombs Jabaliya camp 2009	Sudan bombs Yida camp 2011
42	T-Mobile USA adds new BlackBerry model to portfolio 2009	T-Mobile USA adds new BlackBerry model to portfolio 2012
43	AMD acquires ATI 2006	AMD acquires Seamicro 2012
44	Hewlett-Packard acquires EDS 2008	Hewlett-Packard acquires EYP 2007
45	S. Peterson found guilty of killing pregnant wife L. Peterson CA 2004	C. K. Simpson found guilty of killing pregnant girlfriend K. M. Flynn Mississippi 2013

Table 4.1: Overview of seminal events in ECB & ECB+. The column labeled “T” indicates topic numbers.

Tara Reid's check-in into rehab in 2008. We created an extension to topic number one of the ECB. We extended this topic with texts describing another event instance of the same type, namely Lindsay Lohan going into a rehab facility in 2013. For most topics, originally there was only one seminal event described in the ECB corpus. For each topic we collected descriptions of a second seminal event, adding on average 11 texts to each topic. As a result, we ended up with a corpus that covers at least two seminal events from each topic. TABLE 4.1 shows the complete overview of all 43 seminal events captured by ECB+, taken from the ECB+ annotation guideline (Cybulska and Vossen [2014a]).

ECB+ texts were collected by means of the Google News search. We googled (parts of) descriptions of the 43 seminal events from the ECB corpus to find texts describing other event instances of the same event type. We increased the number of event instances in the corpus from 774 as reported by Lee et al. [2012] to 1958 in ECB+. This is ca. 2.5 times more instances annotated in the corpus.

TABLE 4.2 shows some examples of seminal events broken down into components, as captured per topic in both components of the corpus, the original ECB and in the new component of ECB+. Next to a seminal event per topic, human participants involved with the seminal events as well as their times and locations are listed. Note that the ECB+ extension of the ECB purposefully targets the diversity of event times, locations and participants per event type. This is a data design choice. Artificially diversifying instances of event types described in a coreference corpus is necessary to obtain an empirically valid data set on event coreference in the news that is naturally filled with descriptions of multiple instances of an event type that happened at different times, locations and with different participants.

Topic	Seminal event type	Human participant		Time		Location	
		ECB	ECB+	ECB	ECB+	ECB	ECB+
1	rehab check-in	T.Reid	L.Lohan	2008	2013	Malibu	Rancho Mirage
2	Oscars host announced	H.Jackman	E.Degeneres	2010	2014	-	-
3	inmate escape	Brian Nicols, 4 dead	A.J. Corneaux Jr.	2008	2009	court-house Atlanta,	prison, Texas
4	death	B.Page	E.Williams	2008	2013	LA	
5	head coach fired	Philadelphia 76ers, M.Cheeks	Philadelphia 76ers, J.O'Brien	2008	2005	-	-
6	"Hunger Games" sequel negotiations	C.Weitz	G.Ross	2008	2012	-	-
7	IBF,IBO,WBO titles defended	W.Klitchko, H.Rahman	W.Klitchko, T.Thompson	2008	2012	Germany	Switzerland
8	explosion at a bank	-	-	2008	2012	Oregon	Athens
9	ESA changes	Bush	Obama	2008	2009	-	-
10	eight-year offer	Angels, M.Teixeira	Red Socks, M.Teixeira	2008		-	-

Table 4.2: Overview of seminal events in ECB and ECB+ topics 1-10.

This is why in the ECB+ we model events from news data as a combination of four components (see also section 2.2):

1. an event action component describing what happens or holds true
2. an event time component anchoring an action in time describing when something happens or holds true
3. an event location component specifying where something happens or holds true
4. a participant component that gives the answer to the question: who or what is involved with, undergoes change as result of, or facilitates an event or a state. We divide event participants into human participants and non-human participants.

In the ECB+, we annotated the 43 newly selected seminal events. Event action mentions were annotated in selected sentences together with mentions of their times, locations and participants. Any other events mentioned in the same sentence that describes a seminal event were also annotated. Accordingly every event of a sentence is annotated.

The annotation task required annotators to annotate mentions of event actions, times, locations and participants. In section 4.2, we explain how every event component was annotated in the text. First we discuss actions in section 4.2.1, then we take a closer look at times, locations and participants in section 4.2.2. We look at three aspects of mention annotation per event component: (1) mention extent, (2) mention part of speech and (3) mention typology. Thus for a component, after explaining how the annotators determined the extent of component mentions in text, we give an overview of how a component mention can be expressed in language. Finally, we present the tags that are used to annotate a component. We summarize annotation decisions made with regard to all event components in section 4.2.3.

4.2 ECB+ mention annotation guideline

There are some major differences between the annotation style of the ECB 0.1 corpus (Lee et al. [2012] and Recasens [2011]) and of the ECB+ corpus. In the ECB+ annotation scheme, we made an explicit distinction between action classes and entity types. We do not only have event actions and entities annotated as was done in ECB 0.1, which distinguishes between ACTION and ENTITY, but we also know precisely whether an entity is a human event participant, non-human participant, time or location. We decided to annotate the entity types to make it possible to explore the relationship between event coreference and a particular entity type. Similar reasoning guided our decisions to annotate entity subtypes. We observed that the entity types are rather broad and sometimes differ a lot in their linguistic behavior. Therefore, we have annotated a number of more specific subtypes within the four entity types e.g. HUMAN_PART_PER for human participants of subtype individual person. The same applies to actions that were re-annotated with specific action classes. The complete ECB+ annotation guideline can be found in Cybulska and Vossen [2014a]. In the ECB+ annotation scheme we distinguish 30 annotation tags, taking from Linguistic Data Consortium [2008], Pustejovsky et al. [2003] and Saurí et al. [2005].

The next two sections elaborate on how event components were annotated in text. Section 4.2.1 discusses the annotation of the action component. Section 4.2.2 explains how event times, locations and participants were annotated. Both sections are structured in the same way. First, we describe how the extent of a mention was determined. Then we specify what part of speech a component mention can be expressed by. Next, semantic classes are discussed that were distinguished for the purpose of ECB+ annotation. The classes correspond to annotation tags. Finally, we summarize the annotation decisions in an annotation checklist. Section 4.2.3 concludes section 4.2 with a checklist about ECB+ annotation of all event components with regards to mention extent, mention POS and annotation tags used per component.

In the remainder of this chapter we will underscore words exemplifying how particular aspects of annotation should be annotated. All examples are presented in *italics*. Note that in the given examples not all actions are annotated, but only those that exemplify the construction discussed in the current paragraph.

4.2.1 Annotation of action mentions

An event action mention describes what happens or holds true. Most actions described in the news are instances (or sets of instances) of abstract classes of actions that already happened, are happening, or are expected to happen at a particular time and place, with or without involvement of participants. Thus the majority of action mentions that one encounters in ECB+ are anchored in time and space but one could also come across generic actions in text. Even though we did not expect generic actions, that are not anchored in time and space, to be crucial for the ECB+ annotation task for the sake of completeness the ECB+ annotation guideline does facilitate marking of mentions of abstract, generic events (and coreference between them) if the events occur in a sentence that describes a seminal event.

4.2.1.1 Mention extent

In this section we explain how action mentions are annotated in ECB+. We start by elaborating on how the annotators determined the extent of component mentions in text.

Verbal and nominal actions

Whether an action mention is verbal (like *the earth quaked*) or nominal (like *the earthquake*), the annotators were instructed to always annotate the word or words that are the strongest carrier of the action meaning; i.e. the (semantic) head of an action phrase, as illustrated in EXAMPLES 4.2.1–4.2.4.

Example 4.2.1 *People would rather hear the positive things being talked about than the negatives.*

Example 4.2.2 *The mall gunman may have been shooting at security cameras.*

Example 4.2.3 *FBI did not investigate Fort Hood shooter.*

Example 4.2.4 *This terrible war could have ended in a month.*

EXAMPLES 4.2.1–4.2.4 show how other parts of the action phrases like *would, may have been, did not, this terrible* and *could have* were left unannotated. In verbal phrases, the “auxiliary” verbs, that express, for instance, grammatical tense of a sentence, are not annotated. The same holds for polarity markers applying to actions (e.g.

negation words like *not*). We indicated negation in a different way (as explained in section "Action classes" in 4.2.1.3). With the exception of auxiliary verbs, all other verbs including aspectuals (like *start, stop, continue*) and causative verbs (like *cause*) were annotated as separate action mentions. Consider the extent of action mentions annotated in EXAMPLES 4.2.5 and 4.2.6.

Example 4.2.5 *Another report stated that the fighting started after a high-speed chase with a suspect vehicle in which a Gaddafi loyalist was killed.*

Example 4.2.6 *The earthquake caused ruptures on the surface for a length of 470 kilometers.*

Proper names

Some historically significant event instances have their own name. Writers tend to refer to these events not in a descriptive way, but instead with those so-called proper names. Examples include *9/11, 9/11September 11* or *World War II*. These event descriptions were annotated with all their elements as illustrated in 4.2.7.

Example 4.2.7 *First national memorial dedicated to all who served during World War II.*

Predicative phrases

The same verbs that can express grammatical properties of a main verb (auxiliary verbs) can also be used as main verbs themselves in constructions with predicative phrases. In EXAMPLE 4.2.8, the verb *to be* is used as an auxiliary.

Example 4.2.8 *The mall gunman may have been shooting at security cameras.*

Comparatively, in EXAMPLES 4.2.9 and 4.2.10, this same verb is used as the syntactic main verb.

Example 4.2.9 *Kittens are cute.*

Example 4.2.10 *These people are amazing.*

In EXAMPLES 4.2.9 and 4.2.10, just as in the case of auxiliaries in 4.2.8, we did not annotate the verb *to be* but we only annotate the nominal, pronominal or adjectival part of the predicative phrase, as marked in the two examples above. Let us take a look at two more examples of predicative phrases in 4.2.11 and 4.2.12.

Example 4.2.11 *Gunman in Texas shooting was a marine.*

Example 4.2.12 *Game Five hero David Ross was happy just to be here.*

Marine, happy and *here* should all be tagged as actions (to be specific actions of the class "state", as explained further in the section "Action classes" in 4.2.1.3). At the same time, if the location, time or participant is also part of a predicative phrase, it should also be tagged as such (see for more information section 4.2.2 on time and entity annotation). Copular constructions with predicative phrases are a special case in which the number of annotated mentions might not correspond to the actual number of event participants as in the sentence from EXAMPLE 4.2.13.

Example 4.2.13 *Aaron is my favorite writer.*

In 4.2.13 we would annotate two mentions referring to a single participant referent of the state. With the exception of such copular constructions, in the ECB+ annotation the number of mentions marked per action corresponds to the number of "actual" event participants, times and locations (see section 4.3).

Combination of a verb and a noun

There are a number of verbs (including the so-called "light verbs") that without a noun do not express the full action meaning. If one omits either the noun or the verb of such an action expression, a part of the meaning is lost; for example phrases like *make an offer, witness an attack, interrupt a meeting* or *prevent an assassination*. For action mentions constituted by a combination of a verb and a noun, both parts of the action phrase were annotated separately from each other to preserve the full meaning; the verb as an action and the noun depending on the component that it refers to. In EXAMPLE 4.2.14 we have an action and a human participant entity of type person.

Example 4.2.14 *Congress did not back Barack Obama.*

It could be the case that the noun also refers to an action and then it was also annotated as an action. In EXAMPLE 4.2.15 we have two action mentions.

Example 4.2.15 *Russia has made an offer to Syria.*

4.2.1.2 Mention part of speech

In this section we give an overview of how actions can be presented in text. Note that in the examples not all action mentions are annotated, but only those that exemplify the construction shown in a bullet point.

We annotated action mentions that are expressed by verbs, nouns, present- and past-participles, predicative phrases and pronouns. Below are some examples of action mentions expressed by different part of speech. As illustrated in EXAMPLES 4.2.16–4.2.18 action mentions can be expressed by **verbs**.

Example 4.2.16 *Syrian army fights rebels for control of key Christian town.*

Example 4.2.17 *Indonesia GDP grows less than 6%.*

Example 4.2.18 *At least 17 Taliban militants have been killed by Afghan and coalition security forces during the past 24 hours.*

EXAMPLES 4.2.19–4.2.21 show actions described by **nouns**, including (but not limited to) nominalizations and proper nouns.

Example 4.2.19 *The Civil War ended back in 1865.*

Example 4.2.20 *Fast economic growth across the African continent...*

Example 4.2.21 *Two arrested in the killing of a student.*

EXAMPLES 4.2.22–4.2.23 illustrate action mentions with the attributive use of **present- and past- participles** in modifier position.

Example 4.2.22 *The deceased mens' house was sold yesterday.*

Example 4.2.23 *The crying baby had a high fever.*

Next, EXAMPLES 4.2.24–4.2.25 show action mentions in **predicative phrases** expressed by adjectives, pronouns or nouns, also as part of noun phrases or prepositional phrases (occurring with copular verbs, like constructions in which the verb *to be* is used as the main action verb and not as auxiliary).

Example 4.2.24 *Gunman in Texas shooting was a marine.*

Example 4.2.25 *Game Five hero David Ross was happy just to be here.*

Finally, in EXAMPLE 4.2.26 we see an action expressed by **a pronoun**.

Example 4.2.26 *A small earthquake has hit Japan's eastern coast yesterday. It did not trigger a tsunami.*

4.2.1.3 Action classes

We did not annotate mentions of actions with one general action tag but we specified the class an action belongs to instead. We annotated action mentions with a limited number of classes from the whole set defined in the TimeML annotation guideline (Saurí et al. [2005]). We adopted five event classes from the TimeML specification:

- OCCURRENCE
- PERCEPTION
- REPORTING
- ASPECTUAL and
- STATE (Pustejovsky et al. [2003]).

The action tags that were used in the annotation process, together with explanation of their coverage and examples from TimeML follow.

- The ACTION_OCCURRENCE tag is appropriate for most action mentions in the news, "describing something that happens or occurs in the world" such as *die, crash, build, merge, sell, land, arrive, distribute, eruption, explosion*.
- The ACTION_PERCEPTION tag refers to actions "involving the physical perception of another event" e.g.: *see, hear, watch, feel, glimpse, behold, view, hear, listen, overhear*.
- The ACTION_REPORTING tag was used to annotate reporting actions describing "the action of a person or an organization declaring something, narrating an event, informing about an event" such as *say, report, tell, announce, explain, cite, state*.
- The ACTION_ASPECTUAL tag was used to express "focus on different facets of event history" e.g.: *begin, finish, stop, continue* as in EXAMPLE 4.2.27.

  Example 4.2.27 *The Civil War ended back in 1865.*

The TimeML annotation guideline (Saurí et al. [2005]) distinguishes between five facets of event history: *initiation, reinitiation, termination, culmination* and *continuation* of an event.

- The ACTION_STATE tag "describes circumstances in which something obtains or holds true" such as *(be) on board, hope, love, shortage, (was) an actor, live, the crisis, peace*. This tag is to be assigned to the non-verbal part of predicative phrases (constructions with verb *to be* + nominal/pronominal/ adjectival part), among others.

Additionally we employed two more action classes, one for causal events and one for generic actions.

- The ACTION_CAUSATIVE tag is meant for action mentions such as *cause, lead to, result, facilitate, induce, produce, bring about*.
- The ACTION_GENERIC tag was used to annotate generic events that are not anchored in time or space, see EXAMPLES of generic actions from the TimeML specification (Saurí et al. [2005]) in 4.2.28 and 4.2.29.

Example 4.2.28 *Use of corporate jets for political travel is legal.*

Example 4.2.29 *The rabbi said Jews are prohibited from killing one another.*

These seven classes have seven equivalents to indicate polarity of the event. Polarity provides insight into whether the event did or did not happen. Negation of events can be expressed in different ways, including the use of negative particles with regard to verbs (like *not, neither*), other verbs (like *deny, avoid, be unable*), or by negation of participants involved with an event as in EXAMPLE 4.2.30.

Example 4.2.30 *No soldier went home.*

We annotated negation as a property of sentence actions by means of a set of action classes based on the seven base action classes but with indication of negation through addition of a *NEG_* tag in front of each action class. The following tags were used to indicate negation:

- NEG_ACTION_OCCURRENCE
- NEG_ACTION_PERCEPTION
- NEG_ACTION_REPORTING
- NEG_ACTION_ASPECTUAL
- NEG_ACTION_STATE
- NEG_ACTION_CAUSATIVE
- NEG_ACTION_GENERIC.

4.2.1.4 Action annotation summary

In TABLE 4.3 we outline the main decisions made with regards to annotation of actions in the ECB+ corpus.

In the following section 4.2.2 we discuss how times and entities were annotated in text.

Language phenomenon	**Treatment in ECB+**
Action classes	Five TimeML classes, causative and generic actions & seven negated classes
Auxiliary verbs (incl. auxiliary modals)	Not annotated
Light verbs	Annotated
Phrasal verbs and idioms	All elements annotated also if discontinued
Aspectuals	Annotated with a separate tag
Causative verbs	Annotated with a separate tag
Generic events	Annotated with a separate tag
Event negation	Annotated with a separate tag as an action attribute
NP events	Annotated
Predicative phrases	Annotated
Adjectival predicates	Annotated
Resultative nominalizations	If applicable annotated as participants
Pronominal actions	Annotated

Table 4.3: Overview of decisions made with regards to action annotation.

4.2.2 Times and entity mention annotation

This section elaborates on annotation of time, location and participant mentions in the ECB+ corpus.

For consistency of annotations, we provided the annotators with heuristics they could use to determine which component is mentioned in text. In the event that an annotator found it difficult to identify the appropriate annotation tag for a mention, we recommended that they apply the “substitution test”. We asked the annotators to rephrase a problematic excerpt without changing its meaning. For instance, if it is unclear how to annotate *Hollywood* in the sentence: *Hollywood is getting ready for this year’s Fourth of July BBQ*, one may replace *Hollywood* with a more prototypical location or human participant mention. For example, were one to replace *Hollywood* with *people from Hollywood* the sentence still expresses a similar meaning. It is thus possible to test whether the annotation tag of the equivalent phrase can be used for the original mention. Comparatively, if one were to substitute *Hollywood* with examples of location descriptions such as *in this location, here* or *in the mountains*, the resulting sentence is nonsensical and it is immediately obvious that location tags would be unsuitable.

Below we will look at guidelines with regard to the annotation of time, location and participant mentions in the ECB+.

4.2.2.1 Mention extent

In this section we will explain how the annotators were to determine the mention extent of times and entities described in text.
With regards to times and locations we annotated whole expressions, not only the head of a phrase as shown in EXAMPLES 4.2.31–4.2.37.

Example 4.2.31 *two years ago*

Example 4.2.32 *3 days later*

Example 4.2.33 *in July 1999*

Example 4.2.34 *Portland, Maine*

Example 4.2.35 *5 miles upstream*

Example 4.2.36 *in the capital of Turkmenistan*

Example 4.2.37 *in southern Iraq*

In the case of participants, we annotated only the head of a phrase. By "head" we mean either the pronoun or, for NPs, the nominal part of the NP that is not used as a modifier and that expresses the most specific meaning. For instance, in the case of the NP *the US soldiers* only *soldiers* should be marked as the head of the NP and in the case of *the deceased man*, *man* would be annotated as a human participant and *deceased* as an action. To convey this to the annotators we provided them with some heuristics to ensure consistent coding. For instance, we explained that when one leaves the modifiers out of a NP, the meaning of the phrase becomes more general. If, however, one leaves the head out, the meaning of the phrase changes. Compare:

- *health insurance treaties* vs. *treaties* (the modifiers left out, keeping the head)
- *health insurance treaties* vs. *health insurance* (the head left out).

Consider examples of participant mentions in sentences from 4.2.38–4.2.44.

Example 4.2.38 *Holland has health insurance treaties with a number of countries.*

Example 4.2.39 *Homer the poet*

Example 4.2.40 *The President of the U.S. Barack Obama*

Example 4.2.41 *Sri Lankan politics for several years witnessed a bitter struggle between the president and the Prime Minister.*

Example 4.2.42 *Some of the refugees*

Example 4.2.43 *A group of kids*

Example 4.2.44 *David Cameron, the Prime Minister of UK, said...*

Note that the head might consist of more than one word, in the case of proper names (e.g. *Barack Obama*). With exception of location and time mentions, we did not annotate whole NPs but only their heads and we did not annotate markables within the extent of a longer markable for instance a participant mention within the extent of a bigger participant mention (*U.S. Secretary of State John Kerry*). The participant type which corresponds to the annotation tag is always assigned to the head of a participant mention so, for instance, *the US soldiers* would get the entity type assigned to its head *soldiers*. We did not annotate *US* and its type.

4.2.2.2 Mention part of speech

In this section we give an overview of how times and entities can be described in language. We annotated locations and times expressed by proper names, common nouns (as part of NPs or PPs) and adverbs. Human and non-human participant entities can be expressed by proper names, common nouns (also in NPs or PPs) and pronouns. Below are some examples of times, locations and participant mentions expressed by different part of speech.

Times, locations and participants can be expressed by **a proper name** as the head of the phrase, also as part of a NP or PP. Consider EXAMPLES 4.2.45–4.2.47. Note that in the sentence from 4.2.45 *President* is the head of another person entity, though not one with a proper noun as head, hence not underscored here. And in 4.2.46 *in Warsaw* is a location hence the whole phrase was annotated; the typhoon mention is also a proper name but it refers to an action.

Example 4.2.45 *Barack H. Obama is the 44th President of the United States.*

Example 4.2.46 *UN climate talks in Warsaw darkened by Typhoon Haiyan.*

Example 4.2.47 *In September the debut album by Canadian singer-songwriter Hayden comes out.*

In sentences from 4.2.48–4.2.52 we see examples of times and participants expressed by **common nouns** used as the head of the phrase or used as part of a NP or PP.

Example 4.2.48 *The President of the United States ...*

Example 4.2.49 *All Commission seats and the post of general counsel to the commission are filled by the President of the U.S.*

Example 4.2.50 *The murdered family had stayed for a while in a house where people were previously murdered.*

Example 4.2.51 *This morning the Prime Minister announced she will re-nominate for Leader of the Federal Labor Party in a ballot next Monday morning.*

Example 4.2.52 *The introduction of the euro in 1999 was a major step in European integration.*

Note that in 4.2.50 *in a house* is a location hence the whole phrase was annotated. Next, **pronominal** participants are exemplified in 4.2.53.

Example 4.2.53 *Apple Inc. executive Scott Forstall was asked to leave the company after he refused to sign his name to a letter apologizing for shortcomings in Apple's new mapping service.*

Finally, consider examples of **adverbial** locations and times in 4.2.54–4.2.57.

Example 4.2.54 *The tugboat went 120 miles upstream in 20 hours.*

Example 4.2.55 *The people of Fika got up from Tchad and went east to Dala, and stayed there one year.*

Example 4.2.56 *Structural Heart Program was recently launched at Southcoast.*

Example 4.2.57 *The murdered family had stayed for a while in a house where people were previously murdered.*

Note that actions, times, locations or participants can occur in text as modifiers of heads of nominal phrases as in *Connecticut school shooting, the deceased men, Tuesday's meeting*. If modifiers refer to event components they were also annotated in ECB+ (see section 4.3).

4.2.2.3 Subtypes

We annotated mentions of participants and locations expanding on the ACE entity subtypes (Linguistic Data Consortium [2008]). We annotated time expressions following the types from the TIMEX3 specification (Pustejovsky et al. [2003]). In the following paragraphs we will discuss in detail the procedure for type annotation of time, location and participant mentions.

Times

The time component of events marks explicit time expressions describing when something happens or holds true. We annotated event times following the types from the TIMEX3 specification (Pustejovsky et al. [2003]). When annotating time expressions, the annotators were asked to specify one of the four subtypes: DATE, TIME, DURATION and SET (Pustejovsky et al. [2003]). Four tags were used to annotate times in ECB+: TIME_DATE, TIME_OF_THE_DAY, TIME_DURATION and TIME_REPETITION. Below we specify each tag used in ECB+ together with EXAMPLES from the TimeML specification (Saurí et al. [2005]).

- The TIME_DATE tag refers to calendar time, consider EXAMPLES 4.2.58–4.2.64.

  Example 4.2.58 *June 11, 1989*

  Example 4.2.59 *Yesterday*

  Example 4.2.60 *Summer, 2002*

  Example 4.2.61 *On Tuesday 18th*

  Example 4.2.62 *This summer*

  Example 4.2.63 *The second of December*

  Example 4.2.64 *Last week*

- The TIME_OF_THE_DAY tag corresponds to TimeML's TIME type of a TIMEX and captures expressions referring to a specific time of the day, as illustrated in EXAMPLES 4.2.65–4.2.71.

  Example 4.2.65 *Ten minutes to three*

Example 4.2.66 *At five to eight*

Example 4.2.67 *At twenty after twelve*

Example 4.2.68 *At 9 a.m. Friday, October 1, 1999*

Example 4.2.69 *The morning of January 31*

Example 4.2.70 *(late) Last night*

Example 4.2.71 *Between 8 a.m. and 10 a.m.*

- The TIME_DURATION tag is meant for time expressions denoting durations as exemplified in 4.2.72–4.2.77.

Example 4.2.72 *2 months*

Example 4.2.73 *48 hours*

Example 4.2.74 *Three weeks*

Example 4.2.75 *All last night*

Example 4.2.76 *20 days in July*

Example 4.2.77 *3 hours last Monday*

- The TIME_REPETITION tag corresponds to TimeML's SET (Saurí et al. [2005]) and is used for sets of times describing repeated events as shown in EXAMPLES 4.2.78–4.2.82.

Example 4.2.78 *Often*

Example 4.2.79 *Frequently*

Example 4.2.80 *Every Tuesday*

Example 4.2.81 *Twice a week*

Example 4.2.82 *Every 2 days*

Locations

The location component of events specifies in a sentence where something happens or holds true. We defined event locations in line with ACE's general PLACE attribute, corresponding to ACE's geo-political (GPE), location (LOC) or facility (FAC) entities referring to a physical location. We used three tags to annotate event locations in ECB+: LOC_GEO, LOC_FAC and LOC_OTHER. Below we specify each tag based on definitions from the ACE entity guidelines (Linguistic Data Consortium [2008]). We illustrate the usage of each tag with examples.

- The LOC_GEO tag corresponds to both, ACE's GPE that is geo-political entities i.e. "geographical regions defined by political and/or social groups referencing the territory or geographic position of the GPE" e.g. 4.2.83 as well as ACE's LOC - location entities that is "geographical entities defined on a geographical or astronomical basis such as geographical areas and landmasses, bodies of water, and geological formations", see EXAMPLES 4.2.84–4.2.90.

  Example 4.2.83 *Fighting in Bosnia and Herzegovina came to an end on 11 October 1995.*

  Example 4.2.84 *A 7.2 magnitude earthquake hit in Southern California this afternoon.*

  Example 4.2.85 *Trip around the world*

  Example 4.2.86 *Landing on the moon*

  Example 4.2.87 *On the Vistula river*

  Example 4.2.88 *In the Tatra mountains*

  Example 4.2.89 *In the city*

  Example 4.2.90 *We entered the airspace of Poland.*

- The LOC_FAC tag refers to facility entities i.e. to "buildings and other permanent manmade structures and real estate improvements" referencing where an action happened as illustrated in EXAMPLES 4.2.91–4.2.92.

  Example 4.2.91 *It is the deadliest mass murder in a school in United States history.*

  Example 4.2.92 *On the streets of Singapore*

- Additionally we defined a third location tag: LOC_OTHER for any remaining type of event locations encountered in text as exemplified in 4.2.93–4.2.94.

  Example 4.2.93 *After the Prime Minister sat down on a white wicker chair and greeted the Grade 4 children at St Joseph's primary school, they chorused en masse: "Good morning Prime Minister, may the angels watch over you."*

  Example 4.2.94 *The mall gunman may have been shooting at security cameras.*

Human participants

The participant component of events specifies in a sentence who or what is involved with, undergoes change as a result of or facilitates an event or a state. We defined human event participants similarly to ACE's event participants of entity type person (PER) and organization (ORG) but also metonymically used geo-political (GPE), facility (FAC) and vehicle (VEH) entities when referring to a population or a government (or its representatives). Crucial human participants of events reported in the news are often expressed as syntactic subjects or objects. In ECB+ the following tags were used for annotation of human participants: HUMAN_PART_PER, HUMAN_PART_ORG, HUMAN_PART_GPE, HUMAN_PART_FAC, HUMAN_PART_VEH, HUMAN_PART_MET and HUMAN_PART_GENERIC. Next we describe the tags used to mark human participant mentions accompanied by definitions of the corresponding entity types from the ACE entity guidelines (LDC [2005a], Linguistic Data Consortium [2008]).

- The HUMAN_PART_PER tag refers to person entities and is "limited to humans; it may be a single individual or a group" of individuals; see EXAMPLES 4.2.95–4.2.97 from the ACE entity guidelines (LDC [2005a]).

  Example 4.2.95 *The President of the U.S.*

  Example 4.2.96 *The President of the U.S. Barack Obama*

  Example 4.2.97 *The family.*

- The HUMAN_PART_ORG tag denotes organization entities "limited to corporations, agencies and other groups of people defined by an established organizational structure" as illustrated in EXAMPLES 4.2.98–4.2.99.

  Example 4.2.98 *Air Force helicopters provided air support as the Navy attacked four LTTE boats.*

  Example 4.2.99 *The Free University decided to create a presence in Second Life.*

- The HUMAN_PART_GPE tag is meant for geo-political entities that is "geographical regions defined by political and/or social groups" referring to a population or a government. This tag is also meant for city names used with reference to their inhabitants. Consider EXAMPLES 4.2.100–4.2.102.

  Example 4.2.100 *Poland and the US signed a $34 million deal to modernize the Polish Navy's missile frigate.*

  Example 4.2.101 *Hollywood is getting ready for this year's Fourth of July BBQ.*

  Example 4.2.102 *Boston won from Cleveland today in a short, decisive game that was uninteresting after the first innings.*

- The HUMAN_PART_FAC tag refers to facility entities i.e. “buildings and other permanent manmade structures and real estate improvements” referring to people using or managing them, as shown in EXAMPLE 4.2.103.

  Example 4.2.103 *The school decided to find a new location.*

  But not in EXAMPLE 4.2.104 where *school* is a non-human participant entity and not in 4.2.105 where *school* refers to a location of type facility.

  Example 4.2.104 *The school was totally destroyed.*

  Example 4.2.105 *The blood bath happened in a school.*

- The HUMAN_PART_VEH tag marks vehicle entities which are “physical devices primarily designed to move an object from one location to another”, used in reference to a population or a government usually occurring with GEO adjectives as shown in EXAMPLES 4.2.106–4.2.107.

  Example 4.2.106 *U.S. ships attacked 3 Iraqi patrol boats.*

  Example 4.2.107 *In 1991 Serbian tanks attacked Croatian cities.*

  But not in EXAMPLE 4.2.108 where *ship* is a non-human participant.

  Example 4.2.108 *Somali refugees arrive by ship.*

For the sake of completeness, next to the five subtypes described above we distinguish two additional human participant tags: HUMAN_PART_MET and HUMAN_PART_GENERIC.

- The HUMAN_PART_MET tag is meant for any remaining metonymically expressed human participants of events, see EXAMPLES 4.2.109–4.2.114.

  Example 4.2.109 *30% of households are living from paycheck to paycheck.*

  Example 4.2.110 *The press was present in large numbers and asked a great number of questions.*

  Example 4.2.111 *He has sworn loyalty to the flag.*

  Example 4.2.112 *The crown gave its approval.*

  Example 4.2.113 *That’s not what I’m hearing from the boots on the ground.*

  Example 4.2.114 *The brown shirts marched through the town.*

- HUMAN_PART_GENERIC applies to generic mentions referring to a class or a kind of human participants or their typical representative without pointing to any specific individual or individuals of a class (Linguistic Data Consortium [2008]), for instance generic *you* or *one* as event participants. Consider EXAMPLES 4.2.115–4.2.116.

  Example 4.2.115 *One should treat others as one would like to be treated.*

  Example 4.2.116 *17 year old female seeking employment, loves working with kids.*

Non-human participants

Next to locations, times and human participants, we recognize a fourth entity type - NON_HUMAN_PART which is meant for all remaining entity mentions that contribute to the meaning of an event action. These will often be artifacts expressed as a (direct or prepositional) object of a sentence or as PPs not in object position such as instrument phrases, see EXAMPLES 4.2.117–4.2.120. For example in 4.2.117 both *pencil* and *knife* should be annotated as NON_HUMAN_PART. Note that in 4.2.120 *Mondays* does not refer to the time of an event action but it is also a NON_HUMAN_PART mention.

Example 4.2.117 *sharpen a pencil with a knife*

Example 4.2.118 *Debbie traveled by boat 5 miles upstream to fish in her favorite spot.*

Example 4.2.119 *Samsung signed a deal to be the NBA's official provider of tablets and televisions.*

Example 4.2.120 *I hate Mondays.*

Within the NON_HUMAN_PART type for the sake of completeness we distinguished a special sub-tag: NON_HUMAN_PART_GENERIC for generic mentions referring to a class or a kind of non-human entities or their typical representative without pointing to any specific individual object or objects of a class (Linguistic Data Consortium [2008]). Consider EXAMPLE 4.2.121 where no specific instance of the set of cats is pinpointed.

Example 4.2.121 *Linda loves cats.*

4.2.2.4 Times and entity annotation summary

In TABLE 4.4 we outline the main decisions with regards to annotation of times and entities.

Language phenomenon	Treatment in ECB+
Time mention extent	Whole phrase annotated
Location mention extent	Whole phrase annotated
Participant mention extent	Head of the participant phrase annotated
Pronominal entities	Pronouns annotated
Times	Annotated with TIMEX3 types
Entities	Annotated with distinction of three types: LOC, HUMAN_PART and NON_HUMAN_PART (locations and human participants annotated with a modification of ACEs entity types)

Table 4.4: Overview of decisions made with regards to time and entity annotation.

4.2.3 Mention annotation summary

In section 4.2 we presented the ECB+ mention annotation guideline. The mention annotation scheme is framed by the event model that we employ in this work. We model events as a combination of four components: (1) event actions, (2) times, (3) locations and (4) human- and non-human participants. In the previous sections 4.2.1 and 4.2.2 we looked at three aspects of mention annotation per event component: (1) mention extent, (2) mention POS and (3) mention typology.

In this section we summarize the entire section 4.2. We finalize with an overview of annotation decisions made with regard to all five event components. We outline the ECB+ mention annotation guideline in TABLE 4.5. A stands for action, T for time, L for location. HP refers to human participants and NhP stands for non-human participants. The $\subset$ symbol indicates that token(s) with the POS can be part of a component mention or the other way around. For example adjectives *next* and *last* can be part of a time mention in *next/last Monday* and *southern* can be part of a location mention e.g. in *in southern Iraq*. On the other hand, any entity type: time, location, human- or non-human participant mention can be part of a predicative phrase e.g. *He was a marine* or *The meeting was on Monday.*

Mention	**Component in ECB+**				
Annotation	A	T	L	HP	NhP
Extent	Head (except for idioms and phrasal verbs)	Entire phrase	Entire phrase	Head	Head
POS Form:					
Verbal	+	-	-	-	-
Nominal	+	+	+	+	+
Adjectival	+	ADJ ⊂ T	ADJ ⊂ L	-	-
Predicative phrase (PR)	+	T⊂ PR	L⊂ PR	HP⊂ PR	NH ⊂ PR
Pronominal	+	-	-	+	+
Adverbial	-	+	+	-	-
Typology	(1)OCCURRENCE (2)PERCEPTION (3)REPORTING (4)ASPECTUAL (5)STATE (6)CAUSATIVE (7)GENERIC +7 tags negated	(1)DATE (2)TIME_OF_THE_DAY (3)DURATION (4)REPETITION	(1)GEO (2)FAC (3)OTHER	(1)PER (2)ORG (3)GPE (3)FAC (4)VEH (5)MET (6)GENERIC	(1)NON_HU-MAN_PART (2)NON_HUMAN_PART_GENERIC

Table 4.5: Summary of mention annotation decisions per event component with regard to: mention extent, mention POS form and mention typology.

4.3 “Event-centric” annotation

The ECB+ annotation specification was designed to be “event-centric”. Mentions of event components were annotated in text from the point of view of an event action, marking:

1. participants involved with an action as opposed to any participant mention occurring in a sentence
2. time when an action happened as opposed to any time expression mentioned in text
3. location in which the action was performed in contrast to a locational expression that does not refer to the place where an action happened.

For example *her father* in the sentence *Her father told ABC News he had no idea what exactly was going to happen* refers to the only human participant of the reporting action described in the sentence namely the father of the woman in question. The denotation of *her* does not refer to a participant of the reporting action hence we would leave *her* un-annotated. On the other hand *her* in the sentence *Her stay in rehab is over* does denote a human participant of action *stay*. Similarly *Mondays* in *I hate Mondays* does not refer to the time when the state holds true but in this sentence it should be annotated as a non-human participant.

Event-centric thinking was applied throughout the whole annotation effort and it guided the decision making process with regards to annotation of linguistic phenomena (such as whether to annotate possessive pronouns as human participants or not). It helped us with the identification of the number of location, time and participant markables per action in a sentence. This was especially useful with long component descriptions as in *ABC Entertainment Group prexy Paul Lee* which in ECB+ is annotated as a single human participant mention. The number of markables per action should correspond to the number of actual event participants, times and locations (a special case is the way in which we treat some of the subjects and subject complements in copular constructions, see the paragraph on predicative phrases, section 4.2.1).

In the following section we describe how the coreference relation was annotated among mentions of an event component.

4.4 Coreference annotation

If an event instance is described more than once in one or in multiple texts, we say that its descriptions are coreferent (see also section 2.6). Within the ECB+ annotation task we annotated both cross- and within-document coreference relations (whether anaphoric or not) between mentions of a particular instance of an event component. If an event component instance that is an action or its time, location or participant are described in one or multiple texts more than once, their descriptions were marked as coreferent. Consider EXAMPLES 4.4.1–4.4.2.

Example 4.4.1 *Lindsay Lohan checked into rehab.*

Example 4.4.2 *Ms. Lohan entered a rehab facility.*

These two sentences might refer to the same event instance, although as the human participant has been to rehab multiple times, it may also refer to two different instances. If one can determine based on the context that two event mentions refer to the same event instance, they should be annotated as coreferent. If not, the action mentions should not be made coreferent, but the human participant mentions from our example sentences should be marked as coreferent, as they refer to the same person. One would also need to determine whether *rehab* and *rehab facility* refer to the same facility or not and annotate accordingly.

Coreference relations were established through mentions of actions, times, locations, human participants and non-human participants. Coreference could never be assigned between an action and an entity mention. Coreference should not be assigned between mentions belonging to any two different component types for example between a location and a participant. Two or more time expressions, location or participant mentions corefer with each other if they refer respectively to the same time, place or participants. Two action mentions corefer if they refer to the same instance of an action i.e. an action that happens or holds true: (1) in the same time, (2) in the same place and (3) with the same participants involved. We annotated both, cross- and within-document coreference. Anaphoric coreference was annotated as well.

One often comes across copular constructions with verbs like *be, appear, feel, look, seem, remain, stay, become, end up, get* (copular verbs list taken from OntoNotes annotation guidelines, 2007), see EXAMPLE 4.4.3.

Example 4.4.3 *This boy is James.*

If the subject (*this boy* referring specifically to this particular boy and not any other) and its complement (*James*) both refer to the same entity instance in the world, which in this case is James, coreference between the two should be annotated. If however, the reference of the sentence subject and of the subject complement is not exactly the same, as in EXAMPLE 4.4.4, coreference should not be marked.

Example 4.4.4 *James is just a little boy.*

In EXAMPLE 4.4.4 *James* refers to a particular boy called *James* but the phrase *a little boy* is indefinite and might refer to any little boy in the world, not necessarily to James. The latter phrase refers to a generic property; to a whole set of little boys rather than to a single participant instance. James in this case is just one element of the whole set, hence the reference of the two phrases is not identical.

Both sentences contain predicative phrases, parts of which should be annotated as both: human participants and states. In the sentence from EXAMPLE 4.4.3 *James* should be annotated as human participant of type person and as an action of class state. In the sentence from EXAMPLE 4.4.4 *boy* should be annotated as human participant of type person and generic and as an action of class state.

4.4.1 Coreference annotation summary

TABLE 4.6 outlines how the coreference relation was annotated in ECB+.

In the previous sections (see section 4.2, 4.3 and 4.4) we discussed the ECB+ annotation guideline, designed to mark in text mentions of event components and coreference between them. In section 4.5 we describe the setup of the ECB+ annotation task.

Language phenomenon	Treatment in ECB+
Action anaphora	Annotated
Within document action coreference	Annotated
Cross document action coreference	Annotated
Entity anaphora	Annotated
Within document times and entity coreference	Annotated
Cross document times and entity coreference	Annotated
Coreference between subject and subject complement in copular constructions	Annotated if referring to the same entity

Table 4.6: Overview of decisions made with regards to coreference annotation.

4.5 Setup of the annotation task

The ECB+ corpus (Cybulska and Vossen [2014b]) was annotated in four annotation rounds:

1. Mentions of event instances were annotated in the newly created ECB+ corpus component and within-document coreference relations were established.
2. Modifications were made to the ECB 0.1 annotation (Lee et al. [2012], Recasens [2011]) of the EventCorefBank (Bejan and Harabagiu [2010])
3. Cross-document coreference relations were established for each topic; the new topic scope was extended to include both the ECB texts and the newly added ECB+ texts.
4. Annotations were reviewed and mistakes were corrected. 1840 sentences with highest quality annotations were selected.

Step 1

Two student assistants, Elisa Wubs and Melissa Dabbs, were hired for a period of four months to perform the annotation. They were paid for their work. Both of them are native speakers of English pursuing a degree at VU University Amsterdam (one of them was an exchange student from the UK, both are British nationals). After training the annotators we moved on to the first stage of the annotation.

Firstly, a newly created ECB+ corpus component of 502 news articles was annotated. The annotators were given the task to annotate mentions of event instances together with mentions of their participants, times and locations and within-document coreference between them in the new ECB+ corpus component. The first topic of the new ECB+ component was annotated as warm up by both annotators. The next three topics were also annotated by both annotators. In 55 texts from the first four topics of the ECB+ we pre-selected sentences describing the seminal events. Those sentences were annotated by both annotators and they were used for the calculation of the inter-annotator agreement (see section 4.7). The remainder of the corpus – 447 texts – was divided between the two student assistants and annotated once. The annotators had to chose the sentences describing seminal events and then annotate them.

Step 2

In the second stage of the annotation process, adjustments were made to the 480 texts of the ECB 0.1 annotation (Lee et al. [2012], Recasens [2011]) of the EventCorefBank (Bejan and Harabagiu [2010]) to ensure compatibility of annotations of both corpus components. Each annotator worked on half of the data. There is one major difference between the annotation style of the ECB and that of the new corpus component. In the ECB+ annotation scheme, we make an explicit distinction between action classes and between a number of entity types. We re-annotated the ECB 0.1 texts so that we not only have event actions and entities annotated (ECB 0.1. distinguishes between two tags: ACTION and ENTITY), but can also know precisely whether an entity is a location, time expression or participant. The same applies to action mentions that were re-annotated with specific action classes.

Wherever necessary, adjustments were made with regards to mention extent. For human and non-human participant entities annotated in the ECB 0.1 corpus, we made sure that only the head of a mention was explicitly annotated. With regards to times and locations, we made sure that the whole phrase was marked. Regarding action annotation, wherever necessary we additionally annotated light verbs and adjectival predicates. Finally, adjustments were made to ensure that annotation of the ECB is compatible with the event-centric annotation of the new corpus component.

The re-annotation efforts were focused on sentences that were selected during the annotation of ECB 0.1. This allowed us to speed up the re-annotation process significantly. In principle, we adopted the within-document coreference relations established in ECB 0.1 but wherever necessary we added new chains or adjusted the existing ones.

The within-document annotation in the first two stages of the ECB+ annotation process was performed by means of the CAT - Content Annotation Tool (Bartalesi Lenzi et al. [2012])[3] which we used to annotate mentions of actions, times, locations and participants in text as well as within-document coreference relations between them.

Step 3

The third step in the ECB+ annotation process was to establish cross-document coreference relations between mentions of actions, times, locations and participants from the same topic. Wherever applicable, coreference links were created across both the ECB texts and texts of the newly added ECB+ component of a topic.

Step 4

Finally, we reviewed the annotations. We selected 1840 sentences from the whole corpus that described the seminal events most clearly and that were annotated most extensively by the annotators. In those sentences, missing coreference links and missing mention annotations were added or corrected following the annotation guideline. We discarded multiple sentences for which annotations were incomplete (mainly missing coreference links). The final review was done to increase the quality of the ECB+ annotation.

In the last two steps of the annotation task for marking of cross-document coreference relations and for the review of the annotations, we used a tool called CROMER (CRoss-document Main Event and entity Recognition, Girardi et al. [2014]).

[3] The CAT tool has been previously known as the CELCT Annotation Tool (`http://www.celct.it/projects/CAT.php`).

ECB+	**1840 reviewed sentences**	**All annotations**
No. topics	43	
No. texts	982	
No. annotated action mentions	6833	14884
No. annotated location mentions	1173	2255
No. annotated time mentions	1093	2392
No. annotated human part. mentions	4615	9577
No. annotated non human part. mentions	1408	2963
No. cross-document chains	1958	2204

Table 4.7: ECB+ statistics; including the re-annotated ECB corpus.

CROMER is a Newsreader project (`http://www.newsreader-project.eu/`) extension of a multi-user web interface (Bentivogli et al. [2008]) designed within the Ontotext project (`http://ontotext.fbk.eu/`).

4.6 ECB+ statistics

TABLE 4.7 lists some basic statistics with regards to the newly annotated resource.[4] The ECB+ corpus contains a total of 982 texts that belong to 43 topics following the composition of the ECB corpus. We present the ECB+ annotation statistics after the first three steps of the annotation task were completed, in comparison to statistics after the final review of annotations. In the 1840 reviewed sentences, missing markables or coreference links were added while sentences with missing annotations were discarded. This resulted in a lower amount of annotated mentions in the reviewed sentences. However, there is not such a big difference in the amount of annotated coreference chains. The discarded sentences contained many missing links with regards to coreference annotation as well as incorrectly annotated mentions. In TABLE 4.7 the column with “All annotations” indicates all mention and coreference statistics in ECB+ that contain errors. The column with “1840 reviewed sentences” presents statistics with regard to the reviewed annotations in 1840 sentences. In the reviewed selection of sentences, 6833 action mentions were annotated, 1173 location and 1093 time mentions. Furthermore, 4615 human participant and 1408 non-human participant mentions were marked as well as 1958 cross-document coreference chains. After the final review of annotations, we ended up with an average of ca. 1.8 sentences annotated per document. This is a limitation of the corpus that must be considered when experimenting with ECB+.

The ECB+ corpus is available for download from `http://www.newsreader-project.eu/results/data/the-ecb-corpus/` or `https://github.com/cltl/ecbPlus`.

[4] Note that the coreference chain statistics consider some singleton chains created by the annotators. A total of 28 mentions were mistakenly left tagged with the general annotation tags (ACTION or ENTITY) used in ECB 0.1. These 28 mentions are excluded from mention amounts reported here.

4.7 Inter-annotator agreement

We calculated the inter-annotator agreement scores on topics 1-4 of the new ECB+ corpus component which contains 55 texts.

4.7.1 Mention annotation

We first measured how much agreement there is on the assignment of event component tags per token of a mention. For the purpose of this calculation, a number of sentences describing the seminal events of the first four topics was preselected. Both annotators were asked to annotate the same sentences in all 55 texts of the four topics. To measure the inter-annotator agreement between the annotators we used Cohen's Kappa (Cohen [1960]), a measurement which considers chance agreement. We calculated Cohen's Kappa when distinguishing all 30 annotation tags and also when looking at the main components that is when grouping the specific tags into the five categories: ACTION, LOC, TIME, HUMAN_PARTICIPANT and NON_HUMAN_PARTICIPANT. On the first four topics our two coders reached Cohen's Kappa of 0.74 when assigning all 30 tags. This score can be interpreted as representing substantial agreement (Landis and Koch [1977]). The inter-annotator agreement on the five main event component tags reached 0.79 Cohen's Kappa which also indicates agreement level substantial, although note that in these calculations untagged tokens were considered (for which we automatically assigned the tag UNTAGGED). When disregarding tokens not tagged by any of the annotators, and so only considering tokens tagged by at least one person (5581 out of 10189), Cohen's Kappa of 0.63 was reached on the 30 tag tag set and of 0.68 on the assignment of the main group tags; i.e. substantial agreement. The confusion matrix in TABLE 4.8 shows the distribution of the five main tags in the four topics of the corpus component as coded by the annotators.

C-1	**C-2**										
	A	T	L	H	N	U	A/H	A/L	A/T	H/L	L/N
A	1371	2	6	9	21	95	0	1	0	0	0
T	14	929	1	0	0	94	0	0	3	0	0
L	11	0	646	8	13	55	0	3	0	0	0
H	15	0	9	1118	13	60	2	0	0	0	0
N	16	0	2	2	92	28	0	0	0	0	0
U	447	82	118	196	94	4608	0	0	0	0	0
A/H	1	0	0	0	0	0	0	0	0	0	0
A/L	0	0	0	0	0	0	0	1	0	0	0
A/T	0	0	0	0	0	0	0	0	0	0	0
H/L	0	0	1	0	0	0	0	0	0	0	0
L/N	0	0	0	0	2	0	0	0	0	0	0

Table 4.8: Confusion matrix ECB+ topics 1-4; five component annotation by two coders: C-1 and C-2. A stands for ACTION, T for TIME, L for LOCATION. H stands for HUMAN_PART, N for NON_HUMAN_PART and U for UNTAGGED.

An analysis of the confusion matrix revealed that the annotators had less consensus with the definition of mention extents, annotating whole mention phrases while the guideline specified to only annotate the head or the other way around.

4.7.2 Coreference annotation

We measured how much agreement there is on the assignment of cross-document coreference relation between mentions of an event component. For annotation of cross-document coreference we used CROMER, a tool in which annotators first need to create "instances" (that were assigned human friendly names e.g. *barack_obama*) which uniquely represent collections of corefering mentions (e.g. *Barack Obama, the president of the USA, Obama*). Coreferent mentions from text are linked to one particular instance in CROMER. The set of CROMER instances is shared by annotators of a particular task. We asked our two annotators to establish coreference relations for topics 1-4 of the new ECB+ corpus component. We asked coder A to first work on topics 1 and 2 and coder B to annotate topics 3 and 4. Then coder B was asked to familiarize herself with instances created for topics 1 and 2 (no access to annotations of coder A was possible, only the instances are shared) and then to establish coreference links re-using the instances created for topics 1 and 2 by coder A. A similar procedure was applied for second coder annotation of topics 3 and 4. Because of the CROMER setup, it is clear what the intended instance (referent / denotation) of a coreference chain is. Hence we simply used Cohen's Kappa (Cohen [1960]) to calculate agreement on assignment of the coreference relation. We considered all tokens annotated at least by one annotator as the total number of annotated items. On the first four topics (490 cross document coreference IDs) our two coders reached Cohen's kappa of 0.76 which indicates substantial agreement.

4.8 The ECB+ contribution

In section 4.8 we evaluate the added value of the ECB+ corpus in the context of the coreference research. We examine whether the ECB+ compensates for some of the deficiencies of the ECB 0.1 corpus. In section 4.8.1 we compare the statistical information about the ECB corpus, the ECB 0.1 and ECB+. Next, in section 4.8.2 we will measure the average referential and the lexical diversity of the new resource. Finally, in section 4.8.3 we show how the new resource is used by researchers from the field. We conclude in section 4.9.

4.8.1 Comparison of statistical information

TABLE 4.9 shows the statistical information about the ECB+ corpus, in comparison to the original ECB corpus and the ECB 0.1 annotations.

The ECB+ corpus contains almost double the amount of texts in which many more events were annotated than in the ECB or ECB 0.1. In ECB+ there are 6833 action mentions marked, in the ECB 1744 and in ECB 0.1 2533 mentions. In the ECB+ four entity types were annotated: locations, times, human and non-human participants; in total 8289 (times and) entity mentions were annotated. In the ECB no entities were marked while the ECB 0.1 contains 5447 entity mentions. This is a significant increase in the amount of annotated data for both actions and entities. Furthermore, in the ECB+ many more cross-document coreference chains have been established - 1958 compared to 208 chains in the ECB and 774 (of cross- and within-document) chains in the ECB 0.1.

Moreover, the new texts that were added to the ECB+ corpus describe a second layer of event instances to the ones covered in the ECB data set. This should raise the

		ECB+		**ECB**	**ECB 0.1**
No. topics		43			
No. texts		982		482	
No. annotated action mentions		6833		1744	2533
No. annotated entity mentions	locations	1173	8289 entity mentions, subtypes marked	None	5447 entity mentions, no subtypes marked
	times	1093			
	human participants	4615			
	non-human participants	1408			
No. event coreference chains	within-document	1958		1302	774
	cross-document			208	

Table 4.9: ECB+ statistics in comparison to ECB and ECB 0.1.

average referential diversity of the corpus and potentially also the lexical diversity. In the next section 4.8.2 we will test whether this is the case.

4.8.2 Analyzing lexical and referential diversity of the ECB+

In chapter 3 we examined the ECB 0.1 corpus with regard to referential and lexical diversity of event coreference chains. We calculated that the average referential diversity of the ECB 0.1 corpus is low and amounts to 17%. The ALD of the ECB 0.1 corpus is 18%, which is similarly low. These are low scores that quantify the simplicity of the data set with regard to diversity of event coreference coverage. Due to the poor diversity of coreference resources, event coreference resolvers are not challenged to reason about events with their entities. Coreference can be solved with good scores based on lemma matches of actions. When evaluating event coreference on data sets that lack diversity, it remains unclear whether the tested systems could solve event coreference between events from news articles.

In this section we evaluate whether extending the corpus with a new layer of event instances into ECB+ has increased the average referential and lexical diversity of the coreference resource.

Measuring the increase of the average referential diversity

We calculate the average referential diversity – ARD – as introduced in section 3.3.1 to evaluate the contribution of the newly created ECB+ resource.

We measure the ARD following formula 3.1. To calculate the ARD of the ECB+ corpus we need to know the number of event types and the amount of instances covered by the corpus. We used here the same methodology as the one applied to calculate event type numbers for the ECB 0.1 corpus in section 3.4.1. We calculated the number of event types as represented by WordNet synsets. There are 1481 unique event types covered in the ECB+ corpus. This number indicates synsets unique in the whole corpus. Then we counted event instances as annotated in the ECB+. There are 2741 event instances in the corpus. Following our ARD formula we calculate the average referential diversity of the ECB+ as $|1 - (1481 / 2741)|$ which gives us an ARD of 46% for synsets unique in the whole corpus.

As calculated in section 3.4.1 the ARD of the ECB 0.1 corpus is 17%. This means that our extension of the ECB into ECB+ has increased the ARD of the corpus with

29 points. This increased ARD score reflects the fact that most seminal events are now represented by at least two event instances.

Measuring the increase of the average lexical diversity

Next we will calculate the average lexical diversity (ALD) of the ECB+ corpus to evaluate the contribution of the newly created ECB+ resource.

We define the ALD in section 3.3.2. We use formula 3.2 to calculate the average lexical diversity of the ECB+ corpus. We calculate the average lexical diversity per event instance annotated in the corpus, based on a number of unique lemmas L from mentions M describing an event instance. The formula considers the total number of event instances I annotated in the corpus. To calculate the ALD of the ECB+ corpus we use the same methodology like the one applied to calculate the ALD of the ECB 0.1 corpus in section 3.4.2.

Following our ALD formula we calculate the average lexical diversity of the ECB+ to have the value of 53%. As calculated in section 3.4.2 the ALD of the ECB 0.1 corpus is 18%. The calculations for both corpora disregard singleton instances. These ALD estimates show that transforming the ECB 0.1 into the ECB+ has increased the ALD of the corpus with 35 points. This is a very good result considering that the ECB+ extension was primarily aimed at increasing the referential diversity of the corpus.

4.8.3 ECB+ contribution to the field

The ECB+ corpus has been made freely available for research in 2014. The corpus can be downloaded from `http://www.newsreader-project.eu/results/data/the-ecb-corpus/` or `https://github.com/cltl/ecbPlus`.

The data set has been broadly used in the event coreference community (Wang [2015]) in numerous studies of events and event relations. Since the release of the ECB+, the corpus has received a great deal of scholarly attention, see Cybulska and Vossen [2015a], Cybulska and Vossen [2015b], Friedrich et al. [2015], Yang et al. [2015], Caselli and Vossen [2016], Krause et al. [2016], Minard et al. [2016], O'Gorman et al. [2016], Orasmaa [2016], Postma et al. [2016], Rospocher et al. [2016], Sprugnoli and Tonelli [2016], Vossen et al. [2016], Choubey and Huang [2017], Ferraro [2017], Ribeiro et al. [2017], Segers et al. [2017], Araki et al. [2018], Liu et al. [2018], Lu and Ng [2018], Vossen et al. [2018a], Vossen et al. [2018b], Bugert et al. [2020], Yu et al. [2020] amongst others. This is indicative of the high regard to which this corpus is held by the academic community. Yu et al. [2020] recognize ECB+ as "the largest and most popular dataset for cross-document Event Coreference".

4.9 Conclusion

In chapter 3 we defined two metrics to evaluate a coreference resource: the average lexical and the average referential diversity. We analyzed an existing data set, the ECB corpus, that was frequently used in event coreference studies (see section 3.4). Our experiments showed that, from the point of view of event coreference, the corpus is not very referentially or lexically diverse. To increase the referential diversity of the data set, we augmented it, creating a new ECB+ resource.

In chapter 4 we extended the original ECB corpus with a new corpus component, creating a new ECB+ resource that captures descriptions of double seminal events per

	ECB 0.1	**ECB+**
ARD	17%	46%
ALD	18%	53%

Table 4.10: Evaluation of ECB+ in comparison to ECB 0.1

topic. We did this to increase the referential diversity of the corpus so that it becomes more representative of news available on the web. The corpus was annotated with event classes and with specific types of entities and times as well as with cross- and within-document coreference between them. The coders reached substantial agreement on both mention and conference annotation. We made this newly created resource freely available for research. The ECB+ can be used to develop and test approaches to event extraction and event coreference resolution.

In section 4.8 we measured the average referential diversity and the average lexical diversity of the ECB+ corpus. We compare the ARD and the ALD scores of the ECB+ corpus with the diversity estimates calculated for the ECB 0.1 corpus in chapter 3. TABLE 4.10 shows a significant difference between the ARD and the ALD estimates for the ECB 0.1 and the ECB+. The average referential and lexical diversity have both increased significantly from the ECB 0.1 to the ECB+. The complexity of the data set has increased. It is interesting to see that increasing the ARD of the corpus by adding the coverage of the second event instance per topic has also caused a growth in the ALD. Having a greater referential diversity, the ECB+ is also more diverse lexically. The lexical diversity of the data set increased as a byproduct of the referential extension. We hypothesize that the ALD scores could be higher if the corpus was augmented with the objective to increase its lexical diversity. For example, if one added a new layer (or multiple) of seminal events to the data set whereas the new event descriptions were searched for through synonyms of the event descriptions already covered by the corpus.

By increasing the diversity of the coreference resource, we made it more representative of the population of news articles on the web where one can find descriptions of multiple event instances from an event type. Training and testing coreference resolvers on the ECB+ makes the task of coreference resolution more complex. A system has to distinguish between at least two event instances from an event type. In the ECB, grouping events into coreference chains in most cases would come down to distinguishing between different event types, e.g. between an arrest of a suspect or an earthquake. If working with the ECB+ corpus, a more fine-grained distinction must be made e.g. between arrests of two different suspects (ECB+ topic 35) or between two earthquakes that happened in the same country but at a different time (ECB+ topic 37).

The ECB+ corpus covers at least two event instances per topic. This is an improvement compared to the ECB, but still far from the multitude of event instances described in daily news. Per event type one would want to at least cover descriptions of event instances that differ with regard to every event component to ensure that coreference resolvers learn to distinguish between event instances that happened at different times or places or with different participants involved. The ECB+ is not as diverse referentially as one could wish for. The diversity of event instances in descriptions online could be much higher than two instances per event type. See section 3.3.1 where we discussed the example of hundreds of earthquakes that can happen monthly according to the USGS. Ideally, for coreference experiments one would like to use a corpus that covers multiple layers of event instances from multiple event types. A diachronic corpus across different topics would be ideal. A coreference corpus covering at least more than one instance per event type seems like the absolute minimum for coreference experiments to give meaningful results. The ECB+ is a step in the right direction and could be a starting point for future corpus extensions.

Part III

How can we model the gradable event coreference phenomenon?

In the second part of this dissertation we presented the ECB+, a new resource annotated with event coreference. In the third part, we elaborate on a gradable model of event coreference. This part of the dissertation is dedicated to the following five research questions.

Research questions

- Is there a correlation between the temporal perspective of the writer and the variation in language use?
- Does the time of writing correlate with event granularity?
- How can we model the relationship between granularity and event coreference?
- Can granularity of event times, locations and participants be automatically determined?
- How can we capture and formalize the interplay between different semantic relations and event coreference?

In chapter 5 we analyze how event mentions are realized in different types of text (section 5.1) and what the implications are for modelling events and event relations (section 5.2). We describe our experiments with the Dutch Srebrenica corpus and show how granularity of events described in text, correlates with the temporal perspective of the writer in section 5.3. In section 5.4 we introduce a granularity taxonomy that can be used to determine event granularity. In chapter 6 we formalize a gradable model of event coreference.

Chapter 5

Modeling events with an eye on granularity in the context of event coreference[1]

In this chapter, we research how descriptions of events are realized in different types of text and what the implications are for modeling the event information. There can be many reasons why people may describe similar events in different ways. Here we focus on the temporal factor.

We hypothesize that different temporal perspectives of writers correlate with variation in language use and correspond to some degree with genre distinction. To capture differences between event representations in diverse text types and thus to identify relations between events, we define an event model. We observe clear relations between particular parts of event descriptions - event times, locations and human participants. Texts, written shortly after an event happened, use more specific and uniquely occurring event descriptions than texts describing the same event but written long after the same event transpired. We perform statistical corpus research to confirm this hypothesis. Granularity of event times, locations and human participants seems to play a role in determining event relations. We therefore create a granularity taxonomy that makes it possible to automatically determine granularity of event times, locations and human participants. The ability to automatically determine relations between events and their sub-events over textual data, based on the relations between event times, locations and human participants, has important repercussions for modeling events in the context of coreference resolution. Our use case in this chapter focuses on events described in historical archives.

5.1 Introduction: Event descriptions in texts written with different temporal perspectives

In this section, we report on our research on how descriptions of events are realized in different types of text. We focus on the Srebrenica Massacre, which is a recent event

[1] The contents of this chapter have been published by Cybulska and Vossen [2010] and Cybulska and Vossen [2015b]

with a significant impact in the Netherlands.

Historical archives usually contain a mixture of news articles and historical documents. News articles are written shortly after an event happened; that is, they have a non-existent or shorter temporal perspective. Historical documents are written with a longer temporal perspective. In those two kinds of texts the same events are presented in diverse ways, as exemplified by EXAMPLES 5.1.1–5.1.4.

Example 5.1.1
In de brandende hitte verlieten donderdag meer dan honderd vrachtwagens en bussen volgepakt met vluchtelingen de enclave vanuit de Nederlandse VN-basis Potocari. Een vrouw en een kind kwamen te overlijden tijdens de tocht, aldus de VN. Mannen en jongens van boven de zestien werden uit de mensenmassa gepikt en weggevoerd met onbekende bestemming. Een aantal is naar Bratunac afgevoerd, een stadje in Bosnisch-Servisch gebied ten noorden van de enclave. De Bosnische Serviers willen onderzoeken of zij zich hebben schuldig gemaakt aan 'oorlogsmisdaden'.

News article fragment about the Srebrenica massacre from Volkskrant published on 14 July 1995, immediately after the massacre.

Example 5.1.2
Op 11 juli 1995 forceerden Servische troepen onder bevel van generaal Ratko Mladic zich met tanks de stad binnen en deporteerden en vermoordden ca. 8.000 moslimmannen en -jongens. Het waren Nederlandse troepen van Dutchbat die op dat moment de enclave theoretisch hadden moeten beschermen. Bij voorbaat was echter al bekend dat dit in de praktijk onmogelijk was. Deze actie, die in Nederland bekend staat als het drama van Srebrenica wordt gezien als de ergste daad van genocide in Europa sedert de Tweede Wereldoorlog.

From a Dutch Wikipedia entry: "Het drama van Srebrenica" from November 2009.

Example 5.1.3
On Thursday in the burning heat more than a hundred trucks and busses packed with refugees left the enclave from the Dutch UN base Potocari. A woman and a child passed away during the trip, according to the UN. Men and boys over the age of 16 were separated from the crowd and taken away to an unknown destination. Some of them were transported to Bratunac, a city in Bosnian Serb area to the north of the enclave. The Bosnian Serbs want to investigate if they were guilty of any war crimes.

English translation of a Dutch news article fragment about the Srebrenica massacre, published in Volkskrant on 14 July 1995, immediately after the massacre.

Example 5.1.4
On 11 July 1995 Serb troops under the command of General Ratko Mladic invaded the city with tanks and deported and murdered approximately 8,000 Muslim men and boys. At this time the Dutch troops known as Dutchbat were theoretically supposed to protect the enclave. Actually it was rather clear in advance that in practice it would not be possible. This event, known in the Netherlands as the Srebrenica massacre is seen as the worst act of genocide in Europe since the Second World War.

English translation of a fragment from the Dutch Wikipedia entry: "Het drama van Srebrenica" from November 2009.

Based on these short text examples, it is evident that the Srebrenica Massacre is described in much more detail in the news fragment from 5.1.1 than in the Wikipedia entry from EXAMPLE 5.1.2. In EXAMPLE 5.1.1 some events are presented at a lower granularity level, such as the separation of Muslim boys and men and the fact that they were taken away to a location that is very specifically described by the author (for an introduction to the notion of granularity see section 2.3). These fine granularity events together with other sub-events are part of the coarse granularity event of the Srebrenica genocide. The journalist writing the article at the time did not know that what was happening would eventually be recognized as an act of genocide, so he could not describe the events happening as such. The writers of the Wikipedia entry from EXAMPLE 5.1.2, on the other hand, knew already that the Muslim men were taken away to be murdered. Having a longer temporal perspective on the event they were able to provide more background information and explanation on the event. The authors of the Wikipedia article endeavored to present the event within a broader historical perspective. Sub-events, like the death of a particular woman and child are not included.[2]

For accurate information retrieval across many documents it is crucial to map the different event representations with each other in a uniform way, regardless of how they are expressed in different text genres, allowing for full recall of information related to the same event. Preliminary research was therefore performed on the hypothetical correlation between the temporal perspective and language use. In our study we focused primarily on the differences between two kinds of texts: news articles lacking or having a shorter temporal perspective and texts that are written with a longer temporal perspective, e.g. Wikipedia articles.

5.2 Modelling event mentions in different text genres

We hypothesize that the difference in the temporal perspective of an author corresponds with the diversity in language use. Texts lacking or having a shorter temporal perspective tend to report on fine granularity events while texts written from a longer temporal perspective tend to look for the bigger picture. Texts having none or a shorter temporal perspective tend to present fine granularity events in a relatively neutral way with less explanation of events. Texts having a longer temporal perspective describe coarse granularity events and tend to give more background information, they focus on an interpretation of what happened. Over time with the increasing perspective on an event, the degree of granularity of event descriptions as well as attitude and causality in text become higher. The account of what happened becomes subjective, reflecting writers' interpretation of events.

Furthermore, the different temporal perspective of a writer corresponds with genre distinction. Authors having a shorter temporal perspective or none at all on the described events write e.g. news articles. Descriptions of events from a longer temporal perspective appear in other types of historical or journalistic writing such as educational texts e.g. Wikipedia entries or newspaper articles that are not part of the daily news.

In this work we mainly focused on investigating the correlation between the temporal perspective of the writer and the different degrees of granularity of event descriptions. We leave an investigation of the correlation between temporal perspectives and

[2] Some sub-events (typically the most remarkable ones) can also be found in some texts written from a historical perspective but are much less frequent than in the news and only in the context of a high-level event of which they are a part of.

subjective perspectives as well as temporal perspectives and genre distinction for future research.

Event granularity hypothesis

Our hypothesis is that the closer to the event time a text was written, the more specific and concrete the event descriptions tend to be. The greater the temporal distance from an event the broader the historical perspective on an event and the more general, abstract (and subjective) the way of event presentation. Texts lacking or having a shorter temporal perspective give a detailed account of events at the lower granularity level while texts having a longer temporal perspective on the described events abstract from details and describe events at the higher granularity level using more general and abstract vocabulary but focus more on the explanation of events. With increasing temporal perspective:

1. described event participants change from individuals to group participants
2. event locations change from small to bigger areas
3. event times change from short to longer periods of time.

This tendency could stem from insufficient information. Events described during or immediately after their occurrence may not yet be understood within a broader historical context in the same way that events described far after their occurrence are. At the time of writing not everything is known yet for example the reason why things happen or who is behind what is happening might be unclear. With time passing by, writers have more knowledge about an event and they express this information in their texts, while leaving out descriptive details. Obviously, this distinction is not black-and-white. The change from a detailed to a more general account of events is gradual. Nonetheless, there does seem to be a reliable tendency.

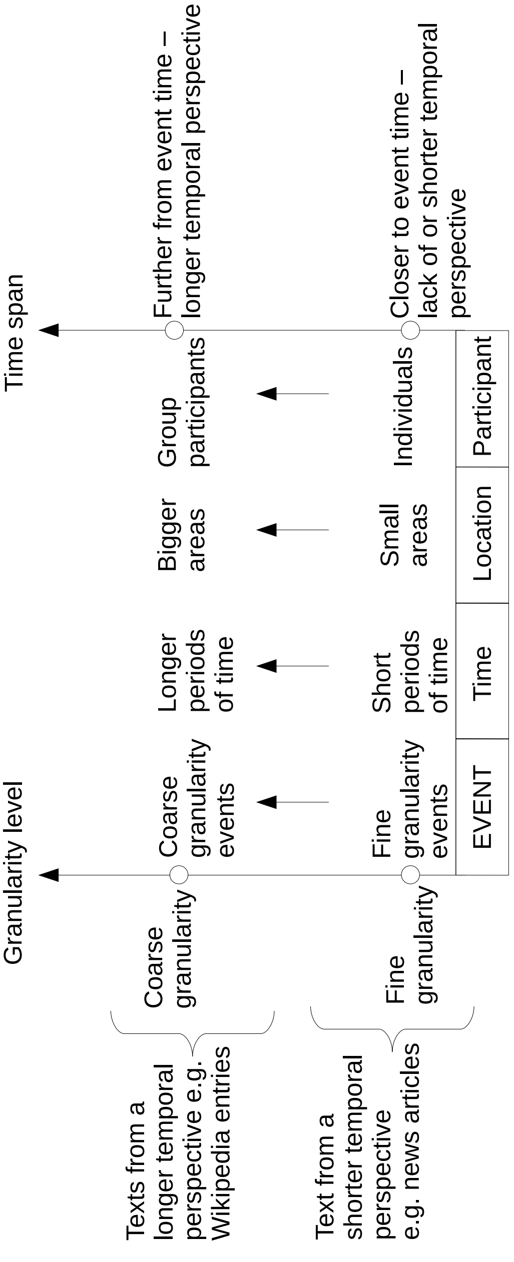

Granularity level
Time span
Coarse granularity
Coarse granularity events
Longer periods of time
Bigger areas
Group participants
Further from event time – longer temporal perspective
Fine granularity
Fine granularity events
Short periods of time
Small areas
Individuals
Closer to event time – lack of or shorter temporal perspective
EVENT
Time
Location
Participant
Texts from a longer temporal perspective e.g. Wikipedia entries
Text from a shorter temporal perspective e.g. news articles

Figure 5.1: Modelling granularity of event descriptions over time.

To capture the gradual change in language use and the differences and relations between events presented in diverse text genres over time, we use an event model that was introduced in section 2.2. The event model consists of four components: an action, a time, location and a participant component (see TABLE 5.1). It is compatible with standard approaches to event modeling. For a comprehensive overview of the history of event modeling in linguistic theory, see Tenny and Pustejovsky [2000]. For resources implementing event models, see FrameNet (Baker et al. [2003]) SIMPLE, BSO (Pustejovsky et al. [2006]) SUMO (Niles and Pease [2001], Niles and Pease [2003], Niles and Terry [2004]) and DOLCE (Masolo et al. [2003]). In this chapter we focus on the three components of the event model that refer to the event context: time, location and human participant. Non-human participants are out of the scope of this research.

Event Component	**Values with coarse vs. fine granularity**
Action	*The Srebrenica Massacre >shooting*
Time	*In the spring of 1995 >today*
Location	*Bosnia >Srebrenica*
Human participant	*UN soldiers >Dutchbat Colonel Karremans*

Table 5.1: Event model with example values disjoint or related through meronymy relation: member-group or part-whole

Event descriptions in text might be at the same granularity level or they can vary with regard to the degree of granularity of the fillers for slots from the event model. Coarse versus fine granularity times, locations and participants from text can be related to each other or not. The significant relation between fillers for slots from the event model in the context of granularity are those on the meronymy axis, member vs. group e.g. *Colonel Karremans* being a member of the group of *Dutch UN soldiers* or part vs. whole relation such as *Srebrenica* being a part of *Bosnia*. Furthermore, in addition to the time of an event we also have to consider another layer of time which is the time of text production that plays a crucial role in our model since it influences the temporal perspective which has a critical impact on granularity of fillers for slots in the model.

The event model after the addition of the time of writing is presented in FIGURE 5.1. The model illustrates the event granularity hypothesis and it captures predictions with regard to granularity change of times, locations and participants that we will test in the next section. FIGURE 5.1 illustrates the two ways in which the relative temporal location of an event impacts its textual description. The first way is that event descriptions become less detailed, which has a predictable consequence for the increasing granularity of times, locations and participants. The second way is in the genre. That is, a text which provides a description of an event soon after that event transpired, or even during its occurrence, is typically attested in news articles. Comparatively, a text which provides a description of event relatively long after it transpired is typically not attested in news articles, but rather in historical analyses or reviews.

5.3 Experiments with the Dutch Srebrenica corpus

5.3.1 Corpus composition

To test the event granularity hypothesis based on the model from FIGURE 5.1 in section 5.2 we created a “Srebrenica corpus” which consists of Dutch texts on the Srebrenica massacre in July 1995. The corpus consists of three components, see TABLE

5.2. The first component contains 26 online texts written about the Srebrenica Genocide from a historical perspective (published some years after the event). The "historical" component will be compared against two news components containing news articles from two Dutch newspapers. The first news component consists of 26 articles from *Volkskrant* and the second one contains 26 articles from *Het Parool*.[3] Srebrenica was captured on July 11, 1995. All news articles included in the corpus were published between July 7 and 17, 1995; shortly before or after the incident.

Corpus component	**No. tokens**	**No. texts**
Volkskrant	16573	26
Texts with historical perspective	20742	26
Het Parool	13481	26

Table 5.2: Structure of the Srebrenica corpus

The news corpus components contain articles representing different newspaper genres. Most are news articles but some belong to other journalistic genres such as profile article, chronology, analysis, opinion, commentary or report.[4] The main condition for the inclusion into this component of the corpus was the short time period between the act of writing and the Srebrenica Genocide.

The component representing texts about the events in Srebrenica from a temporal perspective consists of educational texts,[5] Wikipedia entries and newspaper articles written few years after the massacre happened.[6] Also a number of parliament pieces[7] was included into the corpus component with historical perspective on the events in question.

5.3.2 Corpus processing and contrastive corpus analysis of event mentions

With the event granularity hypothesis presented in section 5.2 we hypothesize that the different temporal perspective of the writer corresponds with the difference in language use.

To validate our hypothesis, we carried out statistical research on the Srebrenica corpus. We conducted a contrastive analysis of event descriptions in the three components of the Srebrenica corpus with tools developed for the KYOTO project.[8] KYOTO is a platform for semantic processing of text according to a uniform conceptual model. It uses a pipeline-architecture of linguistic processors that generates a uniform semantic representation of the text in the so-called Kyoto Annotation Format (KAF).[9] KYOTO has been tested for seven different languages. KAF can be used to represent events with times, locations and participants. For the purpose of our research, the Srebrenica corpus was processed by means of the KYOTO architecture. First, the corpus was tagged with PoS information. It was lemmatized and syntactically parsed by means of

[3] The news articles were acquired through `http://academic.lexisnexis.nl/vu/` .

[4] Text type classification according to the Volkskrant archive.

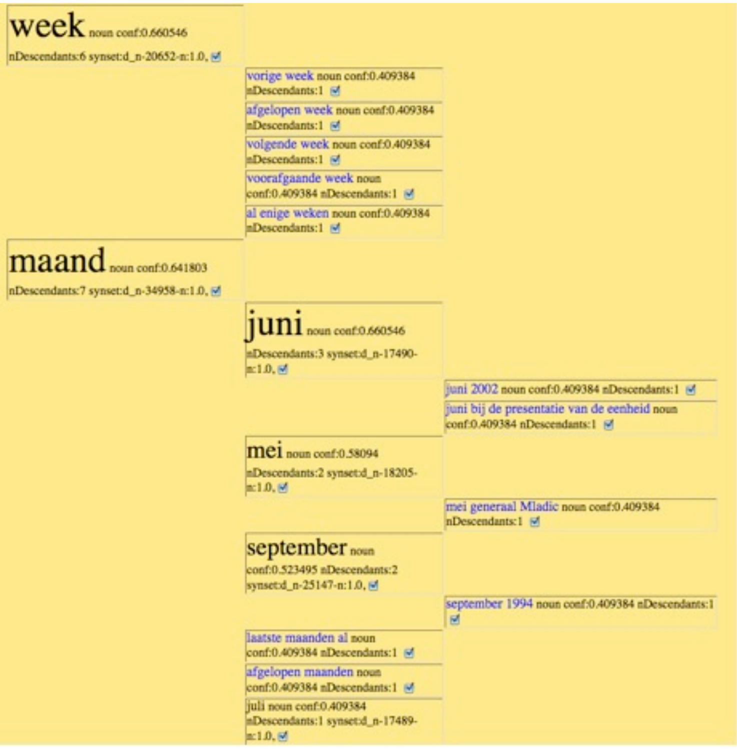
[5] An important source was www.entoennu.nl.

[6] The historical corpus component contains e.g. a dossier on the topic of the Dutch Royal Library from www.kb.nl. Some of the other sources were: www.nos.nl, www.nu.nl, www.anno.nl, www.groene.nl/home.

[7] Acquired at www.parlament.com.

[8] More information about the Knowledge Yielding Ontologies for Transition-based Organization - project can be found at www.kyoto-project.eu . See also Vossen et al. [2008a].

[9] Kyoto fact Annotation Format is described in Bosma et al. [2009].



week noun conf:0.660546 nDescendants:6 synset:d_n-20652-n:1.0, ☑

vorige week noun conf:0.409384 nDescendants:1 ☑

afgelopen week noun conf:0.409384 nDescendants:1 ☑

volgende week noun conf:0.409384 nDescendants:1 ☑

voorafgaande week noun conf:0.409384 nDescendants:1 ☑

al enige weken noun conf:0.409384 nDescendants:1 ☑

maand noun conf:0.641803 nDescendants:7 synset:d_n-34958-n:1.0, ☑

juni noun conf:0.660546 nDescendants:3 synset:d_n-17490-n:1.0, ☑

juni 2002 noun conf:0.409384 nDescendants:1 ☑

juni bij de presentatie van de eenheid noun conf:0.409384 nDescendants:1 ☑

mei noun conf:0.58094 nDescendants:2 synset:d_n-18205-n:1.0, ☑

mei generaal Mladic noun conf:0.409384 nDescendants:1 ☑

september noun conf:0.523495 nDescendants:2 synset:d_n-25147-n:1.0, ☑

september 1994 noun conf:0.409384 nDescendants:1 ☑

laatste maanden al noun conf:0.409384 nDescendants:1 ☑

afgelopen maanden noun conf:0.409384 nDescendants:1 ☑

juli noun conf:0.409384 nDescendants:1 synset:d_n-17489-n:1.0, ☑

Figure 5.2: Term hierarchy for time expressions extracted from the news component *Volkskrant*.

a dependency parser for Dutch - the Alpino-parser.[10] Next, word sense disambiguation was performed[11] and the corpus was semantically annotated with labels from the Dutch WordNet.[12] Finally, from each corpus component a hierarchy of terms was extracted by means of the KYOTO term extractor, Tybot. Tybots are term yielding robots that use patterns of PoS sequences (e.g. NN, N *of* N etc.) to obtain a hierarchy of terms and their hyponymy relations from any collection of texts in the KAF format.[13]

Through the mapping of terms with WordNet synsets it is possible to structure the full list of terms as a text-specific WordNet subtree, see FIGURE 5.2 for the example of time expressions. The words in big fonts (black) are Dutch equivalents for *week* and *month* from the Dutch WordNet, whereas the expressions in smaller font are specific terms and phrases for *weeks* and *months* detected in the news corpus *Volkskrant*. Note that the term extractor of KYOTO also includes general words in the term database. The semantic classification of the terms makes it possible to quickly annotate terms for times, locations and human participants. For example, all terms referring to soldiers of the Dutch troops and the Serb troops are grouped by a few synsets for *troops, army, soldier* etc., similarly for times and locations.

For each labeled term, statistical data is available in KYOTO on their frequency in each corpus component. Thus all terms labeled as times, locations and human participants of the Srebrenica Genocide event were collected per corpus component and statistical information on tokens and concept types from an event component was accumulated per corpus component (see section 2.4 for definitions of a token and of a concept type). We use the results for corpus analysis. First in TABLE 5.3 we look at token and concept type statistics in the three components of the Srebrenica corpus. Next in TABLE 5.4 we look at token and concept type statistics of event descriptions specifically. We consider the three event components, times, locations and human participants, in the three components of the Srebrenica corpus. Note that the statistical data depends on the quality of the word sense disambiguation and on the quality of the Kybots' output.

Validation of the event granularity hypothesis

TABLE 5.3 presents the statistical information on tokens and concept types in the three components of the corpus. The concept-token ratio was calculated per corpus component obtained by dividing the number of concepts by the number of tokens. The general corpus statistics from TABLE 5.3 show that there is a clear difference in the language use between the news lacking temporal perspective on the described event and texts written some time after the described event happened. The concept-token ratio for the

Corpus component	No. tokens	No. concepts	Concept-token ratio
Volkskrant	16573	1863	0.11
Historical texts	20742	1393	0.07
Het Parool	13481	1497	0.11

Table 5.3: Concept type vs. token statistics in Srebrenica corpus components.

[10] http://www.let.rug.nl/vannoord/alp/Alpino/

[11] For word sense disambiguation the UKB system (http://ixa2.si.ehu.es/ukb/) was used. For more information see Agirre and Soroa [2009].

[12] For more information see Vossen et al. [2008b].

[13] For more information on the KYOTO - Tybot see Bosma and Vossen [2010].

historical corpus component amounts to 0.07 and therefore it is remarkably lower in comparison with the concept-token ratio of the both news corpus components which equals to 0.11. Those statistics show that, as expected, in texts written from a historical perspective the number of concept types is much lower than the number of concept types covered in the two news corpus components. A lower proportion of concept types in comparison to tokens indicates less conceptual diversity.

In the previous paragraph we considered statistics generated for all tokens of the three components of the Srebrenica corpus. Next, we will look into statistics only with regard to tokens used in descriptions of the Srebrenica Massacre event. We will consider here three of the four components of events, that is, the time, location and the human participant component. To get some insights into the frequency of tokens and concepts referring to times, locations and human participants in descriptions of the Srebrenica Genocide, the times, locations and human participants were manually labeled as such and statistics were generated per event component for each corpus component.[14] The concept-token ratio was calculated and additionally the absolute and normalized frequencies of concept types and tokens were tabulated. TABLE 5.4 presents the results of the statistical analysis.[15]

Event component	**Measurement**	***Volkskrant***	***Het Parool***	**Historical texts**
Time	Concepts	84	44	36
	Tokens	95	48	75
	Concept-token ratio	0.88	0.92	0.48
	Normalized freq. tokens	0.006	0.004	0.004
	Normalized freq. concepts	0.045	0.029	0.026
Location	Concepts	111	100	71
	Tokens	124	108	91
	Concept-token ratio	0.90	0.93	0.78
	Normalized freq. tokens	0.007	0.008	0.004
	Normalized freq. concepts	0.06	0.067	0.051
Human participant	Concepts	464	379	236
	Tokens	519	412	416
	Concept-token ratio	0.89	0.92	0.57
	Normalized freq. tokens	0.031	0.031	0.02
	Normalized freq. concepts	0.249	0.253	0.169

Table 5.4: Time, location and human participant statistics in the Srebrenica corpus components.

The same trends that we observed when analyzing general token and concept statistics from TABLE 5.3 hold when looking at event descriptions in TABLE 5.4. There is a clear difference in the concept-token ratio for all three event components – time, location and human participant – between news and historical texts. In both news components the concept-token ratio for all components of event descriptions is ca. 0.9 while for historical texts, as expected, the proportion of concepts to tokens is remarkably smaller than in the news. That is especially clear when considering the concept-token

[14] The manual tagging was based on the semantic tagging of all concepts by means of the KYOTO-pipeline. Tybot hierarchies were generated for selected concepts.

[15] The statistical results could be further improved by inclusion of frequency counts for proper names, geo-names and adverbial time and location pointers.

ratio of times: 0.48 and participants: 0.57. The relatively high concept-token ratio of locations can be explained by pragmatic factors. The events in Srebrenica happened in a relatively small geographic area and so the variation possibilities with regard to linguistic expressions are limited. To summarize, around 90% of all event concepts referred to in the news occurs uniquely per corpus component against only ca. 50% of time and human participant concepts and less than 80% of location concepts in historical texts. The percentage of the event concepts, which reoccur per corpus component, in the news amounts only to ca. 10% and is remarkably lower than in the texts written from a historical perspective: ca. 30-50% of event concepts are referred to more than once. The number of concepts covered in the news is thus remarkably higher than the number of concepts used in the corpus component written from a historical perspective when looking at both the general token and concept statistics as well as when looking at event descriptions.

This supports the event granularity hypothesis from section 5.2. News texts, while lacking or having a shorter temporal perspective, give a detailed account of a story describing a higher number of events at a finer granularity level. A detailed account of many fine granularity events is reflected in a relatively high number of concepts in proportion to tokens. On the other hand, texts having a longer temporal perspective on the described events abstract from details and describe a smaller number of events at a coarse granularity level. There is less concept diversity and so the number of concepts in proportion to tokens is lower than in the news as more frequently than in the news different tokens refer to the same concepts.

Finally, in TABLE 5.4 there are no systematic differences between normalized tokens and concept frequency counts in the different corpus components. The normalized counts for the news and for the historical corpus component are nearly identical for times. The most significant difference in counts is within the location component for reasons explained earlier. As mentioned, this arises from the fact that the Srebrenica events happened in a small geographic area. The compatible normalized frequency counts confirm the validity of the concept-token ratio statistics. These counts show that the concept-token ratio statistics cannot be explained by a difference in the text length of the samples.

5.4 Determining event granularity[16]

In the previous section we demonstrated that there is a correlation between granularity of event descriptions depending on the writer's perspective and text types used. Texts written shortly after an event happened tend to describe events at the lower level of granularity while texts written from a longer temporal perspective focus on events with coarse granularity. Our analysis focused on descriptions of times, locations and human participants.

If one could automatically determine whether component mentions are compatible with regard to granularity from the perspective of coreference, one could use this information when solving event coreference (or determining other event relations). The intuition behind this approach is that a coarse granularity event with a longer duration, as occurs on a bigger area and with multiple human participants (for instance *a war between Russia and Ukraine*) might be related to, but will probably not fully corefer with, lower granularity events of shorter duration and with a single participant involved (e.g. *A Russian soldier has shot dead a Ukrainian naval officer*).

[16] The contents of this chapter have been published before as Cybulska and Vossen [2015b]

In this section, in line with our research of the Srebrenica corpus from section 5.3, we focus on granularity of times, locations and human participants. We present a new taxonomy defining granularity levels for times, locations and human participants. There are 15 predefined semantic classes in the taxonomy that represent different granularity levels, which are defined over 434 hypernyms in WordNet, covering 11979 WordNet synsets. The granularity taxonomy is available for further research.

5.4.1 Related work on granularity in NLP

We introduced the notion of granularity in section 2.3. In this section we look at related works on granularity in NLP.

Few researchers looked at granularity in natural language. Vossen studied granualarity of nominal concepts in language (Vossen [1995]). Considering the variation in the degree of specification of word meaning, Mani [1998] suggested development of a knowledge representation that makes the notion of granularity explicit. Mani applied shifts in granularity to problems of polysemy and underspecification of nominalizations. Change in granularity was considered as a special case of abstraction in which elements, which are indistinguishable in a particular context, are collapsed. Mani focused on grain-size shifts amongst polysemous events. Mulkar-Mehta et al. [2011a] describe event granularity as the concept of breaking down a higher-level event into smaller parts, fine-grained events such that each smaller granule plays a part in the higher level whole. Relation types that can exist between the objects at coarse and fine granularity are part-whole relationships amongst entities and events, and causal relationships. Based on annotation of granularity relations in text, the authors conclude that part-whole and causal relations are a good indication of shifts in granularity.

In this work we focus on the notion of granularity in event descriptions. We present a new granularity taxonomy which captures the degree of granularity of event components explicitly for the purpose of usage in NLP applications. We use a taxonomy to distinguish between coarse- and fine-grained granularities of different parts of event descriptions. The intrinsic, conceptual granularity is captured by means of a number of granularity levels defined in the granularity taxonomy. In part IV of the dissertation we determine whether mentions are compatible with regard to granularity for the purpose of event coreference resolution. The motivation behind this approach is an expected correlation between agreement or disagreement in grain-size levels and the notion of coreference. In the prototypical situation, agreement or small granularity differences are expected to indicate coreference. Greater distance in granularity is expected to be a negative indicator of coreference or to indicate other event relations such as scriptal (when events are part of a larger structure, see Schank [1990]) or event membership.

Event component	**Granularity class**	**Description**	**Synset example**
Human participant	*gran_person*	individuals	spokesperson_1
	gran_group	groups or organizations	people_2
Location	*gran_street*	areas up to the size of a building	government_building_1
	gran_city	city districts and cities	city_district_1
	gran_country	size of a country	Upper_Egypt_1
	gran_continent	size of multiple countries	East_Africa_1
Time	*gran_second*	duration up to a minute	sec_1
	gran_min	from a minute to an hour	quarter_4
	gran_hr	from an hour up to 24 hours	hours_2
	gran_day	one to few days, less than a week	day_of_the_week_1
	gran_week	one to few weeks, less than a month	calendar_week_1
	gran_month	indication on the month level	Gregorian_calendar_month_1
	gran_season	few months	season_1
	gran_year	one or multiple years	year_1
	gran_thousands_years	thousands of years	Bronze_Age_1

Table 5.5: Granularity taxonomy classes distinguished per event component.

```
eng-30-08160276-n,gran_group,”citizenry_1,people_2”
eng-30-10638385-n,gran_person,”spokesperson_1,interpreter_3,representative_2,voice_8”
eng-30-15235126-n,gran_second,”second_1,sec_1”
eng-30-15234942-n,gran_min,”quarter_4”
eng-30-15117516-n,gran_hr,”hours_2”
eng-30-15163005-n,gran_day,”day_of_the_week_1”
eng-30-15136147-n,gran_week,”week_3,calendar_week_1”
eng-30-15209706-n,gran_month,”Gregorian_calendar_month_1”
eng-30-15239579-n,gran_season,”season_1”
eng-30-15203791-n,gran_year,”year_1”
eng-30-15231415-n,gran_thousands_years,”Bronze_Age_1”
eng-30-03449564-n,gran_street,”government_building_1”
eng-30-08537837-n,gran_city,”city_district_1”
eng-30-08898002-n,gran_country,”Upper_Egypt_1”
eng-30-08699426-n,gran_continent,”East_Africa_1”
```

Figure 5.3: Example entries from the granularity taxonomy file.

5.4.2 Granularity taxonomy for event times, locations and human participants

We created a granularity taxonomy that consists of 15 predefined semantic classes. The 15 semantic classes represent different granularity levels, which are defined over 434 hypernyms in WordNet, covering 11979 WordNet synsets.

We focus here on partonomic granularity relations (representing granularity through the part-of relation) between event entities. Our 15 semantic classes belong to four relationships from the taxonomy of meronymic relations by Winston et al. [1987]. Granularity levels of the human participant component are contained within the Member-Collection relations of Winston et al. Our temporal granularity levels make part of Winston's Portion-Mass relationships and our locational levels are in line with Place-Area relations in Winston's taxonomy.

FIGURE 5.3 presents a fragment of the granularity taxonomy with synset examples for every granularity class. The taxonomy file is comma separated. In the first column synsets from WordNet 3.0 are indicated. In the second column the granularity levels are captured and the third column indicates the synset IDs as stored in the Natural Language Toolkit (NLTK, Bird et al. [2009]). The choice of the 15 granularity classes was motivated by an analysis of event descriptions in the news. We intended to capture shifts in granularity that seemed meaningful for event coreference resolution on a news corpus such as the ECB (Bejan and Harabagiu [2010]) or ECB+ (Cybulska and Vossen [2014b]). We manually assigned the semantic classes to 434 hypernyms in WordNet which are linked to 11979 synsets. We recognize a number of granularity levels per event component: nine levels for times, four for locations and two for human participants, as presented in TABLE 5.5. The complete taxonomy is presented in appendix 9.

Our granularity taxonomy does not include event actions. Some work on capturing granularity of event actions (in Winston et al. [1987] Feature-Activity relation) was done by Gusev et al. [2011]. To determine granularity of event actions one can use the database of event durations created by Gusev et al. [2011]. The lexicon of event durations (`http://cs.stanford.edu/people/agusev/durations/`) captures durations for events (with or without syntactic objects) inferred by means of web query patterns. Duration distributions were learned with an unsupervised approach. Eight duration levels are considered: *seconds, minutes, hours, days, weeks, months, years* and *decades*. The durations database covers the 1000 most frequent verbs with 10 most frequent grammatical objects of each verb from a newspaper corpus from the New York Times.

5.5 Conclusion

In this chapter we showed that different temporal perspectives of writers correlate with event granularity and correspond with genre diversity. The diversity of language use makes it difficult for a typical search system to find all the information that is semantically connected to an event but formulated in a different way. On the other hand, regularities in the language use within a genre open possibilities for automatic information retrieval from news articles and historical texts.

To capture differences between event representations in diverse text types, we defined an event model that captures event granularity and we carried out some corpus research to confirm our hypothesis on the topic. Granularity of event components correlates with the temporal perspective of the writer. Contrastive corpus analysis made clear that texts lacking or having a shorter temporal perspective include many more specific, uniquely occurring references to times, locations and human participants than texts written from a longer temporal perspective, which remain rather general in their presentation of coarse granularity events. The observed relations between low and high granularity of times, locations and human participants can be used for event coreference resolution and to determine relations between events and their sub-events, across different genres of text.

To facilitate the use of granularity as an indication of event relations, we created a new WordNet-based granularity taxonomy for event times, locations and human participants. Future research could be dedicated to considering extending the granularity taxonomy by learning granularity levels from corpora to overcome the low coverage limitation following from the usage of a WordNet-based taxonomy. One could also augment the taxonomy to cover the non-human participant slot and experiment with other ways to represent event granularity.[17]

In the next chapter, we will investigate the extent to which the insights of this chapter and, more specifically, the granularity taxonomy can enhance approaches to event coreference.

[17] Note that these conclusions are based on a single use case with accounts of the Srebrenica Massacre. This research should be extended to more use cases that also cover different types of events.

Chapter 6

Model of gradable event coreference[1]

In this chapter we formalize our model of gradable event coreference. The model determines whether component mentions are compatible with regard to granularity and hyponymy distance and uses this information as a clue for event coreference resolution. If event component mentions are not coreferent but related through the meronymy or hyponymy relation, this could imply other event relations such as sub-event, causal or temporal relations. That said, the focus of this chapter lies on the identification of semantic relations between event descriptions for the purpose of solving event coreference.

6.1 Theoretical underpinnings

Our approach to event coreference makes two crucial assumptions. First, in accordance with the Quinean theory (Quine [1985]), we assume that semantic relations and coreference between elements of the contextual setting of events are crucial for solving event coreference. As already introduced in section 2.2, the contextual setting of an event as well as its participants cannot be separated from the event itself because they constitute the event. Time and place in which an event happened are crucial for event coreference and so they form the starting point for solving event coreference. Compare *car bombing in Madrid in 1995* with *car bombing in Spain in 2009*. Without time and place information event actions are just denotations of abstract classes of concepts. They need to be anchored in time and space to become instantiated.[2] Coreference thus only makes sense for events within the same time and place. Hence for each event mention in text, one should first try to define time and place and after that, for events occurring within a compatible time and space, search for other linguistic coreference clues. From a practical point of view, determining event time and place should limit the number of candidates for coreferent events and improve the precision of event coreference resolution. Accordingly, to solve event coreference we employ an event model which consists of four components. Our approach determines whether

[1] The contents of this chapter have been published as Cybulska and Vossen [2012].

[2] An exception are event descriptions that depict instances of events that over time have become proper names such as *9/11* or *Srebrenica massacre*.

two events corefer based on the combination of coreference scores calculated for each event component: action, time, location and participant.

Second, we assume that (linguistic) coreference is not an absolute notion. For example, *shooting* and *several shots* can refer to the same event and people may have different or vague intuitions about their identity (for a discussion of full and partial coreference see also Hovy et al. [2013]). A gradable notion of coreference is therefore both operational (for robust automatic detection) and possibly psychologically adequate, compare: *two students taken hostage in Beslanian school* vs. *two people taken hostage in a classroom in Beslan, Russia.* Therefore, for each event pair in the text, we want to calculate a coreference match score as a combination of coreference scores collected for pairs of event components. To obtain the match score for an event component, we will analyze semantic relations and semantic distance between two instances (for instance participants of event A in comparison with participants of event B). Shifts vs. agreement in the level of granularity and distance in hyponymy will play a crucial role in the assignment of the match scores together with other coreference indicators such as identification of repetition, anaphora, synonymy and disjunction. The cumulative coreference match score gathered by an event pair will indicate whether two event descriptions can be considered likely candidates for exhibiting a coreference relation.

The approach used in our work employs a gradable notion of coreference. The probability of coreference is a continuum that goes from non-disjoint, strongly coreferential event mentions to other event relations: scriptal, is-a, and membership relations. Coreference of event action mentions (compare *bombing* vs. *bombing attack*) can gradually transition into other event relations:

- scriptal: event vs. its subevent e.g. *explosion* as a step in the script of a *bombing attack*
- is-a: *bombing* is a kind of *attack*
- membership relations: an *attack* is a member of *series of attacks*.

In a prototypical situation, the gradual probability of coreference inversely correlates with semantic distance between two instances. Semantic distance between instances of an event component can be determined by the kind of semantic relation between them. The model described in this chapter captures the relationship between different semantic relations and coreference on one end of the spectrum and (if not disjoint) other event relations on the other.

In section 6.2 we describe related work with regards to application of semantic shifts in NLP applications. In section 6.3 we present the model that captures the relationship between semantic relations and event coreference. In section 6.4 we draw conclusions.

6.2 Related work[3]

Mulkar-Mehta et al. [2011a] investigated granularity shifts and granularity structures in natural language text. They focused on modeling part-whole relations between entities and events and causal relations between coarse and fine granularities. In their follow-up work (Mulkar-Mehta et al. [2011b]), they described an algorithm for extracting causal

[3] The related work section was written before the publication of the content of this chapter in 2012 (Cybulska and Vossen).

granularity structures from text and its possible applications in question answering and text summarization. Howald and Abramson [2012] successfully used granularity types as features for prediction of rhetorical relations with a 37% performance increase. In our work, we model granularity correlations and hyponymy distance for the purpose of event coreference resolution. To the best of our knowledge, granularity estimates and hyponymy distance have not been used before for this task.

6.3 Proposed model of gradable event coreference based on semantic relations between event components

To capture differences between event descriptions, we apply an event model which consists of four components: action, time, location and participant. As introduced in chapter 5, in textual data one comes across event actions, times, locations and participants at different levels of granularity. Compare for instance actions such as *shooting, fighting, genocide* and *war*, or participants *soldier*, (multiple) *soldiers*, *troops* and *multiple troops*. The same holds for time descriptions, consider *day, week* and *year*, and also for locations such as *city*, *region* and *continent*. Furthermore, there can be different hyponymy distance between component mentions. TABLE 6.1 exemplifies instances of event components related through hyponymy and meronymy. Component mentions are either (partially) overlapping or disjoint.

Component	**Is-a:**		**Inclusion:**
	Class >Subclass	**Class >Instance-of**	**Part-of, Member**
Location	city >capital	country >Bosnia	Bosnia >Srebrenica
Participant	officer >colonel	colonel >Karremans	army >soldier
Time	weekday >Friday	July >July 1995	week >Monday
Action	attack >bombing	genocide >the Srebrenica massacre	series of attacks >attack

Table 6.1: Examples of event components related through hyponymy and meronymy relation.

Typical indicators of coreference are repetition, synonymy, anaphora and disjunction (negative indicator). Next to the indicators of full coreference that are typically used in coreference resolution, significant relations between event components are along a hyponymy axis:

- class vs. its subclass such as *officer* being a subclass of the class *person*,
- instance-of a class such as *Bosnia* being an instance of the class *country*

and along a meronymy axis:

- member vs. group i.e. *Colonel Karremans* being a member of the group of *Dutch UN soldiers* or
- part vs. whole relation such as *Srebrenica* being a part of *Bosnia*.

In addition to different hyponymy distance and different degrees of granularity, words and word combinations at the same level of granularity may differ in terms of pragmatic use, while potentially referring to the same thing. Compare, for example, event participants referred to as *aggressors* and *liberators* or *troops*, *army* and *soldiers*. The same applies to event actions; compare *liberation* with *invasion* or *military intervention*. When solving event coreference and determining event relations, the pragmatic loading has to be accounted for as well. In other words, one has to be able to distinguish between marking of ideological perspectives and proper semantic disjunction.

Our gradable model of event coreference implies that probability of coreference between complete events can be determined by the semantic relations between the event components (below referred to as *Ec*). We thus first define per event component a coreference match (below referred to as *CM*) as a function of the relation type and semantic distance between the instances of components. The highest coreference match (value 1) should be assigned to synonymous items, repetitions, as well as to anaphora in case their number and gender agree, see formula 6.1–6.3.

$$CMrepetition(Ec1, Ec2) = 1 \tag{6.1}$$

$$CManaphora(Ec1, Ec2) = 1 \tag{6.2}$$

$$CMsynonymy(Ec1, Ec2) = 1 \tag{6.3}$$

Similarly, a high match score is used for events with only a difference in perspective (for instance *buy* vs. *sell*), consider formula 6.4.

$$CMperspective(Ec1, Ec2) = 1 \tag{6.4}$$

We further expect that hyponymy relations across event components indicate a probability of coreference. Formula 6.5 expresses that distance inversely correlates with the likelihood of coreference. *Ehl* stands for the estimated hyponymy level within a shared chain of hyponymy relations for *Ec1* and *Ec2* in a resource such as WordNet. By “shared” we mean that the concepts are not disjoint according to the interpretation of the hierarchy. See, for instance, the hyponymy chain from English WordNet connecting the concepts of *hostage* and *person*: *hostage<captive/prisoner <unfortunate <person <being/organism <living thing*).

$$CMhyponymy(Ec1, Ec2) = \frac{1}{(1 + |\Delta(Ehl(Ec1), Ehl(Ec2))|)} \tag{6.5}$$

Meronymy relations between instances are expected to indicate granularity agreement or shifts, where the value of *CM* inversely correlates with the difference in size of the meronymic whole, i.e. the larger the difference in size, the lower the score. This is formalized in formula 6.6. *Eni(Ec)* stands for the estimated number of individuals denoted by *Ec*. *Eni* can be based for instance on predefined levels of granularity, where a large difference in levels correlates with a large difference in the number of denoted individuals. See section 5.4 for a description of our granularity taxonomy. The taxonomy specifies levels of concepts per event component. Granularity levels are distinguished based on the WordNet knowledge base. Furthermore, one must consider multiplications within a level as well: 24 hours make 1 day.

$$CMmeronymy(Ec1, Ec2) = \frac{1}{(1 + |\Delta(Eni(Ec1), Eni(Ec2))|)} \tag{6.6}$$

If two instances are disjoint (for instance human participants of different gender) the match score will equal zero as indicated by formula 6.7.

$$CMdisjunction(Ec1, Ec2) = 0 \tag{6.7}$$

Once the above values have been calculated for every component of an event pair, the collected scores should be combined into a single score for an event pair indicating the likelihood of coreference. Our model predicts that, in a prototypical situation, except for the clear cases resulting in an absolute score of 1 or 0, event components that are far apart in terms of meronymy and hyponymy have an extreme semantic distance and therefore a low likelihood to establish coreference. A participant example would be a *US sergeant* (specific in terms of hyponymy and a single-form) versus *human being*, where the latter does not exclude *US sergeants* but there is a low likelihood that the author is referring to the same referent. For event actions, this could be *a briefing by an US sergeant* versus *strategics*. Through empirical testing, one could determine thresholds for establishing optimal coreference relations across events. Finding optimal thresholds based on data means considering also less prototypical cases when granularity disagreement of a component might not exclude event coreference for example when a human participant of an event could be expressed as a single individual or as a country name but in reference to a leader of the country. Compare *Mark Rutte signed an agreement* vs. *The Netherlands signed an agreement*.

In the prototypical case, within events we observe granularity correlations. If an event action has a coarse granularity (for instance *war*) one can expect the participants of this action to be a multiform and certainly not a single individual. The same holds for the location of a *war* event (one can also expect a location at a higher granularity level for example a territory of a country instead of a small area) and its time span (a longer time period). This observation offers a perspective that it may be possible to determine the granularity level of one event component from those of other components with which it often co-occurs.

In the ideal situation, one has information on all event components. More realistic is the situation where event components are underspecified in the event mentions, for instance in the case of nominalizations (*war, shooting*). Underspecified nominalizations (no time, location and no event participants made explicit) tend to refer to events with coarse granularity that are expected to be described earlier in the text in more detail and so at a lower granularity level. Incomplete events should be analyzed in a separate way. An interesting possibility for future research is to try to learn the missing event information from other knowledge sources (for instance in case of named events from Wikipedia).

Different heuristics can be employed to estimate semantic distance between event mentions in text. Two groups of techniques can be used to define the difference in hyponymy and meronymy: (1) analysis of the text and of the morpho-syntactic properties of event mentions and (2) using background knowledge: either learned from existing resources as WordNet, geo- and temporal ontologies or knowledge based on probability estimates from corpora. Regarding the latter, one could, for instance, try to learn the typical length of duration that is most frequently associated with an action and use this for hyponymy and meronymy estimates, see for example Gusev et al. [2011]. In our experiments with event coreference resolution in part IV we use morpho-syntactic properties (number and multiplications) of mentions and pre-defined WordNet-based granularity levels to determine grain size of event components as well as hyponymy-based semantic distance measures.

6.4 Conclusion

In this chapter, we proposed a model of gradable event coreference that captures the relationship between event granularity and hyponymy distance and event coreference resolution based on semantic relations between mentions of event components. Semantic relations – hyponymy and meronymy – together with other coreference indicators such as repetition, synonymy, anaphora and disjunction are indicative of event coreference. In the fourth part of this dissertation we will experiment with the model of gradable event coreference. In the experiments we use (1) semantic distance estimates based on the hyponymy distance in the WordNet database and (2) agreement in grain size estimates to identify granularity shifts.

Future research could be dedicated to finding the optimal way to implement the gradable model of event coreference. It would be interesting to evaluate different methodologies that can be used to estimate the probability of coreference between mentions of event components. A special treatment should be developed for proper name mentions. Event components or events expressed by proper names can be analyzed with help of background knowledge to look up the parts of event information that are not explicitly expressed in text or to determine grain size of event component mentions expressed by proper names. Finally, an interesting study could be performed to analyze regularities between both granularity agreement and disagreement as indication of event coreference.

Part IV

What is the role of times and entities in event coreference resolution?

In part III of the dissertation we proposed a gradable model of event coreference. In part IV we experiment with the model. In chapter 7 we describe a rule-based experiment. In chapter 8 we experiment with machine learning techniques. The experiments are designed to help us understand the role of times and entities in event coreference resolution. This part of the dissertation is dedicated to the research questions below.

Research questions

- Do times and entities matter in solving event coreference?
- Can we measure the contribution of times and entities to event coreference resolution?
- How is the event information packaged in a typical news text?
- What is the optimal way to make use of times and entities for event coreference resolution?

Chapter 7

Rule-based experiments: contribution of times and entities to event coreference resolution[1]

In this chapter, we experiment with our approach to event coreference resolution presented in section 6.3 which employs both the full and partial linguistic coreference between mentions of events and their times, locations and participants. The first goal of this chapter is to measure the contribution of different components of event descriptions to the task of event coreference resolution. We calculate what event times, locations and participants add to event coreference resolution. Another goal is to evaluate different heuristics that can be used to determine the partial coreference between mentions of event components for the purpose of event coreference resolution. We analyze two techniques: (1) using the hyponymy distance between entity mentions and (2) determining grain size of entities based on pre-defined granularity levels together with an analysis of lexical granularity clues. We will compare results achieved with these two techniques within the participant component.

Considering the goals, we deliberately do not use machine learning as we want to have a clear picture of what the contributions are by different factors. The idea is that coreference of events is calculated from the coreference match scores of each event component. Coreferent action candidates are accordingly filtered based on compatibility of their times, locations, or participants. We report the success rates of our experiments on the ECB 0.1 corpus.[2]

Having an idea of how various event components influence event coreference could guide the feature choice for machine learning. We will experiment with machine learning in chapter 8.

[1] The contents of this chapter have been published as Cybulska and Vossen [2013].

[2] Note that this part of the research was done before the ECB+ corpus was created.

7.1 Gradable approach to event coreference

We argued in chapter 5 that descriptions of one event can differ in granularity (compare: *two students taken hostage in Beslanian school* vs. *two people taken hostage in a classroom in Beslan, Russia*). Coarse granularity events, such as *war*, are more general and abstract with longer time span and group participants; fine granularity events, e.g. a *shooting* event, are rather specific with shorter duration, and individual participants. To capture differences between event representations and to identify relations between events, we apply an event model that consists of four components, namely, action, time, location and participant.

Within our approach to event coreference (as presented in section 6.3), we analyze semantic relations and semantic distance between two instances of each event component to obtain a coreference score per component. We do not only take exact lemma-based matches of event mentions into account but we allow for soft matching based on semantic relations. Our intuition is that semantic relations play a crucial role in establishing coreference relations together with other coreference indicators such as lemma-repetition, anaphora, synonymy and disjunction. Once semantic distance and granularity agreement is calculated for every component of an event pair, the separate scores are combined into a single score for an event pair indicating the likelihood of coreference. Through empirical testing, we determine thresholds for establishing optimal coreference relations across events and their components.

In section 7.2 we position the experiment with our gradable approach to event coreference resolution within related works. In section 7.3 we experiment with our approach to event coreference by means of rule-based techniques. In section 7.4 we discuss evaluation results. We conclude in section 7.5.

7.2 Related work[3]

One of the recent approaches to event coreference resolution was proposed by Bejan and Harabagiu [2010], who experimented with nonparametric Bayesian models. Another one, by Chen et al. [2011], employs support vector machines with tree kernels and spectral graph partitioning. These approaches do not explicitly account for partial coreference of events, where some of the event components are related through the hyponymy or meronymy relationships, which is the focus of our work. Bejan and Harabagiu noted that not accounting for partial coreference is the reason for one of the common errors in their output. The approach of Chen et al. accounts for synonymy between mentions but not for meronymy or hyponymy.

Soft matching was successfully used for entity coreference resolution. Taxonomy based semantic similarity and semantic relatedness (Wikipedia based) were used as features in a machine learning approach to entity coreference by Ponzetto and Strube [2006]. Some semantic features based on synset relations in WordNet are used by Ng and Cardie [2002] and Ng [2005], while Harabagiu et al. [2001] use hyponymy, meronymy and other semantic relations from WordNet for NP coreference. They employ WordNet to distinguish between individuals and groups amongst entities of category person.

Entity coreference has been used explicitly for event coreference resolution in the experiments performed by Lee et al. [2012]; where entities and event clusters are

[3] The related work section was written in 2013, before publication of the content of this chapter by Cybulska and Vossen [2013].

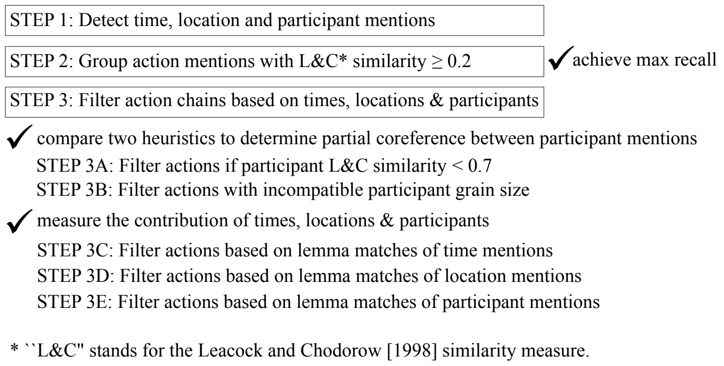
merged by means of linear regression. Partial coreference is incorporated by using distributional similarity as one of features for cluster comparison. Other approaches use entities for event coreference in a more indirect way e.g. Bejan and Harabagiu [2008] and Bejan and Harabagiu [2010] use semantic roles as features for their SVM classifiers. Bejan and Harabagiu [2010] account only for synonymy amongst heads of semantic roles. Chen and Ji [2009a] check for verbal argument compatibility for *Time-Within* and *Place* roles. Their results indicate that features related to event arguments only slightly (ca. +1% MUC and B3) improve event coreference, possibly due to wrong argument labeling. In this work, we measure the influence of times, locations and participants on the task of event coreference resolution.

A theory-oriented discussion about the nature of full-, near- and non-identity of entities and a continuum approach to entity coreference is presented in Recasens et al. [2011]. A discussion of full and partial identity of events, pointing out the significance of partial coreference for coreference resolution, appears in Hovy et al. [2013].

Semantic shifts have been used before in NLP applications. Mulkar-Mehta et al. [2011a] investigated granularity shifts and structures in natural language text. They focused on modeling part-whole relations between entities and events and causal relations between coarse and fine granularities. In their follow-up work (Mulkar-Mehta et al. [2011b]), they described an algorithm for extracting causal granularity structures from text and its possible applications. Howald and Abramson [2012] use granularity types as features for prediction of rhetorical relations. Their results show that granularity types significantly improve the performance of prediction of rhetorical relations amongst clauses. In our work, we measure the contribution of shifts in granularity and hyponymy distance to the task of event coreference resolution.

7.3 Experiment

FIGURE 7.1 outlines the main steps of our experiment.



STEP 1: Detect time, location and participant mentions

STEP 2: Group action mentions with L&C* similarity ≥ 0.2 ✓achieve max recall

STEP 3: Filter action chains based on times, locations & participants

✓ compare two heuristics to determine partial coreference between participant mentions

STEP 3A: Filter actions if participant L&C similarity < 0.7

STEP 3B: Filter actions with incompatible participant grain size

✓ measure the contribution of times, locations & participants

STEP 3C: Filter actions based on lemma matches of time mentions

STEP 3D: Filter actions based on lemma matches of location mentions

STEP 3E: Filter actions based on lemma matches of participant mentions

* ``L&C" stands for the Leacock and Chodorow [1998] similarity measure.

Figure 7.1: Experiment design.

As the goal of the experiment is not to evaluate event detection but the contribution of components to event coreference resolution, we used the stand-off gold standard annotation of event actions by Lee et al. [2012] on top of the EventCorefBank (ECB) corpus[4]. The ECB 0.1 corpus is annotated with cross-document coreference between event mentions. The corpus contains 482 texts from Google News (selected based on inclusion of keywords such as *commercial transaction, attack, death* or *sports*) and grouped into 43 topics.

To be able to experiment with the contribution of times and entities to event coreference resolution, our system extracted times, locations and participants from the ECB corpus, see STEP 1 from FIGURE 7.1. The ECB texts were processed by means of tools developed within the KYOTO project.[5] First, the corpus was lemmatized and tagged with PoS and syntactic information (using the Stanford Parser[6]). Next, word sense disambiguation was performed and the corpus was annotated with synsets from the English Wordnet (version 3.0) and with predefined taxonomy classes. The event taxonomy was manually assigned to 266 hypernyms in WordNet. It consists of four main semantic classes of concepts – one for each event component – action, time, location and participant which altogether cover 53964 synsets. All manually annotated actions from the corpus were used as input in the experiments. To extract times, locations and participants newly created extraction rules for English were used, based on manual annotation of event components in five independent texts. By means of the Kybot module of KYOTO, event times, locations and participants were extracted through rules employing some syntactic clues, PoS and combinatory information together with semantic class definition and exclusion by means of WordNet (Cybulska and Vossen [2011]).

After extracting time, location and participant mentions from the corpus, we move on to generating preliminary chains of coreferring action mentions within a topic. In this step of the experiment we use the gold standard annotations. Action mentions are grouped based on semantic similarity with the objective to ensure maximal recall, see STEP 2 in FIGURE 7.1. Semantic similarity between mentions can be calculated by means of different measures. We employed a taxonomy-based edge counting technique of Leacock and Chodorow [1998], which considers the closest hyponymy path in WordNet between two synsets scaled by the overall depth of the taxonomy. See formula 7.1 where *Si,j* is the similarity between mentions i and j from M (total set of mentions in a topic); where *M(Di,j)* is the minimal distance between two concepts and *Avg(Ddepth)* is the average depth in WordNet for all meanings of all candidates in the topic.

$$(Si, j) = log(M(Di, j)/(2 * Avg(Ddepth))) \tag{7.1}$$

Mentions with relatively short semantic distance between their heads, constitute candidates for coreference chains. For mentions that use the same word, we ignore the synset but consider distance of 1. For synonyms, we use distance of 2. In all other cases, we add the hypernym distance to the initial value of 2. After obtaining the similarity scores for all mentions in a topic, we normalize the scores. We created a matrix between all mentions in a topic and calculated the Leacock and Chodorow similarity (from now on also referred to as L&C) scores. A maximum recall was obtained if we keep equivalence relations for similarity scores of 20% or more of the highest score

[4] http://faculty.washington.edu/bejan/data/ECB1.0.tar.gz, Bejan and Harabagiu [2010]

[5] The ECB corpus texts after processing with the KYOTO tools (a pipeline of linguistic processors) are available at http://www.newsreader-project.eu/results/data/.

[6] http://www-nlp.stanford.edu/software/lex-parser.shtml

within a topic (usually the lemma). For each event mention, we thus keep candidate coreference relations to other mentions if the score is 0.2 or higher.[7]

After generating preliminary event coreference chains based solely on action similarity to ensure high recall, we moved on to the third step of the experiment namely additional filtering of semantically similar actions based on compatibility of their times, locations and participants, see STEP 3 from FIGURE 7.1.

STEP 3A and 3B from FIGURE 7.1 depict our experiments with partial coreference matches of participants in the context of event coreference resolution. We use two different heuristics to determine participant compatibility, one using the distance in hyponymy and another one making grain size estimates based on pre-defined granularity levels together with lexical granularity clues. Note that this participant compatibility is not limited to full identity of participants. Soft matching of participants is more appropriate for the purpose of this task to account for cases of metonymy, e.g. *US aircrafts* instead of *US army*. To generate chains of coreferent participants based on the hyponymy distance, again we use the L&C (the same procedure as in case of action similarity). We determined the optimal coreference threshold for participant mentions on 0.7 normalized L&C score.

Our second heuristic estimates the grain size agreement or disagreement for participant mentions. Event coreference chains are created in case of compatible grain size of participant mentions. To make a grain size estimate, we defined two semantic classes over synsets in WordNet: *gran_person* (e.g. *soldier, doctor*) denoting individual participants and *gran_group* referring to multiple participants (e.g. *army* or *hospital*). These two classes cover 36 WordNet hypernyms which map to 9922 synsets. On top of agreement in grain size levels, we also account for lexical granularity clues within a level such as number and multiplications. At this point we make a rough distinction between one and multiple items within a concept type (e.g. *gran_person*). A difference in grain size level or number is treated as an indication of a granularity shift and is turned into a distance measure. To better handle 43415[8] participant mentions that were POS-tagged as named entities, we decided to add an intermediate *gran_instance* class (for named entity participants that have no synsets such as person or organization names like *John*, or *Doctors Without Borders*) so that we can encourage number matching for our measurements of what grain size exclusively can contribute to event coreference. For agreement in semantic class level, two participant instances can maximally get 3 points. If there is 1 level difference between them (*gran_person* >*gran_instance* or *gran_instance* >*gran_group*) a distance of 2 is determined. In case of participant pairs with *gran_person* and *gran_group*, we have a distance of 1. For number agreement we can assign maximally 2 points. If there is number disagreement we assign 1 point. If there is both level type agreement as well as number agreement, a participant pair is given the maximum of 5 points.

As this chapter aims to measure the influence of different event components on event coreference resolution, in STEPS 3C–3E from FIGURE 7.1 we filter our action chains based on time, location and participant compatibility. We use here lemma matches of time, location and participant mentions. For times and locations we do not use the L&C similarity or granularity estimates to avoid matches like *Monday* and *Tuesday*, sharing a short path in the taxonomy and consequently having a high L&C score. The same holds for the grain size. In line with our theoretical approach, it is crucial to filter action chains on disjoint times and locations. It would be interesting to

[7] The similarity score of 0.2 is a proportional score with respect to the maximum score.

[8] Out of the total of 54236 extracted participant mentions.

consider a different treatment for times, and locations expressed by proper names and apply similarity and granularity measurements to time expressions and locations that are not proper names. In reasoning about proper name mentions employing geo- and temporal ontologies could be helpful but this is out of the scope of this study.

Our current approach, depicted in STEPS 3C–3E from FIGURE 7.1, boosts the score of action coreference for each time, location and participant coreference chain they share, taking the coreference score of each chain as a weight for sharing. We used formula 7.2 in which membership to a coreference set of an event is initially based on the coreference score of the action mention but it is strengthened by the proportion that times, locations or participants are shared with other mentions.

$$Coref(m, E) = MAXLC(m, E) + P(t) \vee P(l) \vee P(p) \tag{7.2}$$

E is the set of mentions in action coreference set, *MAXLC* is the highest similarity score for the mention m in the set *E*. The coreference score of action mention *m* equals the sum of the maximum coreference score *MAXLC* proportion *P* of the overlapping participants *p*, of *m* with the other members of the set and the times *t* or locations *l*, with the other members of the set.

7.4 Evaluating the results

For the evaluation, the manual annotations of actions from the ECB 0.1 corpus were used as key chains and were compared with the response chains generated for each topic using the aforementioned heuristics. Since our goal was to evaluate the importance of coreference between times, locations and participants for the task of event coreference resolution, we compare our evaluation results with system results based on action similarity only, i.e. when disregarding other event components. We also aimed at achieving insights into the contribution of different techniques to consider partial coreference of event components (soft matching). This is why we use a lemma baseline (*LmB*) that assigns coreference relation to all nouns and verbs that belong to the same lemma (strict matching).

Heuristic	**Event component**	**MUC**			**B3**			**CEAFm**	**BLANC**			**CoNLL**
		R	P	F	R	P	F	R/P/F	R	P	F	F
LmB	All N&V	63.8	82.8	71.2	65.3	90.6	75.0	65.9	68.0	84.1	71.1	70.7
L&C	Action	69.4	72.4	69.5	69.4	73.3	68.9	58.7	68.6	71.8	67.5	65.2
Action-L&C Time-Lm	Action Time	66.0	77.7	70.6	66.9	84.2	73.6	63.9	68.4	78.1	70.1	69.4
Action-L&C Location-Lm	Action Location	66.3	77.4	70.6	67.4	83.0	73.4	64.1	68.6	77.3	70.0	69.3
Action-L&C Participant-Lm	Action Participant	66.0	78.4	70.8	67.0	84.9	73.9	64.5	68.6	79.0	70.4	69.7
Action-L&C Participant-L&C	Action Participant	65.2	79.4	70.7	66.8	85.7	74.1	64.9	68.5	79.7	70.4	69.8
Action-L&C Participant-granularity	Action Participant	66.5	77.8	70.4	67.6	81.7	72.2	62.5	68.3	77.9	69.4	68.2

Table 7.1: Coreference evaluation results in MUC, B3, CEAFm, BLANC and CoNLL F (macro averages).

Event component	Number of extracted mentions
Participant	54236
Time	3435
Location	5728

Table 7.2: Extraction statistics.

TABLE 7.1 presents coreference evaluation results achieved by means of the different heuristics: the L&C measure, grain size agreement and lemma match (*Lm*) in comparison to the baseline results (*LmB*) in terms of recall (*R*), precision (*P*) and F-score (*F*), employing the commonly used coreference evaluation metrics: MUC (Vilain et al. [1995]), B3 (Bagga and Baldwin [1998]), mention-based CEAF (Luo [2005]), BLANC (Recasens and Hovy [2011]), and CoNLL F1 (Pradhan et al. [2011]).

Compared to the lemma baseline, our approach using similarity of event actions only (see the second row in TABLE 7.1), across the majority of the evaluation metrics improves *R* with up to 6% while loses more than 10% *P*, which was expected. Note, that the lemma baseline achieves remarkably good results, what could be caused by the fact that the annotators pick up on the most obvious coreference cases. Within narrowly defined topics comprised of news articles of the same day describing the same seminal event, events are usually expressed by the same lemma. See section 3.4 on the diversity of the ECB corpus. The ECB corpus has a low lexical and referential diversity. On one hand, events are described with little lexical variation. On the other hand, the topics of the corpus mostly correspond to single seminal events. This means that in the corpus we usually have only one seminal event described per topic. In such a setup, event entities will not seem to be important for event coreference resolution and so distinguishing between different topics will be enough to achieve good evaluation results.

When comparing the contribution of times, locations and participants (all lemma matches for the sake of comparison) with the approach using exclusively action similarity, we see that the approach combining the action and participant components achieved slightly better results (ca. 1% higher precision scores) than the two other approaches employing the time and location components. Altogether, the differences between the scores are hardly meaningful. However these results show that any of the tested entity types improves event coreference resolution with 4 CoNLL-F points. Furthermore, when analyzing these results one must keep in mind that these evaluation scores are conditioned by the fact that participant mentions occur much more frequently in event descriptions than time and location mentions. See TABLE 7.2 with statistics on items extracted from the ECB corpus.

Out of the two different heuristics used in participant approaches, ca. 1% higher F-scores (a 2-4% improvement of precision) on most evaluation metrics were obtained with L&C similarity. Both participant approaches in most metrics improve the F-scores achieved by the action similarity heuristic. The grain size agreement estimates with ca. 1-4% and participant (semantic) similarity with ca. 1-6%. It is notable that the approach using the L&C similarity of participants, in comparison to the action similarity approach, significantly improves precision with ca. 7 points in MUC, ca. 12 points in B3 and ca. 8 points in BLANC. The improvement in precision reflects the added value of participant similarity when solving event coreference.

Compared to the lemma baseline (*LmB*), our best scoring approach of all, that is action similarity with participant similarity, on most metrics loses ca. 1 point on the

F-scores. It gains less than 2 points in recall, while generating output with ca. 4 points lower precision. Again, the small decline in F-measure can be explained by the low lexical and referential diversity of the corpus, as well as by the fact that we are dealing with within-topic coreference (although cross-document). Corpora, even those annotated with cross-document coreference of events, (intentionally) tend to be composed around a number of seminal events, such as attacks or earthquakes. As argued earlier (see also section 3.4), the diversity of event instances from the same type of event that happened in different time frames, locations and with different participants is much lower in the ECB corpus than in the news. The relatively high scores achieved by the lemma baseline show the need for different event coreference data sets, where cross-document coreference is marked in text across different instances of particular event types, e.g. describing two different wars that take place over longer stretches of time and that include similar types of events. Only then the data will become more representative of the sampled population. Within a more realistic setting, when experimenting with a data set that is more representative of event coreference in the news, we expect component coreference to play a much bigger role.

The list below summarizes evaluation results achieved in related work.

- Bejan and Harabagiu [2010]: 83.8% B3 F, 76.7% CEAF F on the ACE (LDC [2005]) data set and on the ECB corpus 90% B3 F, 86.5% CEAF F-score
- Lee et al. [2012]: 62.7% MUC, 67.7% B3 F, 33.9% (entity based) CEAF, 71.7% BLANC F-score on the ECB 0.1 corpus
- Chen et al. [2011]: 46.91% B3 F on the OntoNotes 2.0 corpus

Compared to the above results, our best scoring approach, using action and participant similarity, coreference between actions was solved with an F-score of 70.7% MUC, 74.1% B3, 64.9% CEAFm, 70.4% BLANC F and 69.8 CoNLL F1. Considering that our approach neither considers anaphora resolution nor syntactic features, there is room for improvement of event coreference resolution with an approach that combines these with semantic matches of event components.

7.5 Conclusion

In this chapter, we experimented with our approach to event coreference that employs the importance of coreference, also partial linguistic coreference, between times, locations and participants for the task of event coreference resolution. Our results show that times and entities play a smaller than expected role in event coreference resolution, especially if the data set has low referential diversity (as described in section 3.4.1). Filtering coreferent action candidates based on compatibility of their participants (our best scoring approach) in comparison to the baseline slightly improves precision of the resolution of coreference between events. The results are promising given the limitations of the approach, such as not performing anaphora resolution and considering the limitations of the data set which has a low referential diversity. One could expect that in a more referentially diverse corpus semantic matches of event times and entities would turn out to be far more important for event coreference resolution.

However, further analysis of event descriptions in the news data set shows an even greater complexity of the problem. An additional challenge of using times and entities for event coreference resolution arises from the fact that event descriptions tend to be scattered over multiple sentences of a document. Pieces of event context that were already mentioned are not repeated. For example, if an article describes a concert or a theater performance most probably the date of the concert will be only named once at the beginning of an article. This phenomenon makes it difficult to analyse the contribution of event times and entities to event coreference resolution at the sentence level. This is another explanation for why the importance of times and entities for event coreference resolution appears to be smaller than expected in experiments from this chapter.

Accordingly, to use event times and entities for event coreference resolution, specifically for cross-document coreference, it is best to consider all event information from a document and not be limited only to event descriptions from single sentences. We will investigate this further in chapter 8 where we will experiment with machine learning as a heuristic to identify cross-topic, cross-document event coreference sets. The approach presented in the next chapter considers all event information from the document. Different event components and semantic relations between them are used as features in machine learning.

Chapter 8

Machine learning experiments: The “Bag of Events” approach to event coreference resolution[1]

In chapter 8 we propose a new robust two-step approach to cross-document event coreference resolution on news articles using machine learning. We trained and evaluated event coreference resolvers on the ECB+ corpus presented in chapter 4. This approach makes use of event times and entities from a document to solve coreference between events in the news.

In section 8.1 we explain why the “Bag of Events” approach employs mentions of event components from a unit of discourse for event coreference resolution. In section 8.2 we show how the approach makes use of event structure to solve event coreference. In section 8.3 we delineate the two-step approach. Section 8.4 reports on the experiments with the new method. We compare the results reached by means of our approach to those from related work in section 8.5. We conclude in section 8.6.

8.1 Using entities and discourse structure for event coreference resolution

It is common practice to use information coming from event arguments for event coreference resolution (Humphreys et al. [1997], Chen and Ji [2009a], Chen and Ji [2009b], Chen et al. [2011], Bejan and Harabagiu [2010], Lee et al. [2012], Cybulska and Vossen [2013], Liu et al. [2014] among others). The research community seems to agree that event context information regarding time and location of an event as well as information about other participants play an important role in resolution of coreference between event mentions. Nevertheless, the contribution coming from event arguments as calculated in some studies does not directly translate into some significant increase of coreference resolution scores. Chen and Ji [2009a] report that features related to event arguments only slightly (+2.4% ECM F) improve within-document event coreference. Cybulska and Vossen [2013] note a ca. 4 point CoNLL F-score improvement of within-topic event coreference resolution based on semantic similarity of event arguments. As

[1] The contents of this chapter have been published in Cybulska and Vossen [2015a].

it was demonstrated in chapter 3.4, this is partly due to the lack of diversity in the ECB corpus.

Using entities for event coreference resolution is complicated by the fact that descriptions of events at the sentence level often lack some pieces of information. As pointed out by Humphreys et al. [1997], it could be the case that a lacking piece of information might be available elsewhere within discourse borders. News articles can be seen as a form of public discourse (van Dijk [1988]). As such, the news follows the Gricean Maxim of quantity (Grice [1975]). Journalists do not make their contribution more informative than necessary. This means that some information previously communicated within a unit of discourse will not be mentioned again unless pragmatically required. This is a challenge for models comparing mentions of events (and their arguments) with one another at the sentence level. One would like to be able to fully make use of information coming from event arguments. Instead of looking at event information available within the same sentence, we propose to take a broader look at event mentions surrounding the event mention in question within a unit of discourse. For the purpose of this study, we consider a document (here a news article) to be our unit of discourse.

Our approach to event coreference resolution makes an explicit use of event and discourse structure thereby compensating for implications of the Gricean Maxim of quantity. News follows the principle of language economy. Information tends not to be repeated within a unit of discourse. This phenomenon poses a challenge for models comparing information about event mentions (and their arguments) at the sentence level. Our approach addresses this challenge by building a knowledge representation per unit of discourse - for present purposes, a document. We collect event information from a single document filling a “document template” and by that creating a “Bag of Events”. We then use supervised classification to determine if pairs of document templates contain corefering event mentions. Then we solve coreference between event mentions from the same document cluster by means of supervised classification of “sentence templates”.

8.2 “Bag of Events” - event template approach

The “Bag of Events” is an “event template” approach that employs the structure of event descriptions for event coreference resolution. In the proposed heuristic, event mentions are examined through five slots, as annotated in the ECB+ dataset used in our experiments. The event slots correspond to different elements of event information: an event action and four types of event arguments: time, location, human and non-human participant slots (see chapter 4 for more information on the annotation of the ECB+ corpus). The Bag of Events approach determines coreference between descriptions of events through compatibility of slots of an event template. EXAMPLE 8.2.1 presents an excerpt from topic 1, text number 7 of the ECB corpus (Bejan and Harabagiu [2010]). Consider two event template examples presenting the distribution of event information over the five event slots in the two example sentences in TABLE 8.1.

Example 8.2.1 *The “American Pie” actress has entered Promises for undisclosed reasons. The actress, 33, reportedly headed to a Malibu treatment facility on Tuesday.*

An event template can be filled from different units of discourse, such as a sentence, a paragraph or an entire document. We propose a two-step classification approach to

Event Slot	**Sentence Template 1**	**Sentence Template 2**	**Document Template**
Action	*entered*	*headed*	*entered, headed*
Time	*N/A*	*on Tuesday*	*on Tuesday*
Location	*Promises*	*to a Malibu treatment facility*	*Promises, to a Malibu treatment facility*
Human Part.	*actress*	*actress*	*actress*
Non-human Part.	*N/A*	*N/A*	*N/A*

Table 8.1: Sentence and document templates ECB topic 1, text 7, sentences 1 and 2.

event coreference resolution. In the first step of the approach, an event template is filled per document; this is a "document template". In the second step of the approach, per action mention, a "sentence template" is filled based on information from the sentence and coreference is solved between event mentions within document clusters created in step 1.

By filling in a document template, one creates a Bag of Events per document. Bag of Events features are then used in supervised classification. This heuristic employs clues coming from discourse structure, implied by discourse borders. Descriptions of different event mentions occurring within a discourse unit, whether coreferent or related in some other way, unless stated otherwise, tend to share their context. In EXAMPLE 8.2.1 the first sentence reveals that an actress has entered a rehab facility called Promises. From the second sentence the reader finds out where the facility is located and when the actress headed there. It is clear to the reader of the text fragment from EXAMPLE 8.2.1 that both event mentions from sentence one and two, happened on Tuesday. Also both sentences mention the same rehab center in Malibu. These observations are crucial for the Bag of Events approach.

Our two-step classification approach replaces the typically used in coreference resolution topic classification step with document template classification. Classifying document templates allows for more specific event context disambiguation also within the same topic. We delineate the two steps of the Bag of Events approach to event coreference resolution in section 8.3.

8.3 Two-step "Bag of Events" approach

We present a robust two-step approach to cross-textual event coreference resolution on news articles that explicitly employs event and discourse structure to account for implications of Gricean maxim of quantity in the news. The first step in this approach is to build a knowledge representation by filling in an event template per unit of discourse, here, a document. We collect all manually annotated event action, location, time, human and non-human participant mentions from a single document and we fill in a document template (as depicted in TABLE 8.1). We determine whether pairs of document templates contain any corefering event mentions by means of supervised classification. In the second step, we use supervised classifiers to solve coreference between pairs of event mentions within clusters of document templates as determined in step 1. For the purpose of this task, an event template is filled again but this time, it is a "sentence template" which gathers event information from the sentence per action mention. Supervised classifiers solve coreference between pairs of event mentions and finally pairs sharing common mentions are chained into coreference clusters. These two steps are described in more detail in 8.3.1 and 8.3.2. FIGURE 8.1 depicts the architecture of the two-step approach when processing the training data. FIGURE 8.2 illustrates the processing of the test set.

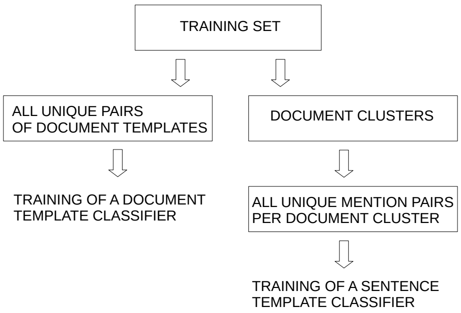

TRAINING SET
ALL UNIQUE PAIRS OF DOCUMENT TEMPLATES
DOCUMENT CLUSTERS
TRAINING OF A DOCUMENT TEMPLATE CLASSIFIER
ALL UNIQUE MENTION PAIRS PER DOCUMENT CLUSTER
TRAINING OF A SENTENCE TEMPLATE CLASSIFIER

Figure 8.1: Training set processing.

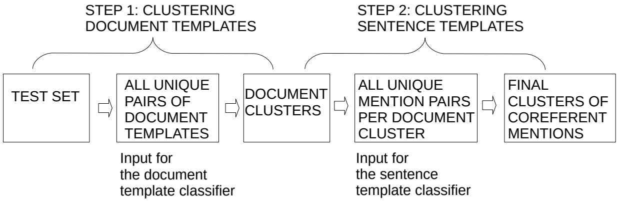

STEP 1: CLUSTERING DOCUMENT TEMPLATES
STEP 2: CLUSTERING SENTENCE TEMPLATES
TEST SET
ALL UNIQUE PAIRS OF DOCUMENT TEMPLATES
DOCUMENT CLUSTERS
ALL UNIQUE MENTION PAIRS PER DOCUMENT CLUSTER
FINAL CLUSTERS OF COREFERENT MENTIONS
Input for the document template classifier
Input for the sentence template classifier

Figure 8.2: Test set processing.

8.3.1 Step 1: clustering document templates

The first step in this approach is to fill in a document template. We create a document template by collecting mentions of the five event slots: action, time, location, human and non-human participant from a single document. In a document template there is no distinction made between pieces of event information coming from different sentences of a document and no information is kept about elements being part of different mentions. A document template can be seen as a Bag of Events. The template stores a set of unique lemmas per event slot.

On the training set of the data, we train a pairwise binary classifier to determine whether two document templates share corefering event mentions. This is a supervised learning task in which we determine "compatibility" of two document templates if any two mentions from those templates were annotated in the corpus as coreferent.

Let *m* be an event mention, and *doc* a collection of mentions from a single document template such that $m\imath : 1 \leq \imath \leq doc\jmath$ where $\imath$ is the index of a mention and $\jmath$ indexes document templates; $doc\jmath : 1 \leq \jmath \leq DOC$ where *DOC* are all document templates from the corpus. Let *m_a* and *m_b* be mentions from different document templates. “Compatibility” of a pair of document templates $(doc\jmath, doc\jmath_{+1})$ is determined based on coreference of any mentions $(m\_a\imath, m\_b\imath)$ from a pair of document templates, see formula 8.1.

$$coref(\exists m\_a\imath \in doc\jmath, \exists m\_b\imath \in doc\jmath_{+1}) \implies \textit{compatibility}(doc\jmath, doc\jmath_{+1}) \tag{8.1}$$

On the training data, we train a binary decision-tree classifier (hereafter *DT*) to find pairs of document templates containing corefering event mentions.

After all unique pairs of document templates from the test set are classified by means of the DT document template classifier, “compatible” pairs are merged into document clusters based on pair overlap.

8.3.2 Step 2: clustering sentence templates

The aim of the second step is to solve coreference between event mentions from document clusters which are the output of the classification task from step 1. We experiment with a supervised decision tree classifier. Pairs of sentence templates are considered in the classification task.

A sentence template (see examples in TABLE 8.1) is created for every event action mention in the data set. All unique pairs of event mentions (and their sentence templates) are generated within clusters of documents sharing corefering event mentions in the training set. Pairs of sentence templates, which translate into features indicating compatibility across five event template slots, are used to train a decision tree sentence template classifier.

On the test set part of the data, after output clusters of the document template classifier from step 1 are turned to mention pairs (all unique action mention pairs within a document cluster), pairs of sentence templates are classified by means of the DT sentence template classifier. To identify the final clusters of coreferent event mentions, within each document cluster, event mentions are grouped into equivalence classes based on corefering pair overlap.

ECB+	**No.**
Topics	43
Texts	982
Action mentions	6833
Location mentions	1173
Time mentions	1093
Human participant mentions	4615
Non-human participant mentions	1408
Coreference chains	1958

Table 8.2: ECB+ statistics.

8.4 Experiments

8.4.1 Corpus

For the experiments we used true mentions from the ECB+ corpus (Cybulska and Vossen [2014b]) described in detail in chapter 4, which is an extended and re-annotated version of the ECB corpus (Bejan and Harabagiu [2010]). The ECB+ corpus contains a new corpus component, consisting of 502 texts, describing different instances of event types that were already captured by the 43 topics of the ECB. As recommended by the authors in the release notes, for experiments on event coreference resolution we used a subset of ECB+ annotations (based on a list of 1840 selected sentences), that were additionally reviewed with focus on coreference relations. TABLE 8.2 presents information about the data set used for the experiments. We divided the corpus into a training set (topics 1-35) and test set (topics 36-45).

8.4.2 Experimental set up

The ECB+ texts are available in the XML format. The texts are tokenized, so no sentence segmentation nor tokenization needed to be done. We POS-tagged (for the purpose of proper verb lemmatization) and lemmatized the corpus sentences. For the experiments we used tools from the Natural Language Toolkit (**?**, NLTK version 2.0.4): the NLTK’s default POS tagger, and WordNet lemmatizer[2] as well as WordNet synset assignment by the NLTK.[3] For machine learning experiments we used scikit-learn (Pedregosa et al. [2011]).

In the experiments, different features were assigned values per event slot (see TABLE 8.3). The lemma overlap feature (L) expresses a percentage of overlapping lemmas between two instances of an event slot, if instantiated in the sentence (with the exclusion of stop words). As the relation between an event and involved entities is not annotated in ECB+, one frequently ends up with multiple entity mentions from the same sentence for an action mention. All entity mentions from the sentence are considered in the overlap calculations. There are two features indicating event mentions’ location within discourse (D), specifying if two mentions come from the same sentence and the same document. Action similarity (A) was calculated for a pair of active action mentions using the Leacock and Chodorow measure (Leacock and Chodorow [1998]). Per entity slot (time, location, human and non-human participant) we checked if there is any coreference relation annotated in the corpus between entity mentions from the sentence for the two compared event actions. We used cosine similarity to express this feature (E).[4] For all five slots a percentage of synset overlap is calculated (S). In case of document templates features referring to active action mentions were disregarded. Instead action mentions from a document were considered. All feature values were rounded to the first decimal point.

We experimented with a few feature sets, considering per event slot lemma features only (L), or combining them with other features described in TABLE 8.3. Before fed to

[2] www.nltk.org/_modules/nltk/stem/wordnet.html

[3] `http://nltk.org/_modules/nltk/corpus/reader/wordnet.html`

[4] We express this feature in the form of the cosine similarity to capture the situation where there are multiple e.g. human participant mentions in one sentence and only one in another as opposed to comparing sentences with multiple matching human participant mentions. Note that the relation between an action mention and the involved entities is not annotated in ECB+ so considering all entity mentions from the sentence, it is common to have multiple entity mentions per action.

Event Slot	**Mentions**	**Feature Kind**	**Explanation**
Action	Active mentions	Lemma overlap (L) Synset overlap (S) Action similarity (A) Discourse location (D) - document - sentence	Numeric feature: overlap % Numeric: overlap % Numeric: L&C Binary: - the same document or not - the same sentence or not
	Sent. or doc. mentions	Lemma overlap (L) Synset overlap (S)	Numeric: overlap % Numeric: overlap %
Time	Sent. or doc mentions	Lemma overlap (L) Entity coreference (E) Synset overlap (S)	Numeric: overlap % Numeric: cosine similarity Numeric: overlap %
Location	Sent. or doc mentions	Lemma overlap (L) Entity coreference (E) Synset overlap (S)	Numeric: overlap % Numeric: cosine similarity Numeric: overlap %
Human Participant	Sent. or doc mentions	Lemma overlap (L) Entity coreference (E) Synset overlap (S)	Numeric: overlap % Numeric: cosine similarity Numeric: overlap %
Non-Human Participant	Sent. or doc mentions	Lemma overlap (L) Entity coreference (E) Synset overlap (S)	Numeric: overlap % Numeric: cosine similarity Numeric: overlap %

Table 8.3: Features grouped into four categories: L-Lemma based, A-Action similarity, D-location within Discourse, E-Entity coreference and S-Synset based. L&C stands for Leacock and Chodorow [1998].

a classifier, missing values were imputed (no normalization was needed for the scikit-learn DT algorithm). All classifiers were trained on an unbalanced number of pairs of document or sentence templates from the training set. We used grid search with ten fold cross-validation to optimize the hyper-parameters (maximum depth, criterion, minimum samples leafs and split) of the decision-tree algorithm.

8.4.3 Baseline

We will consider two baselines: a singleton baseline and a rule-based lemma match baseline. The singleton baseline considers event coreference evaluation scores generated taking into account all event mentions as singletons. In the singleton baseline response there are no “coreference chains” of more than one element. The rule-based lemma baseline generates event mention coreference clusters based on full overlap between lemma or lemmas of compared event triggers (action slot) from the test set. As we have seen in section 4.8.2, the average lexical diversity of the ECB+ is 35% higher than the ALD of the ECB 0.1 (the previous version of the data set) but with an ALD of 53% the ECB+ is still not very diverse lexically so one can expect that the lemma baseline will remain strong as it was the case for ECB 0.1 (see chapter 3.4.3).

Baseline	**MUC**			**B3**			**CEAF**	**BLANC**			**CoNLL**
(BL)	R	P	F	R	P	F	R/P/F	R	P	F	F
Singleton BL	0	0	0	45	100	62	45	50	50	50	39
Action Lemma BL	71	60	65	68	58	63	51	65	62	63	**62**

Table 8.4: Baseline results: singleton baseline and lemma match of event triggers evaluated in MUC, B3, mention-based CEAF, BLANC and CoNLL F.

TABLE 8.4 presents the baselines’ results in terms of recall (R), precision (P) and F-score (F) by employing the coreference resolution evaluation metrics: MUC (Vilain et al. [1995]), B3 (Bagga and Baldwin [1998]), CEAF (Luo [2005]), BLANC (Recasens and Hovy [2011]), and CoNLL F1 (Pradhan et al. [2011]). When discussing event coreference scores it must be noted that some of the commonly used metrics depend on the evaluation data set, with scores going up or down with the number of singleton items in the data (Recasens and Hovy [2011]). Our singleton baseline gives zero scores in MUC, which is understandable due to the fact that the MUC measure promotes longer chains. B3, on the other hand, gives additional points to responses with more singletons, hence the remarkably high scores achieved by the baseline in B3. CEAF and BLANC as well as the CoNLL measures (the latter being an average of MUC, B3 and entity CEAF) give more realistic results. The lemma baseline reaches 62% CoNLL F1. A baseline only considering event triggers will allow for an interesting comparison with our event template approach, employing event argument features.

8.4.4 Evaluation

TABLE 8.5 evaluates the final clusters of corefering event action mentions produced in the experiments by means of the DT algorithm when employing different features.

The best coreference evaluation scores with the highest CoNLL F-score of 73% and BLANC F of 72% were reached by the combination of the document template classifier using feature set L across event slots and the sentence template classifier when employing features LDES (features generated based on L-Lemma overlap, D-location within Discourse, E-Entity coreference and S-Synset overlap; see TABLE 8.3 for feature description). Adding action similarity (A) on top of LDES features does not make any difference on decision tree classifiers with a maximum depth of 5. Our best CoNLL F-score of 73% is an 11 point improvement over the strong (rule-based) event trigger lemma baseline, and a 34 point increase over the singleton baseline.

Step1			**Step2**			**MUC**			**B3**			**CEAF**	**BLANC**			**CoNLL**
Alg	Slot Nr	Feats	Alg	Slot Nr	Feats	R	P	F	R	P	F	F	R	P	F	F
-	-	-	DT	5	L	61	76	68	66	79	72	61	67	69	68	70
DT	5	L	DT	5	L	71	75	73	71	77	74	64	71	71	71	73
DT	5	L	DT	5	LDES	71	75	73	71	78	74	64	72	71	**72**	**73**
DT	2	L	DT	2	LDES	76	70	73	74	68	71	61	74	68	70	70
DT	5	L	DT	5	LADES	71	75	73	71	78	74	64	72	71	72	73

Table 8.5: Bag of Events approach to event coreference resolution, evaluated in MUC, B3, mention-based CEAF, BLANC and CoNLL F on the ECB+ corpus.

To quantify the contribution of document templates, we contrast the results of the two-step Bag of Events approach with scores achieved when skipping step 1, that is, without the initial classification of document templates. The results obtained with sentence template classification only give us some insights into the impact of the document template classification step. Note that the sentence template classification without preliminary document template clustering is computationally much more expensive than the two-step template approach, which ultimately takes into account significantly less item pairs owing to the initial document template clustering. In the one-step approach, the DT sentence template classifier using lemma features (L), when trained on an unbalanced training set, reaches 70% CoNLL F. This is 8% better than the strong lemma baseline disregarding event arguments, but only 3% less than the two-step Bag of Events approach with the two classifiers trained on lemma features (L). The reason for the relatively small improvement by the document template classification step could arise from the fact that in the ECB+ corpus, few sentences are annotated per text. 1840 sentences are annotated in 982 corpus texts, i.e. 1.87 sentence per text. We expect that the impact of document templates would be bigger if more event descriptions from a discourse unit were taken into account than only the ground truth mentions.

We ran an additional experiment with the four entity types bundled into one entity slot. Locations, times, human and non-human participants were combined into a cumulative entity slot resulting in a simplified two-slot template. When using two-slot templates for both, document and sentence classification on the ECB+ 70% CoNLL F score was reached. This is 3% less than with five-slot templates.

8.5 Related work

To the best of our knowledge, the only related study using clues coming from discourse structure for event coreference resolution was done by Humphreys et al. [1997] who perform coreference merging between event template structures. Both approaches determine event compatibility within a discourse representation but we achieve that with a much more restricted template (five slots only) which facilitates merging of all event and entity mentions from a text as the starting point. Humphreys et al. [1997] consider discourse events and entities for event coreference resolution while operating on the level of mentions.

Approach	**Data**	**Model**	**MUC**			**B3**			**CEAF entity**	**BLANC**			**Co-NLL**
			R	P	F	R	P	F	F	R	P	F	F
B&H	ECB annotated by Bejan and Harabagiu [2010]	HDp	52	90	66	69	96	80	71	NA	NA	NA	NA
LEE	ECB 0.1 annotated by Lee et al. [2012]	LR	63	63	63	63	74	68	34	68	79	72	55
BOE-2	ECB annotated by Cybulska and Vossen [2014b]	DT+DT	65	59	62	77	75	76	72	66	70	67	70
BOE-5	ECB annotated by Cybulska and Vossen [2014b]	DT+DT	64	52	57	76	68	72	68	65	66	65	66
BOE-2	ECB+ annotated by Cybulska and Vossen [2014b]	DT+DT	76	70	73	74	68	71	67	74	68	70	70
BOE-5	ECB+ annotated by Cybulska and Vossen [2014b]	DT+DT	71	75	73	71	78	74	71	72	71	72	**73**

Table 8.6: The Bag of Events (BOE) approaches evaluated on ECB and ECB+ in MUC, B3, entity-based CEAF, BLANC and CoNLL-F in comparison with related studies. Note that the BOE approaches use gold and related studies system mentions.

Some of the metrics used to score event coreference resolution are dependent on the number of singleton events in the evaluation data set (Recasens and Hovy [2011]). Thus, for the sake of a meaningful comparison, it is important to consider similar data sets. The ECB and ECB+ are the only available resources annotated with both within- and cross-document event coreference. To the best of our knowledge, no baseline has been set yet for event coreference resolution on the ECB+ corpus. Accordingly, in TABLE 8.6 we also look at results achieved on the ECB corpus. The ECB is a subset of ECB+, and so the closest to the data set used in our experiments but capturing less ambiguity of the annotated event types (Cybulska and Vossen [2014b]). We focus on the CoNLL F measure that was used for comparison of competing coreference resolution systems in the CoNLL 2011 shared task.

The best results of 73% CoNLL F were achieved on the ECB+ by the Bag of Events approach using five-slot event templates (*BOE-5* in TABLE 8.6). When using two-slot templates we get 3 points less CoNLL F on ECB+. For the sake of comparison, we ran an additional experiment on the ECB part of the corpus (annotation by Cybulska and Vossen [2014b]). The ECB was used in related work although with different versions of annotation, so experiments on the ECB are not entirely comparable. We ran two tests, one with the simplified templates considering only the action and entity slot (as annotated in the ECB by Lee et al. [2012]) and one with five-slot templates. The two-slot Bag of Events (*BOE-2*) on the ECB part of the corpus reached results comparable to related works: 70% CoNLL F, while the five-slot template experiment (*BOE-5*) results in 66% CoNLL F. The approach of Lee et al. [2012] (in TABLE 8.6 *LEE*) using linear regression (in TABLE 8.6 *LR*) reached 55% CoNLL F although on a much more difficult task entailing event extraction as well. The component similarity method of Cybulska and Vossen [2013] resulted in 70% CoNLL F but on a simpler within-topic task (not considered in TABLE 8.6). *B&H* in TABLE 8.6 refers to the approach of Bejan and Harabagiu [2010] using hierarchical Dirichlet process (*HDp*). For this study no CoNLL F was reported. In the *BOE* experiments reported in TABLE 8.6, during step 1 only lemma features L were used and for sentence template classification (step 2) LDES features were employed. In all tests with the Bag of Events approach, ground truth mentions were used.

8.6 Conclusion

This chapter presents a two-step Bag of Events approach to event coreference resolution. Instead of performing topic classification before solving coreference between event mentions as is done in most studies, this two-step approach first compares document templates created per discourse unit. Only after does it compare single event mentions and their arguments. In contrast to a heuristic using a topic classifier which might have problems distinguishing between multiple instances of the same event type, the Bag of Events approach facilitates context disambiguation between event mentions from different discourse units. Grouping events depending on compatibility of event context (time, location and participant) on the discourse level allows one to take advantage of event context information, which is mentioned only once per unit of discourse and consequently is not always available on the sentence level. From the perspective of performance, the robust Bag of Events approach using a small feature set also significantly restricts the number of compared items. Therefore, it has much lower memory requirements than a pairwise approach operating on the mention level. Given that this approach does not consider any syntactic features and that the evaluation data set is only annotated with 1.8 sentences per text, the evaluation results are highly encouraging.

Chapter 9

Conclusions and recommendations

This dissertation investigated the topic of cross-document coreference between events in news articles. Modelling the complex phenomenon of event coreference in the news and designing data sets meant for studies of event coreference is not straightforward. The four parts of this dissertation look at the different aspects of the main research problems by defining four research sub-questions. Parts I and II focus on data used to study event coreference and parts III and IV contribute to modelling the event coreference phenomenon. We will now go through research questions addressed in each part of the dissertation.

Part I: Are the existing corpora representative of the event coreference problem?

- Do the data sets annotated with event coreference reflect the natural diversity of news articles?
- What are the requirements for a data set for experiments on unrestricted cross-document event coreference?
- Is there an English data set that fulfils on the requirements?

In part I we investigated how representative of cross-document event coreference in the news are the available data sets. In chapter 3 we evaluated a number of event coreference resources. The only corpus annotated with cross-document event coreference and so the most appropriate for our research was the ECB corpus (Bejan and Harabagiu [2010]).

We first set some requirements for a data set used to experiment with unrestricted cross-document event coreference. An event coreference corpus must be diverse with regards to event instances covered per event type and with regards to formulations used to describe events. We introduced a methodology to quantify the referential and the lexical diversity of a coreference resource.

Then we evaluated whether the ECB fulfilled on the requirements. In the context of the task of event coreference resolution, we analyzed the diversity of the ECB as a sample of the language population of news articles. We calculated the average referential diversity and the average lexical diversity of the ECB 0.1 data. Both the ARD and ALD scores for ECB 0.1 are very low. The ARD of the ECB 0.1 corpus amounts to

17% while the ALD is 18%. These low diversity scores quantify the low complexity of the data set from the point of view of event coreference. In most cases the ECB covers one seminal event per domain, which considerably simplifies the referential diversity. Also the lexical diversity of the ECB is very low. Accordingly, results of tests on event coreference resolution obtained on the ECB corpus cannot be generalized onto the language population of news which is expected to be much more referentially and lexically diverse.

Part II: How can we make a corpus more representative of the event coreference problem?

- How can one obtain an empirically valid data set on event coreference in the news that is representative of the language population of news articles?
- How should a data set that is meant for research on event coreference be organized?
- How should one reflect the lexical and the referential diversity of news articles in a corpus?
- What is the importance of entities for event coreference resolution and how should an event coreference corpus be annotated to facilitate research on the topic?

In part II of the dissertation we explored how a data set can be made more representative of the event coreference problem in the news. We worked with the ECB corpus to increase the referential diversity of the data set so that it is more representative of event coreference in news articles. We extended the corpus with a new corpus component of 502 texts covering new instances from event types covered in the ECB. In chapter 4, we presented the new resource called ECB+. The ECB+ extension of the ECB was purposefully targeting the diversity of event times, locations and participants per event type. We artificially diversified instances of event types described in the ECB corpus so that we could obtain an empirically valid data set on event coreference in the news that is naturally filled with descriptions of multiple instances of an event type that happened at different times, locations and with different participants involved.

We evaluated the contribution of the ECB+ by measuring the referential and lexical diversity and comparing the ARD and the ALD scores with those achieved on the ECB in chapter 3. The average referential and lexical diversity have both increased significantly from the ECB 0.1 to the ECB+. The complexity of the data set has increased. The average referential diversity of the ECB+ is 46% and the average lexical diversity of the ECB+ amounts to 53%. The ARD of the ECB 0.1 corpus is 17%. This indicates that our extension of the ECB into ECB+ has increased the ARD of the corpus with 29 points. The ALD estimate of the ECB corpus is 18%, so transforming the ECB 0.1 into the ECB+ has increased the ALD of the corpus with 35 points. This is a big improvement for both kinds of diversity, especially considering that the ECB+ extension primarily aimed to increase the referential diversity of the corpus. The ALD scores could become higher if the corpus was augmented with the objective to increase its lexical diversity. For example, if one added a new layer (or layers) of seminal events to the data set with new event descriptions searched for through synonyms of the event descriptions already covered by the corpus.

By increasing the diversity of the coreference resource we made it more representative of the population of news articles on the web, where one can find descriptions of multiple event instances from an event type. Training and testing coreference resolvers on the ECB+ makes the task of coreference resolution more complex. A system has to distinguish between at least two event instances from an event type. In the ECB, grouping events into coreference chains in most cases would come down to distinguishing between different event types e.g. between an arrest of a suspect, a bombing, or an earthquake. If working with the ECB+ corpus, a more fine-grained distinction must be made between e.g. arrests of two different suspects (ECB+ topic 35) or between two earthquakes that happened in the same country but at a different time (ECB+ topic 37).

The ECB+ corpus covers at least two event instances per topic. This is an improvement compared to the ECB, but still far from the multitude of event instances described in daily news. Per event type one at least would want to cover descriptions of event instances that differ with regard to every event component to ensure that coreference resolvers learn to distinguish between event instances that happened at different times or places or with different participants. The ECB+ is not as referentially diverse as one may desire. The diversity of event instances in descriptions online could be significantly higher than two or even a few instances per event type. For example, consider hundreds of earthquakes that can happen monthly according to the USGS. Ideally, for coreference experiments one would prefer a corpus that covers multiple layers of event instances from multiple event types. A diachronic corpus across different topics would be ideal. A coreference corpus covering at least more than one instance per event type seems like the absolute minimum for coreference experiments to give meaningful results. The ECB+ is a step in the right direction and is a starting point for future corpus extensions.

Part III: How can we model the gradable event coreference phenomenon?

- Is there a correlation between the temporal perspective of the writer and the variation in language use?
- Does the time of writing correlate with event granularity?
- How can we model the relationship between granularity and event coreference?
- Can granularity of event times, locations and participants be automatically determined?
- How can we capture and formalize the interplay between different semantic relations and event coreference?

In part III we researched how descriptions of events are realized in texts with different temporal perspectives on the described events and what the implications are for modeling the phenomenon of gradable event coreference.

In chapter 5 we performed a study in which we explored differences between event descriptions in news articles lacking or having a shorter temporal perspective and in texts that are written with a longer temporal perspective on an event such as Wikipedia articles. Event descriptions in text differ in granularity. Events at a coarse granularity level, such as *war*, are more general and abstract with a longer time span and with group participants. Events at a fine granularity level, for instance a *shooting* event, are comparatively specific with a shorter duration, and individual participants. A statistical analysis confirmed our hypothesis that news texts written shortly after an event happened give a detailed account of a story describing a higher number of events at a finer granularity level. Event descriptions in the news are more conceptually diverse and formulations used are more unique than in texts describing the same events but written from a longer temporal perspective. With the passage of time, writers have more knowledge about an event. The reason why things happen or who is behind what is happening become clear so events can be explained. Texts having a longer temporal perspective on the described events abstract from details and describe a smaller number of events at a coarse granularity level. There is less concept diversity in event descriptions as the focus shifts to the interpretation of events.

The change from a detailed to a more general account of events is gradual. With the growing temporal perspective, the granularity of event times, locations and human participants grows. The described event participants change from individuals to group participants, event locations transform from small to bigger areas and event times become longer periods of time. Granularity of event times, locations and human participants could be used as a clue in determining event relations automatically. There is an expected correlation between agreement or disagreement in grain-size and the notion of coreference. In the prototypical case, agreement or small granularity differences are expected to indicate coreference. A greater distance in granularity is expected to be a negative indicator of coreference but it could indicate other event relations like scriptal or event membership.

We made an attempt at capturing the degree of granularity of events explicitly for the purpose of usage in NLP applications. We created a granularity taxonomy that makes it possible to automatically determine and to distinguish between coarse and fine granularities of event times, locations and human participants. The intrinsic, conceptual granularity is captured by means of a number of granularity levels defined in the granularity taxonomy. It would be interesting to look at possibilities of extending the taxonomy by learning granularity levels from corpora to overcome the low coverage limitation following from the usage of a WordNet-based resource. One could also augment the taxonomy to cover the non-human participant slot and experiment with ways to represent event action granularity.

In chapter 6 we proposed a model of gradable event coreference. The model captures the relationship between event granularity and hyponymy and event coreference based on semantic relations between mentions of event components. Semantic relations – hyponymy and meronymy – together with other coreference indicators such as repetition, synonymy, anaphora and disjunction are indicative of event coreference. The proposed model (provided in section 6.3) does not see event coreference as an absolute, clear-cut notion. Instead it models event coreference as a gradual phenomenon. It employs the importance of full and partial linguistic coreference between events and their times, locations and participants. In part IV we experimented with the model.

Part IV: What is the role of times and entities in event coreference resolution?

- Do times and entities matter in solving event coreference?
- Can we measure the contribution of times and entities to event coreference resolution?
- How is the event information packaged in a typical news text?
- What is the optimal way to make use of times and entities for event coreference resolution?

In part IV we deliberated about the role of event times and entities for event coreference resolution. To better understand the role of times and entities in event coreference resolution, we performed a rule-based experiment in chapter 7 and we used machine learning techniques in chapter 8.

The main goal of the rule-based experiment from chapter 7 was to measure the contribution of different components of event descriptions to the task of event coreference resolution. Following Quine [1985], one would expect that event context information is crucial for solving event coreference. However one would not see this hypothesis confirmed when looking at test results of computational studies on event coreference resolution (see section 8.1). With our experiment in chapter 7 we tried to calculate the contribution of different entity types to event coreference. Filtering coreferent action candidates based on compatibility of their participants (our best scoring approach) in comparison to the baseline only slightly improves precision of the resolution of coreference between events. The results indicate that entities play a smaller than expected role in event coreference resolution. We hypothesize that the results of such an experiment are not reliable if the corpus is not representative of the researched phenomenon of event coreference resolution. In our case, the data set has a low referential diversity (see section 3.4.1) and so it is not representative of event coreference in news articles. We expect that in a more referentially diverse corpus semantic matches of event times and entities would turn out to be more important for event coreference resolution.

Further analysis of event descriptions in the news data shows an even bigger complexity of the problem. Event information is spread over the entire document. Event descriptions tend to be scattered over multiple sentences of a text. According to the Gricean Maxim of quantity (1975), pieces of information that were already mentioned in a text are not repeated. For example, if an article describes a concert or a theater performance the date of the concert will likely be only named once at the beginning. This phenomenon makes it difficult to use the full potential of information coming from entities for event coreference resolution and it makes it impossible to analyse the contribution of event times and entities to event coreference resolution when operating at the sentence level. This is a possible explanation of why the importance of event entities for event coreference resolution might appear to be smaller than expected. Furthermore, sometimes information is not only missing at the sentence level but it might not be stated explicitly at all in a text. For example, in case of named entity events such as *World War II*, the context information is common knowledge. These are the main challenges that we encountered when trying to estimate the contribution of times and entities to event coreference resolution.

We considered these observations in the design of experiments from chapter 8 where we used machine learning as a heuristic to solve cross-topic, cross-document event coreference on the ECB+ corpus. We proposed a new two-step Bag of Events approach to cross-document event coreference resolution on news articles that aggregates event information per discourse unit before solving event coreference at the sentence level. This way we are able to make proper use of all event entities and times available in a text for event coreference resolution. To solve event coreference, the Bag of Events approach uses all event information available in a document. Different event components and semantic relations between them are used as features for machine learning.

The Bag of Events approach, instead of performing topic classification before solving coreference between event mentions as is done in most approaches, first compares document templates created per discourse unit. Only after does it compare single event mentions and their arguments at the sentence level and that is only if the event mentions come from compatible discourse units. In contrast to a heuristic using a topic classifier, which might have problems distinguishing between multiple instances of the same event type, the Bag of Events approach facilitates context disambiguation between event mentions from different discourse units. Grouping events depending on compatibility of event context (time, location and participants) on the discourse level, allows one to take advantage of event context information, which is mentioned only once per unit of discourse and consequently is not always available at the sentence level.

In chapter 8 we trained and evaluated event coreference resolvers on the ECB+ corpus. Given that the evaluation data set is only annotated with 1.8 sentences per text, the evaluation results are highly encouraging. We expect that in a corpus with a higher amount of sentences annotated per text, aggregating all event information per document will play an even more important role in event coreference resolution.

Event coreference resolution is a complex problem. The four parts of the dissertation contribute to a better understanding of how the diversity of event coreference in the news can be sampled and modelled for the purpose of event coreference resolution. The first two parts of the dissertation set clear directions for creation of corpora annotated with unrestricted cross-document event coreference that would be representative of coreference in the news. The representativeness of event coreference corpora is strictly dependent on sampling of the referential and lexical diversity of event coreference. Ideally, one would like to work towards a diachronic open domain corpus covering a multitude of topics and a multitude of seminal events per topic. The first step in that direction was made with the creation of the ECB+ corpus.

The last two parts of the dissertation deliberate on the gradable phenomenon of event coreference and on the role of event components in event coreference resolution. Event coreference gradually turns to other event relations. A successful approach to event coreference resolution must consider not only the clear-cut cases of full linguistic coreference but also make use of partial coreference between mentions of event times and entities. Determining whether mentions are compatible with regard to granularity and hyponymy can be of help here.

A measurement of the contribution by the different event components to event coreference resolution is not a straightforward task due to the fact that event information is often incomplete, either because it is scattered over the entire document or because it is absent from the text. Additionally, measuring the contribution of the different components to event coreference resolution would be easier on a data set including annotation of links between components of one event, which is not part of any cross-document annotation efforts to date.

The Bag of Events approach to event coreference resolution proposed in this dissertation makes use of partial coreference between event times and entities and it considers the implications following from the Gricean Maxim of quantity. But the approach was not tested in an end-to-end setting that would incorporate extraction of mentions by the system. It would be advisable to evaluate the Bag of Events approach on top of mention extraction and to compare the results with state-of-the-art systems. Additionally, it would be interesting to compare the first step used in the approach, which looks for compatible document templates, with the results achieved by a state-of-the-art topic classifier in the context of event coreference resolution. Such comparisons would be most relevant on a corpus that is representative of the news and covers multiple seminal events per topic, reflecting the complexity of the event coreference phenomenon.

Bibliography

E. Agirre and A. Soroa. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics, (EACL-2009)*, 2009.

James Allan. Topic detection and tracking: Event-based information organization. 2002.

Jun Araki, Lamana Mulaffer, Arun Pandian, Yukari Yamakawa, Kemal Oflazer, and Teruko Mitamura. Interoperable annotation of events and event relations across domains. In *Proceedings of the 14th Workshop on Interoperable Semantic Annotation.*, 2018.

Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 1998.

Collin F. Baker, Charles J. Fillmore, and Beau Cronin. The structure of the framenet database. In *International Journal of Lexicography*, volume 16.3, pages 281–296, 2003.

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. CAT: the CELCT Annotation Tool. In *Proceedings of LREC 2012*, 2012.

Cosmin Adrian Bejan and Sanda Harabagiu. A linguistic resource for discovering event structures and resolving event coreference. In *In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

Cosmin Adrian Bejan and Sanda Harabagiu. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 2010.

Cosmin Adrian Bejan, Matthew Titsworth, Andrew Hickl, and Sanda Harabagiu. Nonparametric bayesian models for unsupervised event coreference resolution. In *Advances in Neural Information Processing Systems 22*, pages 73–81, 2009.

Luisa Bentivogli, Christian Girardi, and Emanuele Pianta. In *Proceedings of the LREC 2008 Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management*, 2008.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media Inc., http://nltk.org/book, 2009.

W. Bosma and P. Vossen. Bootstrapping language neutral term extraction. In *Proceedings of the 7th international conference on Language Resources and Evaluation, (LREC2010), Valletta, Malta*, 2010.

W. Bosma, P. Vossen, A. Soroa, G. Rigau, M. Tesconi, A. Marchetti, M. Monachini, and C. Apiprandi. Kaf: a generic semantic annotation format. In *Proceedings of the GL2009 Workshop on Semantic Annotation, Pisa, Italy*, 2009.

Michael Bugert, Nils Reimers, and Iryna Gurevych. Cross-document event coreference resolution beyond corpus-tailored systems. *arXiv preprint arXiv:2011.12249*, 2020.

Tommaso Caselli and Piek Vossen. The storyline annotation and representation scheme (StaR): a proposal. In *Proceedings of the 2nd Workshop on Computing News Storylines*, pages 67–72, 2016.

Bin Chen, Jian Su, Sinno Jialin Pan, and Chew Lim Tan. A unified event coreference resolution by integrating multiple resolvers. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, November 2011.

Zheng Chen and Heng Ji. Event coreference resolution: Feature impact and evaluation. In *Proceedings of Events in Emerging Text Types (eETTs) Workshop*, 2009a.

Zheng Chen and Heng Ji. Graph-based event coreference resolution. In *TextGraphs-4 Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages Pages 54–57, 2009b.

Prafulla Kumar Choubey and Ruihong Huang. Event coreference resolution by iteratively unfolding inter-dependencies among events. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2124–2133, 2017.

J. Cohen. The coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 21(1):37–46, 1960.

Agata Cybulska and Piek Vossen. Event models for historical perspectives: Determining relations between high and low level events in text, based on the classification of time, location and participants. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, 2010.

Agata Cybulska and Piek Vossen. Historical event extraction from text. In *Proceedings of ACL LaTeCH, Portland, US*, 2011.

Agata Cybulska and Piek Vossen. Using semantic relations to solve event coreference in text. In *Proceedings of the workshop: Semantic relations II. Enhancing resources and applications (SemRel2012, LREC 2012).*, 2012.

Agata Cybulska and Piek Vossen. Semantic relations between events and their time, locations and participants for event coreference resolution. In *Proceedings of recent advances in natural language processing (RANLP-2013)*, 2013.

Agata Cybulska and Piek Vossen. Guidelines for ECB+ annotation of events and their coreference. Technical Report NWR-2014-1, VU University Amsterdam, 2014a.

Agata Cybulska and Piek Vossen. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2014)*, 2014b.

Agata Cybulska and Piek Vossen. "Bag of events" approach to event coreference resolution. Supervised classification of event templates. In *International Journal of Computational Linguistics and Applications (IJCLA).*, 2015a.

Agata Cybulska and Piek Vossen. Translating granularity of event slots into features for event coreference resolution. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, 2015b.

Donald Davidson. The Individuation of Events. 1969.

Pascal Denis and Jason Baldridge. Global joint models for coreference resolution and named entity classification. In *Procesamiento del Lenguaje Natural*, 2009.

Francis Michael Ostrowski Ferraro. *Unsupervised Induction of Frame-Based Linguistic Forms*. Phd dissertation, 2017.

Gottlob Frege. On sense and reference / Uber sinn und bedeutung. In *Zeitschrift fur Philosophie und philosophische Kritik*, volume 100, pages 25–50, 1892.

Annemarie Friedrich, Alexis Palmer, Melissa Peate Sørensen, and Manfred Pinkal. Annotating genericity: a survey, a scheme, and a corpus. In *Proceedings of The 9th Linguistic Annotation Workshop, LAW@NAACL-HLT 2015, June 5, 2015, Denver, Colorado, USA*, pages 21–30, 2015.

Christian Girardi, Manuela Speranza, Rachele Sprugnoli, and Sara Tonelli. CROMER: a Tool for Cross-Document Event and Entity Coreference. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2014)*, 2014.

Paul Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and semantics. 3: Speech acts.*, pages 41–58. New York: Academic Press., 1975.

Andrey Gusev, Nathanael Chambers, Pranav Khaitan, Divye Khilnani, Steven Bethard, and Dan Jurafsky. Using query patterns to learn the duration of events. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS11)*, 2011.

Willem Van Hage, Veronique Malais, Roxane Segers, and Laura Hollink. Design and use of the simple event model (sem). *the Journal of Web Semantics, Elsevier*, 2011.

Sanda M. Harabagiu, Razvan C. Bunescu, and Steven J. Maiorano. Text and knowledge mining for coreference resolution. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 2001.

Jerry R. Hobbs. Granularity. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, 1985.

Yu Hong, Tongtao Zhang, Tim O'Gorman, Sharone Horowit-Hendler, Heng Ji, and Martha Palmer. Building a cross-document event-event relation corpus. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with*

ACL 2016 (LAW-X 2016), pages 1–6, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-1701. URL `https://www.aclweb.org/anthology/W16-1701`.

Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. Events are not simple: Identity, non-identity, and quasi-identity. In *Proceedings of the 1st Workshop on EVENTS: Definitin, Detection, Coreference and Representation, NAACL-HLT 2013.*, 2013.

Blake Stephen Howald and Martha Abramson. The use of granularity in rhetorical prediction. In *Proceedings of the First Joint Conference on Lexi-cal and Computational Semantics (*SEM)*, 2012.

Kevin Humphreys, Robert Gaizauskas, and Saliha Azzam. Event coreference for information extraction. In *ANARESOLUTION '97 Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, 1997.

Catharina Maria Keet. A formal theory of granularity. toward enhancing biological and applied life sciences information systems with granularity. In *Ph.D. thesis, Faculty of Computer Science, Free University of Bozen-Balzano, Italy*, 2008.

Sebastian Krause, Feiyu Xu, Hans Uszkoreit, and Dirk Weissenborn. Event linking with sentential features from convolutional neural networks. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 239–249, 2016.

J. R. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

LDC. ACE (Automatic Content Extraction) English Annotation Guidelines for Events ver. 5.4.3 2005.07.01. In *Linguistic Data Consortium*, 2005.

LDC. Ace (automatic content extraction) english annotation guidelines for entities ver. 5.6.1 2005.05.23. Technical report, Linguistic Data Consortium, 2005a.

LDC. ACE (Automatic Content Extraction) English Annotation Guidelines for Events ver. 5.4.3 2005.07.01. Technical report, Linguistic Data Consortium, 2005b.

Claudia Leacock and Martin Chodorow. Combining local context with wordnet similarity for word sense identification. In *WordNet: A lexical Reference System and its Application*, 1998.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL)*, 2012.

Linguistic Data Consortium. Ace (automatic content extraction) english annotation guidelines for entities, version 6.6 2008.06.13. Technical report, June 2008. http://projects.ldc.upenn.edu/ace/docs/English-Entities-Guidelines_v6.6.pdf.

Zhengzhong Liu, Jun Araki, Eduard Hovy, and Teruko Mitamura. Supervised within-document event coreference using information propagation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2014)*, 2014.

Zhengzhong Liu, Teruko Mitamura, and Eduard Hovy. Graph-based decoding for event sequencing and coreference resolution. In *7th International Conference on Computational Linguistics(COLING 2018)*, 2018.

Jing Lu and Vincent Ng. Event coreference resolution: A survey of two decades of research. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 2018.

Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (EMNLP-2005)*, 2005.

Inderjeet Mani. A theory of granularity and its application to problems of polysemy and under-specification of meaning. In *In Principles of Knowledge Representation and Reasoning: Proceedings of the Sixth International Conference (KR-98)*, 1998.

C. Masolo, S. Borgo, A. Gangemi, N. Guarino, and A. Oltramari. Wonderweb deliverable d18: Ontology library, istc-cnr, trento, italy. 2003.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. MEANTIME, the NewsReader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*, 2016.

Rutu Mulkar-Mehta, Jerry R. Hobbs, and Eduard Hovy. Granularity in natural language discourse. In *Proceedings of International Conference on Computational Semantics*, 2011a.

Rutu Mulkar-Mehta, Jerry R. Hobbs, and Eduard Hovy. Applications and discovery of granularity structures in natural language discourse. In *Proceedings of The Tenth International Symposium on Logical Formalizations of Commonsense Reasoning at the AAAI Spring Symposium, Palo Alto*, 2011b.

Paul Nation and Robin Waring. Vocabulary size, text coverage and word lists. In Schmitt, Norbert, and McCarthy, editors, *Vocabulary: description, acquisition and pedagogy.* Cambridge University Press, 1997.

Vincent Ng. Machine learning for coreference resolution: From local classification to global ranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2005.

Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *in Proceedings of the 40th Annual Meeting of the 162 Association for Computational Linguistics (ACL), Philadelphia*, 2002.

I. Niles and A. Pease. Towards a standard upper ontology. In *Proceedings of FOIS 2001, Ogunquit, Maine*, pages 2–9, 2001.

I. Niles and A. Pease. Linking lexicons and ontologies mapping wordnet to the suggested upper merged ontology. In *Proceedings of the International Conference on Information and Knowledge Engineering, Las Vegas, Nevada*, 2003.

I. Niles and A. Terry. The milo: A general-purpose, mid-level ontology. In *Proceedings of the International Conference on Information and Knowledge Engineering, Las Vegas, Nevada*, 2004.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of 2nd Workshop on Computing News Storylines*, pages 47–56, 2016.

Siim Orasmaa. *Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience*. PhD thesis, 2016.

Fabian Pedregosa, Gal Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and douard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Charles S. Peirce. Collected papers of charles sanders peirce, hartshorne and weiss (eds.). 1931–58.

Charles Sanders Peirce. Prolegomena to an apology for pragmaticism. In *The Monist*, volume 16, pages 492–546, 1906.

Simone Paolo Ponzetto and Michael Strube. Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 192–199, 2006.

Marten Postma, Filip Ilievski, Piek Vossen, and Marieke van Erp. Moving away from semantic overfitting in disambiguation datasets. In *In Proceedings of EMNLP 2016s UBLP (Uphill Battles in Language Processing) workshop*, 2016.

Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. Unrestricted coreference: Indentifying entities and events in ontonotes. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC)*, September 2007.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. CoNLL-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of CoNLL 2011: Shared Task*, 2011.

J. Pustejovsky, C. Havasi, J. Littman, A. Rumshisky, and M. Verhagen. Towards a generative lexical resource: The brandeis semantic ontology. In *Proceedings of the Fifth Language Resource and Evaluation Conference*, 2006.

James Pustejovsky, Jose Castano, Bob Ingria, Roser Sauri, Rob Gaizauskas, Andrea Setzer, and Graham Katz. Timeml: Robust specification of event and temporal expressions in text. In *Proceedings of Computational Semantics Workshop (IWCS-5)*, 2003.

William V. O. Quine. Events and Reification. 1985.

Marta Recasens. Annotation guidelines for entity and event coreference. In *http://www.bbn.com/NLP/OntoNotes*, 2011.

Marta Recasens and Eduard Hovy. BLANC: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510, 2011.

Marta Recasens, Eduard Hovy, and M. Antnia Mart. Identity, non-identity, and near-identity: Addressing the complexity of coreference. In *Lingua, 121(6):1138-1152*, 2011.

Swen Ribeiro, Olivier Ferret, and Xavier Tannier. Unsupervised event clustering and aggregation from newswire and web articles. In *Proceedings of the 2017 EMNLP Workshop on Natural Language Processing meets Journalism, Copenhagen, Denmark, September 7, 2017*, pages 62–67, 2017.

Marco Rospocher, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. Building event-centric knowledge graphs from news. In *Journal of Web Semantics*, volume 37, pages 132–151, 2016.

Roser Saurí, Jessica Littman, Robert Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. Timeml 1.2.1 annotation guidelines, October 2005. http://timeml.org/site/publications/timeMLdocs/ annguide_1.2.1.pdf.

Roger C. Schank. Dynamic memory: A theory of reminding and learning in computers and people. 1990.

R. Segers, T. Caselli, and P. Vossen. The circumstantial event ontology(CEO). In *Proceedings of the Events and Stories in the News Workshop*, pages 37–41, 2017.

John Sinclair. Developing linguistic corpora: a guide to good practice. http://ota.ahds.ac.uk/documents/creating/dlc/ chapter1.htm#section4, 2004.

R. Sprugnoli and S. Tonelli. One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective. In *Natural Language Engineering*, pages 1–22, 2016.

C. Tenny and J. Pustejovsky. A history of events in linguistic theory. In *Events as Grammatical Objects*, 2000.

Teun A. van Dijk. *News As Discourse*. Routledge, 1988.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model theoretic coreference scoring scheme. In *Proceedings of MUC-6*, 1995.

P. Vossen, E. Agirre, N. Calzolari, C. Fellbaum, S. Hsieh, C. Huang, H. Isahara, K. Kanzaki, A. Marchetti, M. Monachini, F. Neri, R. Raffaelli, G. Rigau, and M. Tescon. Kyoto: A system for mining, structuring and distributing knowledge across languages and cultures. In *Proceedings of LREC 2008, Marrakech, Morocco, May 28-30, 2008*, 2008a.

P. Vossen, I. Maks, R. Segers, and H. Van der Vliet. Integrating lexical units, synsets and ontology in the cornetto database. In *Proceedings of LREC 2008, Marrakech, Morocco, May 28-30 May 2008.*, 2008b.

Piek Vossen. Grammatical and conceptual individuation in the lexicon. In *Studies in Language and Language Use, 15*, 1995.

Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, Marco Rospocher, and Roxane Segers. Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*, 110:60–85, 2016.

Piek Vossen, Tommaso Caselli, and Agata Cybulska. How concrete do we get telling stories? In *Topics in Cognitive Science*, volume 10, pages 621–640, 2018a.

Piek Vossen, Marten Postma, and Filip Ilievski. Referencenet: a semantic-pragmatic network for capturing reference relations. In *Proceedings of the 9th Global WordNet Conference, GWC 2018 - Singapore, Singapore*, 2018b.

Ruichen Wang. *Information-based event coreference*. PhD thesis, 2015.

Morton E. Winston, Roger Chaffin, and Douglas Herrmann. A taxonomy of part-whole relations. In *Cognitive Science Volume 11, Issue 4, pages 417 - 444*, 1987.

Bishan Yang, Claire Cardie, and Peter Frazier. A hierarchical distance-dependent Bayesian model for event coreference resolution. *Transactions of the Association for Computational Linguistics*, 3:517–528, 2015.

Xiaodong Yu, Wenpeng Yin, and Dan Roth. Paired representation learning for event and entity coreference, 2020.

Appendix 9: Granularity taxonomy

```
eng-30-00007846-n,gran_person,person_1,individual_1,someone_1,somebody_1,mortal_1,soul_2
eng-30-10476086-n,gran_person,prisoner_1,captive_1
eng-30-09630641-n,gran_person,unfortunate_1,unfortunate_person_1
eng-30-10378412-n,gran_person,operator_2
eng-30-09610660-n,gran_person,communicator_1
eng-30-09774783-n,gran_person,advocate_1,advocator_1,proponent_1,exponent_1
eng-30-10638385-n,gran_person,spokesperson_1,interpreter_3,representative_2,voice_8
eng-30-10638310-n,gran_person,spokesman_1
eng-30-08013845-n,gran_group,al-Qaeda_1,Qaeda_1,al-Qaida_1,Base_15
eng-30-05663671-n,gran_group,government_3
eng-30-08403787-n,gran_group,opposition_5
eng-30-08008335-n,gran_group,organization_1,organisation_2
eng-30-07968702-n,gran_group,interest_6,interest_group_1
eng-30-07969695-n,gran_group,kin_2,kin_group_1,kinship_group_1,kindred_1,clan_1,tribe_4
eng-30-07974850-n,gran_group,fringe_4
eng-30-07991364-n,gran_group,congregation_1_fold_2,faithful_2
eng-30-08215044-n,gran_group,platoon_3
eng-30-08240022-n,gran_group,revolving_door_1
eng-30-08240169-n,gran_group,set_5,circle_2,band_1
eng-30-08245172-n,gran_group,organized_crime_1,gangland_1,gangdom_1
eng-30-08288753-n,gran_group,subculture_1
eng-30-08294395-n,gran_group,nonalignment_1,nonalinement_1
eng-30-08372411-n,gran_group,tribe_1,folk_2
eng-30-08464601-n,gran_group,movement_4,social_movement_1,front_10
eng-30-08479095-n,gran_group,Jewry_1
eng-30-08486306-n,gran_group,wing_8
eng-30-02472987-n,gran_group,world_8,human_race_1,humanity_3,humankind_1
eng-30-07942152-n,gran_group,people_1
eng-30-07950786-n,gran_group,wounded_1,maimed_1
eng-30-07967382-n,gran_group,ethnic_group_1,ethnos_1
eng-30-07967982-n,gran_group,race_3
eng-30-08160276-n,gran_group,citizenry_1,people_2
eng-30-08180190-n,gran_group,multitude_3,masses_1,mass_6,hoi_polloi_1,people_4
eng-30-08306665-n,gran_group,varna_2
eng-30-00827974-n,gran_group,convoy_3
eng-30-03100490-n,gran_group,conveyance_3,transport_1
```

```
eng-30-04566257-n,gran_group,weaponry_1,arms_1,implements_of_war_1
eng-30-15213115-n,gran_month, October_1
eng-30-15213303-n,gran_month,mid-October_1
eng-30-15232899-n,gran_thousands_years,Middle_Paleolithic_1
eng-30-15233047-n,gran_thousands_years,Upper_Paleolithic_1
eng-30-15233239-n,gran_thousands_years,Mesolithic_Age_1
eng-30-15233411-n,gran_thousands_years,Neolithic_Age_1
eng-30-15233614-n,gran_thousands_years,great_year_1
eng-30-15236475-n,gran_season,season_2
eng-30-15237567-n,gran_season,dog_days_1
eng-30-15238074-n,gran_season,midwinter_1
eng-30-15239292-n,gran_week,season_3
eng-30-15239579-n,gran_season,season_1
eng-30-15242955-n,gran_year,long_time_1
eng-30-15232712-n,gran_thousands_years,Lower_Paleolithic_1
eng-30-15232406-n,gran_thousands_years,Paleolithic_Age_1
eng-30-15213669-n,gran_month,mid-November_1
eng-30-15213963-n,gran_month,mid-December_1
eng-30-15226732-n,gran_month,trimester_1
eng-30-15230790-n,gran_thousands_years,Golden_Age_3
eng-30-15231031-n,gran_thousands_years,silver_age_1
eng-30-15231263-n,gran_thousands_years,bronze_age_2
eng-30-15231415-n,gran_thousands_years,Bronze_Age_1
eng-30-15231634-n,gran_thousands_years,iron_age_2
eng-30-15231765-n,gran_thousands_years,Iron_Age_1
eng-30-15231964-n,gran_thousands_years,Stone_Age_1
eng-30-15232236-n,gran_thousands_years,Eolithic_Age_1
eng-30-15248564-n,gran_year,era_1
eng-30-15254550-n,gran_thousands_years,prehistory_1
eng-30-15293931-n,gran_year,indiction_1
eng-30-15294382-n,gran_year,prohibition_3
eng-30-15298283-n,gran_year,Great_Schism_1
eng-30-15141213-n,gran_thousands_years,millennium_1
eng-30-15141375-n,gran_thousands_years,bimillennium_2
eng-30-15157041-n,gran_day,calendar_day_1
eng-30-15163005-n,gran_day,day_of_the_week_1
eng-30-15136147-n,gran_week,week_3,calendar_week_1
eng-30-15155220-n,gran_day,day_1,twenty-four_hours_1
eng-30-15136342-n,gran_week,midweek_2
eng-30-15167349-n,gran_day,night_7
eng-30-15167474-n,gran_day,night_3
eng-30-15205799-n,gran_year,quinquennium_1
eng-30-15206097-n,gran_year,half-century_1
eng-30-15206195-n,gran_year,quarter-century_1
eng-30-15206590-n,gran_month,quarter_6
eng-30-15206744-n,gran_month,phase_of_the_moon_1
eng-30-15210383-n,gran_month,mid-January_1
eng-30-15210765-n,gran_month,mid-February_1
eng-30-15211090-n,gran_month,mid-March_1
eng-30-15211385-n,gran_month,mid-April_1
```

```
eng-30-15211711-n,gran_month,mid-May_1
eng-30-15212070-n,gran_month,mid-June_1
eng-30-15212358-n,gran_month,mid-July_1
eng-30-15212638-n,gran_month,mid-August_1
eng-30-15205719-n,gran_year,quadrennium_1
eng-30-15205532-n,gran_year,century_1
eng-30-15167906-n,gran_day,evening_3
eng-30-15169873-n,gran_week,week_1
eng-30-15170331-n,gran_week,fortnight_1
eng-30-15170504-n,gran_day,weekend_1
eng-30-15201505-n,gran_year,year_3
eng-30-15203791-n,gran_year,year_1
eng-30-15204485-n,gran_month,semester_2
eng-30-15204609-n,gran_month,bimester_1
eng-30-15204907-n,gran_year,lustrum_1
eng-30-15204983-n,gran_year,decade_1
eng-30-15213008-n,gran_month,mid-September_1
eng-30-15236176-n,gran_second,microsecond_1
eng-30-15236015-n,gran_second,nanosecond_1
eng-30-15235853-n,gran_second,picosecond_1
eng-30-15235687-n,gran_second,femtosecond_1
eng-30-15235540-n,gran_second,attosecond_1
eng-30-15235126-n,gran_second,second_1,sec_1
eng-30-15235334-n,gran_second,leap_second_1
eng-30-15117516-n,gran_hr,hours_2
eng-30-15234942-n,gran_min,quarter_4
eng-30-15234764-n,gran_min,minute_1
eng-30-15228267-n,gran_min,quarter-hour_1,15_minutes_1
eng-30-15228162-n,gran_min,half-hour_1
eng-30-15227846-n,gran_hr,hour_1,hr_1
eng-30-15206296-n,gran_month,month_2
eng-30-15167027-n,gran_day,night_1
eng-30-15155747-n,gran_day,night_4
eng-30-15236338-n,gran_second,millisecond_1
eng-30-15164957-n,gran_day,day_4
eng-30-15165289-n,gran_day,morning_1
eng-30-15219351-n,gran_month,Hindu_calendar_month_1
eng-30-15216966-n,gran_month,Islamic_calendar_month_1
eng-30-15214068-n,gran_month,Jewish_calendar_month_1
eng-30-15209706-n,gran_month,Gregorian_calendar_month_1
eng-30-15175640-n,gran_month,Revolutionary_calendar_month_1
eng-30-15203791-n,gran_year,year_1,twelvemonth_1,yr_1
eng-30-08561714-n,gran_country,southwest_4
eng-30-08561835-n,gran_country,west_8
eng-30-08561946-n,gran_country,northwest_4
eng-30-08593262-n,gran_continent,line_11
eng-30-08620061-n,gran_street,point_2
eng-30-08633957-n,gran_city,port_1
eng-30-08520401-n,gran_point_earth,celestial_point_1
eng-30-08547938-n,gran_street,crossing_3
```

```
eng-30-08649345-n,gran_street,side_1
eng-30-08613593-n,gran_street,outside_1,exterior_1
eng-30-08782490-n,gran_country,Achaea_1
eng-30-08783444-n,gran_country,Doris_2
eng-30-09001881-n,gran_continent,Witwatersrand_1,Rand_3
eng-30-08977845-n,gran_country,Sind_1
eng-30-08844923-n,gran_country,Papua_1
eng-30-08814781-n,gran_country,Transylvania_1
eng-30-08790199-n,gran_street,Cynoscephalae_1
eng-30-08588596-n,gran_country,midland_2
eng-30-08509442-n,gran_street,zone_3
eng-30-08509786-n,gran_country,belt_3
eng-30-08612786-n,gran_city,outline_1,lineation_1
eng-30-08516002-n,gran_city,city_line_1
eng-30-08515911-n,gran_city,county_line_1
eng-30-08515817-n,gran_city,district_line_1
eng-30-08514975-n,gran_country,Green_Line_1
eng-30-08515126-n,gran_country,Line_of_Control_1
eng-30-08515457-n,gran_country,state_line_1,state_boundary_1
eng-30-08501114-n,gran_country,frontier_2
eng-30-09433839-n,gran_city,shoreline_1
eng-30-08546183-n,gran_city,county_1
eng-30-08640739-n,gran_city,resort_area_1,playground_1
eng-30-08642145-n,gran_city,block_2,city_block_1
eng-30-08642331-n,gran_city,neighborhood_4
eng-30-08643015-n,gran_street,retreat_2
eng-30-08677628-n,gran_street,venue_1,locale_1
eng-30-08539276-n,gran_city,outskirts_1
eng-30-08641113-n,gran_city,vicinity_1,locality_1,neighborhood_1
eng-30-08523483-n,gran_city,center_1
eng-30-08970833-n,gran_continent,Western_Sahara_1,Spanish_Sahara_1
eng-30-08939562-n,gran_country,French_region_1
eng-30-08939201-n,gran_country,Riviera_1
eng-30-08929722-n,gran_country,Gaul_3
eng-30-08919693-n,gran_country,Phoenicia_1
eng-30-08918944-n,gran_country,Assyria_1
eng-30-08918248-n,gran_country,Sumer_1
eng-30-08917881-n,gran_country,Chaldea_1
eng-30-08917503-n,gran_country,Babylonia_1
eng-30-08916316-n,gran_country,Mesopotamia_1
eng-30-08943601-n,gran_country,Midi_1
eng-30-08944561-n,gran_country,Normandie_1
eng-30-08944818-n,gran_country,Orleanais_1
eng-30-08968677-n,gran_country,Mongolia_2
eng-30-08968390-n,gran_country,Tartary_1
eng-30-08957212-n,gran_country,Lappland_1
eng-30-08954269-n,gran_country,Thule_2
eng-30-08951777-n,gran_country,Frisia_1
eng-30-08951513-n,gran_country,Friesland_1
eng-30-08948155-n,gran_country,Guiana_1
```

```
eng-30-08945277-n,gran_country,Savoy_1
eng-30-08945110-n,gran_country,Lyonnais_1
eng-30-08915784-n,gran_country,Thrace_1
eng-30-08915372-n,gran_country,Macedon_1
eng-30-08915159-n,gran_country,Levant_2
eng-30-08886147-n,gran_country,Northumbria_1
eng-30-08885211-n,gran_country,Yorkshire_1
eng-30-08884806-n,gran_country,Lancashire_1
eng-30-08884673-n,gran_country,East_Anglia_1
eng-30-08882530-n,gran_country,New_Forest_1
eng-30-08881674-n,gran_country,Cumbria_1
eng-30-08857682-n,gran_country,British_Empire_1
eng-30-08845366-n,gran_country,Austria-Hungary_1
eng-30-08830720-n,gran_country,Klondike_1
eng-30-08886277-n,gran_country,West_Country_1
eng-30-08886636-n,gran_country,Wessex_1
eng-30-08888181-n,gran_country,Ulster_1
eng-30-08913242-n,gran_country,Elam_1
eng-30-08910394-n,gran_country,Gulf_States_2
eng-30-08905936-n,gran_country,Maharashtra_1
eng-30-08905751-n,gran_country,Gujarat_1
eng-30-08902894-n,gran_country,Punjab_1
eng-30-08902753-n,gran_country,Kanara_1
eng-30-08902569-n,gran_country,Sikkim_1
eng-30-08902422-n,gran_country,Hindustan_1
eng-30-08891415-n,gran_country,Caledonia_1
eng-30-08821578-n,gran_country,Maritime_Provinces_1
eng-30-09178596-n,gran_country,Dar_al-harb_1
eng-30-09052835-n,gran_country,Carolina_1
eng-30-09052652-n,gran_country,North_1
eng-30-09052100-n,gran_country,Piedmont_1
eng-30-09051898-n,gran_country,Tidewater_2
eng-30-09051726-n,gran_country,Sunbelt_1
eng-30-09051235-n,gran_country,Deep_South_1
eng-30-09050730-n,gran_country,South_1
eng-30-09050244-n,gran_country,Confederacy_1
eng-30-09049599-n,gran_country,Gulf_States_1
eng-30-09053019-n,gran_country,Dakota_2
eng-30-09090389-n,gran_country,Bluegrass_2
eng-30-09166127-n,gran_country,Low_Countries_1
eng-30-09178481-n,gran_country,Dar_al-Islam_1
eng-30-09178310-n,gran_continent,West_Africa_1
eng-30-09178141-n,gran_continent,North_Africa_1
eng-30-09177647-n,gran_continent,Scythia_1
eng-30-09172480-n,gran_continent,Sub-Saharan_Africa_1,Black_Africa_1
eng-30-09166902-n,gran_country,Silicon_Valley_1
eng-30-09166756-n,gran_country,Big_Sur_1
eng-30-09166534-n,gran_country,Silesia_1,Slask_1
eng-30-09166304-n,gran_country,Lusitania_1
eng-30-09049303-n,gran_country,Mid-Atlantic_states_1
```

```
eng-30-09048880-n,gran_country,New_England_1
eng-30-09048460-n,gran_country,Colony_3
eng-30-09012898-n,gran_country,Livonia_1
eng-30-09012101-n,gran_country,Baltic_State_1,Baltic_Republic_1
eng-30-09007471-n,gran_country,European_Russia_1
eng-30-09005712-n,gran_country,Siberia_1
eng-30-09004625-n,gran_country,Chechnya_1,Chechenia_1
eng-30-08995515-n,gran_country,Hejaz_1,Hedjaz_1
eng-30-08995242-n,gran_country,Nejd_1,Najd_1
eng-30-08978821-n,gran_country,Parthia_1
eng-30-09016232-n,gran_country,Donets_Basin_1,Donbass_1
eng-30-09018647-n,gran_country,Iberia_1
eng-30-09022831-n,gran_continent,Latin_America_1
eng-30-09048303-n,gran_continent,West_Coast_1
eng-30-09048127-n,gran_continent,East_Coast_1
eng-30-09042924-n,gran_continent,Ionia_1
eng-30-09039260-n,gran_country,Iraqi_Kurdistan_1
eng-30-09038990-n,gran_country,Kurdistan_1
eng-30-09035305-n,gran_country,Tanganyika_2
eng-30-09029242-n,gran_country,Sudan_2,Soudan_2
eng-30-09028367-n,gran_country,Leon_1
eng-30-09028204-n,gran_country,Galicia_1
eng-30-08975617-n,gran_country,Kashmir_1,Cashmere_3
eng-30-08819883-n,gran_country,Labrador_1
eng-30-08699426-n,gran_continent,East_Africa_1
eng-30-08602650-n,gran_country,Big_Bend_1
eng-30-08597323-n,gran_continent,Maghreb_1,Mahgrib_1
eng-30-08567072-n,gran_country,Enderby_Land_1
eng-30-08628414-n,gran_country,Queen_Maud_Land_1
eng-30-08644722-n,gran_city,country_4
eng-30-08683548-n,gran_city,wilderness_3
eng-30-08682819-n,gran_country,West_3,western_United_States_1
eng-30-08682188-n,gran_country,Wilkes_Land_1
eng-30-08678253-n,gran_country,Victoria_Land_1
eng-30-08675967-n,gran_city,urban_area_1,populated_area_1
eng-30-08673395-n,gran_street,tract_1,piece_of_land_1,piece_of_ground_1
eng-30-08564739-n,gran_country,Pacific_Northwest_1
eng-30-08564307-n,gran_country,Midwest_1,middle_west_1,midwestern_United_States_1
eng-30-08564139-n,gran_country,Northwest_1,northwestern_United_States_1
eng-30-08506347-n,gran_country,semidesert_1
eng-30-08504375-n,gran_country,Nubia_1
eng-30-08503921-n,gran_country,Bithynia_1
eng-30-08503238-n,gran_country,Barbary_1
eng-30-08502797-n,gran_country,Bad_Lands_1,Badlands_2
eng-30-08499840-n,gran_country,colony_5,dependency_3
eng-30-08494782-n,gran_country,Adelie_Land_1,Terre_Adelie_1
eng-30-08493493-n,gran_country,Appalachia_1
eng-30-08493261-n,gran_country,Andalusia_1,Andalucia_1
eng-30-08513718-n,gran_street,place_2,property_3
eng-30-08518940-n,gran_city,river_basin_1,basin_4
```

```
eng-30-08563990-n,gran_country,Northeast_2,northeastern_United_States_1
eng-30-08563627-n,gran_country,Southwest_2,southwestern_United_States_1
eng-30-08563478-n,gran_country,Southeast_2,southeastern_United_States_1
eng-30-08563180-n,gran_country,East_3,eastern_United_States_1
eng-30-08541841-n,gran_continent,zone_2,geographical_zone_1
eng-30-08541454-n,gran_continent,Coats_Land_1
eng-30-08519916-n,gran_continent,Transcaucasia_1
eng-30-08519624-n,gran_continent,Caucasia_1,Caucasus_2
eng-30-08819223-n,gran_country,Dalmatia_1
eng-30-08792083-n,gran_country,Fertile_Crescent_1
eng-30-08791167-n,gran_country,Middle_East_1,Mideast_1
eng-30-08790353-n,gran_country,Arcadia_1
eng-30-08789970-n,gran_country,Thessalia_1,Thessaly_1
eng-30-08788004-n,gran_country,Lydia_1
eng-30-08787878-n,gran_country,Lycia_1
eng-30-08787695-n,gran_country,Laconia_1
eng-30-08787466-n,gran_country,Epirus_1
eng-30-08779830-n,gran_country,Karelia_1
eng-30-08793489-n,gran_country,West_Bank_1
eng-30-08793914-n,gran_country,Galilee_1
eng-30-08794366-n,gran_country,Gaza_Strip_1,Gaza_1
eng-30-08817235-n,gran_country,Montenegro_1,Crna_Gora_1
eng-30-08816969-n,gran_country,Serbia_1,Srbija_1
eng-30-08800911-n,gran_country,Western_Roman_Empire_1,Western_Empire_1
eng-30-08800676-n,gran_country,Byzantine_Empire_1,Byzantium_2
eng-30-08799706-n,gran_country,Philistia_1
eng-30-08799271-n,gran_country,Judea_1,Judaea_1
eng-30-08799123-n,gran_country,Judah_2,Juda_1
eng-30-08798382-n,gran_country,Palestine_2,Canaan_1
eng-30-08794574-n,gran_country,Golan_Heights_1,Golan_1
eng-30-08776320-n,gran_country,Thuringia_1
eng-30-08776138-n,gran_country,Ruhr_2,Ruhr_Valley_1
eng-30-08775784-n,gran_country,Prussia_1,Preussen_1
eng-30-08715110-n,gran_continent,Southeast_Asia_1
eng-30-08713655-n,gran_country,pampas_1
eng-30-08711468-n,gran_continent,Patagonia_1
eng-30-08710535-n,gran_country,Bengal_1
eng-30-08709038-n,gran_country,Caribbean_2
eng-30-08704665-n,gran_country,Illyria_1
eng-30-08701719-n,gran_country,Pontus_2
eng-30-08701410-n,gran_country,Phrygia_1
eng-30-08701296-n,gran_country,Galatia_1
eng-30-08722844-n,gran_country,Manchuria_1
eng-30-08724545-n,gran_country,Turkistan_1,Turkestan_1
eng-30-08731953-n,gran_country,French_Indochina_1
eng-30-08775597-n,gran_country,Brandenburg_1
eng-30-08775297-n,gran_country,Rhineland_1,Rheinland_1
eng-30-08769179-n,gran_country,Saxony_1,Sachsen_1
eng-30-08760510-n,gran_country,Scandinavia_2
eng-30-08760393-n,gran_continent,northern_Europe_1
```

```
eng-30-08758882-n,gran_country,Bohemia_1
eng-30-08758679-n,gran_country,Moravia_1
eng-30-08757569-n,gran_country,Czechoslovakia_1
eng-30-08735564-n,gran_country,Mesoamerica_1
eng-30-08701161-n,gran_country,Cappadocia_1
eng-30-08682575-n,gran_continent,West_1,Occident_1
eng-30-08563085-n,gran_country,southland_1
eng-30-08562990-n,gran_country,northland_1
eng-30-08552138-n,gran_country,district_1
eng-30-08964099-n,gran_country,East_Malaysia_1
eng-30-08892766-n,gran_city,Lothian_Region_1
eng-30-08892058-n,gran_city,Galloway_1
eng-30-08873412-n,gran_country,Lake_District_1,Lakeland_1
eng-30-08858713-n,gran_country,British_West_Africa_1
eng-30-08858529-n,gran_country,British_East_Africa_1
eng-30-08854725-n,gran_country,Acre_2
eng-30-08837864-n,gran_country,Northern_Marianas_1,Northern_Mariana_Islands_1
eng-30-08834916-n,gran_country,Northern_Territory_1
eng-30-08964288-n,gran_country,Sabah_1,North_Borneo_1
eng-30-08964474-n,gran_country,Sarawak_1
eng-30-09090559-n,gran_country,Louisiana_Purchase_1
eng-30-09030096-n,gran_country,Kordofan_1
eng-30-09029884-n,gran_country,Darfur_1
eng-30-09028062-n,gran_country,Catalonia_1
eng-30-09027853-n,gran_country,Castile_1,Castilla_1
eng-30-09027460-n,gran_country,Aragon_2
eng-30-08991878-n,gran_country,American_Samoa_1,Eastern_Samoa_1
eng-30-08971693-n,gran_country,Natal_1,KwaZulu-Natal_1
eng-30-08964647-n,gran_country,West_Malaysia_1
eng-30-08830456-n,gran_country,Yukon_2,Yukon_Territory_1
eng-30-08825664-n,gran_country,Nunavut_1
eng-30-08553535-n,gran_city,residential_district_1,residential_area_1
eng-30-08549070-n,gran_city,development_6
eng-30-08537837-n,gran_city,city_district_1
eng-30-08825477-n,gran_country,Northwest_Territories_1
eng-30-08821187-n,gran_country,Acadia_1
eng-30-08809596-n,gran_country,Papal_States_1
eng-30-08789243-n,gran_city,Boeotia_1
eng-30-08786283-n,gran_city,Attica_1
eng-30-08785132-n,gran_country,Athos_1,Mount_Athos_1
eng-30-08775439-n,gran_country,Palatinate_1,Pfalz_1
eng-30-08631531-n,gran_country,possession_5
eng-30-08540532-n,gran_city,borough_1
eng-30-08626283-n,gran_city,municipality_1
eng-30-08654360-n,gran_country,state_1,province_1
eng-30-08672199-n,gran_city,township_1,town_3
eng-30-08672397-n,gran_city,ward_2
eng-30-08897843-n,gran_country,Lower_Egypt_1
eng-30-08898002-n,gran_country,Upper_Egypt_1
eng-30-08587709-n,gran_city,school_district_1
```

```
eng-30-08587174-n,gran_city,reservation_1,reserve_5
eng-30-08540770-n,gran_city,canton_2
eng-30-08544813-n,gran_country,country_2,state_7
eng-30-08546870-n,gran_country,county_2
eng-30-08553280-n,gran_country,federal_district_1
eng-30-08558488-n,gran_country,principality_1,princedom_2
eng-30-08558155-n,gran_country,kingdom_3,realm_2
eng-30-08557482-n,gran_country,empire_1,imperium_1
eng-30-08557396-n,gran_country,emirate_1
eng-30-08562454-n,gran_continent,Old_World_1
eng-30-08562757-n,gran_continent,Far_East_1
eng-30-08747887-n,gran_country,French_West_Indies_1
eng-30-09386842-n,gran_street,pass_4,mountain_pass_1
eng-30-08561583-n,gran_country,south_4
eng-30-08561462-n,gran_country,southeast_4
eng-30-08561081-n,gran_country,north_4
eng-30-08561230-n,gran_country,northeast_4
eng-30-08561351-n,gran_country,east_5
eng-30-09403734-n,gran_continent,range_4,mountain_range_1
eng-30-09409752-n,gran_city,ridge_4,ridgeline_1
eng-30-09433442-n,gran_city,shore_1
eng-30-09437454-n,gran_street,slope_1,incline_1,side_11
eng-30-09376786-n,gran_city,oceanfront_1
eng-30-09366317-n,gran_city,natural_elevation_1,elevation_4
eng-30-09217230-n,gran_city,beach_1
eng-30-09238926-n,gran_street,cave_1
eng-30-09331251-n,gran_city,lakefront_1
eng-30-09348460-n,gran_city,massif_1
eng-30-09468604-n,gran_city,valley_1,vale_1
eng-30-09461315-n,gran_country,trench_2,deep_2
eng-30-09344198-n,gran_country,lowland_1
eng-30-03542333-n,gran_street,hotel_1
eng-30-03449564-n,gran_street,government_building_1
eng-30-03203806-n,gran_street,diplomatic_building_1
eng-30-03093427-n,gran_street,consulate_1
eng-30-02913152-n,gran_street,building_1,edifice_1
eng-30-03679384-n,gran_street,living_quarters_1,quarters_1
eng-30-03763727-n,gran_street,military_quarters_1
eng-30-02944826-n,gran_street,camp_1,encampment_2,cantonment_1
eng-30-03546340-n,gran_street,housing_1,lodging_1,living_accommodations_1
eng-30-03574555-n,gran_street,institution_2
eng-30-03907654-n,gran_street,penal_institution_1,penal_facility_1
eng-30-03111690-n,gran_street,correctional_institution_1
eng-30-03592245-n,gran_street,jail_1,jailhouse_1,gaol_1,clink_2
eng-30-03297735-n,gran_street,establishment_4
```