Guidelines for ECB+ Annotation of Events and their Coreference

Technical Report NWR-2014-1

Version FINAL

Agata Cybulska, Piek Vossen

VU University Amsterdam
The Network Institute
a.k.cybulska,piek.vossen@vu.nl

BUILDING STRUCTURED EVENT INDEXES OF LARGE VOLUMES OF FINANCIAL AND ECONOMIC DATA FOR DECISION MAKING

ICT 316404

Contents

1 Introduction 7

2 Annotation of event components 9

2.1 Annotation of actions 11

2.1.1 Mention extent 11

2.1.2 Mention part of speech 13

2.1.3 Action classes 13

2.1.4 Action annotation checklist 15

2.2 Times and entity annotation 16

2.2.1 Mention extent 16

2.2.1 Mention part of speech 17

2.2.3 Subtypes 18

2.2.4 Times and entity annotation checklist 22

3 Coreference annotation 23

3.1 Coreference annotation checklist 24

4 Re-annotating ECB 0.1 25

5 Annotation tools 25

Appendix 26

List of Tables

1 Event components……………………………………………………………...8

2 Overview of main decisions with regard to event component annotation in ECB+.10

3 Overview of decisions made with regards to action annotation…………...………15

4 Overview of decisions made with regards to time and entity annotation…….……22

5 Overview of decisions made with regards to coreference annotation…….……….24

6 Overview of seminal events in ECB+ components……………………………..…26

1 Introduction

This document provides guidelines for the annotation of events and coreference between them. This annotation guideline makes a distinction between mentions (descriptions) of events in text and what they refer to, that is, their denotation (e.g. *World War II, WWII* and *the Second World War* all refer to a global war between 1939 and 1945). All mentions of events that refer to the same event should be annotated with coreference relation.

The annotation process consists of two phases. Firstly, a newly created ECB+ corpus component of 502 news articles should be annotated (Cybulska and Vossen, 2014). Secondly, the EventCorefBank (ECB, Bejan and Harabagiu, 2010) of 482 texts will be re-annotated.

In this section, readers will learn how we define events as composed of four components. In section 2, we explain how every event component should be annotated in text. In section 2.1 we discuss actions and in section 2.2 we take a closer look at times, locations and participants. After explaining how to determine the extent of component mentions in text (section 2.1.1 focuses on the extent of action mentions and 2.2.1 on the extent of time and entity mentions), we give an overview of how a component mention can be expressed in language (section 2.1.2 and 2.2.2). Finally, we present tags that should be used to annotate a component (section 2.1.3 and 2.2.3) and we summarize in form of an annotation checklist (section 2.1.4 and 2.2.4). In section 3 we describe how the coreference relation will be annotated amongst mentions of an event component. Section 4 elucidates how the second part of this annotation process is to be performed; that is how an existing corpus that we build upon should be re-annotated. Section 5 presents the tools that will be used for the purpose of the annotation.

First let us take a closer look at events. 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 characterized as “situations that happen or occur”. They can be expressed as punctual, durational, or stative predicates describing “states or circumstances in which something obtains or holds true*”* (Pustejovsky et al., 2003). Expanding the above definitions, we model events from news data as a combination of four components:

1. an event **action component** describing *what* happens or holds true (viz. §2.1)

2. an event **time slot** anchoring an action in time describing *when* something happens or holds true (viz. §2.2.3.1)

3. an event **location component** specifying *where* something happens or holds true (viz. §2.2.3.2)

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** (viz. §2.2.3.3) and **non-human participants** (viz. §2.2.3.4).

The annotation task described in this guideline requires annotators to annotate event actions, times, locations and participants in text. For example in the sentence:

On Monday Lindsay Lohan checked into rehab in Malibu, California after car crash.

1. action		*checked into; crash*
2. time		*On Monday*
3. location		*rehab in Malibu, California*
4. participant	Human	*Lindsay Lohan*
	non-human	*car*

Table 1. Event components.

Lindsay Lohan is a human participant involved with the event action *checked into*. *On Monday* tells us when the event happened and *rehab in Malibu, California* is where the action took place. *Crash* constitutes an action as well and *car* is a non-human participant of that action.

The ECB+ corpus annotation is an **event-centric** annotation task. We annotate mentions of event components in text from the point of view of an event action, marking:

- participants involved with an action as opposed to any participant mention occurring in a sentence
- time when an action happened as opposed to any time expression mentioned in text
- 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 will 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. In this sentence it should be annotated as a non-human participant of action *hate*. Event-centric thinking will guide us through the whole annotation effort and it will condition the decision making process with regard to annotation of particular linguistic phenomena. It will help us with the identification of the number of location, time and participant markables per action in a sentence. This is especially helpful with long component descriptions as in *ABC Entertainment Group prexy Paul Lee* which in ECB+ shall be annotated as a single human participant mention. The number of markables per action should correspond to the number of actual event participants, times and actions (a special case will be the way in which we treat some of the subjects and subject complements in copular constructions, viz. §3).

If an event is described more than once in one or in multiple texts, we say that its descriptions are *coreferent*. The second annotation task consists of marking the inter- and intra-document coreference relation between mentions of actions, participants, times and locations. Consider the following sentences:

Lindsay Lohan checked into rehab.
Ms. Lohan entered a rehab facility.

These two sentences might refer to the same event, although as Ms. Lohan 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 instances refer to the same real world event, they should be annotated as coreferent. If not, the actions should not be made coreferent, but the human participants 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.

2 Annotation of event components

A total of 43 seminal events are to be annotated in 502 texts of the new ECB+ corpus component (see table 6 in the appendix for an overview of seminal events in ECB+, Cybulska and Vossen, 2014). Event actions are to be annotated together with their times, locations and participants involved with them. Any other events named in the same sentence that describes a seminal event, should be annotated as well so that every event of a sentence is annotated.

We will first discuss the annotation of the action component in section 2.1. In section 2.2 we explain how event locations, times and participants should be annotated.

In table 2 we give an overview of the main decisions made with regard to annotation of event components. All these aspects will be described in detail in section 2. Annotators should skip this table in their first reading, though bear in mind that it might come in handy as an annotation checklist later on.

Component Annotation Aspect		Action	Time	Location	Human Participant	Non-human Participant
Mention Extent		Head (except for idioms and phrasal verbs)	Entire phrase	Entire phrase	Head	Head
Mention Form	Verbal	+	-	-	-	-
	Nominal incl. proper names	+	+	+	+	+
	Adjectival	+	ADJ may be part of TIME extent	ADJ may be part of LOC extent	-	-
	Predicative phrase	+	TIME may be part of a pred. phrase	LOC may be part of a pred. phrase	HUMAN_PART may be part of a pred. phrase	NON_HUMAN_PART may be part of a pred. phrase
	Pronominal	+	-	-	+	+
	Adverbial	-	+	+	-	-
Mention typology		OCCURRENCE PERCEPTION REPORTING ASPECTUAL STATE CAUSATIVE GENERIC + all above negated	DATE TIME_OF_THE_DAY DURATION REPETITION	GEO FAC OTHER	PER ORG GPE FAC VEH MET GENERIC	NON_HUMAN_PART NON_HUMAN_PART_GENERIC

Table 2. Overview of main decisions with regard to event component annotation in ECB+. A "+"indicates that a component can be expressed by a phrase or part of speech. A "-" means that a component cannot be represented by a part of speech or phrase.

In the remainder of these guidelines we will underscore words exemplifying how particular aspects of annotation discussed in the following sections should be annotated. All examples are presented in *italics.*

2.1 Annotation of actions

This section elaborates on how to annotate event actions in text. It is divided into four subsections. The first, describes how to determine the extent of a mention; the second presents what part of speech an action can be expressed with. Third subsection shows action classes that we will distinguish for the purpose of this annotation (the classes correspond to annotation tags) and in the fourth subsection we summarize this chapter with an annotation checklist.

2.1.1 Mention extent

In this section we will explain how to annotate an action phrase in text. To make things clearer, we will take a closer look at a number of examples.

Whether an action is verbal (like *the earth quaked*) or nominal (like *the earthquake*), we always annotate the word that is the strongest carrier of the action meaning; i.e. **the head of an action phrase**:

People would rather hear the positive things being talked about than the negatives.
The mall gunman may have been shooting at security cameras.
FBI did not investigate Fort Hood shooter.
This terrible war could have ended in a month.

In the examples above we left other parts of the action phrases like *would, may have been, did not, this terrible* and *could have* 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 will indicate negation in a different way (as explained in section §2.1.3). Besides auxiliary verbs, all verbs including aspectuals (like *start, stop, continue*) and causative verbs (like *cause*) should be annotated as separate actions as exemplified below.

Another report stated that the fighting started after a high-speed chase with a suspect vehicle in which a Gaddafi loyalist was killed.
The earthquake caused ruptures on the surface for a length of 470 kilometers.

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

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

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 the following example, the verb "to be" is used as an auxiliary:

The mall gunman may have been shooting at security cameras.

Comparatively, below, this same verb is used as the syntactic main verb:

Kittens are cute.
These people are amazing.

In these examples, just as in the case of auxiliaries, we will 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:

Gunman in Texas shooting was a marine.
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 §2.1.3). At the same time, if location, time or participant is also part of a predicative phrase, it should also be tagged as such (see for more information section §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 *Aaron is my favorite writer.* In this example we should annotate two mentions referring to a single participant referent of the state.

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 actions constituted by a **combination of a verb and a noun**, to preserve the full meaning both parts of the action phrase are to be annotated separately from each other; the verb as an action and the noun depending on the component that it refers to. It could be the case that the noun refers to an action and then it is also to be annotated as an action.

Congress did not back Barack Obama.
Russia has made an offer to Syria.

Mentions of different actions can be encountered in text: **generic actions** as opposed to actions anchored in time and space. 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. Mentions of abstract, generic events that are not anchored in time or space are also to be annotated and coreference between them should be annotated as well. Below some examples of generic actions from the TimeML specification (Sauri et al., 2006).

Use of corporate jets for political travel is legal.
Businesses are emerging on the Internet so quickly that no one, including government regulators, can keep track of them.

Jews are prohibited from killing one another.
The rabbi said Jews are prohibited from killing one another.

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 given examples not all actions are annotated, but only those that exemplify the construction shown in a bullet point.

We annotate actions that are expressed by:

- verbs

Syrian army fights rebels for control of key Christian town.
Indonesia GDP grows less than 6%.
At least 17 Taliban militants have been killed by Afghan and coalition security forces during the past 24 hours.

- nouns, including (but not limited to) nominalizations and proper nouns

The Civil War ended back in 1865.
Fast economic growth across the African continent…
Two arrested in the killing of a student.

- attributive use of present- and past- participles in modifier position

The deceased mens' house was sold yesterday.
The crying baby had a high fever.

- 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)

Gunman in Texas shooting was a marine.
Game Five hero David Ross was happy just to be here.

- pronouns

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

2.1.3 Action classes

We will not annotate mentions of actions with a general action tag but we will specify the class an action belongs to instead. We annotate actions with a limited number of classes from the whole set defined in the *TimeML Annotation Guidelines 1.2.1* (Sauri et al., 2006). We take over **five event classes from the TimeML** specification:

OCCURRENCE, PERCEPTION, REPORTING, ASPECTUAL and STATE (Pustejovsky et al., 2003).

Below the action tags that are to be used in the annotation process, together with explanation of their coverage and examples from TimeML.

(1) ACTION_OCCURRENCE tag, typically appropriate for most actions in the news, *describing something that happens or occurs in the world* such as *die, crash, build, merge, sell, land, arrive, distribute, eruption, explosion.*

(2) 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.*

(3) ACTION_REPORTING tag should be 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.*

(4) ACTION_ASPECTUAL tag is used to express *focus on different facets of event history* e.g.: *begin, finish, stop, continue* as in: *The Civil War ended back in 1865.* In *TimeML Annotation Guidelines 1.2.1* Sauri et al. (2006) distinguish between five facets of event history: *Initiation, Reinitiation, Termination, Culmination* and *Continuation* of an event.

(5) 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.* The ACTION_STATE tag is (amongst others) to be assigned to the non-verbal part of predicative phrases (constructions with verb *to be + nominal /pronominal/ adjectival part*).

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

(6) ACTION_CAUSATIVE is meant for action mentions such as *cause, lead to, result, facilitate, induce, produce, bring about.*

(7) ACTION_GENERIC tag is used to annotate generic events that are not anchored in time or space (for examples see last paragraph of section 2.1.1).

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 (like *not, neither*), other verbs (like *deny, avoid, be unable*), or by negation of participants involved with an event as in *No soldier went home.* We will annotate negation as a property of sentence actions by means of a set of action classes based on classes 1 - 7 but with indication of negation through addition of a *NEG_* tag in front of each action class. The following tags will be 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.

2.1.4 Action annotation checklist

Language phenomenon	**Treatment in ECB+**
Action classes	Annotated with a limited set of 5 classes from the TimeML specification + 2: causatives and generic actions & 7 negated classes
Auxiliary verbs (incl. auxiliary modals)	Not annotated
Light verbs	Annotated
Phrasal verbs and idioms	All elements annotated, also if discontinued
Aspectuals	Annotated as separate class
Causative verbs	Annotated as separate class
Generic events	Annotated as separate class
Event negation	Annotated as an action attribute
NP events	Annotated
Predicative phrases	Annotated
Adjectival predicates	Annotated
Resultative nominalizations	If applicable annotated as participants

Pronominal actions	Annotated

Table 3. Overview of decisions made with regards to action annotation.

2.2 Times and entity annotation

Similarly like section 2.1, this section is divided into four subsections. The first subsection describes how to determine the extent of a mention; the second presents what part of speech a mention can be expressed by. The third subsection shows component types that we will distinguish for the purpose of this annotation (these types will correspond to annotation tags) and in the fourth subsection again we will summarize this whole chapter with an annotation checklist.

2.2.1 Mention extent

In this subsection we will explain how to determine the extent of times and entities described in text.

With regards to times and locations we annotate **whole expressions**, **not only the head** of a phrase such as *two years ago, 3 days later, in July 1999* or *Portland, Maine*, *5 miles upstream* or *in the capital of Turkmenistan, in southern Iraq*.

In the case of participants we annotate **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* should be annotated as a human participant and *deceased* as an action:

Holland has health insurance treaties with a number of countries.

Homer the poet (most specific nominal part of the phrase)

The President of the U.S. Barack Obama (most specific nominal part of the phrase)

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

Some of the refugees

A group of kids

David Cameron, the Prime Minister of UK, said…

Usually 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).

Note that the head might consist of more than one word, in the case of **proper names** (e.g. *Barack Obama*).

With exception of locations and times, we do not annotate whole NPs but only their heads and we do not annotate markables within the extent of a bigger 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 do not annotate *US* and its type).

2.2.1 Mention part of speech

In this section we give an overview of how times and entities can be described in language.

We annotate 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. Here are some examples of times, locations and participants expressed by different part of speech:

- proper name as head of the phrase; also as part of a NP or PP

Barack H. Obama is the 44th President of the United States. (in this sentence *President* is the head of another person entity, though not one with a proper noun as head, hence not underscored)
UN climate talks in Warsaw darkened by Typhoon Haiyan. (the typhoon mention is also a proper name but it refers to an action)
In September the debut album by Canadian singer-songwriter Hayden comes out.

- common noun as head of the phrase; also as part of a NP or PP

The President of the United States ...
All Commission seats and the post of general counsel to the commission are filled by the President of the U.S.
The murdered family had stayed for a while in a house where people were previously murdered. (*in a house* is a location hence whole phrase was annotated)
This morning the Prime Minister announced she will re-nominate for Leader of the Federal Labor Party in a ballot next Monday morning.
The introduction of the euro in 1999 was a major step in European integration.

- pronominal participants

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.

- adverbial locations and times

The tugboat went 120 miles upstream in 20 hours.
The people of Fika got up from Tchad and went east to Dala, and stayed there one year.
Structural Heart Program was recently launched at Southcoast.
The murdered family had stayed for a while in a house where people were previously murdered.

Note that locations, times and 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 must also be annotated.

2.2.3 Subtypes

We annotate participants and locations expanding on the ACE entity subtypes (LDC, 2008). We annotate times 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 times, locations and participants.

2.2.3.1 Times

The time component of events marks explicit time expressions. When annotating time expressions, the annotators shall specify one of the four major types: DATE, TIME, DURATION and SET (Pustejovsky, et al., 2003).

The following four tags are used to annotate times, accompanied by examples from the TimeML specification (Sauri, et al., 2006).

(1) TIME_DATE tag refers to calendar time:

June 11, 1989
Yesterday
Summer, 2002
On Tuesday 18th
This summer
The second of December
Last week.

(2) 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:

Ten minutes to three
At five to eight
At twenty after twelve
At 9 a.m. Friday, October 1, 1999
The morning of January 31
(late) Last night
Between 8 a.m. and 10 a.m.

(3) TIME_DURATION tag is meant for time expressions denoting durations:

2 months
48 hours
Three weeks
All last night
20 days in July
3 hours last Monday.

(4) TIME_REPETITION tag corresponds to TimeML's SET (Sauri et al., 2006) and is used for sets of times describing repeated events like:

Often
Frequently
Every Tuesday
Twice a week
Every 2 days.

2.2.3.2 Locations

We define event locations in line with ACE's general PLACE attribute, corresponding to entity types GPE, LOC or FAC **referring to a physical location**.

The following three tags are meant for event location annotation, accompanied below by definitions from ACE entity guidelines (LDC, 2008).

(1) LOC_GEO tag corresponds to both, ACE's GPE - geo-political entities i.e. *geographical regions defined by political and/or social groups* ***referencing the territory or geographic position of the GPE***

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

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 the following examples:

A 7.2 magnitude earthquake hit in Southern California this afternoon.
Trip around the world
Landing on the moon
On the Vistula river
In the Tatra mountains
District of the city
We entered the airspace of Poland.

(2) 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.**

It is the deadliest mass murder in a school in United States history.
On the streets of Singapore

We also defined a third location tag:

(3) LOC_OTHER for any remaining type of event locations encountered in text.

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."
The mall gunman may have been shooting at security cameras.

2.2.3.3 Human participants

We define human event participants similarly to ACE's event participants of entity type PER, ORG but also metonymically used GPE, FAC and VEH 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.

The following tags are used to mark human event participants accompanied by definitions of corresponding entity types from ACE entity guidelines (LDC, 2005A, 2008).

(1) HUMAN_PART_PER tag refers to person entities and is *limited to humans; it may be a single individual or a group* of individuals*;* examples from ACE entity guidelines (LDC 2005A):

The President of the U.S.
The President of the U.S. Barack Obama
The family.

(2) HUMAN_PART_ORG tag denotes organization entities *limited to corporations, agencies and other groups of people defined by an established organizational structure.*

Air Force helicopters provided air support as the Navy attacked four LTTE boats.
The VU University Amsterdam decided to create a presence in Second Life.

(3) 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.

Poland and the US signed a $34 million deal to modernize the Polish Navy's missile frigate.
Hollywood is getting ready for this year's Fourth of July BBQ.
Boston won from Cleveland today in a short, decisive game that was uninteresting after the first innings.

(4) 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.** We have an example in the following sentence:

The school decided to find a new location.

But not in examples like:

The school was totally destroyed. (*school* as a non-human participant entity)
The blood bath happened in a school. (*school* as location of type facility)

(5) 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** such as in the following two sentences**:**

U.S. ships attacked 3 Iraqi patrol boats.
In 1991 Serbian tanks attacked Croatian cities.

But not in the example:

Somali refugees arrive by ship. (*ship* as a non-human participant)

In contrast to ACE's guideline we decided to distinguish an additional human participant subtype for human participant mentions, which are ambiguous with regard to their referent.

(6) HUMAN_PART_MET is meant for any remaining metonymically expressed human participants of events, see the following examples.

30% of households are living from paycheck to paycheck.
The press was present in large numbers and asked a great number of questions.
He has sworn loyalty to the flag.
The crown gave its approval.
That's not what I'm hearing from the boots on the ground.
The brown shirts marched through the town.

(7) Our final tag is HUMAN_PART_GENERIC which 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 (LDC 2008), for instance generic *you* or *one* as event participants.

One should treat others as one would like to be treated.
17 year old female seeking employment, loves working with kids.

In the event that the annotator finds it difficult to identify the appropriate annotation tag for a mention, it could be useful to apply the "substitution test". Try to rephrase the 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 logical idea. 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, in the mountains* or something similar, the resulting sentence is nonsensical and it is immediately obvious that location tags would be unsuitable.

2.2.3.4 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 is, besides human participants of events, event times and locations** - that contribute to the meaning of an event action (see examples below). These will often be **artifacts** expressed as a (direct or prepositional) object of a sentence or as PP phrases not in object position such as instrument phrases.

sharpen a pencil with a knife (both *pencil* and *knife* should be annotated as NON_HUMAN_PART)
Debbie traveled by boat 5 miles upstream to fish in her favorite spot.
Samsung signed a deal to be the NBA's official provider of tablets and televisions.
I hate Mondays. (Note that *Mondays* does not refer here to the time of an event action.)

Within the NON_HUMAN_PART type we distinguish 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 (LDC 2008) for instance in the sentence:

Linda loves cats.

2.2.4 Times and entity annotation checklist

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	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 ACE's entity types)

Table 4. Overview of decisions made with regards to time and entity annotation.

3 Coreference annotation

If an event component that is an action or its time, location or participant are described in one or multiple texts more than ones, their descriptions should be marked as coreferent.

Coreference relations can be established through mentions of:

- actions
- human participants
- non-human participants
- locations
- times.

Coreference can never be assigned between an action and an entity. Coreference should not be assigned neither 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

(3) with the same participants involved.

We annotate both, inter- and intra-document coreference.

Anaphoric coreference must be annotated as well.

In text 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) as in:

(1) *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 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:

(2) *James is just a little boy.*

coreference should NOT be marked. In example (2) *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. James* in this case is just one element of the whole set, hence the reference of the two is not identical.

Both sentences contain predicative phrases parts of which should be annotated as both human participants and states. In sentence (1) *James* should be annotated as both human participant of type person and as an action of class state. In sentence (2) *boy* should be annotated as both, human participant of type person and as an action of class state.

3.1 Coreference annotation checklist

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 5. Overview of decisions made with regards to coreference annotation.

4 Re-annotating ECB 0.1

There are some major differences between the annotation style of the ECB corpus (Bejan and Harabagiu, 2010) and 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 will re-annotate ECB 0.1 (Lee et al., 2012 and Recasens, 2011) 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 actions that will be re-annotated with specific action classes.

Wherever necessary, adjustments will be made with regards to mention extent. For human and non-human participant entities annotated in ECB 0.1 we will mark explicitly the heads of entity phrases. With regards to times and locations we will mark the whole phrase if not already done so. Regarding action annotation we need to make sure that light verbs and adjectival actions are annotated.

Finally adjustments might be needed to ensure that ECB 0.1 is compatible with the event centric annotation of the new corpus component.

The re-annotation effort will focus on sentences that were selected during the annotation of ECB 0.1. This should speed up the re-annotation process significantly. We will take over coreference relations established in ECB 0.1 but wherever needed we add new chains or adjust the existing ones.

5 Annotation tools

We annotate mentions of actions, times, participants and locations in text as well as within document coreference between them by means of the CAT - Content Annotation Tool (previously known as CELCT Annotation Tool (http://www.celct.it/projects/CAT.php, Bartalesi Lenzi et al., 2012).

For annotation of cross-document relations we will use a tool called CROMER (CRoss-document Main Event and entity Recognition). CROMER is a Newsreader project extension of a multi-user web interface (Bentivogli et al., 2008) designed within the Ontotext project (http://ontotext.fbk.eu/).

Appendix

Topic	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 sheriff's 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	Mafia member V. Gigante dies in

	prison Sicily 2008	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 6. Overview of seminal events in ECB+ components.

References

Bartalesi Lenzi, V, Moretti, G, Sprugnoli, R, 2012 *CAT: the CELCT Annotation Tool*. In Proceedings of LREC 2012, Istanbul.

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

Cybulska, Agata and Piek Vossen, *Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution,* In Proceedings of LREC 2014

L Bentivogli, C Girardi, E Pianta, *Creating a gold standard for person crossdocument coreference resolution in italian news*, LREC 2008 Workshop on Resources & Evaluation for Identity Matching, Entity Resolution and Entity Management, 2008

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. *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).

LDC. *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities*, ver. 5.6.1 2005.05.23. Linguistic Data Consortium. 2005A.

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

LDC. *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities* ver. 6.6 2008.06.13. Linguistic Data Consortium. 2008.

Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. *Unrestricted coreference: Indentifying entities and events in Ontonotes.* In Proceedings of the IEEE International Conference on Semantic Computing (ICSC), September.

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

Marta Recasens , *Annotation Guidelines for Entity and Event Coreference*, November 10, 2011

R. Sauri, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, J. Pustejovsky. *TimeML Annotation Guidelines, Version 1.2.1*, January 2006.