Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution.

Agata Cybulska, Piek Vossen

VU University Amsterdam
De Boelelaan 1105 1081HV Amsterdam
a.k.cybulska@vu.nl, piek.vossen@vu.nl

Abstract

In this paper we examine the representativeness of the EventCorefBank (ECB) (Bejan and Harabagiu, 2010) with regards to the language population of large-volume streams of news. The ECB corpus is one of the data sets used for evaluation of the task of event coreference resolution. Our analysis shows that the ECB in most cases covers one seminal event per domain, what considerably simplifies event and so language diversity that one comes across in the news. We augmented the corpus with 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, making it more representative of news articles on the web. The new "ECB+" corpus is available for further research.

Keywords: events, event coreference, ECB+

1. Introduction

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, 2005). In the TimeML specification, events are described as "situations that happen or occur" that can be punctual or durational, as well as stative predicates describing "states or circumstances in which something obtains or holds true" (Pustejovsky et al., 2003).
Expanding the above definitions, in this work we model events from news data as a combination of four components:

1. an event action component describing what happens or holds true
2. an event time slot 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.

For example in the sentence:
On Monday Lindsay Lohan checked into rehab in Malibu, California after a car crash.

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

Table 1: Event components.

Lindsay Lohan is a human participant involved with the event, *car* is a non-human participant, *On Monday* tells us when the event happened, *rehab in Malibu, California* is the place where the event happened and *checked into* and *crash* constitute actions (viz. TABLE 1).
We make 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 *Second World War* all refer to a global war between 1939 and 1945). If an event is described more than once in one or in multiple texts, we say that its descriptions are *coreferent*. A coreference relation can be established 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 can be considered as coreferent. If not, the actions should not be seen as coreferent. But the human participant descriptions from our example sentences are coreferent either way, as they refer to the same person. The last question is whether *rehab* and *rehab facility* refer to the same facility but to answer this one would need some extra context information.

Event coreference resolution is the task of determining whether two event descriptions (event mentions), refer to the same event. It is a difficult task that strongly influences diverse NLP applications. Evaluation of coreference resolution is not straightforward. There is no consensus in the field with regards to evaluation measures used to test approaches to coreference resolution. Some of the commonly used metrics[1] are highly dependent on the

[1]Evaluation metrics for coreference resolution include but are

evaluation data set, with scores rapidly going up or down depending on the number of singleton items in the data (Recasens and Hovy, 2011). Researchers tend to use the same data sets for evaluation of coreference resolution so that results of their work can be compared more easily, in consideration of the limitations of the evaluation metrics.

The EventCorefBank (ECB) (Bejan and Harabagiu, 2010) is one of the data sets available for studies of event coreference resolution. The corpus consists of 43 topics (each corresponding to a seminal event), which in total contain 482 texts from GoogleNews archive (http://news.google.com), selectively annotated (amongst other relations) with within- and cross-document event coreference. Events were annotated in accordance with the TimeML specification (Pustejovsky et al., 2003).
The annotation of the ECB, was extended by (Lee et al., 2012), following the OntoNotes annotation guidelines (Pradhan et al., 2007). The re-annotation process resulted in fully annotated sentences and annotation of NP coreference relations (no specific annotation of entity types was performed).

The goal of this paper is to shed light on the notion of representativeness of the ECB corpus as an evaluation data set for event coreference resolution. In follow up to deliberations on ECB's representativeness, this work contributes to "ECB+", a new resource for evaluation of approaches to event coreference resolution, that is more representative of large volume streams of news published over a longer period of time.

After looking at some corpora annotated with coreference of events (chapter 2), in chapter 3 we will analyze the lexical diversity of the ECB corpus. Chapter 4 presents the ECB+ corpus: the way in which the new corpus texts were collected and how the whole corpus was annotated. We conclude in chapter 5.

2. Related Work

Besides the ECB corpus, two preeminent data sets annotated with coreference of events (event identity) are available for English: the ACE 2005 data set and the OntoNotes corpus.

The ACE 2005 data set (LDC, 2005) was used in the 2005 Automatic Content Extraction technology evaluation. The English part of the data is annotated for entities, relations and events. This data set, containing 535 documents, was marked with within document coreference of events. Only a restricted set of 8 event types was annotated as LIFE, MOVEMENT, CONFLICT, JUSTICE.

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

not limited to MUC (Vilain et al., 1995) , B3 (Bagga and Baldwin, 1998) , CEAF (Luo, 2005), BLANC (Recasens and Hovy, 2011) , and most recently, CoNLL F1 (Pradhan et al., 2011) .

3. Lexical Diversity in the ECB

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

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 (Sinclair, 2004). Lets take a closer look at the representativeness of the ECB corpus.

3.1. Experiment 1

To determine the lexical (event mentions) and conceptual (the instances of events) diversity with regards to coreference chains captured by the (latest version of annotation of the) ECB corpus, we performed an experiment. For the purpose of the experiment, we created chains of corefering events based on lemma matches of mentions of event actions. For the experiment 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[2]. TABLE 2 shows the results of experiment 1 achieved by means of lemma matches of event actions in terms of recall (R), precision (P) and F-score (F) by employing the commonly used 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).

The following results were achieved in related work:

- (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
- (Lee et al., 2012): 62.7% MUC, 67.7% B3 F, 33.9% (entity based) CEAF, 71.7% BLANC F-score on the ECB corpus
- (Chen et al., 2011): 46.91% B3 F on the OntoNotes 2.0 corpus.

Compared to the evaluation results achieved in related work by means of the lemma approach coreference between events was solved with an F-score of 54.04% MUC,

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

Experiment\Metric	MUC			B3			CEAFm	BLANC			CoNLL
	R	P	F	R	P	F	R/P/F	R	P	F	F
Experiment 1	54.29	53.80	54.04	60.04	59.05	59.54	40.00	61.56	54.98	56.91	52.52
Experiment 2	51.83	83.16	62.19	59.40	92.75	71.23	61.03	63.10	84.09	65.53	64.76

Table 2: Event coreference resolution based on lemma match of actions in experiment 1 (cross-topic) and experiment 2 (within topic matches), evaluated on the ECB 0.1 in MUC, B3, CEAFm, BLANC and CoNLL F.

59.54% B3, 40.00% CEAFm, 56.91% BLANC F and 52.52% CoNLL F1. Considering that this approach neither performs anaphora resolution nor employs entities or any syntactic features, these are surprisingly good results.

3.2. Hypothesis

The high scores achieved by the lemma baseline make room for (at least) two crucial assumptions. First, it seems that there is relatively little lexical diversity in descriptions of event actions from ECB coreference chains. Second, other event components, besides the action slot, do not seem to play a crucial role in event coreference resolution, at least not if one evaluates on this data.

Based on our linguistic intuitions, we hypothesize that the two assumptions cannot be true in realistic situations (compare event descriptions as *car bombing in Madrid in 1995* with *bombing in Spain in 2009*, or *massacre in Srebrenica* with *genocide in Rwanda*). Our hypothesis is that the ECB corpus, while containing multiple documents describing particular real world events, in most cases captures only single instances of each particular event type. 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. As the number of event instances per topic is limited (in most cases referring to only one event instance of an event type, with exception of few topics like earthquake, acquisition, death and fire), event descriptions from a particular topic tend to share their entities (for a complete overview of seminal events in the ECB see (Cybulska and Vossen, 2014)). By that the event coreference task becomes simplified to topic classification.

3.3. Experiment 2

To illustrate this situation, we repeat our experiment, however this time we use lemma matches of event action mentions to generate event coreference chains within each topic of 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. TABLE 2 shows results of our second experiment.

Restricting lemma matches to actions from a topic bought us a 20-30% increase of precision across the evaluation metrics (8-20% improvement of the F-scores). It is remarkable to see that with the simple lemma match heuristic we obtained results comparable to those achieved by means of sophisticated machine learning approaches in related work.

This experiment to some extent exposed the division of work in a multi-step machine learning approach to coreference resolution. Based on the results of the second experiment, we see that much of the work on a data set like ECB is done with the topic classification step. Based on our intuitions, we make the assumption that the situation looks different if one considers large volumes 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 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 an event type. With this hypothesis in mind, we augmented the ECB corpus with other instances of the already captured event types, creating a new resource called "ECB+".

4. ECB+

4.1. Extending the ECB

With the objective to make the ECB corpus (482 texts[3]) more representative of large volume streams of news, we augmented the topics of the ECB with 502 texts reporting different instances of event types provided in the ECB. For example the first ECB topic consists of texts outlining Tara Reid's check-in into rehab in 2008. We created an extension to topic number one of the ECB, that is constituted by a collection of texts describing another event instance of the same type, namely Lindsay Lohan going into a rehab facility in 2013. ECB+ texts were collected by means of the Google News search. On average we gathered roughly eleven texts per topic. TABLE 3 shows some examples of seminal events, 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

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

Topic	Seminal event type	Human part ECB	Human part ECB+	Time ECB	Time ECB+	Loc ECB	Loc ECB+	Tnr ECB	Tnr ECB+
1	rehab check-in	T.Reid	L.Lohan	2008	2013	Malibu	Rancho Mirage	18	21
2	Oscars host announced	H.Jackman	E.Degeneres	2010	2014	-	-	10	11
3	inmate escape	Brian Nicols,4 dead	A.J. Corneaux Jr.	2008	2009	court-house, Atlanta	prison, Texas	9	11
4	death	B.Page	E.Williams	2008	2013	LA		14	10
5	head coach fired	Philadelphia 76ers, M.Cheeks	Philadelphia 76ers, J.O'Brien	2008	2005	-	-	13	10
6	"Hunger Games" sequel negotiations	C.Weitz	G.Ross	2008	2012	-	-	9	11
7	IBF, IBO, WBO titles defended	W.Klitchko, H.Rahman	W.Klitchko, T.Thompson	2008	2012	Germany	Switzer-land	11-1	11
8	explosion at a bank	-	-	2008	2012	Oregon	Athens	8	11
9	ESA changes	Bush	Obama	2008	2009	-	-	10	13
10	eight-year offer	Angels, M.Teixeira	Red Socks, M.Teixeira	2008		-	-	8	13

Table 3: The overview of seminal events in ECB and ECB+ topics 1-10

times, locations and number of texts per topic are listed. A complete overview of all seminal events captured by ECB+ can be found in the ECB+ annotation guideline.

4.2. ECB+ Annotation Tagset

In ECB+ we focused on annotation of mentions of events with their times and entities as well as coreference between them in text. We made an explicit distinction between specific entity types: human event participants, non-human participants, times, and locations (and a number of more specific subtypes amongst them e.g. HUMAN_PART_PER for human participants of subtype individual person) as well as between a set of action classes. The complete ECB+ annotation guidelines can be found in (Cybulska and Vossen, 2014). In the ECB+ annotation scheme we distinguish in total 30 annotation tags, taking from (Linguistic Data Consortium, 2008), (Pustejovsky et al., 2003) and (Saurí et al., 2005).

We annotated event actions with a limited set of classes from the whole set defined in the *TimeML Annotation Guidelines 1.2.1* (Saurí et al., 2005). We took over five event classes from the TimeML specification (Pustejovsky et al., 2003): "occurrence" (ECB+ tag ACTION_OCCURRENCE), "perception" (ACTION_PERCEPTION), "reporting" (ACTION_REPORTING), "aspectual" (ACTION_ASPECTUAL) and "state" (ACTION_STATE). Additionally we employed two more action classes, one for causal events (ACTION_CAUSATIVE) and one for generic actions (ACTION_GENERIC). These seven classes have seven equivalents, to indicate polarity of the event.[4]

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" (ECB+ tag TIME_DATE), "time" (TIME_OF_THE_DAY), "duration" (TIME_DURATION) and "set" (TIME_REPETITION).

We annotated participants and locations expanding on the ACE entity subtypes (Linguistic Data Consortium, 2008). 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. Three tags were used for event location annotation: (1) LOC_GEO corresponding to both, ACE's geo-political entities as well as ACE's location entities and (2) LOC_FAC meant for facility entities. Our intention was that mentions tagged as both (1) and (2) reference in a sentence where an action happened. We also applied a third location tag: (3) LOC_OTHER – for any remaining type of event locations encountered in text.

We define human event participants similarly to ACE's event participants of entity type "PER" (ECB+ tag HUMAN_PART_PER), "ORG" (HUMAN_PART_ORG) but also metonymically used "GPE" (HUMAN_PART_GPE), "FAC" (HUMAN_PART_FAC) and "VEH" (HUMAN-_PART_VEH) when referring to a population or a

[4]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 an action property by means of a set of action classes based on the seven non-negated action classes but with indication of negation through addition of a negation subtag (NEG_) in front of an action class tag.

government (or its representatives). Besides these five subtypes we also distinguish two additional ones: HUMAN_PART_MET – for any remaining metonymically expressed human participants of events (*He has sworn loyalty to the flag* or *The crown gave its approval*) as well as HUMAN_PART_GENERIC for 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.

Next to locations, times and human participants we recognize a fourth entity type – NON_HUMAN_PART – 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. These are often artifacts expressed as a (direct or prepositional) object of a sentence or as PP phrases not in object position such as instrument phrases. Within the NON_HUMAN_PART type we distinguish a special sub-tag for generic entities: 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) for instance in the sentence: *Linda loves cats*.

Within the ECB+ annotation task we annotated both, inter- and intra-document coreference relations (whether anaphoric or not) between mentions of a particular instance of an event component. 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 that happens or holds true: (1) in the same time, (2) in the same place and (3) with the same participants involved. In case of copular constructions, if the subject and its complement both refer to the same entity in the world, coreference between the two was annotated. If however, the reference of the sentence subject and of the subject complement is not EXACTLY the same as in: *James is just a little boy.* coreference would NOT be marked.[5]

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. 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).

4.4. Setup of the Annotation Task

The ECB+ corpus was annotated in three annotation rounds:

1. First mentions of event components were annotated in the newly created ECB+ corpus component and intra-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. Finally, cross-document coreference relations were established for each topic (the new topics include both, the ECB texts and the newly added ECB+ texts).

Two student assistants 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 the annotators were trained, 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 actions together with mentions of their participants, times and locations and intra-document coreference between them in the new ECB+ corpus component. The first topic of the new ECB+ component was annotated as burn in by both annotators. The next three topics were also annotated by both annotators (in total 55 texts per person annotated by both and used for the calculation of the inter-annotator agreement, see paragraph 4.5) and the remainder of the corpus (447 texts) was divided between the two student assistants and annotated once.

In the second stage of the annotation process, adjustments were made to the ECB 0.1 annotation (Lee et al., 2012; Recasens, 2011) of the EventCorefBank (Bejan and

[5] In the example sentence *James* refers to a particular boy called *James* but the phrase *a little boy* is indefinite and might refer to any little boy, 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.

Number of topics	43
Number of texts	982
Nr of annotated action mentions	14884
Nr of annotated location mentions	2255
Nr of annotated time mentions	2392
Nr of annotated human part. mentions	9577
Nr of annotated non human part. mentions	2963
Nr of intra-document chains	7671
Nr of cross-document chains	2204

Table 4: ECB+ statistics; including the re-annotated ECB corpus.

Harabagiu, 2010) (480 texts) 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 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 annotation 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 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 marked the whole phrase if not already done so. 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 annotation of ECB 0.1. This allowed us to speed up the re-annotation process significantly. In principle we took over the intra document coreference relations established in ECB 0.1 but wherever needed we added new chains or adjusted the existing ones.
The intra-document annotation in the first two stages of the ECB+ annotation process was performed by means of CAT - Content Annotation Tool (Bartalesi Lenzi et al., 2012)[6].

The third and final step in the ECB+ annotation process was to establish cross-document coreference relations between actions, times, locations and participants of a topic. Wherever applicable coreference links were created across both: the ECB texts and texts of the newly added ECB+ component. In this final stage for annotation of cross-document coreference relations we used a tool called CROMER (CRoss-document Main Event and entity Recognition, Girardi et al., 2014)

TABLE 4 lists some basic statistics with regards to the newly annotated resource.[7] The corpus can be downloaded at http://www.newsreader-project.eu/results/data/.

4.5. Inter-annotator Agreement

We calculated the inter-annotator agreement scores on topics 1 - 4 of the new ECB+ corpus component which contains 55 texts. 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 grouping the specific tags into 5 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 substantial agreement (Landis and Koch, 1977). The inter-annotator agreement on the five main event component tags also reached agreement level substantial: 0.79 Cohen's Kappa, although note that in these calculations untagged tokens were considered (for which we automatically assigned 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 (also substantial agreement). The confusion matrix in TABLE 5 shows the distribution of the five main tags in the four topics of the corpus component as coded by the annotators.
An analysis of the confusion matrix revealed that the annotators mainly struggled with the definition of mention extents, annotating whole mention phrases while the guideline specified otherwise that is to only annotate the head (or the other way around). After an additional training the annotators continued with stage one of the annotation.

Furthermore, 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" (assigned human friendly names e.g. *barack_obama*) which 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 coders to establish coreference relations for topics 1- 4 of the new ECB+ corpus com-

[6]Previously known as CELCT Annotation Tool, http://www.celct.it/projects/CAT.php .

[7]Note that the coreference chain statistics consider some singleton chains created by coders and in case of intra document relations some misformed chains as well. A total of 28 mentions was by mistake 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.

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

Table 5: Confusion matrix ECB+ topics 1-4; five component annotation by two coders: B and C. A stands for ACTION, T for TIME, L - LOCATION, H stands for HUMAN_PART tag, N for NON_HUMAN_PART, U for UNTAGGED.

ponent. We asked coder A to first work on topics 1, 2 and coder B to annotate topics 3, 4. Then coder B was asked to familiarize herself with instances created for topics 1,2 (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. Similar procedure was applied for second coder annotation of topics 3 and 4. Because of the CROMER setup it is clear what the intended referent (denotation) of a coreference chain that is of an instance is, hence we simply used Cohen's Kappa (Cohen, 1960) to calculate agreement on assignment of the coreference relation. As the total number of annotated items we considered all tokens annotated at least by one annotator. On the first four topics (490 cross document coreference IDs) our two coders reached Cohen's kappa of 0.76 (substantial agreement).

5. Conclusion

In this paper we analyzed the representativeness of the ECB corpus with regards to large volume news streams. We augmented the original ECB corpus with a new corpus component, creating a new, extended ECB+ corpus that captures descriptions of double event instances per topic, and by that 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 inter- and intra-document coreference between them. The coders reached substantial agreement on both: mention and conference annotation. We make this newly created resource available for research. The ECB+ can be used amongst others to develop and test approaches to event extraction and event coreference resolution.

6. Acknowledgements

This work has been supported by the EC within the 7th framework programme under grant agreement nr. FP7-IST-316040 and the Network Institute, VU University Amsterdam under the Semantics of History project. We are grateful for the contribution of the annotators Elisa Wubs and Melissa Dabbs as well as the feedback from the three anonymous reviewers. All mistakes are our own.

7. References

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

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

Bejan, Cosmin Adrian and Harabagiu, Sanda. (2008). 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.

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

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

Chen, Bin, Su, Jian, Pan, Sinno Jialin, and Tan, Chew Lim. (2011). 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.

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

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

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

LDC. (2005). Ace (automatic content extraction) english annotation guidelines for events ver. 5.4.3 2005.07.01. Technical report, Linguistic Data Consortium.

Lee, Heeyoung, Recasens, Marta, Chang, Angel, Surdeanu,

Mihai, and Jurafsky, Dan. (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)*.

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

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

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

Pradhan, Sameer, Ramshaw, Lance, Marcus, Mitchell, Palmer, Martha, Weischedel, Ralph, and Xue, Nianwen. (2011). Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of CoNLL 2011: Shared Task*.

Pustejovsky, James, Castano, Jose, Ingria, Bob, Sauri, Roser, Gaizauskas, Rob, Setzer, Andrea, and Katz, Graham. (2003). Timeml: Robust specification of event and temporal expressions in text. In *Proceedings of Computational Semantics Workshop (IWCS-5)*.

Recasens, Marta and Hovy, Eduard. (2011). Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Recasens, Marta. (2011). Annotation guidelines for entity and event coreference. In *http://www.bbn.com/NLP/OntoNotes*.

Saurí, Roser, Littman, Jessica, Knippen, Robert, Gaizauskas, Robert, Setzer, Andrea, and Pustejovsky, James. (2005). Timeml 1.2.1 annotation guidelines, October. http://timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf.

Sinclair, John. (2004). Developing linguistic corpora: a guide to good practice. http://ota.ahds.ac.uk/documents/creating/dlc/ chapter1.htm#section4.

Vilain, Marc, Burger, John, Aberdeen, John, Connolly, Dennis, and Hirschman, Lynette. (1995). A model theoretic coreference scoring scheme. In *Proceedings of MUC-6*.