Annotated Data, version 2

Deliverable D3.3.2

Version DRAFT

Authors: Marieke van Erp[1], Piek Vossen[1], Rodrigo Agerri[3], Anne-Lyse Minard[2], Manuela Speranza[2], Ruben Urizar[3], Egoitz Laparra[3], Itziar Aldabe[3], German Rigau[3]

Affiliation: (1) VUA, (2) FBK, (3) EHU

Building structured event indexes of large volumes of financial and economic data for decision making

ICT 316404

Grant Agreement No.	316404
Project Acronym	NEWSREADER
Project Full Title	Building structured event indexes of large volumes of financial and economic data for decision making.
Funding Scheme	FP7-ICT-2011-8
Project Website	http://www.newsreader-project.eu/
Project Coordinator	Prof. dr. Piek T.J.M. Vossen VU University Amsterdam Tel. + 31 (0) 20 5986466 Fax. + 31 (0) 20 5986500 Email: piek.vossen@vu.nl
Document Number	Deliverable D3.3.2
Status & Version	DRAFT
Contractual Date of Delivery	January 2015
Actual Date of Delivery	January 31, 2015
Type	Report
Security (distribution level)	Public
Number of Pages	62
WP Contributing to the Deliverable	WP3
WP Responsible	FBK
EC Project Officer	Susan Fraser

Authors: Marieke van Erp[1], Piek Vossen[1], Rodrigo Agerri[3], Anne-Lyse Minard[2], Manuela Speranza[2], Ruben Urizar[3], Egoitz Laparra[3], Itziar Aldabe[3], German Rigau[3]

Keywords: annotation, data, benchmarking

Abstract: This deliverable describes the annotation efforts of year 2. We present the setup of the annotation tasks (both intra-document and cross-document), the guidelines used (for all four project languages), the resulting gold standard datasets and an evaluation of the current NewsReader system on the manually annotated gold standard datasets. We also present the annotation guidelines and baseline evaluation for the creation of cross-document timelines, for which the NewsReader team is organising a shared task in the SemEval competition.

Table of Revisions

Version	Date	Description and reason	By	Affected sections
0.0	1 December 2014	Init	Marieke van Erp	all
0.1	15 December 2014	Added timelines description	Marieke van Erp	6
0.2	17 December 2014	Added Event coreference Evaluation	Piek Vossen	5
0.3	22 December 2014	Added NERC benchmark English	Rodrigo Agerri	5
0.4	26 December 2014	Added cross-document annotation guidelines and IAA	Anne-Lyse Minard	3, 4
0.5	19 January 2015	Added intra-doc annotation guidelines for Italian and Spanish	Manuela Speranza, Ruben Urizar, Anne-Lyse Minard	2
0.6	21 January 2015	Added temporal processing benchmark English	Anne-Lyse Minard	5
0.7	23 January 2015	Added English nominal coreference intra-document evaluation data	Rodrigo Agerri	5
0.8	29 January 2015	Added English NED intra-document evaluation	Itziar Aldabe	5
0.9	29 January 2015	Internal Review	Egoitz Laparra, Ruben Urizar, Itziar Aldabe	All
0.9	31 January 2015	Check by coordinator	VUA	-

Executive Summary

This deliverable describes the annotation efforts of the second year of the NewsReader project. This deliverable provides an update to Deliverable D3.3.1 Annotated Data v1. This update entails adaptations of the guidelines, in particular those of the attribution element (which was previously called factuality), the expansion to Spanish, Italian and Dutch and their accompanying datasets as well as our progress on cross-document annotation, both for events as well as for timelines.

The intra-document guidelines that were created for English (NewsReader technical report NWR-2014-2. Version 4.1 (Feb 2014)) were used as a mold for the Spanish, Italian and Dutch guidelines (and published as technical reports NWR-2014-6, NWR-2014-7 and NWR-2014-8 respectively). This deliverable will only highlight the changes particular to each language with respect to the English guidelines.

This deliverable also describes the data that was annotated for the four project languages. In order to create a balanced corpus and to facilitate cross-lingual benchmarking, the decision was taken to translate the 120 English Wikinews articles that were chosen for the English benchmarking effort into the other three project languages. These were then annotated and aligned to the English text.

As part of the cross-document annotation effort, we have gone one step beyond cross-document event annotation, as presented in Deliverable 3.3.1, and have added timeline annotations to the NewsReader corpus. The timelines that were generated from the 120 Wikinews articles served as gold standard for the SemEval-2015 Task 4: TimeLine: Cross-Document Event Ordering (pilot task)[1].

The three steps in our annotation effort (intra-document event annotation, cross-document event annotation and cross-document timeline annotation) are necessary steps to structure and organise data in such a way that story lines can be distilled from the sources.

[1] `http://alt.qcri.org/semeval2015/task4/`

Contents

1 Introduction

This deliverable describes the annotation efforts of the second year of the NewsReader project. The goal of the NewsReader project[2] is to reconstruct event story lines from the news by automatically processing daily news streams. For this purpose, an NLP pipeline has been constructed that extracts mentions of events, locations, dates, and participants (see WP04). The results of the extraction phase serve as input to a semantic layer where contradictions and complementary information are reconciled (see WP05) and are ultimately stored in a knowledge base (see WP06). To measure the performance of the automatic event extraction, benchmark datasets need to be developed, which is the focus of WP03. This deliverable is an update of Deliverable D3.2.1 Annotated Data v1. This update entails adaptations of the guidelines, in particular those of the attribution element (which was previously called factuality), the expansion to Spanish, Italian and Dutch and their accompanying datasets as well as our progress on cross-document annotation, both for events as well as for timelines.

In Y1, a core dataset of 120 English Wikinews articles was defined for the creation of the NewsReader gold standard annotated data set and guidelines for both intra-document and cross-document event annotation were defined. In Y2, the English gold standard annotation of the 120 articles was completed, and the articles were translated by professionals to the other three project languages, Spanish, Italian and Dutch. This ensured access to non-copyrighted articles in all project languages on the same topics, and even the option to compare the results of the NewsReader pipeline in the different languages at a finegrained level.

As part of the benchmarking effort, an evaluation of the NewsReader NLP pipeline has been undertaken. At the time of writing, only benchmark results for English have been obtained and are described in this deliverable. As the NewsReader annotations differ from many other annotated datasets in depth and breadth, the NWR team decided to not only report on results on the data annotated within the NewsReader project, but also on gold standard benchmark dataset previously annotated within the NLP community. This provides us with a comparison of the NewsReader NLP pipeline against state-of-the-art benchmarks, as well as of the annotated NewsReader dataset against the benchmark datasets. Our evaluations are accompanied by discussions of the differences between our dataset and the previous benchmark datasets and motivations of why we deem these annotations necessary.

During the project review at the end of February, we shall also provide evaluation results of the Italian, Spanish and Dutch pipelines. Due to the time the annotation effort took, these are somewhat delayed. In Q1 of 2015, an update to this deliverable that includes these results shall be provided.

In Year 1, we experimented with cross-document event annotation on the ECB+ (Cybulska and Vossen, 2014). The results from these experiments led the project to take the cross-document annotation one step further in Y2 and attempt cross-document timeline

[2] `http://www.newsreader-project.eu`

annotation. This annotation was done in the framework of the SemEval challenge[3], providing the project with an opportunity to share the created datasets directly with the research community and obtain feedback on them. As the SemEval campaign is still ongoing, the results will be presented at NAACL-HLT[4] in June.

This deliverable is structured as follows. In Section 2, we describe the updates to the intra-document annotation guidelines (NewsReader technical report NWR-2014-2. Version 4.1 (Feb 2014)). Followed by the update to the cross-document annotation guidelines in Section 3. In Section 4, we describe the dataset that was used in the annotation tasks as well as an overview of the annotation efforts into the different project languages.

In Section 5, the results of the intra-document benchmark evaluations for English are presented. In Section 6, the timeline annotation for SemEval-2015 Task 4: TimeLine: Cross-Document Event Ordering (pilot task) is described.

This deliverable rounds off with a conclusion and pointers for future work in Section 7.

2 Intra-document Annotation

In this section, we detail the updates to the intra-document guidelines for English as put forward in D3.3.1, as well as the intra-document guidelines for the other project languages.

2.1 Updates of the intra-document guidelines for English

During the annotation process, some issues were encountered that led us to update the intra-document guidelines for English. The two main changes concern the definition of the entity class and the annotation of attribution (previously called factuality). The changes are incorporated in an update of the annotation guidelines, published as technical report: Sara Tonelli, Rachele Sprugnoli, Manuela Speranza and Anne-Lyse Minard (2014) *NewsReader Guidelines for Annotation at Document Level.* NWR-2014-2-2. Version FINAL (Aug 2014). Fondazione Bruno Kessler.

2.1.1 Headlines

In addition to the annotation of the first 5 sentences of each document, we decided to also annotate the headlines. The headline annotation gives information about the main event of the news (or main topic). The headlines should be annotated with event and entity mentions, has_participant and refers_to relations. The temporal relations do not need to be annotated because they can be obtained through the events in the document corefering to the events of the headline.

ex: *[Apple]*$_{\text{ENTITY}}$ *[unveils]*$_{\text{EVENT}}$ *[iPod nano]*$_{\text{ENTITY}}$

[3] `http://alt.qcri.org/semeval2015/`
[4] `http://naacl.org/naacl-hlt-2015/`

2.1.2 Entity type PRODUCT

We defined a new entity type, PRODUCT. PRODUCT substituted ARTIFACT to include a wider spectrum of entities.

Product is anything that can be offered to a market that might satisfy a want or need[5]. This includes facilities (i.e. buildings, airports, highways, bridges, etc. as well as other structures and real estate improvements), vehicles (i.e. physical devices primarily designed to move an object from one location to another), weapons (i.e. physical devices primarily used as instruments for physically harming or destroying other entities), food (both human-made and produced by plants), products (including also abstract products such as software), functionalities (or features) of products, services, and trademarks (i.e. elements used for the public recognition of a company, for example logo).

Examples: vehicles, browser, internet access, trademark

2.1.3 Factuality (Attribution)

A profound study on factuality in text carried out at VUA led to new insights into how factuality related values should be annotated. The term factuality was replaced by the term attribution values since we attribute statements to sources in NewsReader and do not make any claims about their factual status. Attribution values of an event include the time it took place, the certainty of the source about it, and whether it is confirmed or denied (polarity). The adaptation of the scheme was led by VUA with the contribution of FBK and EHU. The first observations and insights into how to annotate attribution values were published in van Son *et al.* (2014). This work also presents the first steps towards representing world views by combining attribution and opinions.

In this section we describe the new defined attributes and we give examples of special cases.

2.1.3.1 The *certainty* attribute It expresses how certain the source about an event is: **certain**, **probable** and **possible**. Probable and possible events are typically marked in the text by the presence of modals or modal adverbs:
Markers of probability: *probably, likely, it's probable, it's likely*
Markers of possibility: *possibly, it's possible, maybe, perhaps, may, might, could*

The certainty of events is based on textual properties. We follow the guidelines from FactBank [6] to distinguish between POSSIBLE and PROBABLE events. The idea behind the distinction is that an event can be possibly true or possibly not true at the same time, but something cannot be probably true and probably not true at the same time.

[5]Definition taken from WikiPedia: `http://en.wikipedia.org/wiki/Product_(business)`.
[6]`http://www.cs.brandeis.edu/~roser/pubs/fb_annotGuidelines.pdf`

2.1.3.2 The *polarity* attribute It captures the distinction between affirmative and negative statements. Its values are POS for events with positive meaning (i.e. in most of the affirmative sentences), NEG for events with negative meaning (i.e. in most of the negative sentences), and UNDERSPECIFIED when it's not possible to specify the polarity of an event.

2.1.3.3 The *time* attribute It specifies the time an event took place or will take place, i.e. the semantic temporal value of an event. Its values are NON_FUTURE for present and past events, FUTURE for events that will take place and UNDERSPECIFIED when the time of an event cannot be deducted.

In the specific case of reported speech, the value of the time attribute is always related to the time of utterance and not to the time of writing (i.e. when the utterance is reported). For instance, *leave* in "*John said he would leave for Scotland*" is annotated as FUTURE (because John made a statement about the future) even if, at the time of writing, the leaving might have already taken place.

2.1.3.4 The *special_cases* attribute It captures if the statement has some special status that influences its attribution: general statement (GEN), main clause of a conditional construction (COND_MAIN_CLAUSE) or if clause of a conditional construction (COND_IF _CLAUSE). The default value of this attribute is NONE.

Events that are properties should be marked as general statement. Properties should be distinguisghed from events that are true in the present but have a time span that covers also some portion of the past and of the future.

2.1.3.5 Examples of attribution values annotation We call *attribution values* of an event the information concerning when it took place, the certainty of the source about it, and whether it is confirmed or denied. The *attribution values* consist of the value of attributes certainty, polarity, time and special_cases.

The president forgot to inform the cabinet.

predicate	certainty	polarity	time	special_cases
forgot	CERTAIN	POS	NON_FUTURE	NONE
inform	CERTAIN	NEG	NON_FUTURE	NONE

I don't remember, maybe Obama was born in 1961.

predicate	certainty	polarity	time	special_cases
remember	CERTAIN	NEG	NON_FUTURE	NONE
born	POSSIBLE	POS	NON_FUTURE	NONE

John does not know whether Mary came.

predicate	certainty	polarity	time	special_cases
know	CERTAIN	NEG	NON_FUTURE	NONE
came	POSSIBLE	UNDERSPECIFIED	NON_FUTURE	NONE

If we pollute our planet, future generation will suffer

predicate	certainty	polarity	time	special_cases
pollute	UNDERSPECIFIED	POS	FUTURE	COND_IF_CLAUSE
suffer	CERTAIN	POS	FUTURE	COND_MAIN_CLAUSE

2.1.3.6 No attribution annotation For event mentions referring to actions that are not really used as events in the text (i.e. they do not refer to a specific event and they are not anchored in time), attribution should not be annotated.

Volkswagen did not say how much the XL1 costs to build.

predicate	certainty	polarity	time	special_cases	comment
say	CERTAIN	NEG	NON_FUTURE	NONE	
costs	-	-	-	-	no attribution annotation
build	-	-	-	-	no attribution annotation

2.2 Italian and Spanish annotation task

The corpora used to build a benchmark for Italian and Spanish (as well as Dutch) consist of translations of the English corpus. The alignment between the source corpus (in English) and the corpus in the target language (Italian, Spanish and Dutch) has been done at the sentence level.

We took advantage of the alignment with English corpus and experimented on projecting the English intra-document annotation to the other languages. We provided annotators with files containing both sentences in Italian or Spanish aligned to the sentences in English. The annotation tool used (CAT) allows to visualize the annotation done for English when annotating the Italian or Spanish corpora. Figure 1 shows the annotation task in CAT. In the text panel of the interface, each English sentence (1) is followed by its translation in Italian or in Spanish (2). The instances that had been annotated previously within the English annotation task are displayed in the interface (4).

The annotation of Italian or Spanish sentences consists in four main steps:

- identification and annotation of the extent of a markable (as described in the NewsReader Guidelines);
- alignment of a markable to the corresponding markable in the original English text. This is performed through an attribute of type reference link, the `markId_English` attribute (3), that has been added to all text consuming markables in Italian or Spanish. The value of the `markId_English` attribute is filled through drag and drop of the corresponding markable annotated in English. If there is no equivalence of a

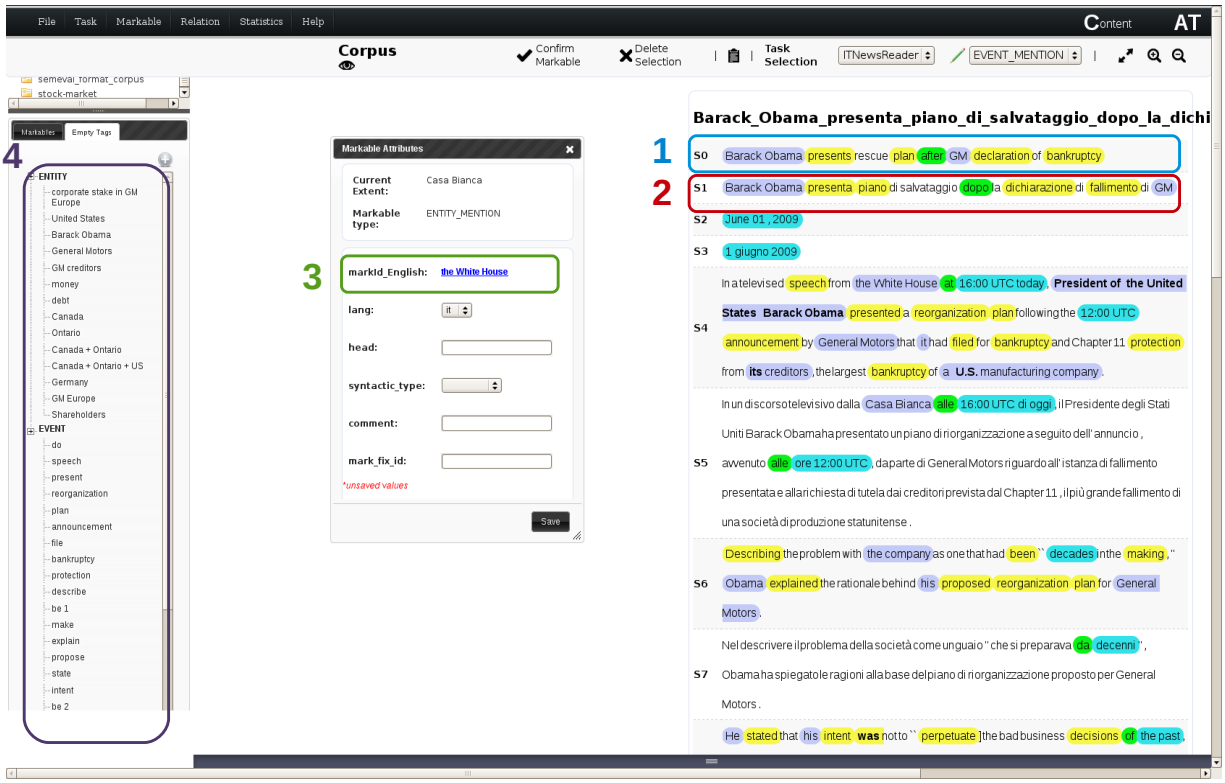

Figure 1: Visualization of the annotation task in CAT

ES/IT markable in English, annotators must create also the needed instances and relations (as for the English intra-doc annotation task);

- check of the relations (REFERS_TO, TLINK, CLINK, HAS_PARTICIPANT, SLINK), the instances (EVENT, ENTITY and empty TIMEX3) and the attributes of the markables (certainty, polarity, timex type, timex value, etc.) imported automatically into Italian or Spanish thanks to the reference links between ES/IT and English markables;

- annotation of missing relations.

2.3 Intra-document Annotation guidelines for Italian and Spanish

For the annotation of the Spanish and Italian guidelines we adopted the NewsReader guidelines defined for English (Tonelli *et al.* (NWR2014-2-2)).

In this section we describe only the extensions needed to adapt them to the specific morpho-syntactic features of Italian and Spanish. The revision and adaptation of the annotation guidelines for events is based on the It-TimeML guidelines (Caselli *et al.* (2011)) and on the Spanish TimeML guidelines (Saurí *et al.* (2009, 2010); Saurí (2010)), while the revision and adaptation of the annotation guidelines for entities is based on the I-CAB guidelines (Magnini *et al.* (2006)).

2.3.1 Contractions of prepositions and definite articles (Italian's articulated prepositions)

In the annotation of entity mentions and time expressions in English, prepositions are excluded from the extent while articles are included (e.g. *to [the family]*, *in [the next months]*). This is problematic for Italian and Spanish which, unlike English, have contractions of simple prepositions and definite articles. This phenomenon, which is common to many prepositions in Italian (e.g. *di*, *a*, *da*, *in*, *su*) and includes both singular and plural (e.g. *al* vs. *agli*) and both masculine and feminine (e.g. *al* vs. *alla*), is limited in Spanish to two contractions, *al* and *del*, in which the prepositions *a* or *de* respectively merge with the masculine singular definite article *el*.
Ex-IT: ***al*** *governo degli Stati Uniti* 'to the US government'
Ex-ES: ***al*** *gobierno de los Estados Unidos* 'to the US government'
Ex-IT: ***dal*** *5 novembre* ***al*** *10 dicembre* 'from November 5 to December 10'
Ex-ES: ***del*** *5 de noviembre* ***al*** *10 de diciembre* 'from November 5 to December 10'

Based on the above mentioned related work, we decided that these contractions should not be split but treated as single units in the annotation process. In particular:

- ENTITY MENTIONs: following the I-CAB guidelines, they should be included in the extent;

- TIMEX3s: following It-TimeML and Spanish TimeML, they should not be included in the extent; when a time expression is introduced by a contraction, this is usually to be marked as temporal SIGNALs.

Ex-IT: ENTITY MENTION *[**al** governo degli Stati Uniti]* '(to) the US government'
Ex-ES: ENTITY MENTION *[**al** gobierno de los Estados Unidos]* '(to) the US government'
Ex-IT: SIGNAL+TIME EXPRESSION *[**dal**] [5 novembre] [**al**] [10 dicembre]* 'from November 5 to December 10'
Ex-ES: SIGNAL+TIME EXPRESSION *[**del**] [5 de noviembre] [**al**] [10 de diciembre]* 'from November 5 to December 10'

2.3.2 Modals

According to the NewsReader guidelines for English (Tonelli *et al.* (NWR2014-2-2)), which are based on TimeML (Pustejovsky *et al.* (2003)), modal verbs are not annotated as events and the `modality` attribute is associated to the main verb (the value of the attribute is the token corresponding to the modal verb). On the other hand, the annotation of modals in NewsReader for Italian and Spanish follows It-TimeML and Spanish TimeML respectively: verbs expressing modality are themselves annotated as events (in particular, in the case of NewsReader, as events of type GRAMMATICAL); in addition, a GLINK (grammatical link) is created between the modal (source) and the main (target) verb (the `modality` attribute associated to the main verb is optional).

For instance, in the Spanish sentence *podemos jugar* 'we can play', two events must be annotated. Both verbs *podemos* 'we can' and *jugar* 'to play' are annotated as events, the verb conveying modality (*podemos*) being marked as an event of type GRAMMATICAL. Then a grammatical link is created between it and the verb *jugar* 'to play'.
Ex-ES: *[podemos] [jugar]* 'we can play'
Ex-IT: *[possiamo] [giocare]* 'we can play'
Ex-ES: *[tendrían] que [mejorar]* 'they will have to improve'
Ex-IT: *[dovranno] [migliorare]* 'they will have to improve'
Ex-ES: *[podrías] [descansar]* 'you could / might take a rest'
Ex-IT: *[potresti] [risposare]* 'you could / might take a rest'

2.3.3 Clitics

For Spanish and Italian, we have devised specific guidelines to handle clitics, which do not exist in English.
Ex-IT: *Aveva già deciso di **parlargli*** 'He had decided to talk to him'
Ex-ES: *Había decidido **hablarle*** 'He had decided to talk to him'

As with contractions of prepositions and definite articles, we have decided to leave the annotation at token level in the case of clitics. In particular, in the case of a token

composed of a verb (i.e. an event) and a clitic (i.e. a pronominal mention of an entity), the whole token will be annotated both as an entity and as an event. As it is important to distinguish the two annotated elements, the `head` attribute of the entity mention (see NewsReader Guidelines, section 3.2) is not optional for clitics as it is for all other types of entity mentions, and the `pred` attribute of the event mention (see NewsReader Guidelines, section 5.2.1) is not optional either.
Ex-IT: EVENT MENTION: [parlargli], pred "parlare"
Ex-IT: ENTITY MENTION: [parlargli], head "gli"

Ex-ES: EVENT MENTION: [hablarle], pred "hablar"
Ex-ES: ENTITY MENTION: [hablarle], head "le"

As far as clitics in pronominal verbs are concerned, we have created specific guidelines for the different classes. Truly reflexive (the object of the action is the same as the subject) and reciprocal pronouns (expressing mutual action or relationship among the referents of a plural subject) are annotated as entities. In the case of benefactive (the focus refers to the person or thing an action is being done for) and pseudo-reflexive pronouns (which occur with intransitive pronominal verbs), we have no entity annotation.
Ex-IT: *[**Mi**] sono ferito in montagna* 'I hurt myself in the mountains'
Ex-ES: *[**Me**] lastimé en la montaña* 'I hurt myself in the mountains'
Ex-IT: *Quelle due persone [**si**] amano* 'Those two people love each other'
Ex-ES: *Esas dos personas [**se**] aman* 'Those two people love each other'
Ex-IT: ***Mi** sono lavato le mani* 'I washed my hands'
Ex-ES: ***Me** lavé las manos* 'I washed my hands'
Ex-IT: ***Si** è mosso troppo velocemente* 'He moved too fast'
Ex-ES: ***Se** movía demasiado deprisa* 'He moved too fast'

When the Spanish "se" and the Italian"si" are used as impersonal pronouns (which corresponds to 'one', 'you', 'we', or 'they' in English) and as passive pronouns, they are not annotated.
Ex-IT: ***Si** dice che sia molto intelligente* 'they/people say he is very smart'
Ex-ES: ***Se** dice que es muy inteligente* 'they/people say he is very smart'
Ex-IT: *Da qui **si** vede il lago* 'from here the lake can be seen'
Ex-ES: *Desde aquí **se** ve el lago* 'from here the lake can be seen'

2.4 Intra-document Annotation guidelines for Dutch

With Dutch being a sister language of English, no major changes were necessary to adapt the intra-document annotation guidelines from English to Dutch, save for a section devoted to the Dutch adverb 'er'. The word can carry a variety of meaning that can be classified into four types of use, namely locative, presentative, prepositional and quantitative (Bennis, 1986). For the NewsReader annotation, only the locative use is deemed important, for

example in the sentence "Hij woont er al jaren" (He has lived there for years) 'er' is to be annotated as a mention of an entity of class LOCATION.

Furthermore, in the Dutch language compounding is more prevalent, which led to changes in some of the examples and difficulties of applying the word count rule to determine whether an entity mention is a NAM (proper name) or a NOM (common noun). Lastly, the annotators were made aware that the Dutch language contains more discontinuous predicates, which affects the event mention annotation layer.

3 Cross-document Annotation Task

Three partner institutions were involved in the NewsReader cross-document annotation task for English: FBK, EHU and VUA. As the leading institution for the annotation effort, FBK produced the annotation guidelines with the collaboration of other partners (Speranza and Minard (NWR2014-9)). After a training phase, in which FBK guided VUA and EHU in using the annotation tool (CROMER Girardi *et al.* (2014)), the annotation of English started with an agreement phase.

Annotation at corpus level consists of two main steps:

- cross-document entity coreference annotation of all entities annotated in the first 5 sentences and the headline of each file;

- cross-document entity and event coreference annotation starting from a set of seed entities. All the mentions that corefer to the seed entities should be annotated as well as the events in which the seed entities are participants.

For the identification of entity mentions, for the definition of their extent, and for the annotation of coreference, we use the NewsReader intra-document annotation guidelines (Tonelli *et al.* (NWR2014-2-2)).

Sections 3.1 and 3.2 describe the annotation of entity mentions and instances and the annotation of event mentions and instances.

In section 3.3, we detail the annotation task, i.e. the different steps of the cross-document annotation.

3.1 Entities

3.1.1 Entity mentions

As far as the extent of entity mentions is concerned, annotators should apply the same guidelines provided for the intra-document annotation (for entity mention extent, see section 3.1 of NWR-2014-2-2). The only exception is that CONJ-mentions (i.e. entity mentions connected by a coordinating conjunction) will not be annotated.

3.1.2 Entity instances

Each entity instance has the following attributes:

- class (compulsory): to be filled according to the guidelines provided for the intra-document annotation (see section 2.1 of NWR-2014-2-2 on Entity types);
- name (compulsory): to be filled according to the guidelines provided for the intra-document annotation (see section 2.4 of NWR-2014-2-2 on Tag Descriptor);
- short description (compulsory): a short description of the entity instance, whose aim is to distinguish it from other entity instances with similar names;
- external reference (compulsory): to be filled according to the guidelines provided for the intra-document annotation (see section 2.3 of NWR-2014-2-2);
- comment (optional).

3.2 Events

3.2.1 Event mentions

The annotation of event mentions is based on the intra-document annotation guidelines (for event mention extent, see section 5.1 of NWR-2014-2-2).

"Event is used as a cover term to identify "something that can be said to obtain or hold true, to happen or to occur" (ISO TimeML Working Group, 2008). This notion can also be referred to as eventuality (Bach, 1986) including all types of actions (punctuals or duratives) and states as well."

Some events annotated following the NewsReader guidelines could not go on a timeline, for example because they didn't happen (counter-factual events) or they are uncertain. In order to annotate only events potentially candidates to participate to a timeline, we have defined criteria based on the intra-document annotation Guidelines.

We annotate verbs, except if they are modified by a modal word, nouns and pronouns. Adjectives generally express a property or attribute of an entity, and anchoring them in time is not simple. So adjectival events will not be annotated.

Events are classified according to semantic features. Those classified as "grammatical" are dependent to a content verb/noun and don't have a time span, so they will not be annotated. We have also decided to leave out cognitive events (i.e. events that describe mental states or mental acts).

The last criterion is based on the factuality and certainty of events. Counter-factual events will not be part of a timeline because they did not take place. Non-factual events are speculative events, so we do not know if they happen or not. If it is certain that they will happen (e.g. "the conference will take place on Monday"), they will be annotated. But if they are uncertain (e.g. "the conference may take place later"), we will not annotate them.

3.2.2 Event instances

CROMER event instances have the following attributes:

- class (compulsory): to be filled according to the guidelines provided for the intra-document annotation (see section 4.2 of NWR-2014-2-2);
- name (compulsory): to be filled according to the guidelines provided for the intra-document annotation (see section 4.1 of NWR-2014-2-2 on Tag Descriptor for Events);
- short description (compulsory): a short description of the event instance, whose aim is to distinguish it from other event instances with similar names;
- time (compulsory for punctual events): a date (maximum granularity day), following the TIMEX3 format (see Section 6.2.2 of NWR-2014-2-2); if not known, add the values XXXX-XX-XX, XXXX-XX or XXXX depending on granularity;
- begin (compulsory for durative events): the starting date of the event (maximum granularity day), following the TIMEX3 format (see Section 6.2.2 of NWR-2014-2-2)); if not known, add the values XXXX-XX-XX, XXXX-XX or XXXX depending on granularity;
- end (if known, compulsory for durative events): the ending date (maximum granularity day), following the TIMEX3 format (see Section 6.2.2 of NWR-2014-2-2);
- external reference (compulsory): to be filled according to the guidelines provided for the intra-document annotation (see section 4.3 of NWR-2014-2-2);
- comment (optional).

In case of repeated events or grouped events the following rules should be applied to fill the time attributes:

- repeated events (e.g. "I go to work every morning")
  - time: if known add the value of the set temporal expression (following the TIMEX3 format);
  - begin (compulsory for repeated events): add the date of the first time the repeated event happened; if not known, add the values XXXX-XX-XX, XXXX-XX or XXXX depending on granularity;
  - end: if known, add the date of the last time the repeated event happened;
- grouped events (e.g. "the explosions caused huge damage") are considered as if they were durations
  - begin (compulsory for grouped events): add the date of the event that happened chronologically first; if not known, assign the values XXXX-XX-XX, XXXX-XX or XXXX depending on granularity;
  - end: if known, put the date of the event that happened chronologically last;

3.3 Annotation task

The cross-document annotation is carried out using a tool called CROMER (CRoss-document Main Event and entity Recognition) (Girardi *et al.* (2014)).

3.3.1 Phase I: Entity annotation

For each entity annotated in the first 6 sentences (included the headline) using CAT, the annotation task consists of the following steps:

- check if an entity instance already exists in CROMER and if not create it;
- assign the mention chains to the entity instance; (if annotators find annotation errors in the import from CAT, they can correct them).
- find other mentions of the entity in the corpus (only in the first 6 sentences of each document annotated with CAT) and assign them to the entity instance.

We manually selected a set of relevant target entities that appeared in at least two different documents and were involved in more than two events. Each partner institution receives the list of the target entities to be annotated in each corpus.

For each selected entity the annotation task consists of the following steps:

- check if an entity instance already exists in CROMER and if not create it;
- assign the mention chains imported from CAT (which refer only to the first 6 sentences of a document) to the entity instance; (if annotators find annotation errors in the import from CAT, they can correct them).
- find all other mentions of the entity in the corpus (i.e. in the remaining part of each document), annotate them with the correct extent (no attributes have to be annotated) and assign them to the entity instance.

3.3.2 Phase II: Event annotation

Annotators should annotate all events having as participant one of the seed entities annotated in phase I.

More specifically, the annotation task consists of the following steps:

- For each selected entity, get the list of all documents in which that entity is mentioned.
- For each document, identify all event mentions having the annotated entity as participant and annotate them as follows:
  - check if the event instance to which it refers already exists; if it does not exist, create it;

- create a HAS-PARTICIPANT relation between the event instance (source) and the entity instance (target);
- if the event mention is one imported from CAT, assign it to the event instance it refers to; otherwise annotate the mention with the correct extent (no attributes have to be annotated) and assign it to the event instance it refers to.

4 Data

4.1 Overview of the data

As was mentioned in D3.3.1, it is important to the NewsReader project to be able to make the annotated data of the NewsReader intra-document annotation task available not only to the project partners, but also to the wider audience of NLP researchers. Therefore, we chose Wikinews[7] as our core corpus for the annotation effort.

Next to the 20 articles concerning "Apple" that were used for defining and finetuning the annotation guidelines in Y1, 10 more "Apple" articles were selected, as well as 30 articles concerning "Airbus-Boeing", 30 articles concerning "GM-Chrysler-Ford" and 30 articles concerning "the stock market". As in Y1, the articles were selected in such a way that the corpus contains different articles that deal with the same topic over time (e.g. launch of a new product, discussion of the same financial indexes). This enables us to benchmark our cross-document event coreference modules, as well as build cross-document time and story lines.

In Table 1 we give some statistics about the intra-document annotation in the Wikinews corpus.

In Table 2 we present some statistics about the cross-document annotation in three subcorpora of the Wikinews corpus.

Since Wikinews does not contain enough overlapping articles between the four project languages to create balanced corpora around the same topics in the different project languages, the decision was made to translate the originally selected English Wikinews articles into Spanish, Italian and Dutch. For the translations, professional translation companies were hired that translated the articles in a sentence-by-sentence manner. The Dutch translations were checked by a NewsReader team member fluent in both Dutch and English.

4.2 Inter-Annotator Agreement in Intra-Document Annotation

4.2.1 English

In deliverable D3.3.1, Section 3.1.1, we reported on the inter-annotator agreement for all aspects (all markables, attributes and relations) of the intra-document annotation. In this section we provide specific data on the inter-annotator agreement on attribution (previously called factuality), for which the guidelines have been changed.

[7] `http://en.wikinews.org/wiki/Main_Page`

	Stock-market	GM-Chrysler-Ford	Airbus-Boeing	Apple	Total
# files	30	30	30	30	120
# sentences	120	118	120	119	477
# tokens	3,332	3,612	3,590	3,407	13,941
EVENT_MENTION	521	567	514	471	2,073
EVENT	455	406	437	392	1,690
ENTITY_MENTION	445	753	661	813	2,672
ENTITY	282	274	339	330	1,225
TIMEX3	164	143	101	118	526
VALUE	282	124	96	48	550
C-SIGNAL	9	3	12	5	29
SIGNAL	92	73	64	60	289
REFERS_TO	732	674	764	697	2,867
TLINK	373	527	408	475	1,783
CLINK	22	4	14	10	50
GLINK	42	76	54	37	209
HAS_PARTICIPANT	364	505	605	509	1,983
SLINK	25	104	68	40	237

Table 1: Intra-document annotation in the Wikinews corpus

We measured the inter-annotator agreement with the Dice's coefficient on 97 event mentions (referring to 83 distinct event instances) annotated in 7 files. Each event mention was annotated with attribution values by two annotators.

The results are presented in Table 3. The agreement is over 0.90 for certainty, time and polarity attributes. For the special_cases attribute, the agreement is over 0.80, with main disagreements on "general statement" events.

4.2.2 Italian

The annotation of the Italian translation of the corpus through the alignment procedure was performed by an expert annotator who is a native speaker of Italian. As no other Italian speaker annotator was part of the consortium during the annotation phase it was impossible to compute data about inter-annotator agreement for Italian. However the data on inter-annotator agreement provided for English can be used as a reference as the guidelines followed for Italian do not differ from those for English, except for a small number of linguistic phenomena that are not present in English (see Section 2.3).

4.2.3 Spanish

The Spanish intra-document annotation was carried out by two native speakers of Spanish. One of the annotators had already taken part in the annotation of English documents and

	Airbus-Boeing	GM-Chrysler-Ford	Stock market	Total
# files	30	30	30	90
# sentences	446	430	459	1,335
# tokens	9,909	10,058	9,916	29,893
# seed entities	13	12	13	38
# event instances	260	248	220	728
# corefering events	70	45	36	151
# cross-doc corefering events	14	9	7	30

Table 2: Cross-document annotation in the Wikinews subcorpora

	certainty	time	special_cases	polarity
Dice's coefficient	0.94	0.90	0.84	0.94

Table 3: Inter-annotator agreement for attribution value annotation

was in charge of training the new annotator as well as reviewing the latter's annotation task. Besides, since Spanish annotation was mostly based on the English guidelines (see Section 2.3), inter-annotator agreement was not measured.

4.2.4 Dutch

The Dutch intra-document annotation effort was split over 9 different annotators each of which only took care of one layer at a time in the annotations[8]. This way, the annotators could focus on one part of the annotation without having to concern themselves with the entire annotation guidelines, speeding up training.

The tasks took place in two phases and were divided as follows:

Phase 1: November 2014

- Entity mentions and instances
- Event mentions and instances
- Temporal expressions
- Numeric expressions, signals and c-signals

Phase 2: December-January 2014

- Has participant relationships
- Attribution
- Temporal links

[8]Some of the annotators took on more than one task

Markables		
	macro-average (markable)	macro-average (token)
SIGNAL	0.666666666667	0.666666666667
VALUE	0.833333333333	0.833333333333
C-SIGNAL	1.0	1.0
TIMEX3	0.974358974359	1.0
EVENT_MENTION	0.779797979798	0.923245614035
ENTITY_MENTION	0.818273976257	0.849865047233
Relations one to one		
	macro-average	macro-average
TLINK	0.391666666667	0.321136173768
CLINK	1.0	1.0
HAS_PARTICIPANT	0.767773892774	0.609925558313
GLINK	1.0	0.333333333333
SLINK	1.0	0.7
Relations many to one		
	global alpha	
REFERS_TO	0.45152158528	

Table 4: Inter-annotator agreement Dutch gold standard annotation

- Slink, glink and clink

The event and entity mentions and instances were shared between two annotators, each annotation the entities and events in 60 articles. For all other tasks, the annotators worked on all 120 articles.

All annotators were asked to annotate three documents from the Apple corpus to compute the inter-annotator agreement against the gold standard created by one of the Dutch trainers. The results are presented in Table 4. Most reported scores are Dice's coefficient, except for the many to one relations in the last row, for which we report the global alpha score. As with the English annotation, the TLINKs provided the greatest challenge for the annotators. On the other markables and relations a high inter-annotator agreement is achieved.

4.3 Inter-Annotator Agreement in cross-document annotation for English

We measure inter-annotator agreement on both mention extents and instances with the Dice's coefficient.

Three annotators have annotated a corpus of 30 documents starting from one seed entity, i.e. they have annotated entity coreferences refering to the seed entity and the events in which the seed entity is a participant. This annotation has been done in the full text. The corpus is composed by articles from WikiNews about *Apple Inc.* and the seed

entity is *iPhone 4*. We first compute the agreement scores by pairs of annotators and then the macro-average on the pairwise scores.

The scores are given in Table 5. The results are satisfactory, with the agreement macro-average above 0.80 for entity mentions and coreference relations (REFERS_TO). For event mentions and event instances annotation the agreement is above 0.65. One reason for this difference is that for entity mentions and entity coreferences annotation the Guidelines are similar to the one used for intra-document annotation, while for event mentions and instances annotation the guidelines are specific to the cross-document annotation task.

	macro-average
ENTITY (product)	0.81
EVENT (speech-cognitive or other)	0.66
REFERS_TO	0.84
EVENT INSTANCES	0.68

Table 5: Inter-annotator agreement on the cross-document annotation task

5 Intra-document Benchmarking

5.1 English

5.1.1 Named Entity Recognition and Classification

The term 'Named Entity', now widely used in Natural Language Processing, was coined for the Sixth Message Understanding Conference (MUC-6) (Grishman and Sundheim 1996). At that time, MUC was focusing on Information Extraction (IE) tasks where structured information of company activities and defense related activities is extracted from unstructured text, such as newspaper articles. In defining the task, people noticed that it is essential to recognize information units like names, including person, organization and location names, and numeric expressions including time, date, money and percent expressions. Identifying references to these entities in text was recognized as one of the important sub-tasks of IE and was called "Named Entity Recognition and Classification (NERC)".

While early systems were making use of handcrafted rule-based algorithms, modern systems most often resort to machine learning techniques. It was indeed concluded in the influential CoNLL 2003 shared task that the choice of features is at least as important as the choice of technique for obtaining a good NERC system Tjong Kim Sang and De Meulder (2003). Moreover, it was also shown that the way NERC systems are evaluated and compared is essential to the progress in the field. Current NERC systems used supervised training over hand-annotated data to learn a statistical model for annotating Named Entities. This means that on the domain and text genre on which the model is trained current systems can obtain high performance scores in terms of phrase-based F1 score, where a named entity is correct if and only if the span and the class of the entity is exactly identified

and classified. At the same, this also means these models performs poorly when applied to a different domain or text genre.

Perhaps unsurprisingly, porting a system to a new domain or textual genre remains a major challenge. Some experiments tested some systems on both the MUC-6 collection composed of newswire texts, and on a proprietary corpus made of manual translations of phone conversations and technical emails Poibeau and Kosseim (2001). They reported a drop in performance for every system (between 20% to 40% of precision and recall). These results have been later confirmed in other more recent works Ratinov and Roth (2009).

Overall, the most studied Named Entity types are three specializations of "proper names": names of "persons", "locations" and "organizations". These types are collectively known as "enamex" since the MUC-6 competition. The type "location" can in turn be divided into multiple subtypes of "fine-grained locations": city, state, country, etc. Fleischman and Hovy (2002). Similarly, "fine-grained person" sub-categories like "politician" and "entertainer" appear in the aforementioned work Fleischman and Hovy (2002). In the ACE program, the type "facility" subsumes entities of the types "location" and "organization", and the type "GPE" is used to represent a location which has a government, such as a city or a country. The type "miscellaneous" is used in the CONLL conferences and includes proper names falling outside the classic "enamex". The class is also sometimes augmented with the type "product" Bick (2004). The "timex" (also coined in MUC) types "date" and "time" and the "number" types "money" and "percent" are also quite predominant in the literature. Finally, the Ontonotes corpus define 18 different named entity types Weischedel *et al.* (2010).

Most approaches rely on manually annotated newswire corpora, namely, in the MUC 6 and 7 (Grishman and Sundheim 1996; Chinchor 1998) conference at the beginning but today mostly on the CONLL 2003 dataset which consists of 4 class entities (person, location, organization and miscellaneous) manually annotated on a subset of the Reuters corpus. Most of the systems consist of language independent systems based on automatic learning of statistical models (for technical details of these approaches see Nadeau and Sekine (2007)). However, this reliance on expensively manually annotated data hinders the creation of NERC systems for most languages and domains.

The 2002 and 2003[9] CoNLL shared tasks provided manually annotated datasets for German, English (2003 edition) and Dutch and Spanish (2002 edition). The data consists of columns separated by a single space. The first item on each line is a word and the last one the named entity tag. An example is shown in Figure 2.

The English data is a collection of news wire articles from the Reuters Corpus[10], in total, 301418 annotated tokens for dev/train/test datasets are provided. Due to copyright issues only the annotations were made available at CONLL 2003 and to build the complete datasets it is necessary to access the Reuters Corpus, which can be obtained from NIST for research purposes. They also provide an official evaluation script[11] which is the one

[9] `http://www.cnts.ua.ac.be/conll2003/ner/`

[10] `http://trec.nist.gov/data/reuters/reuters.html`

[11] `http://www.cnts.ua.ac.be/conll2002/ner/bin/conlleval.txt`

```
Wolff B-PER
, O
currently O
a O
journalist O
in O
Argentina I-LOC
, O
played O
with O
Del I-PER
Bosque I-PER
```

Figure 2: Example of CoNLL format for NERC with the NewsReader system output

used in Newsreader to measure the performance of the NERC taggers with respect to the wikinews gold standard annotated within the project.

In Newsreader we use the *ixa-pipe-nerc* system[12] Agerri *et al.* (2014) off-the-self to train our NERC models; *ixa-pipe-nerc* learns supervised models via the Perceptron algorithm as described by Collins (2002). To avoid duplication of efforts, *ixa-pipe-nerc* uses the Apache OpenNLP project implementation of the Perceptron algorithm[13] customized with its own features. Specifically, *ixa-pipe-nerc* implements basic non-linguistic local features and on top of those a combination of word class representation features partially inspired by Turian *et al.* (2010). The word representation features use large amounts of unlabeled data. The result is a quite simple but competitive system which obtains the best results for English both on the CoNLL 2003 dataset and on Wikinews.

The local features implemented are: current token and token shape (digits, lowercase, punctuation, etc.) in a 2 range window, previous prediction, beginning of sentence, 4 characters in prefix and suffix, bigrams and trigrams (token and shape). On top of them we induce three types of word representations:

- Brown Brown *et al.* (1992) clusters, taking the 4th, 8th, 12th and 20th node in the path. We induced 1000 clusters on the Reuters RCV1 corpus using the tool implemented by Liang[14].
- Clark Clark (2003) clusters, using the standard configuration to induce 600 clusters on the Reuters RCV1 corpus.
- Word2vec Mikolov *et al.* (2013) clusters, based on K-means applied over the extracted word vectors using the skip-gram algorithm[15]; 400 clusters were induced using the

[12] https://github.com/ixa-ehu/ixa-pipe-nerc
[13] http://opennlp.apache.org/
[14] https://github.com/percyliang/brown-cluster
[15] https://code.google.com/p/word2vec/

English Wikipedia.

The implementation of the clustering features looks for the cluster class of the incoming token in one or more of the clustering lexicons induced following the three methods listed above. If found, then the class is added as a feature. The Brown clusters only apply to the token related features, which are duplicated. We chose the best combination of features on the CoNLL 2003 testa dataset, which corresponds to the configuration we have just described.

First we evaluate our NERC system on the CoNLL 2003 official testset. For this evaluation, we added to our system the publicly available gazetteers from the Illinois NER Tagger Ratinov and Roth (2009). The results obtained by *ixa-pipe-nerc* are the best of of any publicly available system up to date Ratinov and Roth (2009), and comparable to the best published results Passos *et al.* (2014) on this dataset, as shown in Table 6.

Table 6: NERC CoNLL 2003 testb results.

	Precision	Recall	F1
Newsreader (ixa-pipe-nerc)	91.64	90.21	**90.92**
Stanford NER	-	-	88.08
Ratinov et al. (2009)	-	-	90.57
Passos et al. (2014)	-	-	90.90

Using the same configuration as the one tested on the CoNLL 2003 dataset, we next evaluated the model using the Wikinews testset annotated within the Newsreader project. This evaluation constitutes a hard *out of domain* evaluation because even though the gold standard is news, the text style is quite different to that of Reuters. Furthermore, and most importantly, the type of named entities annotated in the Wikinews corpus is very different to the type of annotation done in the CoNLL 2003 dataset. In other words, the criteria for a string of text to be considered *the extent of a named entity* greatly differ, which makes the NERC evaluation in terms of F1 quite hard.

Moreover, Wikinews contains annotation for nested entities, so we present in Table 7 the results of the ixa-pipe-nerc best model on the Wikinews corpus in terms of phrase- and token-based F1 for both inner and outer extents for the 3 classes which map from the CoNLL 2003 to the Wikinews datasets, namely, person, organization and location.

Table 7 shows the results of the best ixa-pipe-nerc model obtained as evaluated on the CoNLL 2003 but using as training not only the CoNLL 2003 trainset but also the MUC7 and Ontonotes 4.0 datasets. It also shows the performance of the best model, on the Wikinews dataset, that the Stanford NER system distributes. This model is trained only for three entity types (person, location and organization) on a variety of datasets, including MUC6, MUC7, Ontonotes, Web data and CoNLL 2003.

It can be seen that *ixa-pipe-nerc* outperforms Stanford NER on every dataset and type of evaluation, the differences between them being larger in the standard phrase-based F1 score. The comparatively low scores also confirm the difficulty of adapting supervised

Table 7: NERC Intra-document Benchmarking with Wikinews.

System	mention extent	Precision	Recall	F1
Newsreader (ixa-pipe-nerc)	Inner phrase-based	62.15	76.06	**68.41**
Stanford NER (all english crf distsim)	Inner phrase-based	63.53	68.21	65.79
Newsreader (ixa-pipe-nerc)	Inner token-based	72.17	79.31	**75.57**
Stanford NER (all english crf distsim)	Inner token-based	77.14	71.77	74.36
Newsreader (ixa-pipe-nerc)	Outer phrase-based	53.01	68.03	**59.59**
Stanford NER (all english crf distsim)	Outer phrase-based	52.86	59.51	55.99
Newsreader (ixa-pipe-nerc)	Outer token-based	73.40	67.20	**70.16**
Stanford NER (all english crf distsim)	Outer token-based	78.22	60.63	68.31

models to the Wikinews dataset, although the results for the token based evaluation are higher.

Thus, the results are coherent with the previous assertions on out of domain evaluation. This in particular is not surprising if we consider that the wikinews gold standard was annotated with a completely different guidelines stating as to what a *named entity* is. For example, a frequent source of false positives are the following cases, among others:

- Different criteria to decide a Named Entity is marked: in the expression "40 billion US air tanker contract" the Wikinews gold standard does not mark 'US' as location, where as in the CoNLL 2003 guidelines this is systematically annotated.

- 'the United States' in Wikinews gold standard vs. 'United States' in CoNLL 2003.

- Longer extents containing common nouns: in Wikinews corpus there are many entities such as "United States airframer Boeing" which in this case is considered an organization, whereas in CoNLL 2003 this extent will be two entities: 'United States' as location and 'Boeing' as organization.

- Common nouns modifying the proper name: 'Spokeswoman Sandy Angers' is annotated as a Named Entity of type person whereas in CoNLL 2003 the extent would be 'Sandy Angers' only.

Summarizing, the NERC system integrated in Newsreader obtains the best results in the very competitive CoNLL 2003 evaluation. Furthermore, *ixa-pipe-nerc* clearly outperforms a robust Stanford NER system by around 3-4 F1 scores in the phrase based evaluations and by around 1.5-2 F1 scores in the token based evaluations. These differences are even larger if the CoNLL-only models are used of the multi corpora ones. The results can be reproduced following the procedure explained in the nerc evaluation package[16].

[16] `https://github.com/newsreader/evaluation/tree/master/nerc-evaluation`

5.1.2 Named Entity Disambiguation

In NewsReader, Named Entity Disambiguation is performed using the DBpedia Spotlight technology. More specifically, we use the DBpedia Spotlight probabilistic models. For the evaluation, we will be using the 2010, 2011 English dataset from the TAC KBP editions and the AIDA corpus.

The **AIDA corpus** contains assignments of entities to the mentions of named entities annotated for the original CoNLL 2003 NERC task.[17] The entities are identified by YAGO2[18] entity name, by Wikipedia URL or by Freebase.[19] The CoNLL 2003 dataset is required to create the corpus which in turn requires the Reuters corpus, available from LDC.

The **TAC KBP** 2009 edition distributed a knowledge base extracted from a 2008 dump of Wikipedia and a test set of 3904 queries. Each query consisted of an ID that identified a document within a set of Reuters news articles, a mention string that occurred at least once within that document, and a node ID within the knowledge base. Each knowledge base node contained the Wikipedia article title, Wikipedia article text, a predicted entity type (person, organization, location or misc), and a key-value list of information extracted from the article's infobox. Only articles with infoboxes that were predicted to correspond to a named entity were included in the knowledge base. The annotators favoured mentions that were likely to be ambiguous, in order to provide a more challenging evaluation. If the entity referred to did not occur in the knowledge base, it was labelled NIL. A high percentage of queries in the 2009 test set did not map to any nodes in the knowledge base: the gold standard answer for 2229 of the 3904 queries was NIL.

In the 2010 challenge the same configuration as the 2009 challenge was used with the same knowledge base. In this edition, however, a training set of 1500 queries was provided, with a test set of 2250 queries. In the 2010 training set, only 28.4% of the queries were NIL, compared to 57.1% in the 2009 test data and 54.6% in the 2010 test data. This mismatch between the training and test data show the importance of the NIL queries and it is argued that it may have harmed performance for some systems because it can be quite difficult to determine whether a candidate that seems to weakly match the query should be discarded, in favour of guessing NIL. The most successful strategy to deal with these issue in the 2009 challenge was augmenting the knowledge base with extra articles from a recent Wikipedia dump. If a strong match against articles that did not have any corresponding node in the knowledge base was obtained, then NIL was return for these matches. In the KBP 2012 edition, the reference KB is derived from English Wikipedia, while source documents come from a variety of languages, including English, Chinese, and Spanish.

The evaluation consists of running the NED system on the standard datasets described above, assessing their overall performance. This section presents the results of this evaluation. The performance of the system is measured using the standard precision and recall metrics. Precision is the number of correctly assigned instances divided by the total in-

[17] `http://www.cnts.ua.ac.be/conll2003/ner/`
[18] `http://www.mpi-inf.mpg.de/yago-naga/yago/`
[19] `http://wiki.freebase.com/wiki/Machine_ID`

stances as returned by the system. Recall is the number of correctly assigned instances divided by the number of instances in the gold standard dataset. In our particular setting, we seek to maximize precision, that is, we care more about returning correct links to DBpedia entities than trying to link all possible mentions in the input text. Because we focus our study on NED systems, we discard the so-called NIL instances (instances for which no correct entity exists in the Reference Knowledge Base) from the datasets.

As the module has several parameters, it was optimized in TAC 2010 dataset. Using the best parameter combination, the module has been evaluated on two datasets: TAC 2011 and AIDA. The best results obtained on the first dataset were 79.77 in precision and 60.68 in recall. The best performance on the second dataset is 79.67 in precision and 75.94 in recall.

We have also checked the performance of the NED module on the WikiNews gold standard of the NewsReader project. We have used a subgroup of the 120 files from the dataset: the airbus and stock corpora. We have evaluated the entities disambiguated in the first six sentences of the 60 documents. Table 8 presents the evaluation results, the number of entities manually annotated, the number of entities automatically identified by the NERC module and the number of entities disambiguated by the NED module. The precision and recall are obtained comparing the manually disambiguated entities with the information contained in the entities layer of the NAF files obtained with the NewsReader pipeline.

Corpus	Precision	Recall	Gold	System-NERC	System-NED
Airbus	60.41%	41.42%	420	316	288
Stock	58.75%	28.25%	368	194	177
Total	59.78%	35.28%	788	510	465

Table 8: Performance of the NED module on the WikiNews dataset

The results obtained in the wikinews dataset are lower compared to the ones obtained in the TAC and AIDA datasets. The differences are bigger with respect to the recall values. The number of entities automatically detected by the NERC module is lower than the ones manually annotated and the NED module also fails when disambiguating some of the entities (see Table 8). In addition, we do not take into account the information automatically obtained regarding the nominal coreference for the evaluation.

Finally, we have also evaluated the NewsReader pipeline based on the annotation done for the SemEval-2015 task 4. In this case, we evaluated the pipeline with a set of target entities. Thus, the manually disambiguated entities are not all the entities appearing in the documents but the ones corresponding to the core entities. The evaluation data consist of 3 sets of documents annotated with a set of target entities and each set contains 30 documents. The precision of the system is not possible to obtain because not all the entities were manually disambiguated so we have only measured the recall of our pipeline. Table 9 presents the results. If we compare the results with ones obtained in the wikinews dataset, it seems the NewsReader pipeline obtains less differences between the corpora

in the case of the SemEval entities. The information evaluated corresponds to centroid entities.

Corpus	Recall	Gold
Airbus	34.55%	544
GM	31.80%	585
Stock	32.61%	279
Total	33.02%	1408

Table 9: Performance of the NED module on the WikiNews dataset

We are going to perform an error analysis of the results and an analysis of the information contained in the coreference layer to help in the improvement of the disambiguation task.

5.1.3 Event detection and coreference

Event detection and event coreference are important for NewsReader since they capture the core of the news. Event detection and event-coreference is very different from entity detection and linking. Firstly, events are not as tangible as entities and only exists during a very limited time-frame. What constitutes an event is not easy to define and to some degree also very subjective. Secondly, most events are not registered in a resource as entities are in e.g. DBPedia. As such events are more fluid and temporarily. Likewise, it is not surprising that the way people can refer to an event also varies a lot. We can thus conclude that event detection and coreference is by far more challenging than entity detection, linking and coreference. Event coreference is important because it forms the basis for defining event instances within and across documents. In this section, we describe the evaluation of detecting events mentions and instances within a single document using the Wikinews corpus.

The Semantic Role layer (SRL) is the basic input for the event detection and coreference. We take the predicates that are listed in the SRL as a starting point for creating coreference sets in the coreference layer. Any predicate detected by the SRL is potential event. The coreference layer thus includes both singleton sets (mentions of predicates that do not corefer with other mentions) and multitude set (two or more predicates referring tot the same event). Coreference sets are created using different methods, described below.

Event-coreference is measured in different ways in the literature. We use the method BLANC Pradhan *et al.* (2014) for the intra-document coreference because it also measures singleton coreference sets. In most cases, there is no coreference relation and a mention of an event within a document represents a unique event only mentioned once. Measures that only consider co-referrering event mentions therefore do not consider these non-coreferential event mentions.

For the evaluation of the event-coreference, we used the CorScorer package[20] developed by Luo *et al.* (2014). The CorScorer expects that coreferences are represented in

[20] `https://code.google.com/p/reference-coreference-scorers/`

CoNLL2011/2012 format. We thus developed a package[21] that converts CAT annotations to this format and NAF representations. An example of the output format shown in 3.

```
#begin document (3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners);
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 1 1 Chinese -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 1 2 airlines -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 1 3 agree (9)
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 1 4 purchase (10)
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 1 5 of -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 1 6 Boeing -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 1 7 787 -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 1 8 Dreamliners -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 1 9 worth -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 1 10 US$ -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 1 11 7.2 -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 1 12 bn (11)

3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 17 Officials (12)
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 18 from -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 19 the -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 20 People -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 21 's -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 22 Republic -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 23 of -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 24 China -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 25 have -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 26 agreed (9)
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 27 to -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 28 purchase (10)
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 29 60 -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 30 Boeing -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 31 787 -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 32 Dreamliner -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 33 aircraft -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 34 in -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 35 a -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 36 deal (13)
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 37 worth -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 38 US$ -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 39 7.2 -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 40 bn -
3835_Chinese_airlines_agree_purchase_of_Boeing_787_Dreamliners 3 41 . -
```

Figure 3: Example of CoNLL20111/2012 format for coreference with the NewsReader system output

The NewsReader pipeline annotates all sentences of an article. Since only a few sentences from each article have been annotated, we implemented a function that reduces the CoNLL file for the manual annotation to only those sentences that have an event annotated and we implemented another function that reduces the system output (response) to the same sentences of the manual CoNLL file (key). We developed 4 different systems for the intra-document coreference:

No-coreference baseline Every predicate in the SRL represents a unique event and no coreference relations are created. Since in most cases, there is indeed no co-reference

[21] `https://github.com/cltl/coreference-evaluation`

within a document (events are mentioned only once) this constitutes a good baseline for the majority case.

Lemma baseline All predicates in the SRL that have the same lemma are coreferential, following the hypothesis *one-lemma-one-instance-per-document*.

Wordnet-similarity-version1 The output of the lemma baseline is taken as a starting point but a Wordnet Similarity method defined by Leacock and Chodorow (1998) that is used to compare lemma-based coreference sets. If two different lemmas score higher than 2.0 then the sets are merged and represented by the lowest common subsumer (LCS) synset. Since all lemmas are compared, sets can be chained into longer sets of lemmas with different LCSs. We use WordNet3.0-LMF as a resource and the WordNetTools package[22] to measure the similarity. We extended the relations with 17,739 synset relations based on the Princeton morphosemantic relation file[23]. Through these relations, we can establish cross-part-of-speech similarity in addition to similarity based on solely the hypernym relations in WordNet.

Wordnet-similarity-version2 Takes the output of Wordnet-similarity-version1 as a starting point but merged sets with more than 3 LCS synsets are considered unstable and are dissolved, i.e. only the lemma subsets are maintained. This method tries to undo concept-drift due to chaining events on different senses of the same lemmas, e.g.: Sim (lemma-A, lemma-B) >2.0) AND Sim (lemma-B, lemma-C)>2.0 but Sim (lemma-A, lemma-C) <2.0.

In the next tables you see the results for the 4 systems. The results are assembled by applying the CorScorer8.01 using the BLANC method on the CAT-annotated Wikinews article with the NewsReader pipeline version 2.1 output. We collected the results for each subcorpus and averaged the results. We applied macro averaging by taking the average over the BLANC scores for all documents. We applied micros averaging by collecting all mentions and coreference links (both annotated key and system response) for the whole corpus and calculating the recall, precision and F-measure using these totals.

We give two tables for each method. The first table gives the totals for each subcorpus and the second table gives the macro and micro averaged results. Furthermore, we provide an analysis of the detection of the mentions of events and of the coreference links.

In Tables 10 and 11, we see the results of the baseline where no coreference relations are established in the system output, i.e every event is a singleton set. We first see that the system detects slightly more mentions of events than annotated but also misses some. *Strictly correct identified mentions* indicates the matches. Since there are no links created, BLANC only reports on the non-coreference response links, whereas the coreference response links are zero. All the totals are consistent across the 4 subcorpora.

When we look at the results for the event mentions in Table 11, we see that the overall scores are reasonable: F69.38 macro and F69.99 micro averaged. This is important because

[22] `git@github.com:cltl/WordnetTools.git`

[23] `http://wordnetcode.princeton.edu/standoff-files/morphosemantic-links.xls`

Table 10: Singleton events

	stock market	gm chrysler ford	apple	airbus	Average
Total key mentions	513	549	455	502	505
Total response mentions	560	630	614	596	600
Total missed mentions	176	213	258	206	213
Total invented mentions	129	132	99	112	118
Strictly correct identified mentions	384	417	356	390	387
Total key reference links	4528	5449	3491	4371	4460
Total response reference links	5349	6977	6316	6024	6167
Correct reference links	2523	3012	2085	2578	2550
Total key coreference links	91	246	116	86	135
Total response coreference links	0	0	0	0	0
Correct coreference links	0	0	0	0	0
Total key non-coreference links	4437	5203	3375	4285	4325
Total response non-coreference links	5349	6977	6316	6024	6167
Correct non-coreference links	2523	3012	2085	2578	2550

mention detection has a direct impact on the quality of the coreference detection. Again results are consistent across the 4 sets and differences between macro and micro averaging are small. Recall is a bit higher than precision, which suggest that events could be filtered more.

When we look at the coreference results, we see F24.4 and F47.92 for macro and micro averaging. Both scores are low and solely based on the mentions without coreference relation, which is the majority case. Micro averaged results are higher. Since the amount of the annotation is the same across the articles, this difference may be due to differences in the density of annotation per document. It is not so easy to interpret these differences though.

These results form the basis for the comparison of the other methods that add different degrees of coreference relations to mentions. We first consider the result for the lemma baseline, see Tables 12 and 13. The main difference is in the coreference response links, for which we now do get results from BLANC. We can see that the coreference links only represent a small proportion of the events, compared to the non-coreference links.

We see that the mention scores are the same, which is what we expect. The coreference macro scores are significantly higher: F42 versus F24.4, while micro results are slightly higher: F48.13 versus F47.92. The difference is equally divided over recall and precision.

The next tables show the results using Wordnet Similarity to merge lemma-based coreference sets. Again the results for mentions are similar and differences are mainly found for the coreference relations. Remarkably, the macro-averaged results are lower than the lemma baseline but higher than the no-coreference baseline, while the micro-averaged re-

Table 11: Singleton events, BLANC scores

Macro average mentions	Recall	Precision	F1
airbus	77.91	65.88	70.36
apple	78.14	58.12	65.76
gm_chrysler_ford	77.63	66.59	71.16
stock_market	74.14	67.71	70.24
Average	76.96	64.58	69.38
Micro average mentions	Recall	Precision	F1
airbus	77.69	65.44	71.04
apple	78.24	57.98	66.60
gm_chrysler_ford	75.96	66.19	70.74
stock_market	74.85	68.57	71.58
Average	76.69	64.54	69.99
Macro average coreference	Recall	Precision	F1
airbus	32.38	22.32	24.90
apple	35.09	18.44	23.11
gm_chrysler_ford	31.15	21.18	24.40
stock_market	30.10	23.27	25.21
Average	32.18	21.30	24.40
Micro average coreference	Recall	Precision	F1
airbus	58.98	42.80	49.60
apple	59.73	33.01	42.52
gm_chrysler_ford	55.28	43.17	48.48
stock_market	55.72	47.17	51.09
Average	57.43	41.54	47.92

Table 12: Lemma-baseline

	stock market	gm chrysler ford	apple	airbus	Average
Total key mentions	513	549	455	502	505
Total response mentions	560	630	614	596	600
Total missed mentions	176	213	258	206	213
Total invented mentions	129	132	99	112	118
Strictly correct identified mentions	384	417	356	390	387
Total key reference links	4528	5449	3491	4371	4460
Total response reference links	5349	6977	6316	6024	6167
Correct reference links	2485	3081	2110	2576	2563
Total key coreference links	91	246	116	86	135
Total response coreference links	132	147	112	112	126
Correct coreference links	26	93	44	37	50
Total key non-coreference links	4437	5203	3375	4285	4325
Total response non-coreference links	5217	6830	6204	5912	6041
Correct non-coreference links	2459	2988	2066	2539	2513

sults are even slightly lower than the no-reference baseline. Micro-averaged results are just slight different from the baseline results. We thus focus on the macro-averaged results. The recall for macro-averaged results is higher than the lemma-baseline (R52.22 over R51.15) but precision is much lower (P33.35 against P42.09). We thus recover more coreference relations through the WordNet similarity approach but the concept-drift is too high causing a bad recall.

The 4th method is supposed to correct the concept-drift. An example of such drift is shown in 4, where the lemmas *say*, *record*, *call*, *play* and *hit* are merged. The LCSs for these lemmas are shown in the external references. We have set the maximum number of LCSs to 3. Cases as shown in 4 are resolved by falling back on the lemmas for the coreference sets.

The results for the approach that controls for concept-drift are shown in the tables 16 and 17. We can see in Table that the results recover a bit in precision (P36.02 against P33.35) and also in f-measure (F38.98 against F36.91) but we loose some recall (R51.75 against R52.22). Still, we do not exceed the lemma-baseline. Macro-recall is a bit higher than the lemma approach but precision is still a 6 points lower. Micro-results are also a bit lower than the lemma results: F47.45 against F 48.13.

Table 13: Lemma baseline, BLANC scores

Macro average mentions	Recall	Precision	F1
airbus	77.91	65.88	70.36
apple	78.14	58.12	65.76
gm_chrysler_ford	77.63	66.59	71.16
stock_market	74.14	67.71	70.24
Average	76.96	64.58	69.38
Micro average mentions	Recall	Precision	F1
airbus	77.69	65.44	71.04
apple	78.24	57.98	66.60
gm_chrysler_ford	75.96	66.19	70.74
stock_market	74.85	68.57	71.58
Average	76.69	64.54	69.99
Macro average coreference	Recall	Precision	F1
airbus	53.65	41.40	42.97
apple	50.32	37.38	38.96
gm_chrysler_ford	56.15	51.72	49.23
stock_market	44.49	37.88	36.85
Average	51.15	42.09	42.00
Micro average coreference	Recall	Precision	F1
airbus	58.93	42.76	49.56
apple	60.44	33.41	43.03
gm_chrysler_ford	56.54	44.16	49.59
stock_market	54.88	46.46	50.32
Average	57.70	41.70	48.13

Table 14: Wordnet Similarity, threshold 2.0, cross-part-of-speech relations

	stock market	gm chrysler ford	apple	airbus	Average
Total key mentions	513	549	455	502	505
Total response mentions	557	627	613	595	598
Total missed mentions	174	215	258	204	213
Total invented mentions	130	137	100	111	120
Strictly correct identified mentions	383	412	355	391	385
Total key reference links	4528	5449	3491	4371	4460
Total response reference links	5296	6919	6302	6010	6132
Correct reference links	2352	2899	2044	2491	2447
Total key coreference links	91	246	116	86	135
Total response coreference links	318	433	208	269	307
Correct coreference links	35	124	46	41	62
Total key non-coreference links	4437	5203	3375	4285	4325
Total response non-coreference links	4978	6486	6094	5741	5825
Correct non-coreference links	2317	2775	1998	2450	2385

```
<coref id="coevent23" type="event">
   <!--say-->
   <span><target id="t16"/></span>
   <!--said-->
   <span><target id="t292"/></span>
   <!--said-->
   <span><target id="t366"/></span>
   <!--record-->
   <span><target id="t151"/></span>
   <!--calling-->
   <span><target id="t237"/></span>
   <!--played-->
   <span><target id="t287"/></span>
   <!--hit-->
   <span><target id="t400"/></span>
   <externalReferences>
     <!-- read:8, register:5, show:9, record:3 -->
     <externalRef resource="Princeton WordNet 3.0" reference="eng-30-00922867-v" confidence="2.1972246"/>
     <!-- order:1, tell:4, enjoin:2, say:5 -->
     <externalRef resource="Princeton WordNet 3.0" reference="eng-30-00746718-v" confidence="2.1972246"/>
     <!-- play:1 -->
     <externalRef resource="Princeton WordNet 3.0" reference="eng-30-01072949-v" confidence="2.1972246"/>
     <!-- hit:1 -->
     <externalRef resource="Princeton WordNet 3.0" reference="eng-30-01405044-v" confidence="2.0794415"/>
   </externalReferences>
 </coref>
```

Figure 4: Coreference set with more than 3 Lowest-Common-Subsumers

Table 15: Wordnet Similarity, threshold 2.0, cross-part-of-speech relations, BLANC scores

Macro average mentions	Recall	Precision	F1
airbus	78.19	66.24	70.69
apple	77.94	58.03	65.61
gm_chrysler_ford	76.72	66.09	70.46
stock_market	73.88	67.92	70.23
Average	76.68	64.57	69.25
Micro average mentions	Recall	Precision	F1
airbus	77.89	65.71	71.29
apple	78.02	57.91	66.48
gm_chrysler_ford	75.05	65.71	70.07
stock_market	74.66	68.76	71.59
Average	76.40	64.52	69.86
Macro average coreference	Recall	Precision	F1
airbus	54.59	32.01	35.91
apple	50.64	29.82	34.84
gm_chrysler_ford	57.78	42.05	44.33
stock_market	45.88	29.51	32.56
Average	52.22	33.35	36.91
Micro average coreference	Recall	Precision	F1
airbus	56.99	41.45	47.99
apple	58.55	32.43	41.74
gm_chrysler_ford	53.20	41.90	46.88
stock_market	51.94	44.41	47.88
Average	55.17	40.05	46.12

Table 16: Wordnet Similarity, threshold 2.0, cross-part-of-speech relations but using drift-control max of 3 most-common subsumers per coreference set

	stock market	gm chrysler ford	apple	airbus	Average
Total key mentions	513	549	455	502	505
Total response mentions	558	630	614	596	600
Total missed mentions	176	213	258	206	213
Total invented mentions	131	132	99	112	119
Strictly correct identified mentions	382	417	356	390	386
Total key reference links	4528	5449	3491	4371	4460
Total response reference links	5313	6977	6316	6024	6158
Correct reference links	2413	3058	2084	2549	2526
Total key coreference links	91	246	116	86	135
Total response coreference links	223	240	159	178	200
Correct coreference links	29	112	46	40	57
Total key non-coreference links	4437	5203	3375	4285	4325
Total response non-coreference links	5090	6737	6157	5846	5958
Correct non-coreference links	2384	2946	2038	2509	2469

Table 17: Wordnet Similarity, threshold 2.0, cross-part-of-speech relations but using drift-control max of 3 most-common subsumers per coreference set, BLANC scores

Macro average mentions	Recall	Precision	F1
airbus	77.91	65.88	70.36
apple	78.14	58.12	65.76
gm_chrysler_ford	77.63	66.59	71.16
stock_market	73.71	67.59	69.98
Average	76.85	64.54	69.31
Micro average mentions	Recall	Precision	F1
airbus	77.69	65.44	71.04
apple	78.24	57.98	66.60
gm_chrysler_ford	75.96	66.19	70.74
stock_market	74.46	68.46	71.34
Average	76.59	64.52	69.93
Macro average coreference	Recall	Precision	F1
airbus	53.70	35.49	39.12
apple	50.99	32.31	36.42
gm_chrysler_ford	57.74	45.41	46.79
stock_market	44.59	30.89	33.59
Average	51.75	36.02	38.98
Micro average coreference	Recall	Precision	F1
airbus	58.32	42.31	49.04
apple	59.70	33.00	42.50
gm_chrysler_ford	56.12	43.83	49.22
stock_market	53.29	45.42	49.04
Average	56.86	41.14	47.45

In Table 18, we summarize the results. Although the lemma-baseline gives the best precision and F-measure, we can see that there is room for increasing the recall by grouping events with different lemmas. However, WordNet similarity is not sufficient as a constraint. We can widen the similarity threshold to include candidates but it will decrease precision. The next figures show the graphs for some more WordNet Similarity experiments we ran. We excluded the drift-factor here for comparison and ran the similarity function with the thresholds 1.0, 1.5, 2.0 and 3.0. For comparison, we included the lemma-baseline. We can clearly see in Figure 5 that recall is highest for sim=2.0. This is remarkable since the lower similarity setting will lump together many events in coreference sets. This is due to the fact that BLANC punishes the systems for the singletons that now get wrongly added to coreference sets. Extreme lumping thus results in less singletons being retrieved. When we look at the precision in Figure 6, we see a straight increase up to sim3.0 which is equal to the lemma-baseline. From experience, we know that a Leacock-Chodorow score of 3.0 or higher is only achieved for lemmas and occasionally for synonyms at deep levels of the hierarchy. The f-measure graph in Figure 7 likewise supports the superiority of the lemma-baseline.

Table 18: Macro-averaged results using BLANC for the 4 different methods

BLANC	**Singleton baseline**			**Lemma baseline**			**WordNet Sim-1**			**WordNet Sim-2**		
Macro average	R	P	F1	R	P	F1	R	P	F1	R	P	F1
airbus	32.38	22.32	24.90	53.65	**41.40**	**42.97**	**54.59**	32.01	35.91	53.70	35.49	39.12
apple	35.09	18.44	23.11	50.32	**37.38**	**38.96**	50.64	29.82	34.84	**50.99**	32.31	36.42
gm_chrysler_ford	31.15	21.18	24.40	56.15	**51.72**	**49.23**	**57.78**	42.05	44.33	57.74	45.41	46.79
stock_market	30.10	23.27	25.21	44.49	**37.88**	**36.85**	**45.88**	29.51	32.56	44.59	30.89	33.59
Average	32.18	21.30	24.40	51.15	**42.09**	**42.00**	**52.22**	33.35	36.91	51.75	36.02	38.98

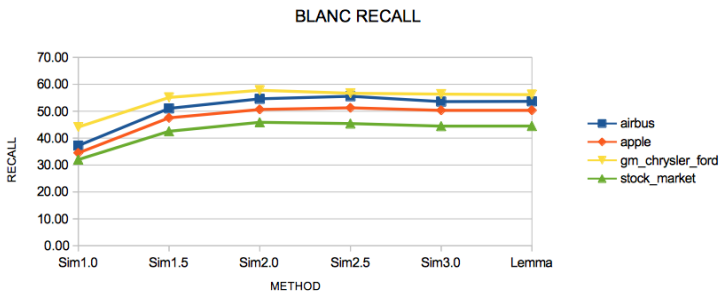

Figure 5: BLANC recall for different Wordnet Similarity values and lemma-baseline

In our future work, we will further experiment with the optimal settings for the WordNet based approach but we will also consider other information than just the predicate to determine coreference. Especially for speech-act events the A0 of events needs to be the same also within the document. This constraint is now only used for cross-document coreference. Furthermore, we can take the sentence distance of predicates into account to

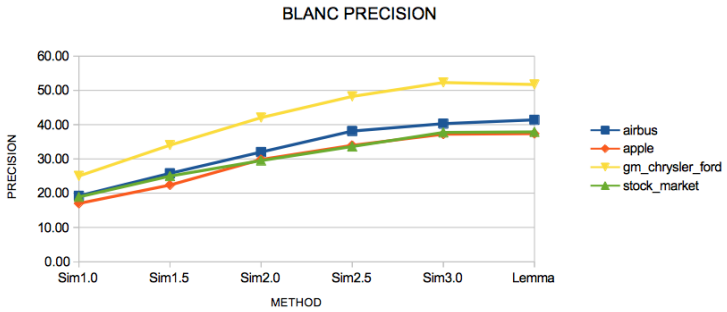

Figure 6: BLANC precision for different Wordnet Similarity values and lemma-baseline

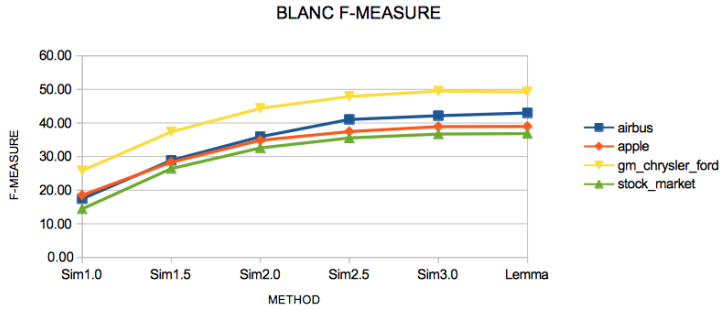

Figure 7: BLANC f-measure for different Wordnet Similarity values and lemma-baseline

establish coreference. Finally, we will try out other clustering methods to group similar predicates.

Overall the results are reasonable given the fact that the mention performance is R76.96, P64.58 and F69.38. The lemma-baseline proportionally performance about 67% recall, 56% precision and 60% F-measure against these totals that provide the maximum possible scores: i.e. we cannot establish correct coreference relations for missed or invented mentions of events.

5.1.4 Nominal coreference

The procedure adopted in Newsreader to perform nominal coreference resolution is based on the Stanford's system Lee *et al.* (2013) and Raghunathan *et al.* (2010); Lee *et al.* (2011). In principle the various versions of the system presented in three different publications consist of several sieve passes, which can be summarized in the 10 passes listed in table 19 Lee *et al.* (2013). The nominal coreference system in the Newsreader pipeline is Corefgraph[24], a loose re-implementation of the Stanford approach for English and Spanish (with ongoing

[24] `https://bitbucket.org/Josu/corefgraph`

development for other languages).

Sieves	Type	CONLL 2011 F1
Mention Detection	NPs, NER and PRP	-
Sieve 1	**Speaker Identification**	**29.2**
Sieve 2	**Exact String Match**	**45.3**
Sieve 3	Relaxed String Match	45.4
Sieve 4	Precise Constructs	45.7
Sieve 5	**Strict Head Match A**	**48.5**
Sieve 6	Strict Head Match B	48.8
Sieve 7	Strict Head Match C	49.3
Sieve 8	Proper Head Noun Match	49.5
Sieve 9	Relaxed Head Match	49.7
Sieve 10	**Pronoun Match**	**59.3**

Table 19: Multi-sieve Pass and CoNLL 2011 dev auto F1 Evaluation.

The Stanford multi-pass sieve coreference resolution (or anaphora resolution) system is described in Lee *et al.* (2013, 2011) and in Raghunathan *et al.* (2010). The approach applies tiers of coreference models one at a time from highest to lowest precision. Each tier builds on the entity clusters constructed by previous models in the sieve, guaranteeing that stronger features are given precedence over weaker ones. Furthermore, each model's decisions are richly informed by sharing attributes across the mentions clustered in earlier tiers. This ensures that each decision uses all of the information available at the time. They implemented all components using only deterministic models. All these components are unsupervised, in the sense that they do not require training on gold coreference links. Furthermore, this framework can be easily extended with arbitrary models, including statistical or supervised models.

This system was the top ranked system at the CoNLL-2011 shared task. The score is higher than that in EMNLP 2010 paper because of additional sieves and better rules (see Lee *et al.* (2013) and Lee *et al.* (2011) for details). Mention detection is included in the package. This is illustrated by the evaluation results of the CoNLL 2011 Coreference Evaluation task Lee *et al.* (2011, 2013), listed in the right hand column of table 19, in which the Stanford's system obtained the best results. The results show a pattern which has also been shown in other results reported with other evaluation sets Raghunathan *et al.* (2010), namely, the fact that a large part of the performance of the multi pass sieve system is based on few of the sieves. Thus, the results show that sieves 1, 2, 5 and 10 provide 97% of the results for that particular evaluation set Lee *et al.* (2011, 2013).

Over the last fifteen years, various competitions have been run to promote research in the field of coreference resolution. The first competition of this kind was MUC, which in its sixth edition (MUC-6, 1995) added a coreference resolution task. The experiment was repeated in the seventh and final edition (MUC-7, 1997). Later, a coreference resolution task was added to ACE from 2002 to the most current competitions. After a few years

without competition in this area, nowadays there is a new wave of interest thanks to the SemEval-2010 Recasens *et al.* (2010), CoNLL 2011 and 2012 tasks Pradhan *et al.* (2012). These last two tasks incorporate all known measures (except ACE- value) and have much larger corpora. In addition, the corpora and participants' output can be downloaded for future comparison.

On the one hand, the main goal of SemEval-2010 task on coreference Resolution in Multiple Languages was to evaluate and compare automatic coreference resolution systems for six different languages (Catalan, Dutch, English, German, Italian, and Spanish). On the other hand, the coreference resolution task of CoNLL-2011 uses the English language portion of the OntoNotes data, which consists of a little over one million words. The main goal was to automatically identify coreferring entities and events given predicted information on the other layers.

Nowadays every English nominal coreference system is evaluated on the CoNLL 2011/2012 paritions of the Ontonotes corpus[25]. The OntoNotes project has created a corpus of large-scale, accurate, and integrated annotations of multiple levels of the shallow semantic structure in text. The idea is that this rich, integrated annotation covering many linguistic layers will allow for richer, cross-layer models enabling significantly better automatic semantic analysis. In addition to co-references, this data is also tagged with syntactic trees, high-coverage verbs, and some noun propositions, verb and noun word senses, and 18 named entity types Weischedel *et al.* (2010). Moreover, OntoNotes 2.0 was used in SemEval Task 1 Recasens *et al.* (2010) and OntoNotes 4.0 (the fourth version of annotations) has been used in the CoNLL 2011 shared task on coreference resolution of which the Stanford's Multi Sieve Pass system was the winner. The English corpora annotated with all the layers contains about 1.3M words. It comprises 450,000 words from newswires, 150,000 from magazine articles, 200,000 from broadcast news, 200,000 from broadcast conversations, and 200,000 web data.

Automatic evaluation measures are crucial for coreference system development and comparison. Unfortunately, there is no agreement at present on a standard measure for coreference resolution evaluation. First, there are two metrics associated with international coreference resolution contests: the MUC scorer (Vilain et al. 1995) and the ACE value (Nist). Second, two commonly used measures, B3 Bagga and Baldwin (1998) and CEAF (Luo 2005), are also used. Finally, an alternative metric called BLANC was presented Recasens *et al.* (2010). B3 and CEAF are mention-based, whereas MUC and BLANC are link-based.

These metrics were all used in the CoNLL 2011 and 2012 tasks, and we will be using the official scorer provided. As or the evaluation of the event-coreference in section 5.1.3, we used the updated CorScorer package[26] developed by Luo *et al.* (2014). The CorScorer expects that coreferences are represented in CoNLL2011/2012 format. We also use the package developed within the Newsreader project[27] to convert CAT and NAF annotations

[25] `http://conll.cemantix.org/2012/introduction.html`
[26] `https://code.google.com/p/reference-coreference-scorers/`
[27] `https://github.com/cltl/coreference-evaluation`

to this format. An example of the output format shown in 3.

Table 20 shows the performance of the English nominal coreference system in Newsreader (Corefgraph) for a direct comparison with the results shown in Table 19.

Table 20: Multi-sieve Pass and CoNLL 2011 dev-auto Evaluation

System	MUC	B^3	CEAF	BLANC	CONLL 2011 F1
Stanford	59.6	68.3	45.5	73.0	59.3
Newsreader (Corefgraph)	51.0	67.2	43.4	69.7	54.8

The results show that Corefgraph still performs lower than the Stanford System in the dev auto dataset. In order to assess if the performance difference is due to the coreference resolution algorithm, we also evaluated Corefgraph in the *supplementary closed track gold boundaries* dataset. The results in table 21 suggest that the differences in performance when evaluated with the dev auto corpus may be related with the mention detection algorithm.

Table 21: Multi-sieve Pass and CoNLL 2011 closed track gold boundaries Evaluation

System	MUC	B^3	CEAF	CONLL 2011 F1
Stanford	80.05	69.70	66.80	72.18
Newsreader (Corefgraph)	79.56	68.09	65.44	71.03

Thus, we obtain mixed results: on the closed track gold boundaries evaluation we obtain results much closer to the Stanford Multi-Sieve Pass system (Lee et al. 2013), and this difference is larger when evaluated on the development auto dataset. Although it is difficult to draw firm conclusions on these results, the performance shown means that the system behaves competitively when compared with other publicly available Coreference system such as the Stanford system.

We have also evaluated the performance on the Wikinews gold standard annotated within the Newsreader project. The results are shown in Table 22.

Table 22: Wikinews Nominal Coreference Evaluation

System	MUC	B^3	CEAF	CONLL 2011 F1
Newsreader (Corefgraph)	19.70	18.34	18.96	19.00

Apart from the usual amount of errors that a coreference evaluation usually produces, the results on the Wikinews dataset are not surprising if we consider a number of issues:

- **Singletons**: In the Wikinews *some* singletons are annotated. The evaluation provided by table 22 removes singletons after annotation. This is related to the next point.

- **Guidelines**: The coreference annotation guidelines for Wikinews are quite different to the Ontonotes and CoNLL 2011/2012 guidelines. In particular, singletons are annotated, but not all of them. If we do not use the singletons for evaluation, we under-generate mentions whereas if leave the singletons after annotation then Corefgraph over-generates. This is due to the fact that in Corefgraph and Ontonotes, every named entity, personal pronoun and, crucially, ever Noun Phrase (except in a few cases) is considered to be a mention. For example, in file 8983 of the airbus segment of Wikinews, there are some singletons annotated such as "50 787s", but many others are left out: "The Air Indian order", "The Air Canada deal", "a deal", "a further US 7 billion", "two last orders for new aircraft", "11 billion of aircraft deals".

- **Cascading errors:** We have seen in section 5.1.1 the misalignment between the corpus used to train the NERC module (CoNLL 2003) and the Wikinews guidelines for named entities. As many of the mentions are named entities, the performance of Corefgraph is suffering due to cascading errors in the pipeline. For example, the NERC module will correctly (according to CoNLL 2003 guidelines) annotate "Heathrow" as named entity and thus as a mention but in the Wikinews annotation the named entity annotation is "Heathrow airport".

The evaluations can be reproduced following the procedure explained in the nominal coreference evaluation package[28].

5.1.5 Semantic Role Labelling

In NewsReader, Semantic Role Labelling for English is carried out using the MATE-tools (Björkelund *et al.*, 2009). This software is a pipeline that includes linguistic processors that performs lemmatization, part-of-speech tagging, dependency parsing, and semantic role labeling of a sentence. The dependency parser had the top score for English for dependency parsing in and SRL on the CoNLL shared task 2009 (Hajič *et al.*, 2009). The performance of the current version of the system on that task is given in Table 23.

Labeled precision	(19137 + 10036) / (22467 + 10818)	87.65%
Labeled recall	(19137 + 10036) / (24748 + 10818)	82.02%
Labeled F1		84.74%
Unlabeled precision	(20697 + 10818) / (22467 + 10818)	94.68%
Unlabeled recall	(20697 + 10818) / (24748 + 10818)	88.61%
Unlabeled F1		91.55%

Table 23: Performance of MATE on the English dataset of CoNLL-2009

For the NewsReader pipeline we have developed a wrapper that includes only the dependency parser and the SRL system of MATE-tools. That means that the rest of the

[28] `https://github.com/newsreader/evaluation/tree/master/nominal-coreference-evaluation`

analysis used as input for this modules, like lemmatization and part-of-speech tagging, are obtained by the tools included in the NewsReader pipeline. In order to evaluate this configuration we have checked the performance of the SRL module on the WikiNews gold standard of the Newsreader project. This dataset contains 120 files with 597 sentences. Applying CoNLL-2009 scorer, we obtain the results in Table 24.

Labeled precision	(975 + 1186) / (5250 + 2416)	28.19%
Labeled recall	(975 + 1186) / (5967 + 1338)	29.58%
Labeled F1		28.87%
Unlabeled precision	(1083 + 1186) / (5250 + 2416)	29.60%
Unlabeled recall	(1083 + 1186) / (5967 + 1338)	31.06%
Unlabeled F1		30.31%

Table 24: Performance of MATE on WikiNews

According to the figures in Table 24 MATE-tools seem to perform very poorly in the WikiNews dataset. However, several considerations must be taken into account. As we were annotating (in WP3) and developing NLP tools (in WP4) in parallel, there are a number of mismatches and misalignments that are now affecting the evaluation of the different NLP modules.

Regarding SRL, it seems that the main problems using the evaluation framework of CONLL09 are:

- The manual annotations do not cover all predicates and mentions.

  MATE detects many more predicates than those annotated (2,416 wrt 1,338). For instance, market.01 is annotated only once while mate annotates 14 mentions. Mate detects 23 mentions to index.01 and none is annotated, etc. Additionally, mate usually provides arguments for these predicates which obviously are also not annotated. This means that we can not correctly calculate the precision of mate. However, if we consider only the 1,338 predicates annotated, mate correctly identifies 1186. That is, 88% recall.

- Some arguments have been manually annotated as CLINKs or SLINKs.

  These are cases of the type:

  "He said he closed the door". While mate considers "he closed the door" as arg1 of "said", the annotation contains an SLINK (subordinate link) between "said" and "closed".

  "I started to run". While mate considers "to run" as argument C-arg1 of "started", the annotation contains a GLINK (Grammatical link) between "started" and "run".

  Thus, in these cases, the current evaluation of mate is wrongly penalized.

- The manual annotations of the arguments correspond to spans and not heads.

  CONLL09 evaluation expects heads instead of spans. So, one head per argument. This is why we only provided the heads. Thus, if we annotate in the gold-standard the full span, let's say five tokens, the scorer evaluates five arguments instead of only one. This is why the system returns a low recall.

Obviously, the combination of the three problems produce quite misleading and unreliable results. For that reason, we have also evaluated the system taking the whole span from NAF instead of using the head of the arguments. The results are show in table 25.

Labeled precision	(3288 + 1186) / (26438 + 2416)	15.51%
Labeled recall	(3288 + 1186) / (5314 + 1338)	67.26%
Labeled F1		25.20%
Unlabeled precision	(4154 + 1186) / (26438 + 2416)	18.51%
Unlabeled recall	(4154 + 1186) / (5314 + 1338)	80.28%
Unlabeled F1		30.08%

Table 25: Performance of MATE on WikiNews with the full span from NAF.

Now, the labeled and unlabeled recall reaches to 67% and 80%, respectively. Nevertheless, this evaluation is still inaccurate because it does not consider the full span as a single argument. This is why the precision results are so low. In order to make a proper evaluation using the CoNLL-2009 scorer, we should annotate the heads in the gold standard and deal with the rest of not annotated predicates (and arguments) and the CLINKs and SLINKs as arguments. Meanwhile, we propose an alternative evaluation taking into account just predicates and arguments postions existing in the gold-standard of WikiNews. Although the results of this evaluation, included in Table 25, still do not show perfectly the performance of MATE in the WikiNews dataset, we believe that they are fairer than the previous ones.

Labeled precision	(3288 + 1186) / (4154 + 1186)	83.78%
Labeled recall	(3288 + 1186) / (5314 + 1338)	67.26%
Labeled F1		74.62%
Unlabeled precision	(4154 + 1186) / (4154 + 1186)	100.00%
Unlabeled recall	(4154 + 1186) / (5314 + 1338)	80.28%
Unlabeled F1		89.06%

Table 26: Performance of MATE over the GS annotations of WikiNews with the full span from NAF.

5.1.6 Temporal processing

For the evaluation of temporal processing modules we developed a scorer based on the evaluation methodology used for the TempEval3 task (UzZaman *et al.* (2013)). It uses the script relation_to_timegraph.py of the TempEval3 evaluation toolkit[29].

The scorer has been used for the EVENTI-Evalita 2014 evaluation campaign [30]. The scorer takes in input files in the CAT labelled format. We thus developed a package that converts NAF layer in CAT labelled format. During this conversion the non text-consuming TIMEX3 representing the document creation time is deleted. The document creation time is extracted from the metadata of a document and not from the text. It is important to annotate it with a non text-consuming TIMEX3 when extracting the temporal relations between events. But in the Wikinews corpus the document creation time is explicitly expressed in texts and during the manual annotation temporal relations have been built using the text-consuming document creation time. This is a specificity of the corpus; this had been done for helping annotators.

For time expression recognition and normalization two evaluations are done: strict matching and relaxed matching. For example, if the gold annotation contains "Tuesday evening" and the system detects "Tuesday", then they will get credit in relaxed matching but not in exact matching. We compute the F1-score of attributes (type and value) by multiplying attribute's accuracy by the F1-score obtained for time expression recognition.

For temporal relation extraction we provide three evaluations: strict matching, relaxed matching and temporal awareness. In the first case two relations match if their sources and their targets strictly match, as well as their types (BEFORE, AFTER, INCLUDES, etc.). In the second case, a relaxed matching is considering between the sources and the targets. The TLINK relations are also evaluated using the evaluation methodology of (UzZaman and Allen (2011)). This evaluation has been used for the TempEval3 task (UzZaman *et al.* (2013)). The metric proposed by UzZaman and Allen (2011) captures the temporal awareness of an annotation in terms of precision, recall and F1 score.

The intra-document annotation Guidelines (Tonelli *et al.* (NWR2014-2-2)) allows the annotation of non text-consuming TIMEX3, for example in order to represent the begin point of a duration. Currently the TimePro module for English does not extract non text-consuming TIMEX3. We decided to do the evaluation without taking into account the non text-consuming TIMEX3. The evaluation will be computed again when TimePro will be extented with new functionalities to annotate non text-consuming TIMEX3.

In the Table 27 we present the results of TimePro on the 4 subcorpora and the micro-average on the whole corpus. The measure uses is the recall, precision and f1-score. We provide the evaluation on three aspects: the recognition of time expression extents, their classification (date, time, set or duration) and their normalization.

The temporal relation extraction module (TempRelPro) extracts relations between two event mentions or between an event mention and a time expression.

To understand better the results it is important to take into account the evaluation of

[29] `http://www.cs.york.ac.uk/semeval-2013/task1/index.php?id=data`
[30] `http://www.evalita.it/2014/tasks/eventi`

	recognition			*classification*	*normalization*
	recall	precision	F1-score	F1-score type	F1-score value
strict match					
Apple	0.805	0.968	0.879	0.831	0.715
Airbus, Boeing	0.761	0.909	0.828	0.817	0.793
GM, Chrysler, Ford	0.714	0.905	0.798	0.782	0.613
Stock market	0.636	0.881	0.738	0.692	0.58
Micro-average	0.72	0.914	**0.805**	0.773	**0.664**
relaxed match					
Apple	0.841	1	0.913	0.856	0.721
Airbus, Boeing	0.848	0.974	0.907	0.86	0.79
GM, Chrysler, Ford	0.789	1	0.882	0.849	0.647
Stock market	0.709	0.982	0.823	0.762	0.646
Micro-average	0.787	0.99	**0.877**	0.827	**0.692**

Table 27: TimePro performance

event detection (see Table 11 in Section 5.1.3) and of time expression (see Table 27). In fact through this evaluation we are evaluating entity pairs extraction and classification, as well as events and time expressions extraction.

In Table 28 we present the scores of the TempRelPro module. The system achieved a micro-average F1 score of 22.9 using the temporal evaluation methodology proposed by UzZaman and Allen (2011). The temporal relation extraction results are reasonable given the fact that the event detection performance is 69.4 F1 score (see Table 11) and the time expression recognition is 80.5 F1 score (see Table 27). In comparison the best system in TempEval 3, ClearTK-2, obtained a F1 score of 30.98 for the task ABC on temporal relation extraction from raw text, achieving a F1 score of 82.71 on time expression recognition and 77.34 on event detection. The results obtained on the subcorpus about "GM, Chrysler and Ford" are similar to those obtained by ClearTK-2 during TempEval 3.

We see two problems in the evaluation of such system that can explain the low results. First of all only a small subset of possible pairs are manually annotated, the most central and obvious relations. In the NewsReader Guidelines the annotation procedure is divided into 5 subtasks: TLINKs between event mentions and the document creation time, TLINKs between main event mentions, TLINKs between main event mentions and subordinated event mentions in the same sentence, TLINKs between event mentions and time expressions in the same sentence and TLINKS between time expressions. Following these subtasks a annotator should be able to annotate the most central relations, but not a complete timegraph between all events, while a system would extract all possible relations.

The system is based on machine learning method and is trained on TimeBank and AQUAINT corpora (corpora distributed for TempEval3). These two corpora have been annotated following TimeML guidelines, which gives the following instruction for the annotation of TLINK: "A TLINK has to be created each time a temporal relationship holding between events or an event and a time needs to be annotated". The resulting annotation

	recall	precision	F1-score
strict match			
Apple	0.174	0.361	0.235
Airbus, Boeing	0.145	0.319	0.199
GM, Chrysler, Ford	0.194	0.474	0.275
Stock market	0.081	0.141	0.103
Micro-average	0.154	0.325	0.209
relaxed match			
Apple	0.177	0.366	0.238
Airbus, Boeing	0.155	0.341	0.213
GM, Chrysler, Ford	0.196	0.479	0.278
Stock market	0.084	0.146	0.106
Micro-average	0.158	0.333	0.214
temporal awareness			
Apple	0.201	0.387	**0.265**
Airbus, Boeing	0.157	0.328	**0.212**
GM, Chrysler, Ford	0.221	0.484	**0.303**
Stock market	0.094	0.155	**0.117**
Micro-average	0.173	0.339	**0.229**

Table 28: TempRelPro performance

of TLINKs differ from the annotation done in NewsReader, that is to say that the training corpus differs from the evaluation corpus, which could partially explain the low results.

This concludes the intra-document benchmark results at the end of Y2 of the NewsReader project. However, we were also curious to see what lies past the horizon, which is why we proposed a shared task on TimeLines as a first step towards Storylines.

6 Timelines

As part of going beyond document-based evaluations, the NewsReader team set up a Timeline evaluation in the context of the SemEval-2015: Semantic Evaluation Exercises.[31] The task, "TimeLine: Cross-Document Event Ordering" was accepted as a pilot task in order to gauge the state-of-the-art in cross-document timeline creation. In this section, we detail the task description, explain the annotation steps, the resulting corpus and the outcomes of the SemEval timeline task.

6.1 SemEval-2015 Task 4. TimeLine: Cross-Document Event

The TimeLine task revolves around ordering events across documents in a timeline around a particular entity. For this task, the English WikiNews articles that were already annotated

[31] `http://alt.qcri.org/semeval2015/`

in the intra- and cross-document annotation task are utilised.

For each of the sub-corpora (Apple, Airbus-Boeing, GM-Chrysler-Ford, Stockmarket), up to 15 entities that are central to the corpus are defined. These are entities that occur in multiple documents and which play a role in different events. For each of the entities, the events in which the entity is a participant in the Arg0 or Arg1 propbank roles (often agent and patient, respectively) are selected and ordered chronologically. In case the event cannot be anchored to a particular date, the date is left blank.

The timelines are represented in a tab separated format in which the first column denotes the ordering, the second one the time anchor, and all following columns are co-referring events. In the events, the documentID and the sentence number are encoded for easy retrieval of the event in text. A timeline may thus look like the following:

```
iTunes
1    2003       11778-3-launch    11778-4-launch
2    2007       11778-4-pass
3    2008-01    11778-7-hold
4    2008-02    11778-2-pass    11778-5-pass
4    2008-02    11778-3-accounts_for
```

The timelines were generated semi-automatically from the manually cross-document annotated text and verified by NewsReader team members.

As this is a new task, the threshold to participate in the task is kept low by offering participants different levels of difficulty in creating the timelines, varying from a subtrack in which the participants are provided with the event mentions and only need to order the events (without temporal anchoring) and a subtrack in which only raw text is provided in which events need to be detected and ordered and anchored temporally. The four tracks offered have the following setup:

Track A (main track):

input data: raw text

output: full TimeLines (ordering of events and assignment of time anchors)

Subtrack A:

input data: raw text

output: TimeLines consist of just ordered events (no assignment of time anchors)

Track B:

input data: texts with manual annotation of event mentions

output: full TimeLines (ordering of events and assignment of time anchors)

Subtrack B:

input data: texts with manual annotation of event mentions

output: TimeLines consist of just ordered events (no assignment of time anchors)

The full task description and annotation guidelines can be found at `http://www.newsreader-project.eu/publications/technical-reports/` as the following techreports:

- Anne-Lyse Minard, Manuela Speranza, Bernardo Magnini, Marieke van Erp, Itziar Aldabe, Ruben Urizar, Eneko Agirre and German Rigau. *TimeLine: Cross-Document Event Ordering. SemEval 2015 – Task 4.* NWR-2014-10. Fondazione Bruno Kessler.
- Anne-Lyse Minard, Alessandro Marchetti, Manuela Speranza, Bernardo Magnini, Marieke van Erp, Itziar Aldabe, Ruben Urizar, Eneko Agirre and German Rigau. *TimeLine: Cross-Document Event Ordering. SemEval 2015 – Task 4. Annotation Guidelines.* NWR-2014-11. Fondazione Bruno Kessler.

6.2 NWR Timelines Dataset

We used as dataset the Wikinews corpus annotated as part of the project. The subcorpus about "Apple Inc." had been used as trial data and the other 3 subcorpora as evaluation data. Based on the cross-document annotation done in these 4 subcorpora, we have built timelines about seed entities. The timelines were automatically built using the cross-document annotation for events, as well as time anchor attributes of event instances and the has_participant relations. Afterwards the timelines have been manually corrected.

In Table 29 we describe the two datasets. The data used for the evaluation consists of 90 documents and 38 timelines.

Dataset	Trial corpus	Evaluation corpora			
	Apple	Airbus-Boeing	GM-Chrysler-Ford	Stock-market	Total eval dataset
# documents	30	30	30	30	90
# sentences	464	446	430	459	1,335
# timelines	6	13	12	13	38
length of timelines	29.3	21.1	20.5	16.9	20.0
# unique event mentions	188	331	305	264	900
average # docs by timeline	5.8	6.2	5.7	9.1	7.1

Table 29: Description of the Timelines Dataset

6.3 Outcomes

Whilst 29 teams had signed up for the TimeLines shared task, only 4 teams submitted results in the end. A paper detailing the results of this effort is in preparation, and the official results of the challenge are presented in Table 30.

	CORPUS1	CORPUS2	CORPUS3		TOTAL	
Participant	F1	F1	F1	F1 score	Precision	Recall
Track A						
SPINOZAVU_1	4.07	5.31	0.42	3.15	7.95	1.96
SPINOZAVU_2	2.67	0.62	0.00	1.05	8.16	0.56
WHUNLP_1	8.31	6.01	6.86	**7.28**	14.10	4.90
SubTrack A						
SPINOZAVU_1	1.20	1.70	2.08	**1.69**	6.70	0.97
SPINOZAVU_2	0.00	0.92	0.00	0.27	13.04	0.14
Track B						
GPLSIUA_1	22.35	19.28	33.59	**25.36**	21.73	30.46
GPLSIUA_2	20.47	16.17	29.90	22.66	20.08	26.00
HeidelToul_1	19.62	7.25	20.37	17.03	20.11	14.76
HeidelToul_2	16.50	10.94	25.89	18.34	13.58	28.23
SubTrack B						
GPLSIUA_1	18.35	20.48	32.08	**23.15**	18.90	29.85
GPLSIUA_2	15.93	14.44	27.48	19.18	16.19	23.52
HeidelToul_1	12.23	14.78	16.11	14.42	19.58	11.42
HeidelToul_2	13.24	15.88	21.99	16.67	12.18	26.41

Table 30: Official results of TimeLine task

7 Conclusions and Future Work

In this deliverable, we described the annotation efforts of the second year of the NewsReader project. There are three main parts to this deliverable, namely updates of the English intra-document guidelines and translations of the guidelines for the other three project languages, Spanish, Italian and Dutch. Then we described the data as well as the intra- and cross-document annotation task and results. In the third part of the deliverable, the benchmark evaluation results of the NewsReader pipeline on the English intra-document annotation are presented and compared to the state-of-the-art results on other gold standard datasets. In the fourth part, we describe the TimeLine SemEval shared task that the NewsReader team is organising as part of furthering the state-of-the-art in cross-document information extraction and towards cross-document storyline extraction.

This year's efforts further consolidated the English annotation guidelines, and we branched out to the other project languages. The annotation effort for English is completed save for a few documents in the cross-document annotation effort. As we took on

the TimeLine SemEval task, we needed to allocate time to annotate cross-document time lines, which was an additional task that was not foreseen originally. However, through the CROMER programme, this could be done in a semi-automatic fashion, which sped up the task greatly. The TimeLine task provided us with insights toward creating storylines, as well as a timelines dataset with 38 cross-document timelines that we intend to use in further research as well as share with the research community.

Furthermore, for Dutch, the intra-document annotation effort is completed, giving us a complete bi-lingual corpus. For Italian and Spanish, the annotation effort is still ongoing, but we intend this to be finished by the end of February 2015, delivering a four-language aligned corpus of 120 news articles with a rich linguistic annotation.

We have started working on evaluating the NewsReader NLP pipeline on our dataset. Although our annotation guidelines are based on state-of-the-art and common practice in our field, our annotations differ from those used in most commonly used gold standard datasets. We have for example defined a "PRODUCT" class for our named entities, since this is an important aspect of our domain and we felt it did not fit into a broader category such as "MISC" in which one would find instances of this class in for example the CoNLL benchmark dataset. As many of our modules are also dependent on annotated data from previous corpora, these differences led to lower scores of our modules on our annotated data, which led us to also report scores of our modules on the standard benchmark datasets. These benchmark datasets are largely independent of each other, e.g. one dataset only containing coreference resolution annotations, another only containing named entity recognition annotations. In the NewsReader dataset, all these types of annotations are layered, creating a rich, and more realistic annotation. In the coming year, we intend to analyse the results of our evaluation efforts and adapt our pipeline to deal with this data better.

In year 3, we will also perform a cross-lingual evaluation effort, to investigate in detail whether the NLP modules in the different languages extract the same information from the texts, and also to see if they can aid each other. We intend to organise another shared task around cross-lingual extraction, as well as around story line extraction for which we build upon the timeline dataset . Furthermore, we will release the annotated datasets, which we intend to accompany with a journal paper submission.

References

Rodrigo Agerri, Josu Bermudez, and German Rigau. IXA Pipeline: Efficient and ready to use multilingual NLP tools. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014), Reykjavik, Iceland*, 2014.

Emmon Bach. The algebra of events. *Linguistics and Philosophy*, 9:5–16, 1986.

A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 79–85, 1998.

Hans Bennis. *Gaps and Dummies*. Amsterdam University Press, 1986. 2005 reprint.

E. Bick. A named entity recognizer for danish. In *Proc. of 4th International Conf. on Language Resources and Evaluation*, page 305–308, 2004.

Anders Björkelund, Love Hafdell, and Pierre Nugues. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL '09, pages 43–48, Boulder, Colorado, USA, 2009.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

Tommaso Caselli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof. Annotating Events, Temporal Expressions and Relations in Italian: the It-TimeML Experience for the Ita-TimeBank. In *Linguistic Annotation Workshop*, pages 143–151, 2011.

Alexander Clark. Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 59–66. Association for Computational Linguistics, 2003.

Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8, 2002.

Agata Cybulska and Piek Vossen. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31 2014.

M. Fleischman and E. Hovy. Fine grained classification of named entities. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, page 1–7, 2002.

Christian Girardi, Manuela Speranza, Rachele Sprugnoli, and Sara Tonelli. Cromer: A tool for cross-document event and entity coreference. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31 2014.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL '09, pages 1–18, Boulder, Colorado, USA, 2009.

ISO TimeML Working Group. ISO TC37 draft international standard DIS 24617-1, August 14 2008. `http://semantic-annotation.uvt.nl/ISO-TimeML-08-13-2008-vankiyong.pdf`.

Claudia Leacock and Martin Chodorow. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283, 1998.

H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, 2011.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, pages 1–54, January 2013.

A. Xiaoqiang Luo, Sameer Pradhan, Marta Recasens, and Eduard Hovy. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, June 2014.

Bernardo Magnini, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi Lenzi, and Rachele Sprugnoli. I-CAB: the Italian Content Annotation Bank. In *Proceedings of LREC 2006 - 5th Conference on Language Resources and Evaluation*, 2006.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. Lexicon infused phrase embeddings for named entity resolution. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 78–86, Ann Arbor, Michigan, June 2014. Association for Computational Linguistics.

T. Poibeau and L. Kosseim. Proper name extraction from non-journalistic texts. *Language and Computers*, 37(1):144–157, 2001.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea, 2012.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. An extension of blanc to system mentions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, June 2014.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *New Directions in Question Answering*, pages 28–34, 2003.

K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501, 2010.

L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, page 147–155, 2009.

Marta Recasens, Toni Martí, Mariona Taulé, Lluís Màrquez, and Emili Sapena. SemEval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, page 70–75, Boulder, Colorado, June 2010. Association for Computational Linguistics.

R Saurí, O Batiukova, and J Pustejovsky. Annotttating events in spanish. timeml annotation guidelines (version tempeval-2010). Technical report, Barcelona Media. Technical Report BM 2009-01, 2009.

R Saurí, E Saquete, and J Pustejovsky. Annotating time expressions in spanish. timeml annotation guidelines (version tempeval-2010). Technical report, Barcelona Media. Technical Report BM 2010-02, 2010.

Roser Saurí. Annotating temporal relations in catalan and spanish. timeml annotation guidelines. (version tempeval-2010). Technical report, Barcelona Media. Technical Report BM 2010-04, 2010.

Manuela Speranza and Anne-Lyse Minard. NewsReader Guidelines for Cross-Document Annotation. Technical report, Fondazione Bruno Kessler, NWR2014-9.

E. F Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147, 2003.

Sara Tonelli, Rachele Sprugnoli, Manuela Speranza, and Anne-Lyse Minard. NewsReader Guidelines for Annotation at Document Level. Technical report, Fondazione Bruno Kessler, NWR2014-2-2.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

Naushad UzZaman and James Allen. Temporal evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 351–356. Association for Computational Linguistics, 2011.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, SemEval '13, pages 1–9, Atlanta, Georgia, USA, 2013.

Chantal van Son, Marieke van Erp, Antske Fokkens, and Piek Vossen. Hope and Fear: Interpreting Perspectives by Integrating Sentiment and Event Factuality. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31 2014.

R. Weischedel, S. Pradhan, L. Ramshaw, J. Kaufman, M. Franchini, M. El-Bachouti, N. Xue, M. Palmer, M. Marcus, and A. Taylor. OntoNotes release 4.0. Technical report, Tech. rept. BBN Technologies, 2010.