NewsReader Guidelines for Annotation at Document Level

NWR-2014-2-2

Version FINAL

Sara Tonelli, Rachele Sprugnoli, Manuela Speranza, Anne-Lyse Minard
satonelli, manspera, sprugnoli, minard@fbk.eu

(1) Fondazione Bruno Kessler
Via Sommarive 18, 38123, Povo (Trento), Italy

Building structured event indexes of large volumes of financial and economic data for decision making
ICT 316404

Contents

1 Overview of annotation guidelines

This document presents the annotation guidelines defined in the NewsReader project.

Annotation consists of two main tasks: the detection and annotation of markables (i.e. entities, events, temporal expressions, numerical expressions and various kind of signals) and the detection and annotation of relations (i.e. coreference, participant roles, causal, temporal, subordinating and grammatical links) between markables.

The guidelines and most of the examples for the annotation of entities and entity mentions are taken from ACE [Linguistic Data Consortium, 2008]. In some cases we adopted a simplified version of those guidelines. On the other hand, the annotation of events is inspired by ISO-TimeML [ISO TimeML Working Group, 2008].

Sections 2 and 3 present two different tags that are used to distinguish between entity instances (i.e. <ENTITY>) and entity mentions (i.e. <ENTITY_MENTION>) in order to handle both the annotation of single mentions and of the coreference chains that link several mentions to the same entity in a text. Links between entity mentions and entity instances are annotated through a link named REFERS_TO described in Section 10.1.

In the sentence *'Qatar Navigation jumped 6.4 percent after the company said it scraped plans for a 20 percent capital increase'* the entity of type organization *Qatar Navigation* is expressed through 3 different textual realizations: *Qatar Navigation* is a mention of syntactic_type NAM, i.e. a proper noun, *the company* is a mention of syntactic_type NOM, i.e. a common noun, and *it* is a PRO mention, i.e. a pronoun.

In Section 2 and 3 square brackets [ ] will indicate the extent of an entity mention and underlining will be used to indicate its syntactic head. Only the part of the extent that illustrates the example being discussed will be annotated.

Sections 4 and 5 describe the annotation of events. *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. Two different tags are adopted to distinguish between instances (i.e. <EVENT>), in Section 4, and instance mentions (i.e. <EVENT_MENTION>) of events, see Section 5 in order to model event coreference.

Temporal expressions, numerical expressions, temporal signals and causal signals are presented in Sections 6, 7, 8, and 9 respectively.

Section 10 is dedicated to the annotation of relations.

2 <ENTITY>

This tag is used to mark entities. An entity is an object or set of objects in the world or the mental representation of an object.

Each entity is described through an empty-element tag with the following attributes:

- id, automatically generated by the annotation tool;
- ent_type;
- ent_class;
- external_ref;
- tag_descriptor;
- comment, a free text field where the annotator can add notes.

BNF of the ENTITY tag
attributes ::= id ent_type [ent_class] [external_ref] tag_descriptor [comment]
id ::= <integer>
ent_type ::= PER | LOC | ORG | ART | FIN | MIX
ent_class ::= SPC | GEN | USP | NEG
external_ref ::= CDATA
tag_descriptor ::= CDATA
comment ::= CDATA

2.1 Entity types

The ent_type attribute specifies the entity type from a semantic perspective. Its possible values correspond to the 5 semantic types explained below:

1. PERSON (PER). Each distinct person or set of people mentioned in a document refers to an entity of type Person. For example, a person entity may be specified by name (*[Barack Obama]*), occupation (*[the CEO]*), pronoun (*[he]*), etc., or by some combination of these.

   Dead people and **human remains** are to be recorded as entities of type Person. So are fictional human characters appearing in movies, TV, books, plays, etc.

   There are a number of words that are ambiguous as to their referent. For example, nouns which normally refer to animals or non-humans, can be used to describe people. If it is clear to the annotator that the

noun refers to a person in a given context, it should be marked as a Person entity, as in *[The political cat of the year]* and *She's known as [the brain of the family]*.

Names of fictional characters are to be tagged; however, character names used as TV show titles will not be tagged when they refer to the show rather than the character name. Thus, we annotate *Batman* in *[Batman] has become a popular icon* but not in *The costume from Batman the TV series*.

Names of animals are not to be tagged, as they do not refer to person entities. The same is true for fictional animals and non-human characters. **Body parts** are taggable ONLY in cases in which the body part can reasonably define the entire person, as in the case of pieces of corpse recovered after attacks, accidents, etc.

2. LOCATION (LOC)[1]. Each distinct geographical region in a document refers to an entity of type Location. We have two types of location entities:

   (a) those which can be defined on a geographical or astronomical basis

   (b) those which constitute a political entity and are thus comprised of a physical location, a population and a government (they are composite locations).

   Location entities of the first type are: geologically designated non artificial locations (e.g *[the Caucasus]*), bodies of water both natural or artificial, celestial bodies, addresses, non-named locations that cross national borders (e.g. *northern Europe*), non-named locations that do not cross national borders (e.g *[southern Germany]*), and borders (e.g. *[borders shared by Turkey, Azerbaijan, and Georgia]*).

   Location entities of the second type are: nations, states, provinces, counties, districts, population centers and entity clusters such as *[the European Union]*, *[the Middle East]*, and *[Eastern Europe]*.

   Portions of location entities constitutes locations entities in their own right (e.g. *[the coast of Britain]*, *[the outskirts of the city]* and *[the center of the city]*[2]).

[1] Our definition of Location entities differs from the original ACE specifications because it includes also those which are annotated as Geo-Political Entities (GPE) in ACE.

[2] Note that each of these examples contains two nested entities.

When general locative phrases like "periphery", "bottom" and "center" are used to pinpoint a portion of a markable location, they are markable locations (e.g. *they live at [the periphery]*).

Annotators should be careful not to interpret all objects as locations. Every physical object implies a location because the space that each physical object occupies is the "location" of that object. However, the expressions in upper case in the sentences *The rabbit is hiding behind that ROCK*, *He dropped the logs on the GROUND* and *He put the lamp back in its PLACE* are not taggable location entities as they do not fall within the classes mentioned above for taggable locations.

Compass points should not be tagged when they serve as adjectives or refer to directions, as in *the ants are heading north* and *they are found as far north as Maine*. Compass points should only be tagged when they refer to sections of a region, as in *[the Far East]*.

In the case of composite locations all mentions of the different aspects (population, physical location and government) are marked as location and coreferenced. In the sentence *The people of France welcomed the agreement* there are two mentions which refer to two different aspects of the same location entity, i.e. *[The people of France]* and *[France]* (notice that we have no entity of type Person in this sentence).

Explicit references to the **government** of a political entity (country, state, city, etc.) are to be treated as references to the same entity evoked by the name of the political entity. Thus *the United States* and *the United States government* are mentions of the same entity. On the other hand, references to a portion of the government (*the Administration*, *the Clinton Administration*) are to be treated as a separate entity (of type Organization), even if they may be used in some cases interchangeably with references to the entire government (compare *the Clinton Administration signed a treaty* and *the United States signed a treaty*).

3. ORGANIZATION (ORG). Corporations, agencies, and other groups of people defined by an established organizational structure. More specifically, we annotate:

   - government organizations, which are of, relating to, or dealing with the structure or affairs of government, politics, or the state[3] (e.g. *[KGB]*, *[Congress]*, *[the US navy]*),

[3]Note that the entire government of an entity, on the other hand, should be tagged as a LOC as explained above

- commercial organizations, which are focused primarily upon providing ideas, products, or services for profit (e.g. *[TechSource Marine Industries in State College, PA.]*),
- educational organizations, which are focused primarily upon the furthering or promulgation of learning/education (e.g. *[NDSU]* and *[University of Minnesota]*),
- entertainment organizations, whose primary activity is entertainment (e.g. *[the Roundabout Theater Company]*),
- non-governmental organizations, which are not a part of a government or commercial organization and whose main role is advocacy, charity or politics in a broad sense, such as (para-)military organizations, political parties, political advocacy groups and think tanks, professional regulatory and advocacy groups, charitable organizations, international regulatory and political bodies (e.g. *[the Red Cross]*, *[the American Bar Association]*),
- media organizations, whose primary interest is the distribution of news or publications (e.g. *[Time magazine]*, *[Associate Press]*),
- religious organizations, which are primarily devoted to issues of religious worship (e.g. *[German Bishops Conference]*),
- medical-science organizations, whose primary activity is the application of medical care or the pursuit of scientific research (e.g. *[Massachusetts General Hospital]*),
- sports organizations, which are primarily concerned with participating in or governing organized sporting events, whether professional, amateur, or scholastic (e.g. *[Saudi Soccer Federation]*).

4. PRODUCT (PRO). Product is anything that can be offered to a market that might satisfy a want or need [4]. 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*).

[4] Definition taken from WikiPedia: `http://en.wikipedia.org/wiki/Product_(business)`.

Examples: *[147,114 vehicles]*, *browser*.

5. FINANCIAL (FIN)[5]. We annotate as Financial the entities belonging to the financial domain which are not included in one of the entity types described above, i.e. Person, Location, Organization, and Product. Examples of Financial entities are *[EGX-30 index]* and *[GDP]*. Notice that a financial market is a financial entity because it is not an organization (e.g *The U.S. market*), whereas stock exchanges defined by an established organizational structure (e.g. *the New York Stock Exchange*) are annotated as Organization entities.

6. MIXED (MIX)[6]. Conjunctions of entities (see Section 3.3) belonging to different types (for example one PER entity and one ORG entity) are annotated as Mixed (e.g. *the CEO and his company*).

2.1.1 Organization Entities vs Person Entities

Whenever an organization takes an action, there are people within or in charge of the organization that one presumes actually made the decision and then carried it out. Thus many organization mentions could be thought of as metonymically referring to people within the organization. However, there seems to be little to be gained in the usual case by thus "reaching inside the organization" to posit a mention of a Person entity. It seems better to adopt the view that **organizations can be agentive, and take action on their own**. For example in *Microsoft said the new tablets would be initially available in Australia*, *Microsoft* is annotated as ORG.

First person plural pronouns are often used by representatives of an organization to refer to that organization. Pronouns are often used in this way by reporters representing a broadcasting station and spokespeople representing organizations. For example, in *our top story*, "our" refers to the broadcasting organization. In these cases, annotators should mark first person plural pronouns as ORG, and not as PER.

Sets of people who are not formally organized into a unit are to be treated as a Person entity rather than an Organization entity. It is often difficult to tell the difference between Organization entities and collections of individuals. Examples of organization-like nouns which are not organizations are "employees", and "crew". Although the members of a company or crew

[5]The original ACE guidelines do not have Financial Entities; we have introduced them because of their relevance in the domain of the NewsReader project.

[6]This entity type does not exist in the ACE guidelines. It has been introduced as a consequence of the introduction of Conjunction entity mentions.

may work together in an organized and even hierarchical fashion, the groups are not organizations by themselves.

We make exception to the above rule for certain **military entities**. The words "troops", "forces", and "police" are commonly used in the same form multiple times in a document, and it is impossible to tell when they are referring to the same group of people, so co-referencing becomes very difficult and inconsistent. Because the same mention string can be used multiple times in different contexts without distinguishing the PER entities involved, these mentions are better tagged as referring to the ORG of which they are members. We will tag any mention of "troops", "police", "forces", etc. as ORG, where it is not explicitly referring to a partial subset of the organization.

[U.S. troops]$_{ORG}$ entered Baghdad yesterday
[A few troops]$_{PER}$ were left to man the base.
[the coalition forces]$_{ORG}$ took the bridge to Basra
[Chicago police]$_{ORG}$ arrested the suspect.

2.1.2 Organization Entities vs Location Entities

Coalitions of governments, as well as the UN, are organizational bodies and should be annotated as Organization entities, as in *[NATO]$_{ORG}$ peacekeepers arrived in the valley before nightfall.*

When the name of a geopolitical entity metonymycally refers to a **sports team**, we annotate it as Organization[7]. Examples:

[America]$_{ORG}$ brought home the gold
[Pittsburgh]$_{ORG}$ won 30-27

2.1.3 Organization Entities vs Products

Entities of type Organization often have a physical entity of type Product associated with them. These two incarnations of the same entity will be tagged as type Organization when the textual reference is directly referring to the organization and as type Product when the mention refers to the physical building. For example, in the following sentence there are two mentions of a hospital: the first mention is referencing the physical building or hospital facility, the second references the organization that runs or administrates the hospital:

[7]This is a typical example of metonymy. There will be other examples of metonymies in this document, but the list is not exhaustive. In all cases when a speaker metonymically uses a reference to one entity to refer to another entity related to it, the entity should be annotated with the type of the entity to which it refers in that context.

Wouters, 42, died an hour later at [St. John Macomb Hospital]$_{ART}$. The suspect died later the same night, [hospital]$_{ORG}$ spokeswoman Rebecca O'Grady said Thursday.

2.1.4 Location vs Person Entities

Population of a political entity (e.g. state, nation, city) is annotated as LOC if it is a reference to the population as a whole. If the phrase to be annotated refers to the population of the entity, or most of the population of a entity, then the annotation should be LOC and the mention is a name mention (see Section 3) because it is the proper name of the LOC. If the phrase refers to a group of people, then PER is the assigned annotation and the mention is nominal (see Section 3) because it does not refer to the proper name of a person.

Examples:

- *[Cubans]$_{LOC}$ have been waiting for this day for a long time.*
- *[A majority of [Americans]$_{LOC}$]$_{PER}$ believe the allegations against Mr. Clinton are true.*
- *[A majority of [Americans surveyed]$_{PER}$]$_{PER}$ believes allegations Mr. Clinton had an affair while he was President are not relevant.*
- *I do think there is a danger that [some Chinese]$_{PER}$ may underestimate American will on the Taiwan issue.*
- *[the [American]$_{LOC}$ people]$_{LOC}$ have a right to get answers.*
- *Yet another cutting edge development by [the French]$_{LOC}$ in their ongoing dealings with their enormous pet population.*

2.1.5 Location Entities vs Products: Government Seat (ART) names standing in for Country's Government

Cases in which the building that is the seat of government (e.g. *the White House*, *the Kremlin*) is metonymically used to refer to the nation's government are annotated as Location Entities, not as Products.

[White House]$_{LOC}$ press secretary Scott McClellan

In the example above, *White House* refers to the government of the USA thus we will tag it as LOC and co-refer it with all other mentions of the USA that might occur in the document. On the contrary, in the following sentence *White House* refers to the seat of USA government thus it is annotated as ART.

[The White House]$_{ART}$ is the official residence of the US President

2.1.6 Entities vs Events

Event names are to be annotated as events and not as entities, even if they refer to events that occur on a regular basis and are associated with institutional structures (e.g. Olympic Games). However, the institutional structures themselves — steering committees, etc. — should be tagged as entities. For example, *the Pan-American Games* is not an entity but an event whereas *the Olympic Committee* is an Organization entity.

Please note that some nominals show a polysemy that involves an eventive meaning versus an entity meaning, for example:

- EVENT / PRODUCT:
  commit oneself to the [collection]$_{EVENT\_MENTION}$ of art works
  inherit [a collection]$_{ART}$ of art works

- EVENT / PERSON:
  He is not involved in the [administration]$_{EVENT\_MENTION}$ of the factory
  Nobody in [the administration]$_{PER}$ would take responsibility

- EVENT / ORGANIZATION:
  The trade group held its [assembly]$_{EVENT\_MENTION}$ in Santiago, Chile
  What did [the French National Assembly]$_{ORG}$ decide about employment?

In such ambiguous cases, annotators can ask themselves the question: "when did this (i.e. *the collection / administration / assembly*) start / happen / end?". If such a question does not seem felicitous, or only makes sense by interpreting it as "when was this created?", then the element in question is probably an ENTITY rather than an EVENT.

Cases in which a nominal event (such as *Conference*) is metonymically used to refer to an entity (as in *Steve Jobs gave his annual opening keynote to the World Wide Developers Conference in San Francisco, California* are annotated as entities (in the example avove, the metonymical use of *Conference* as a group of people is signaled by the use of the preposition *to* instead of *at*).

2.2 Entity classes

The ent_class attribute expresses the definiteness of the entity instance. Its possible values are:

1. SPC (Specific Referential). An entity is SPC when the entity being referred to is a particular, unique object (or set of objects), whether or not the author or reader is aware of the name of the entity or its anchor in the (local) real world.

   Examples:

   [John's lawyer] won the case

   This afternoon, [a crowd of angry people] set fire to [a hotel]

   [At least four people] were injured

2. GEN (Generic Referential). GEN entities do not refer to a particular, unique object (or set of objects), but to a kind or type of entity.

   Notice that the mentions in question are still understood to be referential.

   Examples:

   [Lawyers] don't work for free

   About 231 feet to 264 feet of water is considered shallow for [submarines]

   [extremist groups] have a lot of support these days and a lot of power

   Japan's equivalent of [a naval force] is officially referred to as the Japan Maritime Self-Defense Force

   If the reference is to all members of a set rather than the set itself, the entity is to be tagged as SPC.

   [All the volunteers]$_{SPC}$ work for free
   [volunteers]$_{GEN}$ work for free

3. USP (Under-specified Referential). An entity is USP if it is impossible to determine its referent. Underspecified references include quantified NP's in modal, future, conditional, hypothetical, negated, uncertain, question contexts (in all cases the entity/entities referenced cannot be verified, regardless of the amount of "effort").

   [Many people] will participate in the parade

   I don't know [how many people] came

   Do you know [how many people] came?

   We will elect [five new officials]

   [You] know, I didn't even realize...

In cases where the author makes mention of an entity whose identity would be difficult to locate, and then conflates it with multiple other fuzzy mentions, all mentions are tagged as USP.

[Sources] said...

[Officials] reported...

When used as in the above examples, multiple mentions of *[sources]* and *[officials]* in a document should be coreferenced. Mentions of *[a source]* or *[an official]* (used in singular form) would be SPC instead.

If a GEN or SPC reading is possible, the USP tag should not be used.

4. NEG (Negatively Quantified). An entity is NEG when it has been quantified such that it refers to the empty set of the type of object mentioned.

   Examples:

   [No sensible lawyer] would take that case

   [No one] has claimed responsibility

   There are [no confirmed suspects] yet

   Please note that we do not assign NEG for entities introduced by negated predicates as in *They are not [lawyers]*.

Figure 1 presents the decision tree for the assignment of entity classes.

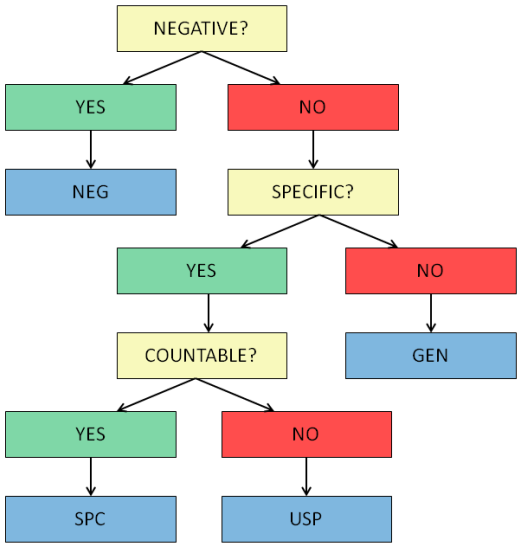

Figure 1: Decision Tree for Entity Class

2.3 External reference

The external reference attribute contains the DBpedia URI identifying the specific entity.

To get the DBpedia URI:

- go on the English Wikipedia `http://en.wikipedia.org/wiki/Main_Page`
- get your search term, have your Wikipedia page displayed
- replace *http://en.wikipedia.org/wiki/* with *http://dbpedia.org/resource/*

In case the page is redirected (in this case it says "Redirected from ..." right below the title), click "Article" on the top of the page. This will re-display the same article content, but with the "canonical" URI in the browser's address bar.

For example, with *[Morsi]'s move* the annotator would search *Morsi* in Wikipedia and find the entry Mohamed Morsi `http://en.wikipedia.org/wiki/Morsi`. This is a redirected page, so it is necessary to click on "Article" to obtain the value of the external reference attribute `http://en.wikipedia.org/page/Mohamed_Morsi`. Thus the correct URI for DBpedia is `http://dbpedia.org/resource/Mohamed_Morsi`.

If no proper DBpedia link is available for an entity, annotators should type the following in the COMMENT attribute: *no er* and leave the external reference attribute empty.

Please note that an entity of class SPC must not be linked to a generic entry in DBpedia. For example, in *Capital is managed on behalf of 93 investors*, the SPC instance *93 investors* should not be linked to `http://dbpedia.org/resource/Investor`. Annotators should leave the external reference attribute empty and type *no er* in the COMMENT attribute. The instance *investors* in *Investors commits money to investment products with the expectation of financial return*, on the other hand, is of class GEN and can therefore have `http://dbpedia.org/resource/Investor` as a value for the external reference attribute.

2.4 Tag Descriptor

It is a human-friendly identifier of the entity (for instance its name), which can be useful to distinguish the Entity in the annotation interface. As for capitalization, annotators should follow the normal rules of the language. Preferred length is one or two words.

3 <ENTITY_MENTION>

Annotators should annotate and coreference all mentions of each entity within a document. An entity mention is the textual realization of an entity, that is the portion of text in which an entity is referenced within a text.

Attributes of the <ENTITY_MENTION> tag are the following:

- id, automatically generated by the annotation tool;
- head;
- syntactic_type;
- comment.

BNF of the ENTITY_MENTION tag
attributes ::= id [head] [syntactic_type] [comment]
id ::= <integer>
head ::= CDATA
syntactic_type ::= NAM | NOM | PRO | PTV | PRE | HLS | CONJ | APP | ARC
comment ::= CDATA

3.1 Entity Mention Extent

The extent of this portion of text is defined to be the entire nominal phrase used to refer to an entity, thus including modifiers (e.g. *a big family*), prepositional phrases (e.g. *the President of the USA*) and dependent clauses (e.g. *John who is working in the garden*).

In case of structures where there is some irresolvable ambiguity as to the attachment of modifiers, the extent annotated should be the maximal extent.

In the case of a **discontinuous constituent**, the extent goes to the end of the constituent, even if that means including tokens that are not part of the constituent. Thus, in *I met some people yesterday who love chess* the extent of the mention is the entire phrase including also the temporal adverb which is not part of the constituent, i.e. *[some people yesterday who love chess]*.

Mentions will frequently be **nested**; that is, they will contain mentions of other entities. For example, in *The president of Ford* we annotate *[The president of Ford]*, a mention of an entity of type Person, which in turn contains the name *[Ford]*, a mention of an entity of type Organization. A series of nested region names are annotated as different entities. Thus, in *Dallas, Texas* we have one entity *[Dallas, Texas]* and one entity *[Texas]* both of type Location.[8]

In general, tokens are broken at white space. However, **possessive endings** ("'s") and **contractions** ("'re" for "are") are treated as separate tokens even if they are not separated from the preceding token by a blank space.

As a general rule we do not include **punctuation** such as commas, periods, and quotation marks in the extent of a mention. However, if words included within the extent continue on after the punctuation mark it is included. For example, in *Joe Smith (KY) who was elected last year* the extent of the mention includes also the parenthesis:

[Joe Smith ([KY]$_{LOC\ NAM}$), [who]$_{PER\text{-}PRO}$ was elected last year]$_{PER\ NAM}$

In the following example, on the other hand, we annotate two different co-referring mentions and the parenthesis are not included[9]:

[European Union]$_{ORG\ NAM}$ ([EU]$_{ORG\ NAM}$)

[8] It is even possible for a noun phrase to contain an embedded mention of the same entity. For instance, the phrase *The people of France* contains two mentions referring to the same location entity, i.e. *[The people of France]* and *[France]* (see location entities in Section 2.1).

[9] In this, our guidelines differ slightly from the ACE guidelines.

3.2 Syntactic head

For each mention, the head of the phrase must be marked. For instance, the head of the mention *the new glass-clad skyscraper* is *skyscraper*.

If the syntactic head of the phrase is a **multiword item**, annotators should mark as head the last word of the multiword. If the head is a **proper name**, however, then the whole extent of the name is considered to be the head. In the following examples, the mention is enclosed in brackets and the head is underlined:

[Fred Smith] became [the new prime minister]
The job fell to [Abraham Abercrombie III]
[TechSource Marine Industries in State College, PA.]
[the Roundabout Theater Company]
[Time magazine]
[University of Minnesota]
[the US navy]
[the United States of America]

Nominal mentions have in general only one word in their syntactic head. The main exception to this are LOC mentions that consist of a LOC name plus a directional modifier such as in the following example: *United States* would be a LOC NAM (i.e. a proper name) but the modifier causes it to become a LOC NOM (i.e. a common noun) (see Section 3.4), anyway *United States* cannot be decomposed, so it must be the entire head of the mention:

[the southwestern United States]

3.3 Syntactic type

Entity mentions are classified according to syntactic categories (e.g. proper names, common nouns, pronouns, etc.) given that entities may be referenced in a text by their proper name, indicated by a common noun or noun phrase, or represented by a pronoun. The syntactic type of a mention is determined by the syntactic category of its syntactic head. Its possible values are:

- NAM (proper name). In the most obvious cases, a NAM is a proper name or nick-name of any entity. For example, *[John]*, *[Defense Secretary William Cohen]*, *[The Jeluzoon Refugee Camp near [Ramallah]]*, *[North Dakota State University in [Fargo]]*, *[The house of representatives]*, *[the US army 101st Airborne Division]*.

  More borderline cases exist, however, such as *[the US Supreme Court]* and *[the US Army]* (see Section 3.4 for more details on the ambiguities between NAMs and NOMs).

- NOM (nominal compound). A NOM is a bare noun or a quantified noun (with a determiner, a quantifier, or a possessive). For example, *[the lawyer]*, *[some American executives]*, *[a crowd of angry people]*, *[these people]*, *[other U.S. officials]*, *[thousands of troops]*, *[this year's Miss America]*, *[offices in [foreign countries]]*, *[Americans] eagerly await the results of the election.*

- PRO (pronoun). All pronouns are PRO including the wh-question words (e.g. who? what? which? where?), relative pronouns and "that". For example, *[he]*, *[no one]*, *[these] have more terminals than the other top-ranking airports*, *[his own]*, *[everyone]*, *[that]'s not mine*, *the Russian navy [which] waited several days before attempting to launch any rescue mission*, *[Who] is the president of Brazil?.*

  As a general rule, a wh-question word corefers with its answer if it is in a question or with the noun it refers to (if it is a relative pronoun). In the following example, the three mentions (a PRO, a NOM and a NAM mention respectively) corefer: *[Who] was [the first president]? [George Washington] was.*

- PTV (partitive). Partitive constructions consists of two elements, the part and the whole (the first element of a partitive construction quantifies over the second element). We will tag the first element as the head of the partitive construction. For example, *[some of the lawyers]*, *[one of the houses]*, *[half of the team]*, *[all of them]*. If the first element consists of more than one word, we will tag as head the right most word of the first element, as in *[sixty percent of the participants]*[10].

  There are some constructions with prepositional phrases that resemble partitives, but are not partitives: two distinct entities (e.g. *two members of the team*), redundant embedded mentions, i.e. two co-referenced mentions (e.g. *[a group of [people]]*, *[the city of [Basra]]*), non quantified nouns (e.g. *[thousands of refugees]* and *[a lot of people]* are NOM mentions)

- PRE.NOM (nominal pre-modifier) and PRE.NAM (proper name pre-modifier). PRE.NOM and PRE.NAM are NOM and NAM entity mentions (respectively) in a modifying position, including both pre-modifiers (the majority of modifiers in English) and post-modifiers (rare).

  [mountain regions] LOC NOM
  [mountain] regions LOC PRE.NOM

[10]The whole is annotated as a nested mention.

[Apple products] PRO NOM
[Apple] products ORG PRE.NAM

Annotation guidelines are the same for PRE.NOM and PRE.NAM mentions, so we will not distinguish between the two types in the remaining of the document, but we will simply refer to the whole class of premodifiers with the short form PRE.

The taggability of PRE mentions can be affected by the context in which they appear (as for all other mentions types). For example, *criminal* in pre-modifying position can either refer to a person who has committed a crime or it can mean "relating to crime". In the first case it is annotated as PER, while in the second case (e.g. in the phrase "criminal charges"), it is not annotated.[11]

If there is ambiguity, we would go with the PER meaning, such as:

[feminist] groups

In order to decide whether a premodifier has to be tagged annotators should further determine whether the modifier would be a noun or an adjective when taken out from the context:

- The modifier would be a NOUN: it should be annotated as an entity mention (for example, we annotate *mountain* in *[mountain] regions* because *mountain* is a noun).
- The modifier would be an ADJECTIVE: annotators should take into consideration the noun from which the adjective derives. If it derives from a proper noun, it should be annotated (for example, *European* should be annotated as PRE.NAM in *[European] traditions*, as it derives from *Europe*), if it derives from a common noun, on the other hand, it should not be annotated (e.g. *mountainous* in *mountainous regions* because it derives from the common noun *mountain*).[12]

Titles, Honorifics, and Positions. In English, titles and most honorifics precede the name. We will not consider these to be part of the

[11] A particular tricky example of pre-modifiers is provided by "Islamic", which refers to the religion of Islam (not to be annotated), and "Muslim", which refers to the people of Islam (to be annotated). "Islamic", however, is sometimes used interchangeably with "Muslim" when modifying PER entities, and then we would tag it PER. However, if "Islamic" is modifying other entity type, it is not taggable.

[12] Note that we do not tag the "X" in "X" department (e.g. [the state department]). We tag "state" in "Secretary of State", because it is referring to the department, and "state" in "state property", because it is referring to the LOC entity, but we consider the "state" too abstract to tag.

name of a Person. We will annotate these as mentions in their own right. For example, in the string *President Obama*, there would be two mentions of the same entity (a PRE and a NAM mention).

The parts of titles are taggable if they refer to entities. For example, in the string *US President Obama*, there would be three mentions of two distinct entities.

[US President Obama] PER NAM

[US President] PER PRE

[US] LOC PRE

Religious titles such as saint, prophet, imam, or archangel are to be treated as titles.

Pope John Paul II

[Pope John Paul II]$_{\text{PER NAM}}$

[Pope]$_{\text{PER PRE}}$

Sometimes job titles refer to an empty position, as in: *She is running for President*. In cases like this, we will not tag *President* because it refers to the job or position and not to the person holding it. Note, however, that this is different from: *Who will be the next President?* In this case, we will tag *[the next President]* as PER USP NOM

- HLS (headless): Headless mentions are constructions in which the nominal head is not overtly expressed. Although these mentions are technically headless, we will assign as head the right most premodifier that falls directly before the spot where the head would be.

  Examples:

  [the toughest]

  [more than 30]

  [many] on both sides

  [60%] said

  [sixty percent] said

  [35] were injured

  They will [each] pay some money

  Bare demonstratives followed by a relative clause (or modified in some way) should be tagged HLS. Annotators should mark the demonstrative as head.

[Those present at the meeting] noticed

We must help [those in need]

- CONJ (conjunction). A conjunction is a construction which consists of two or more full entity mentions connected by a coordinating conjunction (e.g. *and*). The component mentions within the constructions will be tagged with their heads as appropriate. However, the CONJ-mention itself has no head-assignment. The extent of a conjunction includes the extents of all conjoined entities (e.g. *[[Marc]$_{NAM}$ and [John]$_{NAM}$]$_{CONJ}$*).

  The following three examples contain nested mentions:

  20 angry men and women

  [men]$_{\text{PER NOM}}$[13]

  [women]$_{\text{PER NOM}}$

  [20 angry men and women]$_{\text{PER CONJ}}$

  Bill Clinton and Jimmy Carter who are both former presidents

  [Bill Clinton]$_{\text{PER NAM}}$

  [Jimmy Carter]$_{\text{PER NAM}}$

  [who]$_{\text{PER PRO}}$

  [former presidents]$_{\text{PER NOM}}$

  [Bill Clinton and Jimmy Carter who are both former presidents]$_{\text{PER CONJ}}$

  The movers and shakers in Washington

  [movers]$_{\text{PER NOM}}$

  [shakers]$_{\text{PER NOM}}$

  [Washington]$_{\text{LOC NAM}}$

  [The movers and shakers in Washington]$_{\text{PER CONJ}}$

  Argentina, Chile, and Brazil, members of the Andean Group

  [Argentina]$_{\text{LOC NAM}}$

  [Chile]$_{\text{LOC NAM}}$

[13] Please note that modifiers in common between the entity mentions of a CONJ are left outside the extent of each single entity mention.

[Brazil]$_{\text{LOC NAM}}$

[members of the Andean Group]$_{\text{LOC NOM}}$

[the Andean Group]$_{\text{ORG NAM}}$

[Argentina, Chile, and Brazil]$_{\text{LOC CONJ}}$

[Argentina, Chile, and Brazil, members of the Andean Group]$_{\text{LOC APP}}$

- APP (appositional construction). In the case of appositions, the simple mention extent rules do not apply, so we have specific rules. An apposition is a construction which consists of two or more full entity mentions which refer to (or predicate on) the same entity. The component mentions within the APP-constructions will be tagged with their heads as appropriate. However, the APP-mention itself has no head-assignment.

  [[Bill], [John's lawyer]]

  [[Mr. Black, 58], [a victim of the terrorist assault]], told the Associated Press

  [[the show's production company], [Celador]]

  [[the heavy lift ship], [Blue Marlin]]

  [[Sauache County], [home of the Watchtower]]

  [[We] [Richmonders]]

  [[Pittsburgh], [a US airways hub]] recently paid to revamp the airport including shopping malls in the terminals

  The following are examples of annotation of mis-matched appositives:

  Bob, the director, and Bill, the producer, arrived...

  [Bob, the director] APP.PER

  [Bob] NAM.PER

  [the director] NOM.PER

  [Bill, the producer] APP.PER

  [Bill] NAM.PER

  [the producer] NOM.PER

  [Bob, the director, and Bill, the producer] CONJ.PER

  These are **not** appositives:

  [The American Civil Liberties Union]$_{NAM}$ ([ACLU]$_{NAM}$)

[He, [himself]$_{PRO}$]$_{PRO}$ had known it was true.

[[President]$_{PRE}$ George Bush]$_{NAM}$

In the case of appositional constructions, all the components forming these complex constructions and the APP mentions themselves are to be linked with the instance they refer to (see Section 10.1). When appositional constructions are involved relations of a different type, on the other hand, only the APP mentions themselves (and not the single components) are to be linked.

- ARC (apposition with relative clause). As in the case of appositions, the simple mention extent rules do not apply, so we have specific rules. An ARC-construction is an appositional construction with an adjacent relative clause that refers to the initial, referent mention of the entity, rather than the latter, attributive mention(s) of the entity. In ARC-constructions, the component entity mentions and the WHQ mention all are tagged with their heads as appropriate. However, the ARC-mention itself has no head-assignment.

  [[Dennis R. Beresford], [an accounting professor at the University of Georgia] [who] was then chairman of the accounting board]

  In the final example, it is unclear whether the relative clause refers to "John RIchards" or "the party leader". In cases of ambiguity like this, our policy is to tag the string as an ARC-construction. Also, please note that the embedded APP-constructions within ARC's are not tagged. Because an embedded apposition is requisite to ARC's, it is redundant to tag the APP-mention within them.

In the case of appositions with relative clause, all the components forming these complex constructions and the ARC mentions themselves are to be linked with the instance they refer to (see Section 10.1). When appositional constructions are involved relations of a different type, on the other hand, only the ARC mentions themselves (and not the single components) are to be linked.

3.4 NAM vs NOM

Some ambiguities can arise when trying to make a NAM-NOM distinction. It may appear that a NOM is being used to name something, or that a NAM mention may be decomposed into a few NOMs.

A general property of NAMs is that they are defined to pick out one particular entity as a referent. They are unique identifiers, like "*Vladimir Putin*" or "*The United States*".

NOMs, on the other hand, define an entire category. They can pick out a referent which belongs to that category, but only after disambiguating it from all other potential members of the category. If a nominal mention is used as an individual reference in a discourse, the head often has to be "individualized" via quantification and/or qualification with determiners, adjectives, relative clauses, etc.

[Vladimir Putin]$_{NAM}$ sat at the table.

[The man]$_{NOM}$ sat at the table.

References to "God" will be taken to be the name of this entity for tagging purposes. If it is used as a descriptor rather than a name, it will be considered a nominal mention.

if you believe in [God]$_{\text{NAM}}$

he felt like he was [a god]$_{\text{NOM}}$

When NAMs do carry a determiner, the determiner is a definite article and is not separable from the NAM. The definite article cannot be replaced with other determiners, quantifiers, or possessives.

[The New York Times] ran the article.

Ungrammatical:

**A New York Times ran the article.*

**That New York Times ran the article.*

One of the trickiest parts of distinguishing NAMs and NOMs is NOM categories modified by NAMs such that they only have one referent, such as:

the Pakistani army

the Egyptian supreme court

the University of Chicago payroll department

With the LOC/ORG modifying the categories, they pick out a specific referent in each NOM category. It is hard to decide whether the whole string should be treated as a NAM, or as a NOM mention with LOC/ORG PRE.

To annotate this kind of entity, we will follow the steps described below:

- **Step 1: ORG NAM or ORG NOM**, that is decide whether the modified ORG is NAM or NOM.

  Some ORGs are unambiguously NAM, as they automatically pick out one specific entity, not a member of a set. Some ORGs are unambiguously NOM, as they could not be considered the name of an organization, only a type of organization. When it's difficult to decide whether an ORG is NAM or NOM, apply the **ACRONYMS RULE**: ORGs with a corresponding acronym are NAM. Examples:

[American Airlines]$_{NAM}$ *([AA]*$_{NAM)}$[14]

If it's still difficult to decide whether an ORG is NAM or NOM, default to the **WORD COUNT RULE**: ORGs with more than two words are NAMs, and ORGs with one or two words are are most often NOMs.

Thus the following are NOMs:

- *army*
- *supreme court*
- *city council*
- *foreign ministry*
- *health ministry*
- *defense department*

But the following are NAMs:

- *center for disease control*
- *ministry of health*
- *automobile manufacturer's association*
- *communications development office*
- *developmental concepts and doctrine center*
- *census data center*

- **Step 2: NOM with PRE vs. NAM**

  Once it is known whether the modified ORG is NOM or NAM, the next step is to decide whether the entire mention is PRE + NOM or NAM, that is a common noun with a premodifier or a proper noun. For this, we will follow these rules:

  1. **ORG + LOC/ORG $\Longrightarrow$ ORG.NAM**
     [Oxfam Japan]$_{\text{NAM}}$
     [Abbot Laboratories Phillippines]$_{\text{NAM}}$
     [Abbott Laboratories Diagnostic Division]$_{\text{NAM}}$
     Note that the modifying ORG is included as part of the head.

[14] These examples may be considered unambiguously NAM, but there may be more difficult cases which this rule will clarify.

2. **ORG + of/at + LOC/ORG ⟹ ORG.NAM**
   [Agricultural Bank of China]$_{\text{NAM}}$
   [the Payroll Department at University of Pennsylvania]$_{\text{NAM}}$
3. **LOC + ORG**
   (a) **LOC + ORG.NAM ⟹ ORG.NAM**
   [Chinese Center for Disease Control]$_{\text{NAM}}$
   [U.S. Center for Disease Control]$_{\text{NAM}}$
   Note that the LOC is included as part of the head.
   (b) **LOC-adj + ORG.NOM ⟹ LOC.PRE + ORG.NOM**
   [the Pakistani army]$_{\text{NOM}}$
   [Pakistani]$_{\text{PRE}}$
   (c) **LOC-noun + ORG.NOM ⟹ ORG.NAM**
   [Pakistan Supreme Court]$_{\text{NAM}}$
4. **LOC/ORG + 's + ORG ⟹ LOC.PRE + ORG.NOM**
   [Thailand's health ministry]$_{\text{ORG NOM}}$
   [Thailand]$_{\text{LOC NAM}}$

   [Thailand's ministry of health]$_{\text{ORG NOM}}$[15]
   [Thailand]$_{\text{LOC NAM}}$

   [Cytogenetics Laboratory's Diagnostic Division]$_{\text{ORG NOM}}$
   [Cytogenetics Laboratory]$_{\text{PRE}}$

Trumping Rules:

- **possessive/adjective + ORG.NAM ⟹ ORG.NAM**

  Any ORG.NAM preceded by some adjective/possessive is annotated as ORG.NAM

  [The present-day CDC]$_{\text{NAM}}$

- **ORG.USP ⟹ ORG.NOM**

  Any hypothetical, future, etc. organization is automatically tagged NOM regardless of any modifiers or other constructions.

  [the proposed Chinese Development Bank]$_{\text{NOM}}$

  [Chinese]$_{\text{PRE}}$

[15] Please note that the geographical modifier changes the syntactic type: *ministry of health* would be NAM without the modifier.

- **Plural ORGs**

  Any ORGs that are plural are automatically considered NOMs:

  [Chinese and Japanese Centers for Disease Control and Prevention]$_{\text{NOM}}$

  [Chinese]$_{\text{PRE}}$

  [Japanese]$_{\text{PRE}}$

4 <EVENT>

This tag is used to mark instances of events, that is the mental representations of events to which various types of linguistic elements (e.g. nouns, verbs, pronouns) refer within a text.

Each event is described through an empty-element tag with the following attributes:

- id, automatically generated by the annotation tool;
- class, it specifies the event type, whose values correspond to the 5 classes explained below;
- external_ref, it contains the URI used by DBpedia to identify a specific entity instance. This type of attribute would allow the representation of DBpedia entries and others;
- tag_descriptor, a human-friendly description of the event instance;
- comment.

BNF of the EVENT tag
attributes ::= id class [external_ref] tag_descriptor [comment]
id ::= <integer>
tag_descriptor ::= CDATA
class ::= SPEECH_COGNITIVE | GRAMMATICAL | OTHER | MIX
external_ref ::= CDATA
comment ::= CDATA

4.1 Tag Descriptor for Events

It is the nominal identifier of the event instance, which can be useful to distinguish the Event in the annotation interface. As for capitalization, annotators should follow the normal rules of the language. The preferred length of the tag descriptor is one word, that is the most representative token expressing

the event. More words are acceptable in case of ambiguity. For example, in the same document we can have *USA property bubble* and *British property bubble*, two portions of text containing mentions of two different event instances with the same extent *bubble*; in this case it is useful to use more than one word for the tag descriptor, so as to properly distinguish the two event instances.

4.2 Event Classes

Event instances are classified on the basis of 3 classes:

1. SPEECH_COGNITIVE, for speech acts and cognitive events. In particular, this class contains:
   - events that describe the action of a person or an organization declaring something, narrating an event, informing about an event, e.g. *report, say, announcement, deny, explain, explanation*;
   - events that describe mental states and mental acts that involve mental or cognitive processes, e.g. *think, know, remember, perceive, prefer, want, forget, understand, decide, decision.*
2. GRAMMATICAL: events that are semantically dependent on a content verb/noun or on another event:
   - they do not introduce other participants;
   - they have no time span outside the content verb or noun;
   - they do not introduce any change of state that is not already expressed by the governing content verb or noun.

   List of grammatical events:
   (a) light verbs followed by a nominal event or copula verbs, e.g. *be, seem, do, make, get, do, have, take, put, set, let*;
   (b) aspectual verbs and nouns that code information on a particular phase or aspect in the description of an event. They are a grammatical device which code a kind of temporal information and focus on different facets of the event history. In particular, they may signal the initiation, reinitiation, continuation, termination of another event e.g. *stop, beginning, start, end*;
   (c) verbs and nouns expressing causal and motivational relations, e.g. *cause, result, stimulate, enable, stem from, lead to, breed, engender, hatch, induce, occasion, produce, bring about, produce, secure*;

(d) verbs and nouns expressing occurrence, such as *take place, happen, occurrence*.

3. OTHER: all the events in the document not covered by the classes SPEECH_COGNITIVE and GRAMMATICAL.

4. MIXED (MIX): Conjunctions of events belonging to different types (for example one event of type SPEECH-COGNITIVE and one event of type OTHER) are annotated as Mixed (e.g. *he said that he was tired and left*); please note that conjunctions of events are annotated as OTHER/SPEECH_COGNITIVE/GRAMMATICAL if all the conjoined events are of the same type.

4.3 External reference

The external reference attribute contains the DBpedia URI identifying the event.

To get the DBpedia URI:

- go on the English Wikipedia `http://en.wikipedia.org/wiki/Main_Page`
- get your search term, have your Wikipedia page displayed
- replace *http://en.wikipedia.org/wiki/* with *http://dbpedia.org/resource/*

In case the page is redirected (in this case it says "Redirected from ..." right below the title), click "Article" on the top of the page. This will redisplay the same article content, but with the "canonical" URI in the browser's address bar.

For example, if the annotator searches for *'29 crash* in Wikipedia, he/she finds the entry "'29 crash" `http://en.wikipedia.org/wiki/%2729_crash`. This is a redirected page, so it is necessary to click on "Article" to obtain the value of the external reference attribute `http://dbpedia.org/resource/Wall_Street_Crash_of_1929`.

If no proper DBpedia link is available for an event, annotators should type the following in the COMMENT attribute: *no er* and leave the external reference attribute empty.

Please note that a specific event must not be linked to a generic entry in DBpedia. For example the Wall Street Crash of 1929 should not be linked to the generic page `http://dbpedia.org/resource/Stock_market_crash` but to the above mentioned specific page `http://dbpedia.org/resource/Wall_Street_Crash_of_1929`. If a specific page is not available, annotators

should leave the external reference attribute empty and type *no er* in the COMMENT attribute. On the other hand, a generic stock market crash event should have `http://dbpedia.org/resource/Stock_market_crash` as a value for the external reference attribute.

5 <EVENT_MENTION>

This tag encodes different linguistic representations of a given event instance through a set of attributes largely inspired by those used in ISO-TimeML (see 5.2).

5.1 Extension

As for the identification of event mentions, the annotation of their extension is based on the notion of *minimal chunk* inherited by TimeML, because higher constituents may contain more than one event expression. With respect to TimeML, we preferred a more flexible application of the minimal chunk rule for event annotation, which led to the identification of **multi-token event mentions**. In particular, in order to be more informative on the semantic level, we identified a restricted set of exceptions to the minimal chunk rule: the extent of phrasal verbs, idioms and prepositional phrases corresponds to the whole expression if they are entries either in the American or in the British version of the Collins English Dictionary online: `http://www.collinsdictionary.com/dictionary/english`.
The group of tokens which constitute a multi-token event can be discontinuous, that is words can be inserted between the different components of the multi-token event. Also in this case, the components of the multi-token event must all be annotated as a single event, so in this case we have a discontinuous event mention.

Example: phrasal verb, dictionary entry "switch on"
[switch on] the machine
[switch] the light [on]

Example: idiom, dictionary entry "give someone the cold shoulder"
she [gave] him [the cold shoulder]

Example: prepositional phrase, dictionary entry "on board"
they had 150 passengers [on board]

Syntactically, the linguistic elements which may realize an event are the following:

- **Verbs** in finite or non-finite form. When a complex VP is present (i.e. the verb is accompanied by auxiliaries and related particles), the event extent is only the head of the VP. The same is true for phrasal verbs.
  *Israel has been **scrambling** to **buy** more masks abroad*
  *President Clinton **says** he and Blair will **stand** together*

  Please note that auxiliary verbs (*be, have, do*) and modals (*may, might, must, would, should, could, can, ought to, have to, or shall*) are not to be tagged: in constructions of this type, only the main verb, and not the auxiliary form, is to be tagged as event. The information about the presence of a modal is coded in a specific attribute called "modality".

- **Nouns** which can realize eventualities in different ways.
  - through a nominalization process from verbs (i.e. deverbal nouns): e.g. *suspension*;
  - having an eventive meaning in their lexical properties even if they don't derive from verbs: e.g. *crisi1s*;
  - having an eventive reading due to the co-text of their occurrence even if they normally denote objects, locations or time expressions: e.g. *11 September* when it refers to the terrorist attacks that occurred on September 11th, 2001.

  Please note that the event mention tag extends only over the head noun, disregarding any determiners, specifiers, complements, or modifiers.

  Event-denoting nouns and present participle forms acting as prenominal modifiers can be annotated as events if they are not part of a multi-word expression: this means that *election* in *its election defeat* can be annotated whereas *waiting* in *the waiting room* is not to be tagged as an event mention. When in doubt, annotators should refer to the reference on-line dictionary Collins English Dictionary, either to the American or to the British version: `http://www.collinsdictionary.com/dictionary/english` and check if an event mention candidate is part of a multi-word expression or not.

  If a nominal event is the syntactic head of an entity, it is not to be annotated as an event. For example *access* in *The 3G iPhone will provide internet access.* is the syntactic head of the entity *internet access* and will not be annotated as an event.

- **Other elements**:
  - **Adjectives**: Adjectives generally express a property or attribute of an entity, and as such, they denote an event of a stative nature. Event-denoting adjectives will have only their head adjective annotated as the event.

    We do NOT mark up as events adjectives in attributive position, that is adjectives that function as premodifiers of a noun (e.g. *furious reaction, unbearable pain, fair trial, beautiful garden*).

    We ONLY mark up as events adjectives in predicative position, that is adjectives that act as the predicative complement of a verb belonging to one of the types listed below, among others.

    In the examples, the predicative adjective is in bold face.

    * Copulative predicates (e.g., be, seem, etc.) as in *The students seemed **exhausted** after three weeks of classes*;
    * Inchoative predicates (e.g., become, turn into). They express the coming to existence of a situation, e.g. *The Chinese dissident said he left China because his life became **unbearable** there*;
    * Aspectual predicates (e.g., begin, continue, finish, terminate, etc.) as in *Families kept **hopeful** and many did see the return of their loved ones*;
    * Causative predicates (e.g., cause, make, etc.), as in *Dan Hollander, skater and entertainer, really made the audience **happy***;
    * Change of state predicates in general;
    * Predicates of perception (e.g., look, hear, etc.), as in *Ellen DeGeneres and Portia de Rossi looked **ecstatic** as they married in an intimate ceremony on Saturday*;
    * Predicates of evaluation and description (e.g., consider, describe, present, etc.), as in *He is often characterized as **eccentric***.
  - **Prepositional phrases** are to be annotated ONLY when functioning as predicative complements, that is when they are complement of verbs belonging to the grammatical class.

    The annotation of their extent is based on the minimal chunk rule, so only the head preposition is annotated.

    Anyway, prepositional phrases that are entry in the Collins English Dictionary online (the American and the British versions

should be checked): `http://www.collinsdictionary.com/dictionary/english` represent an exception to the minimal chunk rule, so the extent is the whole expression.

In the example below, the grammatical verb is underlined, the prepositional phrase (that is an entry in the dictionary) is in square brackets and the extent of the mention is in boldface:

*They had 150 passengers [**on board**]*

Notice that prepositional phrases are NOT to be annotated if their syntactic head denotes an event; in this case the element to be tagged as event is only the syntactic head (for example, the sentence *It began with a banking crisis* contains the nominal event *crisis* but no prepositional events).

– **Pronouns**, whose annotations is crucial to identify event coreference (*The economic crisis began in 2007: **it** started with a banking crisis*).

With respect to TimeML, we have introduced the annotation of **conjunctions of events**. The extent of a conjunction includes the extents of all conjoined events and is often discontinuous. Please notice that the subparts of a conjunction of events must also be annotated, so in the sentence *he said that he was tired and left* we annotate both the discontinuous event mention *[said left]* and the subparts *[said]* and *[left]*.

Particular attention should be paid to **complex event constructions** in which multiple event mentions (*emX* in the following examples) are present:

- ASPECTUAL CONSTRUCTIONS consisting of an aspectual verb or noun and an event-denoting complement expressed by a VP or an NP: both the predicate (em1) and the complement (em2) are to be annotated as independent event mentions (see boldface text below).
  *Novartis **began**$_{em1}$ the **trading**$_{em2}$ of its American Depositary Shares on the NYSE*
  *The **conclusion**$_{em1}$ of the economic **crisis**$_{em2}$ in the 1970s*

- INCHOATIVE CONSTRUCTIONS expressing the coming to existence of a situation. They generally involve the presence of verbs like *become* and *get*, in addition to their complement, which denotes the resulting situation or process. BOTH the inchoative predicate (em1) and the complement expressing the resulting situation (em2) are to be annotated as events.

The representatives of the biggest Swedish industrial and financial companies ***got***$_{em1}$ ***acquainted***$_{em2}$ *with economic situation in Belarus*

- LIGHT VERB CONSTRUCTIONS involve a verb of very light semantic content (e.g., *make, get, do, have, take, put, set, let*) and a nominal event acting as its selected complement. In these situations, BOTH the verbal (em1) and nominal (em2) elements are tagged as events.
  The manager is ***getting***$_{em1}$ *more* ***support***$_{em2}$ *from the owner*

- COPULATIVE CONSTRUCTIONS are VPs headed by verbs like *be* or *seem*, and which have an NP, AP, or PP as complement. The copulative predicate (em1) is always to be annotated while the predicative complement (em2) are to be marked up as event mention ONLY if it contains a noun with an eventive reading. The NP and the AP are to be annotated according to the rules specified in this guidelines. In *An eminent Indian origin woman* ***is*** *the new head of the British Medical Association* the copulative predicate (*is*) is to be annotated as a mention of a grammatical event, while the predicative complement (*the new head of the British Medical Association*) is an entity mention. On the other hand, in *this seems a long-term crisis*, the copulative predicate (*seems*) and the predicative complement (*crisis*) are both to be tagged as event mentions.

- CAUSATIVE CONSTRUCTIONS: the causal expression (em2), its logical subject (em1) and its event complement (em3) are ALL tagged as independent events.
  Widespread ***floods***$_{em1}$ ***caused***$_{em2}$ *economic* ***losses***$_{em3}$

Metonymy
Nouns can assume an eventive reading due to the co-text of occurrence. For this reason, a mention that in a context represents an implicit event derived by metonymy is to be annotated as an event mention.
Example:
The [bomb] ended the festival three days earlier.
In the example above, *bomb* is metonymic for the attack itself, in addition to representing the explosive device. The physical bomb did not end the festival, rather, its detonation thus *bomb* is annotated as event mention.

5.2 Attributes

The annotation of event mentions includes assigning values to several attributes.

- id, automatically generated by the annotation tool;
- pred, it corresponds to the lemma of the token describing the event;
- certainty, it encodes the distinction between certain, probable and possible events;
- polarity, it distinguishes affirmative (POS) and negative (NEG) statements;
- time, it encodes the statement of events about the future: non-future or future;
- special_cases, it captures if the statement have some special status that influences its attribution: general statement, main clause of a conditional construction or if clause of a conditional construction;
- pos, it specifies the different grammatical categories which may realize an event, i.e. NOUN, VERB, OTHER;
- tense, it captures standard distinctions in the grammatical category of verbal tense, i.e. PRESENT, PAST, FUTURE, NONE, INFINITIVE, PRESPART and PASTPART;
- aspect, it captures standard distinctions in the grammatical category of semantic aspect, i.e. NONE, PROGRESSIVE, PERFECTIVE, and PERFECTIVE_PROGRESSIVE;
- modality, optional attribute that is used to convey different degrees of modality of an event, its value is the lemma of the modal verb modifying the main event, e.g. *may*.
- comment.

BNF of the EVENT_MENTION tag

attributes ::= id [pred] [pos] [tense] [aspect] certainty polarity time special_cases [modality] [comment]
id ::= <integer>
pred ::= CDATA
certainty ::= CERTAIN | POSSIBLE | PROBABLE | UNDERSPECIFIED
polarity ::= NEG | POS | UNDERSPECIFIED
time ::= NON_FUTURE | FUTURE | UNDERSPECIFIED
special_cases ::= NONE | GEN | COND_MAIN_CLAUSE | COND_IF_CLAUSE
pos ::= NOUN | VERB | OTHER

tense ::= FUTURE | PAST | PRESENT | INFINITIVE | PRESPART | PASTPART | NONE
aspect ::= PROGRESSIVE | PERFECTIVE | PERFECTIVE_PROGRESSIVE |NONE
modality ::= CDATA
comment ::= CDATA

5.2.1 The *pred* attribute

It denotes the content related to that event through the indication of a lexical predicate at the lemma level.

5.2.2 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*

Mary came at 8 pm.
came = CERTAIN

Mary might have arrived at 8 pm.
arrived = POSSIBLE

Mary will probably arrive at 8 pm.
arrived = PROBABLE

The certainty of events is based on textual properties. When determining the certainty of a given event mention, annotators should base their assessment **uniquely** on the knowledge available in the sentence expressing the event without using world knowledge or other knowledge taken from the text. In "*I will certainly win the lottery tomorrow*", for example, annotators should not use their knowledge of the world which tells them that winning the lottery is very unlikely; in the text it is presented as a certain event so it should be annotated as CERTAIN. Similarly, if an event is known for certain but is presented as uncertain, it should be annotated as POSSIBLE or PROBABLE. For example *born* in "*I don't remember, maybe Obama was born in 1961*" is POSSIBLE, event if Obama was actually born in 1961.

POSSIBLE vs. PROBABLE We follow the guidelines from FactBank [16] 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. If you can deny the statement using a marker of probability in a context of opposite polarity, that means the statement is POSSIBLE, else it is PROBABLE.

In order to apply this test to our example "*Mary might have arrived at 8 pm.*" we can create a test sentence adding *but probably not*: "*Mary might have arrived at 8 pm* ***but probably not.***". As our test sentence is semantically valid, than the value of the certainty attribute should be POSSIBLE.

The same test can be applied to our example "*Mary will probably arrive at 8 pm.*". In this case, the resulting sentence " **Mary will probably arrive at 8 pm* ***but probably not.***" is not semantically valid, so the value of the certainty attribute should be PROBABLE.

It may not change for a couple of years.
TEST: *It may not change for a couple of years but it most probably will.*
change = POSSIBLE

I think it's not going to change for a couple of years.
TEST: **I think it's not going to change for a couple of years but it probably will.*
change = PROBABLE

Infinitive clauses expressing future events in the past The certainty value of events in final infinitive clauses is determined according to the semantic of the principal clause. The following rules should be applied:

- If the verb of the principal clause expresses an **intention** or an **aim**, the certainty of the infinitive is POSSIBLE.
  I want to help people.

- If the verb of the principal clause expresses a **decision** or an **announcement**, the infinitive clause is CERTAIN.
  The stockholders decided to collect the net income.

- If the verb of the principal clause expresses **orders** or **necessities**, the certainty of the infinitive clause is set to UNDERSPECIFIED.
  John must go to the hospital.

[16] `http://www.cs.brandeis.edu/~roser/pubs/fb_annotGuidelines.pdf`

5.2.3 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.

Mary came at 8 pm.
came = POSITIVE

They didn't read the book.
read = NEGATIVE

The president forgot to inform the cabinet.
forgot = POSITIVE
inform = NEGATIVE

John does not know whether Mary came.
know = NEGATIVE
came = UNDERSPECIFIED

5.2.4 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. The annotators should assign the time value according to the semantic temporal value of events.

In some cases the value of the time attribute of a verbal event can be deduced from the tense of the verb:

- past tense
  Share prices on the Egypt Exchange declined almost 9.5 percent by midday.
  declined = NON_FUTURE

- present tense
  A bank is a financial institution.
  is = NON_FUTURE

- future tense
  Future generation will suffer.
  suffer = FUTURE

Other cases where the value of the time attribut cannot be deduced from the syntactic tense of the event:

- Infinitive verbs:
  The president forgot to inform the cabinet.
  inform = NON_FUTURE

- Verbs preceded by a modal word:
  The StyleSelect USA Index may include a larger percentage of stocks.
  include = NON_FUTURE

  John said he would leave for Scotland on the 21st of May.
  leave = FUTURE

- Nouns:
  A lawsuit in Germany will seek a criminal prosecution of the outgoing Defence Secretary.
  prosecution = FUTURE

Note that, for 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.

The same rule is applied to infinitive expressing a future event with regard to the principal clause.
The stockholders decided to collect the net income.
decided = NON_FUTURE
collect = FUTURE

John wants Mary to go to the hospital.
wants = NON_FUTURE
go = FUTURE

5.2.5 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.

A bank is a financial institution.
is = GEN

If you burn fossil fuels, carbon dioxide is produced.
burn = COND_IF_CLAUSE
produced = COND_MAIN_CLAUSE

Events that are properties should be marked as general statement.
The iPod comes in two colors.
comes = GEN

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.
The company produces cans.
produces = NONE

5.2.6 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

Maybe Ford did not include auto leveling motors on the HID lights.

predicate	certainty	polarity	time	special_cases
include	POSSIBLE	NEG	NON_FUTURE	NONE

A lawsuit in Germany will seek a criminal prosecution of the outgoing Defence Secretary.

predicate	certainty	polarity	time	special_cases
seek	CERTAIN	POS	FUTURE	NONE
prosecution	POSSIBLE	POS	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

5.2.7 Attribution values of conditional constructions

In the following we provide examples of attribution values (i.e. certainty, polarity, time and special_cases attributes) of conditional constructions:

- Zero conditional: *If + simple present, simple present*
  It describes certain consequences rather than hypothetical or possible situations
  e.g. *If you burn fossil fuels, carbon dioxide is produced*

predicate	certainty	polarity	time	special_cases
burn	UNDERSPECIFIED	POS	NON_FUTURE	COND_IF_CLAUSE
produced	CERTAIN	POS	NON_FUTURE	COND_MAIN_CLAUSE

- First conditional: *If + subject + present tense verb, subject + future tense verb*
  It expresses the consequences of a possible future event.
  e.g. *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

- Second conditional: *If + subject + simple past tense verb, subject + conditional tense verb*
  It is used for hypothetical typically counterfactual situations; it can have a present (see example *a*) or a future (see example *b*) time frame.
  a. *If I liked parties, I would attend more of them.*

predicate	certainty	polarity	time	special_cases
liked	UNDERSPECIFIED	POS	NON_FUTURE	COND_IF_CLAUSE
attend	CERTAIN	POS	NON_FUTURE	COND_MAIN_CLAUSE

b. *If I <u>became</u> rich, I would <u>buy</u> this house.*

predicate	certainty	polarity	time	special_cases
became	UNDERSPECIFIED	POS	FUTURE	COND_IF_CLAUSE
buy	CERTAIN	POS	FUTURE	COND_MAIN_CLAUSE

- Third conditional: *If + subject + past perfect, subject + past conditional*
  It expresses counterfactual situations that cannot happen because the window of opportunity has closed; an event has passed that prevents the condition from occurring.
  e.g. *If you had <u>sold</u> the property, you would have <u>realized</u> ordinary gains.*

predicate	certainty	polarity	time	special_cases
sold	UNDERSPECIFIED	POS	NON_FUTURE	COND_IF_CLAUSE
realized	CERTAIN	POS	NON_FUTURE	COND_MAIN_CLAUSE

In the following, we report some examples of conditionals containing event mentions annotated as POSSIBLE or PROBABLE (e.g. due to the presence of modals and modal adverbs indicating possibility or probability, or predicates like *seem* or *appear*).

If we <u>cut</u> trees and <u>make</u> houses in those areas, we may <u>cause</u> environmental <u>pollution</u>

predicate	certainty	polarity	time	special_cases
cut	UNDERSPECIFIED	POS	NON_FUTURE	COND_IF_CLAUSE
make	UNDERSPECIFIED	POS	NON_FUTURE	COND_IF_CLAUSE
cause	POSSIBLE	POS	FUTURE	COND_MAIN_CLAUSE
pollution	POSSIBLE	POS	UNDERSPECIFIED	COND_MAIN_CLAUSE

If I <u>liked</u> parties, I would probably <u>attend</u> more of them.

predicate	certainty	polarity	time	special_cases
liked	UNDERSPECIFIED	POS	NON_FUTURE	COND_IF_CLAUSE
attend	PROBABLE	POS	NON_FUTURE	COND_MAIN_CLAUSE

If I <u>became</u> rich, I would probably <u>buy</u> this house.

predicate	certainty	polarity	time	special_cases
became	UNDERSPECIFIED	POS	FUTURE	COND_IF_CLAUSE
buy	PROBABLE	POS	FUTURE	COND_MAIN_CLAUSE

If you had sold the property, you might have realized ordinary gains

predicate	certainty	polarity	time	special_cases
sold	UNDERSPECIFIED	POS	NON_FUTURE	COND_IF_CLAUSE
realized	POSSIBLE	POS	NON_FUTURE	COND_MAIN_CLAUSE

5.2.8 No attribution annotation

For event mentions refering 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. A comment should be added to specify that the attribution has not been annotated by purpose.

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

John rarely finds a good chance to speak about politics.

predicate	certainty	polarity	time	special_cases	comment
finds	CERTAIN	POS	NON_FUTURE	NONE	
speak	-	-	-	-	no attribution annotation

5.2.9 The *pos* attribute

This attribute captures syntactic distinctions among the expressions that are marked as events. It can have the following values which are distinguished using standard criteria in linguistics:

- VERB: it includes both finite (e.g. *halted*) and non-finite forms (e.g. unraveling);
- NOUN: it includes both common (e.g. *suspension*) and proper nouns (e.g. *Conference* in *at the World Wide Developers Conference in San Francisco, California*);
- OTHER: it includes all other parts of speech, such as pronouns, adjectives and prepositional constructions.

5.2.10 The *tense* attribute

It captures standard distinctions in the grammatical category of verbal tense. This attribute is only of interest for verbal events: events that are other parts of speech receive the value NONE. The tense attribute can have any of the following values:

- PRESENT: for events that occur at the time of the speech act (for events marked with the following verb tenses: simple present, present continuous, present perfect and present perfect continuous).
- PAST: for events that occurred before the speech act (for events marked with the following verb tenses: past simple, past continuous, past perfect, past perfect continuous).
- FUTURE: for events that will occur after the speech act (for events marked with the following tenses: future simple, future continuous, future perfect, future perfect continuous).
- INFINITIVE: for events marked with infinitival to.
- PRESPART: for forms marked with -ing and not preceded by the progressive auxiliary be.
- PASTPART: for past participle forms (many of which take an -ed or -en suffix) which are not preceded by the perfective auxiliary have or the passive auxiliary be.
- NONE: for forms which appear in the bare form, such as immediately following a modal auxiliary like can or would. It is used for nouns, adjectives, PPs and pronouns as well.

5.2.11 The *aspect* attribute

The values assigned to this attribute depends on the surface information of markables only. This attribute is only of interested for verbal events: events that are other parts of speech receive the value NONE.

- PROGRESSIVE: for events which can generally be described as continuous or ongoing, marked with the auxiliary be plus a verb taking an -ing suffix.
- PERFECTIVE: for events which can generally be described as completed, marked with the auxiliary have plus a past participle verb form (often taking an -ed or -en suffix).

- PERFECTIVE_PROGRESSIVE: for events which are marked for both perfective and progressive.
- NONE: for events which are in the simple present, past, or future, with no progressive or perfective marking. It is used for nouns and pronouns as well.

5.2.12 Examples of Tense and Aspect Annotation

Tense and aspect attributes will be established as indicated in the following examples (please note that the extension of the event mention tag is underlined):

1. ACTIVE VOICE

tense="PRESENT"

Verb group	aspect=
sells	NONE
is selling	PROGRESSIVE
has sold	PERFECTIVE
has been selling	PERFECTIVE_PROGRESSIVE

tense="PAST"

Verb group	aspect=
sold	NONE
was selling	PROGRESSIVE
had sold	PERFECTIVE
had been selling	PERFECTIVE_PROGRESSIVE

tense="FUTURE"

Verb group	aspect=
will sell	NONE
is going to sell	NONE
will be selling	PROGRESSIVE
is going to be selling	PROGRESSIVE
will have sold	PERFECTIVE
will have been selling	PERFECTIVE_PROGRESSIVE

2. PASSIVE VOICE

tense="PRESENT"

Verb group	aspect=
is sold	NONE
is being sold	PROGRESSIVE
has been sold	PERFECTIVE

tense="PAST"

Verb group	aspect=
was sold	NONE
was being sold	PROGRESSIVE
had been sold	PERFECTIVE

tense="FUTURE"

Verb group	aspect=
will be sold	NONE
is going to be sold	NONE
will have been sold	PERFECTIVE

3. VERBS PRECEDED BY *have to* or *ought to*

tense="PRESENT"

Verb group	aspect=
has to sell	NONE
has to be selling	PROGRESSIVE
has to have sold	PERFECTIVE
has to have been selling	PERFECTIVE_PROGRESSIVE

tense="PAST"

Verb group	aspect=
had to sold	NONE
had to be selling	PROGRESSIVE

tense="FUTURE"

Verb group	aspect=
will have to sell	NONE
will have to be selling	PROGRESSIVE

4. VERBS PRECEDED BY ANY OTHER AUXILIARY, i.e. *must, may, might, can, could, shall, should, and would.*

tense="NONE"

Verb group	aspect=
could sell	NONE
could be selling	PROGRESSIVE
could have sold	PERFECTIVE
could have been selling	PERFECTIVE_PROGRESSIVE

5. PRESENT PARTICIPLE: to be used only for those cases in which the verb form ending in *-ing* occurs in a subordinate clause and it is not preceded by the verb *be*, e.g. *selling* in the sentence *Talanx AG may in the coming days decide against selling shares to the public.*

tense="PRESPART"

Verb group	aspect=
selling	NONE

6. PAST PARTICIPLE: to be used only for those cases in which the participle occurs in a subordinate clause and it is not preceded by any auxiliary form indicating either passive voice or perfective construction, e.g. the verb sold in the sentence *$91M: Value of Facebook Shares Sold by Sandberg*.

tense="PASTPART"

Verb group	aspect=
sold	NONE

7. INFINITIVE

tense="INFINITIVE"

Verb group	aspect=
(to) sell	NONE
(to) be selling	PROGRESSIVE
(to) have sold	PERFECTIVE
(to) have been selling	PERFECTIVE_PROGRESSIVE

8. NOUN

tense="NONE" aspect="NONE"

Examples
an economic crisis
the beginning
a huge acquisition

9. OTHER ELEMENTS

tense="NONE" aspect="NONE"

5.2.13 The *modality* attribute

The modality attribute is only specified if there is a modal word (i.e. may, might, must, would, should, could, can, ought to, have to, or shall) that modifies the mention. This means that it is used to convey the different degrees of modality nature of an event, mainly epistemic and deontic. Its

values are represented by the lemma of modal verb itself. In the following sentence, for example, the modality attribute should have "might" as value: *Microsoft might **sell** Windows 7.*
Sentences that lack a modal auxiliary will not receive any value for this attribute.

6 Temporal Expressions

The <TIMEX3> tag taken from ISO-TimeML is used to annotate temporal expressions including both durations (e.g. *three years*) and points (e.g. *June 15th 2013, today*). Time points can be either absolute (e.g. *the 15th of June, 2013*) or underspecified expressions (e.g. *today*). Markable expressions can also be event anchored (e.g. *two days before the departure*) or sets of times (e.g. *every month*). The list of attributes selected for NewsReader and shown below is a reduced version of the list described in the ISO-TimeML guidelines:

- id, automatically generated by the annotation tool;
- value, it assigns a normalized value based on the ISO-8601 standard to the temporal expression. For example, the expression *June 15, 2013* would get the normalized form 2013-06-13 (YYYY-MM-DD), and the duration *60 days* would get the normalized form P60D (that means Period of 60 Days).
- type, it specifies the type of the temporal expression through 4 values, i.e. DATE, TIME, DURATION and SET.
- functionInDocument, indicates what is the function of a temporal expression in the document and its function as a temporal anchor for other temporal expressions. In NewsReader we adopt only too values: NONE and CREATION_TIME.
- anchorTimeID, introduces the id value of the temporal expression to which the TIMEX3 marked expression is linked in order to compute its value.
- beginPoint and endPoint to strengthen the annotation of durations.
- comment.

All attributes are required in the annotation of temporal expressions, the only optional attribute is the comment; however, anchorTimeID, beginPoint,

and endPoint can be left empty if no information is available (as shown in the BNF below).

BNF of the TIMEX3 tag

attributes ::= id type value [anchorTimeID] functionInDocument [beginPoint] [endPoint] [comment]
id ::= <integer>
value ::= CDATA
type ::= DATE | TIME | DURATION | SET
anchorTimeID ::= IDREF
beginPoint ::= IDREF
endPoint ::= IDREF
functionInDocument ::= CREATION_TIME | NONE
comment ::= CDATA

6.1 Extension

As a general rule, the extent of a TIMEX3 should be as small as possible. In particular, the annotation of temporal expression is restricted to the expressions that contain a so called lexical trigger, that is a word or a numeric expression whose meaning conveys a temporal unit or concept [Ferro *et al.*, 2005].

POS	**Timex Lexical Triggers**
Nouns	minute, afternoon, midnight, day, night, weekend, month, summer, season, quarter, year, decade, century, millennium, era, semester, [the] future, [the] past
Proper names	week days, names of months, Solstice, New Year's Eve, Washington's Birthday
Adjectives	daily, monthly, biannual, daytime
Adverbs	hourly, daily, monthly, now, recent, future, past, present[17]
Time patterns	8:00, 12/2/00, 1994, 1960s
Time nouns/adverbs	now, today, yesterday, tomorrow
Numbers	3 (as in *He arrived at 3*), three, fifth (as in referring to *the fifth of June*), Sixties (as in referring to the decade *the Sixties*)

The extent of the tag can span across multiple words and it must correspond to one of the following categories:

- Nouns: *today, Thursday*

- Noun Phrases: *the morning, Friday night, the last two years, the beginning of the year*
- Adjectives: *current*
- Adverbs: *recently*
- Adjectives or Adverb Phrases: *half an hour long, two weeks ago, nearly a half-hour*

On the other hand, the extent cannot be a Prepositional Phrase (i.e. the extent cannot begin with a preposition) or a clause of any type (e.g. the extent cannot start with a subordinating conjunction). Thus, the following expressions are to be annotated as follows: *before [Thursday]*, *in [the morning]*, *after the strike ended on [Thursday]* , *over [the last 2 years]*.

Premodifiers (including Determiners) are to be included in the extent of a TIMEX3, whereas postmodifiers (including Prepositional Phrases and dependent clauses) are to be excluded: *[that cold day]*, *[no less than 60 days]*, *[next summer]*, *[the future] of our peoples*, *[months] of renewed hostility*.

As far as more complex temporal expressions are concerned, the following rules apply:
- ONE TIMEX3 TAG when there is no intervening token between temporal terms that express values for a single mention of time, such as *[8:00 p.m.]*, *[Friday]*, *[Tuesday the 18th]*, *[this year's summer]*;
- ONE TIMEX3 TAG when, even if there is an intervening token (i.e. a preposition), the expression specifies a unique date or time value: *[the second of December[*, *[ten minutes to three]*, *[eleven in the morning]*, *[summer of 1965]*;
- ONE TIMEX3 TAG when the temporal expression denotes a duration and the conjunction is expressing a specification relation between the temporal element of the duration: *[2 years, 5 months and 3 days] after the natural disasters*
- TWO TIMEX3 TAGS when there is a sequence of two temporal expressions that are ordered one relative to the other generally using temporal prepositions and conjunctions (e.g. *before*, *after*): *John left [2 days] before [yesterday]*;
- TWO TIMEX3 TAGS when there are two temporal expressions that can be related by a temporal link: *The concert is at [8:00 p.m.] on [Friday]*;
- TWO TIMEX3 TAGS when two temporal expressions are part of a range expression: *from [1992] through [1995]*;
- TWO TIMEX3 TAGS when there is a framed duration, that is a duration located within the scope of a temporal unit which has a precise reference in the calendar: *[the first 6 months] of [the year]*.

6.2 Attributes

6.2.1 TYPE

The **TYPE** attribute has 4 possible values:

1. date, all temporal expressions which describe a calendar date, that is an expression with a granularity equal or greater than day (e.g. days, weeks, months, seasons and business quarters, years and decades, centuries and millennia). E.g. *Friday, October 1, 1999*, *yesterday*, *this year's summer*, *last week*.

2. time, points or intervals of time smaller than a day. Clock times are classified as TIME as well. E.g. *2 p.m.*, *five to eight*, *last night*.

3. duration, periods of time. As a rule, if any specific calendar information is supplied in the temporal expression, then the type of the TIMEX3 must be either DATE or TIME. For instance, the expression *1985* cannot be marked as a DURATION, even if the context may suggest that an event holds throughout that year. Temporal expressions like the former "must always be of type DATE, since they refer to a particular area in the temporal axis even though that area spans over a period of time" [ISO TimeML Working Group, 2008]. E.g. *2 months*, *48 hours*, *three weeks*.

4. set, reoccurring time expressions. E.g. *twice a week*, *every 2 days*.

6.2.2 VALUE

The **VALUE** attribute expresses the meaning of a temporal expression in a way that is strictly dependant upon the assigned type value and following the ISO 8601 format.

- DATE format is YYYY-MM-[WW]-DD (that is Year, Month, Week (optional), and Day).

  ```
  <TIMEX3 id="5" type="DATE" value="2012-12-02">
  the second of December 2012
  </TIMEX3>
  ```

  Use the place-holder character, X, for each unfilled position in the value of a component as in:

  ```
  <TIMEX3 id="5" type="DATE" value="XXXX-12-02">
  the second of December
  </TIMEX3>
  ```

Weeks are assigned the position of Months in the date format and their value corresponds to the week number in the calendar of the corresponding year: W01 refers to the first week of the year and W53 to the last one[18]. The following example refers to the week of the second of December 2012:

```
<TIMEX3 id="5" type="DATE" value="2012-W48">
this week
</TIMEX3>
```

To capture the meaning behind the expression *weekend*, first determine which week is intended, and then place a WE token in the day position of the ISO value.

```
<TIMEX3 id="5" type="DATE" value="2012-W48-WE">
this weekend
</TIMEX3>
```

References to the day of the week (i.e., Monday to Sunday) will be expressed with the more complete format: YYYY-Www-D, where D is the weekday number, from 1 (Monday) to 7 (Sunday). This format will be applied if the text presents the trigger *week*, or if the expression is generic:

```
I hate
<TIMEX3 id="5" type="DATE" value="XXXX-WXX-1">
Monday
</TIMEX3>!
```

If an expression can be encoded equally in a month-based or a week-based format, use the month-based representation.

```
<TIMEX3 id="5" type="DATE" value="2012-12-06">
next Thursday
</TIMEX3>
```

References to **months** are specified as: YYYY-MM, whereas references to **years** are expressed as: YYYY.

```
<TIMEX3 id="5" type="DATE" value="2012-12">
December 2012
</TIMEX3>
```

[18]Some websites provide an interface to automatically calculate the week number: see, for example, the ISO week day calendar at `http://www.personal.ecu.edu/mccartyr/isowdcal.html` or `http://www.tuxgraphics.org/toolbox/calendar.html`.

```
<TIMEX3 id="5" type="DATE" value="2012">
2012
</TIMEX3>
```

Decades will be expressed with the format YYY, **centuries** will follow the format YY, and **millennia** will apply the format Y.

```
<TIMEX3 id="5" type="DATE" value="196">
the '60s
</TIMEX3>
```

```
<TIMEX3 id="5" type="DATE" value="19">
the 20th century\footnote{e.g. the expression refers to 1990}
</TIMEX3>
```

```
<TIMEX3 id="5" type="DATE" value="2">
the third millennium\footnote{e.g. the expression refers to
    2005}
</TIMEX3>
```

Seasons are represented using tokens: SP for spring, SU for summer, FA for fall, and WI for winter.

```
<TIMEX3 id="5" type="DATE" value="XXXX-WI">
the winter
</TIMEX3>
```

```
<TIMEX3 id="5" type="DATE" value="1969-SU">
Summer of '69
</TIMEX3>
```

Tokens are used also to express **quarters/trimesters** (Q1, Q2, Q3, Q4), **halves/semesters** (H1, H2) and **fiscal years**.

```
<TIMEX3 id="5" type="DATE" value="2012-Q4">
the 4th quarter of 2012
</TIMEX3>
```

```
<TIMEX3 id="5" type="DATE" value="FY2012-H1">
the first half of FY2012
</TIMEX3>
```

Fuzzy expressions referring to the past (e.g. *former*, *long ago*), the present (e.g. *now/nowadays*, *the present time*) and the future (e.g. *tomorrow* as a generic reference) are normalized using the values PAST_REF, PRESENT_REF, and FUTURE_REF respectively.

The same lexical expression can have a fuzzy interpretation in a sentence (see the annotation of *today* in in the first example below where it has the "nowadays" meaning) and a precise interpretation in another sentence (see the annotation of *today* in the second example below).

there are a wide number of thematic channels on TV

```
<TIMEX3 id="5" type="DATE" value="PRESENT_REF">
today
</TIMEX3>
there are a wide number of thematic channels on TV
```

today the new Icelandic government will take office

```
<TIMEX3 id="5" type="DATE" value="2012-12-04">
today
</TIMEX3>
the new Icelandic government will take office
```

- TIME format is THH:MM:SS, that is Hours, Minutes and Seconds. If minutes and/or seconds are not specified the alternative formats are THH:MM and THH.

  ```
  <TIMEX3 id="5" type="TIME" value="T18:00">
  6 in the afternoon
  </TIMEX3>
  ```

  If the text includes some reference to the specific date in which the time is anchored, then the value attribute must be completed adding the date:

  ```
  <TIMEX3 id="5" type="TIME" value="2012-12-02T18:00">
  6 in the afternoon
  </TIMEX3>
  ```

  A "Z" at the end of the value indicates that the time is explicitly given in Universal Coordinated Time (UTC) or Greenwich Meridian Time (GMT).

  ```
  <TIMEX3 id="5" type="TIME" value="1996-04-11T11:13Z">
  April 11, 1996 11:13 GMT
  </TIMEX3>
  ```

  If a different **time zone** is specified, the difference in terms of whole-hours is added at the end of the values.

```
<TIMEX3 id="5" type="TIME" value="1994-01-21T08:29-05">
January 21, 1994 08:29 Eastern Standard Time
</TIMEX3>
```

Periods of the day are represented with tokens placed at the hour position: MO for morning, MI for midday, AF for afternoon, EV for evening, NI for night and DT for daytime (morning and afternoon together, working hours).

```
<TIMEX3 id="5" type="TIME" value="2012-12-02TMO">
this morning
</TIMEX3>
```

Please note that if a precise time is present, the tokens are not to be used:

```
<TIMEX3 id="5" type="TIME" value="2012-12-02T11:00">
11 am in the morning
</TIMEX3>
```

- DURATIONS are represented by the format PnYnMnDTnHnMnS or PnW. The *n* is to be replaced with the number of date and time elements that follow it whereas the capital letters in the formula mean:

  the [n] is replaced by the value for each of the date and time elements that follow the [n]. Letter P is the duration designator (historically called period) placed at the start of the duration value, and letter T is the time designator, preceding the time components of the representation.

  - P Period designator, stands for *period of*;
  - ML Millennium designator that follows the value for the number of milleniums;
  - CE Century designator that follows the value for the number of centuries;
  - DE Decade designator that follows the value for the number of decades;

- Y Year designator that follows the value for the number of years;
- M Month designator that follows the value for the number of months;
- W Week designator that follows the value for the number of weeks;
- T Time designator, precedes the time components of the representation (i.e. D, H, M, S);
- D Day designator that follows the value for the number of days;
- H Hour designator that follows the value for the number of hours;
- M Minute designator that follows the value for the number of minutes;
- S Seconds designator that follows the value for the number of seconds.

Date and time elements including their designator may be omitted if their value is zero and decimal fraction can be used.

```
<TIMEX3 id="5" type="DURATION" value="P2Y1M3DT4H5M59S">
two year, one months, three days, four hours, five minutes,
   and fifty-nine seconds
</TIMEX3>

<TIMEX3 id="5" type="DURATION" value="P1M">
1 month
</TIMEX3>

<TIMEX3 id="5" type="DURATION" value="PT1M">
one minute
</TIMEX3>

<TIMEX3 id="5" type="DURATION" value="P0.5Y">
half a year
</TIMEX3>
```

Tokens are to be used to represent durations referring to periods of the day (MO, MI, AF, EV, NI, DT), weekends (WE), seasons (SP, SU, FA, WI), quarters (Q), year halves (H), and fiscal years (FY).

```
<TIMEX3 id="5" type="DURATION" value="PT3NI">
three nights
</TIMEX3>
```

The placeholder X must be employed if the interval denoted by the duration cannot be determined by reasoning due to the presence of beginning and ending points, or if it is not explicitly stated in the expression.

```
<TIMEX3 id="5" type="DURATION" value="PXY">
some years
</TIMEX3>
```

- SET the value attribute expresses the time interval in which the iteration (of events or times) takes place:

```
<TIMEX3 id="5" type="SET" value="P1W">
twice a week
</TIMEX3>
```

```
<TIMEX3 id="5" type="SET" value="XXXX-WXX-1">
every Monday
</TIMEX3>
```

```
<TIMEX3 id="5" type="SET" value="XXXX-10">
every October
</TIMEX3>
```

Please note that in the last example, the value looks like a point and not a duration: in this way it's possible to mark the calendar information (i.e. *October*) present in the temporal expression. The general rule, useful to understand when to use a DATE-like annotation instead of a DURATION-like format, is that if there is no specified calendar date (for example, *October* or *Monday*), then the value for the SET will be like that of a DURATION.

6.2.3 functionInDocument

This attribute indicates the function of a TIMEX3 in providing a temporal anchor for other temporal expressions in the document. For our purposes, the only TIMEX3 that serves a function in the document is the document creation time (see 6.4) thus only 2 values are defined:

- CREATION_TIME: the time the text is created;
- NONE: the default value; a general time without a particular reference to the document's life.

6.2.4 anchorTimeID

Some temporal expressions must be anchored to some other TIMEX3 in the text to fill in the correct value. The ID of this TIMEX is given in the anchorTimeID attribute.

Example:
November 29, 2007
Yesterday about 200 New Zealand residents of all ages marched against a current government bill

To know the calendar date corresponding to *Yesterday* we need to identify its temporal anchor, that is another temporal expression which helps us to recover all the necessary information to identify its Year, Month and Day. Imagine this anchor is the time at which the document has been created (i.e. November 29, 2007), whose id is t1, then we will obtain the following representation:

```
<TIMEX3 id="1" type="DATE" value="2007-11-29"
functionInDocument="CREATION_TIME">
November 29, 2007
</TIMEX3>

<TIMEX3 id="2" type="DATE" value="2007-11-28"
anchorTimeID="t1" functionInDocument="NONE">
Yesterday
</TIMEX3>
about 200 New Zealand residents of all ages marched against
a current government bill
```

Example:
The workshop will resume July 15, 2002. The session will start at 9:00 a.m.

```
The workshop will resume
<TIMEX3 id="1" type="DATE" value="2002-07-15">
July 15, 2002
</TIMEX3>
. The session will start at
<TIMEX3 id="2" type="TIME" value="2002-07-15T9:00" anchorTimeID="t1
    ">
```

```
9:00 a.m.
</TIMEX3>
```

The anchorTimeID attribute plays a relevant role in the annotation of empty TIMEX3 tags, anchoring the empty tag to the non-empty tag it's related to (see 6.5).

6.2.5 beginPoint and endPoint

These two attributes are present in tags of type DURATION that are anchored by another time expression, as well as for range expressions. In particular, they are used when a duration is anchored to one or two temporal expressions which signal(s) its beginning and/or ending point(s).

```
from
<TIMEX3 id="1" type="TIME" value="T18:00" functionInDocument="NONE"
   >
6:00 p.m.
</TIMEX3>
to
<TIMEX3 id="2" type="TIME" value="T20:00" functionInDocument="NONE"
   >
8:00 p.m.
</TIMEX3>
<TIMEX3 id="3" type="DURATION" value="P2H"
beginPoint="1" endPoint="2" functionInDocument="NONE"/>
```

6.3 Culturally-Determined Expressions

The interpretation of some temporal expressions requires cultural or domain-specific knowledge. This is the case of holiday names like *Christmas*. Some of these expressions, like *Thanksgiving Day*, contain lexical trigger words and some other do not. A holiday name is markable but should receive a value only when that value can be inferred from the context of the text, rather than from cultural and world knowledge. Otherwise the XXXX-XX-XX placeholder should be used. Expressions that refer to calendars different from the Gregorian one are tagged when they contain a lexical trigger (like *season* in *baseball season* or *year* in *school year*) as illustrated below:

```
<TIMEX3 id="5" type="DATE" value="XXXX-XX-XX" functionInDocument="
   NONE">
baseball season
</TIMEX3>
```

6.4 Annotation of the Document Creation Time

Please note that if the document creation time (DCT) is explicitly written in the document, it has to be annotated as a TIMEX3. The value of the `functionInDocument` attribute should be set to CREATION_TIME. Example:

```
<TIMEX3 id="1" type="DATE" value="1996-03-27"
functionInDocument="CREATION_TIME">
03-27-96
</TIMEX3>
```

6.5 Empty TIMEX3 tag

TimeML allows the creation of empty, non-text consuming TIMEX3 tags whenever a temporal expressions can be inferred from a text-consuming one.
BNF of the empty TIMEX3 tag
attributes ::= id type value [anchorTimeID] [functionInDocument][beginPoint] [endPoint] tag_descriptor [comment]
id ::= <integer>
value ::= CDATA
type ::= DATE | TIME | DURATION | SET
anchorTimeID ::= IDREF
beginPoint ::= IDREF
endPoint ::= IDREF
functionInDocument ::= CREATION_TIME | NONE
tag_descriptor ::= CDATA
comment ::= CDATA

Anchored durations contain a typical duration expression but refer in fact to a point in time (i.e., a date or time of day). They are so called because the duration is explicitly or implicitly anchored to a further temporal reference, that is an anchor.

- *Anchored durations with **implicit** anchoring reference*: the anchoring element is interpreted from the context. In this case an empty tag will be created referring to the implicit anchoring date.

  two months ago (DCT = 2008-12-02 id="t1")

  ```
  <TIMEX3 id="2" type="DURATION" value="P2M" beginPoint="3"
  endPoint="1" functionInDocument="NONE">
  ```

```
two months ago
</TIMEX3>
</TIMEX3 id="3" anchorTimeID="t2" type="DATE" value="
   2008-10-02"
functionInDocument="NONE">
```

some years ago (DCT = 2008-12-02 id="t1")

```
<TIMEX3 id="2" type="DURATION" value="PXY" beginPoint="3"
endPoint="1" functionInDocument="NONE">
some years ago
</TIMEX3>
<TIMEX3 id="3" anchorTimeID="t2" type="DATE" value="PAST_REF"
functionInDocument="NONE" />
```

- *Anchored durations with **explicit** anchoring reference*: the anchoring element is present in the text, two annotations are possible depending on whether the point in time the duration refers to is a date or a time of day.

  - If the resulting temporal expression refers to a DATE, 3 temporal expressions are annotated: i) the duration (underlined), ii) the date encoding the explicit anchoring reference (in bold), iii) the empty TIMEX3 encoding the resulting date of the full construction.

    the earthquake happened <u>*two years ago*</u> ***today*** (DCT = 2008-12-02 id="t1")

    ```
    <TIMEX3 id="2" type="DURATION" value="P2Y" beginPoint="4"
    endPoint="3" functionInDocument="NONE">
    two years ago
    </TIMEX3>
    <TIMEX3 id="3" anchorTimeID="t1" type="DATE"
    value="2008-12-02" functionInDocument="NONE">
    today
    </TIMEX3>
    <TIMEX3 id="4" anchorTimeID="t2" type="DATE" value="
       2007-12-02"
    functionInDocument="NONE" />
    ```

- If the resulting temporal expression refers to a TIME of date: a single TIMEX3 tag of type "time" is annotated.

He arrived at 10 minutes to 3 p.m.

```
<TIMEX3 id="2" type="TIME" value="T14:50" functionInDocument="
   NONE">
10 minutes to 3 p.m.
</TIMEX3>
```

Range expressions: range expressions involve two temporal expressions either of type DATE or of type TIME, which denote the begin and end points of an implicit duration. In this case an empty TIMEX3 tag expressing the duration will be created.

he was Deputy Prime Minister from 2005 to 2008

```
<TIMEX3 id="1" type="DATE" value="2005" functionInDocument="NONE">
2005
</TIMEX3>
to
<TIMEX3 id="2" type="DATE" value="2008" functionInDocument="NONE">
2008
</TIMEX3>
<TIMEX3 id="3" type="DURATION" value="P3Y" beginPoint="1" endPoint=
   "2"
functionInDocument="NONE" />
```

Framed durations: framing relations denote a time interval and contain a date and a duration expressions, differently to range expressions that denote time intervals using two temporal expressions of type DATE. The date refers to a particular temporal frame within which the duration is located, i.e. it expresses one of the boundaries of the interval. In this cases two empty tags of type DATE will be created to express the begin and end points from which the length of the duration is computed.

Howard Dean raised $1.77 million in the first six months of the year (DCT = 2008-12-02 id="t1")

```
<TIMEX3 id="2" type="DURATION" value="P6M" beginPoint="3" endPoint=
   "4"
functionInDocument="NONE">
the first six months
</TIMEX3>
of
<TIMEX3 id="3" type="DATE" value="2008" anchorTimeID="t1"
   functionInDocument="NONE">
```

```
the year
</TIMEX3>
<TIMEX3 id="4" type="DATE" value="2008-01" anchorTimeID="3"
    functionInDocument="NONE" anchorTimeID="t3"/>
<TIMEX3 id="5" type="DATE" value="2008-06" anchorTimeID="3"
    functionInDocument="NONE" />
```

6.6 Tag Descriptor for Temporal Expressions

It is the nominal identifier of empty TIMEX3 tags, which can be useful to distinguish the temporal expressions in the annotation interface. We suggest to use the content of the "value" attribute as identifier.

7 Numerical Expressions

Given their relevance in the economic and financial domain, a markable VALUE has been created for numerical expressions, i.e. for amounts (distinguishing between monetary and general amounts) and for percentages. Each value tag has the following attributes:

- id: automatically assigned by the tool;
- type: has 3 possible values: PERCENT (e.g. *2.1 percent*), MONEY used for capitals described in terms of currencies (e.g. *20 Euros*), QUANTITY used for numbers of items (e.g. *more than 500*);
- comment.

BNF of the VALUE tag
attributes ::= id type [comment]
id ::= <integer>
type ::= PERCENT — MONEY — QUANTITY
comment ::= CDATA

A numerical expression can have an *indicator*, which is used to express the type of the expression itself: the symbols *%* and *$* are examples of indicators of the PERCENT and MONEY types respectively. An indicator can be expressed either as a symbol or as a string of words (e.g. *$* and *dollars*). Also the number can be either a numeral or a string of words (e.g. *15* and *fifteen*). The extent of a VALUE is the smallest string of words that includes both the number and the indicator (if present) and also any additional quantifiers

that might be present such as *nearly*, *almost*.

5 percent of the organizations
[5 percent]VALUE OF TYPE=PERCENT
[the organizations]MENTION OF AN ORG ENTITY - syntactic_type = NOM
[5 percent of the organizations]MENTION OF AN ORG ENTITY – syntactic_type = PTV

VW had sold nearly 400,000 cars
[nearly 400,000]VALUE OF TYPE=QUANTITY
[nearly 400,000 cars] MENTION OF A PRODUCT ENTITY - syntactic_type = NOM

8 Signals

The tag <SIGNAL>, inherited from ISO-TimeML, is used to annotate all those textual elements which make explicit a temporal relation (i.e. a TLINK, see 10.6) between two event mentions, two temporal expressions, or an event mention and a temporal expression.

The range of linguistic expressions which are to be marked as signals is restricted to:

- Temporal prepositions: *on, in, at, from, to, before, after, during,* etc.;
- Temporal conjunctions: *before, after, while, when*, etc.;
- Temporal adverbs: *meantime, meanwhile*, etc.;
- Special characters: - and /, in temporal expressions denoting ranges (e.g. 26 - 28 September 2006).

The extension is limited to the functional word.

```
<SIGNAL id="1">
on
</SIGNAL>
Monday
```

The tag contains only two attributes: id and comment.

BNF of the SIGNAL tag
attributes ::= id [comment]
id ::= <integer>
comment ::= CDATA

9 C-Signals

The <C-SIGNAL> tag is used to mark-up textual elements that indicate the presence of a causal relation (i.e. a CLINK, see 10.3). More specifically, annotators should identify all causal uses of:

- prepositions, e.g. *because of, on account of, as a result of, due to*;
- conjunctions, e.g. *because, since, so that, hence, thereby, by*;
- adverbial connectors, e.g. *as a result, so, therefore*;
- clause-integrated expressions, e.g. *the result is, the reason why.*

The extent of the tag corresponds to the whole expression, so multi-token extensions are allowed.

The tag contains only two attributes: id (automatically assigned by the annotation tool) and comment.

BNF of the C-SIGNAL tag
attributes ::= id [comment]
id ::= <integer>
comment ::= CDATA

10 Relations

There are six kinds of relations in a NewsReader document: they are used to signal different types of links which may exist between annotated markables (i.e. coreference, syntactic dependencies, causality, temporal ordering, subordination). Two different argument slots are provided for each relation: the first argument is the *source* of the relation and the second argument is the *target*. The arguments taking part to the link are encoded into self-contained elements (i.e. <source .../>, <target .../>). This solution allows for a general, uniform, mechanism for indicating the source and target of any relation between markables, and at the same time allows for handling, if necessary, many-to-many, one-to-many and many-to-one relations. In particular, in NewsReader a relation can have more than one source (see the coreference relation 10.1).

For sake of simplicity, the examples in the following subsections use a non-XML syntax.

10.1 REFERS_TO (Intra-document coreference)

The REFERS_TO relation represents the coreference between an entity mention (i.e. the tag <ENTITY_MENTION>) and an entity instance (i.e. the tag <ENTITY>), and between an event mention (i.e. the tag <EVENT_MENTION>) and an event instance (i.e. the tag <EVENT>). It is a directional, many-to-one relation because many mentions can refer to the same entity (as in 1) or event (as in 2).

1. ***Qatar Navigation*** *jumped 6.4 percent after* ***the company*** *said* ***it*** *scraped plants for a 20 percent capital increase.*

2. *Indonesia's West Papua province was hit by a magnitude 6.1* ***earthquake*** *today, the latest powerful* ***tremor*** *to shake the region.*

As for directionality, the source/s of the relation is/are the mention/s whereas the target is the instance it/they refer to.

BNF of the REFERS_TO relation
attributes ::= id [comment]
id ::= <integer>
comment ::= CDATA

In case of appositional constructions and appositions with relative clause, all the components forming these complex constructions and the APP/ARC-mentions themselves are to be linked with the instance they refer to. In *Bill, John's lawyer, is very well-paid*, the mentions *Bill*, *John's lawyer* and the APP construction *Bill, John's lawyer* are to be annotated as sources of a REFERS_TO link.

Event mentions corefer if their discourse elements (e.g. agents, location, and time) are identical in all respects, as far as one can tell from their occurrence in the text [Hovy *et al.*, 2013]. This means that the mentions are fully identical because there is no semantic difference between them. From the practical point of view, it is possible to replace one mention with the other one, sometimes with just some small syntactic modifications, without any semantic change. Different types of perfect coreference can be identified:

- Lexical identity: The two mentions use exactly the same senses of the same word, including derivational words. E.g. *to acquire* and *acquisition*, *to suspend* and *suspension*

- Synonym: One mention's word is a synonym of the other word. E.g. *to suspend* and *to halt*).

- Wide reading: one mention is a synonym of the wide reading of the other. E.g. *in the attack$_{em1}$ took place yesterday. The bombing$_{em2}$ killed four people*, *em1* and *em2* fully corefers because *bombing* is read in its wide sense denoting the whole attack. The fact that the two mentions corefer is understood from the context.

- Paraphrase: one mention is a paraphrase of the other. Some syntactic changes can be present: e.g. active/passive transformation (*the stock exchange suspended trading / trading was suspended by the stock exchange*) or shifts of perspective but without adding any extra semantic information.

- Pronoun: e.g. (*the party / **that** event*), (*the election went well / **it** went well*).

10.2 HAS_PARTICIPANT (Participant Roles)

The HAS_PARTICIPANT relation links an event mention (source) to an entity mention or to a numerical expression (target) which plays a role in the event. More specifically, it is a directional, one-to-one relation.

The assignment of semantic role labels is encoded through 2 attributes. The first attribute is called *sem_role_framework*: it defines the framework used as a reference and it has three possible choices, i.e. FrameNet, PropBank and KYOTO. The default value is PropBank but the other options are maintained for future possible annotations. The second attribute is named *sem_role*: it encodes the precise value of the semantic role of the participant involved in the relation.

In PropBank there are 5 numbered arguments (corresponding either to the required arguments of a predicate, e.g. agent and patient, or to those arguments that occur with high-frequency in actual usage). In general numbered arguments correspond to the following semantic roles [Bonial *et al.*, 2010]:

1. ARG0: agent;

2. ARG1: patient;

3. ARG2: instrument, benefactive, attribute;

4. ARG3: starting point, benefactive, attribute;

5. ARG4: ending point.

As for modifiers (i.e. ARGM in PropBank) we decided to give special attention to locative modifiers annotated as entity mentions of type LOC. For these modifiers the value ARGM-LOC has been created. All other modifiers are to be annotated using the value ARGM-OTHER. This last value includes, for example, comitatives, goal, and extent modifiers [Bonial *et al.*, 2010].

[Mitt Romney]$_{\text{ENTITY MENTION - PER}}$ *[met]*$_{\text{EVENT MENTION - OTHER}}$ *[the investors]*$_{\text{ENTITY MENTIO}}$
in [Miami]$_{\text{ENTITY MENTION - LOC}}$.
Mitt Romney: ARG0
the investors: ARG1
Miami: ARGM-LOC

[The Dow Jones Industrial Average]$_{\text{ENTITY MENTION - FIN}}$ *[moved]*$_{\text{EVENT MENTION - OTHER}}$
[50 points]$_{\text{VALUE - PERCENTAGE}}$.
The Dow Jones Industrial Average: ARG1
50 points: ARGM-OTHER

In case the participant of an event mention is expressed through a complex mention constructions, i.e. CONJ, APP and ARC, only the longest mention is to be annotated as target of the HAS_PARTICIPANT relation.

Annotators can look up to the list of English frameset that provide many useful examples and the explanation of semantic role attribution for each lemma: `http://verbs.colorado.edu/propbank/framesets-english/`.

BNF of the HAS_PARTICIPANT relation
attributes ::= id sem_role_framework sem_role [comment]
id ::= <integer>
sem_role_framework ::= FRAMENET | PROPBANK | KYOTO
sem_role ::= ARG0 | ARG1 | ARG2 | ARG3 | ARG4 | ARG5 | ARGM-LOC | ARGM-OTHER
comment ::= CDATA

10.3 CLINK (Causal Relations)

In NewsReader, we annotate causal relations between causes and effects denoted by event mentions through a link named CLINK.

As far as directionality is concerned, the source of the relation, that is the first argument, is always the causing event and the target of the relation, that is the second argument, is always the caused event.

BNF of the CLINK tag
attributes ::= id [c-signalID] [comment]
id ::= <integer>
c-signalID ::= IDREF
comment ::= CDATA

There are 3 basic categories of causation: CAUSE, ENABLE, PREVENT. We will annotate all three types of causation but only if there is an explicit causal constructions between two event mentions. In the examples below the causative verbs are in bold and the event mentions involved in the CLINK are underlined:
*the purchase$_{em1\text{-}source}$ **caused** the creation$_{em2\text{-}target}$ of the current building*
*the purchase$_{em1\text{-}source}$ **enabled** the diversification$_{em2\text{-}target}$ of their business*
*the purchase$_{em1\text{-}source}$ **prevented** further changes$_{em2\text{-}target}$ from being made.*

Among all causal expressions [Wolff *et al.*, 2005], only those explicitly asserting a causal relation between two event mentions are to be annotated[19], as detailed below:

- Expressions containing **CAUSE-type verbs** (e.g. *cause, influence, inspire, push, stimulate*), **ENABLE-type verbs** (e.g. *enable, aid, allow, help, permit*), and **PREVENT-type verbs** (e.g. *prevent, block, discourage, dissuade, hinder, impede*).
  *the purchase$_{em1\text{-}source}$ **stimulated** the creation$_{em2\text{-}target}$ of new products*
  *the purchase$_{em1\text{-}source}$ **enabled** the diversification$_{em2\text{-}target}$ of their business*
  *the purchase$_{em1\text{-}source}$ **prevented** the attempt$_{em2\text{-}target}$.*

- Expressions containing **affect verbs**, such as *affect*, *influence*, *determine*, *change* (they can be replaced with *cause*, *enable*, or *prevent*).
  *Ogun ACN crisis$_{em1\text{-}source}$ **affects** the launch$_{em2\text{-}target}$ of the All Progressives Congress*
  Ogun ACN crisis causes the launch of the All Progressives Congress
  Ogun ACN crisis enables the launch of the All Progressives Congress
  Ogun ACN crisis prevents the launch of the All Progressives Congress

- Expressions containing **link verbs**, such as *link, lead, depend on* (thay can be replaced only with *cause* and *enable*).

[19]Lexical causatives such as *break, melt, kill* contain the meaning of causation in their lexical meaning (e.g. *kill* has the embedded meaning of causing someone to die) but are not involved in CLINKs [Huang, 2012].

An earthquake$_{em1\text{-}source}$ *in North America was* ***linked*** *to a tsunami*$_{em2\text{-}target}$ *in Japan*
An earthquake in North America was caused by a tsunami in Japan
An earthquake in North America was enabled by a tsunami in Japan
**An earthquake in North America was prevented by a tsunami in Japan*

- Expressions containing **causative conjunctions and prepositions**, such as:
  - prepositions, e.g. *because of, on account of, as a result of, in response to, due to, from, by*;
  - conjunctions, e.g. *because, since, so that, hence, thereby, by*;
  - adverbial connectors, e.g. *as a result, so, therefore, thus*;
  - clause-integrated expressions, e.g. *the result is, the reason why, that's why.*

  Please note that causative conjunctions and prepositions are annotated as C-SIGNALs (see Section 9) and their ID is to be reported in the c-signalID attribute of the CLINK[20].

  In some contexts, the coordinating conjunction *and* can imply causation; given the ambiguity of this construction and the fact that it is not an explicit causal construction, however, we do not annotate CLINKs between two events connected by *and*. Similarly, also the temporal conjunctions *after* and *when* can implicitly assert a causal relation but should not be annotated as C-SIGNALs and no CLINKs are to be created; temporal conjunctions must instead be annotated as SIGNALs (involved in TLINKs).

- **Periphrastic causatives** are generally composed of a matrix verb that takes an embedded clause or predicate as a complement; for example, in the sentence *The blast caused the boat to heel violently*, the matrix verb (i.e. *caused*) expresses the notion of CAUSE while the embedded verb (i.e. *heel)* expresses a particular result.

  Periphrastic causative verbs fall into three categories:

  1. CAUSE-type verbs: *bribe, cause, compel, convince, drive, have, impel, incite, induce, influence, inspire, lead, move, persuade, prompt, push, force, get, make, rouse, send, set, spur, start, stimulate*

[20]The absence of a value for the c-signalID attribute means that the causal relation is encoded by a verb.

2. PREVENT-type verbs: *bar, block, constrain, deter, discourage, dissuade, hamper, hinder, hold, impede, keep, prevent, protect, restrain, restrict, save, stop*
3. ENABLE-type verbs: *aid, allow, enable, help, leave, let, permit*

The recognition of ENABLE-type causal relations is not always straightforward. The suggestion is to try rephrasing the sentence using the *cause* verb:
a) *The board authorized the purchase of the stocks*
b) *The authorization of the board caused the stocks to be purchased*
The verb authorize proves to be an ENABLE-type verb. In a) a CLINK is annotated between *authorize* and *purchase*; in b) a CLINK is annotated between *authorization* and *purchased*.

Please note that there is an implicit temporal relation between the causing event and the caused one: i.e. the first always occurs before the second. We need not create a TLINK between these two events as the temporal link can be easily inferred.

As for the other relations, also CLINK has two self-contained elements to encode the source (the event mention encoding the cause) and the target (the event mention encoding the effect) of the link.

10.4 SLINK (Subordinating Relations)

Annotation of reported speech leans on TimeML approach, which uses SLINKs (i.e. subordinating links) to connect REPORTING, I_STATE and I_ACTION verbs to their event arguments. In NewsReader, we reduce the scope of SLINKs using it to annotate the subordinating relation between an event mention belonging to the SPEECH_COGNITIVE class (the source of the relation) (see Section 4.2) and the event denoting its complement and expressing the message of reported or direct utterance/thought (the target).

For example, in *The stock exchange suspended trading for half an hour, a source at the exchange told Agence France Presse.* the event mention *told* is linked to the subordinated event mention *suspended* through an SLINK relation.

In a direct speech such as *"It sounded like a jet or rocket" said Eddie Gonzalez.*, the event mention *said* is linked to the subordinated event mention *sounded* through an SLINK relation.

In some cases the same SPEECH_COGNITIVE event mention will introduce more than one SLINK. For instance, in the example below the event said is slink-ed to two events: *listed* and *gave*.
Rita said they correctly listed his name but gave a false address for him.

As for the other relations, also SLINKs have two self-contained elements to encode the source and the target of the link.

BNF of the SLINK relation
attributes ::= id [comment]
id ::= <integer>
comment ::= CDATA

10.5 GLINK (Grammatical Relations)

A GLINK relation is used to link a mention of an event of type GRAMMATICAL (the source of the relation) (see Section 4.2) to the mention of the event encoding its governing content verb or noun (the target). For example, this relation holds between:

- an aspectual verb or noun (em1, source) and its event argument (em2, target) as in *the beginning$_{em1}$ of the crisis$_{em2}$*;
- a verb or a noun expressing occurrence (em1, source) and the occurred event (em2, target) as in *the share drop$_{em2}$ came$_{em1}$ on the same day*;
- a causal verb or noun (em1, source) and the caused event (em2, target) as in *the purchase enabled$_{em1}$ the diversification$_{em2}$ of their business.*

As for the other relations, also GLINKs have two self-contained elements to encode the source and the target of the link.

BNF of the GLINK relation
attributes ::= id [comment]
id ::= <integer>
comment ::= CDATA

10.6 TLINK (Temporal Relations)

The TLINK is used for temporal links between two event mentions, two temporal expressions or between an event mention and a temporal expression.

In order to create storylines, it is important to link each event with (at least) one other event in the text.

For each relation, the following attributes are defined:

- id, automatically generated by the annotation tool;
- reltype, indicating how the two elements are temporally related;
- signalID, it represents the ID of the SIGNAL that explicitly signaled the presence of a TLINK.

As for the REFERS_TO link, also TLINKs have two self-contained elements to encode the source and the target of the link.

BNF of the TLINK tag
attributes ::= id [signalID] relType [comment]
id ::= <integer>
signalID ::= IDREF
relType ::= BEFORE | AFTER | INCLUDES | IS_INCLUDED | SIMULTANEOUS | IAFTER | IBEFORE | BEGINS | ENDS | BEGUN_BY | ENDED_BY | MEASURE

`TLINK` Directionality.

We suggest to create `TLINKs` following the linear order of the sentence: the event mention or TIMEX3 which first appears in the sentence is the source, the one which appears as second is the target. The only strict directionality rule (justified by the fact that MEASURE has no inverse relaton) is given for the `TLINK` of type `MEASURE`, in which the source is always the `TIMEX3` and the target is the Event.

10.6.1 Relation types

The possible values of the `relType` attribute are: `BEFORE`, `AFTER`, `IBEFORE`, `IAFTER`, `INCLUDES`, `IS INCLUDED`, `MEASURE`, `SIMULTANEOUS`, `BEGINS`, `BEGUN BY`, `ENDS`, and `ENDED BY`[21].

Many of the possible values are binary, one being the inverse of the other: `BEFORE` and `AFTER`, `INCLUDES` and `IS_INCLUDED`, `BEGINS` and `BEGUN_BY`, `ENDS` and `ENDED_BY`, `IBEFORE` and `IAFTER`.

For all relTypes except `MEASURE`, each source and target element in the link can involve either an `EVENT` or a `TIMEX3`.

[21] Temporal relations are the same as defined in ISO-TimeML with one exception: we have eliminated the IDENTITY temporal relation, which is not needed in NewsReader as coreferential relations are annotated by using the REFERS_TO link.

For clarity's sake, in the following examples we will present only the annotation of the relevant entities involved in the temporal relations.

1. BEFORE, an event/timex occurs before another, e.g. *She **arrived**<S-EVENT> before his cousin **departure**<T-EVENT>*, *Some young men **robbed**<S-EVENT> the house before **leaving**<T-EVENT> the town*;

2. AFTER, the inverse of BEFORE, e.g. *Some young men **left**<S-EVENT> the town after **robbing**<T-EVENT> the house*;

3. INCLUDES, one event/timex includes the other, *While **washing**<S-EVENT> the dishes John **dropped**<T-EVENT> two plates*;

4. IS_INCLUDED, the inverse of INCLUDES, e.g. *John **left**<S-EVENT> on **Monday**<T-TIMEX3>*;

5. MEASURE, it is used to connect an event and a DURATION TIMEX which provides information on the duration of the related event (i.e. one which answers to the question "how long does/did the event X last?"), e.g. *Marc **worked**<T-EVENT> **one hour**<S-TIMEX3>*, *John **ran**<T-EVENT> for **twenty minutes**<S-TIMEX3>*;

6. SIMULTANEOUS, two events happen at the same time, e.g. *Mary was **watching**<S-EVENT> TV while John was **frying**<T-EVENT> the eggs*, or an event is perceived as happening at a moment (point or interval) in time, e.g. ***Now**<T-TIMEX3> she is **resting**<S-EVENT>*. More specifically, SIMULTANEOUS is assigned to two markables either when they are perceived as happening at the same time, or when they temporally overlap, or when they occur close enough that it is not possible to further distinguish their times. We can have a `SIMULTANEOUS` relation between an `EVENT` and a `DATE/POINT`, but not between an `EVENT` and a `DURATION`. SIMULTANEOUS is also assigned to the event arguments of perceptions verbs. Examples: *When Wong Kwan **spent** 16 million dollars to buy his house, he **thought** it was a good price*, *I **heard** several **explosions***, *Marc **arrived** at **3***.

7. IMMEDIATELY BEFORE (IBEFORE), one event/timex occurs immediately before the other, e.g. *In the **crash**<S-EVENT> all passengers **died**<T-EVENT>*. This relation is a specification of the more general BEFORE relation. It is not very much widespread in documents. Its annotation is subordinated to the presence of specific signals, like *immediately before*, or other discourse elements which indicate that the temporal span between the two entities involved is very short;

8. IMMEDIATELY AFTER (IAFTER), the opposite of IBEFORE, e.g. *One of the eggs* ***broke<S-EVENT>*** *as soon as it* ***touched<T-EVENT>*** *the pan.* This relation is a specification of the more general AFTER relation. It is not very much widespread in documents. Its annotation is subordinated to the presence of specific signals, like *immediately after*, or other discourse elements which indicate that the temporal span between the two entities involved is very short;

9. BEGINS, a timex or an event (the source of the relation) marks the beginning of another timex or event, e.g. *Since he* ***graduated<S-EVENT>*** *he has been* ***looking<T-EVENT>*** *for a job*;

10. BEGUN_BY, the inverse of BEGINS, the beginning of an event is marked by another event or timex(the target of the relation), e.g. *He has* ***worked<S-EVENT>*** *for them since he* ***graduated<S-EVENT>***, *We have been* ***looking<S-EVENT>*** *for a solution since* ***yesterday<T-TIMEX3>***;

11. ENDS, a timex or an event (the source of the relation) marks the ending of another event or timex, e.g. *Until the doorbell* ***rang<S-EVENT>*** *Marc had been* ***sleeping<T-EVENT>***;

12. ENDED_BY, the inverse of ENDS, the end of an event is marked by another event (the target of the relation), e.g. *Marco* ***drove<S-EVENT>*** *until* ***midnight<T-TIMEX>***, *Marc* ***slept<S-EVENT>*** *until the doorbell* ***rang<T-EVENT>***.

In the case of binary relation types, the choice of one of the two depends on the application of the directionality rules. In the following sentence, for example, we have a BEFORE TLINK but not the inverse relation AFTER.

Some young men `<EVENT eiid="ei1">`*robbed*`</EVENT>` *the house* `<SIGNAL id="1">`*before* `</SIGNAL>` `<EVENT eiid="ei2">`*leaving*`</EVENT>` *the town.*

```
<TLINK eventInstanceID="ei1" relatedToEventInstance="ei2"
signalID="s1" relType="BEFORE"/>
```

10.6.2 Subtasks for the Annotation of TLINKs

In order to simplify the explanation, in the following subsections the annotation procedure is divided into a set of subtasks. The complete set of temporal relations can be obtained by merging the different subtasks.

Subtask 1: TLINKs between Event Mentions and the DCT

This task makes explicit the temporal relations between events and the Document Creation Time (DCT) which corresponds to the moment of utterance.
The following guidelines apply:

a.) non verbal events will never be linked to the DCT ;
the auction took place yesterday – no TLINK between *auction* and the DCT

b.) events realized by finite verb forms will be linked to the DCT according to the tense and aspect values of the verb form;
*Spanish 10-year bond yields will **fall** decisively below 5%* – `TLINK="after"`
*Spanish 10-year bond yields are **falling** decisively below 5%* – `TLINK="is_included"`
*Spanish 10-year bond yields **fell** decisively below 5%* – `TLINK="before"`
*Spanish 10-year bond yields have **fallen** decisively below 5%* – `TLINK="before"`

c.) modal verb + verb_INFINITIVE: the verb at the infinitive will be linked to the DCT according to the tense values of the modal verb

d.) copulative predicates (such as"*be/seem*") + eventive NOUN: only the copular verb is linked to the DCT according to the tense;
*it **is** a long-term economical **crisis*** – only *is* has a TLINK with the DCT

e.) verbs at the simple infinitive, ing_form and past participle will not have a temporal relation with the DCT;
*The same interest rate will be maintained while **selling** the bonds this year* – no TLINK between *selling* and the DCT

g.) verb (except modals) + verb_INFINITIVE or ing_form: only the first verb has a relation with the DCT according to the tense.
*Madrid **hoped** to **sell** up to 5 billion of bonds* – only *hoped* has a TLINK with the DCT
*How can the CEO **avoid hurting** his credibility?* – only *avoid* has a TLINK with the DCT

Note that, in general, in case the constructions listed above are modified by a temporal expression, then the TLINK is computed on the basis of the temporal relation between the DCT and the modifying temporal expression.

Subtask 2: TLINKs between Main Event Mentions

Main event mentions correspond to the ROOT element of the parsed sentence. Each sentence may have just one main event.
Some special rules to determine the main events:

- in case the main event mention is realized by a light verb construction or by a copular verb (such as"*be/seem*"), only the verb head must be considered as the main event;
- event mentions in comparative contexts in which the subordinated event mention is introduced by conjunctions such as *as*, *more than*, *less than* cannot be considered as main event;
- mentions of SPEECH_COGNITIVE events used to introduce a reported speech, a direct speech, or a thought are to be considered as main events;
- non-verbal event mentions can be considered as main event mentions only when they are in NP sentences with no verbal elements.

In order to create timelines, the full temporal structure of the document should be annotated thus the annotation of TLINKs across different sentences is particular important to avoid gaps in the temporal graph. On the other hand, however, the annotation of cross-sentential relation is error-prone. To reduce annotation mistakes, the following rules have been defined:

a.) identify the main event mention in sentence A and in its following adjacent sentence (sentence B);

b.) if no temporal relation can be established between the two event mentions, then check if there is a temporal relation between the main event mention in sentence A and the main event mention in the following adjacent sentence (sentence B+1);

c.) iterate the procedure at point b.) until a temporal relation is identified.

In determining the temporal relations between main event mentions, annotators should pay attention to the following elements:

- if each of the main event mention is modified by a temporal expression, assign the temporal relations on the basis of the temporal relation which exists between the two temporal expressions;
- if two main event mention share a participant, it is likely that they stand in temporal relation;
- if there is a signal, assign the temporal relations on the basis of the temporal relations which is coded by the signal;
- tense and aspect may restrict the set of possible temporal relations;

- two event mentions may stand in particular semantic relations which may correspond to possible temporal relations:

  1. entailment relations, i.e. *if event mention A then event mention B* ⟶ `includes`, `is_included`, `before` or `after`;
  2. causative relations ⟶ `before` or `after`.

Subtask 3: TLINKs between Main Event Mention and Subordinated Event Mention in the Same Sentence

The subordinated event is identified on the basis of syntactic relations of dependencies and is restricted to clausal realizations, either finite or non finite (in other words, the subordinated event is a verbal event mention).

No TLINK is established in the following cases:

- MAIN EVENT MENTION + SUB. FINAL CLAUSE (e.g. in *Production was halted to correct the mistake* no TLINK is created between *halted* and *correct*);
- SEEM + PREDICATIVE COMPLEMENT and other copulative constructions: e.g. in *it seems a huge economic crisis* no TLINK is created between *seems* and *crisis*[22];
- MAIN EVENT MENTION + event mention in relative clauses, e.g. in *the stock market, that was halted by automatic curbs, fall 10 percent* no TLINK is created between *halted* and *fall*.

On the contrary, TLINKs are established when the main verb is the verb *want*: e.g. in *I want to leave* and *I want you to leave*, a TLINK is created between *want* and *leave*.

When the subordinated event mention is realized by a finite clause, possible values of the TLINKs are reported in the following schema. Notice that: (i.) the schema described below provides the most likely values but it is not rigid; (ii.) in case there are temporal expressions or signals in the main or in the subordinated clause, annotators should use this information to order the events accordingly.

- The tense of the main event mention is PRESENT (i.e. simultaneous with the DCT):

[22] We annotate a GLINK between *seem* and the predicative complement or between the two parts of the copulative construction (see Section 10.5).

  - the subordinate clause is at the INDICATIVE mood:
    * `TLINK="simultaneous"` = the tense of the subordinated event mention is present
      I know$_{main}$ you are$_{sub}$ tired.
      `TLINK="simultaneous"`

    * `TLINK="after"` = the tense of the subordinated event mention is past
      I know$_{main}$ you were$_{sub}$ tired.
      `TLINK="after"`

    * `TLINK="before"` = the tense of the subordinated event mention is future
      I know$_{main}$ you will be$_{sub}$ tired.
      `TLINK="before"`

- The tense of the main event mention is a PAST (i.e. before the DCT):
  - the subordinate clause is at the INDICATIVE mood:
    * `TLINK="simultaneous|is_included"` = the tense of the subordinated event mention is past perfect or past continuous
      I knew$_{main}$ you were$_{sub}$ tired.
      `TLINK="simultaneous"`
      He told me$_{main}$ that Mary was sleeping$_{sub}$.
      `TLINK="is_included"`

    * `TLINK="after"` = the tense of the subordinated event mention is simple past or past perfect
      I knew$_{main}$ you had been$_{sub}$ sick.
      `TLINK="after"`
      He asked$_{main}$ her what she had seen$_{sub}$.
      `TLINK="after"`
      They told me$_{main}$ you won$_{sub}$.
      `TLINK="after"`
  - the subordinate clause is at the CONDITIONAL mood:
    * `TLINK="before"` = conditional (future-in-the-past)
      I knew$_{main}$ you would arrive$_{sub}$.
      `TLINK="before"`
      He told$_{main}$ her that he would never leave$_{sub}$ her.
      `TLINK="before"`

- The main event is at a tense of the FUTURE (i.e. after the DCT):
  - the subordinate clause is at the INDICATIVE mood:
    * `TLINK="simultaneous` = the tense of the subordinated event mention is present or simple future
      I will tell$_{main}$ him that you are$_{sub}$ tired.
      `TLINK="simultaneous"`
      I will tell$_{main}$ them that you will soon get$_{sub}$ tired.
      `TLINK="simultaneous"`

    * `TLINK="after"` = a PAST tense
      I will tell$_{main}$ him that you were$_{sub}$ sick.
      `TLINK="after"`
      I will ask$_{main}$ him what he had seen$_{sub}$.
      `TLINK="after"`

    * `TLINK="before"` = the tense of the subordinated event mention is present (with future reading) or simple future
      He will promise$_{main}$ that he will go$_{sub}$ to the beach.
      `TLINK="before"`
      Tomorrow I will ask$_{main}$ him where he is spending$_{sub}$ his holyday.
      `TLINK="before"`

Subtask 4: TLINKs between Event Mentions and Timexes in the Same Sentence

The identification of the event mention which is linked to the temporal expression(s) is based on the following rules which differentiate according to the part of speech of the event mention or its context of occurrence.

Non verbal event mentions: it is possible to identify a TLINK between a non-verbal event mention (e.g. a noun) and a temporal expression when the temporal expression "modifies" the non-verbal event mention, e.g.:
The 1992$_{timex}$ evaluation$_{event}$.
The former$_{timex}$ CEO$_{event}$.
Yesterday's$_{timex}$ meeting$_{event}$.

One specific heuristic has been developed for the following context:

- non-verbal event mention + verbal event mention denoting temporal

movement and extension[23] + TIMEX: there is a temporal relation between the non-verbal event and the timex. *The meeting$_{event}$ has been postponed until tomorrow$_{timex}$.*

Verbal event mentions: in case of event mentions realized by verbs, the following rules apply:

- if the timex is in the main clause, then there is a TLINK between the main verb mention and the timex;

  Tomorrow$_{timex}$ the CEO will resign$_{mainevent}$.

  – `TLINK="is_included"`;

- if the timex is in a subordinated sentence, then there is a TLINK between the verb of the subordinated clause and the timex;

  The president thinks that it will be approved$_{event}$ tomorrow$_{timex}$.

  – `TLINK="is_included"`;

- if a coordination relation stands between two or more verbs, then there is a TLINK between the timex and each verb in the coordination (at the same syntactic level);

  The president will announce$_{mainevent}$ that tomorrow$_{timex}$ he will approve$_{subevent}$ the proposal and increase$_{subevent}$ the funding.

  – `TLINK="is_included"`;

  `TLINK = approve`$_{subevent}$ `- tomorrow`$_{timex}$

  – `TLINK="is_included"`;

  `TLINK = increase`$_{subevent}$ `- tomorrow`$_{timex}$

- in case the timex is in the main sentence and the subordinated sentence is a temporal clause introduced by signals (e.g. *when, as, as soon as, once, after, before, etc.*), then there is a TLINK between the main event mention of the main clause and the timex and a TLINK between the verb of the subordinated temporal clause and the timex.

  Today$_{timex}$ the president will resign$_{mainevent}$ before$_{signal}$ he meets$_{subevent}$ his counselor.

[23] We refer to verbs like *postpone*, *put off*, *defer*, *adjourn*, and similar.

– TLINK="is_included";

TLINK = resign$_{mainevent}$ - Today$_{timex}$

– TLINK="is_included";

TLINK = meets$_{mainevent}$ - Today$_{timex}$

– TLINK="before";

TLINK = resign$_{mainevent}$ - meets$_{subevent}$ - before$_{signal}$

The rule described above does not apply if another temporal expression is present in the subordinated clause as in:

Today$_{timex}$ the president will resign$_{mainevent}$ before$_{signal}$ meeting$_{subevent}$, tomorrow$_{timex}$, his counselor.

– TLINK="is_included";

TLINK = resign$_{mainevent}$ - Today$_{timex}$

– TLINK="is_included";

TLINK = meeting$_{mainevent}$ - tomorrow$_{timex}$

– TLINK="before";

TLINK = resign$_{mainevent}$ - meeting$_{subevent}$ - before$_{signal}$

Notice: in case there are two event mentions, one verbal and one non-verbal, and a TIMEX which does not modify directly the non verbal event, the TLINKs are marked up between the verbal event mention and the temporal expression (e.g. *the crisis began in 1929*).

relType **Values for** TLINK **between Event Mentions and Timexes in the same sentence:** the following set of rules have been developed to improve annotators' agreement:

- in the case no signal is present, then the timex can identify either (i.) the temporal localizer of the event or (ii.) a textual temporal anchor. The temporal localizer (in a broad sense) of the event provides the answer to the question "when does/did the event happen?" or "how long does/did the event last?". The correct value is provided by the answer.

- the following TLINK values apply for constructions of the kind "EVENT MENTION + SIGNAL + TIMEX":

  - EVENT MENTION + for + DURATION$_{type}$ simple present or simple past verbal event mentions
    $\longrightarrow$ `relType=''measure''`;
    Mark ran$_{event}$ for 10 minutes$_{duration}$
  - EVENT MENTION + in/during + DURATION$_{type}$
    $\longrightarrow$ `relType=''is_included''`;
    In the last weeks$_{duration}$ many things have happened$_{event}$ here.
  - EVENT MENTION + in + quantified DURATION$_{type}$
    $\longrightarrow$ `relType=''after''`;
    In 5 minutes$_{duration}$ he made$_{event}$ it to the shop.
  - EVENT MENTION + for + DURATION$_{type}$ present perfect or past perfect verbal events
    $\longrightarrow$ `relType=''measure''` + an additional TLINK with `relType=''begun_by''` is created between the EVENT and the beginning point of the duration (it can be a timex of type DATE in the text or an empty timex tag);
    He has had$_{event}$ a strange look for 10 days$_{duration}$.
  - EVENT MENTION + since + DATE$_{type}$
    $\longrightarrow$ `relType=''begun_by''`;
    Since yestersay$_{date}$ she has been sleeping$_{event}$ on the couch.
    Since yesterday$_{date}$ she has had$_{event}$ a strange look.
  - EVENT MENTION + in + DURATION$_{type}$
    $\longrightarrow$ `relType=''after''`;
    He will deliver$_{event}$ his work in three days$_{timex}$
  - EVENT MENTION + within + DURATION$_{type}$
    $\longrightarrow$ NO TLINK; a TLINK with `relType=''ended_by''` is created between the EVENT and the ending point of the duration (it can be a timex of type DATE in the text or an empty timex tag);
    He will deliver$_{event}$ his work within three days$_{timex}$.

- EVENT MENTION + by + DATE$_{type}$
  $\longrightarrow$ `relType=''ended_by''`;
  He will deliver$_{event}$ his work by tomorrow$_{timex}$.
- EVENT MENTION + by + TIME$_{type}$
  $\longrightarrow$ `relType=''before''`;
  He will deliver$_{event}$ his work by 9.00$_{timex}$.

Please note that the lexical aspect of events can affect the annotation of TLINKs. For example, the difference between the two examples below is that in the second one the event mention *arrived* is a punctual/instantaneous event while *concert* in the first one is a durative event and *at 10* means that it *begins* at 10.

Concert$_{event}$ on Wednesday$_{timex}$ at$_{signal}$ 10$_{timex}$
$\longrightarrow$ `relType=''BEGUN_BY''`;

I arrived$_{event}$ at$_{signal}$ 10$_{timex}$
$\longrightarrow$ `relType=''SIMULTANEOUS''`;

Subtask 5: TLINKs between Timexes

TLINKs can be established also between two temporal expressions, typically when two timexes in the same sentence are connected by a SIGNAL.

He worked as a financial advisor for 3 months$_{timex}$ in$_{signal}$ 2010$_{timex}$
$\longrightarrow$ `relType=''IS_INCLUDED''` between *3 months* and *2010*, having *in* as SIGNAL.

11 APPENDIX A - CAT Annotation Task for NewsReader

This appendix presents the CAT annotation task following the guidelines described in this document. The task is written in XML format and can be simply imported in the tool.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<task name="NewsReader">
<markables>
<markable color="#f7f74a" name="EVENT_MENTION">
<attributes>
<attribute default_value="" name="pred" type="textbox"/>
<attribute default_value="" name="pos" type="combobox">
<value value="VERB"/>
<value value="NOUN"/>
<value value="OTHER"/>
</attribute>
<attribute default_value="FACTUAL" name="factuality"
type="combobox">
<value value="FACTUAL"/>
<value value="NON-FACTUAL"/>
<value value="COUNTERFACTUAL"/>
</attribute>
<attribute default_value="CERTAIN" name="certainty"
type="combobox">
<value value="CERTAIN"/>
<value value="UNCERTAIN"/>
</attribute>
<attribute default_value="" name="tense" type="combobox">
<value value="PRESENT"/>
<value value="PAST"/>
<value value="FUTURE"/>
<value value="NONE"/>
<value value="INFINITIVE"/>
<value value="PRESPART"/>
<value value="PASTPART"/>
</attribute>
<attribute default_value="" name="aspect" type="combobox">
<value value="PROGRESSIVE"/>
<value value="PERFECTIVE"/>
<value value="NONE"/>
<value value="PERFECTIVE_PROGRESSIVE"/>
```

```
</attribute>
<attribute default_value="" name="polarity" type="combobox">
<value value="POS"/>
<value value="NEG"/>
</attribute>
<attribute default_value="" name="modality" type="textbox"/>
<attribute default_value="" name="comment" type="textbox"/>
</attributes>
</markable>
<markable color="#6b6b6b" name="ENTITY">
<attributes>
<attribute default_value="PER" name="ent_type" type="combobox">
<value value="PER"/>
<value value="LOC"/>
<value value="PRO"/>
<value value="ORG"/>
<value value="FIN"/>
<value value="MIX"/>
</attribute>
<attribute default_value="" name="ent_class" type="combobox">
<value value="SPC"/>
<value value="USP"/>
<value value="NEG"/>
<value value="GEN"/>
</attribute>
<attribute default_value="" name="external_ref" type="textbox"/>
<attribute default_value="" name="comment" type="textbox"/>
</attributes>
</markable>
<markable color="#c3caf7" name="ENTITY_MENTION">
<attributes>
<attribute default_value="" name="head" type="textbox"/>
<attribute default_value="" name="syntactic_type" type="combobox">
<value value="NAM"/>
<value value="NOM"/>
<value value="PRO"/>
<value value="PTV"/>
<value value="CONJ"/>
<value value="PRE.NAM"/>
<value value="HLS"/>
<value value="APP"/>
<value value="ARC"/>
<value value="PRE.NOM"/>
```

```
</attribute>
<attribute default_value="" name="comment" type="textbox"/>
</attributes>
</markable>
<markable color="#33e2e8" name="TIMEX3">
<attributes>
<attribute default_value="DATE" name="type" type="combobox">
<value value="DATE"/>
<value value="TIME"/>
<value value="DURATION"/>
<value value="SET"/>
</attribute>
<attribute default_value="" name="value" type="textbox"/>
<attribute default_value="NONE" name="functionInDocument"
type="combobox">
<value value="NONE"/>
<value value="CREATION_TIME"/>
</attribute>
<attribute default_value="" name="anchorTimeID"
type="referenceLink"/>
<attribute default_value="" name="beginPoint" type="referenceLink"/
    >
<attribute default_value="" name="endPoint"
type="referenceLink"/>
<attribute default_value="" name="comment" type="textbox"/>
</attributes>
</markable>
<markable color="#09f215" name="SIGNAL">
<attributes>
<attribute default_value="" name="comment" type="textbox"/>
</attributes>
</markable>
<markable color="#ef42ff" name="C-SIGNAL">
<attributes>
<attribute default_value="" name="comment" type="textbox"/>
</attributes>
</markable>
<markable color="#e31010" name="EVENT">
<attributes>
<attribute default_value="OTHER" name="class" type="combobox">
<value value="SPEECH_COGNITIVE"/>
<value value="OTHER"/>
<value value="GRAMMATICAL"/>
```

```
<value value="MIX"/>
</attribute>
<attribute default_value="" name="external_ref" type="textbox"/>
<attribute default_value="" name="comment" type="textbox"/>
</attributes>
</markable>
<markable color="#d1bed1" name="VALUE">
<attributes>
<attribute default_value="PERCENT" name="type" type="combobox">
<value value="PERCENT"/>
<value value="MONEY"/>
<value value="QUANTITY"/>
</attribute>
<attribute default_value="" name="comment" type="textbox"/>
</attributes>
</markable>
</markables>
<relations>
<relation cardinality="one_to_one" color="#cbcbcb"
direction="false" name="HAS_PARTICIPANT">
<attributes>
<attribute default_value="PROPBANK" name="sem_role_framework"
type="combobox">
<value value="PROPBANK"/>
<value value="FRAMENET"/>
<value value="KYOTO"/>
</attribute>
<attribute default_value="Arg0" name="sem_role" type="combobox">
<value value="Arg0"/>
<value value="Arg1"/>
<value value="Arg2"/>
<value value="Arg3"/>
<value value="Arg4"/>
<value value="Argm-LOC"/>
<value value="Argm-OTHER"/>
</attribute>
<attribute default_value="" name="comment" type="textbox"/>
</attributes>
</relation>
<relation cardinality="one_to_one" color="#cbcbcb"
direction="false" name="TLINK">
<attributes>
<attribute default_value="BEFORE" name="reltype" type="combobox">
```

```
<value value="BEFORE"/>
<value value="AFTER"/>
<value value="IBEFORE"/>
<value value="IAFTER"/>
<value value="INCLUDES"/>
<value value="IS_INCLUDED"/>
<value value="MEASURE"/>
<value value="SIMULTANEOUS"/>
<value value="BEGINS"/>
<value value="BEGUN_BY"/>
<value value="ENDS"/>
<value value="ENDED_BY"/>
<value value="IDENTITY"/>
</attribute>
<attribute default_value="" name="signalID" type="referenceLink"/>
<attribute default_value="" name="comment" type="textbox"/>
</attributes>
</relation>
<relation cardinality="one_to_one" color="#080608" direction="false
    " name="CLINK">
<attributes>
<attribute default_value="" name="c-signalID"
type="referenceLink"/>
<attribute default_value="" name="comment" type="textbox"/>
</attributes>
</relation>
<relation cardinality="many_to_one" color="#db1463" direction="true
    " name="REFERS_TO">
<attributes>
<attribute default_value="" name="comment" type="textbox"/>
</attributes>
</relation>
<relation cardinality="one_to_one" color="#cbcbcb"
direction="false" name="SLINK">
<attributes>
<attribute default_value="" name="comment" type="textbox"/>
</attributes>
</relation>
<relation cardinality="one_to_one" color="#157528" direction="false
    " name="GLINK">
<attributes>
<attribute default_value="" name="comment" type="textbox"/>
</attributes>
```

```
</relation>
</relations>
</task>
```

12 APPENDIX B - UML diagram of the annotation scheme

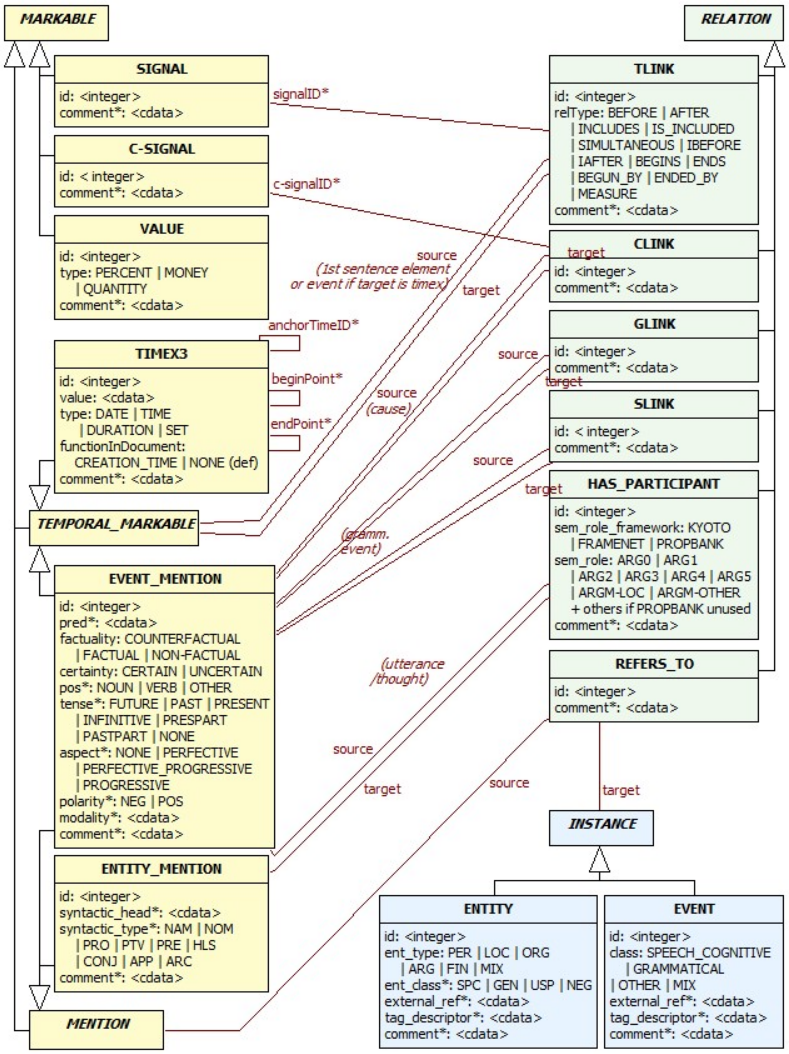

Figure 2: UML diagram of the annotation scheme. Properties and relations are encoded as XML attributes, apart 'source' and 'target' that are elements; '*' denotes optionality.

References

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

[Bonial *et al.*, 2010] Claire Bonial, Olga Babko-Malaya, Jinho D. Choi, Jena Hwang, and Martha Palmer. Propbank annotation guidelines, version 3.0. Technical report, Center for Computational Language and Education Research, Institute of Cognitive Science, University of Colorado at Boulder, Pisa, Italy, 2010. `http://clear.colorado.edu/compsem/documents/propbank_guidelines.pdf`.

[Ferro *et al.*, 2005] Lisa Ferro, Laurie Gerber, Inderjeet Mani, Beth Sundheim, and George Wilson. Tides 2005 standard for the annotation of temporal expressions, September 2005. `http://fofoca.mitre.org/annotation_guidelines/2005_timex2_standard_v1.1.pdf`.

[Hovy *et al.*, 2013] Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. Events are not simple: Identity, non-identity, and quasi-identity. *NAACL HLT 2013*, page 21, 2013.

[Huang, 2012] Li-szu Agnes Huang. The effectiveness of a corpus-based instruction in deepening efl learners' knowledge of periphrastic causatives. *TESOL Journal*, 6:83–108, 2012.

[ISO TimeML Working Group, 2008] 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`.

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

[Wolff *et al.*, 2005] Phillip Wolff, Bianca Klettke, Tatyana Ventura, and Grace Song. Expressing causation in english and other languages. *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin*, pages 29–48, 2005.