Entity Detection and Tracking – Phase 1
ACE Pilot Study Task Definition

1 INTRODUCTION

The objective of the ACE program is to develop automatic content extraction technology to support automatic processing of source language data (in the form of natural text, and as text derived from ASR and OCR). This includes classification, filtering, and selection based on the language content of the source data, i.e., based on the meaning conveyed by the data. Thus the ACE program requires the development of technologies that automatically detect and characterize this meaning.

Ultimately, ACE applications will maintain a database of what is happening in the world. Ideally, this will be in terms of who is doing what, where, and when. As information from source language data is accumulated over time, the database will be updated and maintained. In this way the database becomes a vehicle for tracking the information we are interested in. The database should also maintain pointers into the source data so as to enable more detailed examination of the information represented in the database.

The ACE research objectives are viewed as the detection and characterization of *Entities*, *Relations* and *Events*. The ACE pilot study begins the technology R&D effort by focusing on entity detection in phase 1. This task is being defined so as to support applications as well as to provide a basis for further development in extracting relations and events.

The Entity Detection task requires that selected types of entities mentioned in the source data be detected and that selected attributes of these entities be extracted and merged into a unified representation for each entity. Tracking of entities across document boundaries will be deferred until after the initial phase.

2 TASK DEFINITION

The Entity Detection task is actually a complex of different tasks:

1. ***The detection of entities.*** This is the most basic common task. It is the foundation upon which the other tasks are built, and it is therefore a required task for all ACE technology developers. Recognition of entity type or entity attributes is not part of this task. Note, however, that detection is limited to entities of the specified types.
2. ***The recognition of entity attributes***. This is the basic task of characterizing entities. It includes recognition of entity type. It is a required task for all ACE technology developers. *Performance is measured only for those entities that are correctly detected.* Evaluation of performance will be conditioned on entity and attribute type.
3. ***The detection of entity mentions***. This is an optional task that measures a system's ability to correctly identify the set of mentions that underlie a particular entity, for all entities of the specified types. *Performance is measured only for those entities that are correctly detected.* Evaluation of performance will be conditioned on entity and mention type.
4. ***The recognition of mention extent***. This is an optional task that measures a system's ability to correctly determine the extent of entity mentions. *Performance is measured only for those mentions that are correctly detected.*

In order to be most helpful to research, and to facilitate research progress, these subtasks will be supported separately with performance metrics and evaluation procedures.

The source data will be divided into documents. Since information about mentioned entities will be distributed throughout each document, the entity detection system is required to output information on detected entities for each document as a whole.

Entity output will comprise the following information:

1. Pointers to one or more mentions of the entity. (Only one pointer is required. Pointers to all mentions are required only for the mention detection task.)
2. Entity type and attribute information.
3. Mention extent, in terms of pointers to the beginning and end of each mention (optional – for evaluation of mention extent only).

The type of pointer will depend on the type of source data. For text, the pointer will be a character offset from the beginning of the document. For ASR data, the pointer will be a time offset. For OCR data, the pointer will be a spatial (x,y) offset. (ASR and OCR source data will provide beg/end offsets for each word.)

2.1 ENTITY RESTRICTIONS

Entities to be detected and recognized will be limited to the following types:

1. ***Person***. Person entities are limited to humans. A person may be a single individual or a group if the group has a group identity.
2. ***Organization***. Organization entities are limited to corporations, agencies, governments, and other groups of people defined by an established organizational structure.
3. ***Facility***. Facility entities are limited to buildings and other permanent man-made structure and real estate improvements.
4. ***GSP*** *(A Geographical-Social-Political Entity)*. GSP entities are geographical regions defined by political and/or social groups. A GSP entity subsumes and does not distinguish between a region, its government or its people.
5. ***Location***. Location entities are limited to geographic entities such as geographical areas and landmasses, bodies of water, and geological formations.

Entities may have only one type. Note that it often happens that different entities may be referred to by the same name. Despite this superficial similarity, however, such entities are separate and distinct for the purposes of the ACE program. For example, in the sentence "*Visitors streamed through the White House*", White House is a facility entity, whereas in the sentence "*The White House vetoed the bill*", White House is an organization entity and is distinct from the White House facility entity.

Entities may be either singular or plural, but in order to qualify for detection and recognition, an entity must be specific and not generic. This distinction is often not easy to make. In cases of doubt the decision should be to declare the entity generic and thus exclude it from consideration in the ACE tasks.

There are no limits on the use of inference and world knowledge in determining either the entity type or the entity attributes. The determination should be based on the best judgement of the author's or speaker's intention.

2.2 Entity Attributes

For purposes of the current phase 1 tasks, entity type and entity name are the only attributes to be considered. All proper name mentions of the entity are to be output. Representation and output of the names is in terms of a pair of pointers to the beginning and end of the name, for each name.

2.3 Mention Restrictions

There are no restrictions on how an entity is mentioned. Mentions include both direct and indirect references to the entity and descriptions of the entity. Representation and output of mentions is in terms of a pointer to the head of the mention. If the head comprises multiple words, the pointer should point to the last word in the head. However, if the head is a proper name the pointer may point to any word in the name.

Candidates for mentions are not strictly defined by syntactic or grammatical category. However, mentions are not anticipated in structures other than noun phrases or adjective phrases.

2.4 Mention Extent

Representation and output of mention extent is in terms of a pair of pointers to the beginning/end of the mention.

Detailed guidelines for determining entity types, for determining mentions of them, and for annotating these entities and associated mentions, are given in section 6.

3 Evaluation

The Entity Detection tasks comprise two required tasks and two additional optional tasks. Evaluation will be performed separately for each of the tasks.

3.1 Entity Detection

Entity Detection performance will be measured in terms of missed entities and spurious entities ("false alarms"). Multiple representations of the same entity are considered spurious entities. In order to measure misses and false alarms, each reference entity must first be associated with the appropriate corresponding system output entity. This is done by choosing, for each reference entity, that system output entity with the greatest number of corresponding mentions. Note, however, that a system output entity is permitted to map to at most one reference entity. A miss occurs whenever a reference entity has no corresponding output entity. A false alarm occurs whenever an output entity has no corresponding reference entity.

3.2 Attribute Recognition

Attribute recognition measures the ability of the system to correctly classify the attribute values of the entities, *for all correctly detected entities*. This ability will be measured in terms of classification error rate, which is simply the fraction of reference attributes with values that are not identical to those of the corresponding output attributes.

3.3 Mention Detection

Mention detection measures the ability of the system to correctly detect and associate all of the mentions of an entity, *for all correctly detected entities*. Detection performance will be measured in terms of missed mentions and spurious mentions ("false alarms"). For each mapped reference entity: a miss occurs for each reference mention of that entity without a matching mention in the corresponding output entity, and a false alarm occurs for each mention in the corresponding output entity without a matching reference mention.

3.4 Extent Recognition

Extent recognition measures the ability of the system to correctly determine the extent of the mentions, *for all correctly detected mentions*. This ability will be measured in terms of the classification error rate, which is simply the fraction of all mapped reference mentions that have extents that are not identical to the extents of the corresponding system output mentions.

4 The Pilot Study Corpus

The reference corpus will comprise stories taken from three different sources, namely from newswires, newspapers, and broadcast news programs. Evaluation will be performed on two different versions of newspaper and broadcast news sources, namely both manually and automatically transcribed versions (OCR for newspaper and ASR for broadcast news). The corpus will be made up of the following sources:

	Training 01-02/98	Dev Test 03-04/98	Eval Test 05-06/98
Newswire	30,000 words	15,000 words	15,000 words
Broadcast News	30,000 words	15,000 words	15,000 words
Newspaper	30,000 words	15,000 words	15,000 words

In addition to this primary corpus, a "mini-corpus" will be created to support preliminary task definition and development. This corpus will be taken from broadcast news and newswire, in equal proportion by word, and will total about 10,000 words. The mini-corpus will be taken from the training portion of the pilot study corpus. All participating sites are required to annotate this mini-corpus.

5 Future Directions

The entity detection task outlined here is intended only as an initial R&D direction. There are numerous ACE problem dimensions into which research needs to be extended after the pilot study. These include:

- The addition of more types of entities and more entity attributes.
- The extension of entity detection to tracking across document boundaries.
- The detection and characterization of relations among different entities.
- The detection and characterization of events (which may be viewed as patterned complexes of relations and entities).

6 ENTITY ANNOTATION GUIDELINES

(BY RALPH GRISHMAN)

These guidelines describe how to annotate a document to record the entities mentioned in a document. An entity is

- a person
- an organization
- a facility
- a geographical/social/political (GSP) entity
- a location

or a set of people, organizations, facilities, GSP entities, or locations. A mention of an entity is an explicit reference to the entity in the document. Every entity must have at least one mention. The entity must be specific, rather than generic, in nature.

For each entity, the annotation records the type of the entity, the names of the entity (if any), and the mentions of the entity in the text.

For the present, we make reference to sections in Lisa Ferro's nominal entity recognition task definition (ver. 1.0), which has addressed some of the annotation issues in more detail than we are able to in this initial release. In the future, we may incorporate sections of those guidelines herein.

6.1 TEXT TO ANNOTATE

Only the material between <TEXT> and </TEXT> tags is to be annotated. In newswire documents, material in headlines and slug sections is not to be tagged. In broadcast news, only the transcribed speech is to be tagged; added information, such as that within <TURN> tags, is not to be annotated.

6.2 TYPES OF ENTITIES

6.2.1 PERSONS

Each distinct person or set of people mentioned in a document becomes an entity of type person. People may be specified by name ("*John Smith*"), occupation ("*the butcher*"), family relation ("*dad*"), 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, etc. Groups of people ("*the family*", "*the house painters*", "*the linguists under the table*") are to be considered an entity of type person unless the group meets the requirements of an organization, described below.

6.2.2 ORGANIZATIONS

Each organization or set of organizations mentioned in a document gives rise to an entity of type organization. An organization must have some formally established association. This includes business units, government units, sports teams, and formally organized musical groups. In addition, industrial sectors are treated as organizations (following the nominal entity task definition, section 2.5).

A set of people who are not formally organized into a unit are to be treated as a person entity rather than an organization entity (see the nominal entity task definition, section 2.5).

An organization name may sometimes be used to refer to the members of the organization in aggregate ("*SRI defeated BBN in softball*") or the buildings housing that organization ("*SRI was destroyed by the 2003 earthquake*.") These concepts are subsumed by the organization entity. Thus, in each of these examples "*SRI*" should be considered a mention of (the same) entity of type organization.

6.2.3 GEOGRAPHICAL/SOCIAL/POLITICAL (GSP) ENTITIES

The name of a geographical region which is defined on a political basis, such as "*France*" or "*Boston*", can refer to a range of concepts. It can refer to the physical region itself ("*France has an area of xxx square miles*.", "*France enjoys a temperate climate.*"), the government ("*France signed a treaty with Germany last week.*"), the people ("*France elected a new president.*"), or some other aggregate within this region ("*France produces better wine than New Jersey.*", "*France has a gross domestic product of xxx francs.*") Because it can be difficult to differentiate these concepts (and, indeed, it is not clear if readers and writers always do so), we define a geographical/social/political (GSP) entity which subsumes these concepts. A GSP entity is a geographical region which may also be used to refer to the government and/or populace of that region. Typical GSP entities are countries, states, and cities. Geographical regions which are not associated with a government may be locations (see the discussion of locations, below).

Explicit references to the government of a country (state, city, etc.) are to be treated as references to the same entity evoked by the name of the country. 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 it 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*").

Sometimes the names of GSP entities may be used to refer to other things associated with a region besides the government, people, or aggregate contents of the region. The most common examples are sports teams: "*New York defeated Boston 99-97 in overtime.*" These are to be recorded as distinct entities, not as mentions of the GSP entity. Thus, in this example, both "*New York*" and "*Boston*" would evoke organization entities.

6.2.3.1 NESTED REGION NAMES

A series of nested region names, such as "*Provo, Utah*" evokes one entity for each region. Thus "*Provo, Utah*" evokes one entity for the city (with mention "*Provo, Utah*") and a second one for the state (with mention "*Utah*"). "*Washington, D.C.*" is to be treated as two (nested) region names.

6.2.4 LOCATIONS

Locations defined on a geographical or astronomical basis which are mentioned in a document and do not constitute a political entity give rise to location entities. These include, for example, the solar system, Mars, the continents, the Mideast, the Hudson River, Mt. Everest, and Death Valley.

In general, terrestrial locations must have some two-dimensional extent. Abstract coordinates ("31° S, 22° W") and positions relative to a GSP or location ("30 miles east of Mount Fuji") are not themselves entities. Borders, considered as (one-dimensional) boundaries between two regions, are not entities. Positions distinguished *only* by the occurrence of an event at that position ("the scene of the murder", "the site of the rocket launching") are not entities.

6.2.4.1 SUB-PARTS OF LOCATIONS AND GSPS

Portions of GSP entities or location entities, such as "*the center of the city*", "*the outskirts of the city*", or "*the southern half of New Jersey*" constitute location entities in their own right (in accord with the nominal entity guidelines, section 2.4.1).

Note that location entities may also refer to the population of a region, or other aggregates within that region:

The Deep South *voted for Bush.*

Southern France *drinks more wine than Boston.*

6.2.5 FACILITIES

A facility is a large, functional, primarily man-made structure. These include buildings and similar facilities designed for human habitation, such as houses, factories, stadiums, office buildings, gymnasiums, prisons, museums, and space stations; objects of similar size designed for storage, such as barns, parking garages and airplane hangars; elements of transportation infrastructure, including streets, highways, airports, ports, train stations, bridges, and tunnels. Roughly speaking, facilities are artifacts falling under the domains of architecture and civil engineering.

Individual rooms of buildings are facilities, but other portions of buildings, such as walls, windows, closets, or doors, are not facilities.

In some cases, a facility name is used to refer to an organization (which, typically, operates the facility) or a set of people (the people employed by that organization). These concepts are subsumed by the facility entity. Thus, all of the following instances of "*museum*" (assuming they refer to the same museum) are treated as mentions of a single entity of type facility:

The museum is located on Fifth Avenue.

I walked into the museum.

Mary works for the museum.

The museum insisted that the exhibition was not obscene.

The museum received a gift of $100,000.

6.2.6 DISTINGUISHING TYPES

There are many examples of *systematic metonymy*, some of which have been mentioned above, in which one 'base concept' can refer to a variety of things associated with that concept. For example, an organization name can also refer to the buildings or people of that organization; a country name can also refer to the people or government of that country. In such cases, the EDT annotation seeks to tie all these referents together by linking them to a single entity.

[We may in the future wish to enrich this annotation by tagging each *mention* of the entity by the type of the referent, indicating that some mentions of a country name refer to the government, others to the geographical region, others to the people, and still others to other features of the country.]

Because of this metonymy, it may be difficult to decide on the type of a particular entity. In particular, problems are likely to arise in choosing between people and organizations; between GSPs and locations; and between facilities and organizations. The choice between an organization and a set of people is based on whether the unit is a formally established association (not something which is always easy to determine). The choice between a GSP and a location is based on whether the entity has a government (GSPs have governments; locations do not).

Decisions between facilities and organizations are based, first of all, on the lists of types of entities deemed to be facilities, given at the beginning of the section on facilities. Churches, schools (including universities), embassies, and restaurants are deemed to be organizations. The White House is an organization. These lists will grow in the future as more material is tagged. For items not on the list, decisions can be made based on similarity to listed items. Decisions can also based on whether the "predominant use" of the concept (in general, not in a particular document) is as a reference to the organization or the buildings.

6.3 PLURAL AND CONJOINED ELEMENTS

6.3.1 CONJOINED ELEMENTS

For conjoined elements, we follow the rule adopted for nominal entities (section 3.4): one entity per head noun (heads are discussed further below). Thus "*I met two accountants and three linguists.*" gives rise to two entities, "*two accountants*" and "*three linguists*". (It does *not* give rise to a third entity corresponding to the conjoined phrase.) If there is only a single head, with conjoined modifiers, there is only one entity: "*I know several descriptive and generative linguists.*" In cases of conjoined names, we record two entities: "*Jimmy Carter and Rosalynn Carter*" gives rise to two entities, "*Jimmy Carter*" and "*Rosalynn Carter*". This is the case even if the names share a single last-name token. Thus "*Jimmy and Rosalynn Carter*" gives rise to two entities, the first with mention "*Jimmy*" and the second with mention "*Rosalynn Carter*"; if there is a later reference to "*Jimmy Carter*", it is recorded as another mention of the first entity.

6.3.2 PLURALS

A plural can be an entity:

the injured passengers

Two distinct sets produce separate entities, regardless of whether they have elements in common; so, for example,

ten passengers were injured, six seriously

evokes two entities, one for the ten passengers, one for the six. Distinct sets produce separate entities, even if they have the same string, so

five people like vanilla, five people like chocolate

evokes two entities (the five people who like vanilla and the five who like chocolate). Furthermore, a set is a distinct entity from each of its members;

Fred Smith married Harriet Hope; they lived happily for 6 weeks.

evokes three entities, one for Fred Smith, one for Harriet Hope, and one for the set with members Fred and Harriet. The only mention of the set is the pronoun "they".

6.4 NAMES

For each entity, we record the occurrences of names (if any) used to refer to this entity in the document. We follow the rules for named entities in determining the extent of names. Specifically, honorifics and titles are not considered part of a name (for example, in "*President Fred Smith*" and "*Mr. Fred Smith*", the name is "*Fred Smith*"). Note that we record each occurrence of a name, not just the name itself: if a name appears twice, both instances must be recorded.

Names are atomic ... entity names wholly contained within another name are not annotated. Thus "*The New York Times*" does not evoke a separate entity for the location "*New York*".

6.5 MENTIONS

For each entity, we record all mentions of the entity. Mentions are nominal or prenominal phrases that refer to or describe the entity. For each mention, we record its head and its full extent. An occurrence of a bare name (with no title or modifiers), such as "*Fred Smith*" in "*Fred Smith died.*", will be recorded both as an instance of a name for the entity and as a mention of the entity.

Mentions will frequently be nested; that is, they will contain mentions of other entities. For example, the phrase "*the president of Ford*" is a mention of an entity of type person, and contains the name "*Ford*", a mention of an entity of type organization. It is even possible for a noun phrase to contain an embedded mention of the same entity. For instance, the phrase

the historian who taught herself COBOL

evokes a person entity with two mentions, the entire phrase and the word "*herself*".

Within the left modifiers of a noun, only possessive modifiers (nouns and pronouns) and proper nouns and adjectives can be mentions. Common nouns appearing as the left modifier of other nouns are not mentions. In the examples below, the embedded mentions are shown in boldface:

Mary*'s sister*

his *brother*

my brother*'s keeper*

State Department *spokesman Bob Bumble*

French *movie star LeBon Mot*

the museum guards

state legislature committee members

Relative pronouns do not constitute mentions: the phrase "*the man who came to dinner*" is a mention of an entity of type person, but "*who*" is not a separate mention of that entity.

6.5.1 TYPES OF MENTIONS

We may expect different performance in determining referents and types for pronouns, for names, and for other noun phrases. In order to be able to identify such differences, we distinguish between mentions with a named head ("name-mentions"), those with a noun head ("nominal or nom-mentions") and those with a pronominal head ("pro-mentions"). Mentions with empty heads ("*five of the analysts*") are classified as pro-mentions.

6.5.2 PREDICATE COMPLEMENTS

The mentions should include nominal predicate complements that are affirmatively asserted of a reportable entity, since they describe the entity. Thus

Fred is a real linguist.

evokes an entity of type person with two mentions, "*Fred*" and "*a real linguist*". (Thus, the question of whether the usage is "generic", as discussed below, does not arise in this context.) On the other hand,

Fred is not a real linguist.

evokes an entity of type person with only one mention, "*Fred*". Similarly,

Fred is studying to be a real linguist.

evokes an entity of type person with only one mention, "*Fred*", because the text does not assert that Fred has been, is, or will be a real linguist.

6.5.3 APPOSITION

Appositional modifiers are treated like predicate complements: they are recorded as mentions of the head, without regard to the criteria regarding generic usage. Thus the phrase

Fred, a real linguist, knows ten languages, none fluently

evokes an entity with mentions "*Fred, a real linguist*", and "*a real linguist*".

A decision about whether a phrase is an instance of apposition may depend on subtle clues, such as the presence or absence of determiners, particularly in speech transcripts where (comma) punctuation is absent or unreliable. For example, in a transcript the phrase

the State Department spokesman Uno Little

would be considered an example of apposition, since a name rarely takes the determiner "*the*". It would normally be punctuated

the State Department spokesman, Uno Little,

It would evoke a person entity with two mentions, a nominal mention ("*the State Department spokesman Uno Little*"), with head "*spokesman*", and a name mention ("*Uno Little*"). In contrast,

State Department spokesman Uno Little

would *not* be an example of apposition, since "*spokesman*" normally cannot be the head of a noun phrase without a determiner. In consequence, it would be just a single (name) mention, "*State Department spokesman Uno Little*".

6.5.4 PROPER ADJECTIVES

A proper adjective is to be treated as a mention of the noun from which it is derived. Thus, if "*France*" and "*French*" both appear in a single document, they are to be marked as mentions of the same GSP entity (if only "*French*" appears in a document, it evokes a GSP entity). The adjective is marked as a name mention of the GSP entity.

A noun indicating a national of a given country, such as "*Frenchman*" is a nominal mention of an entity of type person. In many cases — "*Iranian*", "*American*", "*German*", etc. — the same word is used both as a proper adjective and as the name of a national. When used as an adjective ("*I love*

Iranian caviar for breakfast."), it is marked as a name mention of the GSP entity; when used as a noun ("*I met three Iranians.*"), it is marked as a nominal mention of a person entity.

Similar rules apply to adjectives derived from names of organizations. Thus, "*Republican*" in "*Republican platform*" is a name mention of an organization entity, while in "*That Republican likes macaroni and cheese."* it is a nominal mention of a person entity.

6.5.5 QUANTIFIED AND PARTITIVE PHRASES

A partitive construction of the form

quantifier of *ENP*

gives rise to two mentions: one for the entire phrase, and one for the embedded noun phrase *ENP* which is the object of "of". If the entire phrase represents a subset of *ENP*, these will be mentions of distinct entities. Thus in

three of the women

evokes two entities, for "*the women*" and "*three of the women*". Similarly,

some of the women

evokes two entities. On the other hand,

all of the women

has two mentions of *one* entity: "*the women*" and "*all of the women*" (the same set). This is also the case with the partitive-like phrase

a team of five experts

since the team is identical to the set of five experts.

6.5.6 EXTENTS

The extent of a mention consists of the entire nominal phrase. In case of structures where there is some unresolvable 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 is "*some people yesterday who love chess*".

The extent should include all the modifiers of a nominal phrase, including prepositional phrases, relative clauses, appositional phrases, etc. (This is different from the nominal entity guidelines, where appositional modifiers are excluded from the extent of the main noun phrase.) Thus the phrase "*Fred Smith, the noted general*" constitutes two mentions of one entity, "*Fred Smith, the noted general*" and "*the noted general*"; similarly, "*Fred Smith, who is a noted general*" constitutes two mentions, "*Fred Smith, who is a noted general*" and "*a noted general*". Titles, honorifics, and determiners are all treated as modifiers, and are included in the extent.

Tokenization rules follow those used for MUC. Generally speaking, tokens are broken at white space, and each item of punctuation is treated as a separate character. In addition, possessive endings ('s) are treated as separate tokens, and contractions are split (so that "*we're*" becomes the two tokens "*we*" and "*'re*"). Extents must begin at the beginning of a token and end at the end of a token.

6.5.7 HEADS

In addition to marking the entire nominal phrase (the extent), the head of the phrase must be marked. In

The hurricane destroyed the new glass-clad skyscraper.

the mention is "*the new glass-clad skyscraper*" and the head is "*skyscraper*". Except for proper nouns and adjectives, the head is always a single token. If the syntactic head of the phrase is a multi-token item, the last token is marked. 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 bolded and the head is underlined:

<u>Fred Smith</u> *became* ***the new prime <u>minister</u>****.*

The job fell to ***<u>Abraham Abercrombie III</u>****.*

If the phrase is "headless", as in the case of a partitive construction, the last modifier of the empty head is to be marked:

a course in linguistics for ***the <u>young</u>*** *and* ***the <u>restless</u>***

he was introduced to ***<u>five</u> of the analysts***

Note that in the last example, there is a second entity, whose full mention is "*the analysts*" and whose head is "*analysts*".

6.6 GENERIC / SPECIFIC DISTINCTION

Only specific entities should be tagged; generic entities are not marked. Making a consistent distinction in this regard will be difficult, but we give some initial guidelines here, subject to revision and elaboration.

Generic noun phrases refer to a type of thing or a property rather than an actual entity or set of entities. Even if the property or the set is extremely constrained so that there are very few possible members, it should still be considered a generic. Although determiners and combinations of determiners are often clear indicators of generic status ("*any*" ⇒ generic, "*all the*" ⇒ non-generic, "*all five*" ⇒ non-generic), there are many difficult cases. "All" usually indicates a generic noun phrase, but may also appear with non-generic phrases, such as "*all current members of Congress*". Bare plural noun phrases and singular noun phrases with the determiner "*a*" can either refer to a specific, but unspecified entity or set of entities or they can be generic noun phrases. Most occurrences of hard-to-identify generics fall into one of these last two patterns.

Here are some general tests for generics:

1. You should be able to substitute a phrase of the form "*kind of X*" for a generic noun phrase and retain the same meaning:

   I like books = I like the kind of thing called a book.

2. Generics need not actually exist and the statement can still be true:

   I like books about gorillas wearing hats.

   They gave a tax break to consumers under 1 year old.

Generic noun phrases of the type "*a*" + singular noun or bare plurals can be distinguished using tests such as:

1. These noun phrases in negated contexts are generic:

   I didn't see gorillas (a gorilla) here. [generic]

   I saw gorillas (a gorilla) here. [non-generic]

2. These noun phrases in “belief” contexts are generic:

   I want to see gorillas (a gorilla).

I thought I heard a gorilla.

3. These noun phrases in questions are generic:

   Have you seen a gorilla walking by?

   Have you seen gorillas wearing hats?

4. Certain predicates cause bare plural noun phrases to have generic interpretations (these tests do not apply to "*a*" + singular noun). These include

   A. Some quantification predicates

   Gorillas are everywhere.

   B. “kind of” predicates

   Gorillas are rare / common / widespread.

   Gorillas are in short supply.

   Gorillas are indigenous.

   Gorillas are in short supply.

   Gorillas come in many sizes.

   C. Simple present tense predicates

   People with funny hats run fast.

   D. Predicates modified by X times or other adverbials of repetition.

   Gorillas visited me five times today.

   E. Bare plurals with individual-level predicates are generic. Individual-level predicates mark characteristics of individual members of a set, e.g., "*birds have wings*" means that each bird has wings. In contrast, stage-level predicates ("*Gorillas are wrecking my garden*", "*Gorillas are available*") can be either generic or non-generic, depending on context. Thus the subjects are generic in the following sentences:

   Gorillas are intelligent.

   Linguists know French.

   Birds have wings.

5. There are also cases in which noun phrases with "*the*" are generic, even though this is not typically the case. This is when "*the*" plus a singular noun is used to represent a set, e.g.,

   Turing invented the computer. [generic]

   I wrote this on the computer in my office. [non-generic]

   The dodo is extinct. [generic]

   The dodo is dead. [non-generic]

In some cases pronouns such as "*we*", "*you*", and "*they*" should be considered generic. If a pronoun can be replaced by the generic "*one*", its usage is generic and should not be marked:

This is what you might call a great malt liquor.

This is, you know, real hard to understand.

A discourse may well include both instances of "*you*" which are generic and others which are specific.