Entity Detection and Tracking – Phase 1
EDT and Metonymy Annotation Guidelines
Version 2.5 20021205

1 Intro

The objective of the ACE program is to develop automatic content extraction technology to support automatic processing of source language data. 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 ensure 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. ACE Phase 1 begins the technology R&D effort by focusing on entity detection. 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, their sense disambiguated, 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.

This document outlines the ACE Phase 1 annotation tasks (Entity Detection and Tracking, Metonymy Annotation, and Generic/Specific Classification). It is intended to integrate section 6 of the ACE Pilot Study Task Definition v 2.2, EDT Metonymy Annotation Guidelines v 2.4, and various addenda to both documents into up-to-date annotation guidelines. Please refer to NIST's ACE website (www.itl.nist.gov/iaui/894.01/tests/ace/index.htm) for the ACE task definition and evaluation plan.

2 Basic Concepts

An entity is an object or set of objects in the world. A mention is a reference to an entity. Entities may be referenced by their name, indicated by a common noun or noun phrase, or represented by a pronoun. For example, the following are several mentions of a single entity:

Name Mention: *Joe Smith*

Nominal Mention: *the guy wearing a blue shirt*

Pronoun Mentions*: he, him*

For Phase 1 of ACE, entities are limited to the following five types:

- Person - Person entities are limited to humans. A person may be a single individual or a group.
- Organization - Organization entities are limited to corporations, agencies, and other groups of people defined by an established organizational structure.
- Facility - Facility entities are limited to buildings and other permanent man-made structures and real estate improvements.
- Location - Location entities are limited to geographical entities such as geographical areas and landmasses, bodies of water, and geological formations.
- GPE (Geo-political Entity) - GPE entities are geographical regions defined by political and/or social groups. A GPE entity subsumes and does not distinguish between a nation, its region, its government, or its people.

We do not identify mentions of animals or most inanimate objects at this time.

For each entity, the annotation records the type of the entity (PER, ORG, GPE, LOC, or FAC), its class (Generic/Specific), all of the mentions of the entity from the text (Name, nominal, Pronoun), and the role of those mentions if applicable (see section 4.1.5.3 GPE Mention Roles).

3 Text to Annotate

Only 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 or speaker identification tags, is not to be tagged.

4 Entities and Mentions

4.1 Entity Types

4.1.1 Persons

Each distinct person or set of people mentioned in a document refers to 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, 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.

> *He is [a real <u>turkey</u>]*
>
> *[The political <u>cat</u> of the year]*

He was [one of the dark horses]

[The film star]

She's known as [the brain of the family]

[Californian transplants]

He is [a harmonic force]

4.1.1.1 Saints and other religious figures

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

St. Christopher, the patron of transportation

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. Note that capitalization information may not be available in speech transcripts.

If you believe in god, you must…	name mention
Although he felt like he was [a god], he…	nominal mention

4.1.1.2 Fictional characters, names of animals, and names of fictional animals

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.

Batman has become a popular icon

Adam West's 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. These two examples do not yield mentions.

Morris the cat

Snuggle, the fabric softener bear

4.1.1.3 Groups of people

Groups of people are to be considered an entity of type Person unless the group meets the requirements of an organization or a GPE described below.

The family

The house painters

The linguists under the table

4.1.1.3.1 Ethnic, Religious, and Political Groups

Ethnic groups, religious groups and political groups are often referenced by the name of the ethnicity, religion and political party, for example:

African-Americans

Catholics

Democrats

Those groups that have an organizing body are name mentions of the organization. If a mention refers to the members of an organization in general, we consider the mention to refer to the organization.

Democrats support social programs.

Catholics celebrate Lent every year.

Democrats is an organization name because it is used in a context describing the beliefs of the greater organization of the Democratic Party. When a mention refers to an individual person, as in

Mike is a Democrat

or to a small group of individuals, as in

Mike and Bob are both Democrats

the mention is a person nominal and is a mention of the same entity as the person to whom the phrase is attributed.

Ethnic groups do not generally have a formal organization associated with them. As a result, we mark these mentions as names of a person entity.

***{[PER-name]** Cuban Catholics**}** are expecting the Pontiff to preach about the value of religious freedom, something they're just beginning to experience.*

When ethnic designation is given to an individual person or a small group of individuals, the mention is marked as a nominal mention of that person entity.

*Joe is **{[PER-nominal]** a Cuban Catholic**}**.*

In this example, the mentions “Joe” and “a Cuban Catholic” refer to the same entity.

4.1.1.3.2 Family Names

Family names are to be tagged as Person.

The Kennedy’s

The Kennedy family

Please note that the second example contains two mentions of the same entity: one name mention and one nominal mention.

4.1.2 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 and a persistent, established existence. Typical examples are businesses, government units, sports teams, and formally organized music groups. Industrial sectors are also treated as organizations.

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 tagged as person entities. Example organization-like nouns which are *not* organizations are “family,” “employees,” and “crew.” In the latter two cases, although the members of a company or crew may work together in an organized and even hierarchical fashion, the groups are not organizations by themselves.

Some words like “team,” “delegation” and “police” achieve organizational status only in certain contexts. “[The home *team*] flies to Connecticut to meet the Huskies in Hartford” clearly refers to a named sports team and is thus taggable as an organization. However, the “[U.N. weapons inspection *team*]” is less permanent and cohesive, and is thus a person entity rather than an organization. The noun “police” is a person entity in contexts like “[*police*] outnumbered [*demonstrators*]” but an organization entity in “[*police* in East Timor] have arrested [two *men*].”

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.

4.1.2.1 Organization Entities used in Person Contexts

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 though 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 PER metonymy. It seems better to adopt the view that organizations can be agentive, and take action on their own. Only when something in the context draws particular attention to the people within the organization should a separate mention of a PER entity be marked.

4.1.2.2 First Person Pronouns Referring to Organizations

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 mentions, and not as PER mentions.

4.1.3 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 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 GPE 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.

4.1.3.1 Sub-parts of Locations and GPEs

Portions of GPE 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. When general locative phrases like “top,” “bottom,” “edge,” “periphery,” “center,” and “middle” are used to pinpoint a portion of a markable location, they are markable locations.

> *“They tend to live not in [the **center** of [the country]] but at [its **periphery**]”*

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.*

4.1.3.2 Non-Locations

It is easy to start interpreting all objects as locations. Every physical object implies a location because the space that each physical object occupies is the “location” of that object. In addition, our language is full of location modifiers (which are often prepositional phrases) that pinpoint objects and activities, and even abstract concepts:

> *“Your coat is under the dog.”*
>
> *“The rabbit is hiding behind that rock.”*
>
> *“I have an idea in my head.”*

Viewed from a certain angle, “the dog,” “that rock” “my head” become locations. Very “location-ish” nouns make such an interpretation even more tempting:

> *“He dropped the logs on the ground.”*
>
> *“He put the lamp back in its place.”*

However, none of these are taggable location expressions. They do not fall within any of the classes defined above for taggable locations. The annotator must be careful not to fall down this slippery slope.

Do not tag compass points 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 west.”

4.1.4 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.

4.1.4.1 Facility Entities used in Organization Contexts

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).

1. *The museum is located on Fifth Avenue.*
2. *I walked into the museum.*
3. *Mary works for the museum.*
4. *The museum insisted that the exhibition was not obscene.*
5. *The museum received a gift of $100,000.*

Examples 1 and 2 clearly refer to the museum building. Examples 3, 4, and 5 refer to the organization housed in or operating the museum facility. In cases like this, the annotation will reflect both the facility and organization entities. Please see the Metonymy section below for more information.

4.1.5 Geographical/Social/Political Entities (GPE)

Geo-Political Entities are composite entities comprised of a population, a government, a physical location, and a nation (or province, state, county, city, etc.). All mentions of these four aspects of a GPEs will be marked GPE and coreferenced. In this sentence,

The people of France welcomed the agreement.

there are two mentions

[The people of France]	*GPE*
[France]	*GPE*

The mention of the population of France is marked GPE, rather than PER. These mentions would be coreference as they refer to different aspects of a single GPE.

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 GPE 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 GPE entity. Thus, in this example, both "*New York*" and "*Boston*" would evoke organization entities.

4.1.5.1 GPE Clusters to be treated as GPEs

Like GPEs, clusters of GPEs consist of a populace, a well-defined physical territory, and in some cases (like Europe), have an organizing body (the European Union) associated with it. Because of their similarities to GPEs, these entities appear in contexts similar to those of GPEs. For example:

President-elect Kim Dae Jung today blamed much of ***Asia****'s devastating financial crisis on governments that "lie" to their people and "authoritarian" leaders who place economic growth ahead of democratic freedoms. [9801.404]*

Many of the leaders of ***Asian*** *society have been saying that military dictatorship was the way and democracy was not good for their nations," Kim said. [9801.404]*

*They concentrated only on economic development," he said, without singling out any nations but referring to "****Asian****-style democracy," in which governments are built around a strong leader who controls economic policy. [9801.404]*

For this annotation task, named geographical entities that are commonly referred to by those names will be considered GPEs rather than Locations. Following is a non-exhaustive list of entities that were Locations in the Pilot Study, but should be GPEs for this task.

Asia, Europe, Eastern Europe, Western Europe, EU, the Middle East, Palestine, Southeast Asia, New England, South Africa, all continents.

Other, more incidental clusters of GPEs are still considered Locations. For example, *the southern United States* is a Location. On the other hand coalitions of governments, as well as the UN, are organizational bodies and should be marked Organization.

4.1.5.2 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*").

4.1.5.3 GPE Mention Roles

Annotators need to decide for each entity mention in the text which role (Person, Organization, Location, GPE) the context of that mention invokes. This judgment typically depends on the relations that the entity enters into.

France *likes to eat cheese.*	Person Role
France *signed a treaty with Germany last week.*	Organization Role
The world leaders met in ***France*** *yesterday.*	Location Role

In the examples above, the name "France" refers to a range of concepts. Annotators must select the Role which matches the function of the GPE mention.

The GPE role may be used in contexts that highlight the nation (or state or province or city, etc.) aspect of the GPE entity, as distinct from the government, populace, and location, but that it may also be used in contexts referring to an indistinct amalgam of more than one of the aspects of a GPE (government, population, location, and nation).

France *produces better wine than New Jersey.*	GPE Role (whole nation)
France*'s greatest national treasure*	GPE Role (indistinct referent)

Even if more than one aspect of the entity is invoked by the context, only one role should be assigned. This usually occurs in the case of conjoined predicates. For example,

> ***Washington*** *is preparing for potentially massive demonstrations against the World Bank and the International Monetary Fund as ministers from those organizations arrive for Sunday's opening session.*

In the above example, it is the government of Washington (ORG) that is preparing for the demonstrations, but ministers will arrive at the location Washington. In these cases, the annotator should assign a role based on the closest local predicate. In this example, only the ORG role should be assigned to Washington because "preparing…" is the local predicate and invokes an ORG reading.

The following sections give particular guidelines for frequently encountered cases, with examples.

GPEs Modifying People and Artifacts

Pre-modifiers are inherently vague and difficult to decompose. For this reason, all GPE pre-modifiers of people and artifacts will be assigned the role GPE.GPE. For the sake of consistency, the corresponding post-modifiers should also be marked GPE.GPE. For example, ***[[GPE.GPE] French}*** *president* should be marked in the same way as *president of* ***{[GPE.GPE] France}***. More examples of GPEs modifying people include:

{[GPE.GPE] *Israeli* ***}*** *troops*

{[GPE.GPE] *New York* ***}*** *policemen*

Prime Minister of ***{[GPE.GPE]*** *Britain* ***}***

Joe Smith of ***{[GPE.GPE]*** *the United States* ***}***

{[GPE.GPE] *New York* ***}*** *attorney*

{[GPE.GPE] *U.S.* ***}*** *Commander-in-Chief*

GPEs modifying artifacts should also be marked GPE.GPE. Common artifacts modified by GPEs include but are not limited to vehicles, weapons, and flags. Some examples follow:

{[GPE.GPE] *U.S.* ***}*** *surveillance aircraft*

{[GPE.GPE] *Iraqi* ***}*** *flag*

Activities Associated with GPEs

Certain activities are associated with GPEs and therefore invoke a GPE role. For example, in *a pro-Iraq rally, Iraq* is assigned a GPE.GPE annotation. A rally is generally concerned with a nation, rather than exclusively a location or government.

> *The Palestinian Authority has banned pro-***{[GPE.GPE***] Iraq***}** *rallies, but that ban has been widely ignored.*

Military Activity

Similarly, military activities like invasions, military strikes, bombings, etc. are considered to be acts carried out by and directed at entire nations (not distinguishable from the government, people and location of that nation) and therefore are associated with GPEs. Both the aggressors and the victims in these cases are marked GPE.GPE.

> *The city could have used some special protection in nineteen seventy-nine when the* ***{[GPE.GPE]*** *Soviet Union***}** *invaded* ***{[GPE.GPE]*** *Afghanistan***}***.*

Political Communication and Decision-making

On the other hand, ORGs are responsible for decisions to take military actions. ORGs are also responsible for political communication events such as announcements, agreements, statements, denials, expressions of approval and disapproval, etc. So, if *China* agrees to something, *China* is a GPE.ORG.

> Ritter*'s return is seen as something of a test of that agreement, under which* ***{[GPE.ORG]*** *Iraq***}** *agreed to give inspectors full access to eight of Saddam Hussein's presidential palaces.*

Political associations

Political associations hold between people and GPEs. So in *Hillary Clinton (D-NY)*, *NY* is marked GPE.GPE.

> *"This is going to be a brutal fight," said Rep. Thomas C. Sawyer (D-***{[GPE.GPE]*** *Ohio***}***), who has been closely involved in the census and is among those who believes the ongoing debate played a role in Riche's departure.*

Embedding

GPE names embedded in mentions of the government have a GPE role. For example, in *the British government*, *British* is a GPE.GPE. This annotation conveys the relationship between nation and government. Similarly, in cases in which the embedded GPE conveys a political relationship with the location, the GPE is assigned a GPE role, as in the **{[GPE.GPE]**Israeli settlement**}**.

However, in cases in which there is only a locative relationship between the GPE and the LOC, the GPE is assigned a LOC role. For example, in *the heartland of America*, *America* is a GPE.LOC because a locative relation is conveyed.

> *Meanwhile, secretary of state Madeleine Albright, Berger and defense secretary William Cohen announced plans to travel to* ***{[GPE.LOC]*** *an unnamed city in* ***{***

> ***[GPE.LOC]** the **{[GPE.LOC]** US**}** heartland**}** **}** next week, to explain to the American people just why military force will be necessary if diplomacy fails.*

> ***{[LOC]** the**{[GPE.LOC]** Washington**}** area**}***

Athletes, Sports Teams, and GPEs

Athletes and teams are associated with GPE.GPEs as in *Picabo Street of the United States* below. Please note that *Picabo Street* is a person who was a member of the United States Olympic team.

> *Six days into the Nagano Games, one Alpine event _ the women's super-G won on Wednesday by Picabo Street of the **{[GPE.GPE]** United States**}**_ has been completed.*

However, when a GPE name is used as a team name (as in *Boston beat Philly*), the entity is marked as a metonymy, with the Literal mention being the city and the Intended mention being the team.

> ***{[GPE.GPE-Lit] [ORG-Int]** New York**}** had a shot to win but Chris Childs missed a three.*

In addition, because all GPEs are assigned a role, the Literal GPE mention is assigned a GPE role.

GPEs modifying organizations

In cases where GPEs modify organizations, the organizations are considered to be located in that GPE. Those GPEs should be marked GPE.LOC. So, in *New York corporation, New York* gets a GPE.LOC markup.

> *The **{[GPE.LOC]** California**}** company also asked that CAI be ordered to pay restitution to CSC "in an amount to be determined at trial."*

Governments

While the entity type for governments is GPE, the role for governments should always be GPE.ORG.

> *But **{[GPE.ORG]** the Russian government**}** and many politicians will be stridently critical of the United States if they believe they are being ignored.*

(In that particular example, *Russian* would also be marked, so that the full annotation for that phrase would be *{[GPE.ORG] the {[GPE.GPE] Russian} government}*, and the two GPE mentions would be coreferential.)

GPEs and Government Organizations

GPEs modifying government organizations, like *New York police department* and *Kentucky state fire marshall's office,* reflect a relationship between the organizations and the governmental aspect of the GPE, so they are assigned a GPE.ORG markup.

> *The department said Sonabend can appeal to **{[GPE.ORG]** Switzerland**}**'s supreme court.*

GPEs and Populations

As stated above, populations of a GPE are treated as GPE.PER. However, it is sometimes difficult to determine whether a reference to people is a reference to the population.

> *The **Japanese** have a considerable responsibility for the wars of the first half of the century*

In this example, the phrase *the Japanese* may be interpreted as the population of Japan, or the government of Japan, or the Japanese military, or even some part of the Japanese population. If the annotator believes that the phrase in question refers to the population of the GPE, or most of the population of a GPE, then the annotation should be GPE.PER and the mention is a name mention. However, if the annotator believes the phrase refers to a group of people, then PER is the assigned annotation and the mention is nominal because it does not refer to the name of a person. Examples:

> ***{[GPE.PER - name]** Cubans**}** have been waiting for this day for a long time.*
>
> ***{[GPE.PER - nom]** A majority of **{[GPE.PER - name]** Americans**} }** believe the allegations against Mr. Clinton are true.*
>
> *You and th- **{[GPE.PER - nom]** the **{[GPE.GPE - name]** American**}** people**}** have a right to- to get answers.*
>
> ***{[PER - nom]** A majority of **{[PER - nom]** Americans surveyed**} }** believes allegations Mr. Clinton had an affair while he was President are not relevant.*
>
> *Yet another cutting edge development by **{[GPE.PER - name]** the French**}** in their ongoing dealings with their enormous pet population.*
>
> *Butler says those sanctions could end soon if **{[GPE - name]** the Iraqis**}** allow the inspectors to do their job.*
>
> *The Missouri will come to rest near the memorial for the USS Arizona, which was sunk by **{[GPE - name]** the Japanese**}** during the attack on Pearl Harbor.*
>
> ***{[GPE.PER – nom]** The rest of **{[GPE.PER - name]** America**} }***
>
> ***{[PER - nom]** idealistic Europeans**}***
>
> ***{[PER - nom]** Americans who want to come and, and learn, uh, from the communities how to live in a community, how to take decisions among the community**}***
>
> *I do think there is a danger that **{[PER – nom]** some Chinese**}** may underestimate American will on the Taiwan issue.*

4.2 Mentions

For each entity, we record and coreference all mentions of the entity. Mentions are names, nominal phases, or pronominal phrases that refer to or describe the entity. For each mention, we record its full extent and its head.

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*".

4.2.1 Mention Extent

The extent of a mention consists of the entire nominal phrase. 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

[Some people yesterday who love chess]

The extent includes all the modifiers of a nominal phrase, including prepositional phrases, relative clauses, appositional phrases, etc. Thus the phrase

Fred Smith, the noted general

constitutes two mentions of one entity.

[Fred Smith, the noted general]

[the noted general]

Similarly,

Fred Smith, who is a noted general

constitutes two mentions.

[Fred Smith, who is a noted general]

[a noted general]

Generally speaking, tokens are broken at white space, and each item of punctuation is treated as a separate character. As a rule, we do not include punctuation such as commas, periods, and quotation marks in the extent of a mention unless words included within the extent continue on after the punctuation mark. 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.

4.2.2 Mention Head

In addition to the extent of the nominal phrase, the head of the phrase must be marked. In

The hurricane destroyed [the new glass-clad skyscraper].

the full 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 enclosed in brackets and the head is underlined:

> *[Fred Smith] became [the new prime minister].*
>
> *The job fell to [Abraham Abercrombie III].*

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 young] and [the restless]*
>
> *He was introduced to [five 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*.

4.2.3 Markability

4.2.3.1 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".

4.2.3.2 Conjunctions

In conjoined expressions, there should always be one and only one Nominal Entity per head noun. Thus, conjoined noun phrases with no elision of the head noun are to be tagged separately. If a pre-nominal modifier is present it gets included only with the initial noun phrase of the conjunct, and if a post-nominal modifier is present, it gets included only with the final noun phrase of the conjunct.

> *[muslims] and [croats]*
>
> *[many streams] and [rivers]*
>
> *[almost every serb], [croat] and [muslim in bosnia]*

[bus stations], [train stations], and [shopping areas throughout the country]

Note that the task of combining such conjoined expressions into "super-entities" is left for higher levels of processing. For example, one could imagine a preprocess for co-reference analysis in which additional entities are derived from conjoined Nominal or Named Entities:

{[many streams] and [rivers]} are overflowing their banks.

{Jimmy and Rosalyn Carter} donate their time to Habitat for Humanity.

4.2.3.3 Contractions

Ordinarily, we leave the *'s* out of the mention for possessives:

{We}'ll take {{John}'s car}.

For the possessive pronoun *its*, we include the *s* in the mention. Remember that *it's* is conventionally an abbreviation for "it is," while *its* is the correct way to write the possessive. But as a rule, don't trust the punctuation. You may see a sentence like this:

It's a non-profit corporation that gets all its money from donations.

Here the first *its* should be marked *{it}s*, the second *{its}.*

In the expression *Let's* , we mark the *'s* as a pronoun:

Let{'s} go!

4.2.4.3.3 Pronouns Referring to GPEs

Pronouns that refer to GPEs are marked as mentions of the same entity as their antecedent, but are assigned the role invoked by the context of the pronoun, which may not be the same as the role of the antecedent GPE.

*Composite Example: The president flew to **{[GPE.LOC]** Israel**}** to meet with **{[GPE.GPE]** its**}** Prime Minister.*

Similarly, in the case of classic metonymies (where two entities are created), pronoun annotation is determined in part by the link to the antecedent and in part by the context in which the pronoun appears. If the antecedent is a classic metonymy, the pronoun will be a mention of the same entity as either the literal mention or the intended mention of the antecedent.

*Metonymy Example: Thousands of parochial school and college students are joining this year's demonstration, including 1,500 high school students from across the country who spent last night at **{[ORG-Literal][FAC-Intended]** Catholic University**}**. **{[FAC]** It**}**'s in Georgetown.*

In some cases, the antecedent is not a metonymy but the context of the pronoun invokes an entity with a type that is different from that of the antecedent. In such cases, in addition to the mention of the new entity, the annotator should also mark the pronoun as a literal mention of the antecedent entity. (This allows us to maintain the connection between the pronoun and the antecedent.)

*Metonymy Example: **{[FAC]** The museum**}** is located on 45th Street. **{[FAC-Literal] [ORG-Intended]** They**}** just hired a new guard.*

4.2.3.4 Elision

Where elision of the head noun occurs in a conjunction, a single entity is delineated (these could also be viewed as conjoined modifier phrases):

*[the rain-soaked mid-atlantic and new england **states**]*
*[the successful and socially-responsible **manufacturers**]*
*[british and irish **governments**]*

4.2.3.5 Range Expressions and Elision

Components of range expressions are tagged separately if there is no elision of any head noun:

from [the foothills] to [the prairie]
from [the downtown area] to [the suburbs]

However, in examples like the following there is only a single head noun. In these cases we will treat the range expression as a pre-modifier, so that it gets included in the maximum extent of the entity:

ranging from [five to six companies] per day
from [blue collar to white collar workers]

4.2.3.6 Predicate complements

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 two entities: one of type person with only one mention, "*Fred*" and one of type person that is generic with only one mention “*a real linguist*”. Similarly,

Fred is studying to be a real linguist.

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

4.2.3.7 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*".

4.2.3.8 Proper adjectives

A proper adjective is to be treated as a name 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 GPE entity (if only "*French*" appears in a document, it evokes a GPE entity). The adjective is marked as a name mention of the GPE 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 GPE 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.

4.2.3.9 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* that 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.

4.2.4 Types of Mentions

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.

4.2.4.1 Names

For each entity, we record the occurrences of names (if any) used to refer to this entity in the document. For the purposes of ACE, a name is a noun phrase headed by a proper noun. Often the proper noun head is also the full extent of the noun phrase. We record each occurrence of the name of a given entity. If a name appears twice, both instances must be recorded.

Names are atomic. This means that entity names wholly contained within another name are not annotated. For example, in the following phrase only one entity is referenced.

The New York Times

This phrase references the organization of the newspaper. It does not evoke a separate entity for the city of "New York".

4.2.4.1.1 Head and Extent of Names

The following are head and extent rules that are specific to Name mentions.

Definite articles

When a definite article is commonly associated with an entity name, it also must be included in the head of the mention. Here are a few examples.

The Hague

The Rolling Stones

In both of these examples, the determiners are parts of the name of the entities. “The Hague” is an Anglicization of the Dutch “Den Haag” where “Den” is not the Dutch word for “the”. The annotation should include “The” in the head of the mention. “The Rolling Stones” is the name of a rock band. We will include the determiner in the head of the mention, as the band cannot be called *Rolling Stones*. The determiner is part of the head of the group’s name.

Titles and honorifics

Titles such as “Mr.” And role names such as “President” are not considered part of a person name. However, appositives such as “Jr.,” “Sr.” and “III” are considered part of a person name.

Mr Harry Schearer

Secretary Robert Mosbacher

John Doe, Jr.

Mister bettelheim

The revered jackson

Titles, honorifics, and determiners are all treated as modifiers, and are included in the extent of the mention of the person entity.

Multi-modifier Expressions

A single-name expression containing conjoined modifiers with no elision should be marked as a single expression.

U.S. Fish and Wildlife Service

The entire string is to be treated as the name of the organization.

4.2.4.1.2 Markable Names

The following are markability rules that apply specifically to name mentions.

Aliases and Nicknames

Generally, aliases for entities are to be tagged. Taggable aliases will include the following forms of entity names:

Acronyms, formed from the initial letter(s) or syllable(s) of successive or major parts of a compound term. Note that speech examples of acronyms may appear in a non-standard format. For example:

IBM

PACTEL

_a_t and _t

Nicknames and other aliases are tagged as names when they are established alternate ways of referring to an entity; if the annotator does not recognize the status of the nickname, it may be possible to determine from context whether the nickname is “established” or not.

The Big Apple	nickname for New York City
The garden state	nickname for New Jersey

Truncated Names, provided that the resulting form is clearly a proper noun referring to a specific entity, for example in:

Red Sox	alias for the Boston Red Sox
Sears	*alias for Sears Roebuck and Co.*

Entity Names that Modify Persons/Titles

Entity names modifying a person or their title/role are to be tagged.

Microsoft founder Bill Gates

The U.S. Vice-President

Each of the examples above gives us two mentions. Please note that nominal mentions of entities, which modify a person or their title, are not to be tagged.

company chairman James Smith

This example yields only one mention. “company” is not tagged.

4.2.4.2 Nominals

For the purposes of the ACE project, a nominal is a noun phrase headed by a common noun.

4.2.4.2.1 Nominal Left Modifiers

Nominal adjectives and non-possessive common nouns directly modifying other nouns are not markable mentions.

Markable:

I love {French} food.

Not Markable:

I love {prison} food.

4.2.4.3 Pronominals

A pronominal is a word used as a substitute for a noun phrase. Pronominals refer to persons or things that are previously specified or understood from the context.

Pronominals are marked whenever they reference a salient entity. When used as location pronouns, *here* and *there* are markable. Demonstratives *this, that, these,* and *those* are markable when they stand for a noun and not markable when the simply modify a noun. The various forms of *he, she,* and *it* are markable.

Here are examples where the pronoun should be tagged.

- *Northern Idaho is beautiful in the early summer. Motorcycle tourists love to come {**here**} and ride along the snowmelt-rivers.*
- *{**this**} is my grandmother*
- *{**those**} are the guys who stole my car.*
- *The White House and {**its**} surrounding area.*

Here are some examples where the pronoun should **not** be tagged.

- *There's a lot of talk about it.*
- *here is your change*
- *I'm standing here at the crime scene.*
- *those guys stole my car.*
- *this is the last straw!*
- *the dog and her puppies.*

The following are some additional rules that apply to pronominal mentions.

4.2.4.3.1 Headless Mentions

Mentions with empty heads are classified as pro-mentions.

> ***five** of the analysts*

Please note that this example also includes the nominal mention [the analysts].

4.2.5 Coreference of Mentions

If two mentions refer to the same underlying entity, we must indicate this by coreferencing them. In most cases, this is very straightforward. In an article about Osama bin Laden, we want all mentions of Mr bin Laden to be lumped together in the same entity and marked with the base type PER. So, if the following sentences appeared in the same article, we would want to include all the bold mentions in the Osama bin Laden entity.

> *Videos circulated by **Osama bin Laden** have added to the evidence linking **him** and the al-Qaida network to the Sept. 11 terrorist attacks in the United States, the government said Wednesday in an updated dossier on the investigation. The document, published by Prime Minister Tony Blair's office, said **the Saudi dissident** had come "closest to admitting responsibility" for the attacks in an "inflammatory video," allegedly made on Oct. 20, that was not released to the media but circulated to al-Qaida members. "The battle has been moved inside America, and we shall continue until we win this battle, or die in the cause and meet our maker," the document quotes **bin Laden** as saying.*

The name mentions of Osama bin Laden are easy to spot. Please note, however, that we must coreference all mentions that refer to the entity that is Mr bin Laden. This will include nominal mentions such as the Saudi dissident and pronominal mentions such as him.

5 Metonymy

Metonymy occurs when a speaker uses a reference to one entity to refer to another entity (or entities) related to it. For example, in the sentence below *Beijing* is a capital city name that is used as a reference to the Chinese government:

> ***Beijing** will not continue sales of anti-ship missiles to Iran.*

Classic metonymies make reference to two entities, one explicit and one indirect reference. Common examples are cases of capital city names standing in for national governments, as shown above. Other common examples involve facilities and organizations, which are closely related in that organizations typically have facilities, and facilities are typically owned and administered by organizations. Thus when a facility is mentioned, the organization is sometimes also referenced. So, in *the museum announced its new exhibit*, the entity *museum* is a facility that houses artwork, but in this context it is the organization running the museum that is doing the announcing. In cases like this, where both entities are expressed by the same phrase, two entity mentions should be marked, one for each of the corresponding references. If only one entity is expressed, then only one entity mention is marked. In the above example, the annotator would mark mentions of a FAC and an ORG entity for *the museum*.

Classic metonymies are to be annotated with two separate mentions, one for each of the entities referred to. This naturally means that each of those mentions will need to be linked appropriately to any other mentions of that entity in the document. For example, there is a building (a FAC) called the "Holocaust Memorial Museum" but the name of this building is also often used to refer to the organization that runs its business in that building. Thus, in a sentence like the following, "the museum" would be marked as two mentions, one associated with the FAC entity and the other associated with the ORG entity.

> *But Lerman also added that* ***{[FAC][ORG]*** *the museum* ***}*** *would not extend Arafat the formal courtesies that are routine for other world leaders.*

If, elsewhere in the document, a mention of "the museum" occurred in the context "New windows were ordered for the museum", that mention would be marked as an additional mention of the same FAC entity referred to above, but not as an additional mention of the ORG entity.

In cases like the above, where two mentions are marked on the same text, annotators are to specify which of the two mentions is the "literal" one and which the "intended" metonymic one. The Alembic Workbench will support this by allowing the properties "literal" or "intended" to be added to mentions. In examples in these guidelines, the literal mention will always be listed first. Both the literal and the intended mentions, with the entities underlying them, will be counted in the scoring.

The remainder of this section outlines specific annotation guidelines for metonymy in different contexts.

5.1 Capital City for Governmental GPE

Cases in which the capital city is used to refer to the nation's government are marked as true metonyms. (Because two separate GPEs are involved, this is not an exception to the general rule that GPEs are marked as one entity with a role rather than as two entities.)

> *Secretary of Defense William S. Cohen said today that he is satisfied* ***{[GPE.GPE][GPE.ORG]*** *Beijing* ***}*** *will not continue sales of anti-ship missiles to Iran as he wrapped up a four-day visit here that underscored improving Sino American military ties.*

In this example there are two mentions covering the word Beijing. The GPE.GPE is a mention of the city Beijing and the GPE.ORG is a mention of China. The GPE.ORG mention is a mention of the same China entity that would be referred to by other GPE mentions of "China" that might be found elsewhere in the document. Also if there were a later mention of the city of Beijing (for example, *Cohen left the city this morning*), it would be a GPE.LOC mention of the same Beijing entity referred to by the GPE.GPE mention in the above example.

5.2 Metonymies Involving ORG Base Entities

There is a table (see the Pilot Study task definition, Section 6.2.5) that specifies a “base” type for various kinds of entities. Mentions of entities with ORG base types like schools, restaurants, or churches are sometimes used to refer to the organization itself, and sometimes used to refer to the facility that houses that organization. Every mention of such an entity is to be marked (at least) as a mention of an entity of its base type. A second mention of a different type should also be marked if the context invokes a metonymic entity. Thus a mention whose base type is ORG but that is used in a FAC context will have mentions of both of those two entities associated with it.

Below are some examples of ORGs that refer either to a single base type entity, or else to both a base type and metonymic type entity.

Example 1

Universities have an ORG base type so both mentions of the university in 1A and 1B invoke an ORG entity. But 1B also invokes a FAC entity because it refers to the site.

> *Lee Jung Hoon, a political science professor at **{[ORG-1]** Yonsei University**}**… (From 9801.162)*

Thousands of parochial school and college students are joining this year's demonstration, including 1,500 high school students from across the country who spent *last night at **{[ORG-2] [FAC-3]** Catholic University**}**. (From 9801.*175)

Example 2

Embassies have an ORG base type so both 2A and 2B invoke an ORG entity. But 2A also invokes a FAC entity because FACs, not ORGs have gates.

> *…a few hundred ethnic Albanians laid a black wreath at the gate of **{[ORG-4] [FAC-5]** Yugoslavian embassy**}**. (From APW19980308.0201)*
>
> *“Our Ministry of Defense is working very hard with **{[ORG-6]** the U.S. Embassy in Bogota**}** to get the information together," Cano said. (From 9801.382)*

5.3 Metonymies Involving FAC Base Entities

The same approach used for ORG entity mentions that refer to an associated FAC should also be used when a FAC entity mention refers to an associated ORG.

Here are two examples from the same document (9801.266):

> *Competing self-images of victim hood have long prevented Israelis and Arabs from acknowledging the full weight of each other's historical tragedies, and many Arab leaders have resisted efforts to lure them to **{[FAC-7]** the museum**}** and the similar Yad Vashem memorial in Jerusalem.*
>
> *Lerman, reached at his New Jersey home, said the subject of Arafat and Israel's talks with the Palestinian Authority still profoundly divided U.S. and world Jewry and "we believe **{[FAC-8] [ORG-9]** the museum**}** should not get involved in a political dispute where half of the people are for something and half are against it."*

Since museums have a FAC base type, both examples A and B invoke a FAC entity. But example B also invokes an ORG entity because it is the organization that should not get involved in the dispute.

Note in the above examples that FAC mention 7 and FAC mention 8 refer to the same FAC entity, as shown in the following table of entities and mentions:

> *Entity 1:* ***{[FAC-7]*** *the museum* ***}***, ***{[FAC-8]*** *the museum* ***}***
>
> *Entity 2:* ***{[ORG-9]*** *the museum* ***}***

Another common class of FAC metonymies is found when named buildings are used to refer to the organizations based there:

> *It is unlikely* ***{[FAC] [ORG]*** *the White House* ***}*** *would nominate a successor who did not support sampling, and equally unlikely Republican leaders would look favorably on such a candidate.*

5.4 Special Rule for Offices and Branches

Because the term “office” is often used to refer to an organization, as in “the Office of the Attorney General,” the base type for offices will be ORG. When the context suggests a reference to the physical entity, the entity should be marked both ORG and FAC. Examples that are ambiguous as to whether a facility or an organization is intended should be marked metonymically, with both an ORG and a FAC mention. Thus in the following example the office is marked both ORG and FAC because it is unclear whether the context suggests that the investigators are from the physical office or from the organization.

> *Investigators from* ***{[ORG-9] [FAC-10]*** *the Kentucky state fire marshal’s office* ***}****.*

(In that particular example, *Kentucky* would also be marked, so that the full annotation for that phrase would be ***{[ORG-9] [FAC-10]*** *the* ***{[GPE.ORG]*** *Kentucky* ***}*** *state fire marshal’s office* ***}***.)

The same general guidelines apply to other facility terms like “branches” (as in the local branch of a bank).

5.5 Metonymies Involving LOC Base Entities

Entities whose base type is LOC can also be used in metonymic senses. In the following example, “the world” has literal type LOC but intended type PER, and thus is annotated with two separate mentions:

> ***{[LOC] [PER]*** *The whole world* ***}*** *was watching.*

6 Entity Class (Generic/Specific)

An entity is generic when it does not refer to a particular object or particular set of objects in the world. Every entity must be designated as either generic or specific. In some cases this distinction is difficult to make. This section will outline several tests that will help differentiate between the two classes.

6.1 Definition of Generic and Specific

A given common noun (*girl*, *motorcycle*, *bookmark*, *semantic theory*, etc.) denotes a *set of objects*, each of which is an example of the noun in question. In such a system, "boy" would refer to the set BOY whose membership would be precisely *all the boys in the world* (or perhaps: *in the Universe*).

The manner in which NPs refer can be easily explained relative to this backdrop:

1. Some NPs are used to refer to *a particular object in the world*. The set X (the common noun's referents) from which that object is drawn has little significance to the audience, other than to help in the selection of the (particular) object in question.

These NPs say something like: *there is a specific example of X, one that I have in mind, that ...* and are considered to be *non-generic*.

(Note that we will use non-generic and specific interchangeably in the present set of documents. The former is arguably more appropriate, since the annotation conventions adopted here tag the feature GENERIC as either *true* or *false*, but we will let the latter serve as form of shorthand notation.)

2. Other NPs are used to refer to *underspecified objects that may be an example of the set (X) in question, but need not be particular*. Here the set X has a greater degree of significance, since the only constraint on the entity in question is that it be drawn from that set.

These NPs say something like:

> *"Any member of the set X ..."; or*
>
> *"Each member of the set X ..."*

and are considered to be *generic*.

In short, a generic mention is used to refer to *any member of the set in question* rather than *some particular, identifiable member of that set* (which would be picked out by a *Non-generic* mention) and a formal definition seems altogether impossible. As shall soon become clear, we can do little better in providing this notion with a precise definition.

We have therefore allowed the above informal (*folk*) definition --- together with the following discussion of the phenomena; the subsequent taxonomy of common generic-denoting mentions; and the concluding short list of (non-deterministic) tests for the applicability of generic status to a given mention --- to serve as the basis of our tagging decisions with regard to the attribution of generic status.

The (un-)reliability of syntactic or contextual tests here will become clear as the discussion proceeds --- it is helpful to correspondingly consider each of the examples which follow as having a (frequently secondary) role in illustrating this fact, whether or not this expository role is explicitly stated.

6.2 Classes of Mentions Frequently Associated with Generic Entities

We can make some loose generalizations about the classes of NPs, which are likely to refer to generic entities, but it is important to bear in mind the source of our reluctance to offer such categorical (or syntactic) criteria for the assignment of generic status to a given NP.

Typically, generic entities include types of entity, suggested attributes of entities, hypothetical entities, and generalizations across a set or sets of entities.

6.2.1 A Type of Entity

{Mammals} are live bearers.
{Good students} do all the reading.
{Typical firemen} work hard all their lives in dangerous conditions.

6.2.2 A Suggested Attribute of an Entity

John seems to be {a nice person}.
{Misfits} are sometimes {the best employees}.

6.2.3 A Hypothetical Entity

If {a person} steps over the line, {they} must be punished.
Aides say he's plotting a political comeback, even considering a run for president} in two thousand.

6.2.4 A Generalization across a Set of Entities

{Outsiders} think that New Jersey is a different country.
{Purple houses} are really ugly.

Even if the *property* or the *set* underlying the entity in question is extremely constrained (i.e. such that there are very few possible members), that entity should still be considered *generic*.

{People who drive at night in red cars} are likely to get tickets.

The police are looking for {a man who wears green suits and carries a purple briefcase}.

The first of these examples falls into the *Type of Entity* category. The second is a *Hypothetical Entity*. The man in the second example may or may not exist (even though the police are looking for him).

Note that this mention would not be generic if the context went on to say specific things about the man wearing green suits. We have seen several examples of this case above. This is only *generic* if it is unclear if such a person actually exists.

6.3 Tests for Generic-hood

6.3.1 Words that are commonly generic

'anyone', 'most Xs', 'more Xs' tend to be generic, even if the author has someone in mind.

{Anyone who carries a gun} is dangerous.

{Most doctors} are just in it for the money.

{More investigators} are needed for this case.

6.3.2 Determiners

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} here. [generic]

   I saw gorillas {a gorilla} here. [specific]

2. These noun phrases in modal contexts (such as *belief*, *desire*, ...) are *generic*:

   I want to see {gorillas}.

   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?

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.

Occasionally noun phrases with "the" are generic, even though this is not typically the case. We find this 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}. [specific]

{The dodo} is extinct. [generic]

{The dodo} is dead. [specific]

6.3.3 Positive Assertion Test

This test applies to predications such as "*X* is *Y*" (as in the subsequent example). If *X* is specific, then *Y* will be as well, because *Y* is positively asserted of *X*. *Y* is assumed to be coreferential with *X* and therefore specific.

{Joe} is {a nice guy}.

If *X* is *generic* and *Y* is positively asserted of *X*, then *Y* is also *generic*.

{Firemen} are {nice guys}.

This test is less effective when someone other than the author of the story makes the positive assertion. This is just an instance of the case in which a modal context forces a generic reading (as in II-2 above).

> *Mary says that {Joe} is a {a nice guy}.*

This sort of statement falls into the pattern

> *person Z* says/said/thought/*etc. that* X *is* Y

This only counts as a positive assertion if *Y* is not an attribute and person Z is a trustworthy source of information. This case, however, is the exception rather than the rule. Most modal contexts are entirely opaque, and the assertions found inside will not generally hold "in the real world." This means that even the entities at play in such assertions cannot be reliably anchored in "reality;" that there is probably not a specific entity in the world to which the *beliefs*/*desires*/*assertions* of the speaker are linked (via the embedded proposition within which the mention intimating such an entity is located). In the case of:

> *John believes that a gorilla stole his lunch.*

We must assume that "any gorilla will do" (or, at least, that "it could be the case that any gorilla will do").

6.3.4 Negation Tests

1. Common nouns with "no" as a determiner are *generic*.

> *I saw no people in the room.*

2. **Negated pronouns** are *generic*.

> *I saw no one.*
>
> *I saw nobody.*

3. **Negated full NPs** can be *specific*.

> *Who would do that? Not {Joe}.*
>
> *Neither {Joe}, nor {Mary} said anything.*

4. **Common nouns modified by "neither" and partitives with "neither"** can be *specific* (depending on coreference) because the negative properties of "neither" have scope over more than just the NP.

> *{Neither person} left the room.*
>
> *{Neither of {them}} likes to talk much.*

6.3.5 Boiler Plate Test

These are NPs that have a legal-like hypothetical setting. We sometimes call them "*empty shell*" mentions.

> *Each year, we elect {one chairman} and {ten board members}.*
>
> *There can be only one {Miss America} for any given year.*

Given an actual instance of the hypothetical setting, these NPs would be "filled in" by actual entities. All these to-be-instantiated NPs should be marked *generic*.

Notice that this test is not exclusively forward-looking. We also see this phenomenon for classes of *previous* (or *iterative*) "empty shell mentions" serving as the generic entity in question. For example:

> *{Former U.S. presidents} have a hard time finding jobs.*
>
> *{The host} rarely steals the show on Saturday Night Live.*

The first example refers to a *generic entity* for which the entire membership is well defined. Any competent historian of the U.S. governm ent can easily provide an exhaustive list of the members of **FORMER_US_PRESIDENT** --- a trick that does nothing to avert the assignment of *generic* status to the entity picked out by the relevant mention. Rather, we are still compelled to assign generic status by the observation that "former U.S. presidents" is used here to refer to any of a set of objects (**FORMER_US_PRESIDENT**), rather than someone in particular (e.g. Jimmy Carter).

The second example is an iterative case that includes as members both the membership of a (well-defined) set (**FORMER_SNL_HOSTS**) and the membership of a (presently undefined/unpopulated) set (**FUTURE_SNL_HOSTS**). Again, we are not torn by the (partial, extensional) definition of the set. We can see right away that the mention "The host" is being used to pick out *any of* a set of entities (without being particular). By our working definition, the mention is therefore **generic**.

It seems that the **Boiler Plate Test** has been poorly defined above (Test IV). We really intend to distinguish between ***the position itself*** and ***the (current) occupant of that position*** --- where the former is **generic** and the latter **specific**.

Appendix

Sections to be added

Coreference with aliases which refer to more than one entity

> *Kobe Bryant is the next Michael Jordan.*
>
> *Bill Clinton will go down in history as the Jon Bon Jovi of US presidents.*

Job Positions and Titles

Add either as a part of or directly following the Titles and Honorifics section.