

# Simple Named Entity Guidelines

## For Less Commonly Taught Languages

### Version 6.5 – March 28, 2006

Linguistic Data Consortium – LCTL Team  
(Based Largely on the MUC-7 NE Guidelines)

1	Introduction .....	2
2	Entity Types .....	3
2.1	Person Names .....	3
2.1.1	Family Names .....	3
2.2	Titles, roles and appositives .....	4
2.2.1	Titles containing other Entities .....	4
2.2.2	Names broken by Titles .....	5
2.2.3	Sequences of Titles .....	5
2.2.4	Ordinal Suffixes .....	5
2.2.5	Non-Markable Titles .....	5
2.3	Organization Names .....	6
2.3.1	Organizations and “Tag for Meaning” .....	6
2.3.2	Articles in Organization Entities .....	7
2.3.3	Organizations not to tag – Generic references .....	8
2.4	Location Names .....	8
2.4.1	Locations and “Tag for Meaning” .....	8
2.4.2	Extent of Location Names .....	9
2.4.2.1	Articles in Location Entities .....	9
2.4.2.2	Compound expressions .....	9
2.4.2.3	Designators .....	9
2.4.2.4	Location modifiers and "semi-official" place names .....	10
2.5	Deciding among entity types .....	11
2.5.1	ORG referring to LOC, LOC referring to ORG .....	11
3	Difficult Cases .....	12
3.1	Expressions that refer to multiple entities .....	12
3.2	Nested Expressions .....	12
3.3	Entities in pre-modifier position (Adjective Names) .....	13
3.4	Possessives .....	13
3.5	Other types of names .....	14
4	What NOT to tag .....	14
4.1	Events .....	14
4.2	Artifacts and products .....	14
4.3	Generics .....	15
5	Annotation Uncertainty .....	15

# 1 Introduction

An entity is some object in the world – for instance, a place or a person. A named entity is a phrase that uniquely refers to that object by its proper name, acronym, nickname or abbreviation. Some examples of named entities follow:

Coca-Cola Bottling Co.

Bob Austin

the Eiffel Tower

IBM

the Yankees

Uganda

Bowdon, Georgia

Mt. Fuji

the Kremlin

the Kennedys

When annotating Named Entities, there are a number of things to remember at all times:

1. We always tag an expression according to its meaning in the context being evaluated. In other words, the annotation of an expression depends on how it is being used. We will call this rule *Tag for Meaning*.
2. We will only tag names<sup>1</sup>. Names are defined as expressions that uniquely refer to objects directly. The meaning of the parts of names are not typically part of the meaning of the name (i.e. names are not *compositional*) and, therefore, names cannot be broken down into smaller parts for annotation.
3. We will only tag names that refer to objects of the relevant types: PER, ORG, LOC and TTL. These types will be defined in additional detail in Section 2 below.

---

<sup>1</sup> As will be described in Section 2.2, Titles are a special case, since they aren't really names at all.

4. When tagging Named Entity expressions, we will not allow overlapping or embedded annotations. In other words, every annotation must end before another can begin.

## 2 Entity Types

We will identify four types of named entities:

**PERSON (PER):** Person entities are limited to humans identified by name, nickname or alias.

**TITLE/ROLE (TTL):** Personal titles or roles. These are restricted to titles that occur directly before or after the person name they describe.

**ORGANIZATION (ORG):** Organization entities are limited to corporations, institutions, government agencies and other groups of people defined by an established organizational structure.

**LOCATION (LOC):** Location entities include names of politically or geographically defined places (cities, provinces, countries, international regions, bodies of water, mountains, etc.). Locations also include man-made structures like airports, highways, streets, factories and monuments.

Other types of named entities like animals, inanimate objects and monetary units will not be annotated.

Within this document, named entities are indicated by [square brackets].

Note that annotations **may not** overlap or embed in the text. In other words, every annotation must end before another can begin.

### 2.1 Person Names

People may be specified by name, nickname or alias. Names of deceased people, as well as fictional human characters appearing in movies, television, books and so on, should be tagged as PERSON entities. Religious deities should also be tagged as PERSON.

#### 2.1.1 Family Names

Family names should also be tagged as PERSON. In such cases, we will often see the article (*the*) preceding the surname:

The Smiths

These articles should be included in the extent of family names.

## 2.2 Titles, roles and appositives

Titles, roles and honorifics such as "Mr." and "President" are tagged as title entities and are separated from the individual's name. For instance, in the following sentence, there are two separate entities marked:

[Vice President] [Cheney] visited the site.

Some more examples of names and titles:

[GlobalCorp] [Vice President] [John Smith]  
ORG TTL PER

[Treasury] [Secretary] [Jackson]  
ORG TTL PER

the [U.S.] [Vice President], [Dick Cheney]  
ORG TTL PER

[Justice] [Minister] [Giovanni Maria Flick]  
ORG TTL PER

[Mission Control] [Chief] [Vladimir Solovyov]  
ORG TTL PER

[spokesman] [Mary Gillette]  
TTL PER

[US Army] [negotiator] [Harold Norman]  
ORG TTL PER

**NOTE:** Titles are special. In a sense, these elements are not properly names. So it should be no surprise that there will be some specific rules for dealing with them. The most important rule is that we will only tag expressions as TITLE/ROLE elements when they occur directly adjacent to a named PERSON entity. The following subsections will describe these rules in more detail.

### 2.2.1 Titles containing other Entities

If a title contains another taggable entity, then tag each entity separately<sup>2</sup>.

---

<sup>2</sup> Notice that the prohibition against overlapping or nested annotations (we may **never** insert one Named Entity *inside* of another one) forces us to turn the string "Microsoft

For instance:

```
[Microsoft] [Chairman] [Bill Gates] stated that...
  ORG          TTL          PER
```

## 2.2.2 Names broken by Titles

Finally, sometimes the name of the person is split into two pieces by the title. In these cases, we will annotate the two pieces of the PERSON name as two separate PERSON entities:

```
[Alfred] [Lord] [Tennyson]
  PER      TTL    PER
```

## 2.2.3 Sequences of Titles

Sometimes, more than one title will be presented for a single person. In case all of these titles are adjacent to the PERSON name, we will tag each of the titles separately as TITLE/ROLE.

```
[Karachi][Mayor], and [XYZ][Chairman] [Mr] [Anwar Ayub].
  ORG      TTL          ORG      TTL    TTL    PER
```

Notice that the coordinating conjunction *and* is not counted when trying to determine if the proposed TTL element is adjacent to the relevant Person name.

## 2.2.4 Ordinal Suffixes

You may occasionally encounter an ordinal suffix like "Jr.", "Sr.", and "III". These are considered part of a person name and should be marked as part of the name, for instance:

```
[Mr.] [Albert Franklin, Jr.] was part of the research team.
  TTL    PER
```

## 2.2.5 Non-Markable Titles

For this task we define titles and roles as occurring either directly before or directly after a named PERSON entity. Therefore, titles and roles are only

---

Chairman Bill Gates" into a sequence of Named Entities. This is a special case for ORG entities otherwise contained within TITLE/ROLE elements – which are themselves required to be adjacent to a markable PERSON name. In general, we may not break Named Entities into smaller pieces.

tagged when they occur directly next to the person name they modify. In the following example, for instance, the phrase "Vice President" is not considered a title and is not tagged:

The strongest supporter was the Vice President.

## 2.3 Organization Names

Tag all proper name mentions of groups with a defined organizational structure. These include:

### Businesses

[Bridgestone Sports Co.] profits

### Stock exchanges

[NASDAQ] shares

### Multinational organizations

[European Union] representatives

### Political parties

[GOP] hopeful

### Non-generic government entities

[the State Department]

### Sports teams

[the Phillies]

### Military groups

[the Tamil Tigers]

### 2.3.1 Organizations and "Tag for Meaning"

Many other kinds of entities refer to facilities or buildings that are primarily defined by their established organizational structure, and can do things like issue statements, make decisions, hire people, raise money and so on. A

mention of such an entity should be tagged as an ORGANIZATION when it functions like an ORG in the document. These include things like:

### Churches and other religious institutions

[Trinity Lutheran Church]

### Hospitals

[Finger Lakes Area Hospital Corp.]

### Hotels

[Four Seasons Hotel Group]

### Museums

[the Guggenheim Museum]

### Universities

[the University of Chicago]

### Government offices

[the White House]

Since we tag for meaning, these entities are only to be tagged as ORG when they are being used as such. Often the above examples will be tagged as LOC entities. For rules about distinguishing these types, please see Sections 2.4 and 2.5, below.

## 2.3.2 Articles in Organization Entities

The definite and indefinite determiners *the* and *a* should typically be included in the annotation of Organization names:

the State Department

the Phillies

the Tamil Tigers

the Guggenheim Museum

the University of Chicago

the White House

This rule does not apply in cases when they quantify something other than the tagged entity, as in the following examples:

A [Gulshan Hotel] spokesman  
the [U.S.] Vice President

This exception is particularly common when the tagged name is used in the pre-modifier (adjective) position.

### 2.3.3 Organizations not to tag – Generic references

General entity mentions such as "the police" and "the government" should not be tagged, since these are not unique proper name references to specific entities.

## 2.4 Location Names

Examples of place-related strings that are tagged as LOCATION include named heavenly bodies, continents, countries, provinces, counties, cities, regions, districts, towns, villages, neighborhoods, airports, highways, street names, factories, manufacturing plants, street addresses, oceans, seas, straits, bays, channels, sounds, rivers, islands, lakes, national parks, mountains, fictional or mythical locations, and monumental structures, such as the Eiffel Tower and Washington Monument. For instance:

the collapse of the newly-constructed [Teton Dam]  
LOC

the dispute over votes in [Dade County]  
LOC

[The Walt Whitman Bridge] remained closed  
LOC

repairs began on a 10-mile stretch of [the Alaskan Pipeline]  
LOC

[The Garden State] is known for its tomatoes.  
LOC

### 2.4.1 Locations and “Tag for Meaning”

Recall that we will always tag an expression on the basis of how it is being used. As such, there will be a number of cases where an expression typically used to refer to a location will instead be used to refer to an Organization. See section 2.3 and 2.5 for more details.

## 2.4.2 Extent of Location Names

There are several issues surrounding the expression of location names and which parts of a string to tag.

### 2.4.2.1 Articles in Location Entities

The definite and indefinite determiners *the* and *a* should typically be included in the annotation of Location names:

the Guggenheim Museum

the University of Chicago

the White House

This rule does not apply in cases when they quantify something other than the tagged entity, as in the following examples:

A [Gulshan Hotel] driveway

the [U.S.] mainland

This exception is particularly common when the tagged name is used in the pre-modifier (adjective) position.

### 2.4.2.2 Compound expressions

Compound expressions in which place names are separated by a comma in English should be tagged as separate instances of LOCATION.

[Kaohsiung], [Taiwan]

[Washington], [D.C.]

### 2.4.2.3 Designators

When a "designator" is customarily used as a regular part of a place name, that word should also be included in the extent of the LOCATION entity. For example, include in the tagged string the word "River" in the name of a river, "Mountain" in the name of a mountain, "City" in the name of a city, etc., if such words are contained in the string.

```
[Mississippi River]
```

```
[the Himalayan Mountains]
```

#### 2.4.2.4 Location modifiers and "semi-official" place names

Often times place names are modified by words like "Southern", "Lower", "West", "the former" and so on.

When these modifiers are part of a location's official name they should be tagged as part of the LOCATION name. For instance:

```
[Upper Volta]
```

```
[North Dakota]
```

Even if the place name does not have "official" status but has an agreed-upon definition and is in very frequent use, the string should be tagged as a LOCATION, as in:

```
[the Middle East]
```

```
[the West Bank]
```

```
[Eastern Europe]
```

When these modifiers are not the official name of a place, or when the definition of the place might vary from person to person, do not tag the modifier as part of the LOCATION entity name.

```
[Mississippi River] west bank
```

```
former [Soviet Union]
```

```
[Gaul], in present-day [France]
```

```
lower [Manhattan]
```

These place names can sometimes be tricky. If you are not sure whether a modifier is part of an official name, you should include the modifier as part of the place name.

## 2.5 Deciding among entity types

There are some situations where deciding what entity type to assign can be somewhat tricky.

### 2.5.1 ORG referring to LOC, LOC referring to ORG

Many organizations have not only an organizational structure, but a physical location. For instance, museums are primarily organizations but are also housed in a specific building or facility. So while we normally tag museums as ORG entities, there are cases when a particular example might function more like a LOCATION. In cases like this, annotators should tag the named entity based on the way it functions in the sentence. For instance:

[The Guggenheim Museum] announced a new acquisition.  
ORG

[The Guggenheim Museum] was designed by [Wright].  
LOC PER

Thirty people were wounded in the bomb blast in front of the city's [Gulshan Hotel].  
LOC

A [Gulshan Hotel] spokesman called the incident a tragedy.  
ORG

Similarly, city, country, and other place names are frequently used to refer to organizations located in those places rather than the geographical places themselves. For instance:

[Washington] announced a new tax policy today.  
ORG

In this case, the name Washington is used to refer to the US Government, located in Washington. Because Washington is referring to an organization entity in this example, it should be tagged as ORG.

Also, when the name of a unique structure or building (normally a location) is used to refer to the government or other organization housed in that facility, the name should be tagged as an ORG:

[The Pentagon] issued a statement about the incident.  
ORG

The same logic applies to place names referring to sports teams:

[Boston] beat [New York] last night in extra innings.  
ORG                      ORG

In summary, for any cases where a place name is used to refer to an organization, you should tag the name based on function in the sentence:

**ORG:** used when the example primarily refers to the organizational structure, and is acting like an agent (issuing a statement, making a decision, hiring people, raising money, etc.)

**LOC:** used when the example primarily refers to the physical structure, rather than the people/groups who run it.

### 3 Difficult Cases

#### 3.1 Expressions that refer to multiple entities

When a phrase refers to multiple named entities, mark each entity separately.

For instance, this sentence contains two entities:

[China] and [South Korea] signed the agreement.

Similarly,

[Jimmy] and [Rosalyn Carter]

[North] and [South America]

But be careful not to split apart proper names that contain a conjunction. For instance,

[the Fish and Wildlife Service]

is the name of one organization and should be tagged as a single named entity (it's not “the Fish Service” and “the Wildlife Service” as separate names).

#### 3.2 Nested Expressions

Recall that no nested expressions will be marked. When the name of one entity contains within it another entity name, do not pull out the name of the

other entity and mark it separately<sup>3</sup>. Only tag the larger entity. For instance

[Arthur Anderson Consulting] no markup for Arthur Anderson alone

[Boston Chicken Corp.] no markup for Boston alone

### 3.3 Entities in pre-modifier position (Adjective Names)

If an entity name modifies another word (even if that word is not a taggable entity type), you should still tag the entity name.

[Bridgestone] profits

the [Clinton] government

[Treasury] bonds and securities

[U.S.] exporters

[Apple] computers

[Texas] intermediate crude oil

[China] film festival

Similarly, if the entity name occurs in the form of an adjective you should also tag it:

the [American] companies

[Chinese] food

[Enron] documents

a [Midwestern] bank

### 3.4 Possessives

When you encounter a possessive construction, tag the two parts individually as two separate names. For instance:

[Temple University's] [Graduate School of Business]

[Canada's] [Parliament]

---

<sup>3</sup> PERSON entities that are adjacent to TTL and ORG pre-modifiers are a notable exception. Please see Section 2.2.

Keep in mind that annotation requires you to select whole words to tag as names, so you have to include the “s” even though it’s not part of the name.

### 3.5 Other types of names

Aliases, acronyms, nicknames and abbreviations for proper names should be tagged as a name:

IBM	(abbreviation for International Business Machines)
Big Blue	(alias for International Business Machines)
Big Board	(alias for New York Stock Exchange)
Mr. Fix-It	(nickname for candidate for head of the CIA)
the Big Apple	(nickname for New York City)
Red Sox	(alias for the Boston Red Sox)
Sears	(alias for Sears Roebuck and Co.)

## 4 What NOT to tag

### 4.1 Events

Do not tag event names, even if they refer to events that occur on a regular basis and are associated with institutional structures. However, the institutional structures themselves — steering committees, etc. — should be tagged.

the Pan-American Games	(no markup)
vs.	
the Olympic Committee	(ORG)

### 4.2 Artifacts and products

Miscellaneous types of proper names that are not to be tagged as named entities include artifacts, other products, and plural names that do not identify a single, unique entity. For instance,

the Taurus is the latest car model	(no markup)
------------------------------------	-------------

### 4.3 Generics

Generic names that do not refer to a specific entity should not be tagged

`the Campbell Soups of the world` (no markup)

## 5 Annotation Uncertainty

In some cases, you may encounter examples that you don't know how to handle. If so, you should proceed as follows:

Your supervisor might not be available at the moment you have a question or issue, so it's important for you to write down the problem so it can be resolved later.

So that you can keep working even after you have a question about a particular example, we've created one more tag in the annotation tool: "No\_Annotation". When you encounter a problem or have a question about a particular word or phrase and you can't get an immediate answer, label the item "No\_Annotation". That will let us easily find it later when we try to resolve the problem.

You should also use the "No\_Annotation" label in cases where there's some kind of problem with a single word or handful of words in the file – e.g., they're badly translated or the font isn't displaying properly.

If the whole file is problematic (i.e., poor translation, corrupted, font problems), stop working on it and let your supervisor know. If you are using AWS to receive file assignments, you may simply mark the file as "Broken" in the file-assignment interface. If you use this approach, please be sure to include a descriptive comment when AWS asks what is wrong with the file.

- If it's an example not covered in the guidelines, note it in your copy of the guidelines and let your supervisor know about it.
- If it's an example where your language differs from the rules as written for English, note it in your copy of the guidelines and let your supervisor know about it.
- If it's a case where there's something wrong with the file you're working on, stop working on that file and let your supervisor know.