Labeling Manual
for
Broadcast News Data

(dLabel 2.01)

5 November 2003
Columbia University

Table of Contents

1 INTRODUCTION

We are labeling the broadcast news data manually with a set of tags. We want to use the labeled data to train automatic procedures to find a number of things in unlabeled news broadcasts automatically, using Machine Learning techniques. You will be labeling two kinds of things:

- segments (anchor signon and signoffs, headlines, stories, commercials, interviews, and soundbites)
- entities (person names in various categories, organization names, locations, dates, and times)

The person and organization names and the locations you will be labeling will be proper names, e.g. George Jones or IBM or New York City. The person names will be labeled as anchors, reporters, interviewees or just person for all others..

The labeling tool ***dLabel 1.02*** can be used to label the data. A pair of tags <TAG></TAG> is put around the segments or entities to be labeled with that tag. The labeling tool has icons that help you do this

You can follow links in
http://columbia.gaasman.org for answers of questions you may have.

Particularly, you can post your questions at
http://www.gaasman.org/columbiabboard/

You can download the data including this manual at
http://www.cs.columbia.edu/~smaskey/download

You can upload your file at
http://columbia.gaasman.org/file_upload.html

and see the uploaded file at
http://columbia.gaasman.org/uploaded_data/

2 INSTALLATION AND USE OF DLABEL 1.02

dLabel 1.02 is a labeling tool which is still in the build process. Please, post your questions with the software in the web board.

2.1 INSTALLATION

You need to have java installed in your computer before you can run this software. To install java:

1. Download Java 2 Platform, Standard Edition (J2SE) from http://java.sun.com/j2se/1.4.2/download.

2. Download for your operating system and install it.

3. If you are running windows machine then the binary to run java will be at c:\j2sdk(your version)\bin

4. To run the software:

   In Linux, run the following command "java LabelData" in the directory where you downloaded the software.

   For Windows, open the dos prompt by clicking all programs-> accessories -> Command Prompt and change directory where you downloaded the software by using cd command. "cd directory name" then type "java dataLabel"

Problems?

If you can't run "java dataLabel" it is very likely the path is not set right so you can either set the path or you can run by giving full path like " c:\j2sdk(your version)\bin\java LabelData"

2.2 USING THE TOOL

You highlight the text and click on the icon to place a beginning tag <TAG> and an ending tag </TAG> around the text you highlighted. Please remember there can be tags within tags. In order to places tags within tags, tag the text using the first button. Then highlight the text including the first tags and then click on the second button to place second set of tags.

The next version of the tool will include features like button shortcuts and color coding capabilities. If you have any particular feature that you think will be useful in the tool please post your tips at the web board.

3 DATA

3.1 To get the data and upload the annotations once you are done.

1. For each message, you will be highlighting various types of information and clicking on the appropriate button to label that information with its **type.**
2. To do this:
   - Download a news program to be labeled from http://www.cs.columbia.edu/~smaskey/download/data/
   - Fill in the file name, start date, and your name,….in the information table at http://columbia.gaasman.org/info.php.
   - Bring up the labeling tool on your pc.
   - Load in the file you have downloaded.
   - Go through the file, labeling all of the items described below by highlighting them and clicking the appropriate button.
   - When you have finished, upload the file to http://columbia.gaasman.org/fileupload.php.
   - Fill in the end date for this file in the information table (see above).
   - Post any questions that arose during the transcription to the labeling bboard at http://www.columbia.gaasman.org/bboard.php; you may also post questions during the transcription of course, or send them by email to julia@cs.columbia.edu.
   - If you have any problems while labeling, send mail to julia@cs.columbia.edu and smaskey@cs.columbia.edu.

4 ANNOTATION

4.1 General guidelines

Do not label ***incomplete shows***, i.e. any show that does not have an anchor signon and signoff, or in any other way appears to you to be incomplete. Indicate in the info file (http://columbia.gaasman.org/info.php) that this is an incomplete program, so no one else will bother to download it. If in doubt, indicate this in the info file and we will check.

Please make sure when you markup things that you have not deleted or added any characters except for well-formed label markup. Especially if you do not use a button for some label, or if you do not highlight exactly what you mean to before you hit the button, ill-formed input may result, e.g.

Location>Idaho</Location>
I<Location>daho</Location>

Be sure also that you do not delete any character or label in the original file.

Use the Notes button in the transcriber to add notes to a transcript. For example, if you are not sure about which label to assign, use <Person>Rocky</Person><Note>should we label characters in films?</Note> following the labeled entity in question. Any question you have while labeling, you can indicate similarly with a note following the item in question. <Location>Middle Earth</Location><Note> I wasn't sure about this.</Note> Be careful ***not*** to put <Note> labels around other labels, e.g. do ***not*** do <Note> <Location>Middle Earth</Location> I wasn't sure what to do about this.</Note>.

If you have any question that arises more than once, be sure to post a message to the bboard or send mail directly to julia@cs.columbia.edu.

4.2 Segmentation

News broadcasts contain many items of interest. Some represent "parts" of the show. These include:

- Anchor initial **signon**/greeting and identification of the news show
- Listing of **headlines** for the show
- **Commercials** within the show
- **Stories** within the show (already labeled)
- **Interviews** within the show
- **Soundbites** of (uninterviewed) speakers
- Anchor **signoff**/closing from the show

Normally, there will be only one initial greeting, one headline listing, and one closing/signoff. However, there may be more. There can be multiple stories, interviews, commercials, and soundbites within each show. Note that every speaker turn uttered by someone other than an Anchor or a Reporter must be labeled either as a part of an Interview, or as part of a Soundbite.

4.2.1 Anchor signons

Anchor signons will typically occur early in the show, possibly after some initial headlines; exactly where will depend on the show. These are said by the anchor and identify both the show and the anchor name, e.g.

…And a Navy sailor learns about Internet privacy the hard way. <Anchor greeting> <Organization>CNN</Organization> Headline News."
I'm <Anchor>Lyn Vaughn</Anchor>.</Anchor greeting>

For anchor signons and closings, it is obviously important to identify who the anchor is in each show. For some shows there may be more than one anchor. If in doubt, post to the bboard.

4.2.2 Headlines

Headlines also occur early in the show, generally only once. These typically include one sentence on each of the major stories that will be presented in the show and are spoken by the anchor. Begin these with the first headline sentence and end with the last, not including, e.g., the anchor signon:

<Headlines>
<TEXT>

Vice President <Person>Gore</Person> arrives in icy
<Location>Maine</Location> as another northeast storm looms
on the horizon.
</TEXT>
</BODY>
<END_TIME> 01/15/1998 11:31:10.46 </END_TIME>
</DOC>
<DOC>
<DOCNO> CNN19980115.1130.0070 </DOCNO>
<DOCTYPE> MISCELLANEOUS TEXT (manually segmented)
</DOCTYPE>
<DATE_TIME> 01/15/1998 11:31:10.46 </DATE_TIME>
<BODY>
<TEXT>
The <Location>Kentucky</Location> teenager accused of opening fire on
his high school classmates
heads to court.
And a Navy sailor learns about Internet privacy the hard way.
</Headlines>

4.2.3 Stories.

Each news broadcast is primarily made up of a series of news stories. Since these have already been labeled in our data, you do ***not*** have to label story segments.

4.2.4 Interviews

Interviews typically occur within news stories, although sometimes an entire story may be an interview. An interview is an alternation of turns between a reporter or anchor and an interviewee, who is a person who does *not* work for the show. This alternation typically has the appearance of a series of questions or comments by the reporter directed to the interviewee, who may or may not be identified, e.g.

<Interview>
It was a gastrointestinal virus of some sort with flu-like symptoms like
diarrhea and vomiting.
We had 29 people that were ill with it over the course of the weekend.
<TURN>

<ANNOTATION> Reporter: </ANNOTATION>
<Interviewee>Tripp</Interviewee>, who's in charge of the
<Organization>red cross</Organization> Shelter in
<Location>Augusta</Location>, <Location>Maine</Location>, says
people a<TURN>
<ANNOTATION> Reporter: </ANNOTATION>
What happens when people are crowded together in a Shelter for nearly a
week?
Some of them get sick.
<TURN>
We're not sure if it was the flu or not.
re feeling better now.
And the <Organization>red cross</Organization> was able to contain the
spread thanks to some
precautions.
<TURN>
We made sure we isolated them in an area where they're away form the
general population to minimize the spread.
</Interview>

Begin the interview segment with the first reporter turn and end with the last interviewee or reporter turn directly related to this interview, but you should ***not*** include any material from these turns that is not directly relevant to the interview – e.g., do ***not*** include any reporter signoff in the interview segment (e.g. "<Reporter>Elizabeth Cohen</Reporter>,
<Organization>CNN</Organization> reporting. Similarly, do ***not*** include any material from the reporter turn in which the interview begins that is not directly concerned with the interview.

Note above that the interviewee is not explicitly introduced although in this case he is mentioned early in the story. Some interviewees are neither introduced nor mentioned, so you must infer that there is an interview based on the sequence of turns and earlier context. E.g., the preceding interview is directly followed by:
<Interview>
TURN>
There are other health problems at shelters.
<TURN>
There's me and the three kids and we want to go home.
<TURN>
The kids love it, because they have other kids to play with.

Butite been stressful.
<TURN>
</Interview>
Here the reporter is apparently continuing on to interview one of the shelter inhabitants, not the person running the shelter.

Note that it is possible that each of these interviewees has been interviewed by someone else and the network reporter's questions spliced in. That does not change the character of each of these segments as an interview, however.

4.2.5 Soundbites

Soundbites differ from interviews in that there is no sense of dialogue between the non-broadcast person and an anchor or reporter. These are typically turns quoted from a public figure, e.g. an excerpt from a speech or press conference, in which there is no indication that the included turns resulted from someone asking particular questions of the speaker. For example,

<DOC>
<DOCNO> CNN19980115.1130.0389 </DOCNO>
<DOCTYPE> NEWS STORY </DOCTYPE>
<DATE_TIME> 01/15/1998 11:36:29.23 </DATE_TIME>
<BODY>
<TEXT>
President <Soundbite-Speaker>Clinton</ Soundbite-Speaker> says this year's medal of freedom honorees helped <Location>America</Location> to widen the circle of democracy.
He honored Japanese American civil rights activist <Person>Fred Korymatsu</Person>.
Their contributions continue the legacy of another medal honoree.
<Soundbite>
<TURN>
It is fitting that today this ceremony occurs on the birthday of Dr. <Person>Martin Luther king Jr.</Person> who 21 years ago was granted this award by President <Person>Carter</Person> to ensure that his legacy would live on.
Until every child has the opportunity to live up to his or her God-given potential, free from want and a world at peace, Dr. <Person>King</Person>'s work and our work is not yet done.

</Soundbite>
<TURN>
The medal of freedom is the nation's highest civilian honor.
</TEXT>
</BODY>
<END_TIME> 01/15/1998 11:37:32.11 </END_TIME>
</DOC>
However, they may also represent segments of speech from private persons, as when someone's speech is recorded for a story but you do not think the speech seems like the response to a question. If you have doubts about whether something is part of an interview or is a soundbite, include a note after the segment.

Note that when speakers of soundbites are named in the broadcast, their names should be labeled as <Soundbite-Speaker>. See below.

4.2.6 Commercials

Commercials should include only entire turns. Do not label entities within commercial segments. If multiple commercials occur together, label the entire set as one commercial.

> <Commercial><commercial></Commercial>

4.2.7 Anchor signoff/closings:

Anchor closings are spoken by the anchor and typically occur near the end of the program. There should typically be only one per show. Do ***not*** confuse anchor closings with reporter closings, which may occur at the end of stories.

> <Anchor closing>I'm <Anchor>Lyn Vaughn</Anchor>. This is "<Organization>CNN</Organization> Headline News."</Anchor closing>

4.3 Entity Identification

Some general issues that apply to all entity tagging:

4.3.1 Nested expressions:

No nested expressions will be marked within entities. For example, where Location expressions occur within Organizations, only the larger expression

will be marked. Similarly with all other nestings, mark only the larger entity containing the smaller expression.
the <Organization> U. S. Customs Service</Organization>

Of course, there will be entity expressions marked within larger segments, such as interviews.

4.3.2 False starts and repairs:

False starts and repairs should be included inside the entity tags. For example, "…This is <Reporter>George Thomas</Reporter> reporting from <Location>Kab uh from Kabul</Location>."

4.3.3 Personal name tagging:

For all personal names, include only *first name/initial middle name/initial last name*. Do ***not*** include titles or roles (e.g. Mr., President, Sgt.) or appositives (e.g. "<Person>Lee Bollinger</Person>, president of <Organization> Columbia University</Organization>"), except for Jr., Sr., III.. Do not include other following identifying material (e.g. "<Reporter>Mitch Renley, <Organization> CNN</Organization> News").

Mr. <Person>Harry Schearer</Person> died tragically.
Secretary <Person>Robert Mosbacher</Person> died tragically.
<Person>John Doe, Jr.</Person> died tragically.

Family names should be tagged, e.g. "the <Person>Kennedy</Person> family", "the <Person>Kennedys</Person>"

Other uses of personal names that should ***not*** be tagged are:

"the Gramm-Rudman amendment", "the Nobel Prize", "St. Michael"

Personal names are tagged in one of four ways, according to whether or not the person is a participant in the newscast**.** The following are the possible tags:

4.3.3.1.1 Anchor names:

The distinction between anchors and reporters should usually be clear from where and how often each speaks during a broadcast. The person who signs on and off at the beginning and ending of a show is clearly an anchor. Sometimes there will be more than one anchor, however. Other cues to who is an anchor are: they present the headlines, they do not specify a non-studio location,…
**how do they tell?

This is <Anchor>Peter Jennings</Anchor> for <Organization> ABC</Organization> news.
"Yes, <Anchor>Peter</Anchor>, I'm here in <Location>Baghdad</Location>…"

4.3.3.1.2 Reporter names:

Reporters are speakers who are employed by the broadcasting company but are often reporting from locations other than in the main studio. Cues such as "reporting live" or "Mary Lee, CNN, Hong Kong" should tell you the speaker is a reporter.
This is <Reporter>Mitch Renley</Reporter> reporting live…
"Tell me, <Reporter>Mitch</Reporter>, what do you see in the direction of the airport?"

4.3.3.1.3 Interviewee names:

Use with people whose speech actually is recorded in the broadcast but who are not anchors or reporters.

"I am here with <Interviewee>Rudolph Giuliani</Interviewee>, former mayor of <Location>New York</Location>."

(Other) Person names: "It is said that Mayor <Person>Bloomberg</Person> will not run for re-election."

Sometimes you may not know if someone is going to be an interviewee when they are initially mentioned. When you find out that they are, go back and change the <Person> labels to <Interviewee>.

4.3.3.1.4 Soundbite-Speaker:

Label speakers of sound-bites as Soundbite-Speaker, just as you do for Interviewees and Interviews.

4.3.3.1.5 Other personal names:

Use for anyone whose speech is not recorded in the broadcast but who is identified by name. Do ***not*** use when the name is not present (e.g. "a former mayor of <Location>New York</Location> said")

Former <Location>New York</Location> mayor <Person>Rudoph Giuliani</Person> spoke today of his political ambitions.

4.3.4 Organization names:

Organizations to be tagged include named corporate, governmental, or other organizational entities.

<Organization> IBM</Organization> announced layoffs today.
<Organization> Intel's</Organization> profits rose dramatically.
Business executives now follow the <Organization>
GE</Organization> model.

If there are regular words within the title or the name of the organization include such regular word as well. e.g. (<Organization>Boston Chickering Corporation</Organization>).

Corporate Designators: Corporate designators such as "Co." are part of an organization name, e.g. <Organization> Bridgestone Sports Co.</Organization>

4.3.4.1 Miscellaneous Organization-type Entity-Expressions:

These include stock exchanges, multinational organizations, political parties, orchestras, unions, non-generic governmental entity names such as "Congress" or "Chamber of Deputies", sports teams and armies and should be tagged, unless these are designated only by a Location name. For example:

<Organization> NASDAQ</Organization>
<Organization> European Community</Organization>

<Organization> GOP</Organization> presidential hopeful
<Organization> Machinists</Organization> union
the mayor who built Candlestick Park for the <Organization>
Giants</Organization>
<Location>Russia</Location> defeated <Location>France</Location> by a score of…

4.3.4.2 Articles appearing with Organization expressions

These generally should not be tagged, e.g. "the <Organization> University of Chicago</Organization>

4.3.4.3 Proper names referring to facilities

Entities such as churches, embassies, factories, hospitals, hotels, museums, and universities, will be tagged as Organization:

<Organization> Finger Lakes Area Hospital Corp.</Organization>

<Organization> Four Seasons Hotels</Organization>

the <Organization> White House</Organization>

<Organization> Trinity Lutheran Church</Organization>

"the Empire State Building" (no markup)

4.3.4.4 Event-Type Non-Entities:

Do not tag, e.g. "the Pan-American Games". ***Do*** tag institutional structures that are associated with these, e.g. <Organization> U. S. Olympic Committee</Organization>. A location name that is part of an event name should be tagged if the location name is not in adjectival form (as in "the Pan-American Games"); so, "<Location>China</Location> Film Festival"

4.3.4.5 Do not tag names of news shows themselves as organizations

E.g. Do ***not*** tag "Headline News". ***Do*** tag "<Organization>ABC</Organization> News.

4.3.5 Location names:

Location names include the name of politically or geographically defined locations (cities, districts, neighborhoods, villages, airports, highways, street names, street addresses, islands, national parks, fictional or mythical locations, monumental structures that were built primarily as monuments, towns, provinces, countries, international regions, bodies of water, mountains, heavenly bodies, continents).

from <Location>Paris</Location> to
<Location>London</Location>
The <Location>Eiffel Tower</Location>
On the <Location>Ohio River</Location>
The <Location>Gulf of Mexico</Location>
In the <Location>Sierras</Location>
Trucks collided on a rain-soaked <Location>Interstate 5</Location>

Include the smallest contiguous place identifier. When additional modifiers occur (e.g. "<Location>Southampton</Location> in the south of <Location>England</Location>, bracket only the sub segments of the phrase.

If the name of an airport refers to the organization or business of the airport and it is still tagged as Location, e.g. <Location>Massport</Location> owns <Location>Logan Airport</Location>

4.3.5.1 Metonyms

Terms that refer to political, military, athletic and other organizations by the name of a city, country, or other associated location. These would be tagged as Location, not Organization. E.g.

<Location>Germany</Location> invaded <Location>Poland</Location> in <Date> 1939</Date>.

<Location>Baltimore</Location> defeated the <Organization> Yankees</Organization>…

4.3.5.2 Locative Entity-Expressions Tagged in Succession:

Compound expressions in which place names are separated by a comma are to be tagged as ***separate*** instances of Location, e.g.:

<Location>Kaohsiung</Location>, <Location>Taiwan</Location>
<Location>Washington</Location>, <Location>D. C.</Location>
<Location>Brooklyn</Location>,<Location>NY</Location>
the <Location>U.S.</Location> <Location>Virgin Islands</Location>

4.3.5.3 Locative Designators and Specifiers:

Designators that are integrally associated with a place name are tagged as part of the name, e.g.

<Location>Mississippi River</Location>
<Location>Mount McKinley</Location>
<Location>The Hague</Location>

4.3.5.4 Locative Non-Entities:

Do not include common noun phrases functioning as partitive-type locative specifiers directly after Location names, e.g.:

<Location>Mississippi River</Location> west bank

However, due to its political significance the term "West Bank" (of the Jordan River) may be tagged as Location. This is a judgment call.

4.3.5.5 Transnational and Subnational Region Names:

Tag names of continents e.g. "Africa" and regions, e.g. "Middle East", "Pacific Rim". Do not tag names of sub-national regions when reference only by compass-point modifiers, e.g. "the Southwest region", or "the South", since these may refer to multiple locations in different contexts. Do tag names of sub-national regions when they are identifiable even out of context, e.g. "the Ruhr", "the Auvergne", and "Amazonia".

In <Location>North Africa<\Location>
Politics in <Location>Eastern Europe</Location>
The <Location>Pacific Northwest</Location> but…the Northwest…
A vacation in the <Location>Caribbean</Location>

4.3.5.6 Do *not* tag:

Adjectival forms of location names:

"American exporters"
"Caribbean cooking"
"Israeli prime minister"
"African travel"

4.3.6 Date entities:

Tag absolute and relative temporal dates and times. These may be complete or partial. The salient features of the time expressions that are marked is that, whether absolute or relative, they can be anchored on a timeline; unanchored durations, for example, are not marked.

DATE is defined as a temporal unit of a full day or longer. DATEs may be either absolute or relative. Both absolute and relative dates are tagged as Date.

4.3.6.1 Absolute Date Expressions

To be considered an absolute date expression, the expression must indicate a specific segment of time, as follows:

* An expression of days must indicate a particular day, such as "Monday," "10th of October" (not "first day of the month").

* An expression of seasons must indicate a particular season, such as "autumn" (not "next season").

* An expression of financial quarters or halves of the year must indicate which quarter or half, such as "fourth quarter," "first half." Note that there are no proper names, per se, representing these time periods. Nonetheless, these types of time expressions are important in the business domain and are therefore to be tagged.

* An expression of years must indicate a particular year, such as "1995" (not "the current year").

* An expression of decades must indicate a particular decade, such as the <Date>1980s</Date> but not "the last 10 years".

* An expression of centuries must indicate a particular century, such as the <Date>20th century</Date> (but not "this century"). Contiguous date expressions are to be tagged as a single item. E.g.,

<Date>January 1990</Date>

<Date>fiscal 1989</Date>

the <Date>autumn</Date> report

<Date>third quarter of 1991</Date>

<Date>the three months ended Sept. 30</Date> (as referring to the fourth quarter

<Date>the first half of fiscal 1990</Date>

<Date>first-half</Date> profit

<Date>fiscal 1989's fourth quarter</Date>

<Date>4th period</Date> (of a year)

<Date>1975</Date> World Series

Determiners that introduce the expressions are not to be tagged. Words or phrases modifying the expressions (such as "around" or "about") also will not be tagged. Only the actual temporal expression itself is to be tagged.

around the <Date>4th of May</Date>

shortly after the <Date>4th of May</Date>

4.3.6.2 Relative Temporal Expressions

A relative temporal expression (RTE) indicates a date relative to the date of the document ("yesterday", "today", etc. Taggable RTE's include compound temporal expressions containing a deictic marker followed by a time unit, such as "last month" or "next year". If a numeral is included in RTE's of this type, it falls within the scope of the taggable temporal expression ("last two months"). Note that sometimes the deictic marker is postposed, as in "10 years ago" and "four months later". Note also that some RTE's lexicalize deictic markers and time units into a single word, such as "yesterday", which by itself constitutes a taggable expression, and that some RTE's can contain more than one deictic marker, such as "early this year" and "earlier this month." In addition, note that some of the expressions specifically defined as not being absolute temporal expressions are considered markable as relative temporal expressions.

Miscellaneous Temporal Non-Entities: Indefinite or vague date expressions with non-specific starting or stopping dates will ***not*** be tagged. Non-taggable expressions include:

Vague Time Adverbials

"now", "recently", etc.[no markup]

Indefinite Duration-of-Time Phrases

"for the past few years" [no markup]

Time-Relative-to-Event Phrases

"since the beginning of arms control negotiations"[no markup]

Scope of Temporal Expressions Absolute time expressions combining numerals and time-unit designators associated with a single Date tag, are to be tagged as a single item. That is, the subparts (such as numbers and time-units) are not to be tagged separately, even in the case of possessive or partitive constructions. E.g.,

the <Date>first half of fiscal 1990</Date>

Temporal Expressions Containing Adjacent Absolute and Relative Strings: When a time expression contains both relative and absolute elements, the entire expression is to be tagged. The following examples illustrate some of the ways in which elements of relative and absolute time expressions may combine to form taggable time expressions.

<Date>July last year</Date>

the <Date>end of 1991</Date>

<Date>late Tuesday</Date>

Holidays: Special days, such as holidays, that are referenced by name, should be tagged.

because of the observance of <Date>All Saints' Day</Date>

Temporal Expressions Based on Alternate Calendars: Temporal expressions in terms of alternate calendars, such as fiscal years, the Hebrew calendar, Julian dates and "Star Date," will generally be marked up in accordance with the above guidelines for Date.

4.3.7 Cross-entity ISSUES:

The following include issues that apply over all entities or that apply where more than one entity is involved.

Time and Space Modifiers of Locative Entity Expressions: Historic-time modifiers ("former", "present-day") and directional modifiers ("north", "upper", etc.) are taggable only when they are intrinsic parts of a location's official name, e.g. "Upper Volta" or "North Dakota". But…

Former <Location>Soviet Union</Location>
<Location>Gaul</Location> (present-day <Location>France</Location>)
lower <Location>Manhattan</Location>

Entity-Expressions that Modify Non-Entities: Entity names used as modifiers in complex NPs that are not proper names are only to be tagged if it is clear to the annotator from context or world knowledge that the name is that of an organization, person, or location.

The <Person>Clinton</Person> government

<Organization> Treasury</Organization> bonds and securities

<Location>U.S.</Location> exporters

<Organization> Bridgestone</Organization> profits

Entity Expressions that Modify Titles: Entity names modifying person identifiers should be tagged:

<Organization> MIPS</Organization> Vice President <Person>John Hime</Person>

<Organization> Treasury</Organization> Secretary

the <Location>U. S.</Location> Vice President

Entity-Strings Embedded in Entity-Expressions: Multi-word strings that are proper names may contain entity name substrings that are not decomposable; those strings should not be tagged:

<Organization> Arthur Anderson Consulting</Organization>

<Organization> Boston Chicken Corp</Organization>

<Location>Northern California</Location>

<Location>West Texas</Location>

Entity-Expressions that "Possess" Other Entity Expressions: In a possessive construction, the possessor and possessed entity substrings should be tagged separately:

<Organization> Temple University</Organization>'s <Organization> Graduate School of Business</Organization>

<Location>California</Location>'s <Location>Silicon Valley</Location>

<Location>Canada</Location>'s <Organization>
Parliament</Organization>

Entity-Expression Aliases: Aliases for entities should be tagged. Taggable aliases include the following forms:

Acronyms formed from the initial letter(s) or syllable(s) of parts of a compound terms, e.g. <Organization> IBM</Organization>, <Organization> PACTEL</Organization>

Nicknames, e.g. <Organization> Big Blue</Organization>, <Organization> Big Board</Organization> (alias for NY Stock exchange), the <Location>Big Apple</Location>, Mr. <Person>Fix It</Person>

Truncated names, if the result is clearly a proper name referring to a specific entity, e.g. <Organization> Red Sox</Organization> (for Boston Red Sox) or <Organization> Sears</Organization> (for Sears Roebuck and Co.)

Some metonyms, such as " the <Organization> White House</Organization>" and the <Organization> Pentagon</Organization>

Quotation marks around an alias are included if they appear within the entity name, e.g.

<Person>Vito "The Godfather" Corleone</Person>

also known as <Person>"The Godfather"</Person>

The definite article in an alias, as in <Person>The Godfather</Person>

Do ***not*** tag aliases such as:

Common nouns or pronouns such as "the company" in "<Organization> IBM</Organization> announced that the company…"

Aliases that refer to broad industrial sectors, political power centers, etc., such as "the Ivy League", "the Axis", "Iron Curtain countries".

Embedded Locative Entity-Strings and Conjoined Locative Entity-Expressions: The phrase "of ,place-name. Following an organization name may or may not be part of the organization name proper. If there is a corporate designator, it is an organization name; otherwise "of <place-name> is part of the organization name.

<Organization> Hyundai of Korea, Inc.</Organization>

<Organization> Hyundai, Inc</Organization> of
<Location>Korea</Location>

<Organization> McDonald's of Korea</Organization>

Miscellanous non-entities: Things that should not be tagged include:

Artifacts, other products and plural names that do not identify a single, unique entity, such as:

"the Campbell Soups of the world"
"Dow Jones Industrial Average"
<Organization> Ford</Organization> Taurus

Multi-name or multi-number expressions: A conjoined multi-name/number expression, in which there is elision of the head of one conjunct, should be marked up as a single expression, e.g. "<Location>North and South America</Location>"

Multi-modifier expressions: A single-name expression containing conjoined modifiers with no elision also should be marked up as a single expression, e.g. "<Organization> U. S Fish and Wildlife Service"</Organization>

Numeric range expressions: The subparts of date range expressions should be marked up as parts of a single expression, even if there is no elision of the numeric units, e.g. "<Date> from 1990 through 1992</Date>"

In possessive constructions like
<ORGANIZATION>Citibank's</ORGANIZATION>president
<PERSON>Bill Ford</PERSON> tag the organization and the name separately.