

* INTRODUCTION

** Task Overview

The Named Entity task is to identify and classify segments of a text stream which "refer" to objects in ten semantic classes.

The "identification" of a text segment is just the insertion of SGML tags at the beginning and end of the segment in the text stream. The "classification" of a text segment is the use of one of ten different types of SGML tags for the insertion. This insertion of SGML is known as "annotation" or "tagging".

** SGML Description

The guidelines for Mandarin will use abbreviated SGML tags, as described in the following table.

Object Type	Abbreviated SGML	Actual SGML that is to be used in the task	
Person	<PER> </PER>	<B_ENAMEX TYPE="PERSON">	<E_ENAMEX>
Location	<LOC> </LOC>	<B_ENAMEX TYPE="LOCATION">	<E_ENAMEX>
Organization	<ORG> </ORG>	<B_ENAMEX TYPE="ORGANIZATION">	<E_ENAMEX>
Other names	<ONA> </ONA>	<B_ENAMEX TYPE="OTH_NAME">	<E_ENAMEX>
Date	<DTE> </DTE>	<B_TIMEX TYPE="DATE">	<E_TIMEX>
Time	<TME> </TME>	<B_TIMEX TYPE="TIME">	<E_TIMEX>
Duration	<DUR> </DUR>	<B_TIMEX TYPE="DURATION">	<E_TIMEX>
Money	<MNY> </MNY>	<B_NUMEX TYPE="MONEY">	<E_TIMEX>
Measure	<MSR> </MSR>	<B_NUMEX TYPE="MEASURE">	<E_TIMEX>
Percent	<PCT> </PCT>	<B_NUMEX TYPE="PERCENT">	<E_TIMEX>
Cardinal	<CRD> </CRD>	<B_NUMEX TYPE="CARDINAL">	<E_TIMEX>
Other numbers	<ONU> </ONU>	<B_NUMEX TYPE="OTH_NUM">	<E_TIMEX>

Note: The "Other names" and "Other numbers" are not part of the task, but human annotators may use them to identify text segments of interest.

** Simple Examples

*** Person

^邓 ^小平 (deng xiaoping)
 <PER>^邓 ^小平</PER>
 [mc970226]

*** Location

波士顿 (boston)
 <LOC>波士顿</LOC>
 [h4_ma98-v2.utf]

*** Organization

世界贸易组织 (world trade organization)
 <ORG>世界贸易组织</ORG>
 [mc970114.utf]

*** Other names (not evaluated)

汉语 (han language, [i.e., Chinese])
 <ONA>汉语</ONA>
 [mc970114.utf]

*** Date

三月十二日 (march twelfth)
 <DTE>三月十二日</DTE>
 [XIN19980312.0066]

*** Time

九点四十六分 (nine forty-six)
 <TME>九点四十六分</TME>
 [mc970411]

*** Duration

四天 (four days)
<DUR>四天</DUR>
[mk970003]

*** Money

一千三百亿元 (1300,0000,0000 yuan)
<MNY>一千三百亿元</MNY>
[XIN19980312.0077]

*** Measure

三十五公里 (thirty-five kilometers)
<MSR>三十五公里</MSR>
[mv970620a]

*** Percent

百分之十九 (nineteen percent)
<PCT>百分之十九</PCT>
[mk970009]

*** Cardinal

三个经济特区 (three economic special regions)
<CRD>三个</CRD>经济特区
[mc970305.utf]

*** Other numbers (not evaluated)

二零二六一九三一一一 ((202) 619-3111)
<ONU>二零二六一九三一一一</ONU>
[mv970521]

** Discussion

The ten evaluated semantic classes given above can be grouped into three subtasks of the Named Entity task:

Entity names (ENAMEX - persons, locations, and organizations),
Temporal expressions (TIMEX - dates, times, and durations)
Number expressions (NUMEX - money, measures, percents, and cardinal numbers)

For many text processing systems, such identifiers are recognized primarily using local pattern-matching techniques. The TEI (Text Encoding Initiative) Guidelines for Electronic Text Encoding and Interchange cover such identifiers (plus abbreviations) together in section 6.4 of the TEI Guidelines and explain that the identifiers comprise "textual features which it is often convenient to distinguish from their surrounding text. Names, dates, and numbers are likely to be of particular importance to the scholar treating a text as source for a database; distinguishing such items from the surrounding text is however equally important to the scholar primarily interested in lexis."

The system must produce a single, unambiguous output for any relevant string in the text; thus, this evaluation is not based on a view of a pipelined system architecture in which Named Entity recognition would be completely handled as a preprocess to sentence and discourse analysis. The task requires that the system recognize what a string represents, not just its superficial appearance. Sometimes, the right answer is superficially apparent, as in the case of many numerical expressions, and can be obtained by local pattern-matching techniques. In other cases, the right answer is not superficially apparent, as when a single word could represent the name of a location, person, or organization, and the answer may have to be obtained using techniques that draw information from a larger context or from reference lists. Transcriptions of speech lack much of the punctuation found in electronic newswire articles; this missing information makes certain decisions regarding proper names more difficult. Speech recognizers may have trouble identifying numbers accurately, which comprise a large portion of temporal and numerical expressions.

The three subtasks correspond to three SGML tag elements: ENAMEX,

TIMEX, and NUMEX. The subcategorization is captured by the SGML tag attribute called TYPE, which is defined to have a different set of possible values for each tag element.

** Format of Examples in this Document

Examples in this document encompass both text (originally written) and speech (transcribed by humans). Generally each example is given in three lines. The first line is an un-annotated text segment, together with a gloss. The second line is the text segment annotated with the abbreviated SGML tags, and the third line has the source of the segment, in square brackets.

* GUIDELINES FOR MARKUP OF EXCEPTIONAL CONSTRUCTIONS

** Conjunction

Conjoined named entities in general are marked separately. A taggable expression containing conjoined modifiers should still be marked up as a single expression, even if the modifiers could be tagged otherwise (see "Nested Expressions" below).

邮电部 (posts telecommunications ministry)
<ORG>邮电部</ORG>
[MET-2 Guidelines]

国防科学技术工业委员会 (state commission of science, technology, and industry for national defense)
<ORG>国防科学技术工业委员会</ORG>
[MET-2 Guidelines]

担任过工商部长, 教育和文化部长等职 (has served as industry and commerce minister, (and as) education and culture minister)
担任过 <ORG>工商部</ORG>长, <ORG>教育和文化部</ORG>长等职
[mc970421]

A conjoined multi-name, multi-number or multiple-time expression should be marked up as separate expressions (even if there is elision of the head of one conjunct).

中南美 (central and south america, [i.e. "Central America and South America"])
<LOC>中</LOC><LOC>南美</LOC>
[MET-2 Guidelines]

我外交以及防务部门今天重复证实 (our diplomacy and defense departments today both confirmed...)
我 <ORG>外交</ORG> 以及 <ORG>防务部</ORG>门今天重复证实
[mk970001]

An example involving time expressions:

凌晨三四点钟 (AM 3 (or) 4 o'clock)
<TME>凌晨三</TME><TME>四点钟</TME>
[MET-2 Guidelines]

每周一二四六 (every monday, tuesday, thursday and saturday)
每<DTE>周一</DTE> <DTE>二</DTE> <DTE>四</DTE> <DTE>六</DTE>
[mc970114]

If the head of the conjuncts is separated from the conjuncts by the number of conjuncts, mark each conjunct and the number separately, but do not mark the head.

俄哈吉塔四国 (russia, kazakhstan, kyrgyzstan, tajikistan four countries)
<LOC>俄</LOC><LOC>哈</LOC><LOC>吉</LOC><LOC>塔</LOC><CRD>四</CRD>国
[mc970411]

东西两半球出现满月的时间不一样 (east west two hemispheres manifest full moon times are not the same)
<LOC>东</LOC><LOC>西</LOC><CRD>两</CRD>半球...
[mk970005]

美国国会参众两院 (u.s. congress "deliberation" "masses" two houses, i.e., Senate and House)
<ORG>美国国会</ORG><ORG>参</ORG><ORG>众</ORG><CRD>两</CRD>院

[h4_ma98-v2.utf]

解放军陆海空三军 (PLA land sea air three forces, [i.e. the army, navy and air forces of the PLA])
<ORG>解放军</ORG><ORG>陆</ORG><ORG>海</ORG><ORG>空</ORG><CRD>三</CRD>军
[h4_ma98-v2.utf]

** Numeric Range Expressions

The "endpoints" of time, date, and numeric range expressions should be marked up separately. This applies to both TIMEX and NUMEX expressions.

1 9 9 3 年 1 0 月至 1 1 月 (1993 october to november)
<DTE>1 9 9 3 年 1 0 月</DTE>至<DTE>1 1 月</DTE>
[MET-2 Guidelines]

零 下 一 到 十 四 度 (one below zero to fourteen degrees)
<MSR>零 下 一</MSR> 到 <MSR>十 四 度</MSR>
[mc970421]

** Tokenization Conventions

The systems must incorporate certain tokenization conventions.

*** SGML Short References

In speech transcriptions, some SGML annotations and shortrefs are inserted during the transcription process. Except for the shortrefs "." (period), "," (comma), and "?" (qmark), these annotations should be included within the named entity markup wherever they occur within a named entity or there is no white space to separate them from the beginning or end of the name, or if they enclose the beginning or end of the name; otherwise, they should be left outside of the named entity markup.

我 是 ^孙 ^承, (I am Sun Cheng)
我 是 <PER>^孙 ^承</PER>,
[mv970625c]

*** Segmentation of Chinese Text

The transcribers' use of whitespace in their segmentation of Mandarin transcripts should not be taken into account when determining where to place SGML tags. The whitespace-separated "words" may be split with the tags.

延 长 对 <LOC>华</LOC> 最 惠 国 待 遇 (...extending MFN treatment to china)
延 长 对 <LOC>华</LOC> 待 遇
[mc970114]

*** Foreign Terms

Mandarin broadcasts may include foreign terms in the running transcript. Include the foreign terms in tagged phrases, if necessary. (In these examples, the SGML tag <f> is an abbreviation of the longer <b_foreign type="English">.)

<f> American </f> 航 空 公 司 (American aviation company, [i.e., American Airlines])
<ORG><f> American </f> 航 空 公 司</ORG>
[mk970002]

If a foreign name is given together with its Chinese equivalent, tag the two forms separately.

我 是 ^李 ^莉, <f>Lily,</f> (i am lili, Lily)
我 是 <PER>^李 ^莉</PER>, <f><PER>Lily</PER>,</f>
[mk970001]

*** Whitespace and Punctuation

No whitespace or punctuation characters should be included between the inserted begin tags and the first character of the entity, or between the inserted end tags and the last character of the entity. For example, if the text is:

在

<b_foreign language="English">
Stanton
<e_foreign>
市

(in Stanton city)

then the tags should be placed just before "Stanton" and just after 市,
as follows:

在
<b_foreign language="English">
<LOC>Stanton
<e_foreign>
市<LOC>

** Nested Expressions

If one taggable entity contains other taggable entities, the embedded entities are not to be marked.

华沙条约组织 (warsaw pact organization)
<ORG>华沙条约组织</ORG>
[MET-2 Guidelines]

太平洋 亚洲 旅行 协会 (pacific asia travel association)
<ORG>太平洋 亚洲 旅行 协会</ORG>
[mc970421]

See also "Non-decomposable Names", below.

* ENAMEX Specific Guidelines

** Guidelines That Pertain to ALL ENAMEX Types (PERSON, LOCATION, ORGANIZATION)

*** Entity Expressions that Modify Non-tagglables

Modifiers in complex NPs are to be tagged when it is clear to the annotator from context or the annotator's world knowledge that the modifier is a taggable entity.

中国人民 (chinese people)
<LOC>中国</LOC>人民
[tag word for "China"]
[MET-2 guidelines]

中国政府 (chinese government)
<LOC>中国</LOC>政府
[Tag word for "China" ; do not tag "中国政府" as an organization.]
[MET-2 guidelines]

韩国型的轻水反应堆 (korean light-water nuclear reactor)
<LOC>韩国<LOC>型的轻水反应堆
[tag word for "Korea"]
[MET-2 guidelines]

雷锋精神 (lei feng spirit)
<PER>雷锋<PER>精神
[MET-2 guidelines]

巴尔干地区 (the balkan region)
<LOC>巴尔干</LOC>地区
[MET-2 guidelines]

前南地区 (territories of) the former yugoslavia
前<LOC>南</LOC>地区
[MET-2 guidelines]

非洲 维持 和平 部队 (african peacekeeping forces)
<LOC>非洲</LOC> 维持 和平 部队
[mv970626a]

美国小姐 (american girls (not "Miss America"; if
the meaning were Miss America, would not
be taggable))
<LOC>美国</LOC>小姐
[MET-2 Guidelines]

*** Entity-Entity Modification With a "Possessive Particle"

When one taggable expression modifies a taggable expression immediately following it, and the two expressions are separated by a particle such as "的", they should be tagged separately.

美国的纽约 (u.s.'s new york)
<LOC>美国</LOC>的<LOC>纽约</LOC>
[MET-2 Guidelines]

美国的理查德本森 (u.s.'s richard benson)
<LOC>美国</LOC>的<PER>理查德本森</PER>
[MET-2 Guidelines]

前往香港的洪都拉斯领事馆 (went to hong kong's honduras
consulate, [i.e. the Honduran
Consulate in Hong Kong])
前往<LOC>香港</LOC>的<LOC>洪都拉斯</LOC>领事馆
[mv970520a.utf]

*** Modification Without a Subordinate Particle

When one taggable expression modifies a taggable expression immediately following it, and the two expressions are not separated by a particle such as "的", they should be tagged as one entity.

日本足球协会 (japan soccer association)
<ORG>日本足球协会</ORG>
[MET-2 Guidelines]

世界足球协会 (world soccer association)
<ORG>世界足球协会</ORG>
[MET-2 Guidelines]

白宫玫瑰园 (white house rose garden)
<LOC>白宫玫瑰园</LOC>
[h4_ma98-v2.utf]

美国广播公司 (american broadcast corporation)
<ORG>美国广播公司</ORG>
[MET-2 Guidelines]

中共驻南京代表团 (chinese communist (party) representative to nanking)
<ORG>中共驻南京代表团</ORG>
[MET-2 Guidelines]

后来调任美国驻洪都拉斯大使馆 (then transferred to u.s. stationed at
honduras embassy)
后来调任<ORG>美国驻洪都拉斯大使馆</ORG>
[mv970520a]

美国探索电视网 (u.s. discovery network, [i.e., the
"Discovery Channel", of the United States])
<ORG>美国探索电视网</ORG>
[MET-2 Guidelines]

[Include the word for U.S. in the tag zone - even though it isn't really part of the proper name in English - because there is no marker clearly separating the location name from the rest of the string]

法国 国际 政治 杂志 (france international politics magazine)
<ORG>法国 国际 政治 杂志</ORG>
[mc970226.utf]

人数最多的美国国会众议院访华团
(largest u.s. congress house of representatives china tour group)
人数最多的 <ORG>美国国会众议院</ORG> 访<LOC>华</LOC>团
[mc970114]
[Don't separate the "U.S." from "Congress" or the "Congress" from
"House", but do separate the generic descriptor "China tour group"
from the rest]

*** Lexicalized Possessives

If the first person personal pronoun, 我, or 我国 ("my country"), is used as an alias for "China" to modify an ORGANIZATION name, treat the modifier (我 or 我国) as a lexically possessive, separable modifier. However, do NOT tag it (i.e. treat it as a possessive pronoun, not as

a named entity). Hence,

我国共产党 (my country's communist party)
<ORG>我国共产党</ORG>
[MET-2 Guidelines]

as opposed to

中国共产党 (the China Communist Party, [i.e. the Communist Party of China])
<ORG>中国共产党</ORG>
[MET-2 Guidelines]

*** Entity Expression Aliases

**** Taggable Aliases

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:

<f>PATA</f> (acronym for the Pacific Asia Travel Association)
<f><ORG>PATA</ORG></f>
[mc970421]

Nicknames and other aliases are tagged 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. Nicknames and other neologisms that are not commonly used to refer to an entity are not to be tagged (see "Miscellaneous Personal Non-tagables"). Taggable examples include:

沪 (alias for Shanghai)
<LOC>沪</LOC>
[MET-2 Guidelines]

华 (common alias for China/Chinese)
<LOC>华</LOC>
[usually marked up as a LOCATION]
[MET-2 Guidelines]

A non-tagable example is:

东方明珠 ("bright pearl of the orient", referring to Hong Kong)
<LOC>东方</LOC>明珠
[mc970209]
[Tag only the term for "Orient"]

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

港澳 (alias for Hong Kong & Macao)
<LOC>港</LOC><LOC>澳</LOC>
[MET-2 Guidelines]

北约 ("north-treaty", [i.e. NATO])
<ORG>北约</ORG>
[MET-2 Guidelines]

上轮集团 (abbr. of 上海轮胎集团 shanghai tire group)
<ORG>上轮集团</ORG>
[XIN19980423.0058]

朝鲜南北对话 (korea north-south dialogue, [i.e. dialogue/talks between North and South Korea])
<LOC>朝鲜</LOC>南北对话
[Do not tag "north" and "south" as aliases for N.Korea and S. Korea, since these words are abbreviations that can only be understood as "aliases" in this context, and can stand for any number of other countries or regions in other contexts.]
[MET-2 Guidelines]

阿以冲突 (the arab-israel conflict)
阿<LOC>以</LOC>冲突

[This illustrates the common tendency in Chinese to abbreviate country names into one-syllable words. Since "a" 阿 is not an alias for a specific country in this case, it is untagged. the character "ji" 以 is (among other things) an abbreviation-type alias for Israel 以色列. Although some such abbreviations may stand for more than one country name (e.g. 巴 = Palestine/Panama), they nevertheless are clearly proper nouns, unlike 南北 "north"/"south" as used in the above example.]
[MET-2 Guidelines]

前南地区 (the former "yugo" (south) territories, [i.e., territories of the former Yugoslavia]).

前<LOC>南</LOC>地区

[MET-2 guidelines]

[In this example, 南 is a proper noun.]

Certain metonyms, herein designated "proper" metonyms, which chiefly include references to an organization based on the name of a unique structure or facility in which the organization holds office. The association between the name and the organization should be idiosyncratic enough to justify its inclusion in a dictionary definition of the term (in contrast with "common" metonyms, discussed below), as a kind of nickname for the organization.

白宫 官员 表示 (white house officials announced)

<ORG>白宫</ORG> 官员 表示

[mv970520a]

Metonyms, herein designated "common" metonyms, that reference political, military, athletic, and other organizations by the name of a city, country, or other associated location. In these cases, the association between the name's semantic type and the organization is sufficiently predictable and non-idiosyncratic as to preclude a dictionary gloss; hence the name should be tagged as a LOCATION, not as an ORGANIZATION. Some examples of "common" metonyms follow.

美国 不会 将 汇率 作为 贸易 战的 武器. (u.s. would not use the exchange rate as a weapon in a trade war)

<LOC>美国</LOC> 不会 将 汇率 作为 贸易 战的 武器.

[mc970408]

广深铁路以及深圳发展银行部分高层也被免职 (some higher-ups at guangshen railroad and shenzhen development bank were also fired)

<LOC>广深铁路</LOC>以及<ORG>深圳发展银行</ORG>部分高层也被免职

[mv970620a]

*** Non-taggable Aliases

The following forms of entity names will NOT be tagged:

Common nouns, including pronouns, used in anaphoric reference to taggable entity names.

天鹅(Swan)印刷公司...该公司... (swan printing company ... the company...)

<ORG>天鹅(Swan)印刷公司</ORG>...该公司...

[ZBN19980314.0108]

Aliases that refer to broad industrial sectors, political power centers, etc., rather than to specific organizations.

四小龙 ("four little dragons", Taiwan, South Korea, Singapore, and Hong Kong)

[no markup]

[mv970616c.utf]

** Quotation Marks Around an Entity Name

Quotes are included in the tag if they appear within an entity's name, but not if they bound the name.

《星岛日报》的社论说 ("star island daily news" editorial said...)

《<ORG>星岛日报</ORG>》的社论说

[TDT training data, XIN19980101.0052]

** Non-decomposable Non-taggable Names

Complex non-taggable entities that are not to be marked (because the

whole name does not refer to a currently recognized ENAMEX entity)
are not decomposable.

巴拿马运河条约 (panama canal treaty)
[No markup]
[MET-2 Guidelines]

香港脚 ("hong kong foot", an affliction similar to athlete's foot)
[No markup]
[MET-2 Guidelines]

美国小姐 (miss america (not "american girls"))
[no markup]
[MET-2 Guidelines]

第四十六届太平洋亚洲旅行协会年会 (forty-sixth pacific asia
travel association annual
meeting)
[No markup]
[mc970421]

六二八庆回归告别殖民晚会 (the six-two-eight celebrate return (of
hong kong) say goodbye to colonialism
soiree)
[No markup]
[mv970626a]

基督徒 (christ disciple, [i.e., Christian])
[No markup]
[mk970001]

佛教信徒不比基督教少啦 (followers of buddha-religion are not less
numerous than those of christ-religion)
[no markup]
[mk970002.utf]

里氏六点二级. (richter 6.2 "scales")
里氏 <MSR>六点二级</MSR>.
[mv970626c]

复活节临近了,专家们呼吁人们要注意沙门氏杆菌 (as easter approaches,
specialists warn people to
beware of salmon's
bacillus, [i.e.,
Salmonella])
<DTE>复活节</DTE>临近了,专家们呼吁人们要注意沙门氏杆菌

马克思主义 (marx ideology, [i.e., Marxism])
[no markup]
[MET-2 Guidelines]

毛泽东思想 (mao zedong thought)
[No markup]
[MET-2 Guidelines]

邓小平理论 (deng xiaoping theory)
[no markup]
[XIN19980312.0067]

阿伏伽罗定律 (avogadro's law)
[no markup]
[MET-2 Guidelines]

邓小平一片的播出 (deng xiaoping (CL-for-film)'s release, [i.e., the
release of the film "Deng Xiaoping"])
[No markup]
[mc970114]

** Miscellaneous Non-tagables

*** Figures of Speech

Figures of speech include expressions such as metaphors or similes or
devices such as personification or hyperbole. Such expressions which
use an otherwise taggable ENAMEX expression are not to be tagged.

扎伊尔会出现第二个蒙博托 (zaire could manifest a second mobutu)
<LOC>扎伊尔</LOC>会出现第二个蒙博托
[mv970520a.utf]

[tag Zaire, but not Mobuto]

我们常说顾客就是上帝 (we often say the customer is god)
[no markup. See also "Saints and other Religious Figures"]
[mv970618a]

冰雪童话世界 (ice and snow fairy tale world)
[no markup]
[mc970220]

*** Non-taggable Proper Names

Miscellaneous types of proper names that are not to be tagged as ENAMEX include names of events, media (such as TV and radio shows, movies, and books), products and treaties. (For information on the treatment of facilities, see "ORGANIZATION-related Facilities" below.)

参加奥林匹克的足球比赛 (participate in the olympic soccer competitions)
[no markup]
[mv970521b]

水门丑闻 (watergate scandal)
[no markup]
[mv970618a]

二战 (WWII)
[no markup]
mc970114.utf

香港百题第三十集今天为您解答
(Today the thirtieth installment of "Hong Kong 100 Topics" explains to you...)
[no markup]
[mc970421]

一本名为天怒的小说 (a novel entitled heaven's wrath)
[no markup]
[mv970616c.utf]

** Guidelines That Pertain Only to PERSON

This entity type includes not only humans both "real" and "fictional," but also other fictional or real anthropomorphous entities.

*** Titles of PERSONS

**** Titles vs. Generational Designators

Titles and role names are not considered part of a persons name.

^奥尔布莱特 国务卿 (albright state-minister)
<PER>^奥尔布莱特</PER> 国务卿
[mv970620a]

However, generational designators are considered part of a person name.

十四世达赖丹增加措 (fourteenth dalai tenzin gyatso)
<PER>十四世达赖丹增加措</PER>
[XIN19980312.0070]

英国女王伊丽莎白二世 (england's queen elizabeth II)
<LOC>英国</LOC>女王<PER>伊丽莎白二世</PER>
[MET-2 Guidelines]

When a person's title falls between the surname and the given name, include the title.

^李 主席 ^登辉 先生 (li chairman deng-hui mister)
<PER>^李 主席 ^登辉</PER> 先生
[mk970001]

Include the 老 in 老子, "Laozi".
记住老子的话 (remember well laozi's words...)
记住<PER>老子</PER>的话
[mv970521]

**** Entities that Modify Persons/Titles

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

英国女王伊丽莎白 (queen elizabeth of england)
<LOC>英国</LOC>女王<PER>伊丽莎白</PER>
[MET-2 Guidelines]

**** Entities that Modify Organization Chief Titles

Often, terms for titles of the form

<name>-<type> <type>-长, for example 国防部部长

are shortened to

<name>-<type>-长, for example 国防部部长

Decomposition of the first form is straightforward, and is shown in the following examples.

宣传部 部长 ^孙 ^南生 (propaganda ministry ministry-chief sun nansheng)
<ORG>宣传部</ORG> 部长 ^孙 ^南生
[mc970215]

朝鲜人民武装力量部副部长 (korea peoples' armed forces ministry vice ministry-chief)
<ORG>朝鲜人民武装力量部</ORG>副部长

However, the second form brings up a separation issue. It could be argued that the second form is either non-separable or should be separated between the <name> and the <type>. This can be seen in the way the phrases are segmented in the speech transcripts, and by the placement of the word for "assistant" or "vice-" in the following example.

马来西亚副财政部长 (malaysia vice finance ministry-chief)
<LOC>马来西亚</LOC>副<ORG>财政部</ORG>长
[ZBN19980314.0106]

Nonetheless, if there are no intervening words, mark the shortened titles as follows:

国防 部长 ^迟 ^浩田 (defense ministry-chief chi haotian)
<ORG>国防 部</ORG>长 <PER>^迟 ^浩田</PER>
[mc970114]

美国 国防 部长 ^佩里 (u.s. defense department-chief perry)
<ORG>美国 国防 部</ORG>长 <PER>^佩里</PER>
[mc970114]

*** Family Entity Expressions

Family names are to be tagged as PERSON.

蒋氏父子 (the jiang family, father and son)
<PER>蒋</PER>氏父子
[MET-2 Guidelines]

西迪兄弟 ((the) xidi brothers)
<PER>西迪</PER>兄弟
[ZBN19980314.0108]

*** Names of Animals

Names of animals and other non-human characters are to be tagged as PERSON. (No examples found for Mandarin)

*** Saints and other Religious Figures

Although religious titles or specifiers such as "saint," "prophet," "imam," or "archangel" are not be tagged, the proper name will be tagged as a PERSON. This practice becomes more significant in marking up speech transcriptions, due to peculiarities of speech habits or patterns.

^达赖喇嘛 (dalai lama)
<PER>^达赖</PER>喇嘛
[mk970002]

References to God will be taken to be the "name" of this entity for tagging purposes.

上帝似乎已经治好了她丈夫的病 (god seemingly has already cured her husband's illness)
<PER>上帝</PER> 似乎已经治好了她丈夫的病
[mk970010]

The term 天 (Heaven) in the phrase 老天 (Old Heaven) should be tagged as PERSON. It is not always clear whether "heaven" is the name of a person, the name of a place (or, through metonymy, of an organization) or simply a descriptor, meaning "sky". When the respectful title 老 "Old" is used, though, the entity being referred to is clearly being regarded as person-like.

老天原来给我们的这些个波 (these visions that old heaven originally gave us)
老<PER>天</PER> 原来给我们的这些个波
[mk970003]

Other uses of 天 should not be tagged.

山河因之动容,天地为之痛哭 (mountains and rivers are visibly moved, heaven and earth cry bitterly (for the death of Deng Xiaoping))

[No markup]
[mc970226]

*** Fictional Characters

Names of fictional characters are to be tagged.

*** Fictional Animals and Non-human Characters

Fictional animals should be tagged as PERSON.

*** Miscellaneous Personal Non-tagables

Miscellaneous types of proper names that are not to be tagged as PERSON include: individuals identified by their political affiliation, laws named after people, court cases named after people, weather formations, and diseases/ prizes named after people.

和阿尔兹海默氏的 zh- 危险 (and the danger of alzheimers)
[no markup]
[mk970009.utf]

Nicknames and other neologisms that are not commonly used to refer to an entity are not to be tagged. Nicknames and other aliases are tagged only when they are established ways of referring to an entity (see "Entity Expression Aliases").

被称为“改革开放总设计师”的邓小平 (deng xiaoping, who has been called "the architect of reform and openness")

被称为“改革开放总设计师”的<PER>邓小平</PER>
[XIN19980219.0075]

** Guidelines That Pertain Only to LOCATION

The TYPE LOCATION applies to entities representing either geographical, political, or astronomical locations. Examples of strings that are tagged as LOCATION include: named heavenly bodies, continents, countries, provinces, counties, cities, regions, districts, towns, villages, neighborhoods, airports, military bases, railways, railroads, highways, bridges, street names, street addresses, oceans, seas, straits, bays, channels, sounds, rivers, islands, lakes, national parks, mountains, fictional or mythical locations, and certain structures, such as the Eiffel Tower and Washington Monument, that were built primarily as monuments.

在人民大会堂集会 (gather in the great hall of the people)
在 <LOC>人民大会堂</LOC> 集会
[mc970226]

汉江上的圣水大桥 (the songsu bridge spanning the han river)
<LOC>汉江</LOC>上的<LOC>圣水大桥</LOC>
[MET-2 Dry-run data, METID 005]

新亚欧大陆桥 (new asia europe land bridge)
<LOC>新亚欧大陆桥</LOC>
[mc970114]

震中位于北纬三十六点儿零度,东经九十五点儿九度 (epicenter located at
north 36.0 degrees
east 95.9 degrees)
震中位于<LOC>北纬三十六点儿零度,东经九十五点儿九度</LOC>

九泉之下,好好地安息 (under the nine springs rest well in peace)
<LOC>九泉</LOC>之下,好好地安息
[mc970226.utf]

*** Possible Embedded Locative Entity-Strings

A location name immediately preceding an organization name may or may not be part of the organization name proper. The annotation in the answer key will follow these guidelines:

(1) If the organization name already begins with a place name, tag the preceding place name and organization name separately:

日本^关西经济联合会 (japan kansai economic association)
<LOC>日本</LOC> <ORG>^关西经济联合会</ORG>
[mc970403.utf]

(2) If the organization name does not begin with a place name, the place name should be tagged as part of the organization name.

包括美国杜邦,荷兰飞利浦、日本松下、索尼、三菱等 (including
u.s. dupont, netherlands phillips, japan panasonic, sony, mitsubishi
etc.)
包括<ORG>美国杜邦</ORG>, <ORG>荷兰飞利浦</ORG>、
<ORG>日本松下</ORG>、<ORG>索尼</ORG>、<ORG>三菱</ORG>等
[MET-2 dry-run data, METID 002]

[see also "Entity-Strings Embedded in Entity Expressions", above]

*** Locative Entity Expressions Tagged in Succession

Compound expressions in which place names are listed in succession are to be tagged as separate instances of LOCATION.

中国广东
<LOC>中国</LOC><LOC>广东</LOC>
[MET-2 Guidelines]

科伊边境 (kuwait-iraq border (country names abbreviated))
<LOC>科</LOC><LOC>伊</LOC>边境
[MET-2 Guidelines]

吉林省^延边朝鲜族自治州^图门市 (jilin province yanbian korean
autonomous region tumen
municipality)
<LOC>吉林省</LOC> <LOC>^延边朝鲜族自治州</LOC> <LOC>^图门市</LOC>
[mc970209.utf]

*** Locative Designators and Specifiers

Designators that are integrally associated with a place name are to be tagged as part of the name. 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.

约旦河 (jordan river)
<LOC>约旦河</LOC>
[MET-2 Guidelines]

巴拿马城 (panama city)
<LOC>巴拿马城</LOC>
[MET-2 Guidelines]

北京市 (beijing city / municipal area)
<LOC>北京市</LOC>
[MET-2 Guidelines]

马里兰州 (maryland state)
<LOC>马里兰州</LOC>

[not <LOC>马里兰</LOC>州]
[MET-2 Guidelines]

朝鲜半岛 (korea peninsula)
<LOC>朝鲜半岛</LOC>
[MET-2 Guidelines]

长江 流域 (yangtze basin)
<LOC>长江 流域</LOC>
[mc970305]

长江三角洲 (yangtze delta)
<LOC>长江三角洲</LOC>
[MET-2 guidelines]

If the extent of the location name cannot be determined from world knowledge or from the appropriate reference, then the common-noun designator will be included in the location name.

**** Locative Non-tagables: The Postposed Partitive Specifier

Do not include in the tagged string common noun phrases functioning as partitive-type locative specifiers directly after LOCATION names.

密西西比河西岸 (mississippi river west bank)
<LOC>密西西比河</LOC>西岸
[MET-2 Guidelines]

Do not tag internet addresses as LOCATION. Also, do not tag any otherwise-tagable entities that are within internet addresses.

一个 叫做 天堂 之 门 的 网址 (a website called heaven's gate)
[no markup]
[mk970009.utf]

地址的话,我觉得好像是,它是这样的,它是 ETS, dot, ORG.
(as for the address, I think it's something like, it's like this, it's ETS, dot, ORG)
[no markup]
[mv970521]

**** Exceptional Locative Specifiers Used as Entity Expressions

Include partitive-type locative specifiers when they are part of an established name.

约旦河西岸 (jordan river west bank)
<LOC>约旦河西岸</LOC>
[XIN19980118.0045]

台湾 海峡 两岸 (taiwan strait two banks - refers to the PRC and ROC together)
<LOC>台湾 海峡 两岸</LOC>
[mv970520a]

海峡两岸 (strait two banks)
<LOC>海峡两岸</LOC>
[mk970004]

两岸 (two banks)
<LOC>两岸</LOC>
[mk970004]

The term "mainland" 大陆 is frequently used as an alias for "The Peoples Republic of China", to distinguish it from the Republic of China on Taiwan. Tag as follows:

大陆
<LOC>大陆</LOC>
[MET-2 Guidelines]

中国大陆
<LOC>中国大陆</LOC>
[MET-2 Guidelines]

Do not tag "zuguo" 祖国 (ancestor/father/mother-land) as a location.

祖国大陆

祖国<LOC>大陆</LOC>
[XIN19980112.0088]

Do not tag "neidi" 内地 (interior) as a location, even when it is used to distinguish the PRC from Hong Kong.

来自内地和香港特区以及澳门 (coming from interior and hong kong special region as well as macao)
来自内地和<LOC>香港特区</LOC>以及<LOC>澳门</LOC>
[h4_ma98-v2.utf]

*** Transnational and Subnational Region Names

**** Transnational Locative Entity Expressions

Tag names of continents ("Africa"), multi-country sub-continental regions ("Eastern Europe," "Sub-Saharan Africa"), and multi-country trans-continental regions ("the Middle East," "the Pacific Rim").

西非国家领导人 (west africa nation leaders)
<LOC>西非</LOC>国家领导人
[mv970521]

**** Subnational Region Names

Do not tag names of sub-national regions when referenced only by compass-point modifiers. Do not tag "the South" or the "mid-West," analogies to "the Middle East" notwithstanding, because, unlike the latter term, their referential value varies from country to country. For example,

使西南地区的客运 (causing the southwest region's passenger service...)
[no markup]
[mc970331]

Do tag names of sub-national regions if they are identifiable even when the name is disassociated from context. English examples include "the Ruhr," "the Rockies," "the Auvergne," and "Amazonia." Note that these names generally straddle, or lie within, geo-political jurisdictions such as states or provinces.

西北五省区 (the northwest five province region)
<LOC>西北五省区</LOC>
[mc970114]

华南 (south china, an established region name in Chinese)
<LOC>华南</LOC>
[mc970411.utf]

华北 (north china)
<LOC>华北</LOC>
[mc970411.utf]

*** Time Modifiers of Locative Entity Expressions

Historic-time modifiers ("former," "present-day") are not to be included in tagged expressions. These are used as ad hoc modifiers that are readily separable from the name.

前南 (the former yugoslavia)
前<LOC>南</LOC>
[MET-2 Guidelines]

*** Space Modifiers of Locative Entity Expressions

Directional modifiers ("north," "south," "east," "west," "upper," "lower," and combinations thereof) are taggable only when they are intrinsic parts of a location's established name.

南斯拉夫 (yugoslavia)
<LOC>南斯拉夫</LOC>
[mv970521]

北爱尔兰 (north ireland)
<LOC>北爱尔兰</LOC>
[XIN19980115.0055]

湖北 (hubei (province))

<LOC>湖北</LOC>
[mc970127]

中 西伯利亚 (central siberia)
中 <LOC>西伯利亚</LOC>
[mc970403.utf]

If it is not possible to determine either through separability, world knowledge, or an appropriate reference, then the directional modifier should not be included.

*** Miscellaneous Locative Non-tagables

**** Locations Within Names of Languages

Do not tag the names of locations which are in language names of the form X-语 or X-文, where X is a location.

英语 (england language, [i.e. English])
<ONA>英语</ONA>
[mc970415]

中文 (china language)
<ONA>中文</ONA>
[mc970415]

对 西藏 地区的 藏语 广播. (to tibet region tibetan language
broadcasts)
对 <LOC>西藏</LOC> 地区的 <ONA>藏语</ONA> 广播.
[mc970421]

主张 台语 在台 (advocate taiwan language in taiwan)
主张 <ONA>台语</ONA> 在 <LOC>台</LOC>
[mk970010]

汉语 (han language, [i.e. Chinese])
<ONA>汉语</ONA>
[mc970114]

Do tag the location names of the form X-话, where X is a location.

用 四川 话 (using sichuan words)
用 <LOC>四川</LOC> 话
[mc970403]

**** Locations Within Names of Ethnic Groups

Do not tag location names which are part of the names, ending in 族 or 裔, of ethnic groups.

目的是促进塞浦路斯西族与土族的和解 (the intent was to promote peace
and understanding between
cyprus greece-ethnic-group and
turkey-ethnic-group)
目的是促进<LOC>塞浦路斯</LOC><ONA>西族</ONA>与<ONA>土族</ONA>的和解
[mv970520b]

Do not tag the 华 ("China") in 华侨 "Chinese sojourners", persons of Chinese extraction living abroad.

美国华侨 (u.s. persons of chinese extraction, [i.e. Chinese Americans])
<LOC>美国</LOC><ONA>华侨</ONA>
[tag words for "U.S."]
[MET-2 guidelines]

** Guidelines That Pertain Only to ORGANIZATION

*** Corporate or Organization Designators

Corporate or organization designators are considered part of an organization name.

西北航空公司 (northwest airlines corporation)
<ORG>西北航空公司</ORG>
[MET-2 Guidelines]

攻破了波兰队的球门 (breached poland team's goal posts)
攻破了<ORG>波兰队</ORG>的球门

[XIN19980423.0051]

[With sports teams, treat 队 "team" as a corporate designator]

*** Miscellaneous ORGANIZATION-type Entity Expressions

Proper names that are to be tagged as ORGANIZATION include stock exchanges, multinational organizations, businesses, TV or radio stations, political parties, religious groups, orchestras, bands, or musical groups, unions, non-generic governmental entity names such as "congress" or "chamber of deputies," sports teams and armies (unless designated only by country names, which are tagged as LOCATION), as well as fictional organizations (to ensure consistency with marking other fictional entities).

美国海军 (u.s. navy)

<ORG>美国海军</ORG>

[MET-2 Guidelines]

欧共体 (european community)

<ORG>欧共体</ORG>

[MET-2 Guidelines]

生育委员会 (birth control commission)

<ORG>生育委员会</ORG>

[MET-2 Guidelines]

中国奥林匹克队 (china olympic team)

<ORG>中国奥林匹克队</ORG>

[MET-2 Guidelines]

披头四 ("mop-topped four", [i.e. the Beatles])

<ORG>披头四</ORG>

[MET-2 Guidelines]

飞虎队 (flying tigers)

<ORG>飞虎队</ORG>

[MET-2 Guidelines]

[the volunteer American group fighting in China in WWII]

敢死队 (suicide squad)

[no markup, because this is a generic, common noun designator applicable to many different groups]

[MET-2 Guidelines]

但是 共和党人 说

(but republican party people say...)

但是 </ORG>共和党</ORG> 人 说

[mk970001]

(sub-)Committees, delegations, working groups, etc. should be tagged as ORGANIZATIONS, provided they appear to have an institutional structure, or corporate/political objectives.

土耳其 议会 外交 关系 委员会 (turkey parliament diplomatic relations committee)

<ORG>土耳其 议会 外交 关系 委员会</ORG>

[mc970209]

终战 50 周年国会议员联盟 (alliance of [japanese] congressional representatives for the 50th anniversary of the end of world war II)

<ORG>终战 50 周年国会议员联盟</ORG>

**** Broadcasting Stations

Stations, channels, and frequencies are to be tagged as ORGANIZATION.

这是 中央台 报道 的. (this is the central broadcasting station

reporting)

这是 <ORG>中央台</ORG> 报道 的

[mc970114]

这里 是 <f>KAZN</f> <f>AM</f> 一 三 零 零 中文 广播 电台.

(this is KAZN AM one three zero zero chinese broadcast station)

这里 是 <ORG><f>KAZN</f> <f>AM</f> 一 三 零 零 中文 广播 电台</ORG>.

[mk970002]

**** Legislative Bodies

Although congresses can act as either organizations or events they

will be tagged as ORGANIZATION's.

在 国会 发表 国情 咨文 (delivered the state of the union address at
congress)
在 <ORG>国会</ORG> 发表 国情 咨文
[mc970209]

**** Event Organizers

Although event names are not to be tagged, even if they refer to events that occur on a regular basis and are associated with institutional structures, the institutional structures themselves - steering committees, etc. - should be tagged as ORGANIZATION.

**** ORGANIZATION-related Facilities

Proper names referring to meeting places or places where organizational activities occur (e.g., churches, embassies, factories, hospitals, hotels, museums, universities) will be tagged as ORGANIZATION.

北京大学 (beijing university)
<ORG>北京大学</ORG>
[MET-2 Guidelines]

燕京大学 (yanjing university (old name of the above, using antiquated alias
for Beijing))
<ORG>燕京大学</ORG>
[MET-2 Guidelines]

记者来到中山医科大学第一附属医院住院部 (reporter went to zhongshan
medical college number one associated hospital inpatient section)
记者来到<ORG>中山医科大学第一附属医院住院部</ORG>
[MET-2 Dry-run data, METID 010]

**** Treatment of '...军' (...army / ...military...)

The main distinction is between interpreting 军 as an adjective, similar to the English 'military' (i.e. 'not civilian') and interpreting 军 as an 'organization designator'. In order to get the latter interpretation, look for cases in which 军 is preceded by a service 'branch' designator (such as 空 'Air' as in 'Air Force')

一架美军飞机 (a u.s. military aircraft)
一架<LOC>美</LOC>军飞机
[MET-2 Guidelines -- GL-Clarification-Armies]

斯里兰卡空军 ('sri lanka air force')
<ORG>斯里兰卡空军</ORG>
[MET-2 Guidelines -- GL-Clarification-Armies]

In general, do not tag terms ending in 部队 "force" as ORGANIZATION.

西非维和部队 (west africa peacekeeping force)
<LOC>西非</LOC>维和部队
[XIN19980215.0031]

**** Embassies and Consulates

Names of embassies, consulates and other diplomatic missions should be marked as organizations only if both the country they represent and their location can be included in the markup.

后来调任美国驻洪都拉斯大使馆 (then transferred to u.s. stationed at
honduras embassy)
后来调任<ORG>美国驻洪都拉斯大使馆</ORG>
[mv970520a]

If both locations are not present, or are not contiguous within the embassy descriptor, mark any locations separately as LOCATION, and do not tag the embassy as an ORGANIZATION.
美国在通过驻金沙萨大使馆和其他正常渠道 (u.s., going through
stationed at kinshasa embassy
and other normal channels)
<LOC>美国</LOC>在通过驻<LOC>金沙萨</LOC>大使馆和其他正常渠道

*** Decomposable Product Names

In cases where the manufacturer and the product are named, the

manufacturer will be tagged. The product will not be tagged. Products must be defined loosely to include manufactured products (e.g., vehicles), as well as computed products (e.g., stock indexes) and media products (e.g., television shows).

道琼工业平均指数 (dow jones industrial average index)
<ORG>道琼</ORG> 工业平均指数
[mk970001]

Note that only the manufacturer may be extracted from a product name. Other named entities may not.

*** Metonymy

When a publication, regardless of subject matter, "reports," "states," "claims," etc., it is acting as a news source (reporting information). Tag news sources (newspapers, radio and TV stations, and news journals) as ORGANIZATION even when they function as artifacts. This is done because the same name is often used to refer to both the publication and the publisher. To avoid having annotators make the distinction between usage as artifact or organization, both usages are tagged. Moreover, publications often function agentively as ORGs.

人民日报 (peoples' daily)
<ORG>人民日报</ORG>
[mk970006]

人民日报海外版第三版 (people's daily overseas edition page three)
<ORG>人民日报</ORG> 海外版第三版
[mc970114]

这是中央台报道的 (this is central station reporting)
这是 <ORG>中央台</ORG> 报道的
[mc970114.utf]

Note that TV stations differ from TV shows, the latter not being taggable:

边疆行摄制组 (border journeys production crew)
[no markup]
[mc970114]

Similar to publications, metonymy might occur with stock indices and stock markets. Tag these as ORGANIZATION even when they refer to numeric values published by or representing the organization. For example:

今天道琼虽然小跌了 (today even though dow jones fell a little...)
今天 <ORG>道琼</ORG> 虽然小跌了
[mk970001]

*** Generic ORGANIZATION-like Non-tagables

Generic entity names such as "the police" and "the government," are not to be tagged.

中国公安部门 (china public safety department(s))
<LOC>中国</LOC>公安部门
[MET-2 Dry-run data, METID 002]

Do not mark the term 中央 ("center") by itself as an ORGANIZATION.

在中央这个领导下 (under the uh leadership of the center)
[no markup]
[mc970226.utf]

However, do mark 党中央 as an organization.

以江泽民同志为核心的党中央 (party center, with comrade
jiang zeming as its nucleus)
以 <PER>江泽民</PER> 同志为核心的 <ORG>党中央</ORG>
[mc970226.utf]

Mark the more-or-less non-generic names of national legislative bodies and national departments or ministries, even if the name of the nation in question is not present.

当选国会議員 (elected congress member)
当选 <ORG>国会</ORG> 議員
[mc970218]

[said of Portugal's president]

* TIMEX: SPECIFIC GUIDELINES

** Introduction

Only "absolute" time expressions are to be tagged. The TIME type is defined as a temporal unit shorter than a full day, such as second, minute, or hour. The DATE sub-type is a temporal unit of a full day or longer. An additional TIMEX type is DURATION, which captures durations of time.

** Absolute Temporal Expressions - TIME & DATE

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

An expression of minutes must indicate a particular minute and hour.

下午 五点 十分 (pm five oclock ten minutes)

<TME>下午 五点 十分</TME>

[mc970408]

An expression of hours must indicate a particular hour.

早晨 六点 (am six oclock)

<TME>早晨 六点</TME>

[mc970208]

An expression of days must indicate a particular day.

一九九九年 十二月 三十 号 (1999 december thirtieth)

<DTE>一九九九年 十二月 三十 号</DTE>

[mc9703011]

The expressions 上/中/下旬 (first/middle/last 10 days (of a month)) should be treated as absolute time expressions and incorporated into the tagged string:

五月上旬 (may first 10-day-period)

<DTE>五月上旬</DTE>

[MET-2 Guidelines]

An expression of seasons must indicate a particular season.

南极的夏季 (south pole's summer)

<LOC>南极</LOC>的<DTE>夏季</DTE>

[MET-2 Guidelines]

夏 秋 之间 (between summer and autumn)

<DTE>夏</DTE> <DTE>秋</DTE> 之间

[mc970127]

An expression of financial quarters or halves of the year must indicate which quarter or half.

今年上半年的成长 (this year first half growth)

今年<DTE>上半年</DTE>的成长

[ZBN19980226.0003]

在这个 第一 季 结束 以前 (before this first quarter concludes...)

在这个 <DTE>第一 季</DTE> 结束 以前

[mk970001]

An expression of years must indicate a particular year.

一九三六年 (nineteen thirty-six)

<DTE>一九三六年</DTE>

[mc970421]

An expression of decades must indicate a particular decade.

八十年 代 (the eighties)

<DTE>八十年 代</DTE>

[mc970408]

An expression of centuries must indicate a particular century.

十五 世纪 (fifteenth century)

<DTE>十五 世纪</DTE>

[mv970617b]

Temporal expressions are to be tagged as a single item. Contiguous subparts (month/day/year) are not to be separately tagged unless they are taggable expressions of two distinct TIMEX sub-types (date followed by time or time followed by date).

四月 一号 凌晨 四点 三十分 (april first am four thirty)
<DTE>四月 一号</DTE> </TME>凌晨 四点 三十分</TME>
[mc970407]

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.

大约 下午 三点 三十分 (about pm three thirty)
大约 <TME>下午 三点 三十分</TME>
[mc970408]

** Scope of Temporal Expressions

Absolute time expressions combining numerals, time-of-day designators (上午, 中午, 下午, 凌晨, etc.), or other subparts associated with a single TIMEX sub-type, are to be tagged as a single item.

下午 当地 时间 四点 三十一 分 (pm local time four thirty-one)
<TME>下午 当地 时间 四点 三十一 分</TME>
[mc970301]

早上 七点 半 (morning seven oclock (and a) half, [i.e., 7:30 am])
<TME>早上 七点 半</TME>
mc970331

However, time-of-day designators will not be tagged if they stand alone.

今天 早上 一 辆 冷冻车 离开 现场 (this morning a freezer truck left the site)
[no markup]
mk970009

When a time or date unit is expressed as an absolute expression, yet is not anchored, the time expression will be marked.

北京 在 二十三号 (on the twenty-third, beijing...)
<LOC>北京</LOC> 在 <DTE>二十三号</DTE>
[mv970626a]

** Temporal Expressions Containing Adjacent Absolute and Relative Strings

When a time expression contains both relative and absolute elements, only the absolute expression is to be tagged.

本 世纪 六 七 十 年 代 (this century sixties and seventies decades)
本 世纪 <DTE>六</DTE> <DTE>七 十 年 代</DTE>
[mc970421]

下 月 中 旬 (next month middle 10-day-period)
下 月 <DTE>中 旬</DTE>
[mk970001]

在 下 周 二 (on next week tuesday)
在 下 <DTE>周 二</DTE>
[mk970003]

** Holidays and other Named Dates

Dates that are referenced by name, such as holidays and historic or pre-historic periods, should be tagged.

春 节 (spring festival, [i.e., Chinese New Year])
<DTE>春 节</DTE>
[MET-2 Guidelines]

中 秋 时 节 (mid-autumn festival time)
<DTE>中 秋</DTE> 时 节
[mc970114]

中国汉代 (china's han period)
<LOC>中国</LOC><DTE>汉代</DTE>
[MET-2 Guidelines]

旧石器时代中期古人 (paleolithic mid-period man)
<DTE>旧石器时代</DTE> 中期古人
[mv970626a]

Uncommon names for a time period should not be tagged.

中国旅游年 ("year of china tourism", referring to 1997)
<LOC>中国</LOC>旅游年
[mc970209]

所谓的黑色的星期一 (so called black monday)
所谓的黑色的<DTE>星期一</DTE>
[mv970620a.utf]

** Locative Entity-Strings Embedded in Temporal Expressions

Multiword strings that are to be tagged as TIMEX will sometimes contain LOCATION (ENAMEX) substrings. Include these words within the scope of the tagged expression, but do not apply an embedded LOCATION tag.

美国东部时间二月四号晚上 (u.s. eastern time february fourth evening)
<DTE>美国东部时间二月四号</DTE> 晚上
[mc970209]

Sometimes, however, the phrasing is such that the modification and types are non-contiguously arranged as in "Japan time, 19 February, 8:00 A.M." but marking three items separately does not represent the modification accurately. In such cases, mark the entire phrase as a single temporal expression as shown in the following:

北京时间一九九七年二月九号十九点二十八分 (beijing time 1997
february ninth
nineteen hours
twenty-eight minutes)
<TME>北京时间一九九七年二月九号十九点二十八分</TME>
[mc970220]

** Temporal Expressions Based on Alternate Calendars

Temporal expressions in terms of alternate calendars, such as fiscal years, Chinese "five-year plans", the Chinese lunar calendar, etc., will generally be marked up.

提交了1991财政年度预算 (submitted a budget for the 1991 fiscal year)
提交了<DTE>1991财政年度</DTE>预算
[MET-2 Guidelines]

已由"六五"末的百分之八十五点九下降到百分之八十 ("...has already
dropped to 80% since the end of the 6th 5-year plan, when it was
85.9%...")
已由"<DTE>六五</DTE>"末的<PER>百分之八十五点九</PER>下降到<PER>百分之
八十</PER>
[MET-2 Guidelines]

When two alternate ways of expressing the same date or time are given,
tag both expressions separately.

今天是二月二十六号,星期三 (today is february twenty-sixth,
Wednesday)
今天是<DTE>二月二十六号</DTE>, <DTE>星期三</DTE>

各地穆斯林群众喜过肉孜节,开斋节 (everywhere muslim masses celebrate
roza festival, beginning of
"vegetarian" festival, [i.e.,
beginning of Ramadan])
各地穆斯林群众喜过<DTE>肉孜节</DTE>,<DTE>开斋节</DTE>
[mc970209]

Mark the two forms separately even if there are no intervening words
or punctuation.

今天是二月九号农历大年初三 (today is feb 9, lunar calendar "big

今天 是 <DTE>二月 九号</DTE> <DTE>农历 大年 初三</DTE>
[mc970209]

** Times In Sporting Events

For ordinal expressions describing the time in a sporting event, include the ordinal expression, and the period of the sporting event, if it is present.

下半场第六十八分钟 (sixty-eighth minute of the second half)
<TME>下半场第六十八分钟</TME>
[XIN19980423.0053]

在第八分钟首开纪录 (started things off in the eighth minute)
在<F>第八分钟</F>首开纪录
[XIN19980312.0091]

** Non-taggable temporal expressions

*** Anniversaries

Do not tag phrases of the form XX-周年, often translated as "XX-th anniversary"

为 纪念 中国 人民 解放军 建军 七十 周年
to commemorate china people liberation army establish seventy full years [i.e., "to commemorate the seventieth anniversary of the establishment of the PLA"].
为 纪念 <ORG>中国 人民 解放军</ORG> 建军 七十 周年
[h4_ma98-v2.utf]
[Tag the term for the PLA as an organization, but do not tag the "seventy"]

*** Miscellaneous non-taggable temporal expressions

Do not tag the 春 ("spring") in 春联 ("spring couplets").

** DURATION

TIMEX of type DURATION expressions refer to periods of time. They must have the form "[numeral] [time-unit]" where "numeral" is defined as any whole number, fraction, or decimal. Both the number and the time-unit word are included in the scope of the tag. Here is sampling of DURATION expressions:

在 半年 时间 内 (within half a year)
在 <DUR>半年</DUR> 时间 内
[mc970127]

在 水门 丑闻 四 分之 一 世纪 之 际 发表 的 评论 (in the quarter century of discussions since the watergate scandal...)
在 水门 丑闻 <DUR>四 分之 一 世纪</DUR> 之 际 发表 的 评论
[mv970618a]

*** Non-taggable Durations

Generic periods of time, and those without numerals, are not taggable.

好 了, 等 一 会 儿 (ok, wait a while)
[no markup]
[mv970520b]

Durations expressed as orders of magnitude are not taggable.

*** Distinguishing DURATION from DATE/TIME

Absolute times and dates are marked with DATE and TIME types even if they are in a phrase which describes a duration.

二十 一 世 纪 (twenty-first century)
<DTE>二十 一 世 纪</DTE>
[mc970421]

*** Distinguishing TIMEX DURATION from NUMEX MEASURE

People's (or animals) ages, although related to time, are marked with NUMEX type MEASURE. See "Standard Measurement Units", below.

享年 八十九 岁. (dead at the age of eighty-nine)
享年 <MSR>八十九 岁</MSR>.
[mc970116]

* NUMEX SPECIFIC GUIDELINES

The NUMEX portion of the task captures a set of useful numeric expressions categorized by the following TYPES:

MONEY: monetary expression

MEASURE: standard numeric measurement phrases such as age, area, distance, energy, speed, temperature, volume, and weight, plus syntactically-defined measurement phrases

PERCENT: percentage (a fraction expressed in terms of hundredths)

CARDINAL: a numerical count or quantity of some object (in the form of whole numbers, decimals, or fractions)

Note that many of these types could be broken down into subtypes if desired by the end-user. For example, one could envision adding subtypes such as AGE and TEMPERATURE under MEASURE.

** Scope of Numeric Expressions

*** Negative Numbers

If the number in an expression is negative, include in the tagged phrase the part that indicates the number is negative.

北京 晴, 零下 五 到 六 摄氏度. (beijing clear, below zero five to six celsius degrees)
<LOC>北京</LOC> 晴, <MSR>零下 五</MSR> 到 <MSR>六 摄氏度</MSR>.
[mc970127]

*** Numeric Expressions Plus Units of Measure

The entire string expressing a monetary (MONEY) or percentage (PERCENT) value is to be tagged. This same guideline will apply to the MEASURE TYPE as well, so that the numeric value plus the unit of measure is included within the scope of the tag.

Furthermore, in the Mandarin annotation, classifiers will be included in the scope of the tag, since they are syntactically very similar to units of standard measure, as explained in the following quote:

In numeral classifier languages the numeral classifier construction is almost always identical with the measure construction included rules of word order. This is so much the case that many grammars of such languages consider the numeral classifier construction as merely a subvariety of an overall construction type which includes measures. For example, if the order is *five-flat object-book*, a classifier language will almost invariably have the order *five-pounds-cheese*.

((Greenberg, J.H. (1975). Dynamic Aspects of Word Order in the Numeral Classifier. In Charles Li (ed.) Word Order and Word Order Change, University of Texas Press, Austin, TX)

quoted in

(Chinchor, Nancy Ann (1982). Morphological Theory and Numeral Incorporation in American Sign Language. Thesis))

** Numeric Expressions Appearing in Succession

Juxtaposed strings expressing values in two different units of measure are to be tagged separately.

** Approximators and Multipliers in the Modification of Numeric Expressions

*** Approximators

Modifying words that indicate the approximate value of a number or a "relative position" to a number are generally to be excluded from the NUMEX tag if the modifier indicates only some minor imprecision in the known quantity (see also "Multipliers", below).

二十万吨级以上油轮 (200,000 ton over oil tankers)
<MSR>二十万吨级</MSR>以上油轮
[mc970331]

至今已经整整十五年. (to now already fully fifteen years)
至今已经整整 <DUR>十五年</DUR>.
[mc970302]

今早九点整到达北京站. (this morning at nine o'clock sharp arrived
at beijing station)
今早 <TME>九点</TME> 整到达 <LOC>北京站</LOC>.
[mc970331]

If a modifier occurs in the middle of an otherwise taggable numeric expression, the entire expression should be tagged, regardless of whether the modifier seems to be semantically "approximate" or "indefinite."

大约一百多名女学生 (about 100-plus CL female students)
大约 <CRD>一百多名</CRD> 女学生
[mk970005]

在晚上大约七时到达北京. (at pm about seven o'clock arrived beijing)
在 <TME>晚上大约七时</TME> 到达 <LOC>北京</LOC>.
[mk970006]

十几天前 (ten-some days previously)
<DUR>十几天</DUR> 前
[mc970209]

三十几年来 (in the past thirty-some years)
<DUR>三十几年</DUR> 来
[mc970209]

Numeric expressions which give order of magnitude information will be tagged.

亿万人民 (hundreds of millions of people)
<CRD>亿万</CRD> 人民
[mc970303]

*** Multipliers

Modifiers that indicate the multiplied value of a number unit should be included in the tagged string, if the modifier is a substitute for a specific digit within the numeric expression.

几万斤的粘豆包 (several ten-thousand "jin" of bean paste buns)
<MSR>几万斤</MSR> 的粘豆包
[mc970208]

几千万盆, (several ten-million flower-pot-CL)
<CRD>几千万盆</CRD>,
mc970215

几家工厂 (some CL factories)
<CRD>几家</CRD> 工厂
[mc970116]

短短几个月间, (short short a-few CL months' space)
短短 <DUR>几个月</DUR> 间,
mc970208

However, do not mark the constructions which use 多 ("many") rather than 几 ("some"), since it seems less specific.

跟 ^李 ^国能认识多年. (...acquainted with Li Guoneng many years)
跟 <PER>^李 ^国能</PER>v 认识多年.
[mv970521]

Other types of multiplicative cases include actual multiplications. Include the word for "times" in the markup, since it is used like a

classifier or measure. (Note that this type of multiplier is the only type of cardinal number meaning "one" which should be tagged.)

都比以前增加了一倍以上。 (all compared with before increased one time over, [i.e., all at least doubled])

都比以前增加了 <CRD>一倍</CRD> 以上。

[mc970411]

比一九九一年大幅度增长五倍左右。 (compared with 1991 a big increase of five times, more or less)

比 <DTE>一九九一年</DTE> 大幅度增长 <CRD>五倍</CRD> 左右。

[mc970418]

比上年增长了三点七四倍。 (compared with last year an increase of three point seven four times)

比上年增长了 <CRD>三点七四倍</CRD>。

[mc970403]

*** Orders of Magnitude

Numeric expressions which give order of magnitude information will be tagged.

** Named Entity Strings Embedded in Numeric Expressions

Multi-word strings that are to be tagged as NUMEX may contain named entity (ENAMEX) substrings. Include these words within the scope of the tagged expression, but do not apply an embedded ENAMEX tag.

一百二十摄氏度 (120 celsius degrees)

<MSR>一百二十摄氏度</MSR>

[mc970211]

2000美元 (2000 u.s. dollars)

<MNY>2000美元</MNY>

[MET-2 Guidelines]

** TYPE-Specific Guidelines

*** MONEY

Any monetary expression should be assigned the type MONEY, even if it is combined with measured quantities. The entire string expressing a monetary value is to be tagged, both the classifier/measure and the head noun.

五万四千块钱。 (54,000 dollars money)

<MNY>五万四千块钱</MNY>。

[mc970211]

损失三点八亿元人民币。 (... lose three hundred eighty million dollars RMB)

损失<MNY>三点八亿元人民币</MNY>。

[XIN19980105.0003]

2000美元 (2000 U.S. Dollars)

<MNY>2000美元</MNY>

[MET-2 Guidelines]

假如石油价格能够维持在每桶二十+元以上 (if oil prices can stay at over twenty dollars a barrel)

假如石油价格能够维持在每桶<MNY>二十+元</MNY> 以上

[mk970005]

If the currency comes before the numbers, it should also be included in the markup.

捐款一千一百一十四万港元，约合人民币一千一百九十二万元

(donated 11,140,000 HK Dollars, about RMB 11,920,000 dollars)

捐款<MNY>一千一百一十四万港元</MNY>，约合<MNY>人民币一千一百九十二万元</MNY>

Numeric expressions that do not use currency terms directly to indicate money values are still to be tagged if world knowledge indicates that they are monetary values.

以三十三又八分之五收盘 (...closed at thirty-three and five eighths)

以 <MNY>三十三又八分之五</MNY> 收盘
[mk970003]

Expressions for stock mark indices should be marked as MONEY.

道琼工业平均指数惨跌了一百四十点一一 ((dow jones industrial average
fell 140.11)
<ORG>道琼</ORG>工业平均指数惨跌了<MNY>一百四十点一一</MNY>
[mk970009]

*** MEASURE

**** Simple MEASURE's

Taggable MEASURE expressions contain standard units of physical measure, which are defined as measures whose quantity values do not change over time. For example, a "yard" always consists of three feet, so a numeric expression like "ten yards" is taggable. In contrast, in a phrase like "one wave of water after another" the string "one wave of water" is not a MEASURE expression because "wave" does not have a fixed volume. Typical taggable measures are age, area, distance, energy, speed, rate, temperature, volume, and weight.

Age -- the age of a person or thing given in terms of some unit of time plus the phrase including that unit.

一条才数日大的雌鲸鱼 (an only several days old female whale)
一条才<s>数日</s>大的雌鲸鱼
[mk970005]

Indications of age that are non-numeric are not taggable.

Area -- the measure of a two-dimensional space.

Distance - the linear measure of the space between two points.

Energy -- measures of any form of usable power including electricity, heat, work, radiation, light, or sound.

Speed -- a measure of the rate of movement usually in units indicating distance over time.

Temperature -- the degree of hotness or coldness of an object or an environment measured on a standard scale.

Volume - a measure of the amount of space occupied by an object or its capacity.

Weight - a measure of the heaviness of an object using any country's standard measure.

A good test to distinguish a MEASURE expression from other types of numerical expressions is to check if the expression (unless it is a DURATION expression) could be converted into an equivalent expression in standard SI/ISO-31 units (See, for instance, <http://ts.nist.gov/ts/hdocs/200/202/pub814.htm>). Note that some "dimensionless quantities," such as Richter earthquake magnitudes and pH would fail such a test, even though they should be tagged as MEASURES.

**** Ratio expressions

Some taggable MEASURE's are given in terms of ratios of numeric expressions.

***** Simple Juxtaposition of Numeric Expressions

一年一度 (one year one time)
<MSR>一年一度</MSR>
[mc970421]

百年一遇 (one hundred years one event)
<MSR>百年一遇</MSR>
[mc970220]

[Both of the above examples could be seen to be convertible into units of "hertz". This may be a stretch, though, since the terms for "time" and "event" would be lost in the conversion.]

Note that not all juxtapositions are necessarily measures.

一国两制 (one country two systems)

一国<CRD>两</CRD>制

[mv970625c]

[Not "two systems per country", rather a phrase that describes a certain political principle or situation.]

访问团一行二十四人 (tour group one line twenty-four people)

访问团一行<CRD>二十四</CRD>人

[mk970002.utf]

一家五口人 (one family five-CL people)

一家<CRD>五口</CRD>人

[mc970114.utf]

增加到一天二十四小时 (increase to one day
twenty-four hours)

增加到<DUR>一天</DUR><DUR>二十四小时</DUR>

[mv970620a.utf]

***** Ratios expressed with "mei" 每 ("every")

Expressions of the form 每<classifier-or-measure><numeric-expression>
are taggable as measures.

可防御每秒一万立方米流量的洪水 (can prevent every second ten thousand
cubic meters flow of flood water)

可防御<MSR>每秒一万立方米</MSR>流量的洪水

[mc970220]

**** Scope of MEASURE expressions

Dimension words, prepositional phrases, and verbs of measure are not
included within the scope of MEASURE expressions.

海拔 一千八百二十八米 (elevation 1828 meters)

海拔 <MSR>一千八百二十八米</MSR>

[mc970114]

宽五米高五百余米 (wide five meters (,) high over 500 meters)

宽 <MSR>五米</MSR> 高 <MSR>五百余米</MSR>

[mc970114]

里氏六点二级. (richter 6.2 "scales")

里氏 <MSR>六点二级</MSR>.

[mv970626c]

气温可达到摄氏二十四五度 (temperatures can reach celsius
twenty four (or) five degrees)

气温可达到摄氏<MSR>二十四</MSR><MSR>五度</MSR>

[mc970220]

*** CARDINAL

When numerals provide a count or quantity of something that is not a
standard unit of measurement, an amount of money or a percentage, they
are marked by themselves as CARDINAL. The quantified noun is not
included within the scope of the tag, but the classifier is included,
if it is present.

十二亿中国人民 (1.2 billion china people)

<CRD>十二亿</CRD> <LOC>中国</LOC> 人民

[mc970114]

二十二个国家 (twenty-two CL nations)

<CRD>二十二个</CRD> 国家

[mc970421]

四位旅游界人士 (four CL travel world people i.e., four people in
the travel field)

<CRD>四位</CRD> 旅游界人士

[mc970421]

10辆私人小汽车 (ten CL private small cars)

<CRD>10辆</CRD>私人小汽车

[MET-2 Training Data, METID 050]

二十分到六十七分。(twenty points-CL to sixty-seven points-CL (said of scores on a TOEFL test))
<CRD>二十分</CRD> 到 <CRD>六十七分</CRD>.
[mv970521]
[In the above gloss, "points-CL" means the classifier, 分, often translated as "points."]

If the thing counted is implied rather than explicit, the number is still marked.

*** Ratios of Cardinal Numbers

Ratios of cardinal numbers expressed with "bi" 比 ("compare/ratio") should be decomposed into the two parts of the ratio.

今天大盘的下跌股和上涨股的比例是三比二 (today the big board's losers and winners ratio was three to two)
今天<ORG>大盘</ORG>的下跌股和上涨股的比例是<MSR>三</MSR>比<MSR>二</MSR>
[mk970001]

*** Non-taggable Cardinal Number "one"

Terms for the cardinal number "one" should not be tagged, except for the multiplier expression 一倍, which means "onefold", and is almost always used additively, so that "increased onefold" means "doubled." (DURATIONS, MEASURES, MONEY, and PERCENT expressions of one unit should be tagged.)

此外还有一个条件 (besides this there is still one CL condition)
[no markup]
[mc970114]

一座巴勒斯坦城市 (one CL palestine city)
一座 <LOC>巴勒斯坦</LOC> 城市
[mc970116]

一批国家保护动物 (one CL-for-groups nation protect animals)
[no markup]
[mc970415]

我觉得价格就是贵一点儿吧 (i feel the price is expensive a little, eh)
[no markup]
[mc970408]

中国最大的自行车生产企业之一 (china largest bicycle production enterprises, one of)
<LOC>中国</LOC> 最大的自行车生产企业之一
[mc970214]

*** PERCENT

A percentage is a fraction with 100 as the assumed denominator. When tagging a numerical value of percent, the word "percent" or any of its variants is considered part of the number being tagged.

Numeric expressions that do not use percentage terms to indicate percentages are tagged as long as world knowledge indicates that they are expressed in percentages.

每年降低百分之二 (every year decrease two percent)
每年降低<PCT>百分之二</PCT>
[mc970305]

** Idiomatic Numeric Expressions

*** Decomposable Idioms

Except for the expressions given in the following section, idioms containing numbers are decomposable. If the requirements for the specific NUMEX TYPE are met, numerals within idioms can be tagged.

百尺竿头,更进一步 ((from a) hundred foot pole, go even further, ie, "make still further progress")
<MSR>百尺</MSR>竿头,更进一步

[mc970305]

千辛万苦 (thousand fortunes ten-thousand bitternesses)
<CRD>千</CRD>辛<CRD>万</CRD>苦
[ZBN19980314.0104]

再四舍五入. (... and then "four abandon five get" (i.e., "round off"))
再<CRD>四</CRD>舍<CRD>五</CRD>入.
[mv970521]

*** Non-decomposable idioms

Do not tag the numbers in the following expressions:

百姓 ("hundred names", [i.e., "the public", or "the people"])
十分 ("ten parts", when in an adverbial position meaning "very" or "completely")
万分 ("ten thousand parts" in an adverbial position)
万一 ("ten-thousand one", [i.e., "if by some small chance..."])
千万 ("thousand ten-thousand", when it is used to emphasize an injunction, as in 千万要小心地进出股市, "thousand-ten-thousand should carefully enter and leave the stock market")
多半 ("more half", the greater part, most, usually)
十字路口 ("ten ideogram street mouth/opening", [i.e., intersection])
百合 (the lily)

** Non-tagable Numbers

*** Non-tagable Numbers within Names

Organization names are not to be decomposed, i.e., the number should not be marked.

Numbers in street addresses and street names containing a number should not be tagged as CARDINAL (the address as a whole is tagged as ENAMEX LOCATION).

Products and other types of names are also not decomposable, even if a number is given.

*** Non-tagable Numbers within Temporal Expressions

Numbers within temporal expressions are not to be tagged separately within the TIMEX expression.

*** Non-tagable Ordinals

Ordinal number expressions are not to be tagged. An NUMEX of type "ordinal" may be included in future evaluations.

丹麦第二大反对党 (denmark's second largest opposition party)
<LOC>丹麦</LOC> 第二大反对党
[mc970411.utf]

*** Non-tagable Phone Numbers

Phone numbers are not taggable.

二零二六一九三一一一 ((202) 619-3111)
<ONU>二零二六一九三一一一</ONU>
[mv970521]

* SPECIAL FEATURES FOUND ONLY IN TRANSCRIBED SPEECH

The output from human transcriptions and speech recognition systems will differ. Human transcriptions will be marked up by humans, then

snorified and normalized to produce the human-created scoring keys. Speech recognizer output will be marked up by machines, normalized, then scored against keys.

**** Disfluencies**

In the cases of both corrections and partial repeats, a hyphen was previously inserted to show the interruption of the incomplete word. This use of the hyphen has now been replaced by an SGML tag indicating a fragment.

***** Repetition of complete words**

Tag all instances of entities even when they are repeated either for emphasis or correction.

五,六,五,六个 ze- 小时 (five,six,five,six ze- hours)
<DUR>五</DUR>,<DUR>六</DUR>,<DUR>五</DUR>,<DUR>六个 ze- 小时</DUR>
[mv970625c.utf]

***** Partial repetition**

If a fragment of a word or entity name occurs at either the beginning or the end of a complete entity name, the fragment will be left out of the tagged name.

If, however, the fragment occurs within the bounds of the minimal string identifiable as the entity name intended by the speaker, the fragment will be included in the string.

***** Corrections**

If a correction does not fall within a valid Named Entity expression, it will not be tagged.

If however, a correction falls within or comprises a valid expression, it will be tagged even when the annotator can determine that the words were said in error.

跌 破 八千 四 dai- 四百 点, 可
跌 破 <MNY>八千 四 dai- 四百 点</MNY>, 可
[mk9700011]

***** Deletions**

Do not mark up entities which are not there.

过去生活在<b_unclear><e_unclear>因为缺水,人们都在
(used to live at (non-transcribed), because of water shortage, people
all...)
[no markup]
[mc970114.utf]