

Simple Named Entity Guidelines

Version 6.4 – Tamil

March 29, 2006

Linguistic Data Consortium

Developed by Stephanie Strassel
(Based largely on the MUC-7 NE Guidelines)

Adapted to Tamil by David Shree Kumar, Lavanya Anandakumar, Mark Mandel and Christopher Walker

1	Introduction	2
2	Entity Types	2
2.1	Person Names மனித பெயர்கள்	3
2.2	Titles, roles and appositives பெயர் தலைப்பு ,பட்டம், உரிமை.....	3
2.3	Organization Names அமைப்பு, சங்கம்.....	5
2.3.1	General ORGANIZATION-like non-entities.....	6
2.4	Location Names இடம் குறி பெயர்கள்	7
2.4.1	Extent of Location Names.....	7
2.4.1.1	Compound expressions.....	7
2.4.1.2	Designators	8
2.4.1.3	Location modifiers and "semi-official" place names.....	8
2.5	Deciding among entity types.....	9
2.5.1	ORG referring to LOC, LOC referring to ORG	9
3	Difficult Cases	10
3.1	Expressions that refer to multiple entities	10
3.2	Nested Expressions	11
3.3	Entities as modifiers.....	11
3.4	Possessives.....	11
3.5	Other types of names.....	11
4	What NOT to tag	12
4.1	Events.....	12
4.2	Artifacts and products	12
4.3	Generics	13
5	Annotation Uncertainty.....	13

1 Introduction

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

சண்ரைஸ் சோடா கம்பெனி
சங்கர் கோபால்
தாஜ் மஹால்
வாகனம் தொழில்சாலை
மேரினா கால்பந்து அணி
ஊகன்டா
சென்னை ,தமிழ்நாடு
புஜி மலை

2 Entity Types

We will identify four types of named entities:

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

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

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

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

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

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

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

2.1 Person Names *மனித பெயர்கள்*

People may be specified by name, nickname or alias. Family names should also be tagged as PERSON. Names of deceased people, as well as fictional human characters appearing in movies, television, books and so on, should be tagged as PERSON entities. Religious deities should also be tagged as persons. The name of a family should also be tagged when referring to the whole family or members of the family.

மனித பெயர்கள் குடும்ப பெயர், அரச மரபு, இந்து சமயத் தொடர்பான பெயர்கள் . Names of Hindu gods ,Kings Dynasty often appears in Tamil Text to be tagged.

[காந்தியின்] குடும்பம்

PER

மதுரையை [பாண்டியர்கள்] காலம்காலமாக ஆண்டனர்

Per

[பாண்டவர்களுக்கும்], [கவுரவர்களுக்கும்] இடையே போர்

Per

Per

நடந்தது

2.2 Titles, roles and appositives *பெயர் தலைப்பு ,பட்டம், உரிமை.*

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

[துணை ஜனாதிபதி] [கிருஷ்ணகாந்த்] இங்கு வந்தார்.

TTL

PER

For this task we define titles and roles as occurring either directly before or directly after a person name. Therefore, **titles and roles are only tagged when they occur directly next to the person name they modify.** In the following example, for instance, the phrase "Vice President" is not considered a title and is not tagged:

The strongest supporter was the Vice President.

நேற்று துணைஜனாதிபதி இங்கு வந்தார்.

If a title contains within it a taggable entity, tag that entity separately. For instance:

[மைக்ரோஸஃப்ட்] [தலைவர்] [பில் கேட்ஸ்]
ORG TTL PER

Tamil does not use appositives like "Jr.", "Sr.", and "III" with personal names. For a foreign monarch such as "Queen Elizabeth II", the Tamil expression is literally "second Elizabeth queen" so the number is tagged as part of the personal name just as in English.

[Queen] [Elizabeth II] sent a message of greeting.
TTL PER

[இரண்டாம் ஏலிசபத்] [ராணி] வாழ்த்து செய்திகள்
PER TTL
அனுப்பினார்

(This does not happen in Tamil: Finally, sometimes the name of the person is split into two pieces by the title. In these cases, we will annotate the two pieces of the PERSON name as two separate PERSON entities:

[Alfred] [Lord] [Tennyson]
PER TTL PER
In Tamil there is no split in person name with a title.)

Some more examples of names and titles:

[மாருதி கார்] [துணை தலைவர்] [கோபால்]
ORG TTL PER

[நிதி அலுவலகம்] [செயலாளர்] [கோபலகிருஷ்ணன்]
ORG TTL PER

[இந்தியாவின்] [துணைஜனாதிபதி]
LOC TTL

[உச்சநீதிமன்றம்] [மந்திரி] [கருணாகாரன்]
ORG TTL PER

[இந்தியா] [சபாநாயகர்] [நாயுடு]
LOC TTL PER

[விண்வெளி மயத்தின்] [தலைவர்] [அப்துல்கலாம்]
ORG TTL PER

2.3 Organization Names *அமைப்பு, சங்கம்.*

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

Businesses

[காவோரி டயர் கம்பெனி] இலாபம் அடைந்தது

[இந்தியன் ஆயில்] பங்கி

[லாடா] பங்கு

Stock exchanges

[மும்பை பங்கு சந்தை] இன்று விடுமுறை

Multinational organizations

[ஐ. நா சபை] கலந்து கொண்டது

Political parties

[திமுகா] வெற்றி

sports teams

[கல்கத்தா கால் பந்து அணி]

[புதுவை மகளிர் அணி]

Military groups

[விடுதலை புலிகள் இயக்கம்]

[ஆல் மசுத்] என்னும் திவிரவாதிகள்

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

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

Religious organizations

[தென் இந்தியா திருச்சபை]

[இயேசு அழைக்கிறார் சபை]

Hospitals

[அரசாங்க மருத்துவமனை]

[ஆப்போலோ ஆஸ்பத்திரி]

Hotels

[ஹோட்டல் தாஜ்மஹால்]

[லா மேரிடியன்]

Museums

[சென்னை அருங்காட்சியம்]

[சென்னை விலங்குகள் பூங்கா]

Universities

[அண்ணா பல்கலைக்கழகம்]

[அண்ணாமாலை பல்கலைக்கழகம் சிதம்பரம்]

Government offices

[இராஷ்டிரபதி பவன்]

2.3.1 General ORGANIZATION-like non-entities

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

2.4 Location Names *இடம் குறி பெயர்கள்*

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

ஊர்.மாவட்டம் , தொகுதி, தெரு, கடல், தீவு , அணை, விமான நிலையம், தாஜ் மஹால்.

[மேட்டுர் அணையில்] தண்ணீர் நிரம்பியது

LOC

[மண்ணடிபட்டு தொகுதியில்] ஒட்டு எண்ணிக்கையில்

LOC

குழப்பம்.

[சைதாபேட்டை ஆற்றுபாலம்] சீரமைப்பு வேலை நடக்கிறது

LOC

ஐந்து மைல் பிறகு [ஊட்டிமலைபாதை] மூடப்பட்டு

LOC

இருக்கிறது

[ஊத்தாங்கரை] தக்காளிபழத்திற்குபெயர் போனது

LOC

2.4.1 Extent of Location Names

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

2.4.1.1 Compound expressions

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

[புதுடெல்லி] , [இந்தியா]

[மீணம்பாக்கம்] , [சென்னை]

2.4.1.2 Designators

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

[கங்கை ஆறு]
[காஷ்மீர் மலை] பகுதி

2.4.1.3 Location modifiers and "semi-official" place names

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

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

[மத்திய மும்பை]
[வடக்கு டெல்லி]

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

[மத்திய கிழக்கு நாடுகள்] ஆன.
[மேற்கு வங்காளம்]
[கிழக்கு ஐரோப்பிய]

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

[கோதாவரி ஆறு] டேக்கான் பிளட்டோ
முன்னாள் [சோவியத் நாடு]
[மெட்ராஸ்], இப்போது [சென்னை]
[கீழ் திருப்பதி]
கிழக்கு [சென்னை]

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

2.5 Deciding among entity types

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

2.5.1 ORG referring to LOC, LOC referring to ORG

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

[எம் ஜி ஆர் புகைபடம் நகரம்] இடம் மாற்றம் அறிவித்துள்ளது.
ORG

[எம் ஜி ஆர் புகைபடம் நகரத்தை] [கருணாநிதி] திறந்து வைத்தார்.
LOC PER

[ஹோட்டல் தாஜ்] முன் இடத்தில் குண்டு வொடிப்பு
LOC

[ஹோட்டல் தாஜ்] உரிமையாளர் இதை கண்டித்து உள்ளார்
ORG

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

[புது டெல்லி] புதிய வருமான வரி விலக்கு அறிவித்துள்ளது
ORG

In this case, the name புது டெல்லி is the name of New Delhi, the capital of India, used to refer to the Indian Government. Because புது டெல்லி is referring to an organization entity in this example, it should be tagged as ORG.

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

மாணவர்கள் பிரச்சனை,[ராஜ்பவன்] தீர்த்து வைக்கும்
ORG

(For place names referring to sports teams: In Tamil the name of the team is always used, so constructions like this do not occur.)

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

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

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

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

3 Difficult Cases

3.1 Expressions that refer to multiple entities

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

For instance, this sentence contains two entities:

[இந்தியாவும்] [பாகிஸ்தானும்] இராணுவ ஒப்பந்தம்
செய்தார்கள்

Similarly,

[மன்மோகன்] மனைவி [சில்பா சிங்]

[மேற்கு] [கிழக்கு இந்தியாவும்]

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

[கடல் மற்றும் குளம் மீன்துறை] ("Sea and Lake Department")

is the name of one organization and should be tagged as a single named entity: it's not

கடல் மீன்துறை ("Sea Department")

and

குளம் மீன்துறை ("Lake Department")

as separate names.

3.2 Nested Expressions

No nested expressions will be marked. When the name of one entity contains within it another entity name, do not pull out the name of the other entity and mark it separately. Only tag the larger entity. For instance

[சங்கர் சுந்தரம் பைனான்ஸ்] no markup for **சங்கர் சுந்தரம்**
alone

[சுருண கோழிபண்ணை] no markup for **சுருண** alone

3.3 Entities as modifiers

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

[பாம்பே டையிங்] இலாபகரமாக உள்ளது
[கருணாநிதியின்] அரசாங்கம்
[மத்திய அரசாங்கம்] முத்திரை தாள்கள்
[இந்தியா] ஏற்றுமதி செய்கிறது
[டாடா] கார்கள்
[அந்தமான்] கச்சா எண்ணை
[சென்னை] புகைபடம் விழா

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

[இந்தியாவில்] தொழில்சாலைகள்
[இந்திய] குடிமகன்
[இந்தியாவின்] உணவு

3.4 Possessives

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

[தமிழக அரசின்] [கல்வி துறை]

[இந்தியாவின்] [பார்லிமென்ட்]

3.5 Other types of names

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

என்.எல்.சி	abbreviation for நெய்வேலி
அனல்மின் கார்பரஷன்	
நெய்வேலி சுரங்கம்	nickname for நெய்வேலி
அனல்மின் கார்பரஷன்	
மும்பை மார்கெட்	nickname for மும்பை பங்கு
சந்தை	
குண்டர்கள்	nickname for அடிஆட்கள்
பலிங்கிகல் நகரம்	nickname for ராஜஸ்தானை
குறிக்கும்	
மேகான் பேகான்	nickname for வங்காள கால் பந்து
அணி	
நீல்கிரிஸ்	nickname for நீல்கிரிஸ் பெரிய
அங்காடியை குறிக்கும்	

4 What NOT to tag

4.1 Events

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

தமிழ் நாடு அணிக்கு	[no
markup]	
[இந்தியா கால்பந்து அமைப்பு]	[Organi
zation]	

4.2 Artifacts and products

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

இண்டிகா என்பது சிறந்த கார் மாடல்	[no
markup]	

4.3 Generics

Also, generic names that do not refer to a specific entity should not be tagged. (*This is not common in Tamil.*)

உலகில் பெயர் பெற்ற மேகி சாஸ்
markup]

[no

5 Annotation Uncertainty

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

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

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

So that you can keep working even after you have a question about a particular example, we've created one more tag in the annotation tool: "No_Annotation". When you encounter a problem or have a question about a particular word or phrase and you can't get an immediate answer, label the item "No_Annotation". That will let us easily find it later when we try to resolve the problem.

You should also use the "No_Annotation" label in cases where there's some kind of problem with a single word or handful of words in the file – e.g., they're badly translated or the font isn't displaying properly.

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