NCHLT Text Phase II Project: Named Entity Recognition

2013-2015

Protocol: Named entity recognition (TSHIVENDA)

# Introduction

## Aim of this protocol

The aim of this protocol is to guide annotators in the process of annotating the Tshivenda corpus on the named entity layer for the NCHLT Text Phase II project. This document explains some general principles to follow when annotating the corpus, explains the process of annotation, and provides a detailed explanation of the categories and annotations used in the project.

This is a dynamic, “living” document, and should be adapted following the procedure as explained in the “Changing the protocol” section.

## Concepts and definitions

Named entity recognition is the process of identifying rigid designators, such as the name of a person or an organisation, location, brand, product, etc., which references a specific entity with a unique identifier. Identifying these units is an essential tool in the development of complex information extraction and retrieval tools.

The aim of this part of the NCHLT Text Phase II project is to annotate tokens that form part of one of four types of named entities, namely person names, organisation names, location names and miscellaneous named entities that adhere to the three principles above, but do not form part of one of the other three categories. The following are broad definitions of the different types to be annotated in the data:

* PERSON (PER): named person, family, or certain designated non-human individuals.
* ORGANISATION (ORG): named corporate, governmental, or other organisational entities.
* LOCATION (LOC): names of politically or geographically defined locations (cities, provinces, countries, international regions, bodies of water, mountains, etc.) and astronomical locations.
* MISCELANEOUS (MISC): all rigid designators that do not fall into one of the other categories, such as temporal expressions (dates and times), URLs, numerical expressions, publications, names of languages, nationalities, etc.

The annotation of the data will be done according to the standards used in the CoNLL shared task. This scheme assigns an annotation that indicates if the particular token is part of one of the types, or Outside (O) all of the defined types. As an example, the following sentence shows how these tags are assigned according to the PER, ORG, and LOC types:

* **Mr Pierre Vinken** PER
* is O
* chairman O
* of O
* **Elsevier N.V.** ORG
* , O
* the O
* **Dutch** MISC
* publishing O
* group O

## General principles

The following general principles apply:

1. The tokens must be a proper name, most likely written with a capital letter.
2. The tokens must be a unique identifier.
3. The name of the entity must be assigned through some official process, such as birth certification, official registration or assignment through a law or governmental agency.
4. In metonymical expressions where the exact type is unclear, the most prototypical interpretation should be assigned.

Annotation on the NE layer should exceed 80%

The categories set out below are based on:

* Chinchor, N., Brown, E., Ferro, L., & Robinson, P. 1999. 1999 Named Entity Recognition Task Definition. *MITRE and SAIC*. [Available: <ftp://jaguar.ncsl.nist.gov/ace/phase1/ne99_taskdef_v1_4.pdf> ]
* Tjong Kim Sang, E F. 2002. Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. In *Proceedings of the sixth conference on natural language learning at HLT-NAACL. pp. 155-158.* [Available: <http://www.cnts.ua.ac.be/conll2002/ner/>]
* Tjong Kim Sang, E. F., & De Meulder, F. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003-Volume 4.* pp. 142-147. [Available: <http://acl.ldc.upenn.edu/W/W03/W03-0419.pdf>]

# Process

The following process should be followed:

1. The project manager compiles a set of sentences that should be annotated for named entities in Lara3.
2. The annotator receives a Lara3 file with each word that should be annotated with initial rule-based assignments to different named entity types.
3. The annotator verifies and corrects the initial NE assignments according to the categories stipulated in Section 3.
   * If the correct NE assignment is already set in the Type column, leave it “as is”.
   * If there is uncertainty about the NE type the particular word belongs to, select the “Unsure” NE type from the dropdown menu. The language expert needs to make a final decision on these cases.
4. When done, the annotator saves the file and sends it back to the project manager.

# Categories

## Person (PER)

This entity type includes not only humans, both “real” and “fictional”, but also other fictional or real non-human individuals. The reasoning for including these non-humans in the PERSON type is given in the following sections.

We choose not to divide this type into different sub-types, leaving such resolution and sense-evaluation to higher level processing.

* Titles such as “Mr.” and role names such as “President” **are** considered part of a person name. Appositives such as “Jr.,” “Sr.” and “III” are alsoconsidered part of a person name.

"Vho-Shumani Tshikota"

* **Vho-Shumani Tshikota** PER

"Muṅwaleli Robert Mosbacher"

* **Muṅwaleli Robert Mosbacher** PER

"John Doe, Jr."

* **John Doe, Jr.** PER
* In cases where a reference to a specific person is made using their title, the title should be tagged as a PER, but not in cases where the reference is a generalisation of the title.

“kha kha yavho, Muphresidennde o ri”

* kha O
* kha O
* yavho O
* , O
* **Muphresidennde** PER
* o ri O

“musi muphresidennde wa dzangano”

* musi O
* muphresidennde O
* wa O
* dzangano O
* Entity names modifying a person or their title/role should be tagged.

"muṅwaleli wa Telkom Leslie Maasdorp"

* muṅwaleli O
* wa O
* **Telkom** ORG
* **Leslie Maasdorp** PER

"Muṅwaleli wa Gwama"

* **Muṅwaleli wa Gwama** PER
* Surnames are to be tagged as PERSON.

"muta wa ha Mandela"

* muṱa O
* wa O
* ha O
* **Mandela** PER
* Animal names are to be tagged as PERSON.

"Tshamaṱo, mmbwa ya muphresidennde”

* **Tshamaṱo** PER
* , O
* mmbwa O
* ya O
* muphresidennde O
* Religious titles or specifiers such as “saint,” “prophet,” “imam,” or “archangel” are not be tagged, the proper name will be tagged as a PERSON.

"mufhaṱutshedzwa Daswa ndi muhulwane wa kereke"

* mufhaṱutshedzwa O
* **Daswa** PER
* ndi O
* muhulwane O
* wa O
* kereke O
* References to “God”, “Allah”, “Prophet”, “Messiah”, and the like, will be taken to be the “name” of this entity for tagging purposes. If it is used as a descriptor, rather than a name, it will not be tagged.

"arali ni tshi tenda kha Mudzimu..."

* arali O
* ni O
* tshi O
* tenda O
* kha O
* **Mudzimu** PER
* Fictional animals are a specific type of fictional character, and as such should be tagged.

"Sankambe"

* **Sankambe** PER

"muḓivhalea Muzhou"

* muḓivhalea O
* **Muzhou** PER

## Location (LOC)

The entity 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, neighbourhoods, 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.

* A location name following an organisation name may or may not be part of the organisation name proper. The annotation in the answer key will follow these guidelines:

1. If the location name is preceded by “in” or “at,” the location name will be tagged separately from the organisation name, unless world knowledge dictates otherwise (see the note below).
2. If the location name is preceded by “of” or no preposition at all, then determine tagging as follows:
   1. if there is an organisation designator or organisation-designating common or proper noun, it marks the end of the organisation name.

"Hyundai, Inc. of Korea"

* **Hyundai, Inc.** ORG
* ya O
* **Korea** LOC
  1. if there is no organisation designator, or if the organisation is a bank or university, the location name is part of the organisation name.

"Hyundai of Korea, Inc."

* **Hyundai ya Korea, Inc.** ORG
* Note: locatives are often intrinsic to names of universities and should be included in the organisation name when world knowledge dictates, such as in the following:

"Univesithi ya Venda"

* **Univesithi ya Venda** ORG
* “World knowledge” includes the application of using commonly-used acronyms to help determine the extent of the name. For instance:

“Univesithi ya Venda, Khamphasi ya Makhado”

* **Univesithi ya Venda, Khamphasi ya Makhado** ORG
* Abbreviations of organisations that are provided in addition to the full name of the company are annotated as separate entities. For instance:

“Univesithi ya Venda (UNIVEN)”

* **Univesithi ya Venda** ORG
* ( O
* **UNIVEN** ORG
* ) O
* Compound expressions in which place names are listed in succession, with or without a separating comma, are to be tagged as separate instances of LOCATION.

“Makhado, Dzanani”

* Makhado LOC
* , O
* **Dzanani** LOC
* 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.

"Mulambo wa Mutale"

* **Mulambo wa Mutale** LOC
* Do not include in the tagged string common noun phrases functioning as partitive-type locative specifiers directly after LOCATION names, such as:

"Vhupo ha Ṱhohoyanḓou"

* Vhupo O
* Ha O
* **Ṱhohoyanḓou** LOC
* 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”) as LOC.

“Afrika”

* **Afrika** LOC

“Dzingu ḽa Afrika ya Sahara”

* **Dzingu ḽa Afrika ya Sahara** LOC

“Vhubvaḓuvhakati”

* **Vhubvaḓuvhakati** LOC
* Street names that include references to a person should still be tagged as a location

“Tshiṱarata tsha 452 tsha Justice Mohammed”

* **Tshiṱarata tsha 452 tsha Justice Mohammed** LOC
* 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.

“the Southwest region” [no mark-up]

"the northeast" [no mark-up]

* Do tag names of sub-national but specific regions if they are identifiable even when the name is disassociated from context. 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.

“Mavhaleni”

* **Mavhaleni** LOC

“Dzwaini”

* **Dzwaini** LOC

“Sananga”

* **Sananga** LOC
* Directional modifiers (“north,” “south,” “east,” “west,” “upper,” “lower,” and combinations thereof) are taggable only when they are intrinsic parts of a location's official name, as in

“Volta ya Nṱha”

* **Volta ya Nṱha** LOC

or

“Dakota ya Devhula”

* **Dakota ya Devhula** LOC

## Organisations and Products (ORG)

Proper names that are to be tagged as ORGANISATION include stock exchanges, multinational organisations, businesses, TV or radio stations, political parties, religions or 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 organisations (to ensure consistency with marking other fictional entities).

To prevent confusion, especially in the case of metonymical uses of product and organisational names, all products should also be tagged as organisation

* Corporate or organisation designators such as “Co.” are considered part of an ORGANISATION name.

“Khamphani ya Mitambo ya Bridgestone”

* **Khamphani ya Mitambo ya Bridgestone** ORG
* Abbreviations of company names are marked as ORGANISATION.

“JSE”

* **JSE** ORG

"wa u ḓovha muphresidennde wa DA"

* wa O
* u O
* ḓovha O
* muphresidennde O
* wa O
* **DA** ORG
* Although event names will be tagged as MISC, the institutional structures themselves associated with the event – steering committees, etc. – should be tagged as ORGANISATION.

"Komiti ya Lushaka ya Olympiki"

* **Komiti ya Lushaka ya Olympiki** ORG
* Political or religious parties are sometimes referred to as collections of their members, e.g., “the Moonies”, “Democrats”. Such designators are to be tagged as ORGANISATION only when they clearly act as such. If the term is being used to refer to an unspecified collection of members of an organisation, it should not be tagged.

"ANC I tenda uri…”

* **ANC** ORG
* i O
* tenda O
* uri O
* Proper names referring to meeting places or places where organisational activities occur (e.g., churches, embassies, factories, hospitals, hotels, museums, universities) will be tagged as ORGANISATION.

"Mediclinic ya Pretoria"

* **Mediclinic ya Pretoria** ORG

"Hodela ya Four Seasons"

* **Hodela ya Four Seasons** ORG

"Kereke ya Reformed"

* **Kereke ya Reformed** ORG

"ya White House"

* ya O
* **White House** ORG
* Brands should be marked as ORG

“tshinwiwa tsha Coke”

* tshinwiwa O
* tsha O
* **Coke** ORG
* In cases where the manufacturer and the product are named, both the manufacturer and the product will be tagged as ORG. 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).

"Ford Taurus"

* **Ford Taurus** ORG

"Dow Jones ya makwevho"

* **Dow Jones** ORG
* ya O
* makwevho O
* Tag news sources (newspapers, radio and TV stations, and news journals) as ORGANISATION even when they function as artefacts. 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 artefact or organisation, both usages are tagged. Moreover, publications often function agentively as ORGANISATIONS. When a publication, regardless of subject matter, “reports,” “states,” “claims,” etc., it is acting as a news source (reporting information).
* Similar to publications, metonymy occurs frequently with stock indices and stock markets. Tag these as ORGANISATION even when they refer to numeric values published by or representing the organisation.

“Shere dza FNB dzo tsitsela fhasi JSE”

* shere O
* dza O
* **FNB** ORG
* dzo O
* tsitsela O
* fhasi O
* **JSE** ORG
* Generic entity names such as “the police” and “the government,” are not to be tagged.

“this country has never been invaded by an army” [no mark-up]

“mulangi wa Mmbi”

* mulangi O
* wa O
* **Mmbi** ORG

“Mapholisa a Gauteng o wana”

* **Mapholisa a Gauteng** ORG
* o O
* wana O

## Miscellaneous (MISC)

In addition to the specific types defined in the previous three sections, an additional category of unique identifiers should also be marked as Miscellaneous. This category includes all entity types that do not form part of one of the other categories, but still adhere to the four principles defined in 1.3. The following examples provide some guidance on the main types of entities that will fall into this category, but will not necessarily cover all of the different identifiers that may be added to this category.

* Names of publications such as books, acts should be annotated as MISC.

“Mulayo wa 36 wa 1997”

* **Mulayo wa 36 wa 1997** MISC
* Temporal expressions such as date and time should be marked as MISC.

“12 Luhuhi 2006”

* **12 Luhuhi 2006** MISC

“kha ri tangane nga 16:00”

* kha O
* ri O
* tangane O
* nga O
* **16:00** MISC
* Names of languages and ethnic references should be annotated as MISC, e.g. “isiZulu”, “French”, “Germans”.
* Names of animals, plants and illnesses are never NEs and should be marked as outside:

“kha malaria na flu”

* kha O
* malaria O
* na O
* flu O

# Exceptional Cases

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

“Department of Arts and Culture” (which does NOT mean two entities, i.e. “the Department of Arts” and “the Department of Culture”)

* **Muhasho wa Vhutsila na Mvelele** ORG
* A conjoined multi-name expression, in which there is elision of the head of one conjunct, should be marked up as separate expressions.

“Devhula na Amerika Tshipembe”

* **Devhula** LOC
* na O
* **Amerika Tshipembe** LOC
* Proper names used as modifiers in complex noun phrases are to be tagged when it is clear to the annotator from context or the annotator’s world knowledge that the modifier name is that of an organisation, person, or location.

"muvhuso wa Mbeki"

* muvhuso O
* wa O
* **Mbeki** PER

"Bondo na sekhurithisi dza Gwama"

* Bondo O
* na O
* sekhurithisi O
* dza O
* **Gwama** ORG
* In some cases, taggable multi-word strings will contain entity name substrings; such multi-word strings are not decomposable; therefore, the substrings are not to be tagged.

“Arthur Anderson Consulting” [no mark-up for "Arthur Anderson" alone]

* **Consulting ya Arthur Anderson** ORG

“Escort Chicken Corp.” [no mark-up for "Escort" alone]

* **Chicken Corp ya Escort.** ORG
* In a possessive construction, the possessor and possessed named entity substrings should be tagged separately.

“Tshikolo tsha U phasa tsha Vhubindudzi tsha Univesithi ya Pretoria”

* Tshikolo tsha U phasa tsha Vhubindudzi ORG
* tsha O
* **Univesithi ya Pretoria** ORG

“Silicon Valley ya California”

* **Silicon Valley** LOC
* yaO
* **California** LOC
* Generally, aliases for entities are to be tagged. Taggable aliases will include 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.

"IBM" [alias for International Business Machines Corp.]

* **IBM** ORG

"PACTEL" [alias for Pacific Telesys, i.e. Pacific Telephone Systems]

* **PACTEL** ORG
* Nicknames and other aliases are tagged when they are established alternate ways of referring to an entity; if the annotator does not recognise 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.

"Big Blue" [alias for International Business Machines Corp.]

* **Big Blue** ORG

"Big Board" [alias for New York Stock Exchange]

* **Big Board** ORG
* Certain metonyms, herein designated “proper” metonyms, which chiefly include references to an organisation based on the name of a unique structure or facility in which the organisation holds office. The association between the name and the organisation 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 organisation.

"Gwama wo ri..."

* **Gwama** ORG
* wo ri O
* Metonyms, herein designated “common” metonyms, that reference political, military, athletic, and other organisations by the name of a city, country, or other associated location. In these cases, the association between the name’s semantic type and the organisation is sufficiently predictable and non-idiosyncratic as to preclude a dictionary gloss; hence the name should be tagged as a LOCATION, not as an ORGANISATION.

"Germany yo tshotela Poland."

* **Germany** LOC
* Yo O
* tshoṱela O
* **Poland** LOC

"Kaizer Chiefs yo kunda Sundowns”

* **Kaizer Chiefs** ORG
* Yo O
* Kunda O
* **Sundowns** ORG
* When a definite article is commonly associated with an entity name, it also must be tagged. In most cases, the definite article will be excluded from the annotated entity.

"kha Hague"

* **kha Hague** LOC

# Changing the protocol

Since this protocol is considered to be a dynamic document, it is subject to change from time to time. For purposes of good governance, the following principles and procedures should be followed:

* The language expert (in consultation with the project leader) has the final say in changing this procedure document, and is the only authoritative person who might change the protocol.
* If an annotator (or other project member) wishes to change a specification, s/he should log a request on Bugzilla, indicating the nature of the change request. The language expert then considers the request, and adds the change to the protocol document. The protocol document is saved with a new version number. All other annotators are then informed of the change, and requested to change previous annotations, where applicable.
* If an annotator wishes to add specific examples to the protocol, s/he sends an email to the language expert with the suggested examples. The language expert then adds the examples to this document, and when a significant number of examples have been added, saves the document with a new version number. All other annotators are then informed of the new version of the protocol document.