NCHLT Text Phase II Project: Named Entity Recognition

2013-2015

Protocol: Named entity recognition (SEPEDI)

# Introduction

## Aim of this protocol

The aim of this protocol is to guide annotators in the process of annotating the Sepedi 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.

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”, religious names 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.
* **Mna Willemse** PER
* **Mna Tienie Willemse** PER
* **Mdi Nicholsom** PER
* **Morena Modulasetulo** PER
* **Morena S.J. le Roux** PER
* **Kgošigadi Mokope Modjadji** PER
* **Maloko ao a hlomphegago** PER
* **Modimo**  PER
* **Jesu Kriste** 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.

“Hyundai, Inc. ya Korea”

* **Hyundai, Inc.** ORG
* ya O
* **Korea** LOC
* **Hyundai yaKorea, 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:

"University of Pretoria"

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

“North-West University, Potchefstroom Campus”

* **North-West University, Potchefstroom Campus** ORG
* Abbreviations of organisations that are provided in addition to the full name of the company are annotated as separate entities. For instance:

“Yunibesithi ya Pretoria (UP)”

* **Yunibesithi ya Pretoria** ORG
* ( O
* **UP** 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.

“Johannesburg, Gauteng”

* **Johannesburg** LOC
* , O
* **Gauteng** 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, “Dam” in the name of a dam, “Mountain” in the name of a mountain, “City” in the name of a city, etc., if such words are contained in the string.

"Letamo la Dap Naude le Letamo la Ebenzer"

* **Letamo la Dap Naude** LOC
* le O
* **Letamo la Ebenzer** LOC
* **Noka ya Limpopo** LOC
* Do not include in the tagged string common noun phrases functioning as partitive-type locative specifiers directly after LOCATION names, such as:

"tikologo(ng) ya Greater Cape Town area"

* tikologo(ng) O
* ya O
* **Greater Cape Town** 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

”Afrika Borwa”

* **Afrika Borwa** LOC

”Repabliki ya Afrika Borwa”

* **Repabliki ya Afrika Borwa** LOC

”Repabliking ya Afrika Borwa”

* **Repabliking ya Afrika Borwa** LOC

“Tunisia le Chili”

* **Tunisia** LOC
* le O
* **Chili** 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.

“Bridgestone Sports Co.”

* **Bridgestone Sports Co.** ORG
* Abbreviations of company names are marked as ORGANISATION.

“JSE”

* **JSE** ORG

“Pretoria Mediclinic”

* **Pretoria Mediclinic** ORG

"Four Seasons Hotels"

* **Four Seasons Hotels** ORG

"Kereke ya Gereformeerd"

* **Kereke ya Gereformeerd** ORG
* Brands should be marked as ORG
* **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

"palogare ya Dow Jones"

* palogare O
* ya O
* **Dow Jones** ORG
* 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.

“makala a FNB”

* makala O
* a O
* **FNB** 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] but “Gauteng Police” is marked as ORG.

“Maphodisa a Gauteng”

* **Maphdisa a Gauteng** ORG

“Kgorotshekong ya Maseterata”

* **Kgorotshekong ya Maseterata** ORG

## 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.

“Molao 36 wa 1997”

* **Molao 36 wa 1997** MISC

“Kuranta ya Mmušo”

* **Kuranta ya Mmušo** MISC

“Disetifikeiti tša Maemo a Lenyalo”

* **Disetifikeiti tša Maemo a Lenyalo** MISC
* Temporal expressions such as date and time should be marked as MISC.

“12 February 2006”

* **12 February 2006** MISC

“31 Mopitlo 2009”

* **31 Mopitlo 2009** MISC

“re tlo kopana ka 16:00”

* re O
* tlo O
* kopana O
* ka O
* **16:00** MISC
* Names of languages and ethnic references should be annotated as MISC, e.g. “isiZulu”, “French”, “Germans”.
* ba O
* ithuta O
* **Sepedi** MISC
* le O
* **isiZulu** MISC

# 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.