Guidelines for annotating NDT-NER

1. What is an annotation?

An annotation is a markup of a text span in a specific format that indicates a feature or features of the text within the span." (MUC-7)

Each annotation has one feature, and one possible future extension:

1. All annotations must have one **type**: e.g. person, organization, movie
2. *Possible extension:* Some annotations will have one **identifier**: a Wikipedia URL that globally and uniquely identifies the named entity.

2. What should be annotated?

In the <Norwegian NER> project we annotate named entities. All names of people (fictional and real), organizations, places, products and events that are found in the texts should be annotated.

The main rules of thumb for annotating something as a name are:

- The word is **capitalized**
  Exceptions:
  - Most words in sentence-initial position
  - Names correctly spelled with initial lowercase, e.g. 'de Soto' or 'iPhone'
  - Spelling errors, e.g. 'charlie chaplin'
- The word has a **unique reference**, i.e. it points to a single thing in the world
  Exceptions:
  - Brands (e.g. 'Audi'), refers to a single brand but a range of objects/object types
  - Some titles, e.g. the reference of 'Kongen av Danmark' will change over time
- The reference is **constant over time**
  - 'Rosenborg-keeperen' is not a name, as it's reference will change over time
  - 'O.J.-saken' is a name, as it's reference does not change
- The word is or contains a **proper noun**
  - 'Per Hansen' are two consecutive proper nouns with a single reference
  - Names may include words which are not proper nouns, e.g. prepositions, as in 'Kongen **av** Danmark'
- The word is not inflected, and not possible to inflect (except for genetive form)
  - 'Per' can not be inflected: 'Perer*' or 'Peren*'
  - 'Pers' is the genitive form, we do not treat genitives as inflections in our context

    - ‘Toyotaer’ or ‘Toyota’er’ : A few names may be inflected. We annotate the inflected forms as *derived* from the original name, but not as a name in itself.

Some examples which are regarded names:

- Names: E.g., ‘Harry Hole’, 'Hole', 'Harry'
- Initials: E.g. 'H.H.'
- Spelling mistakes: E.g. 'Hary Hole'
- All orthographic variations E.g. 'harry hole', 'Harry HOLE'
- Names of fictional characters, but only when capitalized: ‘Den Grønnkledde’ and ‘Fiskerkona’.

Some examples which are regarded non-names:

- Subject headings (except when they are a proper name like described above)
- Pronouns (han, hun, De, Dem, Deres)
- Expressions of time (exceptions are time expressions that refer to specific events, such as “9/11” and “22. juli”)
- Currencies (Euro, Kroner, etc.)

2.1 Entity types

There are 6 types of entities:

- Person (PER)
- Organisation (ORG)
- Location (LOC)
- Product (PROD)
- Event (EVT)
- Miscellaneous (MISC)

In addition, there are three “special” entity types:

- Geo-political entity, GPE,with sub-types:
    - GPE_LOC (mainly a locative sense), can be mapped to LOC
    - GPE_ORG (mainly a non-locative sense), can be mapped to ORG
    - GPE_XXX (uncertain about the sense)
- Words derived from names (DRV)
- Uncertain (XXX)

Entities are not organized hierarchically now, but this is also a possible future extension.

The same proper name can belong to different types due to different contexts. Thus, the same name can refer to both the organization and the place where it is:

- Hun jobbet for Theatercaféen$_{\text{(ORG)}}$, men de møttes på Kafe Løve$_{\text{(LOC)}}$.

Stortinget is a location when something happens there, and an organisation when they decide something.

Newspapers and tv-shows are products when we refer to the product, but organisations if we refer to them as workplaces.

- Han jobber i Aftenposten$_{\text{(ORG)}}$.
- Hun fikk Aftenposten$_{\text{(PROD)}}$ på døra.

Newspapers are organizations when they are actors:

- Klassekampen$_{\text{(ORG)}}$ skriver at...

...but products when the physical newspaper is referred to:

- I Klassekampen$_{\text{(PROD)}}$ står det at...

2.1.1 Entity type: Person

The person name category includes names of real people and fictional characters. Names of animals and sometimes non-living things are also annotated as Person.

- Den kjente TV-hunden Lassie$_{\text{(PER)}}$
- Bo Obama$_{\text{(PER)}}$ is a pet dog of the Obama$_{\text{(PER)}}$ family.

Family names should be annotated as person, even if they refer to several people.

Gods are also annotated as Persons, when capitalized and having a single reference

- Gud$_{\text{(PER)}}$ spiller en stor rolle i livet.

When sentence-initial and not having a single reference, it is not tagged:

- Gud_ eller ikke, Zlatan$_{\text{(PER)}}$ er best

2.1.2 Entity type: Organization

Include any named collection of people, such as firms, institutions, organizations, pop groups, sports teams, unions, political parties etc.

Organization also includes names of places when they act as administrative units:

- Vålerenga$_{\text{(ORG)}}$ tapte mot Tromsø$_{\text{(ORG)}}$.

Corporate designators like AS, Co. and Ltd. Are to be included as part of the name.

The term “& co” in “Bjørgen & co” should not be annotated because it is a designator for unnamed persons, and not an organization or a company:

- Bjørgen$_{\text{(PER)}}$ & co gjorde rent bord

In contrast to e.g.:

- Advokatfirmaet Lie & Co$_{\text{(ORG)}}$ representerer Hansen$_{\text{(PER)}}$.
- Jeg leser ofte Donald Duck & Co$_{\text{(PROD)}}$

2.1.3 Entity type: Location

Includes geographical places, buildings and facilities. Examples are airports, churches, restaurants, hotels, tourist attractions, hospitals, shops, street addresses, roads, oceans, fjords, mountains, parks and also fictional locations.
Postal addresses are not annotated, but the building, town, county, country within the address are to be annotated, all as LOC entities.
E.g.:

- Øvregaten 2a$_{(LOC)}$, 5003 Bergen$_{(LOC)}$

2.1.4 Entity type: Products

All things artificially produced are regarded products. This may include more abstract entities, such as speeches, radio shows, programming languages, contracts, laws and even ideas (if they are named).

Brands are products when they refer to a product or a line of products, but organisation when they refer to the acting or producing entity:

- Jeg har kjøpt meg Audi A9$_{(PROD).}$ (Single reference)
- Audi$_{(PROD)}$ er den beste bilen. (Brand reference)
- Audi$_{(ORG)}$ lager de beste bilene. (Actor reference)
- Audi$_{(ORG)}$ hevder de ikke jukset med utslippene. (Actor reference)

2.1.5 Entity type: Event

Includes names of festivals, cultural events, sports events, weather phenomena and wars. Events always have a time span, and often a location where they take place as well.

An event and the organization that arranges the event can share a name, but should be annotated with different categories: event and organization.

- Quartfestivalen$_{(ORG)}$ gikk konkurs i 2008.
- Rolling Stones$_{(ORG)}$ fikk dessverre aldri spilt på Quartfestivalen$_{(EVT)}$.

2.1.6 Entity type: Miscellaneous

Includes all proper names that do not belong in the other categories. Examples are animals species, latin names and names of medical diseases and medical disorders. Things that are manufactured or produced are generally Products, whereas thing naturally or spontaneously arising are in general Miscellaneous.

2.1.7 Special entity type: Geo-Political Entity

Geo-Political Entity (GPE) - GPE entities are geographical regions defined by political and/or social groups. A GPE entity subsumes and does not distinguish between a nation, its region, its government, or its people. Geo-Political Entities are composite entities comprised of a population, a government, a physical location, and a nation (or province, state, county, city, etc.).

An entity that is a GPE, is always tagged as a GPE, but further with a sub type. We further divide GPEs into two different senses, LOC and ORG, by the following principle:

- If a sense is mostly locative, annotate as GPE_LOC
- If a sense is *not* mostly locative, annotate as GPE_ORG
- If in doubt, annotate as GPE_XXX (temporary annotation), and we'll discuss it and improve the guidelines

A GPE must be one of the following: Nation, City, Region with a parliament-like government
The following are not GPEs: Parts of cities, Roads,

Examples:

- Finlands$_{\text{(GPE\_ORG)}}$ kvinnelige president, Tarja
- Presidenten i Finland$_{\text{(GPE\_ORG)}}$
- Norge$_{\text{(GPE\_ORG)}}$ reagerer på politivolden i Catalonia$_{\text{(GPE\_LOC)}}$

Sometimes the names of GPE entities may be used to refer to other things associated with a region besides the government, people, or aggregate contents of the region. The most common examples are sports teams:

- Norge$_{\text{(ORG)}}$ slo Sverige$_{\text{(ORG)}}$ i Sveits$_{\text{(GPE\_LOC)}}$

These entities should be marked as teams (ORG), as they do not refer to any geo-political aspect of the entity.

2.1.8 Special entity type: Derived

Special category for words with proper derived from names, with the following traits:

- Contains a full name
- Is capitalized
- Is not a name itself

Examples are:

- Oslo-mannen$_{\text{(DRV)}}$ sa at …
- Høyre-leder$_{\text{(DRV)}}$ Jan Petersen$_{\text{(PER)}}$

Names that are inflected and used as nouns, are also tagged derived (DRV):

- Nobel-prisene

2.1.9 Special entity type: Unknown

If you do not know which entity type to use, mark the entity as unknown. This is important in order to collect and document all difficult cases.

Example:

- Kong Harald$_{\text{(XXX)}}$ og dronning Sonja$_{\text{(PER)}}$ (should title be part of the name here)

When using XXX, always add a note about the reason for the uncertainty

2.2 Ambiguity

Ambiguity is a frequent source of doubt when annotating. We choose to solve ambiguity in general as follows:

- Always use local context (the document) to resolve ambiguity.
- Choose the entity type based on the local context.
- We assume that every entity has a base, or literal, meaning.
- When there is ambiguity, either because of lack of context or genuine ambiguity, always choose the literal meaning of the word(s).

Examples:

- Vietnam er flott (unclear reference, choose literal meaning GPE_LOC)
- Den kalde krigen var aldri det samme etter Vietnam (choose EVT here, as it is clear that the reference is the Vietnam war)
- Ifølge Aftenposten er… (ambiguous, choose literal meaning PROD)
- Stjal krokodille fordi dyrene ved Akvariet lider (ambiguous, choose literal meaning LOC)

2.3 Examples

2.3.1 Nested names

Always annotate the whole name, never nested parts.

Annotate like this:

- Høgskolen i Oslo og Akershus$_{\text{(ORG)}}$

And not like this:

- Høgskolen i Oslo$_{\text{(GPE)}}$ og Akershus$_{\text{(GPE)}}$

2.3.2 Possession/genitive

Annotate the genitive marker as part of the name

- Harry Holes$_{\text{(PER)}}$ kontor.

Note that the possessor and possessed named entity substring should be tagged separately:

- Mortens$_{\text{(PER)}}$ Citroën$_{\text{(PROD)}}$.

2.3.3 Titles

Most titles do not have an initial capital letter in Norwegian. The exception are some instances of royal titles. We never annotate titles as a name or part of a name:

- Deres Majestet.
- Ærede Kong Harald$_{\text{(PER)}}$
- Mullah Omar$_{\text{(PER)}}$

Official work titles like ‘Fylkesmannen’ and ‘Rådmannen’ should be annotated as organizations because they refer to official institutions:

- I går ble det klart at Fylkesmannen$_{\text{(ORG)}}$ har vedtatt ...

When the same words refer to the person/occupation, they are not considered names:

- Fylkesmannen kjørte i grøfta sør for...

2.3.4 Names connected by conjunctions

Annotate names connected by conjunctions as multiple distinct entities:

- Eli$_{\text{(PER)}}$ og Carl I. Hagen$_{\text{(PER)}}$

2.3.5 Names that include numbers

Include numbers when they are part of the entity name:

- Pilestredet 48$_{\text{(LOC)}}$
- Rema 1000$_{\text{(ORG)}}$
- 1001 natt$_{\text{(PROD)}}$
- 22. Juli$_{\text{(EVT)}}$

2.3.6 Quotations around a name

Exclude quotations and other punctuation characters as part of the name:

- "Forføreren$_{\text{(PROD)}}$" av Jan Kjærstad$_{\text{(PER)}}$.
- Oslo$_{\text{(GPE\_LOC)}}$-Bergen$_{\text{(GPE\_LOC)}}$ er ganske langt…

2.3.7 Words derived from names

Words that contain names, but are not in themselves a name, and are not capitalized, should not be annotated:

- Han der newyorkeren
- New York-baserte$_{\text{(DRV)}}$ ACLU$_{\text{(ORG)}}$.
- Bergensbølgen er en betegnelse på populærmusikalske strømninger fra bergensområdet.
- Bergensbølgen er en betegnelse på populærmusikalske strømninger fra Bergen-distriktet$_{\text{(DRV)}}$.
- Folk sier at bergensbølgen er en betegnelse på populærmusikalske strømninger fra Bergen$_{\text{(GPE)}}$.

The reason for this is that these words do not have a unique entity as a reference, but rather exploit an entity as part of their semantics. Also, the derived word itself belongs to a word class not being a Proper Noun.

2.3.8 Proper names that include hyphens

All names that include hyphens should be annotated, if and only if they constitute a new name:

- Lillehammer-saken$_{\text{(EVT)}}$ (named event)

- Lillehammer-advokaten (not a name)

2.3.9 - og, - eller and slashes

Names that are combined using various conventions, should be given type by the combination of the names:

- Arbeids- og inkluderingsdepartementet$_{\text{(ORG)}}$
- HIV/AIDS-katastrofen$_{\text{(DRV)}}$
- Per- eller Arne-saken$_{\text{(EVT)}}$ (fictional example)

2.3.10 Names containing lowercased words

According to Språkrådets advice concerning capitalization, many names should only be capitalized in their first word. We still treat the following words as part of the name:

- Oslo rådhus$_{\text{(ORG)}}$ vedtok i dag.
- Oslo rådhus$_{\text{LOC)}}$ ble innviet i 1950.

2.3.11 Names and metonymical usage

There are subtle differences between metonymy, derived senses and actual new names. A few examples from the data:

- I dag skal Fr.p.-gruppen$_{\text{(DRV)}}$ si ja eller nei til en ny regjering Bondevik$_{\text{(PER)}}$.

Here, we choose the literal sense.

- Fr.p.$_{\text{(ORG)}}$ baner veien for Bondevik II$_{\text{(ORG)}}$.

In this example, however, we are talking about a specific government. It even has it’s own name, given the “II” suffix.

- I dag skal Fr.p.-gruppen$_{\text{(DRV)}}$ ta stilling til om de vil felle Bondevik-regjeringen$_{\text{(ORG)}}$. (fictional)

In this example, the Fr.p.-gruppen is interpreted as a generic term, whereas Bondevik-regjeringen has a clear reference, even though the reference is not constant.The boundaries here are not clear, however.

2.3 Questions

In cases where the locative meaning applies for an organizational area, there is sometimes doubt to whether the appropriate annotation should be GPE_LOC/ORG, or ORG.

- I år innføres ordningen i EØS
- Politiadvokat XX ved Helgeland politidistrikt sier én person er pågrepet og siktet
- Det har hittil i år vært tre pågripelser i Hordaland politidistrikt, og to i politidistriktene Hedmark og Rogaland

In cases where a PROD (typically a law or an official policy paper) is followed by a concretization (typically a number), should the concretization be included in the name?

- Som en konsekvens av forslaget, oppheves Grunnlovens § 2 annet ledd

There is often doubt related to whether a sentence-initial word should be annotated, given that many words can both be used as a proper name and a generic term.

- Sysselmannen på Svalbard har den øverste offisielle myndigheten i området
- Kirken har egne organer som kan overta de oppgaver som Kongen (regjeringen) i dag ivaretar
- Regjeringen har enda ikke uttalt seg om hendelsen
- Politiet sier at det kan være en sammenheng mellom ugjerningene

Sometimes the name is related to the context

- Mammadamen (name of a blog) sendte meg et spørsmål i forrige uke

In cases where a name is connected to a hyphen and quotation marks, should the whole construction be annotated as DRV? The punctuation marks in these cases are separated by spaces in the corpus.

- « A Game of Thrones » –rollespillet utgis neste uke
- “ Jurassic Park “ –regissøren planlegger ny film

Sometimes nicknames can contain a proper name. The question in these cases is whether the whole construction should be annotated, or just the proper name.

- Spørsmålet er hvem som vinner de 11 valgmannsstemmene i Arizona$_{\text{(GPE\_LOC)}}$, the Grand Canyon state

3. Previous work, basis of the guidelines

These guidelines are heavily based on “Guidelines for annotating NRK SIFT content fields (version 3b)”, with a few modifications:

- Added GPE for entities which are both Organisation and Locations
- Entity type WORK replaced by PRODUCT, a more general type
- Renamed PLACE to LOCATION
- Renamed OTHER to MISCELLANEOUS
- Several entities belonging to OTHER in NRK SIFT, now belongs to PRODUCT
- Words derived from names should not be annotated (e.g. “Oslo-baserte” and “bergensbølgen”)
- Fictional characters with no capitalization should not be annotated (e.g. “den grønnkledte”)
- These guidelines do not cover hierarchical entity types (out of scope for project)
- These guidelines do not cover linking of mentions (out of scope for project)
- Special entity type UNKNOWN to use for borderline cases, where discussion is needed, etc.

The guidelines are also partially based on:

- Howard, May-Helen (2014). Manuell annotering i NRK-prosjektet: Hvilke utfordringer må det tas stilling til? Bacheloroppgave, ABI, HiOA
- Jónsdóttir, A. B. (2003). ARNER, what kind of name is that?: an automatic rule-based named entity recognizer for Norwegian.
- ACE (Automatic Content Extraction) English Annotation Guidelines for Entities Version 6.6 2008.06.13
- CONLL NER shared task: http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt
- Europarl: http://web.mit.edu/andreeab/www/corpus/annotationGuidelines.pdf