Annotatierichtlijnen voor (Un)named Entities in Nederlandse Literatuur

Algemeen
Named entities in een tekst zijn instanties die verwijzen naar specifieke entiteiten in de werkelijkheid. Voorbeelden hiervan zijn: personen, locaties en organisaties. Verder zijn er unnamed entities, dat zijn instanties in een tekst die **niet** verwijzen naar specifieke entiteiten c.q. Die verwijzen naar abstracte entiteiten. Voorbeelden hiervan zijn: (persoonlijke, verwijzende en wederkerende) voornaamwoorden en algemene omschrijvingen van entiteiten (de vrouw, de dief, het gebouw en het instituut).
Een entiteit is een object of een groep van objecten in de wereld of een mentale representatie van een object.

De SoNaR annotatierichtlijnen dienen als basis voor onze annotatierichtlijnen. Ook gebruiken we de NewsReader annotatierichtlijnen als inspiratie. Het grote verschil met SoNaR is dat we het Product (PRO) label niet gebruiken en het label OBJ toevoegen aan de richtlijnen.
Verder gebruiken we minder (specifieke) sublabels.

De hoofdsoorten named entities die we onderscheiden zijn:

- Persoon (PER)
- Locatie (LOC)
- Organisatie (ORG)
- Persoonskenmerken (PKM)
- Object (OBJ)
- Miscellaneous (MISC)

We maken verder in tegenstelling tot SoNaR geen onderscheid tussen letterlijk- en figuurlijk taalgebruik in de labels. Bij figuurlijk taalgebruik wordt het label geannoteerd waar het taalgebruik opslaat. [*Dat is het brein van de organisatie.*] wordt zodoende geannoteerd als PER en [Den Haag heeft besloten om …] als ORG. Dit is alleen mogelijk als de context bekend is.

De omschrijvingen van de hoofdsoorten met hun subcategorieën zijn als volgt:

Annotatie van personen (PER)
Elk onderscheidbaar persoon of groep van personen die genoemd wordt in een tekst refereert aan een entiteit van het type Persoon. Een persoon entiteit kan bijvoorbeeld gespecificeerd worden als: naam [*Sherlock*], functie/beroep [*detective*], voornaamwoord [*hij*] en meer. Een combinatie van het eerder genoemde is ook mogelijk. Hieronder vallen ook: dode en fictieve personen, goden, families en dierennamen. We onderscheiden de volgende drie subcategorieën:

- PER.fic → fictief
- PER.nfic → Niet fictief, goden vallen hier ook onder.
- PER.misc → De rest, zoals figuurlijk taalgebruik, maar ook personen die niet expliciet genoemd worden c.q. het is niet precies duidelijk om wie het gaat (vaak fm label bij gender). Voorbeelden hiervan zijn: ‘iemand’,’mijn geest’ en ‘een gezworene’.

Annotatie van locaties (LOC)
Elk onderscheidbare geografische plaatsbepaling die genoemd wordt in een tekst refereert aan een entiteit van het type Locatie.

We onderscheiden de volgende subcategorieën:

- LOC.heelal → hemellichamen (planeten, asteroïden, de maan)
- LOC.water → natuurlijke of kunstmatige watermassa’s (oceanen, (stuw)meren, kanalen)
- LOC.bc (bevolkingscentra) → steden, dorpen, wijken, gehuchten, buurtschappen en gemeenten
- LOC.land → vermeldingen van (voormalige) landen of naties in hun geheel
- LOC.regio → toegewezen geografische niet kunstmatige locaties (gebergten/heuvelruggen) niet genoemde landsgrensoverschrijdende locaties [*West-Europa*], niet genoemde niet landsgrensoverschrijdende locaties [*Noord-Nederland*], provincies/states, districten en samengestelde entiteiten die een regio weergeven (*[de Europese Unie]*, *[het Midden-Oosten]*)[1].
- LOC.lijn (lijnlocaties) → alle eindimensionale locaties, zoals straten, tunnels, bruggen, rivieroevers (vs. de [*Kust*] = regio), autosnelwegen en specificaties van landsgrenzen. Eendimensionale monumenten zoals de Chinese of Berlijnse Muur vallen onder LOC.punt.
- LOC.punt (puntlocaties) → specifiek omschreven locaties zoals: residenties, adressen, marktpleinen, luchthavens, stations, havens, dokken, sportstadia, parken, bossen, monumenten, kerken en tempels. Pretparken, hotels, ziekenhuizen, musea en scholen vallen hier ook onder wanneer die als locatie worden genoemd (anders is het ORG).
- LOC.fic → alle fictieve locaties (de [*Gouw*], Het [*Land van Oz*])

[1] Het verschil tussen samengestelde entiteiten die een regio weergeven en niet genoemde niet landsgrensoverschrijdende locaties is dat bij de eerste er een lidwoord voor kan komen.

Annotatie van organisaties (ORG)

Een organisatie is een doelgerichte samenbundeling van kennis, vaardigheden en kracht tussen enkele personen die primair middelen en activiteiten aanwendt om te voorzien in de behoefte aan producten en/of diensten in haar omgeving[2].

We onderscheiden de volgende drie subcategorieën:

- ORG.gov (overheid) → overheidsinstanties die onderdeel zijn van, gerelateerd aan of omgaan met de structuur of zaken van de overheid, politiek of de staat. De [*Tweede Kamer*], het [*leger*], de [*AIVD/MIVD*].
- ORG.com (commercieel) → organisaties die primair gefocust zijn op het leveren/verschaffen van ideeën, producten of diensten met als doel winst te maken c.q. met een winstoogmerk. Hieronder vallen ook organisaties/bedrijven waar de overheid een aandeel in heeft, beurzen en andere handelsplaatsen. De [*Nationale (Postcode) Loterij*], [*Rabobank*], de [*KFC*].
- ORG.misc → alle andere organisaties. Die dus niet zozeer door de overheid worden gereguleerd of die niet uit zijn op zoveel mogelijk winst. Hieronder vallen onder andere: musea, bibliotheken, sportverenigingen, ziekenhuizen, jeugd-, studie- en studentenverenigingen, politieke partijen en vakbonden, paramilitaire groepen, religieuze groepen en historische organisaties/bewegingen die indertijd/achteraf een naam hadden/hebben gekregen. De [Tachtigers], [Romantici], [Barok].

Annotatie van persoonskenmerken (PKM)

Onder persoonskenmerken vallen alle innerlijke en uiterlijke kenmerken van mens en dier, die (expliciet) toebehorend aan een persoon worden genoemd (vaak gekenmerkt door een bezittelijk voornaamwoord). Entiteiten die slaan op een persoonskenmerk in het algemeen worden dus geannoteerd als MISC, bijvoorbeeld: ‘De kennis der gevolgtrekking’, ‘een permanente zwakte’, ‘een kort voorbijgaand genoegen’ en ‘een specialiteit’. Beroepen vallen niet onder persoonskenmerken.

‘Zijn been’ zal dus altijd als PKM geannoteerd worden, onafhankelijk van of het been toebehoort aan een mens, of een paard.

We onderscheiden de volgende subcategorieën:

- PKM.in → alles wat met het innerlijke van de mens te maken heeft, zoals kenmerken, eigenschappen en organen. Voorbeelden hiervan zijn: *‘mijn gestel’*, ‘*zijn egoïsme’* ,‘*haar talent’* en *‘uw brein’*. Ook kenmerken die niet gedefinieerd kunnen worden als fysieke ledematen, organen, of andere delen van het lichaam, vallen hieronder. Voorbeelden hiervan zijn: ‘*haar schelle stem*’, *‘zijn gevoel voor humor’* en ‘*haar goede reukvermogen*’.
- PKM.uit → alles wat met het uiterlijke van de mens te maken heeft, zoals fysieke kenmerken. Denk hierbij aan ledematen, organen, etc. Voorbeelden hiervan zijn ‘*rood , borstelig haar’, ‘een zeer groot hoofd’ , ‘zijn boze blik’ en ‘zijn gele tanden’.*

[2] https://nl.wikipedia.org/wiki/Organisatie

Of een entiteit de tag PKM krijgt is afhankelijk van de context. Zo zal ‘de voet’ getagd worden als PKM wanneer het de voet van een persoon betreft, maar niet als het om ‘de voet’ van een berg gaat.

Annotatie van objecten (OBJ)
Onder objecten vallen enkel fysieke voorwerpen of groep van voorwerpen (welke je aan kunt raken). Hieronder vallen dus simpelweg alle woorden die slaan op een gebruiksobjecten, voorwerpen, etenswaren/drinkwaar (zowel direct als indirect) en drugs. Voorbeelden hiervan zijn:

- *zijn flesch*
- *de winkelramen*
- *de dolk*
- *een aardappel*
- *mijn lunch*
- *den krachtigen wijn*

We onderscheiden geen subcategorieën.

Annotatie van miscellaneous (MISC)
Wanneer een entiteit niet onder een van de eerder vijf genoemde hoofdcategorieën valt, of het onduidelijk is bij welk hoofdsoort een entiteit hoort, dan annoteren we die als MISC.
We onderscheiden geen subcategorieën.