Turku NER corpus annotation guidelines

Maria Pyykönen, Miika Oinonen, Veronika Laippala, Sampo Pyysalo

Version 1.0

Contents

1 Introduction

This document details the guidelines for the annotation of the Turku NER corpus, a broad-coverage corpus for Finnish named entity recognition. The annotation follows the guidelines of the FiNER corpus Ruokolainen et al. (2019), and this document should be understood as accompanying and further detailing the existing FiNER guidelines. Some of those guidelines are restated here, but this document focuses on their application to the Turku NER corpus rather than providing an exhaustive documentation of the annotation. For any detail not found here, we refer to Ruokolainen et al. (2019) and related FiNER documentation (see Section 6).

2 General guidelines

2.1 Representation

Entity mentions are annotated as **continuous, non-overlapping spans** of text that are assigned **exactly one type** from the following categories:

- PER: person
- ORG: organization
- LOC: location
- PRO: product
- EVENT: event
- DATE: data

The start and end of each annotated span must coincide with the start and end (resp.) of a *syntactic word*. That is, each word must be either annotated as a whole or not at all. For the specific definition of syntactic word, we follow the guidelines for Universal Dependencies for Finnish.[1] (In the corpus annotation process, annotators are provided with texts with gold tokenization so that space-separated spans correspond to syntactic words per this definition.)

2.2 Annotation span

Excepting for DATE, annotated mentions typically involve one or more **proper nouns**, which are annotated when they refer to an entity of one of the targeted types. Common **head nouns** that further identify or specify an entity of an annotated type are included in the span of annotations. For example, for *Turun yliopisto*, *Suomen valtio*, and *Helsingin hovioikeus*, the nouns *yliopisto*, *valtio*, and *hovioikeus* are included in the span of the annotation.

By contrast, head words are **excluded** from the span of annotation when the noun phrase does not refer to an entity of a targeted type. For example, for *Turun murre*, *Suomen talous*, and *Helsingin tieverkko*, only the proper nouns (i.e. *Turun*, *Suomen*, and *Helsingin*) are annotated. Modifiers such as adjectives that are not relevant to identifying the entity or its type are likewise excluded from the span of annotations.

2.3 Affixes

Inflectional affixes are **included** in the span of annotations. This generally follows from the requirement that annotations either include or exclude syntactic words as a whole, but is followed also in cases where the affix is separated by space (e.g. due to tokenization or non-standard usage). For example, for *EU␣:n* and *USA␣:ssa* (where "␣" denotes space) the affixes *:n* and *:ssa* are included in the span of the annotations.

[1] https://universaldependencies.org/fi/

2.4 Coordination with ellipsis

When two or more mentions of the same type appear in a coordinated construct where part of a mention is elided, the entire expression is marked as a single entity mention, as in for example *Pohjois- ja Latinalaisessa Amerikassa* ("North and Latin America"), and *Euroopan parlamentin ja neuvoston* ("the European Parliament and the Council of Europe").

2.5 Quotation marks

When a mention or part of it appears enclosed in quotes, the quotation marks are **included** in the span of annotations. From this follows that expressions such as *"Sotareppu"*, *"Yliopisto palaa" -tilaisuuden*, and *Kirill "Kirka" Babitzinin* are all annotated in their entirety.

2.6 Hyphenated compounds

In accordance with the rule of annotating only complete (syntactic) words, hyphenated compounds in which both words refer to the same entity are annotated as a whole. Thus, for example *Oscar-palkinnon* ("Academy Award") is annotated, as the common noun *palkinnon* is used to specify the same entity that is referred to by the proper noun *Oscar*. Other examples of this include for instance *Martta-tädin* ("aunt Martta") and *BRIC-maat* ("the BRIC countries"). This practice is also followed in cases in which the hyphen is preceded by a space, and thus for example the expression *Damn Seagulls␣-yhtyeen* is annotated in its entirety.

The rule is also further extended to cover cases in which the expression *-niminen* ("named"), *-merkkinen* ("of brand"), or similar is used to attach a name to a noun that further specifies it, such as in *Accenture-niminen firma* ("a firm named Accenture") or *Volvo-merkkinen taksiauto* ("a taxi of the brand Volvo"). In cases such as these, the complete expression is included in the span of annotation.

On the other hand, hyphenated compounds in which the respective parts do not refer to the same entity are not annotated at all. Thus, for example *Oscar-ehdokkuus* ("Oscar nomination"), *Suomi-loma* ("a vacation in Finland"), and *EU-maat* ("the EU countries") are out of the scope of annotation, as they do not refer to an entity that would belong to any of the targeted types. The difference between *EU-maat* and for instance *BRIC-maat* here is that while *BRIC* is an acronym used to refer to the countries (i.e. *maat*) Brazil, Russia, India and China, *EU* refers to an organization (i.e. The European Union), and thus does not strictly refer to the same thing as *maat*.

2.7 Abbreviations and acronyms

Abbreviations and acronyms appearing as aliases and immediately following a mention are marked as a single annotation together with the full form, as in for example *Research in Motion (RIM)*. In cases where the mentions are separated by text that is not part of the name, separate annotations are created; this would be the case for for example *Research in Motion, lyhennettynä RIM*, where *Research in Motion* would constitute one annotation, and *RIM* another.

2.8 Foreign names and non-latin characters

Non-Finnish names are annotated similarly to Finnish names, even when written in non-latin characters.

2.9 Fictional entities

The names of fictional entities are annotated similarly to real-world entities, without distinction (e.g. *Galaktinen Tasavalta* and *Yavinin taistelu*).

2.10 Typographical and other errors

Text containing errors is annotated according to the annotator's best estimate of the author's intent. For example, proper nouns written in lowercase are annotated as if correctly capitalized.

3 Type-specific guidelines

3.1 PERSON

- Titles (*mr, mrs, president*, etc.) are **not included** in the span of PER annotations.
- Online handles, pseudonyms, nicknames and similar are **annotated** as PER (e.g. *@digikim, Hymy-Late*).
- Animate beings (e.g. animals) are **annotated** as PER when named (e.g. *Molla* [a cow]).
- Inanimate things are **not annotated** as person even when referred to by a human name or otherwise personified, such as *Annelin (minun autoni)* ("Anneli (my car)").
- Nationalities, ethnic groups and similar are **not annotated** (e.g. in *Kreviinit* ["Crevins"]).

3.2 ORGANIZATION

- Educational, religious, legal and similar institutions (along with their faculties and/or departments) are **annotated** as ORG when the institution in question is specifically named (e.g. *Turun yliopiston humanistinen tiedekunta* and *Suomen ortodoksinen kirkko*).
- The standalone words *yliopisto* or *(ortodoksinen) kirkko* (and similar) are **not annotated** if the expression does not name the institution. Similarly, *Poliisi* without further specification is not annotated.
- Committees and similar are **not annotated** if the organization they belong to is not named: for example *Norjan Nobel-komitea* is annotated, while *elintarvikekomitea* is not.
- Named ministries are **annotated** as ORG even when mentioned without a country, e.g. *Ulkoministeriössä* or *oikeusministeriön*, but the standalone word *ministeriö* is **not annotated**.
- *[maan] hallitus* ("[a country's] government") is **annotated** as ORG.
- The names of political parties are **annotated** also when written without an initial capital letter (e.g. *kokoomus* and *keskusta*)

3.3 LOCATION

- Head words identifying non-specific locations are **not included** in the span of LOC annotations (e.g. *Turun alue*, "the area of Turku").
- Geopolitical regions such as *Schengenin alue*, *BRIC-maat*, and *EMEA-alue* are **annotated** as LOC.

3.4 PRODUCT

- Brands related to food (e.g. *Oivariini*) are **annotated** as PRO, while types of food (e.g. *Quattro Stagioni-pizza*) are **not annotated**.
- Theories (e.g. *Einsteinin suhteellisuusteoria*), strategies (e.g. *Fischerin puolustus*) and similar are **not annotated** as PRO.
- The identifiers of EU directives (e.g. *90/426/ETY*) are **not annotated**.
- Disease names (e.g. *Hullun lehmän tauti*) are **not annotated**.
- Chemicals names (e.g. *mykotoksiini* and *kokkidiostaatti*) are **not annotated** unless specifically used to refer to an intentionally created named product (e.g. *Risperidon* and *Risperdal*®).
- Types of trains are **annotated** as PRO (e.g. *Intercity-juna*) but train lines are not annotated (e.g. *S-juna Kirkkonummelle*)
- The word *Internet* is **annotated** as PRO when capitalized; when the lowercase spelling *internet* is used, the word is **not annotated**.
- Stock indices (e.g. *OMXH* and *Stoxx Europe 600 -indeksi*), credit ratings (e.g. *BBB+*), and certification marks (e.g. *CE-merkintä*) are **annotated** as PRO.
- Named agreements, treaties and laws are **annotated** as PRO (e.g. *Euroopan yhteisön perustamissopimus*, *Kioton sopimus*, and *Suomen perustuslaki*)

3.5 EVENT

- Years are **included** in the span of EVENT annotations (e.g. *Yhdysvaltain Grand Prix 1973*).
- Named battles are **annotated** as EVENT (e.g. *Guadalcanalin laivastotaistelu* and *Samarin saaren taistelu*).
- Elections are **not annotated** as EVENT (e.g. *Belgian vaalit* and *USA:n kongressivaalit*).

Annotation guidelines for sports events and comparable competitions are detailed in Finnish in the FiNER tagger guidelines.[2] In summary,

- Events that have a proper name are always **annotated** (e.g. *Kalevan kisat*).
- Olympics are **annotated** when the reference is unambiguous, either by being unique (e.g. *Helsingin olympiakisat*) or by identifying both 1) the year and 2) either the location or summer/winter.
- Other repeated events are **annotated** if 1) the sport, 2) the year, and 3) the scope (e.g. *MM-kisat*) or the location are identified.

Examples:

- *jääkiekon MM-kisat* and *suunnistuksen MM-kisoissa* are **not annotated** (sport and scope, but no year).
- *Lahden MM-kisoissa 1989* and *Münchenin EM-kisoissa 2002* are **not annotated** (location and year, but no sport).
- *Ateenan olympialaisissa* is **not annotated** (ambiguous reference to olympics).
- *Vuoden 1998 jalkapallon maailmanmestaruuskilpailut* is **annotated** (sport, year and scope).

These guidelines are applied regardless of the scope and formality of the event; for example *Vuoden 2008 MM-kyykkä* is **annotated**.

[2] `https://github.com/Traubert/FiNer-rules/blob/master/info/annotation_guidelines.md#7-urheilukilpailut`

3.6 DATE

- The word *vuosi* (“year”) is **included** in the span of DATE annotations (e.g. *vuonna 1985*).
- Seasons (e.g. *kevät, kesä*), days of the week (e.g. *maanantai, tiistai*) and times of the day are **not annotated** and **not included** in the span of DATE annotations.
- Decades are **not annotated** (e.g. *1960-luku*).
- The words *lopussa, alussa, aikana, puolivälissä* and similar are **not included** in the span of DATE annotations (e.g. *vuoden 1990 alussa*).
- For date range expressions, the word *välillä* ("between") is **included** in the span (e.g. *vuosien 2010 ja 2011 välillä*).
- Periods that are part of dates are **included** in the span of DATE annotations also when sentence-final.

3.7 Out of scope

- Language names are **not annotated** (e.g. *suomi*, *suomen kieli*, *Basic english*).
- Species names are **not annotated** (e.g. *E. coli*).
- Currencies are **not annotated** (e.g. *Islannin kruunu*).

4 Ambiguities

4.1 ORGANIZATION vs. LOCATION

- Buildings, facilities and similar referred to by the name of an organization are annotated as **ORG**, such as *Stockmann* in *menin Stockmannille* (“I went to Stockmann”).
- Sports teams are annotated as **ORG** also when referred to by the name of a location, such as in *maailmanmestari [Brasilia] voitti [Skotlannin] 2–1* (“the world champion Brazil defeated Scotland 2–1”).
- Following the FiNER guidelines, municipalities are annotated as **ORG** (e.g. *Karkun kunta*), while provinces (e.g. *Turun lääni*), dioceses (e.g. *Memphisin hiippakunnan*), and similar are annotated as **LOC**. In addition, we extend the annotation **ORG** from municipalities to rural municipalities as well (e.g. *Mikkelin maalaiskunta*).

4.2 ORGANIZATION vs. PRODUCT

The names of magazines, websites and similar are annotated as **ORG** rather than **PRO** when used in a context that attributes intention or action to the entity (e.g. *Sekä [Yandex] että [Google] painottavat* and *[File-lehti] tuntee vastuunsa*) or identifies other properties of organizations such as having staff (e.g. *[Aamulehden] toimittaja*). By contrast, the **PRO** label is used if the entity is referred to as an inanimate physical object or abstract information entity (e.g. *Kuva löytyy [GitHubista]*).

4.3 ORGANIZATION vs. EVENT

The names of sports organizations (and similar) that share a name with a series of repeated events (such as some leagues and cups) are annotated as **ORG** rather than **EVENT** when used generally without reference to a specific timed instance of an event (e.g. *Veikkausliiga* and *Englannin Valioliigassa*). Similarly, names shared between organizations and associated regular events such as *Eurooppa-neuvosto* are annotated as **ORG** when not referencing a specific timed instance of an event.

5 Examples

This section provides annotated examples from the Turku NER and FiNER corpora organized by type. Common patterns are also identified, such as *[GPE] parlamentti* (e.g. *Euroopan parlamentti*) annotated as organization.[3] The grouping and labels follow the FiNER tagger documentation,[4] and fine-grained FiNER tagger types (e.g. LocPpl) are provided for reference.

5.1 LOCATION

Politically defined locations (LocPpl)

- **countries, states**
  - *Afganistan*, *Amerikka*, *Australia*, *Belgia*, *Brasilia*, *Britannia*, *Bulgaria*, *Egypti*, *Espanja*
- **federal states and self-governing territories**
  - *Arizona*, *Florida*, *Kalifornia*, *Texas*, *Utah*
- **provinces (historical and modern)**
  - [GPE] lääni (*Turun lääni*, *Uudenmaan lääni*, *Hämeen lääni*)
- **administrative subdivisions (cantons, prefectures, municipalities, districts; dioceses, electorates...)**
  - *Satakunta*
  - [GPE] hiippakunta (*Münsterin hiippakunta*, *Viipurin hiippakunta*)
- **settlements, i.e. cities, towns, villages**
  - *Barcelona*, *Espoo*, *Helsinki*, *New York*, *Tampere*, *Turku*
- **neighborhoods, residential areas**

Geography (LocGpl)

- **geographical, geopolitical and cultural regions**
  - *Baltia*, *BRIC-maat*, *AKT-maat*
- **continents, landmasses**
  - *Aasia*, *Afrikka*, *Etelä-Amerikka*, *Etelämanner*
- **islands, archipelagoes**
  - *Mikkelinsaaret*, *Senkakusaaret*
- **mountains, mountain ranges, summits**
  - *Appalakit*, *Kalliovuoret*, *Mammoth Mountain*
- **bodies of water (oceans/seas, lakes, rivers, springs, gulfs...)**
  - *Atlantti*, *Välimeri*, *Inarijärvi*, *Pyhäjärvi*, *Aurajoki*
- **deserts, wastelands**
  - *Saharan autiomaa*
- **forests**
- **national parks**
  - *Liesjärven kansallispuisto*, *Yosemiten kansallispuisto*, *UKK-kansallispuisto*

[3] [GPE] stands for geopolitical entity.

[4] `https://github.com/Traubert/FiNer-rules/blob/master/finer-readme.md`

Streets & Roads (LocStr)

- **street names**
  - *Tehtaankatu*, *Pensaskatu*, *Revontulentie*, *Simolankatu*
- **city squares and plazas**
  - *Elielinaukio*, *Taivaallisen rauhan aukio*
  - [GPE] ...tori (*Helsingin rautatientori*)
- **roads, highways**
  - *E70-tie*, *valtatie 2*, *seututie 955*
- **addresses**
  - *Tyynenmerenkatu 11*, *Stenbäckinkatu 11*

Structures, facilities, areas (LocFnc)

- **buildings (city halls, stadiums, temples, castles...)**
  - *Lasipalatsi*, *Kytäjän kartano*, *Halikon pappila*
  - [GPE] ...stadioni (*Sydneyn jalkapallostadioni*)
  - [GPE] linna (*Turun linna*, *Kajaanin linna*)
  - [GPE] tuomiokirkko (*Uppsalan tuomiokirkko*, *Turun tuomiokirkko*)
  - [GPE] suurkirkko (*Helsingin suurkirkko*)
- **infrastructure (bridges, tunnels, canals, dams...)**
  - *Linnansilta*, *Aninkaisten silta*
- **fortfications (city walls, gates...)**
  - *Suomenlinna*
- **other structures, large monuments and landmarks**
- **facilities (factories, power plants...)**
  - *Luonnontieteiden talo 1*, *Hanasaari B-voimalaitos*
- **rooms and spaces (auditoriums, halls...)**
  - *Louhisali*
- **designated areas and zones (military bases, garrisons, cemeteries...)**
  - [GPE] lentotukikohta (*Cagliarin lentotukikohta*)
- **harbors, airports, railway and bus stations**
  - [GPE] satama (*Futianin satama*)
  - [GPE] lentoasema ("Fornebun lentoasema", *Larnakan lentoasema*)
  - [GPE] rautatie (*Jokioisten rautatie*)
  - [GPE]-[GPE] rautatie (*Hyvinkään–Karkkilan rautatie*)

Astronomy (LocAst)

- ***Earth*, *Sun*, *Moon* when capitalized**
  - *Maa*, *Aurinko*, *Kuu*
- **other celestial bodies: planets, planetoids, moons/satellites, asteroids, comets etc.**
  - *Jupiter*, *Gliese 581 e*
- **solar systems**
- **stars and suns, constellations**

- **galaxies**
  - *Andromedan galaksi*
- **nebulae**
- **other regions and parts of the universe**

5.2 ORGANIZATION

Political organizations (OrgPlt)

- **Political parties**
  - *Keskusta*, *SDP*, *Vasemmistoliitto*, *Republikaaninen puolue*, *Piraattipuolue*
  - [GPE] ...puolue (*Norjan työväenpuolue*)
- **Political youth organizations**
  - *Naši-nuorisojärjestö*
- **Legislatures (parliaments)**
  - [GPE] kongressi (*Yhdysvaltain kongressi*)
  - [GPE] parlamentti (*Euroopan parlamentti*, *EU-parlamentti*, *Wallonian parlamentti*)
  - [GPE] senaatti (*Belgian senaatti*, *Yhdysvaltojen senaatti*)
  - [GPE] komissio (*Euroopan komissio*)
- **Governments**
  - [GPE] hallitus (*Yhdysvaltain hallitus*, *Venäjän hallitus*)
  - [GPE] hallinto (*Ukrainan hallinto*)

Cultural organizations (OrgClt)

- **Bands, choirs, orchestras**
  - *Radiohead*, *Rage Against The Machine*, *Queen*
  - [GPE] ...orkesteri (*Lontoon sinfoniaorkesteri*)
- **Theatre, ballet, and opera companies**
  - [GPE] ...teatteri (*Malmön kaupunginteatteri*, *Korvenkylän kesäteatteri*)
  - [GPE] ...baletti (*Viron baletti*)
- **Other perfoming groups and troupes**
- **Museums, galleries**
  - *Kansallismuseo*, *Ateneum*
  - [GPE] museo (*Kainuun museo*)
  - [GPE] taidemuseo (*Helsingin taidemuseo*)

Media (OrgTvr)

- **News agencies**
  - *Reuters*
- **Broadcasting companies**
  - *Yle*, *BBC*
- **Television channels**
  - *Yle TV1*, *MTV3*
- **Radio stations**

  - *Iskelmäradio*, *YleX*
- **Newspapers, magazines, journals, periodicals and other publications**
  - *Helsingin Sanomat*, *Taloussanomat*, *The Wall Street Journal*, *Der Spiegel*
- **News portals and sites**
  - *Cnet News.com*, *The Hacker News*, *Yahoo News*

Financial organizations (OrgFin)

- **Banks**
  - *Nordea*, *Ålandsbanken*, *Handelsbanken*, *Deutsche Bank*
  - [GPE] pankki (*Suomen Pankki*)
  - [GPE] keskuspankki (*Euroopan keskuspankki*)
  - [GPE] kansanpankki (*Kiinan kansanpankki*)
- **Funds**
  - *Valto Takalan rahasto*
  - [GPE] ...rahasto (*Suomen Kulttuurirahasto*, *Varsinais-Suomen rahasto*)
- **Stock exchanges**
  - [GPE] pörssi (*Helsingin pörssi*, *New Yorkin pörssi*)

Educational organizations (OrgEdu)

- **Schools and educational institutes, including universities**
  - *Aalto-yliopisto*, *Massachusetts Institute of Technology*, *MIT*,
  - [GPE] ...koulu (*Kouran koulu*, *Someron yhteiskoulu*)
  - [GPE] ...korkeakoulu (*Turun kauppakorkeakoulu*, *Turun ammattikorkeakoulu*)
  - [GPE] työväenopisto (*Helsingin työväenopisto*)
  - [GPE] lukio (*Viipurin lukio*, *Tallinnan lukio*)
  - [GPE] lyseo (*Savonlinnan lyseo*)
  - [GPE] ...opisto (*Rauman merenkulkuopisto*)
  - [GPE] ...yliopisto (*Turun yliopisto*, *Tokion teknillinen yliopistosto*, *Pietarin valtionyliopisto*)
  - [GPE] akatemia (*Turun akatemia*, *Ranskan akatemia*)
  - [GPE] ...akatemia (*Yhdysvaltain tiedeakatemia*, *Yhdysvaltain laivastoakatemia*)
- **Faculties and departments within schools**
  - [GPE] ...yliopiston ... laitos (*Turun yliopiston hoitotieteen laitos*, *Michiganin teknisen yliopiston fysiikan laitos*)
- **Seminaries**

Athletic organizations (OrgAth)

- **Sports clubs**
  - *Arsenal*, *Espoon Blues*, *HIFK*, *Jokerit*, *Kärpät*, *San Jose Sharks*
- **Racing teams**
  - *McLaren*
- **Sports leagues (not competitions)**
  - *SM-liiga*, *Veikkausliiga*, *Bundesliiga*, *NBA-liiga*
  - [GPE] ...liiga (*Venäjän liiga*, *Englannin liiga*, *Englannin Valioliiga*)

Corporations & Miscellaneous (OrgCrp)

- **Corporations, companies, businesses**
  - *Apple*, *British Gas*, *Google*, *Huawei*, *Intel*, *Motorola*, *Siemens*, *Volvo*, *Yandex*
- **Societies, associations, fraternities/sorotities, orders of chivalry etc.**
  - [GPE] kauppakamari (*Turun kauppakamari*)
- **Boards, councils, committees**
- **Comissions**
  - [GPE] ...komissio (*Yhdysvaltain viestintäkomissio*, *Yhdysvaltain arvopaperikomissio*)
- **Judiciaries, courts**
  - [GPE] käräjäoikeus (*Espoon käräjäoikeus*)
  - [GPE] hovioikeus (*Turun hovioikeus*)
  - [GPE] korkein oikeus (*Yhdysvaltain korkein oikeus*, *Uuden-Seelannin korkein oikeus*)
  - [GPE] ...tuomioistuin (*Turkin perustuslakituomioistuin*, *Euroopan unionin tuomioistuin*, *Euroopan ihmisoikeustuomioistuin*)
- **International/supranational organizations and unions**
  - *Euroopan unioni*, *EU*, *Yhdistyneet kansakunnat*, *YK*, *Punainen Risti*
- **Public administration and authorities (ministries, bureaus, agencies, offices etc.)**
  - *Ulkoministeriö*, *oikeusministeriö*, *liikenne- ja viestintäministeriö*
  - *Supo*, *FBI*, *NSA*, *Nasa*
  - [GPE] ...ministeriö (*Yhdysvaltain puolustusministeriö*, *Suomen ulkoministeriö*)
  - [GPE] ...virasto (*Yhdysvaltain tiedusteluvirasto*, *Yhdysvaltojen turvallisuusvirasto*, *Yhdysvaltain patenttivirasto*, *Turkin viestintävirasto*)
  - [GPE] ...hallinto (*Yhdysvaltain avaruushallinto*, *Yhdysvaltain ilmailuhallinto*, *Venäjän avaruushallinto*)
  - [GPE] avaruusjärjestö (*Euroopan avaruusjärjestö*)
  - [GPE] kaupunginvaltuusto (*Someron kaupunginvaltuusto*)
- **Various groups, alliances, leagues etc.**
- **Dynasties**
- **States and municipalities (as organizations)**
  - [GPE] valtio (*Suomen valtio*)
  - [GPE] kunta (*Vimpelin kunta*, *Nurmon kunta*)
- **Criminal, terrorist, and paramilitary organizations**
  - *Taleban*
  - [GPE] punakaarti (*Tikkurilan punakaarti*)
- **Law enforcement**
  - [GPE] poliisi (*Suomen poliisi*, *Australian poliisi*, *Kalifornian liikkuva poliisi*)
  - [GPE] poliisilaitos (*New Yorkin poliisilaitos*)
- **Military, armed forces**
  - [GPE] laivasto (*Yhdysvaltain laivasto*)
  - [GPE] ilmavoimat (*Puolan ilmavoimat*, *Suomen ilmavoimat*)
  - [GPE] asevoimat (*Puolan asevoimat*)
  - [GPE] puolustusvoimat (*Suomen puolustusvoimat*)
  - [GPE] suojeluskunta (*Vimpelin suojeluskunta*)
- **Religious organizations (churches, congregations, cults, sects...)**
  - [GPE] ortodoksinen kirkko (*Suomen ortodoksinen kirkko*)
  - [GPE] ...seurakunta (*Tampereen helluntaiseurakunta*)

5.3 PERSON

(Human) persons (real or fictional) (PrsHum)

- **Personal names (including given names, family names, patronymics etc.)**
  - *Ahtisaari*, *Barack Obama*, *Bushin*, *Jokinen*,*Väyrynen*, *Vladimir Putin*
  - *Libuše*, *Hypatia*, *Darth Vader*
- **Families and family names**
  - *Baudelairet Marx-veljekset*
- **Aliases, pseudonyms, nicknames, usernames**
  - *Bono*, *Kim Dotcom*, *Stilsu*, *Marko "Fobba" Forss*, *Q*, *@digikim*

Animals (PrsAnm)

- **Pets, domestic animals etc. with names**
  - *Molla*

Mythical beings (PrsMyt)

- **Deities**
  - *Jumala*, *Poseidon*
- **Mythical and fictional creatures**

Note: the FiNER tagger category PrsTit (titles) marking titles that precede personal names is **not annotated** in the Turku NER corpus.

5.4 PRODUCT

- **Software**
  - *Linux*, *Skype*, *Windows*
- **Services and websites**
  - *Facebook*, *Google*, *Pirate Bay*, *Twitter*, *Youtube*
- **Hardware, consumer electronics**
  - *iPhone*, *Windows Phone*, *Galaxy S5*, *Nokia Lumia 520*, *MacBook Air*
- **Literature and poetry**
  - *Raamattu*, *Psalttari*
- **Artwork**
  - *Mona Lisa*, *Teuvo , maanteiden kuningas*
- **Films, plays, television programs**
  - *The Blues Brothers*, *Blade Runner*, *Kesäyön unelma*, *Simpsonit*
- **Video games**
  - *Angry Birds*, *Minecraft*, *Saints Row IV*
- **Pharmaceuticals and narcotics**
  - *Viagra*, *Risperidon*
- **Agreements and treaties**
  - [GPE] sopimus (*Kioton sopimus*, *Pariisin sopimus*)

  - [GPE] perustamissopimus (*Euroopan yhteisön perustamissopimus*)
  - [GPE] ...sopimus (*Kielin rauhansopimus*, *Varsovan yleissopimus*)

- **Legislation (laws, acts...)**
  - *Patriot Act*, *Obamacare*
  - [GPE] perustuslaki (*Suomen perustuslaki*, *Puolan tasavallan perustuslaki*)
- **Projects, operations**
  - *Atlas-projektin*, *OpenSSH-projekti*, *Open Computer -projekti*
- **Weapons (mostry firearms and explosives)**
  - *Sig Sauer Mosquito -pistooli*
- **Awards, prizes, trophies**
  - *Grammy*
- **Vehicles and vessels (cars, trains, ships, aircraft, space shuttles...)**
  - *Tesla*, *Model S*, *Peugeot 205*, *Opportunity-mönkijä*, *Kansainvälinen avaruusasema*
- **Food & beverages**
  - *Kevyt olo -kivennäisjuoma*
- **fruit and vegetable cultivars (capitalized)**
- **rare instances or relics and artifacts**

5.5 EVENT

- **Wars, conflicts, battles**
  - *Toinen maailmansota*, *kansalaissota*, *talvisota*, *kylmä sota*
  - [GPE] ...sota (*Suomen sisällissota*)
  - [GPE] ...taistelu (*Portinhoikan taistelu*, *Guadalcanalin laivastotaistelu*)
- **Uprisings, revolutions**
  - [GPE] vallankumous (*Ranskan vallankumous*, *Venäjän vallankumouksen*)
- **Crises**
  - [GPE] kriisi (*Krimin kriisi*)
- **Concerts**
  - *Zoo TV*, *Idols tekee hyvää-konsertti*
- **Exhibitions, biennals, and other cultural events**
  - , *Shanghai EXPO-maailmannäyttelyn*, *Grammy-gaala*
- **Sports competitions, Olympic games and other sporting events**
  - *vuoden 1988 talviolympialaiset*, *Amsterdamin olympiakisojen*, *Euroopan GP*
- **Festivals, fairs, conventions, expos**
  - *Flow-festivaalit*, *CES-messut*, *Google I/O*
- **Conferences, meetings, summits**
  - *Build-konferenssi*, *YK:n ilmastonmuutoskonferenssi*

5.6 DATE

- **Years (not decades, centuries or millenia)**
  - *1994*, *2008*, *vuonna 2007*, *vuonna 85 eaa.*, *vuosina 2005 – 2009*
- **Months (of a certain year)**
  - *tammikussa*, *helmikuussa*, *loka–joulukuussa*, *vuoden 1943 toukokuussa*
- **Days of said months**
  - *9. joulukuuta*, *9. tammikuuta 2018*, *7 päivänä marraskuuta 2005*
- **Combinations of above (full or partials dates)**
- **Date formats DD.MM.YYYY, YYYYMMDD, YYYY-MM-DD, YYYY/MM/DD**
  - *9.4.*, *1.1.2010*, *4.–5.11.2004*

6 Additional references

- Ruokolainen et al. 2019: `https://arxiv.org/pdf/1908.04212.pdf`
- FiNER tagger documentation:
  `https://github.com/Traubert/FiNer-rules/blob/master/finer-readme.md`
- FiNER annotation guidelines (in Finnish):
  `https://github.com/Traubert/FiNer-rules/blob/master/info/annotation_guidelines.md`

References

Ruokolainen, T., Kauppinen, P., Silfverberg, M., and Lindén, K. (2019). A finnish news corpus for named entity recognition. *Language Resources and Evaluation*, pages 1–26.