# Finnish toxicity annotation guidelines

## Annotation guidelines for “threat”

**Perspective definitions**:

* **THREAT:** Describes an intention to inflict pain, injury, or violence against an individual or group.
* **THREATENING:** Language that is threatening or encouraging violence or harm, including self-harm.

**Guidelines for Finnish**:

* Expressions of intent to self-harm are annotated as threat.
  + Example: “kukaan ei välitä musta, tapan itteni” → threat.
* Incitement to violence is annotated as threat.
  + Example: “Uusnatseille kuula kalloon vaan.” → threat.
* Incitement to self-harm is annotated as threat.
  + Example: “tehkää itsemurha.” → threat
* Only threats toward people should are annotated as threat.
  + Example: “KUOLEMA kakkiville KOIRILLE” → **not** threat.
* Descriptions of others’ intent to do harm are not annotated as threat.
  + Example: “Hägglund haluaa läskit kaasukammioihin” → **not** threat
  + Example: “mies huuteli: minä tapan sinut!” → **not** threat
* Threats phrased as hypotheticals are annotated as threat.
  + Example: “jos tuollainen tulisi vastaan niin vetäisin turpaan” → threat
* Counterfactuals are not annotated as threat.
  + Example: “jos olisin ollut tälläinen silloin, olisin lyönyt sen tajuttomaks!” → **not** threat
* Unless explicitly marked as fiction or sarcasm, threats are annotated as threat even if they are implausible.
  + Example: “Minun tehtäväni on kuohita julkisesti Punaisella torilla tsaari Vladimir Putin.“ → threat

**TODO:**

* “Wishes” of harm without stated intent to cause or incite, e.g. “toivottavasti hukut.”

## Annotation guidelines for “insult”

**Perspective definition**:

* **INSULT:** Insulting, inflammatory, or negative comment towards a person or a group of people. Such comments are **not** necessarily identity specific.

**Guidelines for Finnish**:

* Comments that use terms that are insulting towards groups of people are annotated as insult.
  + Example: “Mistähän voisi saada lisää matuja” → insult
* Negative comments towards self are **not** annotated as insult.
  + Example: “olen liian ruma jotta naiset tekisivät aloitteita” → **not** insult
* Negatives clearly marked as hypothetical are **not** annotated as insult.
  + Example: “Jos olet liian lihava, sinun täytyy maksaa” → **not** insult
* Negative comments towards targets other than people **are** not annotated as insult.
  + Example: “LG on surkea televisio” → **not** insult

**Disagreements:**

* Unclear or hypothetical target**,** e.g. “Oliko nää paskanaamoja ?” or “Ja pelkureita.”
* Hys hys korkki kiinni ja tutimaan. Mitäs suvakki tähän?
* Monta syytä lentää kanarialle. Sinulla yksi syy, ettet lennä sinne, olet persaukinen.
* [...] Hävetkää te keskustan lihasyöjät.
* moukkamaisuutenne huittislaiset. Ars longa,vita brevis.

**Majority unsure**:

* Jobbari on muuten yksityinen kauppias ilman y tunnusta eli kusettaja eli taas ja taas YKSITYINEN.
* Oletko säkin innostunut sensuroinnista? Mä tossa olen pyytänyt putsaamaan pari tökeröä seksivammaisen viestiä ja niin lähti ja enemmän kuin pyysinkään.

**Questions**:

* How to label negative comments without insulting terms, e.g. “Koska te "suvaitsevaiset" koko ajan haukutte ja pilkkaatte suomalaisia olette te rasisteja. Joka harjoittaa rasismia on rasisti.”

## Annotation guidelines for “identity attack”

**Perspective definition**:

* **IDENTITY ATTACK:** Negative or hateful comments targeting someone because of their identity.

**Guidelines for Finnish**:

Negative statements towards groups of people that nobody self-identifies with are not labeled as identity attack

* Example: “Pummit hirteen” → **not** identity attack
* Example: “Vihaan sinunlaisiasi natsi-idiootteja.” → not identity attack

TODO:

* “Suomalainen nainen on poispilattu” -> identity attack
* “Kusessa olisimme islamistien kanssa ilman USA,ja.” -> identity attack
* Thinly veiled negative comments are negative even if no explicit negative language is used, e.g. “Totta on, että vähemmistönä ollaan oltu varmaan koko foorumin olemassaolon ajan. Jakuten aloittaja totesi, homoudessa on tulevaisuus. Toivon, että kun 20-30 vuoden päästä ihminen laskeutuu Marsin pinnalle ovat nämä urhoot kokeneita homomiehiä ja julistavat uuden planeetan homojen planeetaksi. Tulevaisuudessa Mars asutetaan 100% homoilla ja kuten aloittaja meinasi, tarvittaessa joitain valioyksilö naisia säilytetään steriilikopeissa. Joka tapauksessa yriteään päästä heteroudesta eroon mahdollisimman pian maapallon pelastamiseksi. Käsittääkseni homot eivät myöskään ole kovin sotaisaa kansaa.”

Open question:

* Is political identity an “identity” for the purposes of this labeling?

## Annotation guidelines for “toxicity”

**Perspective definitions**:

* **TOXICITY:** A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.

**Guidelines**:

* Mild and/or humorous use of swearwords is not necessarily toxic (TODO: example)
* Neutrally or positively presented sexually explicit content is **not** annotated as toxicity

(TODO: example)

* Unreasonably expressed negative Comments can be toxic regardless of the target, e.g. organizations and people not present in the conversation

**Examples**:

* “Suomen ei pitäisi ryhtyä imperialistisen roistovaltion sätkynukeksi.” -> toxicity (unreasonable)
* “kuinka väitetään että exät kuuluvat ex miesten elämään LASTEN takia, heillä oikeus vaikka mihin LAPSEN takia, ja hups, sitten kun se lapsi kasvaa, eikä hänen nimissään enään syytä pitää yhteyksiä,alkaa kyttäys!!!!kyllähän nyt exä saa häiritä ja terrorisoida, mutta nykysen vika!!!=) exällä on oikeus tuhota ex miehen perhe jos haluaa, sillä hän on miehen lapsen ÄITI ja lapsen edun nimissä voi tehä ihan mitä vaan!!!sairasta!!!!! kyllä luulis olevan exillä parempaaki tekemistä kun puuttua toisten elämään, ja häiritä jatkuvasti exää!! nimim.kokemusta on!!! jos minun exääni törmäisin meidän lähisiwassa, niin varmaan sekoisin, koska todellakin hänenkin maitopurkin hakureissulleen tulisi matkaa 100km..=)” -> toxicity (unreasonable “style”)
* “otetaanko huomioon myös raiskatut ja ylihyväillyt eläimet.kelpaako tyhjät bensakannut.” -> toxicity (unreasonable)
* “syönyt olet kissaa ja siitä taudin sait se kun ei kuumennettaessa se basilli kuoleuvu. paappa osote tähän sivustolle, niin huomenna et enää turhaan pilaa ilmastoa pahalla henkitykse hajulla pyyntös on kuultu ja käsky on käynyt” -> toxicity (unreasonable)
* “Entisajan kylähullut eivät tappaneet ihmisiä, koska niitä pidettiin herranhuhteessa eikä syötetty helvetillisiä mömmöjä sekoittamaan päätä. Sellaiset kylähullut jotka eivät pystyneet huolehtimaan itsestään, laitettiin sitten hullujenhuoneelle (ns. mielisairaala) Nykyään vähän väsähtänyttä tyyppiä jolla on sydänsuruja tai vastaavia, palvotaan jumalana ja pidetään kuin kukkaa kämmenellä. Ja tyyppiin lisäksi upotetaan ns. onnellisuuspillereitä (SSRI- tai vastaavia mömmöjä) niin että varmasti polla on sekaisin. Johan siinä menee vahvempikin tyyoppi sekaisin. No, lopputulos on nähtävissä, 2-kymppisenä pannaan eläkkeelle ja sitten elämä onkin ihan varmasti sekaisin. Saa syödä onnellisuuspillereitä ihan tosissaan että elämä "tuntuu elämältä". Ja työtätekevät maksavat tuonkin kaiken …”. -> toxicity (unreasonable)
* “Rullalaudat liekeissä Alkoi ottamaan päähän se ristuksen kolina meidänkin pihassa, ja niin ryhdyin operoimaan asian hyväksi. Viime viikonloppuna otin hiukan hapanta ja aamuyöstä rakentelin sitten kunnon roihun telineistä. Käristelin parit HK:n siniset sinapin kera ja ongelmat oli ratkaistu. Seuraavana päivänä kävin naureskelemassa ja kusin nuotioon.” -> not toxic

**TODO**:

* What to do about negativity toward unknown parties, e.g. “Metelöivät palstalla, kun eivät muuta uskalla, pelkäävät joutumista naurunalaiseksi.” or “​​Minäkin toivoisin jonkun noista " läskimahat- sossuperikset " jne. kirjoittajista kertovan mitä heillä liikkuu pääkopassa oikeasti , Saavat ilmeisesti hyvää mieltä ja onnellisen tunnetilan saatuaan purkaa pahaa oloaan.”

## Annotation guidelines for “severe toxicity”

**Perspective definition**:

* A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words.

**Guidelines**:

* TODO: rewrite: “not to consider sexually explicit content alone as sufficient for severe toxicity, regardless of how explicit”
* TODO: A single insult, threat, etc. can be severely toxic; multiple toxic “elements” are not required to assign the label
* TODO: A comment should be labeled severely toxic if it contains content that is severely toxic according to the above criteria, regardless of whether the comment *also* contains substantive content.
* TODO: The label should be assigned regardless of whether the person or group that is the target of toxicity is assumed to be present in the discussion.

**Examples**:

* “Lestaatiolaiset ovat typeriä ja itsekkäitä paskanjauhajia, jotka tunkevat pitkät nenänsä myös toisten asioihin. Viisas haistattaa paskat moiselle roskasakille. Se on vain kärpäsenpaskan kokoinen lauma häiriintyneitä sekopäitä jossain korvessa.” -> severe toxicity

## Annotation guidelines for “obscene”

**Perspective definition**:

* **PROFANITY: Swear words, curse words, or other obscene or profane language.**

**Guidelines**:

* Also “mild” swear words such as “hitto”, “kusessa”, etc. are annotated as obscene
  + Example: Hitto millä tollaiset saa kuriin! → obscene
* Mispelled or other variations of swear words
  + Example: PERKULE! Ei tällainen ole sopivaa → obscene
* Swear words where some letters have been “masked” are also annotated as obscene
  + Example: “hän voisi mennä vaikka hevon p....see...n.” → obscene
* References to sex or sexual topics are **not** automatically annotated as obscene, but comments using more explicit sexual terminology than topically or contextually appropriate **are** annotated as obscene.

TODO:

* Sadehousut jalassa Vaimoni ei tykkää sadeasufetissistäni mutta vietän ihanaa vapaapäivää sadehousut jalassa ja käyn suihkussa koko ajan sadepöksyissä. Pvc vasten ihoani on ihaninta mitä tiedän.
* Tapahtui henkinen ylöstempaus. Tämä tarkoittaa, että pelastustie on nyt sulkeutunut. Mitenköhän tämä miekkonen kehtaa katsoa itseään peilistä??? Sen verran paksua tuubaa äijä päästele. Vai onko sillä metodina että rahat pois holmöiltä, ja niitähän rittää.
* “sontaa”