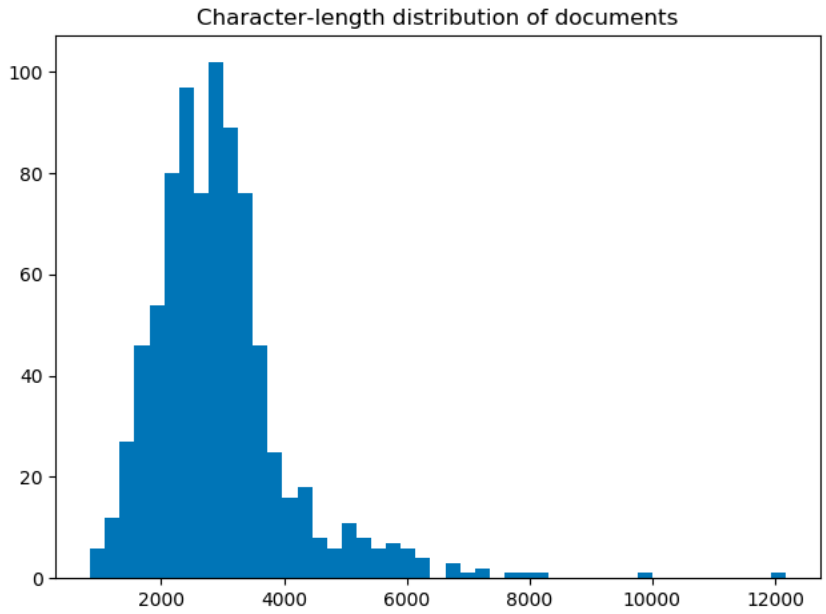
# Navodila za označevanje korpusa SentiCoref 1.0

Verzija 1.0 20190923

## 1 Uvod

Podatkovna množica SentiCoref 1.0 vsebuje besedila iz korpusa *SentiNews 1.0 (Bučar, 2017)*. Korpus vsebuje 10.427 dokumentov, kjer je sentiment označen na nivoju dokumenta, odstavka in posameznega stavka.

Korpus *SentiCoref 1.0* vsebuje 837 dokumentov, izbranih iz korpusa *SentiNews 1.0*. Dokumenti so bili izbrani na podlagi števila imenskih entitet, ki jih odkrije orodje Polyglot in le ti vsebujejo med 50 in 73 imenskih entitet. Porazdelitev dolžine člankov glede na število znakov vidimo na sliki Slika 1.




*Slika 1: Porazdelitev dolžin dokumentov v korpusu SentiCoref 1.0 glede na število znakov.*

### 1.1 Postopek označevanja

Osnovni korpus že vsebuje avtomatsko označene imenske entitete prepoznavalnika Polyglot.

Postopek označevanja v okolju Webanno naj poteka sledeče:

1. Označimo vse imenske entitete v dokumentu in popravimo napačno označene entitete avtomatskega označevalnika.
2. Označimo vse koreferenčnosti.

3. Ponovno preberemo besedilo dokumenta in zadnji omenitvi vsake izmed entitet priredimo ustrezen sentiment.
4. Označimo dokument kot končan.

Pri označevanju sledimo spodnjim navodilom. V primeru vprašanj se posvetujemo s celotno skupino označevalcev preko skupine na Slack-u.

# 2 Označevanje imenskih entitet

V osnovno pripravljenem korpusu SentiCoref 1.0 so pred-označene entitete s prepoznavalnikom entitet Polyglot. Tipi imenskih entitet, ki jih v korpusu označujemo, so:

- osebno ime,
- organizacija in
- zemljepisno ime.

Prvi slovenski korpus z označenimi imenskimi entitetami je korpus ssj500k *(Krek in sod., 2019*), ki pa ne vsebuje dovolj dolgih besedil za smiselno označitev koreferenčnosti. Nad tem korpusom je bil tudi razvit prepoznavalnik imenskih entitet za slovenščino *(Štajner, Erjavec in Krek, 2013)*.

Problematični primeri (Projekt »Sporazumevanje v slovenskem jeziku« specifikacije za učni korpus):

- kot lastna imena označujemo kratična imena:
    - geografskih oz. političnih enot: EU, RS, ZR, itd. (označimo kot zemljepisno ime)
    - političnih strank: CDU, SD, SDS, itd. (označimo kot organizacijo)
    - podjetij in organizacij (tudi njihovih delov): BBC, BTC, DURS, DZ, RTV, SFOR, itd. (označimo kot organizacijo)
    - kratice zakonov: ZDR, ZJU, ZSPJS, itd. (stvarno ime, ne označujemo)
- kot lastnih imen ne označujemo delov kratičnih poimenovanj blagovnih znamk ali imen programskih orodij: M30X, Pocket PC, Palm OS, Android OS, itd.

# 3 Označevanje koreferenčnosti

Odkrivanje koreferenčnosti je združevanje vseh omenitev na posamezno entiteto. Za slovenski jezik je bil v letu 2018 označen koreferenčni korpus coref149 *(Žitnik in Bajec, 2018).* Spodnja navodila so posodobljena navodila za označevanje koreferečnosti v korpusu SentiCoref 1.0.

Osnovni koncepti:

- *Entiteta* predstavlja enega ali več objektov. V besedilu so lahko omenjene (t.j. referencirane) s svojim imenom, različnimi samostalniškimi besedami/besednimi zvezami ali zaimki. V slovenskem jeziku je lahko omenitev vsebovana tudi v povedku.
- *Omenitev* je omemba oz. referenca na posamezno entiteto v besedilu. Omenitve niso le imenske entitete, ampak vsak del besedila, ki se sklicuje na določeno entiteto.
- *Koreferenčnost* je povezava med dvema omenitvama. Zaporedne povezave tvorijo koreferenčne verige, ki jih pretvorimo v skupine omemb, kjer vsaka skupina predstavlja entiteto.

## 3.1 Smernice za označevanje

Pri označevanju bomo označili dele besedila, ki se nanašajo na entitete tipov:

- oseba ali skupina oseb (*[Marijan Schiffrer]*, *[poslanec SKD]*, *[zamejci]*),
- organizacija (*[Biotehniška fakulteta], [BTF], [Kmetijski poskusni center Jable]*),
- zemljepisno ime - mesto, kraj (*[Jable], [Rodica], [Maribor]*).

Označili bomo vse možne omenitve, ki se sklicujejo na entiteto zgornjega tipa in so izrecno omenjene v besedilu. Če entiteta v celotnem dokumentu ni omenjena poimensko oz. je sklic nanjo le z zaimkom, je ne označimo (npr. omenitve, ki kažejo na pripovedovalca v prvoosebnem besedilu). Če se neka entiteta pojavi še v drugi, razširjeni entiteti, sta obe entiteti predstavljeni kot popolnoma svoji entiteti (npr. oseba, ki nastopa v besedilu sama, in skupina oseb, ki vključuje tudi to osebo). Skupin entitet ne označujemo.

V korpusu označujemo samo obstoj koreferenčnosti, ne definiramo pa različnih tipov koreferenčnih povezav. Npr. *[nacionalni TV hiši [ZDF]$_2$ in [ARD]$_3$]$_1$* smo označili s tremi omenitvami, saj *ZDF* in *ARD* pomenita vsaka svojo entiteto, celoten primer pa novo – združeno entiteto, ki predstavlja obe televizijski hiši skupaj.

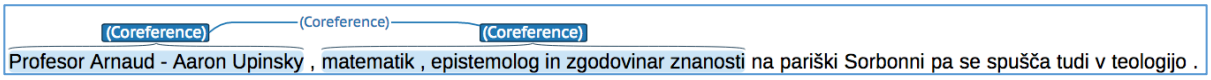
Pri označevanju omenitev, ki so v besedilu dopolnjene z levim prilastkom (angl. *predicate complement*, *premodifier*), označimo celotno omenitev. Npr. v besedilu *[strokovnega direktorja [KC]$_2$ Zorana Arneža]$_1$* označimo le eno omenitev, ki se nanaša na osebo *Zoran Arnež*. V tem besedilu je niz *Zorana Arneža* označen tudi kot imenska entiteta. Tretja omenitev se sklicuje na *Klinični center*. V primeru pristavkov in desnih samostalniških prilastkov do tega ne prihaja, saj je omenitev navadno dopolnjena z dodatnim besedilom, ki je pri pristavku ločena še z ločilom. Primera tovrstnih omenitev sta *[Mag. Franc Avberšek]$_1$ , [direktor [velenjskega podjetja]$_3$]$_1$* in *[Olimpijskega komiteja [Slovenije]$_2$]$_1$ ( [OKS]$_1$ )]*, pri čemer *velenjsko podjetje* in *Slovenija* nista povezani z ostalima dvema omenitvama.

Poleg omejitve na omenitve, ki so v besedilu izrecno zapisane, posebno pozornost namenjamo še elipsi (izpustu) osebka. Če je omenitev na entiteto v vlogi osebka v povedi izpuščena (Holozan 2015: 62), kot omenitev označimo del povedka, iz katerega se lahko razbere koreferenčnost. Primeri povedkov in povedkovih vezi so *[je lepa]*, *[je bila] lepa*, *[postal je] učitelj* ali *[mora] delati*. V povedi označimo le prvo povedkovo določilo oz. povedek, če ni določilo zapisano ločeno. Če se v povedi nahaja še drugi tip omenitve na entiteto, v tej povedi označimo le to omenitev in ne dodatno še povedka. Dobesedni navedek obravnavamo kot novo poved.

### 3.1.1 Apozicije

Razširitev omenitve s specifičnim nazivom.

Primera:




*Slika 2: Druga omenitev predstavlja specifično razlago, ki se nanaša na prvo omenitev.*

*Slika 3: Specifična razlaga predhodne koreferenčne omenitve.*

# 4 Označevanje sentimenta

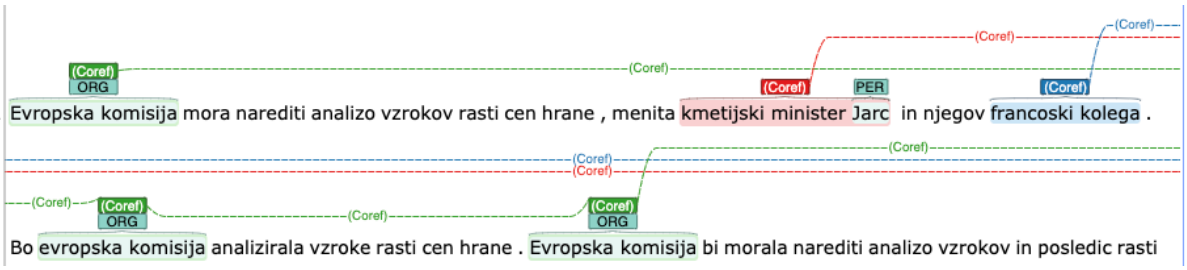
Podobno, kot je označen korpus SentiNews 1.0, bomo v korpusu SentiCoref 1.0 označili entitete po pet-stopenjski Lickertovi lestvici:

- 1 - Zelo negativno,
- 2 - Negativno,
- 3 - Nevtralno,
- 4 - Pozitivno ali
- 5 - Zelo pozitivno.

V vsakem dokumentu bomo eno izmed oznak sentimenta priredili vsaki izmed označenih entitet.

# 5 Primer označevanja

Spodaj je primer označevanja za dokument 1.txt.



cen hrane po posameznih sektorjih , sta po današnjem srečanju v Ljubljani povedala minister za kmetijstvo Iztok Jarc in

njegov francoski kolega Michel Barnier . Sicer imata po njunih besedah Slovenija in Francija podobna stališča glede skupne

kmetijske politike in podobne pomisleke glede reforme vina . Poudarila sta pomen dobrega sodelovanja med državama tudi na

področju kmetijstva v času predsedovanja EU prihodnje leto . Med možnimi ukrepi EU je Barnier omenil obnovo zemlje

v prahi . Gre za štiri milijone hektarjev zemlje , obnova bi omogočila večjo proizvodnjo in s tem znižanje cen . O tem bo

Evropska komisija odločala septembra . Možen ukrep so tudi dodatne kvote na področju mleka , je dejal . Zakaj cene hrane

naraščajo Strinjala sva se , da bi morala Evropska komisija narediti analizo vzrokov po posameznih sektorjih in analizo posledic

takšnega dogajanja , je Barnier odgovoril na vprašanje , kaj lahko EU naredi za umiritev cen hrane . Izpostavil je visoke

stroške za pridelovalce mesa oz . perutnine in dejal , da bi bilo treba uravnotežiti posledice . Po Barnierjevem mnenju gre za

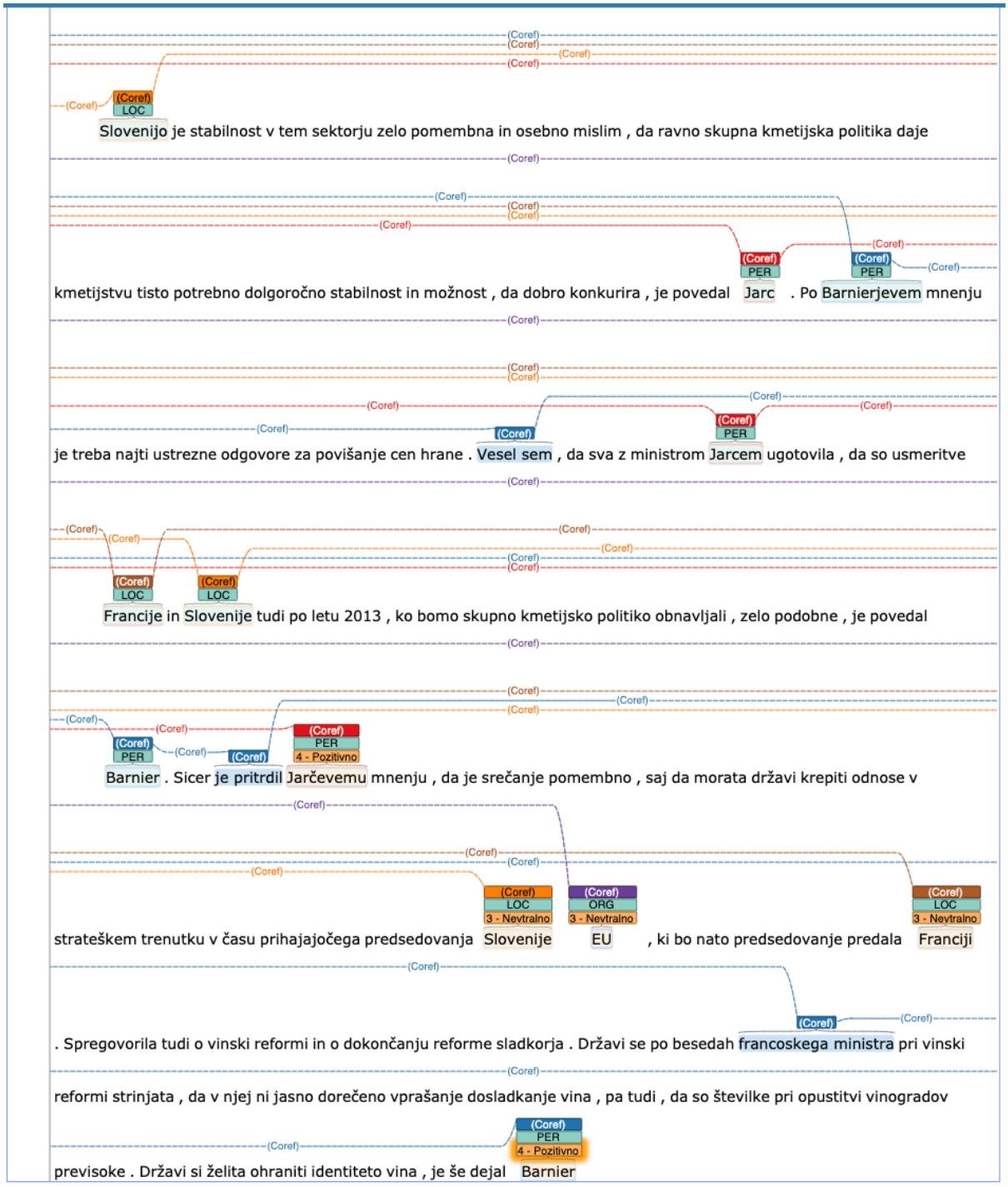
trajnejše poviševanje cen in je treba najti ustrezne odgovore . Kmetijski minister Iztok Jarc meni , da je treba odpreti razpravo

o visokih cenah hrane tudi na evropski ravni . Tudi Jarc meni , da treba odpreti razpravo o visokih cenah hrane tudi na

evropski ravni . Sicer je , kot pravi , težko napovedati , ali se bodo napovedi o rasti cen v prihodnje uresničile , bo pa

ministrstvo v vsakem primeru pozorno spremljalo dogajanje in poskušalo ustrezno reagirati . Podobne usmeritve Slovenije in

Francije Jarc je dejal , da imata Francija in Slovenija precej podobna stališča na skupno kmetijsko politiko . Za




# 6 Reference


Žitnik, S., in Bajec, M. (2018): Odkrivanje koreferenčnosti v slovenskem jeziku na označenih besedilih iz coref149. Slovenščina 2.0, 6 (1): 37–67.
DOI: http://dx.doi.org/10.4312/slo2.0.2017.1.37-67.

Bučar, J. (2017): Manually sentiment annotated Slovenian news corpus SentiNews 1.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1110.

Štajner, T., Erjavec, T., Krek, S. (2013): Razpoznavanje imenskih entitet v slovenskem besedilu. Slovenščina 2.0, 1 (2): 58–81.
URL: http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_04.pdf.

Krek, S, in sod. (2019): Training corpus ssj500k 2.2, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1210.