“My Croatian is better than yours”

Sentiment Analysis for Language Attitude Research

Barbara Kovačić
Ludwig-Maximilians-Universität München
Sveučilišta u Zagrebu

Stereotypes about Kajkavian Speakers

Language Attitude

(Garett, 2006)

- Attitude is a psychological construct
- defined as a predisposition to respond to objects positively or negatively
- are often learned through socialization
- comprised of cognitive (beliefs and stereotypes), affective (evaluations), and behavioral components
- connection between behavior and other components remains unclear
- Stereotypes serve various functions and structure the social world and help explain relationships between groups
- early-acquired attitudes tend to be less changeable later in life

Methods to research Language Attitude

(Dragojevic et al, 2021)

Societal Treatment Approach	Direct Approach	Indirect Approach
Researchers do not interview participants directly, but instead observe or analyze artifacts to infer attitudes	Explicitly asking respondents about their language preferences through surveys or interviews	Respondents are subtly asked about their language preferences

“From Fringe to Infrastructure”

(Fišer et al, 2021)

- research focus: users' attitudes toward their national languages, especially
  - prestige
  - standard language ideology
  - social status related to language skills
- dataset
  - manually annotated twitter dataset in Slovene, Croatian and Serbian
  - automatically collected tweets based on headwords such as “language”, “orthography”, “grammar”, “dictionary” and the name of the language
  - *content:* discussion of correct and incorrect language use in the respective language

“From Fringe to Infrastructure”

(Fišer et al, 2021)

- annotation:
  - attitudes towards the standard language and its status (e.g. nationalist)
  - attitudes towards the rules, deviations, and errors (e.g. lamenting)
  - attitudes towards people who use language in a certain way (mostly in an unsuitable way, as considered by the person expressing a specific attitude) (e.g. dismissive)
  - the types of discourse (or their function) in which language-related matters are featured (e.g. idiomatic, informative).

“From Fringe to Infrastructure”

(Fišer et al, 2021)

Structure of Croatian Dataset

dataset filtered on the keywords “jezik”, “pravopis”, “gramatika”, “rječnik” and “hrvatski”, resulting in 11,845,710 tokens

Headword	Absolute Frequency
jezik	8138
pravopis	510
gramatika	405
rječnik	540
hrvatski	24261

“From Fringe to Infrastructure”

(Fišer et al, 2021)

Structure of Croatian Dataset

50 tweets, containing one or more sentences, per keyword have been annotated based on their language attitude → 750 tweets in total

inquisitive	Pazite li na pravopis i gramatiku na društvenim mrežama? *Do you pay attention to orthography and grammar on social media?*
informative	Relevantni pravopis je online i jednostavan je za upotrebu http://t.co/oFvpzNyDci pa nema više izlike za nepismenost:) The relevant orthography guide is online and easy to use http://t.co/oFvpzNyDci , so there is no excuse for illiteracy anymore :)

“From Fringe to Infrastructure”

(Fišer et al, 2021)

Results for Croatian Dataset

- tweets often express uncertainties about lexis, highlighting differences between Croatian and similar languages like Serbian and Bosnian
- frequent discussions about new orthography guides in Croatia, often expressing dissatisfaction or frustration within the community
- concerns or comments about various regional dialects, a topic not observed in the Serbian dataset
- high engagement in conversations about distinguishing Croatian from other languages, emphasizing its uniqueness

“From Fringe to Infrastructure”

(Fišer et al, 2021)

Our goal

1. Sentiment annotate the 750 LA/stance annotated tweets on sentence level

2. Analyse if there is a correlation between sentiment and the LA/stance

3. Sentiment annotate other tweets based on keywords (e.g language) on sentence level

4. Analyse if there is a correlation between sentiment and keyword

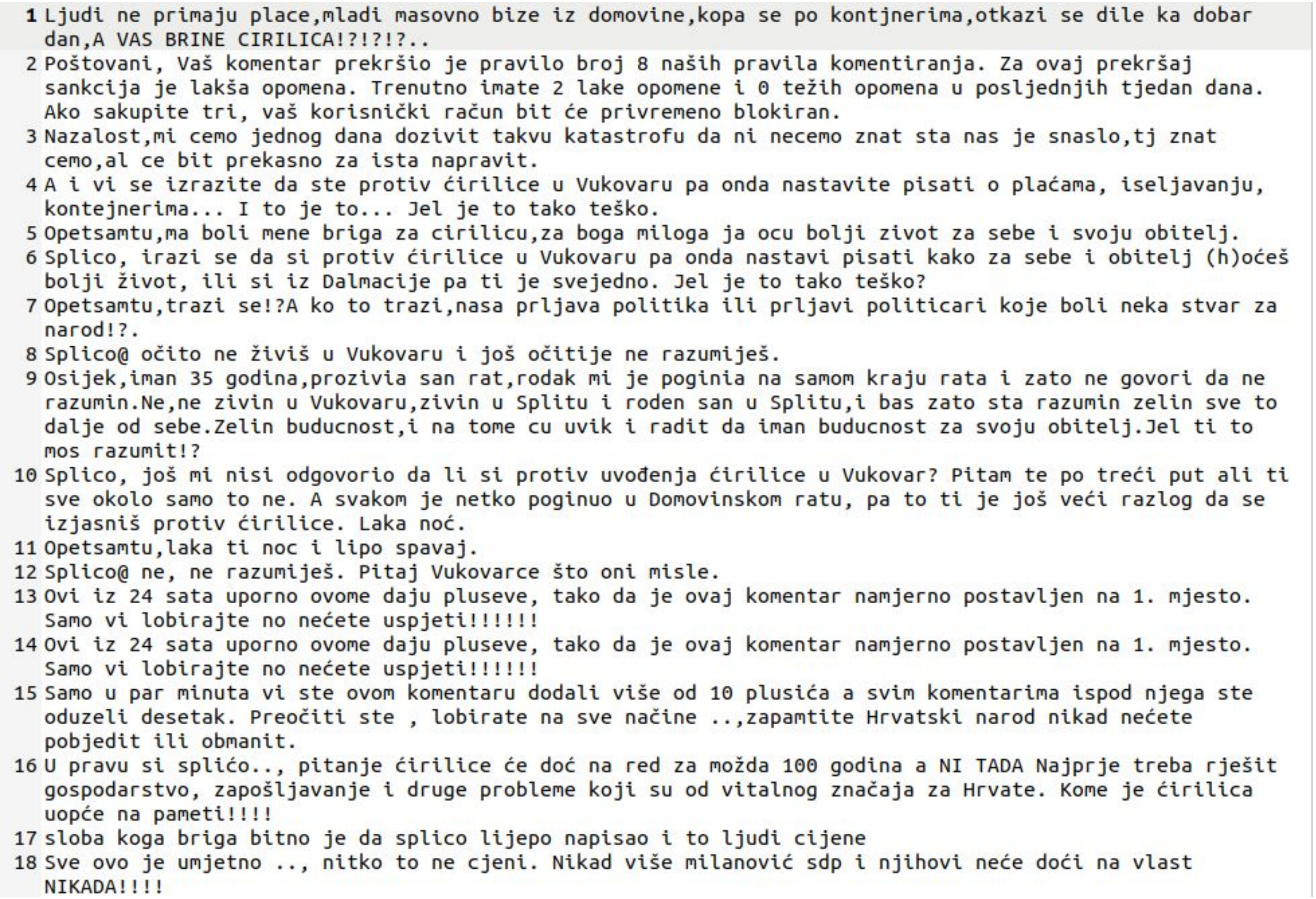
5. Train and finetune transformer model based on sentiment

6. Optional: if there is still time left and annotation guidelines available, LA/stance annotate more tweets

MetaLangNEWS-COMMENTS-Hr

(Bogetić & Batanović , 2020)

- corpus of user comments on online news articles on the topic of language from major Croatian daily newspapers and news portals
- published in the five-year period of January 1, 2015 - January 1, 2020
- designed to facilitate research on metalanguage (‘language about language’)
- insights about
  - linguistic ideologies
  - language policy and planning
  - contemporary debates on language defining, naming, and standardisation
- 738 articles, 21533 texts, 823459 tokens



1 Ljudi ne primaju place,mladi masovno bize iz domovine,kopa se po kontjnerima,otkazi se dile ka dobar dan,A VAS BRINE CIRILICA!?!?!?..

2 Poštovani, Vaš komentar prekršio je pravilo broj 8 naših pravila komentiranja. Za ovaj prekršaj sankcija je lakša opomena. Trenutno imate 2 lake opomene i 0 težih opomena u posljednjih tjedan dana. Ako sakupite tri, vaš korisnički račun bit će privremeno blokiran.

3 Nazalost,mi cemo jednog dana dozivit takvu katastrofu da ni necemo znat sta nas je snaslo,tj znat cemo,al ce bit prekasno za ista napravit.

4 A i vi se izrazite da ste protiv ćirilice u Vukovaru pa onda nastavite pisati o plaćama, iseljavanju, kontejnerima... I to je to... Jel je to tako teško.

5 Opetsamtu,ma boli mene briga za cirilicu,za boga miloga ja ocu bolji zivot za sebe i svoju obitelj.

6 Splico, irazi se da si protiv ćirilice u Vukovaru pa onda nastavi pisati kako za sebe i obitelj (h)oćeš bolji život, ili si iz Dalmacije pa ti je svejedno. Jel je to tako teško?

7 Opetsamtu,trazi se!?A ko to trazi,nasa prljava politika ili prljavi politicari koje boli neka stvar za narod!?.

8 Splico@ očito ne živiš u Vukovaru i još očitije ne razumiješ.

9 Osijek,iman 35 godina,prozivia san rat,rodak mi je poginia na samom kraju rata i zato ne govori da ne razumin.Ne,ne zivin u Vukovaru,zivin u Splitu i roden san u Splitu,i bas zato sta razumin zelin sve to dalje od sebe.Zelin buducnost,i na tome cu uvik i radit da iman buducnost za svoju obitelj.Jel ti to mos razumit!?

10 Splico, još mi nisi odgovorio da li si protiv uvođenja ćirilice u Vukovar? Pitam te po treći put ali ti sve okolo samo to ne. A svakom je netko poginuo u Domovinskom ratu, pa to ti je još veći razlog da se izjasniš protiv ćirilice. Laka noć.

11 Opetsamtu,laka ti noc i lipo spavaj.

12 Splico@ ne, ne razumiješ. Pitaj Vukovarce što oni misle.

13 Ovi iz 24 sata uporno ovome daju pluseve, tako da je ovaj komentar namjerno postavljen na 1. mjesto. Samo vi lobirajte no nećete uspjeti!!!!!!

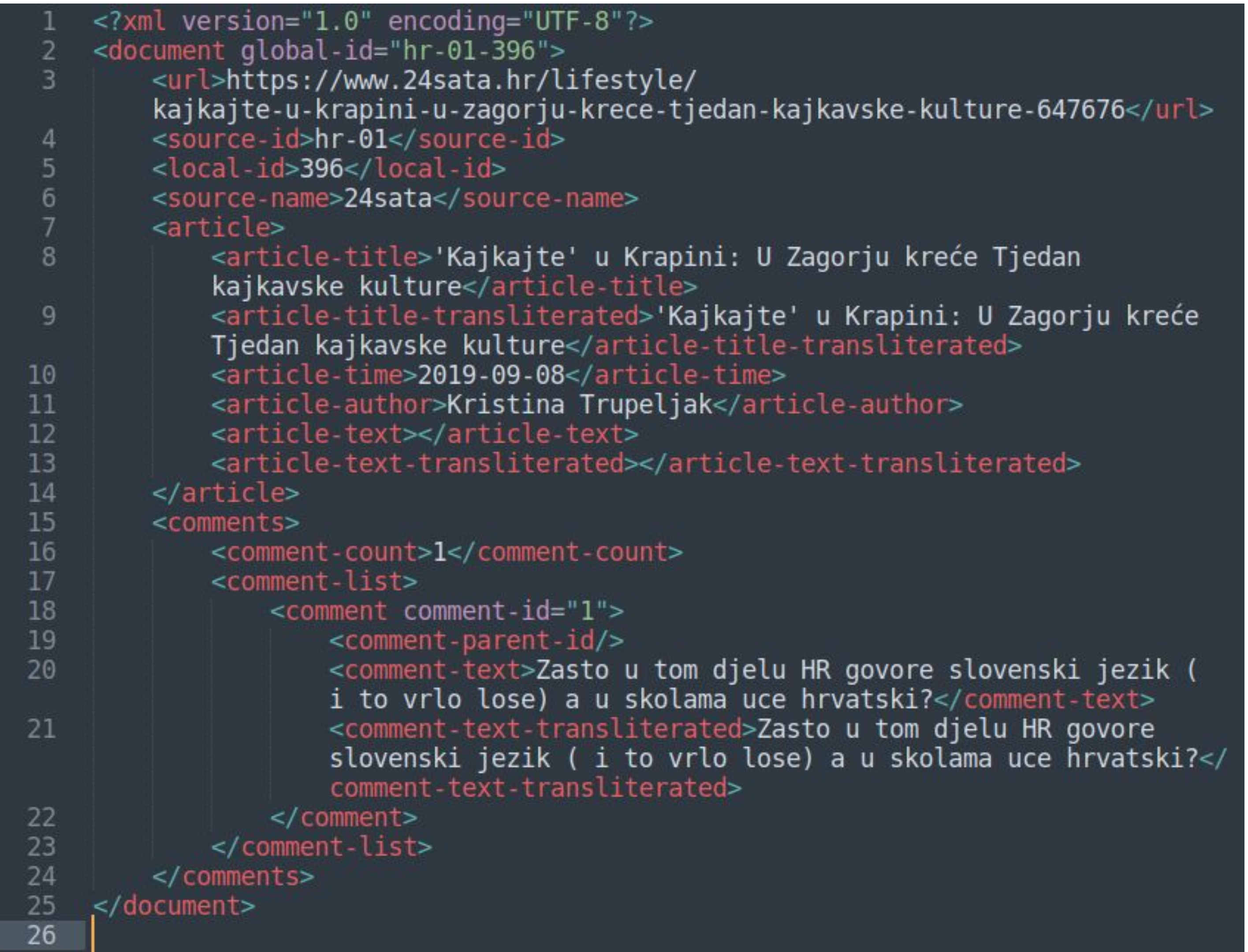
14 Ovi iz 24 sata uporno ovome daju pluseve, tako da je ovaj komentar namjerno postavljen na 1. mjesto. Samo vi lobirajte no nećete uspjeti!!!!!!

15 Samo u par minuta vi ste ovom komentaru dodali više od 10 plusića a svim komentarima ispod njega ste oduzeli desetak. Preočiti ste , lobirate na sve načine ..,zapamtite Hrvatski narod nikad nećete pobjedit ili obmanit.

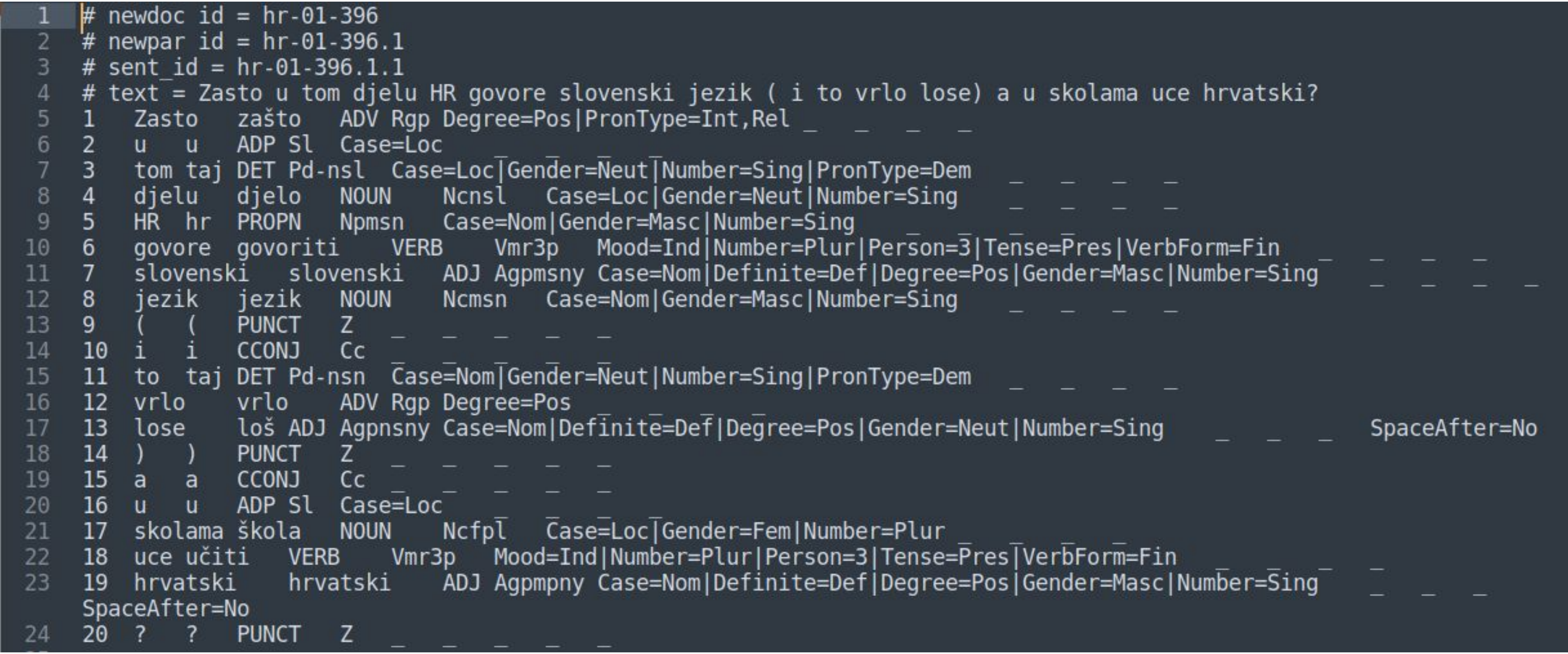
16 U pravu si splićo.., pitanje ćirilice će doć na red za možda 100 godina a NI TADA Najprje treba rješit gospodarstvo, zapošljavanje i druge probleme koji su od vitalnog značaja za Hrvate. Kome je ćirilica uopće na pameti!!!!

17 sloba koga briga bitno je da splico lijepo napisao i to ljudi cijene

18 Sve ovo je umjetno .., nitko to ne cjeni. Nikad više milanović sdp i njihovi neće doći na vlast NIKADA!!!!



```xml
<?xml version="1.0" encoding="UTF-8"?>
<document global-id="hr-01-396">
    <url>https://www.24sata.hr/lifestyle/kajkajte-u-krapini-u-zagorju-krece-tjedan-kajkavske-kulture-647676</url>
    <source-id>hr-01</source-id>
    <local-id>396</local-id>
    <source-name>24sata</source-name>
    <article>
        <article-title>'Kajkajte' u Krapini: U Zagorju kreće Tjedan kajkavske kulture</article-title>
        <article-title-transliterated>'Kajkajte' u Krapini: U Zagorju kreće Tjedan kajkavske kulture</article-title-transliterated>
        <article-time>2019-09-08</article-time>
        <article-author>Kristina Trupeljak</article-author>
        <article-text></article-text>
        <article-text-transliterated></article-text-transliterated>
    </article>
    <comments>
        <comment-count>1</comment-count>
        <comment-list>
            <comment comment-id="1">
                <comment-parent-id/>
                <comment-text>Zasto u tom djelu HR govore slovenski jezik ( i to vrlo lose) a u skolama uce hrvatski?</comment-text>
                <comment-text-transliterated>Zasto u tom djelu HR govore slovenski jezik ( i to vrlo lose) a u skolama uce hrvatski?</comment-text-transliterated>
            </comment>
        </comment-list>
    </comments>
</document>
```



```
# newdoc id = hr-01-396
# newpar id = hr-01-396.1
# sent_id = hr-01-396.1.1
# text = Zasto u tom djelu HR govore slovenski jezik ( i to vrlo lose) a u skolama uce hrvatski?
1   Zasto   zašto   ADV Rgp Degree=Pos|PronType=Int,Rel _   _   _   _
2   u   u   ADP Sl  Case=Loc    _   _   _   _
3   tom taj DET Pd-nsl  Case=Loc|Gender=Neut|Number=Sing|PronType=Dem   _   _   _   _
4   djelu   djelo   NOUN    Ncnsl   Case=Loc|Gender=Neut|Number=Sing    _   _   _   _
5   HR  hr  PROPN   Npmsn   Case=Nom|Gender=Masc|Number=Sing    _   _   _   _
6   govore  govoriti    VERB    Vmr3p   Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin   _   _   _   _
7   slovenski   slovenski   ADJ Agpmsny Case=Nom|Definite=Def|Degree=Pos|Gender=Masc|Number=Sing    _   _   _   _
8   jezik   jezik   NOUN    Ncmsn   Case=Nom|Gender=Masc|Number=Sing    _   _   _   _
9   (   (   PUNCT   Z   _   _   _   _   _
10  i   i   CCONJ   Cc  _   _   _   _   _
11  to  taj DET Pd-nsn  Case=Nom|Gender=Neut|Number=Sing|PronType=Dem   _   _   _   _
12  vrlo    vrlo    ADV Rgp Degree=Pos  _   _   _   _
13  lose    loš ADJ Agpnsny Case=Nom|Definite=Def|Degree=Pos|Gender=Neut|Number=Sing    _   _   _   SpaceAfter=No
14  )   )   PUNCT   Z   _   _   _   _   _
15  a   a   CCONJ   Cc  _   _   _   _   _
16  u   u   ADP Sl  Case=Loc    _   _   _   _
17  skolama škola   NOUN    Ncfpl   Case=Loc|Gender=Fem|Number=Plur _   _   _   _
18  uce učiti   VERB    Vmr3p   Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin   _   _   _   _
19  hrvatski    hrvatski    ADJ Agpmpny Case=Nom|Definite=Def|Degree=Pos|Gender=Masc|Number=Sing    _   _   _
SpaceAfter=No
20  ?   ?   PUNCT   Z   _   _   _   _   _
```

MetaLangNEWS-COMMENTS-Hr

(Bogetić & Batanović , 2020)

Our goal

1. Sentiment annotate the comments on sentence level
2. Second layer of annotation
    a. sentence consists language attitude? yes / no
    b. language attitude overt or covert?
    c. who is the target of the sentiment?
    d. …
3. Analysis of correlation between sentiment and second layer of annotation

References

BOGETIĆ, Ksenija, BATANOVIĆ, Vuk. Annotated corpus of Croatian language-related news comments MetaLangNEWS-COMMENTS-Hr. [Ljubljana]: ZRC SAZU; [Beograd]: Regional Linguistic Data Initiative Centre, 2020. 1 spletni vir. CLARIN.SI data & tools. https://www.clarin.si/repository/xmlui/handle/11356/1370. [COBISS.SI-ID 35287299]

Dragojevic, M., Fasoli, F., Cramer, J., & Rakić, T. (2021). Toward a Century of Language Attitudes Research: Looking Back and Moving Forward. Journal of Language and Social Psychology, 40(1), 60–79. https://doi-org.emedien.ub.uni-muenchen.de/10.1177/0261927X20966714

Fišer, D., Ljubešić, N., & Popič, D. (2021). From Fringe to Infrastructure: A Researcher’s Journey through South Slavic Language Attitudes on Social Media. *Modern Languages Open*.

Garrett, P. (2006). Language attitudes. In *The Routledge companion to sociolinguistics* (pp. 136-141). Routledge.

Backlog

“From Fringe to Infrastructure”

(Fišer et al, 2021)

Language Attitude Annotation

inquisitive	Pazite li na pravopis i gramatiku na društvenim mrežama? *Do you pay attention to orthography and grammar on social media?*
informative	Relevantni pravopis je online i jednostavan je za upotrebu http://t.co/oFvpzNyDci pa nema više izlike za nepismenost:) *The relevant orthography guide is online and easy to use http://t.co/oFvpzNyDci , so there is no excuse for illiteracy anymore :)*
lamenting	Meni hrvatski pravopis je sve gori. Užas *I think the Croatian orthography is getting worse and worse. Horror*

“From Fringe to Infrastructure”

(Fišer et al, 2021)

Language Attitude Annotation

jocular	Inace ne volim viceve, ali novom pravopisu u Hrvata, kao laik, povremeno moras priznati komicni momenat. *I normally don’t like jokes but, to a layman, the new Croatian orthography is hilarious in certain places.*
dismissive	Ne znam prema kojim to kriterijima se danas zapošljavaju novinari koji fejlaju već na pravopisu. *I don’t know which criteria are used to employ journalists who fail already at orthography.*
defensive	Ne vjeruj ženi s lošim pravopisom. *Don’t trust a woman with poor orthography.*

“From Fringe to Infrastructure”

(Fišer et al, 2021)

Language Attitude Annotation

apologetic	nemojte mi o pravopisu na rano jutro, nisam nepismena samo lijena *spare me the orthography first thing in the morning, I’m not illiterate, just lazy*
idiomatic	»Pravopis sa zvijezdama« bih gledao. *I’d watch “Orthography with the stars”.*
nationalist	Spasimo spomenik Ljudevitu Gaju jer je bio ustaša i kao ustaša stvorio je ustaški pravopis 1830 *Let’s save the monument to Ljudevid Gaj because he was a Ustashe terrorist and as such created the Ustashe orthography in 1830*

“From Fringe to Infrastructure”

(Fišer et al, 2021)

Language Attitude Annotation

purist	Zašto nitko ne provjeri pravopis prije tiska? Sramočenje *Why doesn’t somebody check the Orthography guide before sending it to print? Disgrace.*
praising	Meni je pravopis bio najdrazi predmet u skoli. *Orthography was my favourite subject at school.*
anti-purist	Danas ne postoji jedinstven pravopis – ima ih čak četiri. *A single orthography doesn’t exist this day and age – there’s as many as four.*

→ no examples for “neutral” and “anti-nationalist”