https://doi.org/10.31820/f.33.1.7

Ksenija Bogetić

METALANGCORP: PRESENTING THE FIRST CORPUS OF MEDIA METALANGUAGE IN SLOVENE, CROATIAN AND SERBIAN, AND ITS CROSS-DISCIPLINE APPLICABILITY[1]

dr. sc. Ksenija Bogetić, Inštitut za kulturne in spominske študije, Ljubljana
ksenija@zrc-sazu.si orcid.org/0000-0002-7731-6324

pregledni članak
UDK 811.163'322
rukopis primljen: 23. listopada 2020; prihvaćen za tisak: 14. siječnja 2021.

Growing interest in meta-language, in linguistics and other disciplines, has highlighted a gap in metalanguage corpora and analytical resources, which remain among the scarcest in corpus-linguistic developments so far. This paper is aimed at making a step towards filling this gap, both by presenting our own metalanguage corpus resource and using it in a short sample analysis to discuss the applications of such resources in linguistics and social sciences. Specifically, the paper presents for the first time MetaLangCORP, a multi-element corpus of contemporary media metalanguage in languages of three post-Yugoslav states, linguistically annotated and made available open-access at the CLARIN repository of linguistic resources. To put the corpus in context, the meaning and relevance of metalanguage research is outlined, the existing

[1] Author statement: The research is funded by the Slovenian Research Agency (ARRS), as part of the research project "(Re-)imagining language, nation and collective identity in the 21st century: Language ideologies as new connections in post-Yugoslav digital mediascapes" (N6-0123).

efforts at compiling corpora of metalanguage are reviewed, and a sample preliminary analysis of MetaLangCORP keywords is presented to open a broader discussion on the potential applicability of metalanguage corpora. More broadly, it is hoped that making this kind of data available will prompt more nuanced analyses of metalanguage, as well as more corpus-building efforts along similar lines in Slavic and other linguistic scholarship.

Key words: *metalanguage; language corpora; language in (new) media*

1. Introduction

Metalanguage, or 'talk about talk', long hovered on the margins of linguistic enquiry as non-scientific, "often noted but rarely interrogated" (Squires 2010: 483). The turn of the century, however, is said to have marked the beginning of unprecedented public and scholarly interest in language debates (Cameron, 2014), happening in light of technological transformation, demographic change, and political crisis. In parallel, the understanding of the mass media as major producers and sustainers of all our beliefs about language (Androutsopoulos, 2011), has resulted in proliferation of analyses of metalanguage specifically in (new) media contexts, promising fresh insights on how the shifting and increasingly multi-voiced debates on language may structure our perceptions of communication, identity and society in the 21st century. Whether and how conceptions of language are changing in times of digital networks and flows of people, what these conceptions mean for academic and political elites and what they mean for language users themselves, or how they play into the core sociolinguistic questions of language variation, change, standardisation, identity and ideology, are all questions awaiting deeper insights from a wider range of lingua-political contexts.

In light of this growing interest, however, there is a lingering demand for more empirically and critically grounded research , as well as for the development of corpora and analytical resources in order to achieve a better understanding of real-world metalanguage and its wider societal meanings. This is especially true if we are to try and capture not only the established expert discourses, but also the bottom-up, citizen and grassroots discourses that are all contributing to the polyphony of contemporary metalanguage debates (Krzyżanowski & Tucker, 2018). On another level, insights on structural properties based on real-life metalanguage remain a notable gap in areas ranging from pragmatics to natural language semantics and language

processing (Wilson, 2010). Still, aside from individual, mostly small-scale analyses, systematically built corpora of metalanguage are among the scarcest in corpus-linguistic developments so far.

This paper is aimed at taking a step towards filling this gap, both by presenting our own metalanguage corpus resource and using it in a short sample analysis to discuss the applications of such resources in linguistics and social sciences. After a brief outline of the meaning and relevance of metalanguage (section 2), the existing efforts at compiling corpora of metalanguage are reviewed (section 3), and MetaLangCORP – a multi-element corpus of contemporary media metalanguage in three post-Yugoslav languages – is presented for the first time (section 4). Section 5 demonstrates the possibilities of analysis through sample corpus-linguistic findings, discusses their implications, and the potential of corpus use for broader lingua-political analysis in the humanities. Observations from corpus compilation and analysis are summed up in the final section, with a view to future research.

2. Metalanguage

It has become a popular truism in linguistics to argue that human language is unique in being able to represent itself. Metalanguage is indeed an essential linguistic mechanism which allows us to communicate explicit information about language itself. For language scholars from the 2000's onwards, awareness of these properties of language has led to a broader awareness that to study language in use is to study not just what people do with language, but also what they believe and feel about language, and how both are part of larger structures of society and power. This realisation has so far had different resonance in different areas of linguistics, but has contributed to developing richer models of language as a contextualised and contextualising phenomenon, and as a set of strategic, often reflexive, socially imbued practices. Metalinguistic processes lie at the basis of all these models in one way or another (Jaworski et al., 2012).

Going back in time, the history of the concept of *metalanguage* can be traced to early functionalist approaches to language, which were later taken up in linguistic pragmatics and anthropology, variationist sociolinguistics, as well as in research on language attitudes in social psychology. To be sure, these different approaches mean that metalanguage has been variously defined in different areas and periods of linguistics, but its overall meanings

are well captured in the tripartite model described by Dennis Preston (2004). In this model, Metalanguage 1 consists of *language about language,* i.e. those instances where people explicitly comment on some aspect of language, e.g., when discussing how a word is correctly pronounced (2004: 75). Metalanguage 2, on the other hand, refers to less explicit linguistic commentary with solely some *mention of talk* itself, e.g., when asking someone to clarify what they said, or whether they understand what you said (cf. the phatic function of language in Jakobson's (1960) terms). Metalanguage 3 refers to the more general set of shared beliefs and attitudes about language structure and use within a given speech community, which can in turn motivate the kinds of commentary of Metalanguage 1 and 2. Importantly, the definition of Metalanguage 3 closely taps into today's major field of study on metalanguage and society, by now a separate discipline: that of language ideology studies (e.g. Gal & Woolard, 2014; Cavanaugh, 2020). Language ideology explores how beliefs about society' impact beliefs about language and vice versa, acknowledging that these ordinary beliefs are behind much language change, behind trajectories of language policy, planning or education, and behind people's perceptions of community and identity (Bogetić, 2017). It also opens a more holistic view of language by linking the multiple contexts – discursive, cultural, technological, economic, political and historical – in which linguistic processes are embedded (Johnson & Ensslin, 2007). As Jaworski et al. (2012) put it, all concerns about language ideology are ultimately concerns over metalanguage, since much of what we can learn about the ideological underpinnings of language use is derived from what people actually have to *say* about language.

A different motivator of interest in metalanguage has more recently come from computational linguistics. Namely, recognizing a wide variety of metalinguistic constructions is a skill that humans take for granted in communication with their interlocutors (Wilson, 2010). However, applications of natural language processing generally lack the ability to recognize and interpret metalanguage (Anderson et al., 2002). At the same time, resources of real metalanguage in use that could be used as a basis for training in machine learning applications are very scarce.

Finally, the growth of digital media has contributed a specific slant in the interest in metalanguage. In many parts of the world, metalanguage discussions have ranged from popular anxieties over "the breakdown of communication" and supposedly declining language standards (Cameron, 2014; Bogetić, 2016; Garley, 2019), to a broader interest in the reflexivity

of the digital media era, where metalinguistic sensitivity becomes a "hallmark of contemporary social life" (Jaworski et al. 2012). In practical terms, online media texts provide both a particularly useful and widely accessible source of data for observing and analysing such (language-) ideological processes (Johnson & Ensslin, 2006). However, the popularly praised availability of data thanks to the digital media has not contributed much to their use as resources for empirical study of metalanguage, which has largely been based on sparse or cherry-picked datasets and anecdotal evidence. Despite some not-so-recent calls (Wilson, 2010), corpora and language resources on metalanguage in its own right remain among the least developed in corpus linguistic repositories, especially in non-Anglophone contexts. The present corpus compilation was motivated by one such context, that of vehement language debates in post-Yugoslav nation-states, where empirical data concerning contemporary language in use has traditionally represented a major gap in sociolinguistics, despite exciting developments in corpus-building and tools (see the fast growing repository of linguistic data at clarin.si).

3. Metalanguage corpora in linguistics research

Corpora of metalanguage have started to be developed comparatively late, as a result of the comparatively late development of interest in metalanguage itself. They can most basically be characterised as compilations of Metalanguage 1, i.e. sets of texts where language is explicitly thematised, but by their very nature they will include instances of Metalanguage 2, and depending on motivation behind collection, may facilitate insights into Metalanguage 3 and language ideologies. The rare examples of existing corpora follow different approaches to corpus compilation, following different motivations for study.

One approach is to create strings of 'mentioned language', including sentences or sentence segments that explicitly thematise aspects of language. Compilation of such corpora is motivated mainly by efforts towards formal description of metalanguage, and its utilization in language technologies and natural language processing applications. Such is the set of first tagged and delineated corpora of English metalanguage developed by Wilson between 2010 and 2013, designed to allow empirical examination of metalanguage and to study the feasibility of automatic identification. Wilson's approach is based on articles from the English Wikipedia, from

which instances were mined using a combination of automated and manual efforts, based on a predetermined set of cues for selecting candidate instances. At the point of writing this article, none of the compiled corpora appear openly accessible online.

A different approach includes corpora that can be labelled as thematic. They concern text collections that either share the *core theme* of language / a language / an aspect of language, or those built to contain texts with *instances* of metalanguage commentary, usually reflecting interest in specific kinds of metalanguage (thus not metalinguistically-thematic in the real sense of the word), as well as those that fall somewhere in between. Such corpora are more useful to sociolinguists and other humanities scholars exploring conceptions of language in particular genres or social contexts. Examples of thematic corpora are the specialized corpora of newspaper texts made for early studies of British metalanguage in news media, based on automatic search of predetermined language-related concepts in selected newspapers. They were used in Johnson et al.'s work (2003) on metalinguistic expressions relating to 'political correctness', or Johnson and Ensslin's (2006) work on expressions relating to 'Englishness'. Another example of a metalanguage corpus based on specific terms of interest is the corpus compiled by Charlotte Taylor (2015) for her research on metalanguage of mock politeness, sarcasm and irony in English and Italian. Her corpus is search-term specific, which is to say that only texts that contained selected search terms of interest (to do with mock politeness) are included. These key mock politeness indicators were used in order to compile corpora from the target discourse types, in this case, from two online forums, one based in the UK and one in Italy.

In the Slavic linguistic context, resources of metalanguage, and research on metalanguage or language ideology in general, are practically absent to date. An interesting exception is the corpus built by Adnan Ajšić (2015) for the purposes of his research on language ideologies and ethnonationalist discourse in the mainstream press in Serbia. Specifically, he compiled a specialized research corpus designed to represent general language-related discourse in mainstream Serbian press, i.e. a thematic metalanguage corpus in the real sense. The corpus consists of articles focusing on language, published in four national newspapers in the period between 2003 and 2008. However, it is important to note that the resource, just as all the above mentioned corpus resources, were used for purposes of individual research and not shared as open-access or access-on-demand

tools, despite the trend towards resource sharing taking place in corpus linguistics within the past decade. In this respect, the absence of comprehensive, systematic, readily usable repositories of metalanguage remain a gap in Slavic resource building and far beyond.

4. Corpus motivation, compilation and structure

The MetaLangCORP presented in this paper was designed for purposes of a project entitled (*Re-)imagining language, nation and collective identity in the 21st century: Language ideologies in post-Yugoslav digital mediascapes*. The project explores conceptions of language and nation in Yugoslavia's successor states, to span six states (so far focused on three: Slovenia, Croatia, Serbia) and the most recent period of five years (Jan. 2015 – Jan. 2020). It also aims to address broader gaps in sociolinguistics and social sciences, where understanding ideas of nationhood in relation to language has been identified as a central gap, a kind of tacit knowledge rarely investigated in any empirical way (Kamusella 2018); in the (post-)Yugoslav context, the gap appears even more notable amidst little systematic investigation and often reductionist perspectives on language and conflict. Therefore, the empirical dimension and a corpus approach are central to the project ambitions in addressing existing gaps.

The corpus we need for this kind of exploration is of thematic type, focused on media discourse, and composed of texts that centre on the core theme of language. The texts, composed of articles from major national newspapers and newspaper portals (more details below) follow the dual choice of data:

(a) news media articles, i.e. online versions of newspapers/news portal articles, given the long established role of the media as the major producers of language ideologies (e.g. Androutsopoulos, 2011[2])

(b) reader commentary, in below-text comments sections, given the massive ongoing convergence of traditional and user-generated media transforming public discourse as we know it (Lenihan, 2018).

[2] As particularly explored in Anglophone contexts; see earlier analyses in Lippi-Green 1997, Silverstein 1999.

Capturing the voices from the latter source of data, as yet less explored in this local context and elsewhere[3], is an important aspect of the project that again requires an empirical outlook and careful corpus compilation.

Hence, while the whole collection of multi-lingual data is labelled as MetaLangCORP, this corpus consists of separate components. For each language, there are two separate corpora: MetaLangCORP-NEWS and MetaLangCORP-NEWS-COMMENTS, labelled based on language as MetaLangCOPR-NEWS-Slo, MetaLangCORP-NEWS-COMMENTS-Slo, MetaLangCORP-NEWS-Hr etc. They are stored, annotated and shared as separate datasets, given that the two subtypes for each language represent different genres, based in different sources, and would rarely be justifiably quantitatively analysed as one dataset (though corpus files allow merged use if desired).

Corpus compilation was based on automated text collection based on a pre-defined set of language-related terms, in the most widely read newspapers and news portals in each state (information available via alexa.com[4]). Data scraping was performed using machine learning techniques, in collaboration with experts for the ReLDI centre (Regional Linguistic Data Initiative). All materials were collected from freely available newspaper archives online, and readers' comments accompanying articles were all available on the articles pages.

For each language subcorpus, the compilation resulted in datasets of over half a million words. Details on corpus size are given in Table 1. It is interesting that the size of the comments (sub)corpora for the Serbian and Croatian sources exceeds that of the news articles corpora. This reflects generally high levels of readers' engagement with the news texts on lan-

[3] Of course, this is not to take news comments as fully representative of popular opinion. People who read and comment on news may be different from those who do not (Hermida and Thurman 2008) Other difficulties are the traditional ones associated with CMC, such as not knowing the poster's true identity, or having online discussions include only a small core of very active users. Still, news comments are a prime site in which to approach language ideologies as emergent within the amplified heteroglossia of the digital media; designing corpora to include all of the reader comments on all the topic-related articles produces a material representative of the comments discourse observed.

[4] Using freely available information from Alexa Rank metrics, which rank websites in order of popularity, by looking at the estimated average daily unique visitors and number of pageviews over a several month period.

guage, where comparatively short articles are often followed by a long list of passionate comments, sometimes amounting to hundreds. It is of note that the Slovenian sources in general tend to have fewer comments when compared to the Serbian and Croatian ones.

Table 1: Corpus details

Corpus	MetaLangCOPR-NEWS-Slo	MetaLangCOPR-NEWS-Hr	MetaLangCOPR-NEWS-Sr
Articles	555	817	1088
Tokens	501,634	535,445	559,989
Types	55,673	53,455	51,672

Corpus	MetaLangCORP-NEWS-COMMENTS-Slo	MetaLangCORP-NEWS-COMMENTS-Hr	MetaLangCORP-NEWS-COMMENTS-Sr
Articles	/	/	/
Tokens	36,808	743,949	732,655
Types	10,760	93,558	79,207

The corpus, i.e. each of its components, is tagged using models for morpho-syntactic annotation and lemmatisation. Specifically, the annotation includes lemmatisation, part-of-speech (POS) tagging, including Universal Dependency POS tags and MULTEXT-East-standard tags. The corpus is available in several formats, including plain .txt files which facilitate simple analysis in concordancing tools, as well as fully qualitative analysis; there are also .xml files containing all relevant metadata, and files in the derived format of CoNLL-U. Figure 1 shows the corpus encoding used; Figure 2 shows a sample of the CoNLL-U annotated format.

hr-01-1393.xml

```
<?xml version="1.0" encoding="utf-8"?>
<document global-id="hr-01-1393">
 <url>
  https://www.24sata.hr/news/hrvatski-i-srpski-jedan-jezik-strucnjaci-potpisali-
deklaraciju-517733
 </url>
 <source-id>
  hr-01
 </source-id>
 <local-id>
  1393
 </local-id>
 <source-name>
  24sata
 </source-name>
 <article>
  <article-title>
   'Zajednički jezik': Deklaraciju potpisalo trideset stručnjaka
  </article-title>
  <article-time>
   2017-03-28
  </article-time>
  <article-author>
   HINA
  </article-author>
  <article-text>
   Osnovna namjera ovog projekta je problematiziranje pitanja postojanja četiri
"politička" jezika na prostoru nekadašnjeg hrvatsko-srpskog/srpsko-hrvatskog. Nakon niza
regionalnih konferencija projekta "Jezici i nacionalizmi", više od trideset stručnjaka iz
Bosne i Hercegovine, Hrvatske, Crne Gore i Srbije sastavilo je "Deklaraciju o zajedničkom
```

Figure 1: Example of .xml format metadata of MetaLangNEWS-CORP

```
1       Agensi  agens   NOUN    Ncmpn   Case=Nom|Gender=Masc|Number=Plur                 _       _       _       _
2       nisu    biti    AUX     Var3p   Mood=Ind|Number=Plur|Person=3|Polarity=Neg|Tense=Pres|VerbForm=Fin      _
3       agenti  agent   NOUN    Ncmpn   Case=Nom|Gender=Masc|Number=Plur         _       _       _       _
4       Nadiranje       nadiranje       NOUN    Ncnsn   Case=Nom|Gender=Neut|Number=Sing        _       _       _
5       stranih stran   ADJ     Agpmpgy Case=Gen|Definite=Def|Degree=Pos|Gender=Masc|Number=Plur        _       _
6       (       (       PUNCT   Z       _       _       _       _       SpaceAfter=No
7       pogotovo        pogotovo        ADV     Rgp     Degree=Pos      _       _       _       _
8       engleskih       engleski        ADJ     Agpmpgy Case=Gen|Definite=Def|Degree=Pos|Gender=Masc|Number=Plur
9       )       )       PUNCT   Z       _       _       _       _
10      reči    reč     NOUN    Ncfpg   Case=Gen|Gender=Fem|Number=Plur _       _       _       _
11      u       u       ADP     Sa      Case=Acc        _       _       _       _
12      svetske svetski ADJ     Agpmpay Case=Acc|Definite=Def|Degree=Pos|Gender=Masc|Number=Plur        _       _
13      jezike  jezik   NOUN    Ncmpa   Case=Acc|Gender=Masc|Number=Plur         _       _       _       _
14      nije    biti    AUX     Var3s   Mood=Ind|Number=Sing|Person=3|Polarity=Neg|Tense=Pres|VerbForm=Fin      _
15      ništa   ništa   PRON    Pi3n-n  Case=Nom|Gender=Neut|PronType=Neg        _       _       _       _
16      novo    nov     ADJ     Agpnsny Case=Nom|Definite=Def|Degree=Pos|Gender=Neut|Number=Sing        _       _
17      i       i       CCONJ   Cc      _       _       _       _       _
18      nepoznato       nepoznat        ADJ     Agpnsny Case=Nom|Definite=Def|Degree=Pos|Gender=Neut|Number=Sing
19      .       .       PUNCT   Z       _       _       _       _       _
20      Tako    tako    ADV     Rgp     Degree=Pos|PronType=Dem _       _       _       _
21      i       i       CCONJ   Cc      _       _       _       _       _
22      srpski  srpski  ADJ     Agpmsny Case=Nom|Definite=Def|Degree=Pos|Gender=Masc|Number=Sing        _       _
23      „       „       NOUN    Ncmsn   Case=Nom|Gender=Masc|Number=Sing        _       _       _       SpaceAfter=No
```

Figure 2: Example CoNLL-U format with annotations of MetaLangNEWS-CORP

The full corpus is made available open access at the clarin.si repository. The results are searchable in the sub-repositories for the individual languages (Serbian, Croatian and Slovene), and the two subcorpora for each language are linked appropriately. The comments corpora are fully anonymised, meaning that even user acronyms were not preserved in the files, which precludes ethical and legal problems in data sharing. MetaLangCORP is available under a *Creative Commons Attribution-ShareAlike 4.0 International Public License* (https://creativecommons.org/licenses/by-sa/4.0/legalcode).

5. Preliminary analysis and metalanguage research potential

5.1. Keywords analysis

The comprehensive set of media texts dealing with the topic of (Serbian/Croatian/Slovene) language provides a representative sample in which to investigate different aspects of meta-language comments and debates. In this section we present some simple preliminary analyses, as a way into the data and into the discussion of further analytical possibilities.

On the most basic level, quantitative analyses of the corpora can provide snapshots of the discourse on language from each country, show the key concepts in the discourses to be explored further, or other unexpected patterns for in-depth analysis, depending on researchers' interests. For purposes of the present project, they also allow comparisons between countries, or comparisons between news media and citizens' discourses. Preliminary analysis thus departed from a common technique in corpus linguistics called *keywords analysis (KW)*: a statistical approach to word frequencies used to identify words occurring with unusual frequency in a given text. Keywords provide insights into central concepts in a discourse, showing the 'aboutness' of a material (Scott, 2009). Keywords are nowadays easily obtainable though standard and often freely available concordancing software, that also allow their meanings to be automatically presented together within concordance lines, i.e. broader immediate context. For the present analysis, *LancsBox* (Brezina, 2020) was used.

The top 20 lexical keywords for each subcorpus obtained in preliminary analyses[5] are listed and discussed below.

Table 2: Top lexical keywords in the MetaLangCOPR-NEWS-Hr / Sr / Slo

Corpus	KW
MetaLangCOPR-NEWS-Hr	jezik, hrvatski, Hrvatska, Vukovar, manjina, srpski, pismo, nacionalni, zakon, ustavni, ćirilica, predsjednik, vijeće, Penava, Hrvat, narod, Srbija, Srbin, Deklaracija, ministar
MetaLangCOPR-NEWS-Sr	jezik, srpski, pismo, Srbija, narod, bosanski, ćirilica, jezički, Srbin, crnogorski, hrvatski, standardizacija, upotreba, profesor, SANU, Beograd, fakultet, zakon, latinica, nacionalni
MetaLangCOPR-NEWS-Slo	jezik, slovenski, sloveščina, Slovenec, Slovenija, dr, univerza, angleščina, tuji, jezikovni, SAZU, manjšina, spol, knjižni, FRAN, Ahačič, materni, profesor, znakovni, šola

Table 3: Top lexical keywords in the MetaLangCOPR-NEWS-COMMENTS Hr / Sr / Slo

Corpus	KW
MetaLangCOPR-NEWS-COMM.-Hr	jezik, hrvatski, ćirilica, Vukovar, Hrvat, pismo, Srbin, narod, pravo, HDZ, problem, srpski, komentar, danas, oni, opomena, manjina, vlast, mi, pitanje
MetaLangCOPR-NEWS-COMM.-Sr	jezik, srpski, pismo, Srbin, Srbija, oni, narod, danas, mi, svet, ćirilica, ti, latinica, hrvati, hrvatski, engleski, bosanski, reč, jedan, slovo
MetaLangCOPR-NEWS-COMM.-Slo	jezik, slovenski, slovenščina, Slovenija, angleščina, šola, slovar, univerza, mi, danes, Slovenec, zgodovina, pravilno, Trubar, država, tuji, oni, praznik, politik, hvala

[5] Refinements may include statistical measures and evaluation of reference corpora.

Some indications on the 'aboutness' of the news articles is obtained already at this level. Not surprisingly, the top lexical keywords in all three data sets include corpus compilation terms, such as *jezik*, *hrvatski*, etc. Further, when we look at the news keywords in the three datasets, the most prominent shared keywords include ethnonyms and names of states. In addition, there are also notable references to the institutional and legal authority (*zakon* 'law', *ustavni* 'constitutional'; keywords denoting 'professors', 'universities', 'faculties', *SANU* 'Serbian Academy of Sciences and Arts', *SAZU* 'Slovenian Academy of Sciences and Arts'), pointing to aspects of the article sources and the voices cited.

The results also suggest some commonalities and differences that would merit further examination. In Serbian and Croatian articles, several keywords denote the 'other' from the former Yugoslav sphere, including the descriptive and language-naming adjectives (*srpski* 'Serbian' in the Croatian corpus, *hrvatski* 'Croatian' in the Serbian corpus , *bosanski* 'Bosnian', *crnogorski* 'Montenegrin' in the Serbian corpus). Concordance analysis in the Serbian data shows that these are used mainly in discussions of (il)legitimacy of national language standards, including the most recent one, Montenegrin; in the Croatian data, this use is more about in-group-out-group oppositions of Croatian and Serbian languages and influences. Still, it is of note that *Deklaracija* 'Declaration' makes it to the top keywords in the Croatian dataset, reflecting topics of language definition and differentiation (referring to publishing of the 'Declaration on the Common Language' *Deklaracija o zajedničkom jeziku* in 2017, stating that Croats, Bosniaks, Serbs and Montenegrins share a common standard language of the polycentric type). The Croatian keywords overall point to somewhat more thematic uniformity, as testified by high-ranking keywords of *Vukovar* (site of tensions concerning two-script street signs in Latin and Cyrillic script), *Penava* (the mayor of Vukovar), *manjine* 'minorities' and *Srbin* 'Serb', most frequently found precisely in these numerous articles on Vukovar and minority language rights.

By comparison, while Slovenian keywords do not contain ethnonyms of former-Yugoslav neighbours, they reflect clear foci on the national, and on the 'outside' influence more broadly — referenced as *tuji* 'foreign', and also *angleščina* 'the English language'. Concordance line analysis, however, in this case shows more thematic focus on anxieties over the status of Slovenian in educational contexts in particular, and over the influences of English as a global language and increasing language of instruction (cf. also the keywords of *šola* 'school', *univerza* 'university'). Partly related keywords

may be those referring to developments of Slovene language resources (FRAN, a portal of Slovene dictionaries and language resources, fran.si), and one personal name of an established linguist, *Ahačič*. A point of note in the Slovenian data, finally, concerns the keywords of *spol* 'sex/gender', reflecting the growing presence of debates on gender-sensitive language in the public sphere in Slovenia in the past few years.

The comments keywords, when taken overall and compared to the news articles keywords, show some overlap in the key concepts discussed, including the specific metalanguage concepts, ethnonyms and names of states, one's own and other languages. Overall, they also exhibit more diversity compared to the news articles. One point to note, for instance, is the lexeme *danas*/*danes* 'today', which is, less expectedly, among keywords in all three datasets; concordance analysis suggests that it occurs in contexts of comparisons of present-day and past language situations, or in mostly negative references to ideological changes taking place 'today'. The Slovenian keyword *zgodovina* 'history' in part overlaps with such contexts of use. Another point of interest is the presence of pronominal forms as keywords, as exemplified by : *you* in the Serbian and Croatian (sub)corpus and the oppositional *we* and *they* in all three (sub)corpora. The keyness of pronominal forms, rarely observed in other analyses of (digital) media discourse, may suggest the interactional nature of comments discourses more broadly, but also some oppositional, essentialising elements of discourses on language more specifically.

5.2. Further analytical possibilities

For purposes of projects similar to ours, i.e. following an interest in language ideologies in the geopolitical and historical context, a range of techniques is offered specifically by corpus linguistics to produce the analysis, some departing from keywords findings, others taking different routes. One finding that is particularly useful in recent discourse-oriented research is that of grouping keywords into wider *semantic macrostructures* (McEnery et al. 2016) — i.e. similar categories of meaning and function, in order to identify wider, 'global meanings, topics or themes' (van Dijk, 2009: 68). Semantic macrostructures are not pre-given relationships of individual words, which would be identifiable automatically in different types of text, but emerge as standout groupings carrying meaning and function in one specific discourse. For example, *pismo* 'script', *ćirilica* 'cyrillics' and *latinica* 'latin script' found to

be keywords above would combine into one semantic macrostructure of SCRIPT, which apparently works as a major point of discussion in the Serbian and Croatian media contexts. Independent of the keywords function, another common corpus-linguistic technique is that of *collocation analysis* — used to identify the words that tend to collocate with the core terms analysed. In the case of metalanguage corpora the core terms will usually be 'language', or language-related concepts, but can of course include any aspect of interest. Systematic co-occurence of words can often provide insights into associations, discourses and ideologies not observable from qualitative analysis alone. One example may be the collocation of *latinica (N.)* 'latin script' and *hrvatski (Adj.)* 'Croatian' found in preliminary analysis of the Serbian comments corpus, as in *hrvatska latinica* 'Croatian Latin script', *hrvatsko pismo latinica* 'Croatian script Latin', *uvođenje latinice pod hrvatskim uticajema* 'introduction of the latin script under Croatian influence', pointing to a discursive narrowing of meaning of the globally common Latin script.

Quantitative findings on their own, as presented in the above illustration, tend to be somewhat dry, especially when analysing multi-layered phenomena like metalanguage. On the one hand, I would fully agree with Mautner (2009) that corpus numerical counts can work to reify rather than problematize simplistic social representations. On the other, when used appropriately, they are a fast-working and invaluable well of empirical data that can be used as a departure point for further analysis. For socially-oriented research they are best combined with qualitative methods, such as Critical Discourse Analysis (Wodak, 2011), or in combination with methods of anthropological linguistics, linguistic ethnography and language ideology studies. Typically, the results of quantitative, corpus-based analysis are used to identify patterns in the data which are worth pursuing further, and the analysis proceeds in a hermeneutic circle, allowing the researcher to go back and forth in checking qualitative observations for reliability using quantitative methods.

For linguistic research in the narrow sense, a corpus of real-life metalanguage can contribute to varied searches of pragmatic signals of metalinguistic commentary, framing of correct and incorrect language use, mentioned language, etc. The rich annotation in MetaLangCORP can assist in searches for particular syntactic structures, parts of speech or lexical combinations. For research in pragmatics, the (sub)corpus of citizen online language can provide useful insights into metalanguage interactional patterns within unmonitored contexts of often non-standard language use.

Finally, the corpus is envisaged to contribute to far broader types of analysis on language, media and politics in post-Yugoslav successor states. It allows a comprehensive look at the media (and concomittantly, expert and institutionalised) representations concerning particular events (e.g. the Declaration on the Common Language, or proposed changes on gendered language), particular aspects of language (e.g., language in education, or language minority rights), particular language authority or particular stakeholders (e.g. language academies, or certain citizen groups). These topics have been — and hopefully will continue to be, with more depth and empirical grounding — scrutinised as part of varied research traditions in social sciences and the humanities, including sociology, anthropology, political science, culture and memory studies and nationalism studies. Crucially, it is hoped that making this kind of data available will prompt more nuanced analyses of metalanguage, and prompt more corpus-building efforts along similar lines in Slavic and other linguistic scholarship.

6. Concluding remarks

This paper introduced the MetaLangCORP, the first corpus of contemporary Slovene, Croatian and Serbian media metalanguage texts. Throughout, the underlying focus has been on the relevance of studying metalanguage from various perspectives, including the language-ideological focus on post-Yugoslav language politics, which motivated the wider research project under which MetaLangCOPR was developed. In parallel, the paper has highlighted the need for more empirical research, and hence, more real-life data resources, for studying metalanguage more systematically. The preliminary analysis of keywords and discussion of other possible directions of analysis shows the productivity of using corpus materials and corpus linguistic techniques for metalanguage analysis, but also points to the benefits of combining quantitative and qualitative methodologies, especially when one's research is socially or discursively oriented.

Beyond this project motivation, it is hoped that the first South Slavic corpus of 'language about language' can facilitate a wide range of research, from contemporary metalanguage debates, through aspects of new media language and language change, to the specific pragmatic and semantic aspects of metalanguage and metalanguage signalling. As to the wider interest in various fields of social sciences concerning aspects of language, ideology and politics, corpus-based research is expected to prompt more

empirically grounded analyses, departing from systematically collected data and verifiable insights into main concepts. The main contribution of metalanguage corpora, especially thematic ones, lies in this openness to different kinds of analysis of the many communicatively and/or socially relevant aspects of language.

In the future, we plan to complement MetaLangCORP with datasets from three more states, Macedonia, Bosnia & Hercegovina, and Montenegro, thus finalising a comprehensive resource of contemporary media metalanguage from the former Yugoslav area. Additionally, the work of collecting metalanguage posts from social media sites of Twitter and Facebook is underway, oriented towards language use and language commentary in social networking digital environments.

References

Ajšić, A. (2015) Language ideologies, public discourses, and ethnonationalism in the Balkans: A corpus-based study. *Unpublished doctoral dissertation. Northern Arizona University, Flagstaff, AZ.(UMINo. 3705442)*.

Anderson, Michael L., Andrew Fister, Bryant Lee & Danny Wang. (2004) On the frequency and types of meta-language in conversation: A preliminary report. In Proc. of the 14th Annual Conference of the Society for Text & Discourse.

Androutsopoulos, Jannis. (2011) From variation to heteroglossia in the study of computer-mediated discourse. In Crispin Thurlow & Kristine Mroczek (eds.), Digital discourse: Language in the new media, 277–298. Oxford: Oxford University Press.

Blommaert, Jan (ed.). (2009) Language ideological debates, Vol. 2. Berlin: De Gruyter Mouton.

Bogetić, K. (2016) Metalinguistic comments in teenage personal blogs: Bringing youth voices to studies of youth, language and technology. *Text & Talk*, *36*(3), 245–268.

Bogetić, K. (2017) Language is a 'beautiful creature', not an 'old fridge': Direct metaphors as corrective framing devices. *Metaphor and the Social World*, *7*(2), 190–212.

Cameron, D. (2014) 14 Gender and Language Ideologies. *The handbook of language, gender, and sexuality*, 281.

Cavanaugh, J. R. (2020) Language ideology revisited. *International Journal of the Sociology of Language*, *2020*(263), 51–57.

Ensslin, A., & Johnson, S. (2006) Language in the news: Investigating representations of 'Englishness' using WordSmith Tools. *Corpora*, *1*(2), 153–185.

Gal, S., & Woolard, K. (2014) *Languages and publics: The making of authority*. Routledge.

Garley, M. (2019) "Do they know the normal language?". *Critical Multilingualism Studies*, 7(3), 93–128.

Hermida, A., & Thurman, N. (2008) A clash of cultures: The integration of user-generated content within professional journalistic frameworks at British newspaper websites. *Journalism practice*, 2(3), 343–356.

Jakobson, R. (1960) Linguistics and poetics. In *Style in language* (pp. 350–377). MA: MIT Press.

Jaworski, Adam, Nicholas Coupland & Dariusz Galasinski (eds). (2012) Metalanguage: Social and ideological perspectives. Berlin: Mouton de Gruyter.

Johnson, S., & Ensslin, A. (Eds.). (2007) *Language in the media: Representations, identities, ideologies*. A&C Black.

Lenihan, A. (2018) Language Policy and New Media: An Age of Convergence Culture. In *The Oxford handbook of language policy and planning* (pp. 654–674). Oxford University Press.

Kamusella, T. (2018) The politics of language and nationalism in modern Central Europe. Springer.

Krzyżanowski, M., & Tucker, J. A. (2018) Re/constructing politics through social & online media: Discourses, ideologies, and mediated political practices. *Journal of Language and Politics*, *17*(2), 141–154.

Lippi-Green, R. (1997) English with an Accent: Language. Ideology and Discrimination in the United States. Routledge.

Milroy, James & Lesley Milroy. (1999) Authority in language: Investigating standard English (3rd ed.). London: Routledge.

Mautner, G. (2009) Corpora and critical discourse analysis. *Contemporary corpus linguistics*, 32–46.

Preston, D. R. (2004) Folk metalanguage. *Language Power and Social Process*, *11*, 75–104.

Scott, M. (2009) In search of a bad reference corpus. Available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.167.2638.

Squires, Lauren. (2010) Enregistering internet language. Language in Society 39(4), 457–492.

Taylor, C. (2015) *Mock Politeness in English and Italian: A Corpus-assisted Study of the Metalanguage of Sarcasm and Irony*. Lancaster University.

Van Dijk, T. A. (2009) Critical discourse studies: A sociocognitive approach. *Methods of critical discourse analysis*, *2*(1), 62–86.

Wilson, S. (2012) The creation of a corpus of English metalanguage. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 638–646).

Wodak, R. (2011) Critical linguistics and critical discourse analysis. *Discursive pragmatics*, 50–69.

SAŽETAK

Ksenija Bogetić

MetaLangCORP: PREDSTAVLJANJE PRVOGA KORPUSA MEDIJSKOGA METAJEZIKA NA SLOVENSKOM, HRVATSKOM I SRPSKOM I MOGUĆNOSTI NJEGOVE MEĐUDISCIPLINARNE PRIMJENE

Sve veći interes za metajezik, kako u lingvistici, tako i u drugim disciplinama, naglasio je prazninu koja postoji u metajezičnim korpusima i analitičkim izvorima koji spadaju među neke od najrjeđih u sklopu suvremenih dosega korpusne linvistike. Ovaj je rad usmjeren ka popunjavanju te praznine na način da u njemu predstavljamo naš metajezični korpus te ga potom koristimo u kratkoj analizi koja služi kao primjer na temelju kojega raspravljamo o mogućnostima primjene takvih izvora u lingvistici i društvenim znanostima. U radu se prvi put predstavlja MetaLangCorp, višeelmentni korpus suvremenoga medijskog metajezika prisutnoga u jezicima triju država nastalih raspadom Jugoslavije, koji je lingvistički anotiran i dostupan u slobodnome pristupu u sklopu repozitorija lingvističkih resursa CLARIN. Kako bismo korpus smjestili u kontekst, dajemo kratki prikaz značenja i značaja metajezika, kratki osvrt na postojeće napore u sastavljanju metajezičnih korpusa te predstavljamo preliminarnu analizu ključnih riječi iz MetaLangCORP-a s ciljem otvaranja šire rasprave o mogućim primjenama metajezičnih korpusa. Nadamo se da će dostupnost ovih podataka potaknuti iznijansiranije analize metajezika kao i daljnje slične napore usmjerene na stvaranje korpusa kako za slavenske, tako i za jezike koji pripadaju drugim jezičnim porodicama.

Ključne riječi: *metajezik; jezični korpusi; jezik u (novim) medijima*

Appendix 1: List of news corpus sources used for MetaLangCORP (period: 2015-2020)

Language	Source
Serbian	Politika
	Blic
	Kurir
	Danas
	Alo
	Večernje Novosti
	B92
	Srbija Danas
Croatian	24 sata
	Jutarnji list
	Večernji list
	Slobodna Dalmacija
	Novi list
	Index.hr
	Net.hr
Slovenian	Delo
	Slovenske Novice
	Dnevnik
	Večer
	Svet24
	24ur