Znanstveni radovi

UDK 811.163'42:070
811.163'322:070
Izvorni znanstveni članak
Prihvaćeno za tisak: 5. listopada 2022.
https://doi.org/10.22210/suvlin.2022.094.01

Ksenija Bogetić
Research Centre of the Slovenian Academy of Sciences and Arts, Slovenija
ksenija@zrc–sazu.si

Vuk Batanović
Innovation Center of the School of Electrical Engineering, University of Belgrade, Srbija
vuk.batanovic@ic.etf.bg.ac.rs

Nikola Ljubešić
Jožef Stefan Institute, Ljubljana
Faculty of Computer and Information Science, University of Ljubljana, Slovenija
nikola.ljubesic@ijs.si

Corpus compilation for digital humanities in lower–resourced languages: A practical look at compiling thematic digital media corpora in Serbian, Croatian and Slovenian

The digital era has unlocked unprecedented possibilities of compiling corpora of social discourse, which has brought corpus linguistic methods into closer interaction with other methods of discourse analysis and the humanities. Even when not using any specific techniques of corpus linguistics, drawing on some sort of corpus is increasingly resorted to for empirically–grounded social–scientific analysis (sometimes dubbed 'corpus–assisted discourse analysis' or 'corpus–based critical discourse analysis', cf. Hardt–Mautner 1995; Baker 2016). In the post–Yugoslav space, recent corpus developments have brought table–turning advantages in many areas of discourse research, along with an ongoing proliferation of corpora and tools. Still, for linguists and discourse analysts who embark on collecting specialized corpora for their own research purposes, many questions persist – partly due to the fast–changing background of these issues, but also due to the fact that there is still a gap in the corpus method, and in guidelines for corpus compilation, when applied beyond the anglophone contexts.

In this paper we aim to discuss some possible solutions to these difficulties, by presenting one step–by–step account of a corpus building procedure specifically for Croatian, Serbian and Slovenian, through an example of compiling a thematic corpus from digital media sources (news articles and reader comments). Following an overview of corpus types, uses and advantages in social sciences and digital humanities, we present the corpus compilation possibilities in the South Slavic language contexts, including data scraping options, permissions and ethical issues, the factors that facilitate or complicate automated collection, and corpus annotation and processing possibilities. The study shows expanding possibilities for work with the given languages, but also some persistently grey areas where researchers need to make decisions based on research expectations. Overall, the paper aims to recapitulate our own corpus compilation experience in the wider context of South–Slavic corpus linguistics and corpus linguistic approaches in the humanities more generally.

1. Introduction

The growing interest in corpus building in social sciences and digital media studies can in part indeed be attributed to affordances of the internet, but has crucially been going hand in hand with the developments in corpus linguistic scholarship and social science perspectives more broadly. For a start, having become a prime empirical approach to language data, corpus linguistics has by now stepped beyond its long–lingering links to anglophone language science and lexicography. If we consider the linguistic scholarship of the post–Yugoslav space, it is no exaggeration to say that recent corpus developments have brought table–turning advantages; the ongoing proliferation of corpora and tools is creating invaluable sources for linguistic analysis and linguistic description across varieties, dialects, registers or levels of standardness.

In parallel, however, from the past decade onwards, a major turn concerning corpus linguistics, that is felt locally too, has been the awareness of the *assistive* potential of corpora in all forms of social research. In the post–Yugoslav area, we are seeing a range of studies starting to use corpus methods to explore important social issues that have locally lacked systematic empirical study (e.g. nationalism and news media in Serbia, Bajić 2018; discourses on sexually marginalised groups in Slovenia, Gorjanc and Fišer 2020; ideas of climate change in the UK and Croatia, Bašić et al. 2020). Even when not using any specific techniques of corpus linguistics, drawing on some sort of 'corpus' is increasingly resorted to for empirically grounded social–scientific analysis. In this perspective, the general growth of digitisation truly has unlocked unprecedented possibilities of corpus compilation. Perhaps most ubiquitously, digital media text archives, citizen journalism, and social networking sites now appear to offer a mine of data on social discourses and social movements, precious for the fast–developing field of digital humanities in particular. The Web may then indeed appear "a fabulous linguists' [or any social scientists'] playground", as corpus analysts saw it at the start of the century (Kilgarriff and Grefenstette 2003: 333).

Still, if one tries to apply these possibilities in practice, to compile their own corpus for their own analysis, the quotes may lose a lot of their appeal. Figuring out what one can use on this 'playground', whether one is allowed to enter it, or able to enter it at all, will require some not–so–obvious decisions, and can lead researchers both to unrealistic plans or discouragement at the outset. In available publications it will be hard to find the answers, partly due to the fast–changing background of these issues, but also due to the fact that most CL introductory guidebooks are oriented towards linguists, and rarely discuss initial compilation issues for thematic corpora as used by humanities scholars who are not necessarily (corpus) linguists at all. Moreover, a major proportion of existing research involves English language corpora, and there is a gap in the corpus method being applied to non–western and non–English contexts. The languages of former Yugoslavia are in this sense still

considered low–resourced languages, concerning the available tools and electronic resources, but also the available guidelines or examples of good practice. The availability of data and utility of analysis via corpus tools thus get clouded when one tries to envisage corpus compilation for particular languages and particular purposes.

In this paper we present a step–by–step account of a corpus building procedure specifically for Serbian, Croatian and Slovenian, through an example of compiling a thematic corpus, intended as a brief state–of–the art snapshot of options for any researcher weighing up their approach. The focus is on using a set of popular sources: online news texts, and citizen online comments, connected by one thematic frame (in our case that of 'language', but with the method applicable to any other theme). To set compilation against its background, we very briefly outline the corpus types, uses and advantages in social sciences and digital humanities (Section 2). Subsequently, we move on to discuss the corpus compilation possibilities and solutions adopted in our own work, including data scraping options i.e. methods of automatically collecting and extracting website data, permissions and ethical issues, and the factors that facilitate or complicate automated collection from these specific sources (Section 3). We then turn to corpus annotation and processing decisions, and present the use of tools found to easily and successfully deal with Serbian, Croatian and Slovenian data, in what we will show are budding, but to date unparalleled ongoing developments that promise easy preparation of data for various types of analyses (Section 4). Finally, we evaluate the available reference corpora, and present a state–of the art list of the resources and their advantages and disadvantages (Section 5). The discussion section recapitulates our own corpus compilation experience in the wider context of South–Slavic corpus linguistics and corpus linguistic approaches in the humanities more generally.

2. Using corpora in discourse analysis, social sciences and digital humanities: Types, uses, and the example corpus

The synergy of corpus linguistic and socially oriented inquiry, while in the post–Yugoslav area still considered a 'new' endeavor, is being vastly proven productive for overcoming the potential weaknesses of qualitative and quantitative approaches. For one, the field of discourse analysis has widely adopted corpus methods to study social discourses, social changes and ideologies, with a "firmer grip on the data" (Baker 2014: 3). The same advantage, in fact, holds for all humanities research; if the major 'scientific' criticism directed against qualitative social science is the possibility of cherry–picking evidence to suit researchers' intuitions or positions (Partington 2003), drawing on a representative corpus provides a safeguard against over–relying on one's own expectations. In turn, the socially oriented shift in using corpora addresses some criticisms of corpus linguistics itself, mainly with regard to inadequate treatment of social context.

Given these central uses and advantages, before turning to the corpus compilation process, we should briefly point to some distinctions in corpus types. What most linguists will have in mind when they speak of corpora are the large, *general corpora*, or *reference corpora*, composed to be representative of a particular language or language variety. On the other hand, it is the smaller, *specialised corpora* that will be compiled and used by analysts exploring social discourse – specialised corpora do not aim to be representative of an entire language, but usually cover specific domains or genres (Brezina and McEnery 2020). Their distinct advantage is that they allow a much closer link between the corpus and the contexts in which the texts in the corpus were produced (Baker 2014); the corpus compiler is often also the analyst, so knowledge of the context allows balancing quantitative and qualitative findings.

Specialized corpora have been applied to varied ends[1], in ways that have also reflected epistemic shifts in scholarship. At the outset, in (critical) discourse analysis, smaller representative corpora on particular samples of discourse have been combined with corpus tools, for example, to identify key concepts in a discourse, find associations between social actors, or conduct synchronic and diachronic comparisons. Perhaps the best–known examples are the pioneering studies on UK media representations of migrants and asylum seekers, which uncovered patterns in anxiety–inducing images of migration as natural disaster (Baker et al. 2008, Gabrielatos and Baker 2008); some similar metaphorical representations are identified in more recent Slovenian data, in corpus–assisted analysis of a specialised corpus of online news migration discourse (Fijavž and Fišer 2020). Arguably, over time, it is the digital humanities and (digital) media studies that have become the prime demanders for corpus grounding and specialised corpus building. Much of this effort has of course been directed toward practical language technology solutions, but a part of it has evolved theoretically in fusion with social science foci. This latter interest needs to be understood in light of the broader shift of interest towards the changing nature of public discourse and social participation propelled by digital technologies today – mainly, the shift towards including user–generated data, and the emic, bottom–up, citizen discourses and citizen media. As the public sphere is increasingly pluralised and fragmented, new modes of analysis become necessary. Corpora of 'unreal' media discourses, such as Tweets, citizen journalisms, or microblogging, are then bringing precious insights into how society is framed by not only those in power, but also those with less opportunity to get their voices heard.

What is common when considering any such specialised corpora is that an appropriate corpus will rarely be available as a ready usable resource, and will need to be compiled by the analyst(s) themselves with concrete demands in mind. Our own current and past projects have used different types of digital sources, like posts from a platform of online anti–feminist groups (The Manosphere, Author et al.

1 For more exhaustive accounts on the benefits of combining corpus–linguistic and various social–scientific methodologies see e.g. Baker (2020).

2021), blogs from a teenage dating portal (Mylol.net, Bogetić 2016), texts specifically on one topic (newspaper articles on proposals of gender–sensitive language in Serbia, Bogetić forthcoming) or containing specific content features (online hate speech, de Maiti et al. 2019). For the former, e.g. blogs of one portal, collecting a random balanced sample of data may be sufficient. For the latter, e.g. texts on one particular issue of interest, compilation will be somewhat more demanding, though we find it less discussed in existing guidebooks.

In this paper we focus on and present a case study of building what will be called *thematic corpora,* as one type of specialised corpora; still, many of the steps we describe will be similar to other corpora compiled for study in social sciences and digital humanities. Thematic corpora are built for the specific purpose of investigating the discourse on a particular theme or topic of interest to the researcher, such as migration in the above, or gender–sensitive language. They can be compiled from one source (e.g. newspaper articles on *X*), but also multiple sources (e.g. newspaper articles on *X*, digital media posts on *X,* spoken interviews on *X*). In other words, they are founded on two axes, where one is the theme, issue or event, and another is the choice of discourse type and genre. The latter will depend on the former, on the theme and how the analyst wishes to approach it. Their compilation itself raises some distinct issues compared to other types of specialised corpora, from identification of topic to scraping on different platforms. Still, as we will show, it leaves several acceptable options available, sufficient tools for the languages in question, and notable possibilities for further analysis. Their nature will be clearer when considered with an example.

2.1. Case study: The project and the corpus

The corpus used for illustration here was designed for 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, spanning six states (in the present analysis only three: Slovenia, Croatia, Serbia) and the most recent period of five years (2015 through 2019). It also aims to address broader gaps in sociolinguistics and social science, 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 2019); in the (post–)Yugoslav context, the gap appears even more notable amidst little systematic investigation and often reductionist perspectives of language and conflict. Therefore, the empirical dimension and a corpus approach are central to the project ambitions in addressing existing gaps. The choice of data is also important for the goals, and hence for the corpus compilation sources: (a) media texts, 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. Milroy 2001) and (b) citizen commentary, in below–text comments sections, given the massive ongoing convergence of tradition-

al and user–generated media transforming public discourse as we know it (Lenihan 2018). Capturing the voices from the latter source of data, as yet less explored in this local context and elsewhere, is an important aspect of the project that again requires an empirical outlook and careful corpus compilation.

The corpus we need for this kind of exploration centres on one theme (that of language). The analysis can then nevertheless zoom in on specific aspects of interest, in this case language and nationhood in particular, but also any other linguistic or non–linguistic concepts that turn out to be salient in the discourse. Defining a set of major newspapers and collecting all texts that deal with a topic of (Serbian/Croatian/Slovenian) language, from any perspective, provides a representative sample in which to investigate the quantitative and qualitative patterns of meta–language discussion in the discourse selected for analysis (news media). The corpus needs to be compiled in such a way to allow quantifiable insights, searches, and basic analysis in standard concordances that use CL techniques (more on this in Section 3). In our case, we will primarily want to have insights into the stand–out concepts and groups of concepts, or into the statistical strength of connections between concepts such as those related to language and those related to the nation. Such insights are easily obtainable from once–compiled, comprehensive, representative corpora, to give us snapshots of the discourse on language from each country, to allow comparisons between countries, or comparisons between official media and citizens' discourses. Topic identification for the news texts is then a central step for the compilation of any thematic corpus of this kind, as the other types of data (comments, shares/comments) are subsequently obtained as URL–linked to these.

3. Corpus compilation: steps and decisions

3.1. Getting–in: Ethics, permissions, feasibility

Corpus work invariably yields questions relating to collecting and sharing texts, from ethics, through copyright matters, to the feasibility of collecting data at all. When embarking on our project with the three lower–resourced languages, however, the questions appeared even more acute given the scarcity of available guidelines or examples of good practice. Regardless of language, these matters are even more of a challenge when working with digital sources, the analysis of which is still a relatively young and changing sphere of work.

While digital media present unprecedented possibilities for sociolinguistic research and for corpus builders, by virtue of being public and free to access, arguably this need not mean they are public forums (Taylor and Pagliari 2018; D'Arcy and Young 2012). For one thing, the text authors, commenters and sharers will not be aware that their posts are used for scientific research. While some scholars have advocated trying to obtain permission in smaller fora or platforms (King 2009), this practice has generally been abandoned as both practically unfeasible in most cases (Jones 2015), and as compromising the aim of creating a representative corpus in

cases where only posts with permissions are retained (Milani 2017). One approach has been that for online social spaces local norms of access and visibility should be taken into account, the major issue being whether the information posted can be considered public or not (see Buchanan 2011; Zimmer 2010). To some extent it makes sense to acknowledge that in writings such as newspaper comments, or public tweets, users share thoughts with *any* strangers with *any* browsing purposes (Solberg 2010); this is different from, e.g. those sites that require extensive information upon registration, and assume a private, in–group audience (e.g. new mothers' forum, or a dating site; even though these are a continuum, see Buchanan 2011).

Overall, a 'pragmatic' (Anders 2018) and increasing approach in thinking about the ethics of internet–based research, that largely makes sense in the context of our project as well, means recognising the difficulties of conducting 'covert' research, while centrally respecting the anonymity of 'participants'. Starting from the premise that all comments analysed were posted in publicly available space, with the intention of having their views heard, it is anonymization that became our remaining and central concern for protecting the users' identities, rather than seeking permissions. With this in mind, all the pseudonyms were eliminated, as they were not needed for our study purposes; in the resulting examples in publications they can simply be changed in ways that resemble the pseudonym styles of the sites.

Aside from ethical issues, feasibility is a major consideration to evaluate at the very beginning. When it comes to newspaper articles, generally, we find that building corpora from South Slavic language sources tends to be less problematic, given the availability of newspapers' online archiving today (more on this in the following section); still, not all newspapers will archive all content online, and this needs to be checked in advance. Also, the practice is likely to vary from one country to the next – a preliminary evaluation of our further data collection from Montenegro shows several newspapers do not offer access to earlier articles online; for our currently discussed corpora from Serbia, Croatia and Slovenia, this did not pose an issue. The data scraping process can also be sped up when using available online news archives, e.g. ebart.rs, which store news from multiple newspapers and therefore eliminate the need to build separate scrapers for each news website. Nevertheless, at the time of writing most such archives tend not to be inclusive enough for researchers who wish to include, like we did, all the major newspapers from one country. Similarly to news articles, we found news article comments relatively easy to scrape, due to being archived on newspaper sites online. Finally, if one is to go further and collect e.g. social media data, as will be the case in the second part of our project, different limitations need to be considered, not only pertaining to ethics and copyright, but also to feasibility. For example, our initial intention to use Facebook data proved far more complicated, given that Facebook API allows practically no direct data collection any longer. Twitter is more accessible, and upon registering and obtaining

a developer account from the site (rather fast and uncomplicated), allows building own datasets.

Finally, to facilitate all these decisions and practical steps, from our experience we can strongly recommend what is becoming more common in discourse–oriented CL: joint work of humanities researcher/corpus linguist/programming expert, with the latter possibly assisting or advising throughout the process. A profile fitting all these areas is useful of course, though in reality increasingly rare. Some familiarity with the basics and possibilities of data collection and desired analysis is nevertheless often a prerequisite for such collaborations to indeed be fruitful.

In the rest of this section and in Section 4, our focus emerging from such collaboration will be twofold: (i) presenting the possibilities and our own choices in corpus compilation and annotation, and (ii) presenting details of computation and tool operation that may help with practical choices, and also be of use to those with more corpus technologies and programming knowledge.

3.2. Text collection and scraping

In order to construct news article corpora focused on the topic of language, we considered the following sources from Serbian, Croatian, and Slovenian media:

1. From Serbia: "Politika", "Večernje novosti" and "Danas" (major daily newspapers), "Blic", "Kurir" and "Alo!" (popular tabloids), as well as "B92" and "Srbija Danas", two well–known online media portals.
2. From Croatia: "Jutarnji list", "Večernji list", "Slobodna Dalmacija", "Novi list" (major daily newspapers), "24sata" (popular tabloid), as well as "Index.hr" and "Net.hr", two well–known online media portals.
3. From Slovenia: "Delo", "Dnevnik", "Večer" (major daily newspapers), "Slovenske novice", "Svet24" (popular tabloids), "24ur.com" online media portal.

A natural way of approaching the task of topic–focused data collection of this sort would be to access online archives of the chosen media sources, download/scrape all the articles within the selected timeframe, and perform content filtering afterwards. However, we found such an approach to be impossible, since many media websites provide no publicly available archives, effectively prohibiting us from browsing the articles published on a certain date or within a certain time scope.

Nevertheless, older articles do remain accessible on all source websites, but they can usually be reached only via a website's search engine, which is in some cases internal to the website, and in others merely an interface for performing a website–localized Google Search. We therefore decided to construct our corpora by searching the source websites using several specific queries of interest, collecting the articles returned by the search engines, and subsequently filtering them. For the Serbian and Croatian sources we used all inflections of the noun "jezik" (language) and the adjective "jezički" (linguistic) as search queries. For the Slovenian sources we used the same words, translated to Slovenian (i.e. "jezik" and "jezični"),

but we also added the noun "Slovenščina" (Slovenian), which is often found in Slovenian texts on language[2].

The scraping procedure, i.e. the collection and extraction of data from the chosen websites, was implemented in Python using the BeautifulSoup[3] and scrapy[4] libraries. In addition to the textual content of news articles, we also collected article titles, publication dates, URLs, and article author names.

In parallel to the scraping of articles, visitor comments for each article were also collected. These comments were organized in a tree structure in order to preserve their positioning in the comment discourse. Many websites, however, do not provide the option of replying to a particular comment, transforming the comment organization into a simple list. In addition, on some websites, such as B92, the links between comments and comment replies are not explicit, but instead embedded into the comment texts in the form of a quotation template. Due to such issues, preserving the proper ordering of the collected comments can be a non–trivial aspect of data collection – and one that should definitely be taken into consideration when using the data for discourse analysis.

The use of the script (Cyrillic/Latin) was considered a matter of interest, especially for discourse analyses of language ideology, where script choices may both be telling in themselves, and, as our data showed, subject of explicit meta–linguistic commentary, or even deliberate and creative script play. Thus, we kept both the original texts as well as their transliterations into the Latin script.

Each comment was assigned a unique numeric ID describing its position in the comment tree of its respective article. Similarly, each article was assigned a standardized unique ID identifying the language of the article via its ISO code, its source website via a numeric value, and its ordinal value within the source website.

3.3. Text filtering

In the text collection phase, we accumulated all articles from the chosen source websites which were possibly related to the topic of language. However, only a subset of these articles were actually focused on language issues, while in others the search keywords we used appeared in different, non–linguistic contexts. For instance, the noun "jezik" in Serbian, Croatian and Slovenian is polysemous, similarly to the noun "tongue" in English – it can either refer to language or to the tongue as an organ. For this reason, it was necessary to filter out the topically irrelevant articles (and their respective visitor comments) from the dataset.

2 Of course, selecting the scope of queries is a complex methodological question in itself (e.g. including a wider set, such as *dialect* etc., may be productive in some cases; our test checks, however, showed that our narrowed choice was adequate for capturing relevant texts); other methodological questions, such as the discursive scope of texts, are also relevant to this research but remain outside the scope of the present paper.

3 http://pypi.org/project/beautifulsoup4/

4 http://scrapy.org/

Achieving this goal effectively required performing an automated separation of article texts into two classes – relevant and irrelevant, with regard to the topic of language. After initial experimentation and a literature review, we decided to tackle this task as one of semi–supervised classification, using iterative bootstrapping. In this setup, we hand–picked several example articles for both classes in all three languages, which we used to train three binary SVM (Support Vector Machines) classifiers, one per language. These classifiers used the bag–of–n–grams approach, in which each document is treated as an unordered set of individual words and sequences of two or three consecutive words within it. The remaining articles were then placed in test sets (one test set per language) and automatically classified using the SVMs. Test set articles for which SVM classification decisions were the most certain were manually examined to verify their class, and were then used to expand the training sets, being simultaneously removed from the test set. This process was repeated as long as we were able to find new relevant articles in each iteration. In practice, the bootstrapping usually stopped once the classifier could no longer find any test set articles for which the probability of belonging to the relevant class was over 50%.

The critical resource in this filtering process was the number of truly relevant articles, since the classification problem was highly imbalanced. In other words, the texts automatically collected in the previous stage contained far more irrelevant than relevant articles. For this reason, manual checks of classifier outputs, although time–consuming, were invaluable in securing high levels of precision. Without them, it is likely that many false positive articles would end up included in the relevant set.

To illustrate the scarcity of relevant articles and the difficulty in isolating them, our initial text collection produced a corpus of 34,573 articles in Serbian, while only 1,088 of them were retained after filtering (3.15%). Similarly, out of the 32,990 collected articles in Croatian, 738 were retained as relevant (2.24%). Finally, out of the 29,105 articles in Slovenian, 555 were retained in the filtered corpus (1.91%). Detailed, per–source statistics of this type for all three languages are presented in Table 1.

Language	Source	Collected articles	Relevant articles	Relevance percentage
Serbian	Politika	3,760	245	6.52%
	Blic	5,303	174	3.28%
	Kurir	5,365	80	1.49%
	Danas	1,982	130	6.56%

	Alo	2,019	43	2.13%
	Večernje Novosti	2,782	213	7.66%
	B92	2,854	118	4.13%
	Srbija Danas	10,508	85	0.81%
	Total	34,573	1088	3.15%
Croatian	24 sata	5,670	48	0.85%
	Jutarnji list	7,117	142	2.00%
	Večernji list	7,217	225	3.12%
	Slobodna Dalmacija	446	29	6.50%
	Novi list	3,121	87	2.79%
	Index.hr	7,163	140	1.95%
	Net.hr	2,256	67	2.97%
	Total	32,990	738	2.24%
Slovenian	Delo	8,528	192	2.25%
	Slovenske Novice	4,487	29	0.65%
	Dnevnik	4,808	114	2.37%
	Večer	6,706	179	2.67%
	Svet24	1,935	15	0.78%
	24ur	2,641	26	0.98%
	Total	29,105	555	1.91%

Table 1. Collected and relevant articles per source and country/language.

As seen from the table, the percentage of relevant articles is substantially higher in major traditional media, such as Politika or Slobodna Dalmacija. On the other hand, the relevance percentage is lower in media of a more tabloid/popular nature, particularly those in the form of web portals, such as Srbija Danas and 24ur.

3.4. Disseminating the corpora: Possibilities and legal limitations

After the evaluation of options and constraints for collecting the data, and the collection process, came the decision of whether and how we may want to share our data with other researchers. In our own work, we were guided by a general commitment to the growing efforts of open data sharing, and made the corpus available to other researchers online. Still, the decision to disseminate one's own specialized corpus requires considering the options and legal limitations.

Overall, two commonly confused ways of free corpus dissemination need to be distinguished, since they have implications for the legal and ethical questions. One includes the 'free access' corpora: corpora that offer access free of charge, whose content can be viewed by using an online concordancer, allowing access to segments of the data searched, but not texts in their entirety. Many of the globally most used corpora are today disseminated in the free access mode, including the congregation of english–corpora.org with the field's most deployed sources, such as the British National Corpus (BNC) or the Corpus of American English (COCA). The practice is increasingly common for other languages in Eastern/Central Europe as well (see e.g. Erjavec 2013 for Slovenian) and different corpus types, including the more specialised ones. The second mode of dissemination is one that allows the complete takeover of material, in the 'open access corpora': corpora that allow for a user to download them to their own computer, in their entirety, free of charge. For corpus linguists, and those interested in particular item frequencies or particular lexical items, free access via concordancers is often sufficient. For scholars in the humanities interested in broader discursive patterns[5], however, open access download and analysis of texts on one's own computer is often far more advantageous.

Our own corpus set was disseminated as open access under the Creative Commons – Attribution–NonCommercial–ShareAlike 4.0 International (CC BY–NC–SA 4.0) license, published on the CLARIN.SI repository (Bogetić and Batanović 2020), following evaluation and approval. Generally, this mode of dissemination can legally be more problematic, since it allows third parties to take entire collections of texts and potentially share them in ways that we as corpus builders no longer have control over. Still, from a practical angle, for those planning to compose and disseminate corpora of texts in similar ways, it should be noted that the undefined legal practice means this kind of dissemination is as yet acted upon. An important point to also bear in mind is that upon complaints, one can always remove the text or a set of texts from the corpus. In our case, with the corpus available open access online, none of this has proven problematic to date.

5 Or those requiring any further pre–processing, such as selection of relevant parts or metadata via means of machine learning, or manual selection or labeling.

4. Linguistic annotation of corpora: The processes and tools

To put the corpus to use, some linguistic processing is typically needed to make the corpus more searchable and easier and more informative to analyse. Given the type of data we discuss in this paper, the central task of linguistic processing is to (1) lemmatise the text given its rich inflectional morphology and (2) enrich the text with basic linguistic information, mostly part–of–speech and morphosyntactic description. Annotation is in our case important and complex, as we work with a group of Slavic languages of rich inflectional morphology, and also, as we work with user–generated data that does not follow the linguistic norm. Luckily, while annotation of such material was a very hard endeavour just several years ago due to lack of freely available tools, researchers working with texts in Serbian, Croatian or Slovenian today have at their disposal a range of tools to perform the tasks for them, designed not to demand great technical competence (cf. eg. Ljubešić et al. 2016), and being freely available under very permissive licences. The process does not require the user to be familiar with the underlying approaches that the tools are based on (though some familiarity with the functions is an advantage). Note, also, that a corpus can be stored in multiple formats, as was done in our case: so, a text can be viewed as just a plain, 'clean text' or it can also be viewed with the tags, providing access to the versions that best meet our needs at different steps of analysis.

This section will cover four main topics: (i) a very basic description of the logic of supervised machine learning, the driving force behind the technologies presented here, together with thus based taggers that we chose to use (ii) the various approaches to text annotation, together with our own choices and (iii) the limitations of the machine–learning–based automatic processing of language data each user should be aware of.

4.1. Machine–learning–based linguistic processing

Since the mid–1990s, the dominant paradigm in developing technologies for linguistic processing, also called language technologies, has been machine learning. This paradigm allows computers to solve language–related problems (machine translation, text normalization, part–of–speech tagging, etc.) by learning from examples, i.e., from data instances in which the task at hand has already been solved by humans. For lemmatization, such examples meant for learning would be sentences broken up into tokens (words and punctuation), with each token being mapped to the canonical, dictionary form of that token. For part–of–speech tagging, such data would be sentences with each token having a manually assigned part–of–speech tag. Such datasets are called "manually annotated" or "training" datasets, as they are used for training computer programs called language tools that automate the task initially performed manually by humans.

While machine–learning–based language technologies serve a great role – the automatic annotation of large quantities of text that would be absolutely infeasible

if done manually, they do have a series of limitations, the most prominent being that they cannot learn much beyond what they have seen data–wise during their training. While they have a generalization capability, and a similar sentence will be correctly annotated, if the text moves too far away from what examples the computer has seen during its training, performance will decrease drastically.

One older option for linguistic processing of our languages of interest is the freely available ReLDI tagger (Ljubešić et al. 2016). The tagger is known to have high success rates for working with all three of our languages of interest, and was found suitable in our own work. Limitations are nevertheless occasionally observed, and need to be borne in mind, depending on one's research purposes. As an example of a failure, we can use a sentence from our Croatian comments corpus, a sentence written in the Zagreb vernacular, where the Kajkavian variant of the interrogative pronoun "što", namely "kaj" is used. The example sentence "Kaj si ti konzumiralo, dijete?", results in a wrong part–of–speech tagging result, namely the following tags: "VERB AUX PRON VERB PUNCT NOUN PUNCT". While the remainder of the sentence is correctly tagged, the interrogative pronoun at the first position is tagged as a verb. This happened for a very simple reason – in the Croatian hr500k training dataset, whose aim is to be representative of standard Croatian language, the interrogative pronoun "kaj" occurs very infrequently, and not always in the role of a pronoun.

Despite usability and common use of the ReLDI tagger, our own choice was hence different, and involved the CLASSLA pipeline (Ljubešić and Dobrovoljc, 2019)6, as a newer tool version. We will briefly discuss its machine–learning based approach here, to help sketch the operation of the technologies in very basic terms and the advantages of the recently developed pipeline.

To help language technologies to deal better with language variation, and the limitations of training datasets, in the last decade there was a very important development in the area of machine learning, especially in the realm of language processing, called model pre–training. While we regularly train our models on the training datasets, that are of limited size, we have much more texts of various languages available, that are not manually annotated in any way. Given the distributional hypothesis "You shall know a word by the company it keeps" (Firth, 1957), the idea is to numerically represent the meaning of a specific word given all the contexts that word has occurred in inside a large collection of texts.

While there are different approaches to calculating the numerical representations of meaning of words nowadays, in the CLASSLA pipeline the concept of word embeddings is used, where each word is represented through 100 latent semantic numerical variables, i.e., numbers. The word embeddings for the Croatian language (Ljubešić, 2018) are learned from 1.7 billion words of texts harvested from the Croatian web and Croatian news portals. Inspecting the representation of the word "kaj" in comparison to representations of other words, one can observe that the

6 https://pypi.org/project/classla/

closest neighbours of that word are, inter alia, the standard interrogative pronoun "što", its Serbian variant "šta", their variants without diacritics "sto" and "sta", as well as the Čakavian variant of the interrogative pronoun, "ča". Given that the tool of our choice, the CLASSLA pipeline, has the described embeddings available, employing this tool to perform part–of–speech tagging of text written in Croatian, our example sentence "Kaj si ti konzumiralo, dijete?" is correctly part–of–speech tagged as "PRON AUX PRON VERB PUNCT NOUN PUNCT", regardless of the fact that it was trained on the hr500k dataset only, same as the ReLDI–tagger, which did not manage to deal with this level of linguistic variation. The main reason for the success of the CLASSLA tool is the fact that it uses the word embedding collection that was pre–trained on a collection of raw, non–annotated Croatian texts in which the word "kaj" occurred more than 250 thousand times.

4.2. Text annotation

The first type of annotation we needed for data analysis is lemmatization. Lemmatization deals with the inflectional morphological variation, e.g., mapping of the 3rd singular present verb form "radi" to its dictionary infinitive form "raditi". Lemmatization is especially relevant for highly inflected languages, such as the three languages that we are dealing with here, as it enables a much simpler lookup of specific lexemes, regardless of the morphological inflectional form they are in. Even more importantly, when analysing a discourse in terms of keywords and key concepts, as will be done as part of the present investigation on language–ideological themes in media texts, lemmatised data give much more accurate results (for example, individual words *radi* 'to work 3p sg', *radiš* '2p sg' may not reach statistical keyness, but the overarching lemma *raditi* 'to work' may do so).

The second annotation step involved part–of–speech tagging and morphosyntactic description. In our corpus work, we have used the part–of–speech tagset of the Universal Dependencies project (De Marneffe et al. 2021) and the MULTEXT–East morphosyntactic tagset (Erjavec, 2012). These are closely connected, the latter for the most part being only more detailed. For instance, the MULTEXT–East tagset has a category for abbreviations, while the Universal Dependencies tagset does not, requiring the word to be simply labeled with the part–of–speech of the expanded form. If our analysis was to focus, for instance, on abbreviations themselves, or on metalinguistic comments surrounding abbreviations, the MULTEXT–East tagset will allow fast identification of all such instances. On the other hand, the Universal Dependencies tagset may appear more simple for interpretation, and in some cases, more compatible with existing tools and concordancers. This also highlights a very important feature of any linguistic annotation process: each follows a specific formalism with its specificities, and the user should be very well acquainted with that formalism in advance.

In Table 2 we can observe the levels of token, lemma, part–of–speech information and morphosyntactic description. While the first column contains the tokens

of the sentence ("znači", "su"), the second column consists of manually assigned lemmas ("značiti", "biti"), the third one part–of–speech information (e.g. verb), with the last column containing the token's manually assigned morphosyntactic description (e.g. main verb, present tense, third person, singular).

Token	Lemma	Part–of–speech	Morphosyntactic description
to	taj	DET	Pd–nsn
ne	ne	PART	Qz
znači	značiti	VERB	Vmr3s
da	da	SCONJ	Cs
su	biti	AUX	Var3p
posljednje	posljednji	ADJ	Agpfpny
tri	tri	NUM	Mlc
riječi	riječ	NOUN	Ncfpn
„	„	PUNCT	Z
nepravilne	nepravilan	ADJ	Agpfpny
“	“	PUNCT	Z

Table 2. An example of an annotated sentence from the Croatian corpus

Finally, for the kind of data used in our project it is important to decide how to approach the abundance of non–standard, computer mediated language. User–generated internet content such as online news comments that we collected is especially known to contain a great amount of noise and non–standard writing, such as abbreviations, erratic punctuation, misspellings, colloquial and dialectal expressions, sometimes jointly described as cyberorthography (King 2009).

There are two approaches to processing non–standard text; for work with Slavic languages, the decisions will depend both on the type of corpora and analyses planned, and the resources available for the particular language.

One possibility is to use non–standard text normalizers, i.e., tools that work to turn non–standard forms into their normalized variants, and once the whole text/dataset is normalized, to process the text as any standard text. An example of an instance of this type of normalization would be the mapping of the non–standard "kaj" variant of the interrogative pronoun into the standard "što" form. For this approach, training datasets consisting of non–standard texts and their normalizations are required. Another option is to skip this step of normalization, but to rely on annotation tools that can recognize and identify non–standard forms and tag them accurately. There have been increasing efforts in this direction for South Slavic languages, which involve adapting standard language tools to non–standard

language by training them also on examples of non–standard text (Ljubešić et al. 2017). This approach requires additional non–standard manually annotated training data. Research has shown (Zupan et al., 2019) that in cases where few resources are available for producing manually annotated training data, the normalization approach achieves better results, but that with reasonable resources for producing training data, better results are obtained with the approach of adapting the whole toolchain to non–standard language. This approach was used in the CLASSLA pipeline, where non–standard models are trained on a combination of standard and non–standard data. In other words, our choice in the end involved relying on the CLASSLA pipeline to adequately work with standard and non–standard data, and to skip normalization of the whole dataset. Initial assessment of results shows the approach to be quite productive.

Finally, of course, when designing a thematic corpus for discourse–analytical purposes, the degree of corpus annotation can vary greatly, depending on one's research goals, and on whether the corpus is intended for sharing as an open resource for future research. Considering one's goals, it is possible that analysts find less detailed annotation quite sufficient for their study purposes; based on our work, we would nevertheless recommend at least lemmatisation when working with online Slavic writing. Concerning corpus sharing, it may be beneficial to use as much annotation as possible to make the most of a corpus usability in future studies, though this will to a great extent depend on research costs, infrastructure availability and external, technical collaborators in the corpus compilation process.

4.3. The limitations of machine–learning–based linguistic processing

While the linguistic processing presented in this section has a great positive impact on the usefulness of a thematic or any other type of corpus, there are limitations. We have tried to put some of them forward already, and hope to suggest a systematisation of those below.

1. Automatic linguistic processing is based on models trained on the manually annotated training data. While significant resources are invested in producing these datasets, these annotations are still not perfect. Human annotators do make mistakes and there are always errors to be expected already in the data that we teach computers on. It is expected that some of these errors will be propagated to automatic annotations in large corpora.
2. The machine learning models are trained on a limited amount of data. There is capacity in those models for generalization (e.g., using the fact that adjectives are frequently followed by nouns etc.), but this generalization capacity does not fully parallel that of humans.
3. We do have an estimate of the amount of error that computers produce on unseen texts. For part–of–speech tagging and lemmatisation roughly 2% (1 in 50) of the tokens are wrongly annotated. The level of morphosyntac-

tic description carries a level of error of around 5%, meaning that 1 in 20 tokens will be erroneously annotated on that level.

4. The error in automatic annotation of texts in the most important layers is generally low, two to five percent; it should nevertheless be understood that the errors are not random, but are mostly present in phenomena that are either not well handled in the formalism applied, are badly annotated in the training data, or are infrequent in the training data. If the phenomena of interest to a researcher are those that are badly annotated, relying on such annotations could prove to be disastrous for one's research.
5. The training data is not only limited in size, but also in the representativeness of all the possible phenomena that can occur in language. While for standard language mostly newspaper data is used, and these do quite a good job at representing the language in general, one can easily assume that such data will not perform as well on very different genres, such as lyrical texts.
6. Each linguistic formalism is an approximation of the linguistic phenomena described. Not every user of every corpus will be satisfied with the solutions in specific formalisms, and some formalisms will not serve well specific research questions.

Overall, despite these limitations – which may be somewhat more notable for low–resourced languages, but are an unavoidable feature of machine–learning–based linguistic processing regardless of language – data annotation is becoming a more accurate and less demanding task, especially given the development of tools that help with it. In our case, the use of the CLASSLA pipeline was found useful and easy to apply; tools of this kind are useful even when one does not have in–depth knowledge of the machine–learning based processing and the mechanisms behind it. In this sense, a discourse analyst working with one's own corpus may resort to the use of CLASSLA, ReLDIAnno[7] (a web service using the ReLDI tagger) and similar tools, regardless of programming skills, though they should bear the above limitations in mind when planning their own research. Still, a collaboration with programming and corpus technologies experts is what we recommend as particularly beneficial when available.

5. Choosing a reference corpus

After all compilation questions are resolved, the corpora are ready for use. Still, using them in practice for discourse analytical purposes will likely require using them with a reference corpus – a large corpus of general language, which is used to compute patterns in the corpus of study, such as keywords.

7 http://clarin.si/services/web/query

Namely, if one is to use the basic CL analysis techniques, most notably the *keywords* function, this last step merits consideration. Keyword identification – a statistical approach to word frequencies to identify words occurring with unusual frequency in a given text, by comparing word frequencies in the compiled specialized corpus with those of a larger reference corpus – is a useful first step in most quantitative analysis of a social discourse. Keywords provide insights into central concepts in a discourse, showing the 'aboutness' of a material, and are subsequently analysable in different ways. It was important to the present project, and likely to be useful for other analysts, obtainable easily through standard software. Hence, finding an appropriate reference corpus to use is an important decision that will follow one's own corpus compilation, and not always easy for lower–resource languages. For the languages we worked in, there are currently fast ongoing developments in this direction, so we present a state–of–the–art list below (Table 3).

Corpus	**Link**	**Language**	**Sources / genre**	**Size**	**Annotation**	**Access**
GigaFida	https://viri.cjvt.si/gigafida/ http://hdl.handle.net/11356/1320	Slovenian	daily news, magazines, web texts, and different types of publications (fiction, schoolbooks, and non–fiction)	1.1 billion words	Morpho-syntactically annotated and lemmatised	freely available for search
Croatian web corpus hrWaC	http://hdl.handle.net/11356/1064 https://www.clarin.si/noske/run.cgi/corp_info?corpname=hrwac	Croatian	texts from the Croatian top–level web domain (.hr)	1.4 billion tokens	Morpho-syntactically annotated and lemmatised	freely available for download (CC–BY–SA) and search
Croatian language corpus Riznica	http://hdl.handle.net/11356/1180 https://www.clarin.si/noske/run.cgi/corp_info?corpname=riznica	Croatian	8% of fiction texts and 72% of specialized texts	102 million tokens	Morpho-syntactically annotated and lemmatised	freely available for download (CC–BY–SA) and search

Korpus savremenog srpskog jezika SrpKor2013	http://korpus.matf.bg.ac.rs/prezentacija/korpusi.html	Serbian	Literature, popular science, news	122 million words	Morphologically annotated (website notes the annotation as incomplete)	upon request at korpus@matf.bg.ac.rs
Lematizirani korpus savremenog srpskog jezika (SrpLemKor)	http://www.korpus.matf.bg.ac.rs/SrpLemKor/	Serbian	Literature, science, news, law	3,7 million words	Lemmatised and PoS Annotated	upon request under the terms of CC_BY–NC licence
Serbian web corpus srWaC	http://hdl.handle.net/11356/1063 https://www.clarin.si/noske/run.cgi/corp_info?corpname=srwac	Serbian	texts from the Serbian top–level web domain (.rs)	555 million tokens	Morpho-syntactically annotated and lemmatised	freely available for download (CC–BY–SA) and search

Table 3. An overview of reference corpora for Slovenian, Croatian, and Serbian

Corpus choice will of course depend on study aims, the size of one's own reference corpus, genre of data, etc. We are still evaluating the different corpora available for our own work, but must point out to possible difficulties, such as the time scope – reference corpora that are even just a decade older are bound to e.g. yield keywords that simply reflect new words or concepts, e.g. "covid", rather than discourse foci. Another option that has recently been suggested for lower–resourced languages is to compile one's own '–hoc reference corpus' (Kania, 2021) from the time frames and genres suitable for comparison. In this sense, lower–resourced languages are sometimes described as a "blessing in disguise" (ibid.) as they call for more careful consideration of limitations of reference corpora.

6. Discussion and conclusions

Language corpora have presented great opportunities for social science research beyond linguistics, and are attracting increasing interest in post–Yugoslav scholarship, both from perspectives of corpus building and corpus use. Still, corpus compilation and use includes a range of steps that are typically fuzzy to researchers of social discourse, especially given the absence of publications that deal with it explicitly in this language context. We have used our own project to systematise

these explicitly by bringing to spotlight the compilation of specialised corpora, specifically thematic corpora, which may be of greatest use in the humanities–disciplinary approaches.

Topic–focused data collection of media content can take different paths, which we have touched upon. Still, for the languages in question we found that, unfortunately, it is not possible to simply use online archives of selected media sources, as is the common practice in many other languages, given that the local media texts are not adequately congregated at any such archive. We have presented the alternative approach of querying search engines and subsequent filtering, which proved successful in our own research. Additionally, issues of ethics and copyright have nevertheless presented dilemmas, as a grey area in the South Slavic space (but also more broadly), requiring consideration of corpus dissemination. Our own approach is committed to open access sharing, which appears to be growing in Slavic corpus–building. We hope our corpora to be a contribution in this direction.

The step of corpus annotation and processing was found to be relatively straightforward for this kind of data, including the reader comments (sub)corpora that are more complex by virtue of non–standard language. There are increasingly available tools for this kind of task in South Slavic languages, though each will have limitations that are important to understand and that we have striven to point out. In this respect, collaboration of discourse analysts and corpus/programming experts is becoming more common; we found it to be very productive in our own work, and can recommend it when project capabilities allow.

Finally, this account merits a word of caution. We have tried to point out the limitations of corpus preparation and use that we have encountered in the course of our own project, such as limitations of machine–learning–based annotation, or limitations regarding suitable reference corpora). In addition, however, we find it is important for researchers in social sciences to be aware of limitations of corpus–based discourse analysis more broadly. For analysing social ideologies, identities and relations of power, as Motschenbacher (2018) points out, unreflected use of CL can have limited destabilising and de–essentialising potential. Grounding social analysis *primarily* in numbers can be misleading, both given the limitations of corpora and the complexities of the social discourse encapsulated in a corpus. In this respect, a careful synergy of an empirical, corpus driven approach with critical analysis in social context is key to avoiding little–revealing or even reductionist conclusions granted only by numerical patterns – and certainly a door to great possibilities of research where corpus use is indispensable, and hopefully to grow in the future of Slavic scholarship.

7. References

Baker, Paul (2020). Corpus–assisted discourse analysis. Hart, Christopher, ed. *Researching Discourse. A Student Guide*. Routledge, 124–142

Baker, Paul (2014). Bad wigs and screaming mimis': Using corpus–assisted techniques to carry out critical discourse analysis of the representation of trans people in the British press. Hart, Christopher and Piotr Cap, eds. *Contemporary critical discourse studies*. London: Bloomsbury, 211–235.

Bašić, Ivana, Marina Grubišić, Snježana Veselica–Majhut (2020). Diskursno oblikovanje klimatskih promjena u anglofonim i hrvatskim izvorima informiranja. *Suvremena lingvistika* 46(89): 1–23, https://doi.org/10.22210/suvlin.2020.089.01

Brezina, Vaclav and Tony McEnery (2020). Introduction to Corpus Linguistics. Tracy–Ventura, Nicole and Magali Paquot, eds. *The Routledge Handbook of Second Language Acquisition and Corpora*. Routledge, 11–22

Bogetić, Ksenija (2016). Metalinguistic comments in teenage personal blogs: Bringing youth voices to studies of youth, language and technology. *Text & Talk* 36(3): 245–268, https://doi.org/10.1515/text–2016–0012

Bogetić, Ksenija and Vuk Batanović (2020). Annotated corpus of Slovenian language–related news articles MetaLangNEWS–Sl. *Slovenian language resource repository CLARIN.SI*, ISSN 2820–4042, http://hdl.handle.net/11356/1360

Bogetić, Ksenija and Vuk Batanović (2020). Annotated corpus of Slovenian language–related news comments MetaLangNEWS–COMMENTS–Sl. *Slovenian language resource repository CLARIN.SI*, ISSN 2820–4042, http://hdl.handle.net/11356/1362

Bogetić, Ksenija and Vuk Batanović (2020). Annotated corpus of Croatian language–related news articles MetaLangNEWS–Hr. *Slovenian language resource repository CLARIN.SI*, ISSN 2820–4042, http://hdl.handle.net/11356/1369

Bogetić, Ksenija and Vuk Batanović (2020). Annotated corpus of Croatian language–related news comments MetaLangNEWS–COMMENTS–Hr. *Slovenian language resource repository CLARIN.SI*, ISSN 2820–4042, http://hdl.handle.net/11356/1370

Bogetić, Ksenija and Vuk Batanović (2020). Annotated corpus of Serbian language–related news articles MetaLangNEWS–Sr. *Slovenian language resource repository CLARIN.SI*, ISSN 2820–4042, http://hdl.handle.net/11356/1371

Bogetić, Ksenija and Vuk Batanović (2020). Annotated corpus of Serbian language–related news comments MetaLangNEWS–COMMENTS–Sr. *Slovenian language resource repository CLARIN.SI*, ISSN 2820–4042, http://hdl.handle.net/11356/1372

Buchanan, Elizabeth A. (2011). Internet Research Ethics: Past, present, and future. Consalvo, Mia and Charles Ess, eds. *The handbook of Internet studies*. Blackwell Publishing, 83–108, https://doi.org/10.1002/9781444314861.ch5

D'Arcy, Alexandra and Taylor Marie Young (2012). Ethics and social media: Implications for sociolinguistics in the networked public. *Journal of Sociolinguistics* 16(4): 532–546, https://doi.org/10.1111/j.1467–9841.2012.00543.x

de Marneffe, Marie–Catherine, Christopher D. Manning, Joakim Nivre and Daniel Zeman (2021). Universal dependencies. *Computational linguistics* 47(2): 255–308, https://doi.org/10.1162/COLI_a_00402

Erjavec, Tomaž (2012). MULTEXT–East: Morphosyntactic resources for Central and Eastern European languages. *Language resources and evaluation* 46(1): 131–42

Erjavec, Tomaž (2013). Korpusi in konkordančniki na strežniku nl. ijs. si. *Slovenščina 2.0: empirical, applied and interdisciplinary research* 1(1): 24–49

Goh, Gwang–Yoon (2011). Choosing a reference corpus for keyword calculation. *Linguistic Research* 28(1): 239–256

Gorjanc, Vojko and Darja Fišer (2018). Twitter in razmerja moči: diskurzna analiza kampanj ob referendumu za izenačitev zakonskih zvez v Sloveniji. *Slavisticna Revija* 66(4): 473–495

Fijavž, Zoran and Darja Fišer (2020). Corpus–assisted analysis of water flow metaphors in Slovene online news migration discourse of 2015. Fišer, Darja and Philippa Smith, eds. *The Dark Side of Digital Platforms.* Ljubljana University Press, Faculty of Arts, 56–84

Gabrielatos, Costas and Paul Baker (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996–2005. *Journal of English Linguistics* 36(1):5–38, https://doi.org/10.1177/0075424207311247

Kania, Ursula (2020). Marriage for all ('Ehe fuer alle')?! A corpus–assisted discourse analysis of the marriage equality debate in Germany. *Critical Discourse Studies* 17(2): 138–155, https://doi.org/10.1080/17405904.2019.1656656

Kamusella, Tomasz D. (2019). The fallacy of national studies. Fellerer, Jan Robert Pyrah and Marius Turda, eds. *Identities in–between in East–Central Europe*. Routledge

Kilgarriff, Adam and Gregory Grefenstette (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics* 29(3): 333–347, https://doi.org/10.1162/089120103322711569

King, B. (2009). Building and analysing corpora of computer–mediated communication. Baker, Paul, ed. *Contemporary Corpus Linguistics*. London: Continuum, 301–320

de Maiti, K. Pahor, Darja Fišer and Nikola Ljubešić (2019). How haters write: analysis of nonstandard language in online hate speech. *7th Conference on Computer–Mediated Communication (CMC) and Social Media Corpora (CMC–Corpora 2019)*, Cergy–Pontoise, France. 37–42

Ljubešić, Nikola, Tomaž Erjavec and Darja Fišer (2017). Adapting a State–of–the–Art Tagger for South Slavic Languages to Non–Standard Text. *Proceedings of the 6th Workshop on Balto–Slavic Natural Language Processing*. 60–68

Ljubešić, Nikola (2018). Word embeddings CLARIN.SI–embed.hr 1.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1205

Milani, Tommaso M. (2017). Language and sexuality. *The Oxford Handbook of Language and Society*. Oxford University Press, 403–422

Milroy, James (2001). Language ideologies and the consequences of standardization. *Journal of Sociolinguistics* 5(4): 530–555, https://doi.org/10.1111/1467–9481.00163

Motschenbacher, Heiko (2018). Corpus linguistics in language and sexuality studies: Taking stock and looking ahead. *Journal of Language and Sexuality* 7(2): 145–174, https://doi.org/10.1075/jls.17019.mot

Stubbs, Michael (2001). *Words and Phrases: Corpus Studies in Lexical Semantics*. Oxford: Blackwell

Taylor, Joanna and Claudia Pagliari (2018). Mining social media data: how are research sponsors and researchers addressing the ethical challenges?. *Research Ethics* 14(2): 1–39, https://doi.org/10.1177/1747016117738559

Zimmer, Michael (2010). "But the data is already public": on the ethics of research in Facebook. *Ethics and information technology* 12(4): 313–325, https://doi.org/10.1007/s10676–010–9227–5

Kompiliranje korpusa u digitalnim humanističkim znanostima u jezicima s ograničenim resursima: o praksi kompiliranja tematskih korpusa iz digitalnih medija za srpski, hrvatski i slovenski

Digitalno doba otvorilo je nove mogućnosti za sastavljanje korpusa društvenog diskursa, što je korpusnolingvističke metode približilo drugim metodama analize diskursa te humanističkim znanostima. Čak i kada se ne koriste nikakve specifične tehnike korpusne lingvistike, danas je za empirijski utemeljenu društveno–znanstvenu analizu sve učestalije korištenje neke vrste korpusa ('korpusno–asistirana analiza diskursa' ili 'kritička korpusna analiza', Hardt–Mautner 1995; Baker 2016). U postjugoslavenskom prostoru, nedavni razvoj korpusne lingvistike donio je prednosti u mnogim područjima istraživanja. Ipak, za lingviste i analitičare diskursa koji se upuštaju u prikupljanje specijaliziranih korpusa za vlastite istraživačke svrhe, i dalje ostaju otvorena mnoga pitanja – djelomično zbog pozadine korpusne lingvistike koja se brzo mijenja, ali i zbog činjenice da još uvijek postoji rascjep u poznavanju korpusnih metoda, kao i metodologije sastavljanja korpusa izvan anglofonskog konteksta. Ovim radom pokušavamo smanjiti spomenuti rascjep predstavljajući jedan postupni prikaz postupka izgradnje korpusa za hrvatski, srpski i slovenski, kroz primjer sastavljanja tematskog korpusa iz digitalnih medija (novinski članci i komentari čitatelja). Nakon pregleda tipova korpusa, korištenja i prednosti u društvenim znanostima i digitalnim humanističkim znanostima, predstavljamo mogućnosti sastavljanja korpusa u južnoslavenskim jezičnim kontekstima, uključujući opcije preuzimanja podataka s mreže, dozvola i etičkih pitanja, čimbenika koji olakšavaju ili otežavaju automatizirano prikupljanje i označavanje korpusa i mogućnosti obrade. Studija otkriva sve veće mogućnosti za rad s danim jezicima, ali i neka uporno siva područja u kojima istraživači trebaju donositi odluke na temelju istraživačkih očekivanja. Općenito, rad ima za cilj rekapitulirati vlastito iskustvo sastavljanja korpusa u širem kontekstu južnoslavenske korpusne lingvistike i korpusnih lingvističkih pristupa u humanističkim znanostima općenito.

Keywords: corpus linguistics, corpus compilation, corpora and discourse analysis, digital media
Ključne riječi: korpusna lingvistika, kompilacija korpusa, korpusi i analiza diskursa, digitalni mediji