V zeleni drželi zeleni breg

Studies in Honor of Marc L. Greenberg

Edited by

Stephen M. Dickey &
Mark Richard Lauersdorf

Bloomington, Indiana, 2018

SLAVICA

ISBN 978-0-89357-491-8

Technical Editors: Phillip Weirich and Renata Uzzell

Cover Design by Tracey Theriault

Cover Image: Aleš Maučec (www.facebook.com/ales.maucec). Slovenia (Prekmurje, Goričko), Vučja Gomila—a small village (pop. 269) in the northeastern part of Slovenia near the border with Hungary.
© Aleš Maučec, permission granted by the author.

Library of Congress Cataloging-in-Publication Data

Names: Dickey, Stephen M., editor. | Lauersdorf, Mark Richard, editor. | Greenberg, Marc L., honoree.
Title: V zeleni drželi zeleni breg : studies in honor of Marc L. Greenberg / edited by Stephen M. Dickey,   Mark Richard Lauersdorf.
Description: Bloomington, Indiana : Slavica Publishers, 2018.
Identifiers: LCCN 2018037963 | ISBN 9780893574918
Subjects: LCSH: Slavic languages. | Slavic languages, Southern.
Classification: LCC PG14.G54 V33 2018 | DDC 491.8/1--dc23
LC record available at https://lccn.loc.gov/2018037963

Slavica Publishers
Indiana University
1430 N. Willis Drive
Bloomington, IN 47404-2146
USA

[Tel.] 1-812-856-4186
[Toll-free] 1-877-SLAVICA
[Fax] 1-812-856-4187
[Email] slavica@indiana.edu
[www] http://www.slavica.com/

Contents

Evaluating the Effects of Language Planning Efforts in Croatia: Evidence from Corpus Data

Keith Langston

Abstract: Croatian language planning has been the subject of a substantial amount of research, especially in the period since Croatia declared its independence in 1991. This paper discusses language planning efforts in Croatia from the 1990s to the present and uses corpus data from the Croatian Web Corpus (HrWaC, v. 2.2) to evaluate the acceptance of various changes that have been advocated for the Croatian standard language. The data suggest that language planning efforts in Croatia have had some effect on patterns of usage, but that the results are uneven, varying for different lexical items as well as for orthographical or other changes that have been proposed.

1. Introduction

Croatian language planning has been the subject of a substantial amount of research, due to the specific historical development of the standard language and political changes that took place in the 20th century. Leading intellectual figures in the 19th century promoted the creation of a modern standard language that could unify the Croatian and Serbian peoples, and official policies in the 20th century advanced the idea of a single Serbo-Croatian language, both in the interwar years in the Kingdom of Yugoslavia and after World War II in the Socialist Federal Republic of Yugoslavia (SFRY). However, many Croats felt that the goal of a unified language was being pursued at the expense of traditional Croatian patterns of language usage, with Serbian forms officially preferred in many instances, and that not just Croatian linguistic norms but also other expressions of Croatian national identity and autonomy were being unfairly suppressed. During the tumultuous years leading up to and following the collapse of the SFRY in the early 1990s, many people felt that it was necessary to reassert the status of Croatian as a separate language, distinct from Serbian, as part of their struggle for the recognition of the Croats as a distinct nation with the right to political self-determination. This assertion of Croatian linguistic identity has been manifested in efforts to purify the language, to reverse the effects of previous language policies, and to further

Stephen M. Dickey and Mark Richard Lauersdorf, eds. *V zeleni drželi zeleni breg: Studies in Honor of Marc L. Greenberg*. Bloomington, IN: Slavica Publishers, 2018, 157–98.

differentiate the language from Serbian. However, these efforts have also been controversial and have been met with varying degrees of resistance.[1]

Research on language planning is largely "future-oriented" (Eastman 1983: 3), with the goal of understanding how to bring about language change and to assist in the creation and implementation of effective policies. It often focuses on the initial selection of a variety to serve as the basis for standardization, the codification of norms, efforts to gain acceptance for this emerging standard, and its elaboration to allow it to function in all spheres of society, as in the model of language planning processes proposed by Haugen (1983). However, the theoretical models that have been proposed are largely descriptive rather than explanatory in nature (Haugen 1983: 276), and are not directly applicable to the Croatian situation, which can instead be seen as an attempt at restandardization of an already-existing norm, undertaken for purely symbolic reasons (Langston and Peti-Stantić 2014: 276). Furthermore, assessments of the effects of recent language planning efforts in Croatia have been largely anecdotal, with some exceptions (see below).

The present work uses corpus data from the large (1.2 billion word) Croatian Web Corpus (HrWaC, v. 2.2) in an attempt to gauge the acceptance of various proposed changes to the Croatian standard language. These results are compared with earlier patterns of usage, as attested in the frequency dictionary by Moguš, Bratanić, and Tadić (1999), based on a corpus of texts dating from the 1930s to mid 1970s, and with data from the Serbian Web Corpus (SrWaC, v. 1.2, approximately 480 million words; see Ljubešić and Klubička 2014 for more information on these two web corpora). The remainder of the article is organized as follows. Section 2 discusses language planning efforts in Croatia since 1990 and some of the changes that have been proposed. Section 3 describes previous corpus-based research on Croatian language planning, and section 4 presents the results of the current research. Additional discussion and conclusions are given in section 5.

[1] Greenberg 2011 describes 19th-century Croatian efforts to promote South Slavic unity, in the Illyrian Movement and its aftermath, and Greenberg 2015 discusses the rise and fall of Serbo-Croatian in the context of global perspectives on sociolinguistics. For a more detailed account of the complex standardization processes in the former Yugoslavia and language planning in Croatia, see also Langston and Peti-Stantić 2014. As pointed out by a reviewer, the idea of Croatian as a distinct language is (to all appearances) broadly accepted within Croatia; it is designated as the official language in the constitution and recognized as an official language of the EU, so its status is secure. However, the desire for recognition of the language under its own distinct name does not necessarily translate into support for efforts to modify the norms of standard usage.

2. Croatian Language Planning in the 1990s and Beyond

Linguistic purism is not a new phenomenon in Croatia. According to Skelin Horvat (2004: 99): "Due to the unfavorable socio-political situation in Croatia throughout the 20th century, a specific type of purism developed, which aimed to eliminate Serbisms from the Croatian language and to clearly differentiate Croatian from Serbian. Since 1990, again due to the specific socio-political situation, Serbisms, alongside Anglicisms, have attracted the greatest attention from purists."[2] Borrowings from English have attracted a growing amount of attention, due to the influence of English as a global language, but the elimination of foreign borrowings is also seen as a means of differentiating Croatian from Serbian, since the latter is viewed as being historically more open to loanwords.[3] Despite a long tradition of grammars, dictionaries, and other normative works, dating back to the 19th century and earlier, there remains a sense in some quarters that the Croatian standard is deficient in some way:

> Despite its rich tradition of language handbooks, Croatian even today is not a completely standardized language. The instability of the Croatian norm is attested by the foundation of the *Vijeće za normu*,[4] the often conflicting opinions of Croatian linguists, public debates in the media and in roundtable discussions, different orthographic practices, as well as many questions about the Croatian language that remain unresearched, poorly researched, or incorrectly described. (Hudeček, Mihaljević, and Vukojević 2011: 89)

As a result, the 1990s saw a rapid growth in the publication of language handbooks of various types, with the goal of creating a "pure" Croatian standard language, which would be indisputably independent from Serbian. The majority of this advice concerned the usage of individual lexical items, suggesting "better" or "purely Croatian" forms to replace those seen as being more typical of Serbian usage. Notably, several so-called "differential dic-

[2] All translations are my own.

[3] Because of the importance of purism in the development of the contemporary standard language, Croatian linguists have written extensively on this topic; see Langston and Peti-Stantić (2014: 175–80) for a discussion. Thomas 1988 provides a detailed study of the development of the Croatian lexicon in the 19th century and the role played by purism; see also Thomas 1991 for a broader investigation of linguistic purism.

[4] The Council for the Norms of the Croatian Standard Language, established in 2005 by the Ministry of Science, Education, and Sports and dissolved in 2012. The collected proceedings of its meetings were published in a special issue of the journal *Jezik* (Vijeće za normu 2013).

tionaries" of Croatian and Serbian were published (Brodnjak 1991 being the most prominent example), but even some traditional monolingual or bilingual dictionaries would identify certain forms as Serbian and refer users to Croatian equivalents (e.g., Bujas 1999).[5] Some of these works providing advice on proper usage have appeared in more than one edition, and their publication has continued up to the present with little abatement.[6]

Although most attention has been focused on the lexicon (including different patterns of word-formation and loanword adaptation in Croatian and Serbian), other grammatical differences that are seen as being distinctively Croatian have also been promoted by language planners. One of the more notable examples is the use of longer variants of certain definite adjectival endings: masculine/neuter genitive singular *-oga* versus *-og*, dative-locative singular *-omu*, *-ome* versus *-om*;[7] and to a much lesser extent, dative-locative-instrumental plural (all genders) *-ima* versus *-im*. The longer endings are often recommended as a stylistic variant for the first in a sequence of adjectives (e.g., *gramatika hrvatsk**oga** književnog jezika* 'the grammar of the Croatian literary language') or when an adjective stands alone without an accompanying noun, but they have also been promoted as being more typical of traditional Croatian usage and stylistically better in general (e.g., Težak 1991: 91–92, Tanocki 1994: 115–16, Frančić, Hudeček, and Mihaljević 2006: 119). Some writers or editors now exhibit a strong preference for the longer endings. For example, in the editor's introduction to issue 15 of *LAHOR*, a journal aimed at teachers of Croatian as a first or second language, apart from two quoted titles only 2 of 91 total instances of masculine/neuter genitive or dative-locative singular adjectives use the shorter endings, and there are no occurrences of shorter genitive singular endings; e.g., *u drug**omu** dijelu (iz)govorn**oga** prostora* 'in the

[5] Other dictionaries consistently refer users to recommended Croatian forms with an arrow or some other symbol, although they may not explicitly label the deprecated forms as "Serbian"; e.g., Šonje 2000, Jojić 2015.

[6] For partial lists of language handbooks published since 1990, see http://ihjj.hr/stranica/jezicni-savjeti/27/ and http://www.hrvatskiplus.org/article.php?id=1795&naslov=vazniji-hrvatski-jezicni-savjetnici (both last accessed 2 July 2018).

[7] Some sources recommend a distinction here, with *-omu* prescribed for the dative, *-ome* for locative, but this does not appear to have any basis in historical patterns of usage (see Tafra 1995). Current usage of these endings is not consistently differentiated according to case, although *-ome* appears to be less common in the dative. For example, in the data cited in Table 9 below, the first 20 occurrences of *novome* in the corpus comprise 3 dative vs. 17 locative contexts, while the first 20 for *novomu* comprise 9 dative and 11 locative contexts. To cite one specific text, Jelaska 2014 (in the journal *LAHOR*, see below) uses both *-omu* and *-ome* throughout the article in the locative case, even in very similar phrases; e.g., *u engleskome govornome području*, *u tome smislu* vs. *na nekomu području*, *u društvenomu smislu*, etc. There is no obvious pattern here for the choice between these two endings.

second part of the articulatory space' (Babić 2013: 2). This is typical for this journal (and the intention here is presumably to model the "best" usage for Croatian teachers), but is strikingly different both from 20th-century patterns of written usage and from the spoken language.

Proposed changes to the spelling of certain forms have been particularly controversial, with the public and scholarly debates on this topic sometimes referred to as "orthographic war (*pravopisni rat*)."[8] As one example of the intensity of opinions on this topic, after the publication of a new orthographic manual by the Institute for the Croatian Language and Linguistics (Institut za hrvatski jezik i jezikoslovlje (IHJJ), see below), the director of the IHJJ filed a lawsuit against Nataša Bašić for an allegedly libelous review published in the journal *Jezik*, and demanded the removal of the journal's editor, Sanda Ham (Ham 2015; Bagdasarov 2015: 155).[9] Political figures have also entered into the controversy, as when Prime Minister Ivo Sanader publicly declared that he would continue to write *neću* 'I will not' as a single word, rather than separately (*ne ću*), despite the recommendation of the Council for the Norms of the Croatian Standard Language (Maretić Žonja 2005), or when the newly appointed minister Željko Jovanović declared a symbolic change of *šport* 'sports' in the name of his ministry back to *sport* shortly after the 2011 elections: "I will be the Minister of Education, Science, and Sports [*sporta*], which I want particularly to emphasize. *Sporta*. Thus, also on this symbolic level great changes are beginning, not just in my ministry but also in the entire government" (Pavić 2012).[10]

Competing orthographic manuals have been published in multiple editions in Croatia over the last few decades, beginning with the republication in 1990 of the 1971 orthographic manual (*pravopis*) by Stjepan Babić, Božidar Finka, and Milan Moguš, which had been banned by the Yugoslav authorities. In addition to subsequent revised editions of this handbook in 1994, 1995, 1996, 2000, 2002, 2003, 2004, and later editions published under the names of

[8] See, for example, Babić and Ham 2004, the first of a series of articles in the journal *Jezik* giving an annotated bibliography of articles in the press about orthographic issues.

[9] Bagdasarov (2015) gives a detailed history of the development and publication of the IHJJ orthographic manual in the context of language policy, from his perspective as another critic of this work.

[10] The form *šport* was originally borrowed into Croatian from German, but was replaced by *sport*, like the form used in English and other languages, after World War II. Croatian language planners have argued for a return to *šport* (e.g., Dulčić 1997: 124–25). The existing *Ministarstvo prosvjete, kulture i sporta* (Ministry of Education, Culture, and Sports) was renamed the *Ministarstvo prosvjete, kulture i športa* under the conservative HDZ (*Hrvatska demokratska zajednica*/Croatian Democratic Union) government in 1992, and the spelling *šport* remained (under different organizations and names of this government ministry) up until the end of 2011.

just Babić and Moguš after Finka's death (Babić and Moguš 2010, 2011), others include a version of the Babić-Finka-Moguš *pravopis* adapted for use in the school system (Babić, Ham, and Moguš 2005, 2008, 2009, 2012); later editions of the *pravopis* by Vladimir Anić and Josip Silić, first published in 1986 (Anić and Silić 1990, 2001); an orthographic manual issued by the cultural organization Matica hrvatska (Badurina, Marković, and Mićanović 2007, 2008); and the previously mentioned *pravopis* published by the IHJJ (Jozić et al. 2013).

Three proposed changes to the existing orthographic norms have been the topic of most discussion and disagreement.[11] One has already been mentioned above, a proposal to spell the negated future auxiliary as two separate words (e.g., *ne ću* '(I) will not' as opposed to *neću*). Another involves the spelling of the reflexes of the Common Slavic vowel **ě*. According to the established norm, the spelling is *ije* when long, *je* when short; e.g., *riječ* 'word' vs. *rječnik* 'dictionary'. Since some speakers are unsure of the quantity of vowels in related word forms such as these, this variation in spelling is a frequent source of mistakes. Furthermore, the spelling of the long reflex of **ě* as *ije* implies a disyllabic pronunciation, which is not the norm in standard Croatian or most local štokavian varieties spoken by Croats; instead, the normal pronunciation is a diphthong in both instances. However, rather than trying to address these more fundamental problems, language planners have focused their attention on the rule for spelling the short reflex of **ě* after a cluster of consonant + *r*, which according to the established norm omits the *j*, corresponding to the most common pronunciation of these forms; e.g., *grijeh* 'sin', *grešnik* 'sinner' (cf. *riječ*, *rječnik* cited above). Various proposals have been made to introduce *j* in the spelling of these forms, but they all allow variants and exceptions, which serve only to complicate rather than simplify the situation. For example, the 2010 edition of the Babić-Moguš *pravopis* requires *j* in some forms, such as *pogrješka* 'mistake'; allows variants in others, such as *brjegovi/bregovi* 'hills'; and prescribes no *j* in forms like *vrijeme* 'time', GEN.SG *vremena*, ADJ *vremenski*).

The final change involves the spelling of dental stops before the affricates *c* [ts], *č* [tʃ]. According to the so-called "phonetic" spelling principles that derive from the work of the Serbian language reformer Vuk Karadžić in the 19th century, and which were the norm throughout most of the 20th century, phonological processes of voicing assimilation and the reduction of consonant clusters are reflected in the spelling; e.g., *sudac* 'judge', NOM.PL /sudts-i/ → [suttsi] → [sutsi], spelled *suci*. According to proposed changes, retention of the dental stop should be the norm in forms with so-called "mobile *a*" (e.g.,

[11] More radical changes, such as writing both the long and short reflexes of Common Slavic **ě* as *ie* or returning to a morphophonological spelling system, similar to spelling practices in the 19th century, have been advocated by some authors (e.g., László 1994, who used his own spelling system in many publications; cf. also the preface to Šimundić 1994), but have not been promoted in any of the mainstream orthographic handbooks.

dodatak 'addition', NOM.PL *dodatci*; *očevidac* 'eye-witness', NOM.PL *očevidci;* NOM.PL *sudci*; with the exception of *otac* 'father', GEN.SG *oca,* etc.). Although this principle arguably makes the relationship between different inflected forms of the same word more apparent, it introduces new inconsistencies since spellings with *d* apply only to the position before a voiceless affricate, and not before a voiceless stop. As a result, the prescribed spelling for forms like *predak* 'ancestor' would be with *t* in the oblique singular forms, with voicing assimilation indicated orthographically (GEN-ACC.SG *pretka,* DAT-LOC.SG *pretku*, INS.SG *pretkom*), but with *d* in the vocative singular and the plural (VOC.SG *predče,* NOM.PL *predci,* DAT-INS-LOC.PL *predcima*).

Apart from language handbooks and discussions of proper usage in the media, efforts to restandardize Croatian have to some extent been supported by governmental policy. This is seen mainly in the adoption of official terminology in areas directly controlled by the government; e.g., the replacement of *pasoš* 'passport' by *putovnica, oficir* 'officer' by *časnik*, etc. In many cases, these changes represent the replacement of terms in official use during the Yugoslav period by political, military, and legal terminology that was used in Croatia during the 19th century. Official language policies can also be seen in the actions of government agencies to approve certain handbooks for use in the schools or in requirements for the evaluation of textbooks, in a law requiring the names of businesses to be in Croatian (which does not appear to be strictly enforced), and so forth (see Langston and Peti-Stantić 2014: 130–46 for more information). However, calls for a comprehensive law on language, which would broadly mandate the use of specific forms or spellings in all public or official contexts, have been rejected.

The participants in public debates on language policy since 1990 can be broadly characterized as falling into one of two camps. There are those who are advocates for purism and for a recodification of the Croatian language, who see the usage of distinctively Croatian forms as an expression of Croatian national identity and patriotism. They typically feel that the government is not doing enough to regulate language usage, by implementing policies that would conform to their views. On the opposing side are those who believe that any far-reaching attempts to make changes to the norm, especially if implemented in a rigid, top-down manner, would only serve to destabilize the standard language. Although the heated nature of discussions on this topic has to some extent subsided since the politically turbulent period of the early 1990s, these questions continue to be debated two decades later. For example, in a newspaper article written in connection with the disbanding of the Council for the Norms of the Croatian Language and other actions by the Milanović government in 2012, we read:

> Twenty years later, the story about language is still in the repertoire, but it's pretty clear that the language reform of the 90s didn't have

> the desired success. On the contrary, linguists today contend that the changes that were insisted on resulted in a fear of using the language among ordinary citizens, or as the prominent linguist Nives Opačić explained in her book *Croatian language guideposts*, what we got were 'uncertain, stammering, tongue-tied, and miserable speakers of their native – Croatian – language,' and instead of a means of communication, language has become a means of political qualification and disqualification. (Piteša and Kalogjera 2012)

3. Previous Corpus-Based Research on the Effects of Language Planning

Several researchers have previously used corpus data to attempt to quantify the effects of Croatian language planning efforts.[12] They focus on pairs of words that are synonymous or nearly synonymous, where one is viewed by purists as "foreign", "Serbian", or otherwise undesirable in some way, and the other is the recommended Croatian form. However, there are limitations to this work due to the size and composition of the corpora that were used. Grčević (2001) presents data on lexical usage from the Mannheim Croatian Corpus, consisting at the time of this research of approximately 14 million words from several (mainly right-leaning) newspapers from 1997–99, but the corpus itself is not publicly available. Rittgasser (2003) describes changes in patterns of lexical usage based on his own corpus of more than 5 million words from media texts, but does not provide the actual frequency data. Both of these sources conclude that the changes in usage in the 1990s were not as dramatic as they had sometimes been impressionistically described, and they examine factors influencing the usage of specific lexical items. Czerwiński's (2005) corpus is based on smaller samples from five media sources dating from January 2002, ranging from right to left on the political spectrum. He finds that there are notable differences in usage between the more conservative/nationalist media and those with a more liberal political orientation. More recently, Schou Madsen (2017) compares the adaptation and usage of loanwords and loan translations in Croatian and Serbian, based on data from the Croatian National Corpus (*Hrvatski nacionalni korpus*, version 2.5, containing approximately 100 million words primarily from texts published between 1990 and 2005), and the Corpus of Contemporary Serbian (*Korpus savremenog srpskog jezika*, 2013, containing approximately 122 million words primarily

[12] Some research has also investigated attitudes towards language planning; e.g., Jahn 1999, Langston and Peti-Stantić 2014: 242–45. As indicated by Jahn (1999), and also suggested by an anonymous reviewer, attitudes towards language planning/language purism may vary in different regions of the country; however, I am not aware of any comprehensive study of this question.

from texts published between 1990 and 2013; see Schou Madsen 2017: 46–53 for a description of these sources).[13]

In Langston and Peti-Stantić 2014 we analyze patterns of usage based mainly on data from subcorpora contained in version 2.0 of the Croatian National Corpus,[14] supplemented with additional data from corpora compiled specially for this research (see Langston and Peti-Stantić 2014: 259). These corpora contain texts from various media outlets, dating from approximately the same period as the corpora discussed above (1997-2005). Earlier usage is represented by frequency data in Moguš, Bratanić, and Tadić (1999) and another frequency dictionary based on a smaller sample of texts from the newspapers *Vjesnik* and *Večernji list* from 1980 (Šojat 1983). For some pairs of words, we also found differing patterns of usage based on the political orientation of the sources, with the more liberal media being closer to earlier patterns of usage reflected in these frequency dictionaries. However, usage for other lexical items varied: some recommended forms were widely used in all sources, while others had not been broadly adopted. In addition to the traditional media sources, we also compared usage in a one-million word corpus of Croatian blog texts (from www.blog.hr, 2004–05), which represents more casual, unedited language. Here usage was generally found to be closer to that attested in Moguš, Bratanić, and Tadić (1999), which indicates some resistance to language planning efforts; however, at least in some instances it is also possible to discern a trend in the blog corpus towards increased usage of "purely Croatian" forms that have been promoted since the 1990s (Langston and Peti-Stantić 2014: 268).

4. Data from Contemporary Web Corpora

The present work analyzes data from much larger web corpora that have recently become available: the Croatian Web Corpus (HrWaC, v. 2.2; approximately 1.2 billion words) and the Serbian Web Corpus (SrWaC, v. 1.2; approximately 480 million words; Ljubešić and Klubička 2014). These sources should be more representative of contemporary usage than the corpora used in previous research, and can help us determine the effects of Croatian language planning efforts today.[15] The Serbian Web Corpus serves as a comparable ref-

[13] See also Thomas 1978 for an earlier study comparing usage in newspapers from Zagreb, Belgrade, and Sarajevo.

[14] Version 3.0 and earlier versions are now available at http://filip.ffzg.hr/cgi-bin/run.cgi/first_form?corpname=HNK_v30;align= (last accessed 3 July 2018).

[15] Besides containing a much wider variety of texts, HrWaC consists largely of texts dating after the period 1997–2005 covered in the corpora used in previous research (the texts in this corpus were compiled in two separate crawls of the top-level .hr domain in 2011 and 2013). However, the exact date of many web texts cannot be de-

erence point to confirm differences in Croatian and Serbian patterns of usage. For earlier Croatian usage, we must still rely on the *Hrvatski čestotni rječnik* (*Croatian frequency dictionary*, henceforth HČR) by Moguš, Bratanić, and Tadić (1999). HČR is based on a one million-word corpus, evenly divided between drama, prose, poetry, textbooks, and newspaper texts. The literary texts in this corpus date from about 1930–75; the textbook portion consists of samples from 58 textbooks on various topics that were in use in Croatia in the mid 1970s, and the newspaper texts are taken from editions of *Vjesnik, Slobodna Dalmacija, Novi list, Glas Slavonije*, and the Zagreb edition of *Borba* from 1975 and 1977 (see Moguš, Bratanić, and Tadić 1999: 6–10).

When comparing different corpora, the goal is generally to determine whether a given element (e.g., a word, a collocation, a grammatical construction) occurs significantly more frequently in one corpus than in another. If we find significant differences, this allows us to draw the conclusion that the corpora represent samples from different populations (in some sense), with different patterns of language use. However, basic statistical tests that have often been used in earlier corpus research (such as the chi-squared test), which are intended to determine whether or not we can reject the null hypothesis (in this case, that there is no difference between two corpora) with a sufficiently high degree of certainty, have been shown to be inadequate. With such tests, the null hypothesis is based on the assumption of randomness, but language is not random; therefore, "when we look at linguistic phenomena in corpora, the null hypothesis will never be true" (Kilgarriff 2005: 263). Especially when large corpora are involved, even a comparison of word frequencies between subparts of the *same* corpus typically results in findings of statistically significant differences. For example, the results of a chi-squared test indicate that the frequencies of the most common 100 words are significantly different in the 2011 and 2013 subcorpora of HrWaC ($\chi^2 = 3189100$, df = 99, $p < .001$). However, there is no reason for us to suspect that there is any meaningful difference in the usage of high-frequency words such as *biti* 'to be', *i* 'and', *u* 'in', etc. in these two samples.

As a result, linguists disagree about the utility of significance testing when comparing different corpora; at the very least, information about effect sizes and the dispersion of the data within the corpus should be taken into account. A variety of more sophisticated techniques for hypothesis testing have been proposed, but no single approach is universally accepted (see Gries 2005, Lijffijt et al. 2016). Given the very different nature of the corpus data here, particularly the limited information provided by the frequency dictionary representing earlier usage, we will focus instead on descriptive statis-

termined (e.g., for websites that are only periodically updated), and media, internet forum, and blog sites may also include earlier texts, to the extent that they maintain archived postings. The texts in SrWaC were collected in 2013.

tics and examples of the dispersion of forms within the corpora, to the extent that this information is available and relevant. Such data are still sufficient to determine overall trends in patterns of usage, especially when we find large differences over time.

4.1. Lexical Pairs

The analysis here will focus on the usage of selected pairs of synonymous or nearly synonymous words, one of which is considered more typical of traditional Croatian usage, and the other of which is viewed by purists as "non-Croatian" in some sense, usually a loanword or a word that is felt to be more typical of Serbian (see the Appendix for a complete list). While there is no single, unambiguous criterion for selection here, these are words that have been included in various handbooks or discussed in language advice features in the media, and which have also been selected for study in the earlier research described above. They can be considered representative examples of lexical items that have been the focus of language planning efforts, although they do not represent an exhaustive list of such forms. The forms studied here can be categorized into a number of broad groups, based on their patterns of usage. For purposes of comparison, I also include representative pairs that have long been cited as examples of differences between Croatian and Serbian and whose differing spheres of usage have remained relatively stable. These forms can serve as a gauge of the representativeness of the corpora used in this research.

In order to compare the frequency of different lexical items, given the different sizes of the corpora used here, we will focus on the relative frequency of each member of a lexical pair, expressed as a percentage of the total number of occurrences of both lexical items. This assumes that both members of a pair are equally acceptable in all contexts, which is not true except in the case of complete synonymy, and also ignores the possibility of stylistic differentiation. However, this is a necessary oversimplification for purposes of comparison, since it is not possible to examine all of the individual contexts in which these lexical items occur. Raw frequency counts are also given for all forms, and for nouns referring to people these generally represent totals including feminine derivatives,[16] although only the masculine form is cited for reasons of space (i.e., the figures given for a form like *sekretar* 'secretary' include counts for *sekretarica*/*sekretarka*). In all instances, the form that is considered by some

[16] Unless there is no regular feminine correspondent. For agent nouns in *-lac* (Table 6), there is no regular derivational process for creating feminine forms, so these are compared only with masculine forms in *-telj*, and not the regularly derived feminine forms in *-teljica*.

authorities to be “foreign” or not characteristic of “good”/“cultivated” Croatian usage is listed first.

4.1.1. Stable Differences in Usage between Croatian and Serbian

Despite the close relationship between Croatian and Serbian, data such as those in Tables 1 (opposite page) and 2 (on page 179) show that there are clear differences in lexical usage, as is already well known. Many forms that are considered typical of Serbian are not attested at all in the corpus used for HČR. The data from the large web corpora used here almost always include some attestations of both forms, but when the total number of attestations is small, they can usually be explained in some way. For example, the occurrences of *bioskop* ‘movie theater’ in HrWaC are almost exclusively in reference to Serbia; e.g., several such instances are from a 2010 article with the dateline Beograd, titled “Zatvara se bioskop Balkan (the movie theater Balkan is closing)”.[17] Other occurrences of *bioskop* are largely in blogs hosted in the .hr top-level domain, but at least some of which are presumably written by ethnic Serbs, since the texts use ekavian forms.[18] Because of the way the web corpora were compiled, by automatically crawling sites within the top-level Croatian and Serbian web domains, and the possibility of direct quotations or other references to original Croatian or Serbian sources, there are inevitably some data in each corpus that cannot be considered typical of Croatian or Serbian usage, respectively. In other instances both forms are used to a significant degree in a given language, but sometimes in different contexts (e.g., although *vaspitanje* and *odgoj* are characteristic of Serbian and Croatian usage, respectively, in reference to the upbringing or education of children, *odgoj* occurs in SrWaC in the meaning ‘raising/breeding livestock’ [*odgoj krava, kokoši* ‘raising cows, chickens’, etc.];[19] *vaspitanje* would not be normal in Serbian usage here, while Croatian would use *uzgoj* in this context). Therefore, despite some expected variation in usage and a certain residue of forms that can be attributed to other factors, both the web corpora used here and HČR conform to numerous earlier descriptions of differences between Croatian and Serbian; these

[17] http://banke.com.hr/zatvara-se-bioskop-balkan/ (last accessed 3 July 2018).

[18] Standard Croatian is ijekavian, so called because of the reflexes (*ije ~ je*) of Common Slavic **ě*. Standard Serbian is predominantly ekavian, with the reflex *e*; although ijekavian variants are also considered standard, such forms are used almost exclusively by Serbs living in areas outside of Serbia. Forms in the following tables are cited in their ijekavian (Croatian) spellings. The data from SrWaC include counts for both ekavian and ijekavian spellings.

[19] The form *odgoj* also occurs in reference to children in SrWaC, but *vaspitanje* is clearly preferred.

Table 1. (Near-)categorical usage of different lexical items in Croatian and Serbian, no substantial change in Croatian from the frequency distribution in HČR

	HČR		**HrWaC**		**SrWaC**	
	N	%	N	%	N	%
bioskop	0	0.0	559	0.8	14036	87.6
kino	24	100.0	69061	99.2	1983	12.4
‘cinema’						
fabrika	8	5.6	4005	4.4	62913	98.5
tvornica	135	94.4	86370	95.6	986	1.5
‘factory’						
ostrvo	6	5.5	2087	1.1	35959	86.3
otok	103	94.5	182917	98.9	5699	13.7
‘island’						
pantalone	0	0.0	122	0.5	5476	94.9
hlače	36	100.0	25669	99.5	293	5.1
‘trousers’						
pozorište	0	0.0	3207	3.1	45402	97.8
kazalište	128	100.0	101223	96.9	1024	2.2
‘theater’						
supa	0	0.0	574	2.2	5207	95.3
juha	17	100.0	25158	97.8	258	4.7
‘soup’						
uslov	12	2.9	1644	0.5	209458	98.6
uvjet	409	97.1	344230	99.5	2933	1.4
‘condition’						
vaspitanje	0	0.0	392	0.6	16100	91.2
odgoj	81	100.0	70349	99.4	1548	8.8
‘upbringing, education’						

Table 2. Phonological and morphological differences between Croatian and Serbian, no substantial change in Croatian from the frequency distribution in HČR

	HČR		HrWaC		SrWaC	
	N	%	N	%	N	%
demokratija	1	4.5	3027	4.0	38009	97.2
demokracija	21	95.5	71732	96.0	1078	2.8
‘democracy’						
hemija	0	0.0	428	1.9	9753	99.1
kemija	28	100.0	22435	98.1	88	0.9
‘chemistry’						
informisati (se)	0	0.0	324	0.8	14708	98.0
informirati (se)	20	100.0	39281	99.2	295	2.0
‘to inform (be informed)’						
organizovati (se)	0	0.0	1555	0.7	120705	98.2
organizirati (se)	92	100.0	227848	99.3	2241	1.8
‘to organize (be organized)’						
savremen	0	0.0	2283	1.9	80913	98.8
suvremen	159	100.0	120441	98.1	964	1.2
‘contemporary’						
srećan	11	2.5	1802	1.0	41028	88.4
sretan	427	97.5	184324	99.0	5374	11.6
‘happy’						

corpora are representative of what is known about traditional patterns of usage in these two language varieties.

4.1.2. Stable Variation in Croatian Usage

In other pairs of words, we see a pattern of mixed usage in Croatian, but with no striking change in the relative frequency of these forms between the periods covered by HČR and HrWaC. In the data in Table 3 on the opposing page, the second member of each pair in most cases is clearly not characteristic of

Table 3. Mixed usage in both HČR and HrWaC, with no major change

	HČR		HrWaC		SrWaC	
	N	%	N	%	N	%
greška	46	52.3	91240	58.1	60111	97.9
pogreška	42	47.7	65692	41.9	1296	2.1
'mistake'[20]						
grupa	200	69.7	305620	51.3	256657	98.5
skupina	87	30.3	290576	48.7	3989	1.5
'group'						
originalan	20	55.6	63889	60.0	25359	69.4
izvoran	16	44.4	42545	40.0	11163	30.6
'original'						
penzija	9	25.7	14696	13.7	40822	98.9
mirovina	26	74.3	92465	86.3	436	1.1
'pension'						
period	40	23.4	85158	28.6	173257	95.3
razdoblje	131	76.6	212250	71.4	8465	4.7
'period'						
porijeklo	20	66.7	34181	48.7	40658	99.1
podrijetlo	10	33.3	36011	51.3	359	0.9
'origin'						
sport	15	88.2	160282	87.5	82972	99.9
šport	2	11.8	22915	12.5	97	0.1
'sport(s)'						

1992, Protuđer 1998 for other forms listed here). Although there appears to be some shift in HrWaC towards recommended Croatian forms in some instances, the relative frequency is not dramatically different (defined here as a change in frequency of less than 20%) from HČR.

4.1.3. Substantial Shifts in Usage towards Recommended Croatian Forms

A comparison between the frequency data in HČR and HrWaC for the forms in Table 4 on the following page indicates a more substantial shift in usage

Table 4. Mixed usage in HČR, substantial shift towards recommended forms in HrWaC

	HČR		**HrWaC**		**SrWaC**	
	N	%	N	%	N	%
advokat	27	87.1	2745	2.8	15511	97.1
odvjetnik	4	12.9	95887	97.2	460	2.9
‘lawyer’						
budžet	26	47.3	21198	13.0	89566	94.3
proračun	29	52.7	142389	87.0	5415	5.7
‘budget’						
firma	7	43.8	110750	17.8	148968	98.0
tvrtka	9	56.3	512450	82.2	3086	2.0
‘firm’						
muzika	78	40.8	43350	16.4	78848	98.7
glazba	113	59.2	221757	83.6	1026	1.3
‘music’						
nivo	59	51.8	68018	17.1	195932	98.4
razina	55	48.2	329644	82.9	3167	1.6
‘level’						
sekretar	111	82.2	7642	8.2	44440	98.8
tajnik	24	17.8	85901	91.8	531	1.2
‘secretary’						
sistem	303	68.4	102841	14.7	349517	98.9
sustav	140	31.6	597977	85.3	3767	1.1
‘system’						

Serbian usage, and these forms have often been recommended by Croatian purists for this very reason. For example, *podrijetlo* is originally a regional variant ultimately derived from the same root as *porijeklo*, but is seen as preferable by some because it is exclusively Croatian: “But this form [*podrijetlo*] is *ours alone*. Now when we can freely choose, when we can weigh what better suits us, when from the riches of the Croatian lexical stock we can choose for the norm that which is only ours and thus create differences from our eastern neighbors—let us choose what is ours.” (Dulčić 1997: 123; see also Krmpotić towards the recommended Croatian forms in these pairs of lexical items. The fact that the second member of each pair was already in use earlier in the 20th century (as shown by attestations in HČR) no doubt facilitated its increase in frequency, but this does not seem to be the case with similar forms in Table 3 above, even though they have also been promoted by purists as the "better", "more authentically Croatian" variants. This indicates that the outcome of such language planning efforts is very much dependent on the specific lexical items in question. For example, the pairs *originalan/izvoran* 'original' in Table 3 and *nivo/razina* 'level' in Table 4 both consist of a foreign borrowing and a native Slavic form and had similar relative frequencies in HČR, but only *razina* and not *izvoran* exhibits a large increase in usage. A number of factors may be at play here; *originalan* is of classical origin, while *nivo* is a more recent borrowing from French and is not fully adapted to Croatian; its declensional pattern (GEN.SG *nivoa*, etc.) marks it as foreign. In addition, at least according to some sources, *izvoran* does not capture all the possible meanings of *originalan* and speakers may feel that it cannot be substituted in all contexts.[20] Still, changes in usage may be facilitated when the dispreferred form is recognizably foreign in origin, since speakers will have no doubt about which form is recommended. When both forms are Slavic in origin and both have a history of usage in Croatian (e.g., *greška/pogreška* 'mistake' in Table 3), there is no obvious reason for a contemporary speaker to identify one form as somehow being the more authentically Croatian variant.

For the pairs of words in Table 4, although the relative frequency of the recommended Croatian forms has increased, their usage is not necessarily evenly distributed throughout HrWaC. For example, as shown in Figures 1 and 2 on p. 174, *muzika* 'music' is relatively more frequent, and the synonym *glazba* correspondingly less frequent, in the same positions in the corpus, which consist primarily of texts from blogs and discussions on forum.hr. This suggests that *muzika* is still preferred in more informal contexts. (There is a large spike in frequency for both forms in the portion of the corpus containing texts from the website muzika.hr.) Similarly, although *budžet* 'budget' is fairly evenly distributed throughout HrWaC, *proračun* 'budget' is markedly less frequent in the portions containing blogs and forum discussions. This is in keeping with Ritgasser's (2003: 9) observation that *proračun* is more frequent in official contexts (e.g., *državni proračun* 'state budget'), while *budžet* tends to be used in everyday contexts (e.g., *obiteljski/kućni budžet* 'family/household budget').

[20] Bujas 1999 and the Hrvatski jezični portal (hjp.znanje.hr, last accessed 3 July 2018) give only meanings of the type 'first or earliest; authentic' for *izvoran*, while *originalan* has the additional meanings of 'special, unique; unusual'. However, Šonje 2000 gives both sets of meanings for *izvoran*, and purists have argued that the meanings of these two words are identical and the Croatian form should be preferred over the borrowing (e.g., Dulčić 1997: 182).

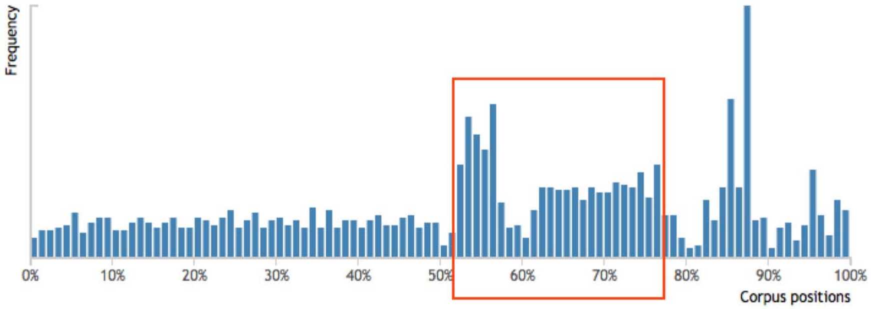

Figure 1. Frequency distribution of *muzika* 'music' in HrWaC

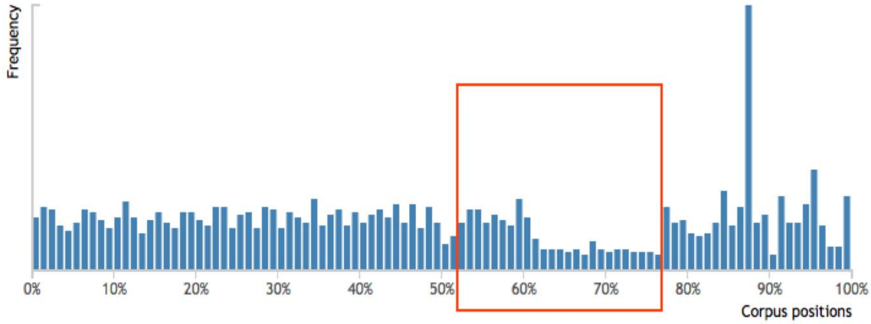

Figure 2. Frequency distribution of *glazba* 'music' in HrWaC

In the examples in Table 5 on p. 176 we see an even more dramatic shift towards the usage of forms that were previously uncommon, although probably not completely unfamiliar to most Croatian speakers. The first member of each pair was a normal, unmarked form widely used in a variety of contexts throughout most of the 20th century (*avion*, *izvještaj*, *pažnja*, *raskršće* are attested in all subcorpora in HČR; *opozicija* and *upotreba* are absent only from the poetry subcorpus; and the remainder are attested in at least three of the five types of texts), while the second member of each pair is either unattested or infrequent in HČR. However, the increased usage of many of the recommended Croatian forms in HrWaC seen in the data here seems to be concentrated in news reports or other "official" sources. For example, *avion* and *zrakoplov* 'airplane' are now almost equally frequent in HrWaC as a whole, but while *avion* is reasonably well represented throughout the corpus (Fig. 3 on page 177), usage of *zrakoplov* is markedly lower in the highlighted portion of the corpus, which consists almost entirely of posts from forum.hr (Fig. 4 on page 177). The portion of the corpus with the lowest usage of *zrakoplov* contains posts from a different internet forum (cafe.mobil.hr) and a blog site (mojblog.hr). Conversely, the two highest spikes for the usage of *avion* are in portions of HrWaC consisting of texts from the tabloid *24sata* and forum.hr.

Similar patterns are seen for pairs such as *pažnja/pozornost* 'attention' (Figs. 5 and 6 on p. 178) or *upotreba/uporaba* 'use'. Other pairs of words in Table 5, such as *izvještaj/izvješće* 'report' are both more likely to be used in official contexts rather than in everyday situations.

We also find a similar shift in frequency with occurrences of two competing agentive suffixes, -*lac* and -*telj*, the former of which has been branded as more typical of Serbian usage. As shown by examples in Table 6 on p. 179, -*lac* was previously common in Croatian usage (and remains standard for certain forms; e.g., *ronilac* 'diver'), but -*telj* has been promoted as more typical of Croatian (Babić 1995: 140–41). One particular phonological change in the adaptation of a word of foreign origin, from the previous *Evropa, evropski* 'Europe, European' to *Europa, europski,* has been almost universally adopted in the texts in HrWaC; this particular form has apparently acquired an especially symbolic value as a marker of "Croatian-ness".

4.1.4. Recommended Croatian Forms That Have Not Been Widely Adopted

Finally, there are a number of forms that have been promoted by language planners but which have gained little, if any, traction in usage; see Table 7 on p. 179. Although discussions of Croatian language planning efforts in the 1990s often referred to the creation of "new" words, most of the actual changes in usage involve forms that were already employed to some significant degree in the 20th century or that were at least familiar to many speakers. True neologisms that have been proposed as replacements for established

Table 5. (Near-)categorical usage in HČR, substantial shift towards the recommended Croatian form in HrWaC

	HČR		HrWaC		SrWaC	
	N	%	N	%	N	%
ambasador	21	100.0	10710	26.3	29324	99.4
veleposlanik	0	0.0	30038	73.7	170	0.6
‘ambassador’						
avion	100	95.2	69065	53.1	42504	99.3
zrakoplov	5	4.8	60939	46.9	298	0.7
‘airplane’						
delegacija	154	100.0	18680	52.9	24733	98.8
izaslanstvo	0	0.0	16642	47.1	310	1.2
‘delegation’						
direktor	101	99.0	209660	62.1	180490	99.8
ravnatelj	1	1.0	128067	37.9	420	0.2
‘director’						
izvještaj	55	100.0	61035	36.9	89793	99.1
izvješće	0	0.0	104282	63.1	785	0.9
‘report’						
oficir	68	90.7	6837	24.6	11428	97.7
časnik	7	9.3	20998	75.4	271	2.3
‘officer’						
opozicija	26	92.9	16193	33.8	33982	99.6
oporba	2	7.1	31783	66.2	138	0.4
‘opposition’						
pažnja	121	95.3	115922	62.3	98681	99.3
pozornost	6	4.7	70055	37.7	718	0.7
‘attention’						
raskršće	43	97.7	3115	20.1	1667	96.4
raskrižje	1	2.3	12380	79.9	63	3.6
‘intersection’						
upotreba	128	100.0	71450	51.3	78529	99.3
uporaba	0	0.0	67950	48.7	531	0.7
‘use’						

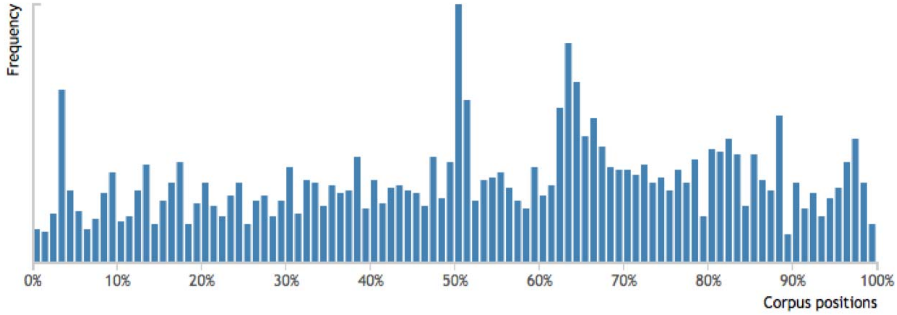

Figure 3. Frequency distribution of *avion* 'airplane' in HrWaC

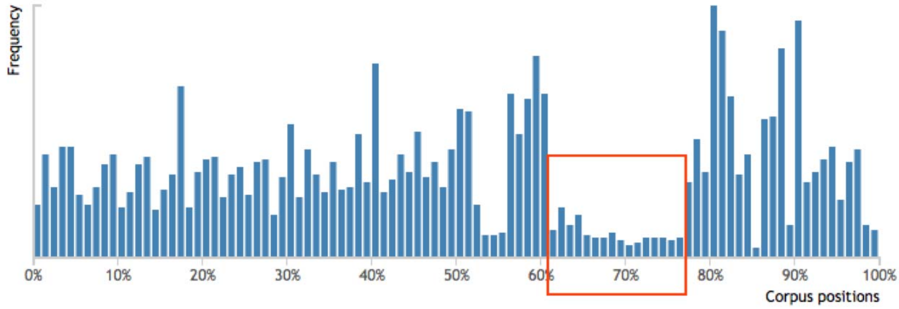

Figure 4. Frequency distribution of *zrakoplov* 'airplane' in HrWaC

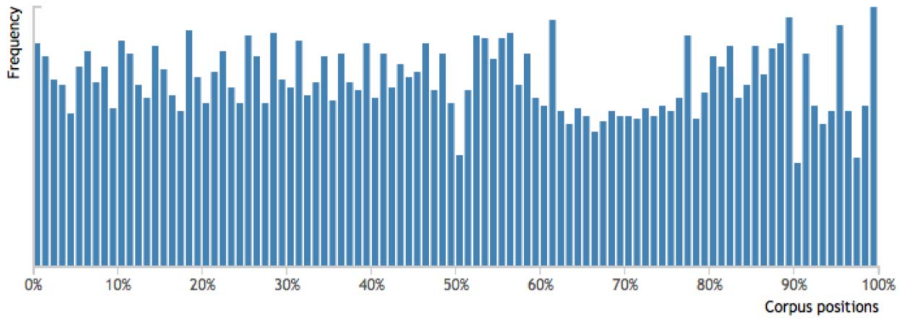

Figure 5. Distribution of *pažnja* 'attention' in HrWaC

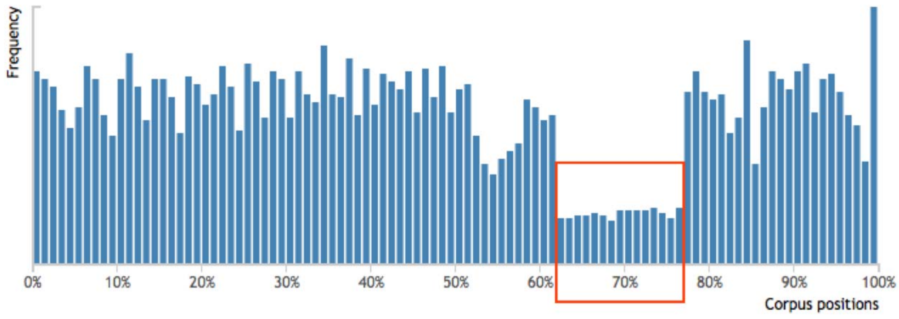

Figure 6. Distribution of *pozornost* 'attention' in HrWaC

Table 6. (Near-)categorical usage in HČR, substantial shift towards the recommended Croatian form in HrWaC: morphological and phonological features

	HČR		HrWaC		SrWaC	
	N	%	N	%	N	%
čitalac	24	88.9	1084	1.7	34259	98.0
čitatelj	3	11.1	64026	98.3	694	2.0
‘reader’						
gledalac	48	100.0	1035	1.3	10332	97.5
gledatelj	0	0.0	77248	98.7	261	2.5
‘viewer’						
posjetilac	24	100.0	4717	4.5	43734	98.7
posjetitelj	0	0.0	99021	95.5	555	1.3
‘visitor’						
evropski	106	100.0	20789	3.7	285757	98.4
europski	0	0.0	538114	96.3	4574	1.6
‘European’						

Table 7. (Near-)categorical usage in HČR, no substantial change towards the recommended form in HrWaC

	HČR		HrWaC		SrWaC	
	N	%	N	%	N	%
analiza	107	100.0	128796	98.7	74872	100.0
raščlamba	0	0.0	1669	1.3	10	< 0.1
‘analysis’						
civilizacija	24	96.0	36323	97.0	17845	99.9
uljudba	1	4.0	1133	3.0	23	0.1
‘civilization’						
datum	16	100.0	78813	96.8	40815	100.0
nadnevak	0	0.0	2641	3.2	5	< 0.1
‘date’						
kandidat	49	100.0	166675	98.4	71009	100.0
pristupnik	0	0.0	2755	1.6	11	< 0.1
‘candidate’						

loanwords have generally not been accepted. Forms such as *očvrsje* for *hardver/hardware, mekušje* or *napudbina* for *softver/software* occur only a handful of times in HrWaC, although others, such as *oporaba* (0.6 per million words) for *reciklaža* 'recycling', *uspornik* (0.2 per million words) for *ležeći policajac* 'speed bump', are marginally better attested, and some such as *preglednik* 'web browser' have been more widely adopted (12,124 occurrences in HrWaC, versus 6,111 for *browser/brauzer*).

4.2. Orthographic and Other Changes

Despite all of the discussion and controversy surrounding certain proposed changes to the existing orthographic rules (see section 2 above), it should be noted that they affect only a limited number of forms in the language, and even in sources that consistently follow the new spelling rules, one often has to search for some time to find any examples. Data from HrWaC show that these changes are not broadly attested in texts online, although there is variation among individual forms that are subject to the same rules. Regarding the retention of *t, d* before *c, č* (e.g., *podatak* 'piece of data', NOM.PL *podaci* [established spelling]/*podatci* [recommended spelling in some recent sources]), for the more common words listed as examples of this rule in the IHJJ orthographic manual (pravopis.hr), the newer spelling rarely occurs more than 10% of the time (Table 8, a. opposite). For the spelling of *j* after sequences of consonant + *r*, usage varies (Table 8, b.). The most frequent word affected by this change is *(po)greška* [established spelling]/*(po)grješka* [recommended spelling in some recent sources] 'mistake', where the spelling with *j* is uncommon. However, the innovative spelling *grješan* 'sinful' is somewhat more frequent relative to *grešan*, even though this word is based on the same root as *(po)gr(j)eška*, and for the imperfective of the verb *spriječiti* 'to prevent', the spelling with *j* (*sprječavati*) is slightly more common than without (*sprečavati*). Other forms, where the length of the vowel varies in inflection, are too infrequent to allow us to draw any firm conclusions. For example, for *crijep* 'tile', we find the following spellings for the nominative plural form: *crepovi* (36 occurrences), according to the established norm; *crjepovi* (138), which conforms to the newer rules; but the most common spelling is *crijepovi* (156), which is prescriptively incorrect. For *brijeg* 'hill', the results are different; the most common spelling of the plural follows the established norm (*bregovi*, 168 occurrences); *brjegovi* is attested 45 times and *brijegovi* 37. These latter examples may indicate that attested spellings with Crj may represent confusion about spellings with *ije* ~ *je/e* in related forms, rather than the adoption of proposed changes to the orthographic norm. Finally, there appears to be strong resistance against the proposed spelling of the negated future auxiliary as two separate words, based on the data in HrWaC (Table 8, c.).

Table 8. Adoption of proposed orthographic changes

	HrWaC	
	N	%
a. retention of *t*, *d* before *c*:		
ostaci	14463	89.1
ostatci	1771	10.9
‘remnants’		
podaci	70205	93.6
podatci	4808	6.4
‘data’		
preci	7447	95.7
predci	334	4.3
‘ancestors’		
suci	24562	95.6
sudci	1131	4.4
‘judges’		
zadaci	7499	91.6
zadatci	686	8.4
‘tasks, assignments’		
b. C*rj* sequences		
grešan	4837	81.5
grješan	1100	18.5
‘sinful’		
(po)greška	154389	98.3
(po)grješka	2618	1.7
‘mistake’		
sprečavati	13835	47.5
sprječavati	15322	52.5
‘prevent’		
unapređivati	3519	69.7
unaprjeđivati	1530	30.3
‘advance’		
c. negated future auxiliary		
neće	684040	98.8
ne će	8466	1.2
‘will not (AUX.3SG)’		

As indicated earlier, some publications appear to have an editorial policy favoring longer definite adjectival endings, and these have been recommended as a feature of refined Croatian usage. If we compare forms for two common adjectives in Table 9, we see that the longer forms are clearly more characteristic of Croatian rather than Serbian usage, in accordance with previous descriptions. The longer endings are also clearly more frequent for the genitive rather than dative-locative forms. But the corpus data do not indicate any strong trend towards the general usage of the longer forms.

Table 9. Use of longer adjective endings

	HrWaC		**SrWaC**	
	N	%	N	%
novog	196421	89.4	118927	99.0
novoga	23206	10.6	1167	1.0
novom	128652	96.0	102597	99.6
novome	5185	3.9	403	0.4
novomu	178	0.1	1	< 0.1
‘new’				
starog	46899	84.6	20803	96.5
staroga	8536	15.4	753	3.5
starom	33446	95.0	14537	99.0
starome	1665	4.7	145	1.0
staromu	80	0.2	6	< 0.1
‘old’				

5. Conclusion

The data analyzed here suggest that language planning efforts in Croatia since 1990 have had some effect on patterns of usage, but that the results are uneven, varying for different items. A number of the forms examined here show a marked increase in frequency in HrWaC, compared to patterns of earlier usage attested in HČR, but other forms have not been widely adopted. Previous research, as well as some of the distributional data available in HrWaC, indicate that there is also a component of stylistic differentiation, with recommended "purely Croatian" forms preferred in more formal or official contexts, where they may be seen as a mark of refined or cultivated usage, and other forms retaining their hold in more casual contexts. An additional factor is that the use of particular forms has been imbued with a political meaning: even seemingly trivial differences, such as the use of *europski* instead of *evropski* or the masculine-neuter genitive singular adjective ending *-oga* instead of *-og* are viewed by some as symbolic markers of Croatian national identity. However, proposed changes to established orthographic practices have been more difficult to implement.

The list of forms examined here is selective, focusing primarily on more frequent words that are attested in HČR and that have received some particular attention in discussions of Croatian language planning and purism. There are many other forms that have been proposed as replacements for foreign borrowings that have largely been ignored and are not widely used; e.g., *raspit* for *anketa* 'survey, poll' (6 versus 55,003 occurrences in HrWaC, respectively) or *slikokaz* for *kino* 'cinema' (72 vs. 68,555 occurrences in HrWaC, and it should be noted that in these few attestations *slikokaz* is primarily used in other meanings, such as 'presentation' or 'slideshow').[21] Attention is still being paid to borrowings, and replacements continue to be proposed for more recent loans in particular (see, for example, the website bolje.hr hosted by the IHJJ, which allows users to suggest their own replacements for borrowings; entries here include *višezadaćnost* for 'multitasking' and *sebić* or *samoslika* for 'selfie'). However, there are also numerous other borrowings that are well entrenched in the language (some of which are no longer recognizable by the average speaker as being foreign in origin) and which nobody has seriously proposed replacing. Modern Croatian language planning efforts are selective in nature, focusing largely on forms that have attracted the attention of language planners in some way. Rather than advocating a thorough purification of the language (which is impossible in practical terms), purists concentrate on the symbolic replacement of certain forms.

Previous corpus research has focused largely on journalistic texts, since (often for practical reasons) these were the predominant or only types of texts in the corpora that were used. While these earlier corpora provide valuable information about changing patterns of usage, particularly since the media play an influential societal role as linguistic models, journalistic texts are subject to the editorial policies of the individual publications included in these corpora, which are limited in number and are not necessarily representative of Croatian usage more generally. Journalistic texts are also heavily represented in HrWaC, but this corpus includes a much broader range of published sources (which, of course, may also be subject to editorial policies), in addition to containing blog and forum posts by individuals and a wide range of

[21] Both *raspit* and *slikokaz* are recommended by Šimundić (1994) and Protuđer (1998).

texts from other websites. However, the abundance of material here creates another problem: for forms with tens or hundreds of thousands of attestations over thousands of different sites, it is not practical to do a detailed analysis of factors that may influence the choice of one form over another, although computational methods may allow us to take some of these factors into account. In addition to more sophisticated analyses of currently available text corpora, a large-scale, representative corpus of Croatian texts from the 20th century is needed to provide a better point of comparison for earlier patterns of usage. Data on spoken language and the attitudes of Croatian speakers towards language planning efforts are also needed.

Other authors have expressed similar views to the quote from the newspaper article by Piteša and Kalogjera cited above, that Croatian language planning has had a destabilizing effect on the standard language, making speakers uncertain of which forms they should use. Even as an outsider, I have personally heard people speaking in public (on television, at academic conferences or other events) correct themselves in mid-sentence, switching from a dispreferred form to its recommended Croatian replacement. However, there is no evidence to document any drastic, large-scale changes in Croatian usage over the past few decades. Certain forms have increased in frequency, at least in some contexts, but these represent only a small fraction of the lexical stock and thus seem unlikely to threaten the stability of the norm. Language planning efforts have succeeded to the extent that people feel a need to use certain variants, especially in more formal contexts, but they have not (yet) dramatically reshaped the language in any fundamental way.

References

Anić, Vladimir, and Josip Silić. (1990) *Pravopisni priručnik hrvatskoga ili srpskoga jezika*. Zagreb: Školska knjiga.

———. (2001) *Pravopis hrvatskoga jezika*. Zagreb: Novi Liber–Školska knjiga.

Babić, Stjepan. (1995) *Hrvatski jučer i danas*. Zagreb: Školske novine.

Babić, Stjepan, Božidar Finka, and Milan Moguš. (1990) *Hrvatski pravopis*. Zagreb: Školska knjiga. [Subsequent editions 1994, 1995, 1996, 2000, 2002, 2003, 2004.]

Babić, Stjepan and Sanda Ham. (2004) "Pravopisni rat: Komentirana bibliografija publicističkih članaka o hrvatskome pravopisu objavljenih u 2000. i 2001. godini". *Jezik* 51(3): 130–53.

Babić, Stjepan, Sanda Ham, and Milan Moguš. (2005) *Hrvatski školski pravopis*. Zagreb: Školska knjiga. [Subsequent editions 2008, 2009, 2012.]

Babić, Stjepan and Milan Moguš. (2010) *Hrvatski pravopis*. Zagreb: Školska knjiga. [2nd ed. 2011.]

Babić, Zrinka. (2013) "O člancima i njihovu znanstvenomu vrjednovanju". *LAHOR* 15: 1–8.

Badurina, Lada, Ivan Marković, and Krešimir Mićanović. (2007) *Hrvatski pravopis*. Zagreb: Matica hrvatska. [2nd ed. 2008.]

Bagdasarov, Artur R. (2015) "*Hrvatski pravopis* Instituta za hrvatski jezik i jezikoslovlje kao objekt jezične politike". *Filologija* 64: 147–78.

Brodnjak, Vladimir. (1991) *Razlikovni rječnik srpskog i hrvatskog jezika*. Zagreb: Školske novine.

Bujas, Željko. (1999). *Veliki hrvatsko-engleski rječnik*. Zagreb: Globus.

Czerwiński, Maciej. (2005) *Język – ideologija – naród. Polityka językowa w Chorwacji a język mediów*. Cracow: Scriptum.

Dulčić, Mihovil, ed. (1997) *Govorimo hrvatski: Jezični savjeti*. Zagreb: Hrvatski radio–Naprijed.

Eastman, Carol M. (1983) *Language planning: An introduction*. San Francisco: Chandler & Sharp.

Frančić, Anđela, Lana Hudeček, and Milica Mihaljević. (2006) *Normativnost i višefunkcionalnost u hrvatskome standardnom jeziku*. 2nd ed. Zagreb: Hrvatska sveučilišna naklada.

Grčević, Mario. (2001) "Über die kroatischen Sprachveränderungen der 90-er Jahre zwischen Information, Desinformation und Sprachpolitik". *Die slawischen Sprachen* 67: 43–77.

Greenberg, Marc L. (2011) "The Illyrian movement: A Croatian vision of South Slavic unity". Joshua A. Fishman and Ofelia García, eds. *Handbook of language and ethnic identity: The success-failure continuum in language identity efforts*. Vol. 2. Oxford: Oxford University Press, 364–80.

———. (2015) "The Slavic area: Trajectories, borders, centres, and peripheries in the Second World". Dick Smakman and Patrick Heinrich, eds. *Globalising sociolinguistics: Challenging and expanding theory*. London–New York: Routledge, 164–77.

Gries, Stefan T. (2005) "Null-hypothesis significance testing of word frequencies: A follow-up on Kilgarriff". *Corpus linguistics and linguistic theory* 1–2: 277–94.

Ham, Sanda. (2015) "Pravopisna previranja u nas uvijek su vezana uz smjenu vlasti". Interview with Irena Kustura. *Večernji list*, 12 August. https://www.vecernji.hr/vijesti/pravopisna-previranja-u-nas-uvijek-su-vezana-uz-smjenu-vlasti-1018949 (last accessed 4 December 2017).

Haugen, Einar. (1983) "The implementation of corpus planning: Theory and practice". Juan Cobarrubias and Joshua A. Fishman, eds. *Progress in language planning: International perspectives*. Berlin–New York–Amsterdam: Mouton, 269–89.

HrWaC. http://nlp.ffzg.hr/resources/corpora/hrwac/ (last accessed 22 December 2017; as of July 2018 the corpus is available on a new server, https://www.clarin.si/noske/).

Hudeček, Lana, Milica Mihaljević, and Luka Vukojević. (2011) "Nekoliko aktualnih problema hrvatske jezične norme". *Tabula* 9: 88–103.

Jahn, Jens-Eberhard. (1999) "New Croatian language planning and its consequences for language attitudes and linguistic behavior—the Istrian case". *Language and communication* 19: 329–54.

Jelaska, Zrinka. (2014) "Vrste nasljednih govornika". *LAHOR* 17: 83–105.

Jojić, Ljiljana. (2015) *Veliki rječnik hrvatskoga standardnog jezika*. Zagreb: Školska knjiga.

Jozić, Željko et al. (2013) *Hrvatski pravopis*. Zagreb: Institut za hrvatski jezik i jezikoslovlje. [Online version: http://pravopis.hr, last accessed 3 July 2018.]

Kilgarriff, Adam. (2005) "Language is never, ever, ever random". *Corpus linguistics and linguistic theory* 1–2: 263–75.

Krmpotić, Marijan. (1992) *Jezični priručnik*. Zagreb: Hrvatska televizija, Služba za razvoj i izobrazbu.

Langston, Keith and Anita Peti-Stantić. (2014) *Language planning and national identity in Croatia*. Basingstoke: Palgrave Macmillan.

László, Bulcsú. (1994) "Tvorbeni pravopis". *Dometi* 11: 77–82.

Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolomäki, and Heikki Mannila. (2016) "Significance testing of word frequencies in corpora". *Digital scholarship in the humanities* 31(2): 374–97.

Ljubešić, Nikola and Filip Klubička. (2014) "{bs,hr,sr}WaC – Web corpora of Bosnian, Croatian, and Serbian". Felix Bildhauer and Roland Schäfer, eds. *Proceedings of the 9th Web as Corpus Workshop (WaC-9) @ EACL 2014*. Red Hook, NY: Curran, 29–35.

Maretić Žonja, Petra. (2005) "Sanader: I dalje ću pisati 'neću' zajedno". *Večernji list*, 19 December. https://www.vecernji.hr/vijesti/sanader-i-dalje-cu-pisati-necu-zajedno-813929 (last accessed 3 December 2017).

Moguš, Milan, Maja Bratanić, and Marko Tadić. (1999) *Hrvatski čestotni rječnik*. Zagreb: Zavod za lingvistiku Filozofskog fakulteta–Školska knjiga.

Pavić, Snježana. (2012) "RAT ZA JEZIK 'Ako kažemo špedicija, onda je šport, a ne sport! "Ministarstvo zdravlja" čista je BESMISLICA'". *Jutarnji list*, 5 January. https://www.jutarnji.hr/vijesti/hrvatska/rat-za-jezik-ako-kazemo-spedicija-onda-je-sport-a-ne-sport-ministarstvo-zdravlja-cista-je-besmislica/1630374/ (last accessed 4 December 2017).

Piteša, Adriana and Ivana Kalogjera. (2012) "Hrvatski jezik, 20 godina poslije: Što su jezični puristi devedesetih htjeli napraviti od hrvatskog? I u čemu su uspjeli". *Jutarnji list*, 1 August. http://www.jutarnji.hr/vijesti/hrvatski-jezik-20-godina-poslije-sto-su-jezicni-puristi-devedesetih-htjeli-napraviti-od-hrvatskog-i-u-cemu-su-uspjeli/1630764/ (last accessed 26 August 2017).

Protuđer, Ilija. (1998) *Pravilno govorim hrvatski: Praktični jezični savjetnik*. Split.

Rittgasser, Stjepan. (2003) "Mijene u leksiku hrvatskoga jezika". *Jezik* 50(1): 6–14.

Schou Madsen, Martin. (2017) *Planning against change: Serbian and Croatian reactions to contact-induced linguistic innovation*. Ph.D. dissertation, University of Copenhagen.

Skelin Horvat, Anita. (2004) "Posuđivanje u hrvatski jezik u dvama razdobljima". *Suvremena lingvistika* 57–58: 93–104.

SrWaC. http://nlp.ffzg.hr/resources/corpora/srwac/ (last accessed 22 December 2017; as of July 2018 the corpus is available on a new server, https://www.clarin.si/noske/).

Šimundić, Mate. (1994) *Rječnik suvišnih tuđica u hrvatskomu jeziku*. Zagreb: Barka.

Šojat, Zorislav. (1983) *Čestotni rječnik Večernjeg lista i Vjesnika*. Zagreb.

Šonje, Jure, ed. (2000) *Rječnik hrvatskoga jezika*. Zagreb: Leksikografski zavod Miroslav Krleža–Školska knjiga.

Tafra, Branka. (1995) "Obilježja hrvatske gramatičke norme do kraja 19. stoljeća". *Filologija* 24–25: 349–53.

Tanocki, Franjo. (1994) *Hrvatska riječ: Jezični priručnik*. Osijek: Matica hrvatska.

Težak, Stjepko. (1991) *Hrvatski naš svagda(š)nji*. Zagreb: Školske novine.

Thomas, George. (1978) "The origin and nature of lexical purism in the Croatian variant of Serbo-Croatian". *Canadian Slavonic papers* 20(3): 405–20.

———. (1988) *The impact of the Illyrian movement on the Croatian lexicon*. Munich–Berlin–Washington, DC: Otto Sagner. [*Slavistische Beiträge*, 223.]

———. (1991) *Linguistic purism*. London–New York: Longman.

Vijeće za normu. (2013) "Zapisnici sjednica Vijeća za normu hrvatskoga standardnog jezika". *Jezik* 60 (2–4): 62–149.

University of Georgia
langston@uga.edu

Appendix: Lexical Data

Pairs of (near) synonyms, where the second member of each pair is the recommended Croatian form, according to various sources. For nouns referring to people, feminine derivatives in *-ica*, *-ka*, *-kinja* are included in the totals, except as indicated above (see ftn. 16).

	HČR		HrWaC		SrWaC	
	N	%	N	%	N	%
advokat	27	87.1	2745	2.8	15511	97.1
odv(j)etnik	4	12.9	95887	97.2	460	2.9
‘lawyer’						
aerodrom	47		29834	57.2	23646	99.5
zračna luka	—[22]		22288	42.8	127	0.5
‘airport’						
ambasada	5	100.0	7112	26.2	19720	99.0
veleposlanstvo	0	0.0	20021	73.8	204	1.0
‘embassy’						
ambasador	21	100.0	10710	26.3	29324	99.4
veleposlanik	0	0.0	30038	73.7	170	0.6
‘ambassador’						
analiza	107	100.0	128796	98.7	74872	100.0
raščlamba	0	0.0	1669	1.3	10	<0.1
‘analysis’						
armija	94	38.7	18236	11.7	15336	15.8
vojska	149	61.3	138295	88.3	81951	84.2
‘army’						
atmosfera	48	96.0	95759	83.2	40116	99.5
ozračje	2	4.0	19346	16.8	183	0.5
‘atmosphere’						
avion	100	95.2	69065	53.1	42504	99.3
zrakoplov	5	4.8	60939	46.9	298	0.7
‘airplane’						

biografija	10	83.3	17185	59.0	12579	95.3
životopis	2	16.7	11922	41.0	618	4.7
‘biography’						
bioskop	0	0.0	559	0.8	14036	87.6
kino	24	100.0	69061	99.2	1983	12.4
‘cinema’						
browser/brauzer	—		6111	33.5	4263	94.0
preglednik	—		12124	66.5	293	6.0
‘(web) browser’						
budžet	26	47.3	21198	13.0	89566	94.3
proračun	29	52.7	142389	87.0	5415	5.7
‘budget’						
centar	71	48.6	412504	82.6	268358	96.6
središte	75	51.4	87054	17.4	9310	3.4
‘center’						
civilizacija	24	96.0	36323	97.0	17845	99.9
uljudba	1	4.0	1133	3.0	23	0.1
‘civilization’						
čitalac	24	88.9	1084	1.7	34259	98.0
čitatelj	3	11.1	64026	98.3	694	2.0
‘reader’						
datum	16	100.0	78813	96.8	40815	100.0
nadnevak	0	0.0	2641	3.2	5	<0.1
‘date’						
delegacija	154	100.0	18680	52.9	24733	98.8
izaslanstvo	0	0.0	16642	47.1	310	1.2
‘delegation’						
delegat	91	94.8	10333	45.6	6214	61.7
izaslanik	5	5.2	12313	54.4	3864	38.3
‘delegate’						

demokratija	1	4.5	3027	4.0	38009	97.2
demokracija	21	95.5	71732	96.0	1078	2.8
‘democracy’						
direktor	101	99.0	209660	62.1	180490	99.8
ravnatelj	1	1.0	128067	37.9	420	0.2
‘director’						
ekonomija	13	5.0	65803	27.5	70632	48.8
gospodarstvo	14	5.4	159304	66.6	1661	1.1
privreda	234	89.7	14221	5.9	72353	50.0
‘economy’						
Evropa	—		29792	8.6	201048	96.7
Europa	—		316329	91.4	6830	3.3
‘Europe’						
evropski	106	100.0	20789	3.7	285757	98.4
europski	0	0.0	538114	96.3	4574	1.6
‘European’						
fabrika	8	5.6	4005	4.4	62913	98.5
tvornica	135	94.4	86370	95.6	986	1.5
‘factory’						
faktor	125	98.4	60072	64.3	45835	99.4
čimbenik	2	1.6	33417	35.7	290	0.6
‘factor’						
finale	10	83.3	111492	78.8	36432	82.9
završnica	2	16.7	30016	21.2	7537	17.1
‘finale, final(s)’						
firma	7	43.8	110750	17.8	148968	98.0
tvrtka	9	56.3	512450	82.2	3086	2.0
‘firm’						
geografija	6	40.0	6548	62.9	4812	97.0
zemljopis	9	60.0	3860	37.1	148	3.0
‘geography’						

glasanje	3	100	12928	34.1	19584	98.6
glasovanje	0	0	25022	65.9	276	1.4
‘voting’						
gledalac	48	100.0	1035	1.3	10332	97.5
gledatelj	0	0.0	77248	98.7	261	2.5
‘viewer’						
godišnjica	56	71.8	28465	37.1	11882	96.6
obl(j)etnica	22	28.2	48339	62.9	424	3.4
‘anniversary’						
gr(j)eška	46	52.3	91240	58.1	60111	97.9
pogr(j)eška	42	47.7	65692	41.9	1296	2.1
‘mistake’						
grupa	200	69.7	305620	51.3	256657	98.5
skupina	87	30.3	290576	48.7	3989	1.5
‘group’						
hardver	—		8382	93.5	6821	99.8
očvrsje	—		21	0.2	0	0.0
sklopovlje	—		564	6.3	12	0.2
‘hardware’						
hapšenje	6	100.0	3748	13.1	14872	98.9
uhićenje	0	0.0	24940	86.9	169	1.1
‘arrest’						
hemija	0	0.0	428	1.9	9753	99.1
kemija	28	100.0	22435	98.1	88	0.9
‘chemistry’						
hiljada	50	22.5	12637	3.5	109677	97.6
tisuća	172	77.5	350439	96.5	2655	2.4
‘thousand’						
informisati (se)	0	0.0	324	0.8	14708	98.0
informirati (se)	20	100.0	39281	99.2	295	2.0
‘to inform’						

izuzetno	23	76.7	89247	46.9	76063	98.8
iznimno	7	23.3	101203	53.1	901	1.2
‘exceptionally’						
izv(j)eštaj	55	100.0	61035	36.9	89793	99.1
izv(j)ešće[23]	0	0.0	104282	63.1	785	0.9
‘report’						
kandidat	49	100.0	166675	98.4	71009	100.0
pristupnik	0	0.0	2755	1.6	11	<0.1
‘candidate’						
klavir	15	100.0	13177	76.0	4943	99.0
glasovir	0	0.0	4161	24.0	49	1.0
‘piano’						
komisija	75	98.7	108304	59.2	88008	99.6
pov(j)erenstvo	1	1.3	74593	40.8	340	0.4
‘commission’						
kompjuter	0	0.0	23988	14.2	21875	93.8
kompjutor	1	33.3	9505	5.6	91	0.4
računalo	2	66.7	135985	80.2	1362	5.8
‘computer’						
kompozicija	25	83.3	14224	32.2	10562	98.4
skladba	5	16.7	29930	67.8	169	1.6
‘composition’						
ležeći policajac	—		861	75.3	135	94.4
uspornik	—		282	24.7	8	5.6
‘speed bump’						
muzika	78	40.8	43350	16.4	78848	98.7
glazba	113	59.2	221757	83.6	1026	1.3
‘music’						
nauka	126	59.2	19201	11.6	83541	98.4
znanost	87	40.8	146760	88.4	1392	1.6
‘science’						

nesumnjivo	16	69.6	10642	42.1	9018	97.8
nedvojbeno	7	30.4	14607	57.9	201	2.2
‘undoubtedly’						
nivo	59	51.8	68018	17.1	195932	98.4
razina	55	48.2	329644	82.9	3167	1.6
‘level’						
oficir	68	90.7	6837	24.6	11428	97.7
časnik	7	9.3	20998	75.4	271	2.3
‘officer’						
opozicija	26	92.9	16193	33.8	33982	99.6
oporba	2	7.1	31783	66.2	138	0.4
‘opposition’						
organizovati (se)	0	0.0	1555	0.7	120705	98.2
organizirati (se)	92	100.0	227848	99.3	2241	1.8
‘to organize’						
originalan	20	55.6	63889	60.0	25359	69.4
izvoran	16	44.4	42545	40.0	11163	30.6
‘original’						
osnivanje	54	87.1	60389	80.2	30166	99.4
osnutak	8	12.9	14929	19.8	190	0.6
‘founding, establishment’						
ostrvo	6	5.5	2087	1.1	35959	86.3
otok	103	94.5	182917	98.9	5699	13.7
‘island’						
pantalone	0	0	122	0.5	5476	94.9
hlače	36	100	25669	99.5	293	5.1
‘pants’						
pasoš	1	100.0	2276	10.2	11685	98.3
putovnica	0	0.0	20077	89.8	201	1.7
‘passport’						

patrola	11	73.3	4864	56.7	3241	99.3
ophodnja	4	26.7	3709	43.3	22	0.7
‘patrol’						
pauza	72	34.8	35565	60.5	20465	91.0
stanka	135	65.2	23218	39.5	2036	9.0
‘pause’						
pažnja	121	95.3	115922	62.3	98681	99.3
pozornost	6	4.7	70055	37.7	718	0.7
‘attention’						
penzija	9	25.7	14696	13.7	40822	98.9
mirovina	26	74.3	92465	86.3	436	1.1
‘pension’						
period	40	23.4	85158	28.6	173257	95.3
razdoblje	131	76.6	212250	71.4	8465	4.7
‘period’						
por(ij)eklo	20	66.7	34181	48.7	40658	99.1
podr(ij)etlo	10	33.3	36011	51.3	359	0.9
‘origin’						
pos(j)etilac	24	100.0	4717	4.5	43734	98.7
pos(j)etitelj	0	0.0	99021	95.5	555	1.3
‘visitor’						
postepeno	32	61.5	11046	35.3	12956	90.1
postupno	20	38.5	20258	64.7	1426	9.9
‘gradually’						
pozorište	0	0.0	3207	3.1	45402	97.8
kazalište	128	100.0	101223	96.9	1024	2.2
‘theater’						
prilika	278	94.2	361015	82.6	183071	99.5
prigoda	17	5.8	76125	17.4	1007	0.5
‘opportunity, occasion’						

prisutan	104	98.1	114427	78.2	63946	99.4
nazočan	2	1.9	31848	21.8	405	0.6
‘present’						
propaganda	19	95.0	16596	56.6	9527	99.1
promidžba	1	5.0	12732	43.4	82	0.9
‘propaganda, advertising’						
protest	11	84.6	12567	15.7	34579	98.8
prosv(j)ed	2	15.4	67592	84.3	418	1.2
‘protest’						
radnik	381	100.0	254217	67.8	144712	99.3
djelatnik[24]	0	0.0	120829	32.2	1055	0.7
‘worker’						
raskršće	43	97.7	3115	20.1	1667	96.4
raskrižje	1	2.3	12380	79.9	63	3.6
‘intersection’						
reciklaža	—		3548	80.3	6756	100.0
oporaba	—		872	19.7	1	<0.1
‘recycling’						
sala	7	4.6	6234	4.1	13342	54.1
dvorana	146	95.4	144097	95.9	11303	45.9
‘hall, auditorium’						
saučešće	9	39.1	655	7.0	1827	95.9
sućut	14	60.9	8668	93.0	78	4.1
‘sympathy’						
savremen	0	0.0	2283	1.9	80913	98.8
suvremen	159	100.0	120441	98.1	964	1.2
‘contemporary’						
sekretar	111	82.2	7642	8.2	44440	98.8
tajnik	24	17.8	85901	91.8	531	1.2
‘secretary’						

sistem	303	68.4	102841	14.7	349517	98.9
sustav	140	31.6	597977	85.3	3767	1.1
‘system’						
slušalac	13	92.9	215	1.5	2572	92.9
slušatelj	1	7.1	14251	98.5	197	7.1
‘listener’						
softver	—		40040	97.4	33184	99.5
mekušje	—		3	<0.1	0	0.0
napudbina	—		9	<0.1	0	0.0
programska oprema	—		1070	2.6	164	0.5
‘software’						
spisak	6	16.2	7241	5.7	27460	60.5
popis	31	83.8	120873	94.3	17950	39.5
‘list’						
sport	15	88.2	160282	87.5	82972	99.9
šport	2	11.8	22915	12.5	97	0.1
‘sport(s)’						
srećan	11	2.5	1802	1.0	41028	88.4
sretan	427	97.5	184324	99.0	5374	11.6
‘happy’						
supa	0	0.0	574	2.2	5207	95.3
juha	17	100.0	25158	97.8	258	4.7
‘soup’						
taksa	8	72.7	2615	18.1	11460	99.5
pristojba	3	27.3	11863	81.9	60	0.5
‘fee, charge’						
talas	63	30.1	1024	1.7	22799	87.7
val	146	69.9	59985	98.3	3212	12.3
‘wave’						

učesnik	7	15.6	10233	6.9	87440	98.9
sudionik	38	84.4	138439	93.1	994	1.1
‘participant’						
uniforma	49	77.8	13573	55.9	6926	88.7
odora	14	22.2	10697	44.1	880	11.3
‘uniform’						
upotreba	128	100.0	71450	51.2	78529	99.3
uporaba	0	0.0	67950	48.7	531	0.7
poraba	0	0.0	249	0.2	6	<0.1
‘use’						
uputstvo	5	13.2	6931	11.3	20482	88.6
uputa	33	86.8	54389	88.7	2647	11.4
‘instruction(s)’						
uslov	12	2.9	1644	0.5	209458	98.6
uv(j)et	409	97.1	344230	99.5	2933	1.4
‘condition’						
uticaj	0	0.0	5987	3.3	84734	98.1
ut(j)ecaj	136	100.0	177413	96.7	1678	1.9
‘influence’						
utisak	17	27.0	9377	7.6	50485	97.8
dojam	46	73.0	114315	92.4	1136	2.2
‘impression’						
vasiona	5	4.6	215	0.4	2044	15.0
svemir	103	95.4	59289	99.6	11621	85.0
‘universe’						
vaspitanje	0	0.0	392	0.6	16100	91.2
odgoj	81	100.0	70349	99.4	1548	8.8
‘education, upbringing’						
verzija	12	100.0	139871	83.6	66868	99.6
inačica	0	0.0	27397	16.4	275	0.4
‘version’						

v(ij)ek	114	38.5	50708	18.9	167202	95.6
stol(j)eće	182	61.5	217284	81.1	7707	4.4
‘century’						
voz	41	31.3	2213	3.6	18974	97.3
vlak	90	68.7	59415	96.4	523	2.7
‘train’						
zakletva	9	69.2	3564	47.5	4716	99.0
prisega	4	30.8	3934	52.5	50	1.0
‘oath’						
zvanično	1	3.3	2714	3.8	27675	88.3
službeno	29	96.7	68350	96.2	3673	11.7
‘officially’						

[22] HČR gives frequencies only for individual words, not lexicalized phrases.

[23] The total from SrWaC is approximate, since it is not possible to automatically distinguish all instances of *izvešće* (noun) from *izvest(i)* + *će* → *izvešće* ‘will lead out; will perform’ in a search.

[24] *Djelatnik* had the primary meaning ‘someone active in, a participant’, but in Croatia has been increasingly used in the meaning of ‘worker’. Many of the instances in HrWaC are in contexts where *radnik* was previously the norm (e.g., *Naša tvrtka periodično traži djelatnike sljedećih profila* ‘Our firm periodically seeks workers with the following profiles’).