A large number of publications have been devoted to questions of Croatian linguistic identity and language purism in the period since the collapse of the former Yugoslavia. A few of these have used corpus data, most often drawn from mass media sources, to investigate patterns of lexical usage (e.g., Rittgasser 2003), and studies such as Langston (1999) and Czerwiński (2005) have shown that while words that have been promoted as “pure Croatian” replacements for forms that were previously used in both Croatia and Serbia are attested everywhere, puristic tendencies are clearly stronger in media outlets that have a more right-wing/nationalist orientation. One must assume that these differences reflect the different editorial policies of these media organizations. While these results provide evidence of language policy decisions in the independent Croatian state, the influence of these policies on Croatian speakers in general remains uncertain. It would be useful to compare the results of previous studies with an analysis of texts that have not been subjected to official lektoriranje (language editing) or the self-editing that individuals usually apply in formal written contexts. This paper examines a 1,000,000-word corpus of texts from 30 different Croatian blogs, covering a period from May 2004-March 2005. Not surprisingly, this corpus shows a much lower frequency for many of the forms in question, although a number of characteristically Croatian words have achieved a certain degree of acceptance. These data will also be compared with the results of a survey of teachers in Croatia about the acceptability of various lexical items and with the usage in a comparable corpus of texts drawn from Serbian blogs.

References: