

## **Croatian Corpus Processing: State-of-art and Perspectives**

*dr. Marko Tadić* (marko.tadic@ffzg.hr)

Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu

(<http://www.ffzg.hr/zzl/zzl-home.htm>)

This article tries to give a survey of Croatian corpus processing. It lists the most important projects since first Croatian computer corpus (Gundulić's *Osman*) up to the present time. The article focuses on the Croatian National Corpus which is the central project in the field of corpus linguistics in Croatia today. The Croatian National Corpus consists of two parts: 1) representative 30-million Corpus of Contemporary Croatian Language and 2) Croatian Electronic Text Archive. The 30-million Corpus covers the first phase of Croatian National Corpus while the effort in the second phase will be concentrated to the widening the contents of Croatian Electronic Text Archive. The 30-million Corpus, which is now at the stage of advanced planning and software and pilot corpus (8.5 million of running words) testing, is targeted to be finished in the year 2000.