Automatic Recognition of Czech Derivational Prefixes

Alfonso MEDINA Urrea[1] and Jaroslava HLAVÁČOVÁ[2]

[1] GIL IINGEN UNAM
Apartado Postal 70-472, 04510 Coyoacán, DF, MEXICO
amedinau@iingen.unam.mx
[2] ÚFAL MFF UK
Malostranské náměstí 25, 11800, Praha, CZECH REPUBLIC
hlava@ufal.mff.cuni.cz

Abstract. This paper describes the application of a method for the automatic, unsupervised recognition of derivational prefixes of Czech words. The technique combines two statistical measures — Entropy and the Economy Principle. The data were taken from the list of almost 170 000 lemmas of the Czech National Corpus

1 Introduction

Our contribution concerns only those languages where words are created by means of affixes. Usually, there exists a quite stable vocabulary, but it is possible to create entirely new words adding suffixes and/or prefixes to already existing ones. If the derivation follows common rules for word creation, everybody understands them, even if they have never seen them before.

The Czech language belongs to the group of languages that derive their vocabulary mainly by means of adding affixes. While the set of suffixes is very stable and does not change during long periods of time, prefixes are much more vivid. Of course, there is a set of old, traditional prefixes, that have been used for a very long time and do not change. But one can very easily add a morph, usually borrowed from other languages, in front of an existing word and create an entirely new word. The old prefixes can be found in every grammar, but the new ones cannot.

Everybody who understands the language understands new prefixes. Everybody but computers. And for any analysis of language, it is very important to know them. Without a sufficiently large list of prefixes, we cannot run successfully enough a morphological analysis, which usually stands on the basis of all automatic language processing.

To be specific — the morphological analyzer always encounters unknown words; that is, words for which it does not recognize their basic forms nor their morphological categories. It is possible to design a "guesser" that uses special properties of the language which could help to guess those basic features of the unknown word. In Czech, we usually take suffixes as the basis [1].