Tracking Lexical Changes in the Reference Corpus of Slovene Texts

Vojko Gorjanc

University of Ljubljana, Faculty of Arts, Aškerčeva 2, 1000 Ljubljana, Slovenia

vojko.gorjanc@guest.arnes.si

Keywords: corpus linguistics, corpus-based lexicology, lexical semantics, lexical changes, Slovene language

Introduction

With the help of a corpus, we can track lexical changes quickly and reliably, and also observe the response of a selected language to new lexical items introduced into it from other languages, e.g. English, or some other language with which the selected language has direct contact, in the case of Slovene, these are Italian, German, Hungarian or Croatian. The present paper focuses on lexical items introduced into English in the last decade of the 20th century. The starting point for the comparison with the state in the corpus of the Slovene language is John Ayto's list of lexical items. We will try to determine how the Slovene language reacts to lexical items introduced into Slovene from English as the language of global society. Using the corpus, we can track the new lexical items through the last decade, and observe their characteristics in the corpus.

The Corpus

The Corpus of the Slovene Language, called FIDA, is a reference corpus of Slovene. It is composed of contemporary Slovene texts, the majority of which were published in the 1990's. The corpus contains just over 100 million words, encompassing a broad variety of language variants and registers. It is composed of written texts and texts originally produced as written-to-be-spoken; the transcripts of Slovene parliamentary proceedings are the only spoken component of the corpus.

Methodology

Using a FIDA wordlist, we will obtain information on the lexical items from Ayto's list which are relevant to the Slovene language. By means of corpus analysis, we will determine when a word occurs in the Slovene language and how it establishes itself in the language, and by means of statistical analysis, we will determine the possible collocations of the selected word and their changes from the first occurrence of the word until the end of the decade. Since pairs of synonyms or strings often occur with new expressions, we will try to determine how they occur in the corpus and how they disappear. With the help of markers of semantic relations already identified for the Slovene language by corpus analysis, we will identify pairs of synonyms and strings within the corpus. We will focus on the context of synonyms, their distribution within the corpus regarding the time of their occurrence and the genre, and on when and why one of the synonyms becomes dominant in the language while other variants disappear. For extracting collocations, the MI^3 value with information on the probability of a word pair occurring together or separately will be used. The MI^3 value has turned out to be sound information for content words in Slovene. On the other hand, with this value it is hard to determine function words as part of collocations; in particular, in the case of collocators of verbs and nouns with prepositional phrases. For instance, to detect propositional words, raw statistics provide more valuable information. After detection, a noun + propositional word pair, for example, the MI^3 value for the whole pair is calculated to extract the string of collocators.

Conclusion

In its observation of the process of accepting new lexical items into Slovene, corpus analysis reveals the great creativity of Slovene language speakers; in addition to loan words, original Slovene expressions occur almost invariably. Corpus data shows a great deal of variability, linked above all to the desire for original expressions, and, after a few years, the data begin to reveal the prevailing variant within the entire corpus or a selected genre. The question of which variant is eventually fully accepted, however, remains open. Full acceptance of a variant is conditioned by a series of linguistic as well as non-linguistic factors; in order for a variant to be fully accepted in Slovene today, it must be also sexy and cool.