ABSTRACT

Corpora and Language Education:
Exploitation potentials in teaching Greek and construction of pedagogically relevant corpora

Ph.D. thesis

Maria E. Giagkou

This thesis pertains to the field of Applied Corpus Linguistics, a relevantly new field that aims at contributing to the objectives of Applied Linguistics, by adopting the principles and methodological tools of Corpus Linguistics. In particular, the thesis focuses on language education and seeks to promote the discussion for the use of corpora in Greek as mother tongue teaching and to identify the advantages that derive from exploiting corpora in the language lesson.

In this context, the ways in which the methods and findings of Corpus Linguistics can update and enrich the pedagogical language description, the content of the language lesson and the teaching practice are investigated. On this basis, two main issues are highlighted that constitute the primary research objectives:

a) the inductive inferencing of pedagogically useful conclusions from findings of data-driven or data-based linguistic inquiry and

b) the construction of pedagogically relevant corpora, i.e. corpora that are suitable for integration in the teaching practice and exploitable by pupils

Towards the first objective, the thesis formulates specific proposals for updating the content of the Greek language textbooks and curricula in primary and secondary education to accommodate for information from actual language use. More specifically, the phenomena that were analysed are collocations, learned (in contrast to colloquial) participles, the use of final –n in the negative adverb *δεν* and the conditionals. Language use data for each of the above phenomena were drawn from Greek corpora and analysed. By contrasting these data to the language lesson content, a set of issues were revealed, making evident the insufficient or inaccurate coverage of these phenomena in the pedagogical language description.

The second objective of this thesis pertains to mother tongue teaching practice. It is rooted in the fact that, in order to obtain learning outcomes through the introduction of corpora in teaching practice, the corpora should be pedagogically relevant. An important dimension of pedagogical relevance is the reading difficulty / readability level of the texts that comprise a corpus. The quantitative estimation of reading difficulty / readability of Greek texts was used as the means to construct a pedagogically relevant corpus.

Existing readability formulas and text classification techniques were used as a basis for building a statistical text classification model for Greek.To this end, a wide set of text features that quantify vocabulary difficulty, sentence structure complexity and text coherence were investigated, so as to identify a subset of features that could be used as indicators to collectively determine the level of reading difficulty. The proposed model takes into account the values of a number of the above mentrioned textual characteristics and decides whether a text is readable or not for junior high school students. The model can be exploited for the extraction of readable texts from a Greek reference corpus, thus constructing a pedagogically relevant sub-corpus.

The model was validated in two complementary experiments. The results of the automatic classification were compared to education experts' judgment (as reflected in their selection of texts included in the Greek language textbooks) and also to students' own appraisal. For the second part of the validation an extensive field research was carried out involving 913 junior high school students that were asked to fill in cloze tests for a number of texts. In the validation results the model achieved a high overall percentage of correct classifications, providing clear evidence of its suitability for the problem at hand. This was further attested by comparing the model to the existing readability formulas for Greek, which revealed a significantly higher correct classification rate in the order of 10%.

The contribution of the current thesis is summarised in the following:

- It extends the scope of Applied Corpus Linguistics, limited to second and foreign language teaching and learning in the current literature, to first language teaching

- It provides evidence on the value of data-driven linguistic inquiry for language education through specific examples which demonstrate that the educational exploitation of corpora can bridge the gap between language that is taught and actual language use, thus leading to a more thorough, accurate and "realistic" language description
- It covers a significant gap in Greek readability research that had been limited to the adaptation of existing readability formulas, originally developed for languages other than Greek. The text classification model proposed in this thesis offers a robust solution for the automatic construction of pedagogically relevant corpora comprising readable texts.