ORIENTAL MANUSCRIPTS AND NEW INFORMATION TECHNOLOGIES

P. Zemanek

CORPUS LINGUISTICS AND ARABIC

The corpus linguistics can be characterized as a computer-aided analysis of large amounts of texts stored in a machine readable form, which provides empirical data on the language that can be used for further interpretation. The number of corpora (text and speech) and lexical databases available is constantly increasing, as well as the number of institutions that are active in this field. It is of course natural that corpus linguistics is going to witness a fast growth in the near future. That is why it is certainly going to affect the Arabic studies. In this article, we would like to have a look at the possibilities, problems and perspectives of corpus linguistics and Arabic. At the current stage, most of the remarks will be connected with the construction of a corpus.

The fast developments in this field have been so far limited mostly to European languages, where the number of corpora available and those under construction is considerably high. Nowadays, almost every European language has got its own corpus or has such a corpus under preparation. Projects like the Bank of English [1], British National Corpus [2], and many others show the direction in which the corpus linguistics goes today, i.e. first of all quantitative growth, offering researchers more reliable statistical data.

The corpora that are available today can be divided into several types, according to the text type, annotation type and according to their use.

The corpora according to text type are:

1) Balanced corpora that consist of different genres of size proportional to the distribution of a certain text type within the language in question. An example of an attempt to construct a balanced corpus is the Brown Corpus.

2) Pyramidal corpora range from very large samples of a few representative genres to small samples of a wide variety of genres.

3) Opportunistic corpora: their method of the texts acquisition can be characterized by "take what you can get". This makes their construction easier, but, on the other hand, can have consequences for the reliability of the results. It is believed that a huge size of such a corpus avoids the problems with the representativeness of the sample. Sometimes, they are also called "monitoring corpora".

Corpora divided according to annotation type are:

1) Raw, i.e. that text is only tokenized and cleaned [3], no additional tagging is done.

2) PoS tagged: Raw text is annotated with syntactic category at word level (part-of-speech tagging).

3) Treebanks: PoS tagged text is annotated with skeletal syntactic structure. Typically a parse grammar is defined. Corpora are automatically parsed. Parse trees are selected and if necessary corrected by human annotators. Word strings for which no parse tree is found by the grammar are either omitted or manually annotated.

4) Linguistically interpreted corpora: this type of corpora aims at deliberate annotation of various kinds of linguistic information. In a sense the treebanks can be considered a subtype to the linguistically interpreted corpora.

The third criterion that can be used for the corpora classification is their use, where we get the corpora used for training, mostly statistical models for natural language processing and speech processing; corpora used for testing, i.e. for evaluation of statistical models after training.

Besides, there are also corpora used for speech recognition or speech generation. Such a type of corpora is of minor importance for Modern Standard Arabic as a primarily written language. The corpora of speech in Arabic will be rather limited to the dialects, as is the case with the CALLHOME corpus of Egyptian Arabic speech [4].

The developments of corpus linguistics in connection with Arabic are not that many at present. There are some corpora that are used for research, but most of them are only in a raw form, i.e. they are not tagged for the morphological, syntactical or other type of linguistic information. According to my knowledge, the only corpus so far which has been announced to be fully tagged for both morphological and syntactical information is currently not available for research [5].

On the other hand, it seems that there is time for a start in the Arabic corpus linguistics. There are possibilities of obtaining large amounts of Arabic texts in electronic form. There are several Arabic newspapers that offer their data on CDs (*al-Ḥayāt*, London, etc.) or on the Internet (*al-Rāya*, Qatar; *al-Waṭan*, Qatar, etc.), and several other products where Arabic texts can be obtained. Besides, the Arabic OCR has reached an acceptable standard for cleanly printed texts in modern, computer-generated fonts [6]. This means that the primary condition necessary for a computer-aided analysis of Arabic texts is fulfilled.

For analysis of such a type of data, there is currently no specialized linguistic program available, but there is