

---

# 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-Hayāt*, 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

a number of linguistic software available that can be used for analysis. The basic requirements for a program to be used for analysis of Arabic texts are in fact limited to the use of the whole set of upper ASCII, and preferably a possibility of defining one's own sorting order [7], but there might be a problem, especially with the DOS-based programs, in viewing the results. This may lead to using the software for the analysis, but for interpretation of the results, a software capable of viewing Arabic texts has to be used. This may not be the most comfortable way of work but it meets the second necessary condition for a computer-aided analysis of a language.

The Semitic languages like Arabic present for a computer aided analysis of texts a special challenge. The difficulties lie in several points which to a considerable degree influence the type of an ideal corpus of Arabic. These problems are mainly in the special character of the graphemic representation, which is limited mostly to consonants and long vowels. The vocalization signs for short vowels, gemination, case endings, etc., are used only occasionally. This considerably restricts the information in the text and increases the ambiguity of such a type of a text. Another problem is morphotactic, i.e. that certain types of synsemantic words can be added to an autosemantic word in its traditional definition, and these form in its graphemic representation a single string without explicitly marked morphological boundaries.

It is relatively difficult to determine an exact proportion between the graphemes of a vocalized and non-vocalized text, since the ideal cases, i.e. the texts completely free of vocalization signs, are relatively rare. Almost every text has at the most ambiguous places at least some indication of the way the text should be read, especially indications of a passive reading or gemination. On the other hand, it can be argued what a fully vocalized text is. In the so-called substandard norm we witness frequent omission of the case endings and elision of the indefinite article. This fact leads to unclear frontiers on both sides. Nevertheless, when confronting completely non-vocalized text with the fully vocalized one we get the proportion of 1 : 1.584, i.e. the non-vocalized text contains about 63% of the information comprised in the fully vocalized text [8].

Such an increased ambiguity has its consequences on what type of information should be tagged. For dealing with the vocalization, there are several ways of solving this problem. First, it is possible to fully or partially vocalize the text, which would bring the processing of Arabic close to the natural language processing of other types of languages, but, on the other hand, would take the analysis away from what is a basic characteristics of Arabic graphemic representation [9]. Secondly, it is possible to add a fully vocalized form of the token as a tag. It is as laborious as the first solution, and the two solutions are quite close to each other. On the other hand, the solution with only tagging the grammatical information together with a root information would be sufficient for a construction of a vocalized form.

Regarding the type of the Arabic morphology, this ambiguity is even more deepened. The concept of it is based on the so-called consonantal root, which forms the semantic base, and an actual word is derived from it by addition of the vocalic pattern and affixes. For example, the root *drs* (درس) is connected with the concept of study and vocalized forms like *darasa* "to study", *durisa* "to be studied", *darsun* "lesson, lecture", *madrasatun* "school" are exam-

ples of the actualization of this root. The first three words are moreover in the non-vocalized text represented by the string *drs* (درس). The root is also used in European dictionaries of Arabic as the sorting criterion and the real words are ordered under this morphologico-semantic abstraction. In the real text, these consonants are usually surrounded (and in some cases even divided, in case of infixes) by other graphemes. The root consonants can be further changed by assimilation or in case of the so-called weak radicals even elided. This further impedes the identification of the root and its look-up in the dictionary and demands a thorough knowledge of the Arabic derivation system.

According to recent estimations [10], there are about 5,000 roots used in the current Arabic texts, and about 400 derivational patterns, most of them are further ambiguous. On the other hand, there is no root that would make use of all the derivational possibilities. Every root combines only with a smaller group of these patterns, in average 17—18.

Almost every form based on the root is further ambiguous. Only very little number of patterns are fully unambiguous and most of the forms have more possibilities of vocalization. The number of these possibilities usually varies from 2 to 5, but, in extreme cases, it can reach a considerably high numbers. For example, the sequence *y'd* (يعد) can be interpreted as belonging to several roots:

— root *'dd* (عد) "to count": verb forms: indicative, subjunctive, apocopate;

— root *'wd* (عود) "to return": verb forms: apocopate, apocopate of the 4th verbal stem;

— root *w'd* (وعد) "to promise": verb forms: indicative, subjunctive, apocopate.

Altogether, this sequence has 8 possible vocalizations, and this number can be doubled by the use of the passive form (i.e. 16 possibilities). Even more possibilities has the sequence *t'd* (تعد), where thanks to the fact that Arabic does not formally distinguish in the imperfect verb forms between the 2nd person masculine singular and the 3rd person feminine singular, the number of the possibilities would then be again doubled, i.e. 32 possible forms altogether.

Regarding to the problems mentioned above we assume that the root information is also one of the essential types of information to be contained in a corpus of Arabic. It would certainly be very useful to get some tools that would be able of a (successful) root analysis, but the fulfilment of such a requirement is not met today.

As it has been pointed above, there are also synsemantic words, like particles, prepositions and pronouns, that are prefixed or suffixed to the autosemantic words, based on the root. This means that the concept of a word as one string is seriously changed in the Arabic script. A string can not only form a word, but can consist of several morphological units, like prepositions, the actual word, and suffixed personal or possessive pronouns.

The words that can be prefixed to the word are first of all: the definite article, prepositions (*bi-*, *li-*, *wa-*, etc.), various types of particles (*fa-*, *la-*, *sa-*, etc.). As it was the case above, also here we have a possibility of ambiguity. Sometimes the question whether the first letter belongs to the word or is a prefix to the word can have several solutions, as there is a number of biliteral roots in Arabic that can be identical with the rest of the roots with initial *b* for example [11].

The suffixed words are limited to personal pronouns only, that can be suffixed both to nouns and verbs. The ambiguity here is not as big as in the preceding case, but still is present, especially in the singular form of the pronouns. The roots beginning with *b* and ending with *h* are four in Hans Wehr's dictionary [12], and three of them can be ambiguous, i.e. interpreted as both having *h* as the final radical or having *h* suffixed as a pronoun.

Such a situation shows that it would be very useful to have some kind of morphological information available so that these difficulties can be overcome. One possibility is mentioned in Beesley 1996 [13], and it is an automatic morphological analyzer, that should be able to provide the information on the morphological boundaries of the strings and the root. Another possibility, which can be used in cor-

pus linguistics, is to tag the corpus also for morphological information in such a way that it can be used when necessary.

Let us now have a look at basic quantitative data. The figures given here are deducted from a corpus of 100,000 words, from a newspaper news, and most of them are only in a raw form, without more sophisticated analysis, and thus are to be taken only for orientation.

The text corpus consisting of 100,000 words contains only 21,059 tokens [14], and of them, 12,165 occur only once, i.e. 57.7% of the whole corpus consists of isolated tokens. The frequencies between 2 and 10 form another 35.4% of the tokens, i.e. altogether 93.1% of the tokens. These data are summarized in the following table:

Table 1

The distribution of tokens frequency lower than 10

Frequency	Number	%	Frequency	Number	%
1	12,165	57.7	6	445	2.1
2	3,179	15.1	7	300	1.4
3	1,471	7	8	238	1.1
4	873	4.1	9	188	0.9
5	632	3	10	163	0.7

We have tried to count the number of verbs in the tokens appearing more than 10 times, i.e. of 1,558 tokens. The whole number of verbs appearing in this set is 196 [15], but this number contains also various representations of verbs, there are, for example, 9 forms

of the verb *كان*, 8 forms of the verb *قام*, etc., which means that the actual number of various verbs will be considerably lower. The following table shows the 10 most common verbs together with their various manifestations:

Table 2

The 10 most frequent verbs

Nos.	Verb.	Frequency	Manifestations
1.	قال	588	(15) تقول, (27) وتقول, (36) يقول, (37) ويقول, (53) قالت, (116) قال, (292) وقال, (12) قالوا
2.	كان	378	(20) سيكون, (24) كانوا, (30) يكن, (38) يكون, (47) وكانت, (51) تكون, (141) كانت, (11) ستكون, (16) تكن
3.	أكد	155	(11) واكدت, (12) تؤكد, (12) تؤكد, (14) اكدت, (51) أكد, (55) واكد
4.	اشار	146	(11) وتشير, (12) تشير, (16) يشار, (32) واشارت, (75) واشار
5.	تم	144	(13) سيتم, (22) تتم, (41) يتم, (68) تم
6.	قام	133	(25) يقوم, (27) قام, (29) قامت, (52) تقوم
7.	عمل	108	(11) عملت, (19) يعمل, (23) تعمل, (55) عمل
8.	ليس	94	(23) وليس, (25) ليست, (46) ليس
9.	يمكن	73	(73) يمكن
10.	اضاف	69	(11) وازضاف, (58) اضاف



Plate 1

The table clearly shows that it is only the most common words that appear in a 100,000 words text with frequency big enough to draw some conclusions on their behaviour [16]. It is to be expected that less common verbs will need much bigger corpus to provide enough data on their use in the language.

These types of difficulties more or less determine the shape of a corpus of Arabic. It is obvious that for more sophisticated analysis, the corpus should be tagged, and the minimum requirements for the tags types are: (i) tagging morphological boundaries; (ii) part-of-speech tags; and (iii) providing the root information. The size of the corpus has to be relatively big, as showed the analysis of some characteristics of a 100,000 words text, which obviously provided enough information only on the most common words. The example of the Brown corpus of English (1 million words) shows that even such a size is not big enough for a proper analysis of a language, and in case of Arabic as a flectional language it is clear that the frequencies of especially verbs would be much less. It is quite

probable that, e.g., for a lexical studies, even a corpus consisting of 10 million words might not be big enough.

This lead us to the decision to start work on a corpus of Arabic [17], aimed at modern standard Arabic, especially from the last 30 years. The projected size of the corpus is now 30 million words, and we assume that this size might be big enough even for lexical studies. The basic characteristics of the corpus would be: a balanced corpus with tags for morphological boundaries, part-of-speech, and root.

As the corpus is projected as a balanced one, we will try to cover as many varieties of Arabic as possible, i.e. we will gather texts from all major regions of Arabic, i.e. the Arabic Maghreb, Mashreq, and the Gulf area. It will cover both texts from periodicals (newspapers, magazines) and books, and will try to find a balance between various language styles.

Below, there is one of possible shapes of the corpus, certainly not free of problems and points that have to be further discussed.

Table 3

Number of the token	Token	Morphological boundaries	PoS tag [18]	Root
0001	وكان	و-كان	VPBe	كون
0002	الخلفاء	ال-خلفاء	NNP	خلف
0003	من	من	Prep	—
0004	الجهة	ال-جهة	NPS	وجه
0005	الآخرى	ال-آخرى	NAs	ءخر
0006	يعرفون	يعرفون	VIP3m	عرف
0007	حاجة	حاجة	NNS	حوج
0008	لأمراء	ال-أمراء	NNP	ءمر
0009	المسلمين	ال-مسلمين	NNP	سلم
٠٠١٠	الى	الى	Prep	—
0011	رضاهم	رضا-هم	NNsP	رضو

## Notes

1. A constantly growing commercial project of a monitoring corpus of English. Available at the University of Birmingham. A number of words in the corpus announced in summer 1996 was 320 million.
2. A project directed by the Oxford University Press, a balanced 100 million words corpus.
3. I.e. only the control characters are eliminated, only headlines and paragraphs are possibly marked.
4. The CALLHOME Egyptian Arabic corpus of telephone speech, available from the Linguistic Data Consortium, University of Philadelphia, consists of 120 unscripted telephone conversations between native speakers of Egyptian Colloquial Arabic. For more details, cf. the LDC Home page (<http://www ldc.upenn.edu>).
5. This corpus has been developed by the Sakhr Company (Egypt, Saudi Arabia, (<http://www.sakhr.com>)). According to my knowledge, it is available only internally for the company.
6. E.g., the 3rd version of Sakhr's Automatic Reader offers acceptable results even without the necessity of training the fonts. Besides, there are products offered by Caere (Arabic OmniPage) and TexPert for Macintosh. In the reviews that appeared in the electronic discussion lists (especially ITISALAT), the Sakhr's product seems to be superior to the other ones. According to my own experience, with quality printouts the success rate can reach 99%, requiring only very little postprocessing.
7. The last requirement is not really serious, since the character sequence on both DOS/Windows and Macintosh platforms more or less retain the character order of the Arabic alphabet.
8. The completely non-vocalized text in the extent of 1,000 graphemes resulted in our analysis in 1,584 graphemes of its fully vocalized counterpart, i.e. with the representation of all the short vowels, endings, and geminated consonants.

9. This might not be that serious for a linguist, but it is impractical in two aspects. First, the acquisition of new data would be very laborious, and secondly, any practical applications might fail to analyse real Arabic texts.

10. Kenneth R. Beesley, "Arabic finite-state morphological analysis and generation". Paper read at COLING-96, Copenhagen, August 1996, 6 pp.

11. The ambiguous cases can be quite numerous, for example, in Hans Wehr's dictionary, the roots beginning with *bj* are 8 and of them, 6 can be interpreted as consisting of the preposition *bi-* and a biradical root.

12. Hans Wehr, *A Dictionary of Modern Written Arabic*. An enlarged and improved version of Hans Wehr's *Arabisches Woerterbuch für die Schriftsprache der Gegenwart*, English translation by J. M. Cowan (Wiesbaden 1961—1994).

13. Beesley, "Arabic finite-state morphological analysis and generation".

14. The "token" here is understood as any string between two spaces. This certainly means that there are strings that contain more than one word, i.e. there are strings that consist of prefixes (prepositions, particles, etc.), word and suffixes (suffixed pronouns), as it has been described here above. Another fact worth of attention is that these tokens do not distinguish between various types of parts of speech, i.e. one token can represent both verbs and nouns. This has also been mentioned here above.

15. This number is a number of various verb forms appearing in the set. There are certainly strings that can be interpreted as both verbs or nouns, but since they can be interpreted as both, it can be assumed that these strings, at least to some extent, represent also verbs.

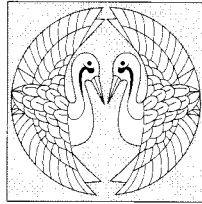
16. It is obvious that the types of verbs here correspond very strongly with the type of the text used for the collection of data. Most of the verbs are typical for a political news type of text.

17. From 1997, this project is supported by the Grant Agency of the Czech Republic, under the name *Thesaurus Linguae Arabicae*.

18. The tags used here are only provisional, there are still problems to be discussed. E.g., there is little difference between names and adjectives in Arabic, quite often a word can serve both as a noun or an adjective. Another problem is the representation of affixed words, and there are many other issues that will need a careful consideration.

---

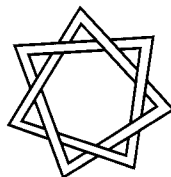
RUSSIAN ACADEMY OF SCIENCES  
THE INSTITUTE OF ORIENTAL STUDIES  
ST.PETERSBURG BRANCH



# **Manuscripta Orientalia**

*International Journal for Oriental Manuscript Research*

Vol. 3 No. 3 November 1997



**ТБЕСА**  
**St. Petersburg-Helsinki**

T 4338  
D.L.G.

## CONTENTS

<i>TEXTS AND MANUSCRIPTS: DESCRIPTION AND RESEARCH</i> . . . . .	3
E. Kychanov. "The Altar Record on Confucius' Conciliation", an Unknown Tangut Apocryphal Work . . . . .	3
I. Kulganek. Manuscripts and Sound Records of the Mongol-Oirat Heroic Epic "Jangar" in the Archives of St. Petersburg . . . . .	8
<i>TEXT AND ITS CULTURAL INTERPRETATION</i> . . . . .	11
E. Rezvan. The Our'ān and Its World: III. "Echoings of Universal Harmonies" (Prophetic Revelation, Religious Inspiration, Occult Practice) . . . . .	11
S. Klyashtorny. About One Khazar Title in Ibn Faḍlān . . . . .	22
<i>PRESENTING THE COLLECTIONS</i> . . . . .	24
O. Yastrebova. Reconstruction and Description of Mīrzā Muḥammad Muqīm's Collection of Manuscripts in the National Library of Russia . . . . .	24
<i>MANUSCRIPTS CONSERVATION</i> . . . . .	34
M. Blank, N. Stavisky. Conservation of Medieval Manuscripts in the Library of the Jewish Theological Seminary of America . . . . .	34
<i>ORIENTAL MANUSCRIPTS AND NEW INFORMATION TECHNOLOGIES</i> . . . . .	46
P. Zemanek. Corpus Linguistics and Arabic . . . . .	46
<i>PRESENTING THE MANUSCRIPT</i> . . . . .	54
L. N. Menshikov. An Album of Illustrations to the Famous Chinese Novels . . . . .	54
<i>BOOK REVIEWS</i> . . . . .	69

### Front cover:

"Ni Heng (173—198), a poet in the service of Cao Cao". Illustration No. 31 to the Chinese novel *Three Kingdoms* from the Album H-13 preserved in the manuscript collection of the St. Petersburg Branch of the Institute of Oriental Studies (early 19th century), 15.6 × 19.6 cm.

### Back cover:

- Plate 1.** "A high-spirited stone, a divine oriole". Illustration No. 46 to the Chinese novel *A Dream in the Red Chamber* from the same Album, 15.5 × 19.6 cm.
- Plate 2.** "Shi Ziang-yun falling asleep on the stone bench". Illustration No. 58 to the Chinese novel *A Dream in the Red Chamber* from the same Album, 15.2 × 19.6 cm.
- Plate 3.** "Lin Dai-yu speaking to a parrot". Illustration No. 57 to the Chinese novel *A Dream in the Red Chamber* from the same Album, 15.5 × 19.5 cm.