Sherds from an Arabic Treebanking Mosaic

Otakar Smrž and Petr Zemánek

Abstract

This paper would like to introduce the reader into those aspects of the Arabic language which require some special treatment compared to languages Europeans are more familiar with. In spite of having fresh experience in building the Prague Arabic Dependency Treebank, the authors try to take a broader view of the problems encountered under way. The topics discussed include linguistic data retrieval, morphology and morphotactics modelling, and description of the language on the analytical level.

1 Introduction

Let us assume a background knowledge of the motivation and concepts of the Prague Arabic Dependency Treebank project (cf. Smrž, Šnaidauf, Zemánek 2002). Its idea is to follow the practice set up by the Prague Dependency Treebank for Czech, as long as analogy between the two languages allows.

There are points, however, which we would like to draw attention to, since they defy the usual considerations rather than yielding to them, or simply have not been dealt with before. We shall focus on these without much stress on the complete and overall look in the frame of their application, and thus become free to mention also other approaches not necessarily realized by our team.

2 Characteristics of the Arabic Language and Script

Arabic is, together with the Northwestern Semitic, a branch of the Central group of West Semitic languages (for further details, cf. Faber 1997). The literary language should be the same throughout the Arabic speaking countries, the local dialects can be considerably distinct from each other. Generally, Arabic is the mother tongue of about 300 million people.

Arabic is usually characterized as a highly introflective language, i.e. a language, where—apart from other standard instruments of flexion (mainly desinential in case of Arabic)—there is a system of inner flexion, based on different roles of the consonantal root (mostly tri-radical), its vocalization and various affixes. This system works mainly in the word-building, where the root is considered a semantic base of the word, the vocalization together with the root forms a stem (lexical and partially also morphological "actualization" of the root), which is closer to the actual meaning of the word. The word is then "finalized" by the use of various affixes.[1] In the non-concatenative approach (cf., e.g., McCarthy 1985), this scheme is represented in several tiers, where the respective morphemes (root, vocalization, affixes) occupy one such tier, and the word is then built by the junction of these tiers. This has been also used in the NLP domain (cf. works by Kiraz, e.g. 1998, 1999 and 2000, and Beesley 1999).[2] The system is also used in the organization of dictionaries of Arabic (esp. those produced in the West), which means that for dictionary look-up it is necessary that the user be capable of full morphological analysis of a given word form, or potentially a string of word forms (see below).

---

[1] E.g., for the root `[ktb]`, the vocalization **katab** represents the stem of the verb to write in perfect tense, **kotub** the stem of the same verb in imperfect (both active voices). **katabato** then means she wrote, **yakotubu** is he writes, the prefixes/suffixes and the stems being independent of each other. The notation of Arabic is treated in Section 3.

[2] In his works, Kiraz uses a consonantal skeleton and a vocalization as a template, from which the actual word form is generated. E.g., a template CVCVC with the root CCC=`[ktb]` and vocalization VV=`[aa]` gives **katab** as the result. Beesley, on the other hand, rather works with so-called patterns, i.e. CaCaC in this case.

For the NLP, it is mainly the Arabic script that presents a challenge. It uses two types of graphemes: "real" graphemes which represent mainly consonants (Huru~f, "letters" in Arabic), and additional marks, used mainly for vocalization (in English generally referred to as "diacritics", in Arabic as Haraka~t, "movements"). In the script—as a principle—we find mostly the marks from the first group, representing the consonantal skeleton of a word. Rarely does the script comprise the so-called vocalization marks, which include marks for vowels and other characters (e.g., a mark for gemination of a consonant). The virtual absence of "diacritics" certainly increases ambiguity of Arabic texts.

The degree, density and quality of vocalization differ according to the kind of the text, being then described as fully vocalized, partially vocalized or non-vocalized. The distinction may sometimes be fuzzy since omission/enforcement of diacritics can happen locally in contrast to the global style. Table 1 offers a rough view of the classification reasoned by the distribution of graphemes:

Components	Fully Vocalized		Partially Vocalized	Non-vocalized
	Bible	Quran	Fiction	Other
graphemes	3 743 329	610 644	36 860 639	151 840 850
letters	55.73 %	55.49 %	97.53 %	99.94 %
diacritics	44.27 %	44.51 %	2.47 %	0.06 %

Table 1: Percentage of letters vs. diacritics in chosen Arabic texts. The CLARA corpus (Zemánek 2001) provides canonized religious texts as the only reliable source of fully vocalized data. Fiction features partial vocalization, while the other subcorpora are non-vocalized, settling closely to the average values shown above.

Another drawback of the notation of Arabic is that words can be clustered into one string of characters, within which, for the sake of further analysis, the morphotactic borders have to be set. These clusters can consist of one autosemantic word, definite article and a number of functional words, such as prepositions (especially uniliteral), conjunctions and pronouns (objective at verbs, possessive at nouns, mixed at prepositions). Also some other markers (the future tense marker etc.) can appear. Thus, e.g., the string **fsyktbwnhA** so they will write it/them can be divided into at least four words—cf. Table 6. This means that for analyses of Arabic on the analytical and the tectogrammatical levels, it is necessary to provide a sequence of tokens resulting from the morphologically disambiguated language.

Most of morphological analyzers of Arabic offer segmentation on the level of morphemes. For further treatment of the output of these analyzers, a decision has to be made in respect of how this information is to be represented in the morphological annotation and how the strings of characters are to be divided into tokens needed on the analytical level.

The preceding sentence implicitly gives the answer to this problem. This means that in case there is a syntactic type of government within the respective string, this government has to be reflected on the analytical level, and these "new" units have to be generated by the splitting of the original string.[3] Else, there are connections that do not have to be separated, although they do not belong to the original form of the autosemantic word.[4]

In the syntax, Arabic can be characterized as a language with prevailing VSO word order. This, however, holds mainly for sentences where the role of the predicate is played by a verb. Besides, there are sentences with non-verbal predication, where the predicate is expressed by a noun, prepositional phrase or by other means. In nominal sentences, the word order is mostly inverse, i.e. subject, predicate and no object. In addition, Arabic has a number of topicalizers, which make the word order far from fixed, and thus increase the number of instances differing from the VSO order.

[3] In some cases, also other changes than splitting are necessary. E.g., the preposition **li_** combines with the definite article **{al_** merging into **lil_** instead of **li{al_**. During tokenization, missing graphemes must be restored.

[4] Here belongs e.g. the definite article **{al_**, which certainly does not form a part of the lexeme. However, it does not have to be represented as a unit on the analytical level, and it is better to keep the article's value in a morphological tag.

3 Representation of Arabic

Since the times when Unicode came into general use, persistent encoding of the Arabic script has not been a problem to talk about. Nevertheless, in recognized cases, adopting alternative transliteration or transcription systems may prove more convenient.

The Arabic script is suited for recording individual phonemes of the language. Written from right to left, the strokes continuously cross the boundaries of letters, the shapes of which conform to the adjacent glyphs or letter forms (initial, medial, final, isolated). Irrespective of the presence or absence of short vowels and other optional marks (altogether referred to as diacritics), the algorithm determining the glyphs given the letters is well-defined.

This regularity (provided that the original script is correct) makes it possible to encode just the letters and let the forms be computed at the very moment of script rendering. While Unicode charts Arabic Presentation Forms-A (0xFB50–0xFDFF) and Arabic Presentation Forms-B (0xFE70–0xFEFF) ensure fidelity by remembering every single ligature of a sequence of glyphs, the more common systems like Unicode Arabic (0x0600–0x06FF), Windows CP 1256, ISO 8859-6 or lower ASCII Buckwalter transliteration introduce one-to-one mappings of distinct graphemes, i.e. letters and diacritics.

Unlike these graphemic transliteration concepts, the typesetting system of ArabTeX (Lagally 1999) defines its own notation, which covers both contemporary and historical orthography in an excellent way. Moreover, the encoding is human-readable, and thus comes in handy wherever the script were too difficult to display or edit. The point is that ArabTeX has to evaluate a larger context of each lower ASCII character to generate the corresponding Arabic representation. Real-time conversions become however less efficient then.

We will use Buckwalter transliteration in examples emphasizing the actual manner of vocalization, e.g. in morphology analyses, whereas ArabTeX notation will restore the complete word forms. An approximate phonetic transcription shall be enough to engage in the dependency trees later on. Table 2 demonstrates these three encodings on an Arabic sentence asking you to “read this text carefully”.

Aiqora>o h`*aA {ln~aS~a bi{notibaAhK	Buckwalter graphemic transliteration
iqra’ h_a_dA an-na.s.sa bi-intibAhiN	ArabTeX transliteration encoding
iqra’ ha~Va~ an-naSSa bi- intiba~hin	phonetic transcription of tokenized text

Table 2: Comparison of lower ASCII transliterations and a phonetic transcription. The fully vocalized text implies use of various diacritical marks, even those echoing an empty vowel (Buckwalter). Original orthography is preserved, though it disguises for the sake of readability (ArabTeX). Understanding all the graphemes and the proper pronunciation of the symbols in our transcription, quite vague indeed, is not essential for this paper.

4 Linguistic Data Retrieval

The resources being exploited in the treebanking project count LDC Arabic Newswire A Corpus (ANAC), Corpus Linguae Arabicae (CLARA) and Ummah Arabic-English Parallel News Corpus (UAEC). After characterizing each data set, we shall explain our method of document topic analysis which helped retrieve a domain-specific language resource.

4.1 Resource Information

Arabic Newswire A Corpus was collected by the Linguistic Data Consortium (LDC), University of Pennsylvania, from the news which appeared on the Agence France Presse (AFP) Arabic Newswire in the period from May 1994 to December 2000. The corpus contains roughly 80 million words in about 384 thousand documents of a wide thematic scope. Although renowned information retrieval experi-

ments have been performed on these data (cf. Sawaf et al. 2001, Brants et al. 2002, Oard et al. 2002), there is no reusable topic identification associated with the data yet.

Corpus Linguae Arabicae is, on the other hand, a topically classified corpus of Modern Standard Arabic which has been compiled at the Institute of Comparative Linguistics, Charles University in Prague (Zemánek 2001). Out of the total of 53 million words, 13 million constitute a subcorpus of the language of economics, business and finance (mostly from the Hayat newspaper of the years 1995–1997), the rest being news in general, expert materials, fiction, and scientific literature.

The last corpus to mention is Ummah Arabic-English Parallel News Corpus, based on various Arabic newspapers digests issued weekly by Ummah Press Service in Cairo. The news stories in this collection, gathered by the LDC, date from January 2001 to March 2002 (reported by Xiaoyi Ma of the LDC, July 31st 2002). There are 3,039 story pairs giving 13,027 sentence pairs, or 765,492 words altogether, 352,759 in Arabic and 412,733 in English.[5]

4.2 Document Topic Analysis

The topically distinguished subcorpora of CLARA can be utilized for building reference models of each particular language domain. For an arbitrary document, some measure of conformity or similarity to the given model may be studied to see whether both the document and the subcorpus fall in the same thematic class. Out of the set of all ANAC articles, possible candidates to treat economics, business or finance (to be found also in legal and industrial texts) were extracted like this. Still, before including them in the new resource of the desired property, humans must have confirmed their relevance.

4.2.1 The method and its application

While diverse techniques may be employed (Oard et al. 2002), we resorted to statistical modelling. The choice of the method was conditioned by the size of the data in question. The documents to classify comprised just 200 words on average, spanning say from 50 to 500 words. That is why distribution of unigrams occurring in a text was taken as our modelling criterion. It would have been hopeless to follow any bigger elements, once having such sparse testing data.

According to CLARA subdivision, reference models of these domains were established: economics and finance, law, industry, agriculture, traffic, politics, humanities, sports, medicine, science, arts, fiction, as well as a complementary non-economics model (fields from law to fiction). Furthermore, global models for both CLARA and ANAC corpora were prepared.

Even for every single ANAC document, a unigram model can be constructed. Its resemblance to the reference models gets quantified, for instance, by the value of the correlation coefficient between the respective distribution functions (normalized to integrate to one). In order to enhance sensitivity to linguistic nuances of the domains, unigram frequencies beyond a certain interval were clipped to zero. Empirically, the reliability range of <0.002 %, 0.200 %> was set for reference models, while <0.002 %, 100.0 %> imposed no upper bound on the models being tested.

Those documents which identified best with one of the first three reference models, or which identified with them on the second position and whose correlation coefficient there scored at least 90 % of the winning value, proceeded to manual verification. Such texts provided in total more than 1 million words. Humans themselves had difficulties in appointing sharp and unbiased criteria for topic assignment, anyway, their judgements disqualified one third of the pool.

There are probably many ANAC articles which were never recognized and yet should have been, as there are those which did suit the method and were rejected by humans. Recalling our initial intention of building a domain-specific language resource, reducing but not eliminating the manual effort, we dare declare our solution successful.

[5] Word counts for ANAC and CLARA share the definition of a word (strings delimited by boundaries between incompatible characters), which is different from that used with UAEC (splitting on whitespace only). In neither case are the words real linguistic units, as justified by the tokenization problem.

4.2.2 Discussion and remarks

The method of correlation coefficient evaluation may be interpreted equivalently in terms of vector calculus. Let us imagine that every distinct word form (or n-gram) generates one dimension of a vector space. A language model over these forms then renders as a vector the co-ordinates of which correspond to the frequencies of the n-grams. In such a case, the correlation coefficient equals the cosine of the angle contained by the vectors of the models being compared. Naturally, the classification process likens to finding which of the reference vectors deviates least from the vector being studied.

The notion of vectors makes it easy to consider relations among the reference models, too.[6] Our experiment in Table 3 tells about orientation of the domain vectors relative to the vectors of CLARA and ANAC, and indicates some prevailing character of the topics in the corpora.

The Table also shows discrepancies in the quality of both resources. While ANAC data are robust and uniform, CLARA models do not seem representative enough due to the low ratio of words to word forms. Its subcorpora may not be formatted consistently, and feature different typographic conventions. Definitely, ANAC transcribes all foreign proper names and abbreviations into Arabic, while the Hayat newspaper keeps Roman characters intact. It would therefore be necessary to improve the language models prior to tuning-up the aspects of the method.

Topic Domain	Word Count	Form Count	W/F	Correlation Coef.		Deviation Angle	
				CLARA	ANAC	CLARA	ANAC
economics and finance	12 722 560	272 378	46.7	0.737	*0.573*	42.5	***55.0***
law	1 121 202	84 097	13.3	0.709	0.492	44.8	60.5
industry	2 507 161	111 648	22.5	0.646	0.372	49.8	68.2
agriculture	671 601	43 261	15.5	0.542	0.372	57.2	68.2
traffic	532 440	59 651	8.9	0.679	0.389	47.2	67.1
politics	9 928 893	284 848	34.9	*0.820*	*0.695*	***34.9***	***46.0***
humanities	9 053 453	481 989	18.8	0.754	0.392	41.1	66.9
sports	1 240 809	91 823	13.5	0.662	*0.566*	48.5	***55.5***
medicine	1 649 972	148 004	11.1	0.757	0.464	40.8	62.4
science	1 710 542	144 332	11.9	*0.846*	0.506	***32.2***	59.6
arts	713 117	80 621	8.8	0.748	0.434	41.6	64.3
fiction	10 812 730	652 273	16.6	0.692	0.334	46.2	70.5
non-economics	39 948 703	1 159 126	34.5	*0.898*	0.539	***26.1***	57.4
CLARA	52 671 263	1 227 361	42.9	1.000	*0.609*	0.0	***52.5***
ANAC	79 872 381	555 973	143.7	0.609	1.000	52.5	0.0

Table 3: Reference models and their relation to CLARA and ANAC. Characteristics of the data sets prompt questions about their reliability. The discussion above explains why the deviation angles (given in degrees) are better for CLARA (union of subcorpora) than for ANAC (uneven data type).

5 Morphological Analysis and Disambiguation

Given an unresolved string of Arabic characters, morphological analyzers commonly spell out a word stem and all underlying morphemes, clitics etc. with their appropriate labelling. The systems usually differ in the implementation of the parsing process and in the method of stem decomposition, if any, into the root and the pattern (cp. Beesley or Kiraz or Cavalli-Sforza et al.). Let us have a closer look on those the performance of which has been tested during our project.

[6] In fact, there is no need for vector discrimination in the space.

Xerox Arabic Morphological Analyzer (XAMA) is based clearly on the two-level morphology reusing finite-state tools developed for language independent processing by the Xerox Research Centre Europe (cf. Beesley 2001). Though its analyses offer valuable information on roots, patterns and case and mood endings, some intricate derivational schemes are missing and the vocabulary cannot be easily extended by end-users.

Tim Buckwalter's Arabic Morphology Analyzer (cf. Maamouri and Cieri 2002) does not go in detail about root and pattern nor does it discover all imaginable readings. It works with a lexicon of stem entries to which input strings are reduced while obeying Arabic morphotactics rules. The system, being utilized in the PENN Arabic Treebank as well as in the Prague Dependency Treebank projects, is iteratively refined according to real corpus evidence and comments from annotators.

5.1 Ambiguity and Tokenization Problems

The orthographical conventions of ignoring diacritics in writing and of tying words together increase the number of interpretations of a string in an extraordinary way. There are, of course, ambiguities caused by morphonological transformations applying widely to weak Arabic consonants, or by other systematic or incidental language tricks.

Examples shall support such claims. Table 4 summarizes all existing readings for a string **fhm**. The first column identifies the solutions for reference, the last two provide the full Arabic forms and their translations into English. Explicit linguistic information rests in the analysis strings, the format of which comes from the XAMA tool but has been modified to seem more intuitive. Besides, five solutions in the Table were inferred from other sources (Wehr 1974), therefore the use of the XAMA$^+$ heading.

ID	XAMA$^+$ String ~ [**Root**&**Pattern**]+**Morpheme**+Label	Full Form	Gloss
1	[**fhm**&**CaCiC**]+Verb+FormI+Perf+Act+**a**+3P+Masc+Sg	fahima	he understood
2	[**fhm**&**CuCiC**]+Verb+FormI+Perf+Pass+**a**+3P+Masc+Sg	fuhima	he was understood
3	[**fhm**&**CaC~aC**]+Verb+FormII+Perf+Act+**a**+3P+Masc+Sg	fahhama	he made understand
4	[**fhm**&**CuC~iC**]+Verb+FormII+Perf+Pass+**a**+3P+Masc+Sg	fuhhima	he was made to understand
5	[**fhm**&**CaC~iC**]+Verb+FormII+Impv+**o**+Masc+Sg	fahhim	make [sg.m.] understand
6	[**fhm**&**CaCoC**]+Noun+**N**+Indef+Nom	fahmuN	understanding [1.indef.]
7	[**fhm**&**CaCoC**]+Noun+**K**+Indef+Gen	fahmiN	understanding [2.indef.]
8	[**fhm**&**CaCoC**]+Noun+**u**+Def+Nom	fahmu	understanding [1.] *to/of*
9	[**fhm**&**CaCoC**]+Noun+**i**+Def+Gen	fahmi	understanding [2.] *to/of*
10	[**fhm**&**CaCoC**]+Noun+**a**+Def+Acc	fahma	understanding [4.] *to/of*
11	**fa**+Conj+[**hmm**&**CaCaC**]+Verb+FormI+Perf+Act+**a**+3P+Masc+Sg	fa-hamma	so he commenced
12	**fa**+Conj+[**hmm**&**CuCiC**]+Verb+FormI+Perf+Pass+**a**+3P+Masc+Sg	fa-humma	so he was commenced
13	**fa**+Conj+[**hmm**&**{uCoCuC**]+Verb+FormI+Impv+**i**+Masc+Sg	fa-hummi	so commence [sg.m.]
14	**fa**+Conj+[**hmm**&**CaCoC**]+Noun+**N**+Indef+Nom	fa-hammuN	and interest [1.indef.]
15	**fa**+Conj+[**hmm**&**CaCoC**]+Noun+**K**+Indef+Gen	fa-hammiN	and interest [2.indef.]
16	**fa**+Conj+[**hmm**&**CaCoC**]+Noun+**u**+Def+Nom	fa-hammu	and interest [1.] *in/of*
17	**fa**+Conj+[**hmm**&**CaCoC**]+Noun+**i**+Def+Gen	fa-hammi	and interest [2.] *in/of*
18	**fa**+Conj+[**hmm**&**CaCoC**]+Noun+**a**+Def+Acc	fa-hamma	and interest [4.] *in/of*
19	**fa**+Conj+[**hym**&**{iCoCiC**]+Verb+FormI+Impv+**o**+Masc+Sg	fa-him	so be [sg.m.] in love
20	**fa**+Conj+[**whm**&**{iCoCiC**]+Verb+FormI+Impv+**o**+Masc+Sg	fa-him	so imagine [sg.m.]
21	**fa**+Conj+**hum**+Funcwa	fa-hum	and they [pl.m.an.]
22	[**wfy**&**{iCoCiC**]+Verb+FormI+Impv+**o**+Masc+Sg+**hum**+Pron+DO+3P+Masc+Pl	fi-him	fulfil [sg.m.] them [pl.m.an.]

Table 4: Possible readings of the fhm string. Notice the clustering of solutions (1,2), (3–5), (6–10), (11–13), (14–18), (19), (20), (21), (22) which groups together words of the same lexical unit. Derivations of imperatives 13, 19, 20 and 22 from the canonical forms of the verbs are quite adventurous for an Arabic grammarian. So adventurous that the **{uCoCuC** or **{iCoCiC** patterns no longer appear on the surface. Nonetheless, all of the transformations are present and frequent in today's Arabic.

The ordering of the solutions tries to reflect their lexical relationship, thus splitting the list into nine disparate clusters. While the top analyses treat the input string as one word, later on, two separable tokens, varied in nature, are identified. This is the point why morphological disambiguation is a prerequisite to operations on the analytical level.

Partly by chance, partly by the power of the word-merging phenomenon, even the notorious example of a **ktb** sequence can be tokenized in two ways, and read in seventeen. Table 5 gives explanation.

ID	XAMA+ String ~ [Root&Pattern]+Morpheme+Label	Full Form	Gloss
1	[**ktb**&**CaCaC**]+Verb+FormI+Perf+Act+**a**+3P+Masc+Sg	kataba	he wrote
2	[**ktb**&**CuCiC**]+Verb+FormI+Perf+Pass+**a**+3P+Masc+Sg	kutiba	he was written
3	[**ktb**&**CaC~aC**]+Verb+FormII+Perf+Act+**a**+3P+Masc+Sg	kattaba	he made write
4	[**ktb**&**CuC~iC**]+Verb+FormII+Perf+Pass+**a**+3P+Masc+Sg	kuttiba	he was made to write
5	[**ktb**&**CaC~iC**]+Verb+FormII+Impv+**o**+Masc+Sg	kattib	make [sg.m.] write
6	[**ktb**&**CuCuC**]+Noun+**N**+Indef+Nom	kutubuN	books [1.indef.]
7	[**ktb**&**CuCuC**]+Noun+**K**+Indef+Gen	kutubiN	books [2.indef.]
8	[**ktb**&**CuCuC**]+Noun+**u**+Def+Nom	kutubu	books [1.] *of*
9	[**ktb**&**CuCuC**]+Noun+**i**+Def+Gen	kutubi	books [2.] *of*
10	[**ktb**&**CuCuC**]+Noun+**a**+Def+Acc	kutuba	books [4.] *of*
11	[**ktb**&**CaCoC**]+Noun+**N**+Indef+Nom	katbuN	writing up [1.indef.]
12	[**ktb**&**CaCoC**]+Noun+**K**+Indef+Gen	katbiN	writing up [2.indef.]
13	[**ktb**&**CaCoC**]+Noun+**u**+Def+Nom	katbu	writing up [1.] *of*
14	[**ktb**&**CaCoC**]+Noun+**i**+Def+Gen	katbi	writing up [2.] *of*
15	[**ktb**&**CaCoC**]+Noun+**a**+Def+Acc	katba	writing up [4.] *of*
16	**ka**+Prep+[**tbb**&**CaCoC**]+Noun+**K**+Indef+Gen	ka-tabbiN	like destruction [2.indef.]
17	**ka**+Prep+[**tbb**&**CaCoC**]+Noun+**i**+Def+Gen	ka-tabbi	like destruction [2.] *of*

Table 5: Possible readings of the ktb string. The lexical clustering gives (1,2), (3–5), (6–10), (11–15) as related units of the same root, while (16,17) resolve a prefixed preposition and a root of a totally different semantic content.

5.2 Lemma and Positional Tag in Arabic

Disambiguation of a set of morphological analyses applicable to a string in question does not only yield tokens for the upper levels of linguistic description, but in itself relates the tokens (word-forms) to their actual canonical forms (lemmas). If a word-form derives from its lemma as taking on certain morphological properties, then they are to be revealed by the analysis and pronounced in some labelling (tag).

Token	XAMA+ Tokenized String	Lemma	Positional Tag	Interpretation
fa-	**fa**+Conj	**fa_**	P---------	particle *of consequence*
sa-	**sa**+Fut	**sa_**	P---------	particle *of future tense*
yaktubUna	**ya**+VPref+[**ktb**&**CoCuC**]+Verb+FormI+Imperf+Act+**Una**+Indic+3P+Masc+Pl	[**ktb**&**CoCuC**]	VI1-AMP--3	verb of the 1st stem in indicative, active, masculine plural, 3rd person
-hA	**hA**+Pron+DO+3P+Fem+Sg	**_hA**	NZ---FS4-3	pronoun, feminine singular, accusative, 3rd person
.sabA.ha	[**SbH**&**CaCAC**]+Noun+**a**+Def+Acc	[**SbH**&**CaCAC**]	NO---MS4--	non-derived noun, masculine singular in accusative
al-.gadi	**{al**+Art+**gad**+Stem+**i**+Def+Gen	**gad**	NO---MS2D-	non-derived noun, masculine singular in genitive, prefixed definite article

Table 6: Review of the approach to morphological analysis. Three input strings **fsyktbwnhA SbAH Algd** are disambiguated into six tokens, literally so will write [pl.m.an.] it/them [sg.f.] morning [acc.] the-tomorrow [gen.]. In the Table, tokens use ArabTeX notation, lemmas keep to Buckwalter's style. Positional tags are shown and discussed next. Some dictionaries associate words like .gaduN tomorrow with the root&pattern lemma (in our case [**gdw**&**CaCxX**]) rather than just isolating the stem (i.e. **gad**).

Unfortunately, neither of our analyzers declares clearly what a lemma is. Having so far defined lemmas equal to stems, we lack means to unite singulars and broken plurals (mismatching patterns) under one lexical unit, unless we duplicate the work of an analyzer and build some coupling lexicons. This, though not so desperate, kind of problem comes with perfect and imperfect patterns of a verb, too. Tim Buckwalter's Analyzer is however expected to undergo improvements in this regard.

As to the format of a tag, the XAMA-like output is intelligible, but yet somewhat ineffective in terms of its automated processing. The information may be recorded as a bit vector in which mutually exclusive values of a morphological category map into a fixed position. The system of positional tags for Czech (cf. Hajič 2002) inspired a preliminary design of such a scheme for Arabic.

6 Selected Syntactic Structures

The principles which we follow on the analytical level are strongly influenced by the conclusions taken for the representation of Czech (cf., e.g., Böhmová et al. 2001). As both Czech and Arabic are languages with a rich inflection and a relatively free word order, many of the solutions for Czech are applicable also to Arabic. However, it is obvious that in Arabic, we will find phenomena that will not fit into the guidelines drawn in the Czech annotation manual. Some of these will be treated here, namely:

- non-verbal predication in Arabic,
- co-reference matching,
- verbal characteristics of certain nominal formations,
- figura etymologica.

6.1 Non-verbal Predication

Beside the standard type of predication expressed by a verb, Arabic possesses a number of other types of predication. This set of predication types is traditionally grouped under the heading of "nominal sentence". However, in this set, there are several other types of predication that do not fit easily under such a heading. Therefore, we will distinguish between verbal predication and other types which we label as non-verbal predication.

As predication expressed by a verb presents no crucial problems for the dependency type of syntactic representation, we will not treat it here. The main focus here will be the non-verbal predication which can be divided as follows:

- pure nominal sentence,
- "clausal" (conjunctional) predication,
- impersonal predication with a prepositional phrase (locative and possessive constructions),
- existential predication.

From the point of view of the dependency approach, a verb is the governing node of the sentence. As in these sentences no verb is used, we transfer this role to the highest node of the predicate, which then becomes the highest node of the sentence.

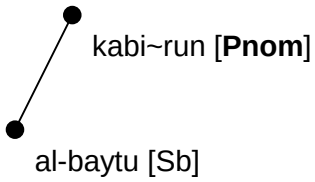
Such a solution is quite smooth and expectable in case of a nominal sentence (Example 1 on the next page). The nominal predicate (labelled as Pnom) without a verbal conjecture can be found in several other languages (e.g., Russian), and can easily take over the role of the governing node.

A slightly different picture occurs if the (nominal) predicate is represented by a clause (Example 2 therein), because in such a case, the sentence is governed by a conjunction as its highest node. Following the principle given above, it is the conjunction that has to receive the role of the predicate in the tree structure (labelled as PredC), while Pnom prepends to the particular function of the head of the clause.

(1) al-baytu kabi~run.

the-house big.

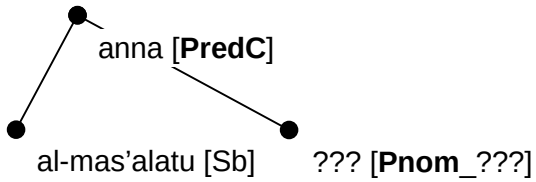
The house is big.



(2) al-mas'alatu anna ...

the-problem that …

The problem is that …



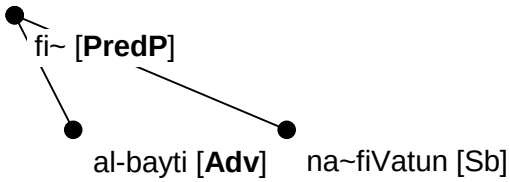
The next two types of sentences already have their traditional solution on the surface syntax level (especially in the Arabic environment), where the prepositional phrase could be perceived as the subject. However, in these cases, there is no explicit preference about the predicate, and a choice has to be made which part of the sentence will assume that role. For such cases, we have decided to suggest a solution that is closer to the underlying, tectogrammatical level.

The first type, impersonal predication with a prepositional phrase, can be illustrated by the two following sentences, expressing locative and possessive types of constructions:

(3) fi~ al-bayti na~fiVatun.

in the-house[gen.] window.

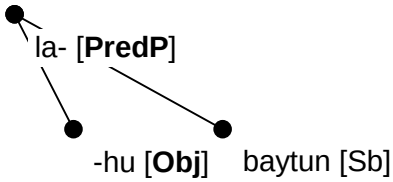
There is a window in the house.



(4) la- -hu baytun.

for him a-house[nom.].

He has a house.



As it has been pointed out before, there are voices according to which the first (prepositional) part of the sentence should be treated as subject and the second part (na~fiVatun, baytun) as predicate. However, there are other concepts that appeared in the linguistic theory. In deciding our own approach, we were inspired by the work by Freeze 1992, who argues that locative and possessive constructions have the same manifestation on deeper syntactic levels, and both these constructions are treated as predicative. When this point of view is adopted, we have to change also the manifestation on the surface level, where the roles of the parts of the sentence are exchanged and the role of the predicate is played by the prepositional phrase. Then, as the dependency governing in Arabic is respected, the preposition becomes the head of the predicate. In order to distinguish it properly from other sentence structures, we label it with PredP. The difference between the locative and possessive constructions is expressed by the function right after the preposition, as emphasized in Examples 3 and 4 above.

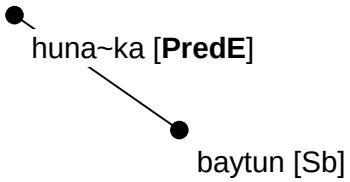
In spite of the fact that this solution can seem contradictory to the traditional approach, evidence in favour of the prepositional phrase as predicate is given already in one of the most classical and respected grammars (cf. Wright 1875, esp. 271–276) and found also recently (Moutaouakil 1989:87).

The other type, which we call existential predication, is represented by Examples 5 and 6. For Freeze (1992), there are two types of existential sentences on the surface level—those with a locative-phrase subject, and the proform existential. Arabic (which he also treated in his study) would fall into the second group, the proform being locative—both lexically and syntactically—and thus non-subjective.

(5) huna~ka baytun.

there a-house[nom.]

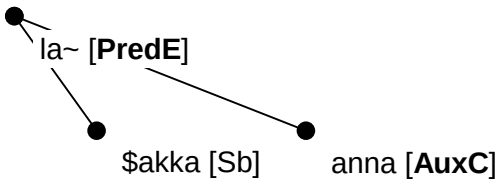
There is a house.



(6) la~ $akka anna ...

no doubt[acc.] that …

There is no doubt that ...



This led us to a solution which is in principle identical with the one outlined above, i.e. we ascribe the predicative role to the existential part of the sentence, which is huna~ka and la~ respectively. As this construction is somewhat different from the other types of predication, we label it with a somewhat different function PredE. Note that in Example 6, the head of the clause coming after anna would be annotated in the manner of Example 2.

6.2 Co-reference Matching

As in other languages, pronouns are not trivial to associate with the entity they represent. Resolving these bonds is important for true interpretation of the text on the tectogrammatical level. In annotation, linking the appropriate co-references is done by means of the so-called "lines across the graph", pointing from the pronouns to the expression being substituted.

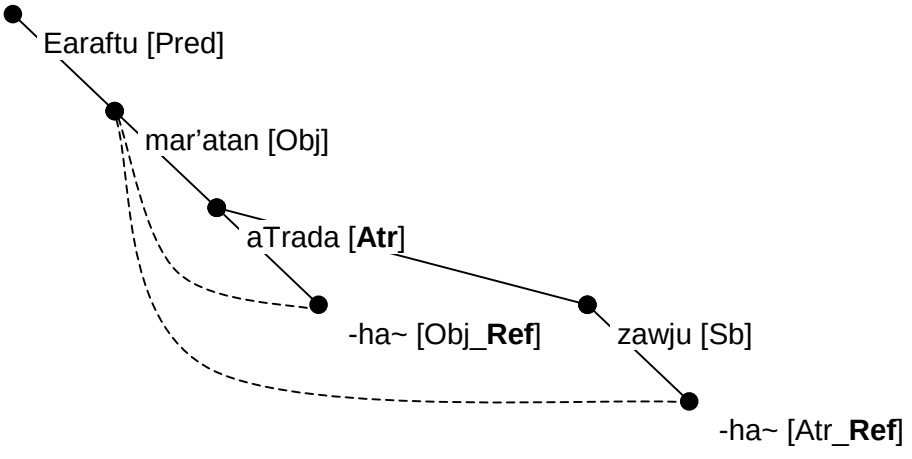
There are however pronouns for which their match need not be marked explicitly since it results clearly from the syntactic structure (i.e. cases of grammatical co-reference). In relative clauses, attributive pseudo-clauses or when anteposition takes place, it is enough to attach a suffix _Ref to the analytical function of such a pronoun, implying a certain algorithm shall be put in force to determine the node the referential corresponds to.

For example, in Arabic relative sentences, we find a very explicit expression of traces that are left after a movement of an element. A referential pronoun can be found at such places with the only exception when the movement concerns the subject of the relative clause. Our approach is shown in Example 7. Other kinds of traces and links between nodes are treated below in Sections 6.3 and 6.4.

(7) Earaftu mar'atan aTrada -ha~ zawju -ha~.

I-knew a-woman[acc.] chased-away her[acc.] husband[nom.] her[gen.].

I knew a woman whose husband chased her away.

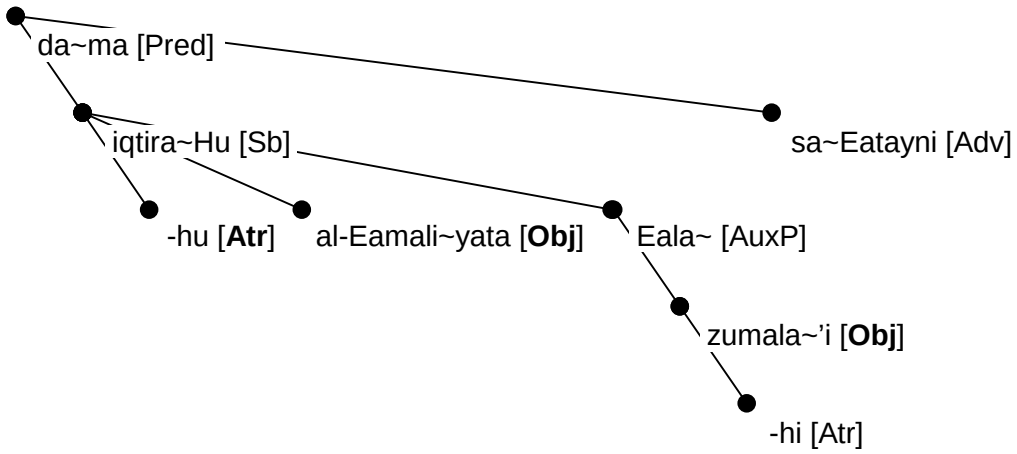


6.3 Verbal Characteristics of Certain Nominal Formations

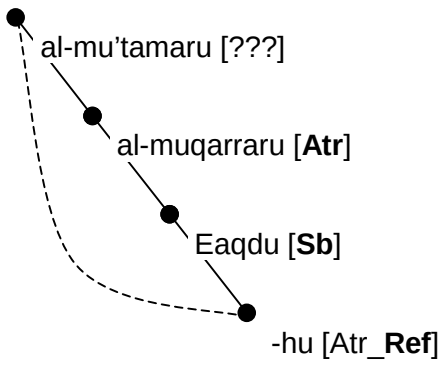
The Arabic word-derivational system is very heavily dependent on the verbal system. The purely nominal part of Arabic lexicon is relatively small and most words in Arabic are generated by the verbal system (known as verbal nouns and active/passive participles). These deverbatives can preserve both nominal and verbal syntactic attributes. This can sometimes lead to a special sort of constructions, where words traditionally classified as nominal in nature can exert a verbal type of government over some part of the sentence. This fact does not change the structure of the dependency tree on the analytical level, but substantially changes the usual picture of relations between the analytical functions and the morphological attributes of nodes. Below, we give some examples that cover the following types:

- verbal noun with predicative function (Example 8)
- participle with predicative function (Example 9)
- sequence of such nominal forms (Example 10)

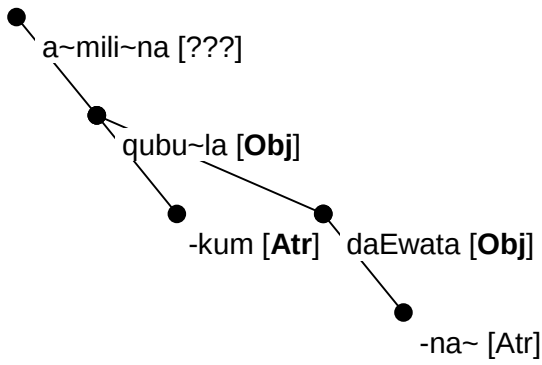
(8) da~ma iqtira~Hu -hu al-Eamali~yata Eala~ zumala~'i -hi sa~Eatayni.
lasted proposal his the-operation[acc.] on colleagues his two-hours[acc.].
It took two hours when he proposed the operation to his colleagues.



(9) al-mu'tamaru al-muqarraru
the-congress the-decided
Eaqdu -hu
convening its
congress whose convention is decided



(10) a~mili~na qubu~la -kum
hoping accepting[acc.] your
daEwata -na~
invitation[acc.] our
hoping that you will accept our invitation

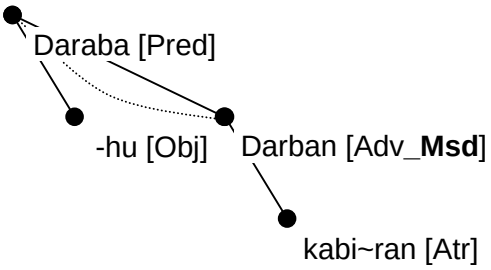


In Example 9, we find a non-clausal construction resulting from a transformation of a relative clause into a nominal attributive phrase, which we call an attributive pseudo-clause. The referential pronouns may be dealt with as if they occurred in a proper relative clause, though.

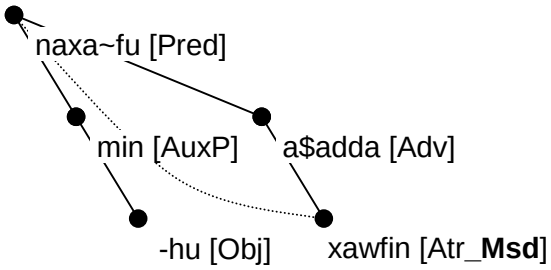
6.4 Figura Etymologica

As an intensifying construction, Arabic can (and often does so) use the so-called accusative of the inner object, where the verbal noun of the same root as the verb is used. The verbal noun can also be a part of an attributive phrase, and then replaces what might be an adverbial of mood in English.

(11) Daraba -hu Darban kabi~ran.
he-hit him hitting[acc.] big.
He gave him a big blow.



(12) naxa~fu min -hu a$adda xawfin.
we-fear from him/it strongest[acc.] fear.
We are afraid of him/it the most.



In such cases, there is a need to indicate the semantic adherence of a masdar (verbal noun) to its verb, in order to avoid possible mechanical translations ("He hit him big hitting."). Therefore, all the words will keep their usual positions in the tree as will their analytical functions, and the form of a masdar will receive a suffix _Msd to its analytical function. The "line across the graph" will go from the verbal noun to the upper-closest verb. Unlike co-reference matching, now the relation between the two nodes is semantic rather than syntactic.

7 Conclusion and Perspectives

This short overview cannot list all the problems and interesting cases which our team have encountered when working on the Prague Arabic Dependency Treebank. However, we hope that from the points mentioned here, one can get an idea of the issues dwelling in the description of Arabic on morphological and analytical levels.

Apart from the annotation procedure and the design of the guidelines for the tectogrammatical level, the team pursues problems concerning automation and pre-processing. These involve preparatory tree-building and function/functor assignment based on a set of observed rules, transformation of phrase-structure trees into dependency trees (cf. Žabokrtský and Kučerová 2002) or automated assignment of case and mood endings (in analogy to Žabokrtský et al. 2002).

8 Acknowledgements

The results presented here would not have been achieved without a fruitful collaboration of the following colleagues: Jan Hajič, Ivona Kučerová, Jan Šnaidauf, Ondřej Beránek, Petr Pajas, Monika Kolbová, Martin Špáta, Pavel Ťupek, Jiří Hana and Daniel Zeman.

The research reflected in this paper was supported by the Ministry of Education of the CR, projects LN00A063 and MSM113200006, and also by the Czech Grant Agency, GACR 405/02/0823.

References

Beesley, K. R. (1999): Arabic Stem Morphotactics via Finite-State Intersection. Benmamoun, E. (ed.): Perspectives on Arabic Linguistics XII. Papers from the Twelfth Annual Symposium on Arabic Linguistics. Benjamins, Amsterdam, pp. 85–99.

Beesley, K. R. (2001): Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. Association for Computational Linguistics. 39th Annual Meeting and 10th Conference of the European Chapter. Workshop Proceedings on Arabic Language Processing: Status and Prospects, July 6th 2001. CNRS – Institut de Recherche en Informatique de Toulouse, and Université des Sciences Sociales, Toulouse, France, pp. 1–8.

Böhmová, A. – Hajič, J. – Hajičová, E. – Hladká, B. (2001): The Prague Dependency Treebank: A Three-Level Annotation Scenario. Abeille, A. (ed.): Treebanks: Building and Using Syntactically Annotated Corpora. Kluwer Academic Publishers. In press.

Brants, T. – Chen, F. – Farahat, A. (2002): Arabic Document Topic Analysis. Proceedings of the Post Workshop of LREC 2002 on Arabic Language Resources and Evaluation: Status and Prospects. ELRA, Las Palmas de Gran Canaria.

Cavalli-Sforza, V. – Soudi, A. – Mitamura, T. (2000): Arabic Morphology Generation Using a Concatenative Strategy. Proceedings of NAACL 2000, Seattle, WA, pp. 86–93.

Faber, A. (1997): Genealogical Subgrouping of the Semitic Languages. Hetzron, R. (ed.): The Semitic Languages, Routledge, London, pp. 1–15.

Freeze, R. (1992): Existential and Other Locatives. Language 68, No. 3, 1992, pp. 553–595.

Hajič, J. (2002): Disambiguation of Rich Inflection (Computational Morphology of Czech). Habilitation Thesis, Charles University in Prague, Faculty of Mathematics and Physics. Karolinum, Charles University Press, Prague, 334 p.

Kiraz, G. A. (1998): Arabic Computational Morphology in the West. Proceedings of the 6th International Conference and Exhibition on Multi-lingual Computing, Cambridge. `http://www.bell-labs.com/project/tts/icemco-98.ps`

Kiraz, G. A. (1999): Computational Tool for Developing Morphophonological Models for Arabic. Benmamoun, E. (ed.): Perspectives on Arabic Linguistics XII. Papers from the Twelfth Annual Symposium on Arabic Linguistics, Benjamins, Amsterdam, pp. 101–110.

Kiraz, G. A. (2000): Multi-Tiered Nonlinear Morphology Using Multi-Tape Finite Automata: A Case Study on Syriac and Arabic. Computational Linguistics 26 (1), pp. 77–105.

Lagally, K. (1999): ArabTeX: A System for Typesetting Arabic. User Manual Version 3.09. Technical Report 1998/09. Institut für Informatik, Universität Stuttgart.

Maamouri, M. – Cieri, C. (2002): Resources for Natural Language Processing at the Linguistic Data Consortium. Proceedings of the International Symposium on Processing of Arabic, April 18th–20th 2002. University of Manouba, Tunisia, pp. 125–146.

McCarthy, J. (1985): Formal Problems in Semitic Phonology and Morphology, Outstanding Dissertations in Linguistics Series, Garland Publishing, New York, 430 p.

Moutaouakil, A. (1989): Pragmatic Functions in a Functional Grammar of Arabic. Dordrecht – Providence, Foris, 156 p.

Oard, D. W. – Gey, F. C. – Dorr, B. J. (2002): Evaluating Arabic Retrieval from English or French Queries: The TREC-2001 Cross-Language Information Retrieval Track. Proceedings of the Post Workshop of LREC 2002 on Arabic Language Resources and Evaluation: Status and Prospects. ELRA, Las Palmas de Gran Canaria.

Sawaf, H. – Zaplo, J. – Ney, H. (2001): Statistical Classification Methods for Arabic News Articles. Association for Computational Linguistics. 39th Annual Meeting and 10th Conference of the European Chapter. Workshop Proceedings on Arabic Language Processing: Status and Prospects, July 6th 2001. CNRS – Institut de Recherche en Informatique de Toulouse, and Université des Sciences Sociales, Toulouse, France, pp. 127–132.

Smrž, O. – Šnaidauf, J. – Zemánek, P. (2002): Prague Dependency Treebank for Arabic: Multi-Level Annotation of Arabic Corpus. Proceedings of the International Symposium on Processing of Arabic, April 18th–20th 2002. University of Manouba, Tunisia, pp. 147–155.

Soudi, A. – Cavalli-Sforza, V. – Jamari, A. (2001): A Computational Lexeme-Based Treatement of Arabic Morphology. Association for Computational Linguistics. 39th Annual Meeting and 10th Conference of the European Chapter. Workshop Proceedings on Arabic Language Processing: Status and Prospects, July 6th 2001. CNRS – Institut de Recherche en Informatique de Toulouse, and Université des Sciences Sociales, Toulouse, France, pp. 155–162.

Wehr, H. (1974): A Dictionary of Modern Written Arabic. Arabic-English. Wiesbaden, Harrassowitz, 1110 p.

Wright, W. (1875): A Grammar of the Arabic Language. Vol. II. London, F. Norgate, 483 p.

Zemánek, P. (2001): CLARA (Corpus Linguae Arabicae): An Overview. Association for Computational Linguistics. 39th Annual Meeting and 10th Conference of the European Chapter. Workshop Proceedings on Arabic Language Processing: Status and Prospects, July 6th 2001. CNRS – Institut de Recherche en Informatique de Toulouse, and Université des Sciences Sociales, Toulouse, France, pp. 111–112.

Žabokrtský, Z. – Kučerová, I. (2002): Transforming Penn Treebank Phrase Trees into (Praguian) Tectogrammatical Dependency Trees. PBML 78, Charles University, Prague.

Žabokrtský, Z. – Sgall, P. – Džeroski, S. (2002): A Machine Learning Approach to Automatic Functor Assignment in the Prague Dependency Treebank. Proceedings of LREC 2002, ELRA, Las Palmas de Gran Canaria, pp. 1513–1520.