Prague Dependency Treebank for Arabic: Multi-level Annotation of Arabic Corpus

Otakar Smrž, Jan Šnaidauf, Petr Zemánek[1]
Charles University, Prague, Czech Republic

Abstract

This contribution describes a project of Multi-level annotation of a corpus of Arabic with the following levels of annotation: morphological, analytical (surface syntax) and tectogrammatical (roughly, deep syntax). Here, a description of the procedures of annotation at respective levels is given. Major attention is given to the analytical level, where a number of examples of surface dependency relations in Arabic sentences is given.

1 Introduction

The project of Multi-level Annotation of Arabic Corpus is, in larger scale, a part of the research concerning machine-translation from Arabic to English. The activity has been initiated by the linguistics departments of Charles University (Czech Republic) and Johns Hopkins University (USA). Viewed stand-alone, multi-level annotation of language corpora gives structure to the data, and thus results in an invaluable resource of linguistic information.

The general aim of the work in progress is to prepare the representation of Arabic sharing the conception of the Prague Dependency Treebank for Czech (PDT, [1,2]), where the nature of the language was studied in three levels. The lowest level comprises a full morphological annotation (morphological level), the next level deals with surface syntactic annotation using dependency syntax (analytical level). The highest level of annotation is the level of linguistic meaning, called the tectogrammatical level. The linguistic concept is based on a dependency approach, where the verbal part of the sentence plays a major role. A significant part of the whole corpus will be annotated manually, and on this part, statistically-based annotation tools will be trained.

The team consists of the following members: Ondřej Beránek, Monika Kolbová, Ivona Kučerová, Otakar Smrž, Jan Šnaidauf, Martin Špáta, Pavel Ťupek and Petr Zemánek.

In this contribution, we comment on the corpus data retrieval and collection, procedures of annotation at the morphological, analytical and tectogrammatical levels. At the end, an appendix of an alternative representation of Arabic auxiliary verbs is given.

2 Corpus Data Retrieval/Collection

For the purposes of the project, an Arabic Corpus of the language of economics, business, law and industry has been created. We took advantage of the currently available corpora, CLARA and LDC Arabic Newswire A Corpus, and performed machine-assisted extraction of the relevant data. The resultant corpus counts about 12,000,000 words.

CLARA (Corpus Linguae Arabicae) is a topically classified corpus of Modern Standard Arabic which has been compiled at the Institute of Near Eastern Studies, Charles University. (c.f. [7]). CLARA contains roughly 50,000,000 words, out of which 9,000,000 represent the written language of economics, business and finance. Data have been collected from a series of newspapers and other printed resources from the second half of the nineties.

The Arabic Newswire A Corpus consists of articles from the Agence France Presse (AFP) Arabic Newswire, which date from May 1994 to December 2000. The corpus has been created by Linguistic Data Consortium, University of Pennsylvania, and its size is almost 80,000,000 words. The newswire

[1] smrz@ufal.mff.cuni.cz, snaidauf@yahoo.com, petr.zemanek@ff.cuni.cz

transcripts provide another style of Arabic, which is closer to a spoken language. The documents are however not organized with respect to topic, and a special analysis had to be carried out to extract the data of our interest. Subcorpora of CLARA were utilized for modelling the particular language domain, and every single document of LDC Newswire Corpus was tested against each of the models. The best scoring candidates were extracted into preliminary collections, contents of which were examined by humans. Incident inadequate articles were filtered out.

The computational method used was determined by the size of the data in question. The articles to classify contained only 200 words on average, spanning say from 50 to 500 words. Occurrence histograms of unigrams were taken as the criterion, since sensible bigram or trigram models could hardly be constructed on such poor samples.

Correlation coefficient between the histogram of every article and the histograms of the subcorpora was computed. The higher the result, the better the conformance of the sample to the model.[2]

More technical details about the extraction are to appear in (Prague Bulletin of Mathematical Linguistics, Autumn 2002). See also other studies on LDC Arabic Newswire Corpus A classification (of different motivation and results, however - [6]).

The Arabic Corpus for the Annotation Project was compiled from the relevant subcorpora of CLARA (~9,000,000) and from our extraction of LDC Newswire Corpus (~3,000,000). This rate has been reasoned by word per form ratio indicators (~45 and ~140 respectively).

The data manager is Pavel Ťupek with assistence from Monika Kolbová and Martin Špáta, additional works on this part of the project are carried out also by Otakar Smrž and Petr Zemánek.

3 Morphological level of annotation

On this level, a standard morphological analysis is being done. Here, also the complexity of Arabic flexional system has to be dealt with, as well as the vast ambiguity caused by the graphemic representation of Arabic. Besides, the strings in Arabic script can contain more words in one string (usually an autosemantic word together with clitics like particles, conjunctions and suffixed pronouns), where at least some of them have to be analyzed for morphotactical boundaries. E.g., a string *fsyktbwnh* is read as *fa-sa-yaktubu~na-hu*, which can be analyzed as *and/then-FUT-they will write[masc.]-it*. In other words, a morphological analysis of Arabic has to be accompanied by morphological disambiguation and tokenization.

The morphological analysis is being done semi-automatically. Each string of Arabic characters is assigned a value by a morphological analyzer, which also provides possible solutions for tokenization (morphotactical analysis). The analyzer produces a suggestion of lemmas and roots and a combination of values of individual morphological categories.

Once tokenization and morphological disambiguation is done, the respective morphological information refering to each token can be transformed to a fine-grained "positional tag", where exclusive values of a category map to the reserved position. This notation is useful for its fix length and tabularity, and easy machine interpretation.

With respect to Arabic morphology, the structure consisting of 15 slots can be designed. We distinguish such categories, as tense, mood, gender, number, case, definiteness, voice, person, etc., together with the underlying template and explicit root (up to five radicals). The system has some inner conditioning - in some cases, the meaning of a "letter" in a position is correctly read only together with some preceding tag. This means that e.g. a P on the second position is to be read as Passive Participle in case of N (Noun) as a preceding tag, and as Perfect after V (Verb). In this contribution, we do not provide the complete system for readibility reasons. The following shows only the basic principles of tag construction.

---

[2] This value may be interpreted equivalently in terms of vector analysis. Imagine that N-dimensional space gets generated by N word-forms. The histograms of the sample and the models will then be rendered as vectors of this space. The best candidates are those pointing in similar directions as the desired models do. Cosine of the difference of directions is equal to the correlation coeficient.

Table 1: Example of first slots in a positional tag

Position	Previous Tag	Category	Value	Meaning
1	no condition	Part of speech	N	Noun
			V	Verb
			P	Particle
2	1: N	Primitive / Derived	0	Primitive
			A	Active Participle
			P	Passive Participle
			M	Masdar
			L	Location or Time
			T	Tools
			E	Elative
			C	Color
	1: V	Tense & Mood	P	Perfect
			I	Imperfect Indicative
			S	Imperfect Subjunctive
			J	Imperfect Jussive
			R	Imperative
			E	Energicus '-an'
			N	Energicus '-anna'

Such data are then processed by a team of morphology annotators, who choose the appropriate combination of lemma and tag in a given context (disambiguation). We use the two-level FST system developed by Xerox (Xerox Arabic Morphological Analyzer - cf. e.g. [5]), alternatively we will use the morphological analyzer being developed by the LDC.

The manual disambiguization will be done on standard sized subcorpus (1,500,000 – 2,000,000 words). On these data, statistical analyzers will be trained. This procedure will be applied also to other levels of annotation. The manual annotation is carried out by Monika Kolbová and Martin Špáta under the supervision of Otakar Smrž and Petr Zemánek.

For the manual annotation we use the DA (DisAmbiguator) editor, which has been developed at the Faculty of Mathematics and Physics, Charles University, by Jiří Hana.

4 Analytical level of annotation

At present, the analytical level represents the main level for structural annotation. Here, linear annotation is abandoned, where each word was taken separately irrespective of context, and it is the structure of the sentence that is introduced into the annotation of the text (however, not the structure of the text). All the original words of the text are preserved and obtain their functions in the resulting structure.

This level is based on the dependency approach, where the syntactic relations between “words” (after morphotactical analysis) in a sentence are represented. The structure is based on the dependency relation between a governor and its dependent node. A type of dependency represents one of the attributes of each node and is, understandably, oriented "upwards", i.e., towards its "governing" node. Each word and punctuation mark has its node, these nodes are connected by edges (arcs) to form a single (acyclic) tree graph. No other nodes are added, with the exception of the "master" sentence node (AuxS).

This level of annotation is in fact rather a technical level, covering the surface level of syntax, i.e. a level which covers functions of actual words as they appear in a surface shape of a sentence. Each word (including synsemantic words) is assigned a main syntactic function ('analytic functor', 'Afun'); an example of such functions are Sb, Obj, Adv, etc. (standing for Subject, Object and Adverb respectively).

On the analytical level, it is necessary to work with morphotactically analyzed data. This is different from the case of Czech, where both types of annotation ran simultaneously and were merged later.

The morphologically disambiguated data are processed by a team of syntax annotators. The annotation is done mostly manually up to the size which makes it possible to use statistically based annotation tools.

For this level of analysis, we had to develop a systematic specification which applies the principles of dependency syntax to Arabic. These guidelines have to deal with problems specific to Arabic, especially the nominal sentence without a copula or auxiliary verb, the preferable parent node for representation of dependency relations. These guidelines have been prepared by a team consisting of Ondřej Beránek, Ivona Kučerová, Otakar Smrž, Jan Šnaidauf and Petr Zemánek. Ondřej Beránek and Jan Šnaidauf also carry out the manual annotation of this level of analysis.

For the manual annotation, TrEd (Tree Editor) is used. The system has been developed at the Faculty of Mathematics and Physics, Charles University, by Petr Pajas. TrEd interface has been extended for our convenience, and fully complies with the right-to-left Arabic script (CP 1256).

5 Model Sentences Analysis

In the following, a few examples illustrating our approach will be given. In the first part, there is a list of analytical functions used in the examples, the second part offers analysis of model sentences. In the examples, we use a slightly adapted transliteration system developed by Tim Buckwalter, which is also being used by the Xerox Arabic Morphological Analyzer[3].

5.1 List of analytical functions:

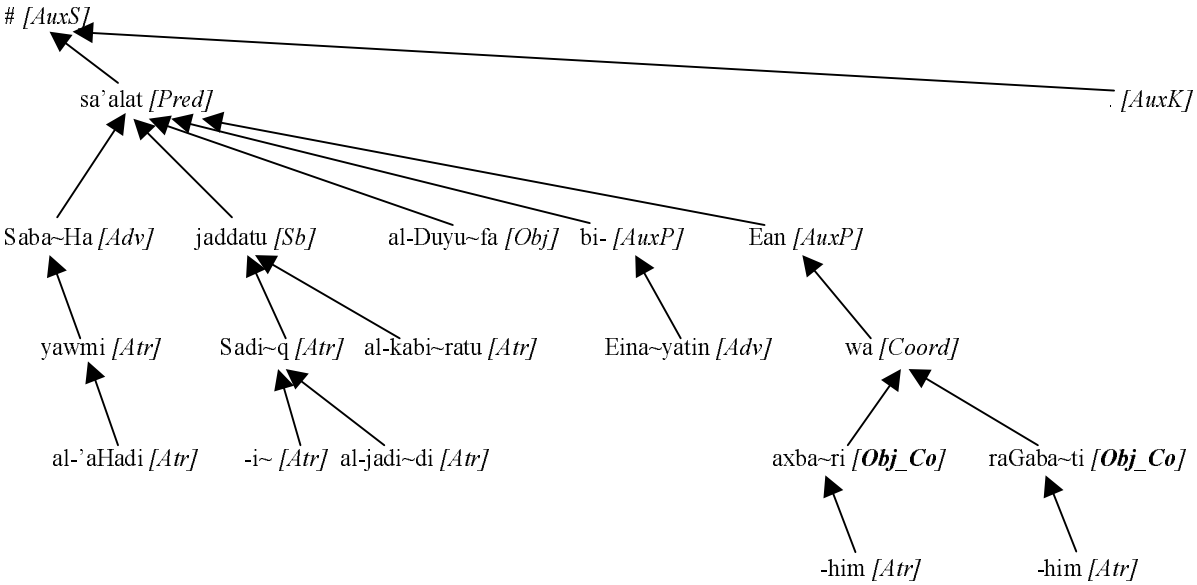
Table 2: List of analytical functions

Function	Description
Pred	Predicate
PPred	Predicative preposition
Pnom	Nominal part of predicate
Sb	Subject
Obj	Object
Adv	Adverb
Atv	Complement
Ante	Sentence member in anteposition
Coord	Coordination
AuxP	Preposition
AuxV	Auxiliary verb
AuxM	Particles modifying the verb
AuxC	Conjunction
AuxE	Emphasis
AuxY	Others
AuxS	Sentence beginning
AuxK	Sentence end
_Co	Member of coordination
_An	Reference to anteposition
_Ms	Masdar adverb

[3] The Buckwalter's system has been preserved with the exception of the two following changes: "G" stands for *ghain* (instead of "g") and "V" stands for *tha~* (instead of "v"). The system is available e.g. at http://www.xrce.xerox.com/research/mltt/arabic/info/buckwalter-about.html

5.2 Verbal sentence

Saba~Ha yawmi al-'aHadi sa'alat jaddatu Sadi~q-i~ al-jadi~di al-kabi~ratu al-Duyu~fa bi-Eina~yatin Ean raGaba~ti-him wa axba~ri-him. – Sunday morning, the old grandmother of my new friend asked the guests with concern about their news and their wishes.

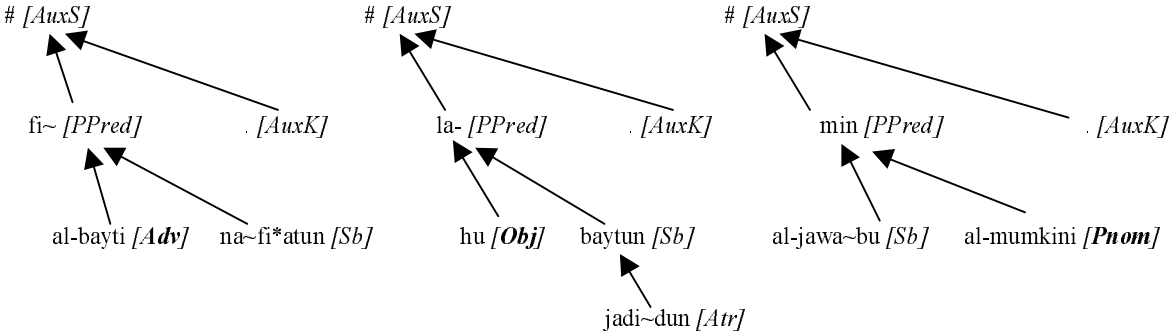


Comments: Note the separation of suffixed pronouns and the way coordination is used. Furthermore, prepositional phrases resulting from verbal government are considered indirect objects.

5.3 Nominal sentence

5.3.1 Prepositional nominal sentence (present tense)

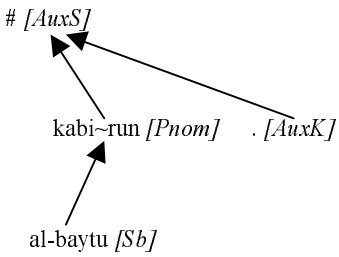
a) *fi~ al-bayti na~fi*atun* – In the house there is a window.
b) *la-hu baytun jadi~dun* – He has a new house.
c) *al-jawa~bu min al-mumkini* – It is possible to reply.

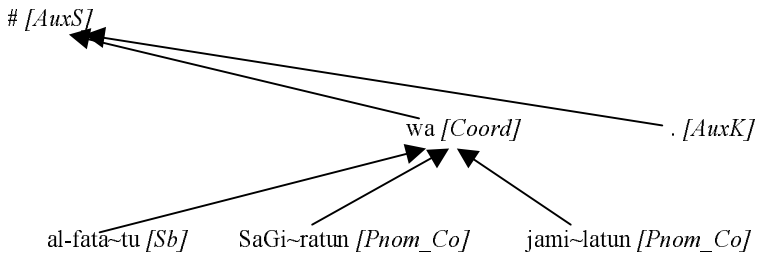


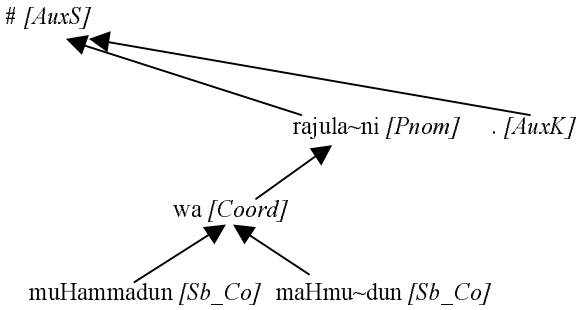
Comments: Note the different analytical functions of second-part expressions in prepositional phrases with a predicative preposition.

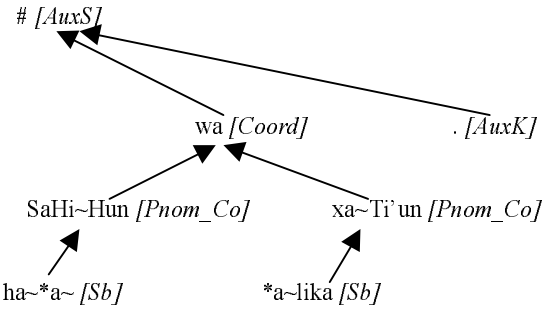
5.3.2 Pure nominal sentence (present tense)

a) *al-baytu kabi~run* – The house is big.
b) *al-fata~tu SaGi~ratun wa jami~latun* – The girl is young and nice.
c) *muHammadun wa maHmu~dun rajula~ni* – Muhammad and Mahmud are men.
d) *ha~*a~ SaHi~Hun wa *a~lika xa~Ti'un* - This is right and that is false.









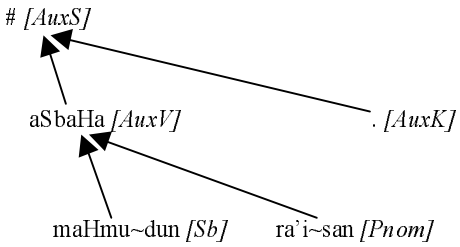
Comments: In order to treat these four examples consistently, only the given type of analysis was found acceptable (i.e. the nominal part of predicate governing the subject in the tree). Thus, it is considered technically necessary (no empty nodes can be added) while not exactly rendering the relations in a nominal sentence.

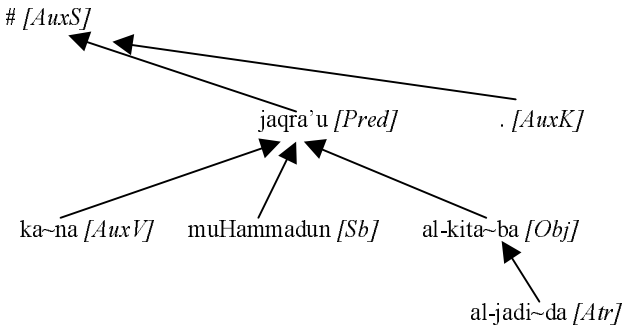
5.3.3 Nominal sentences in the past – see the next section (Sisters of ka~na)

5.4 Sisters of ka~na and compound verbs

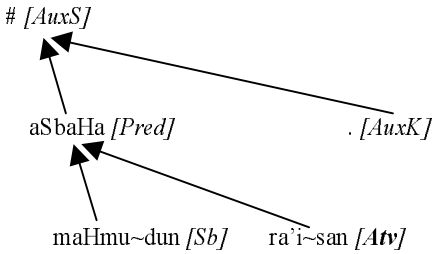
a) *aSbaHa maHmu~dun ra'i~san* – Mahmud became director
b) *ka~na muHammadun jaqra'u al-kita~ba al-jadi~da* – Muhammad was reading the new book.

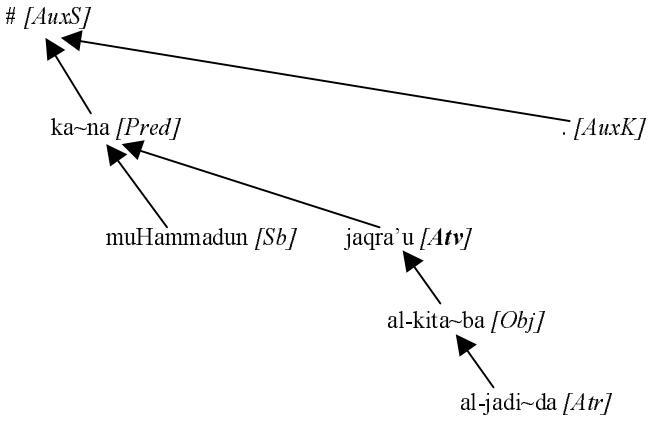
5.4.1 The traditional view





5.4.2 A new proposal for an alternative solution

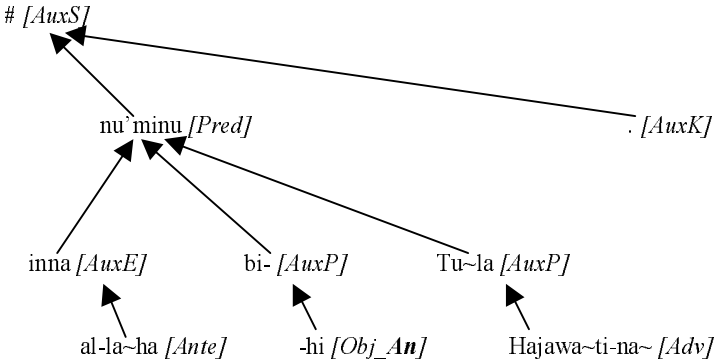




Comments: For details on this approach to sisters of ka~na and compound verbs see the **Appendix**. There you will find a study treating this problem both from structural and semantic viewpoints. An effort is made to integrate these seemingly different verbal classes within *one single system*.

5.5 Sentence elements in anteposition

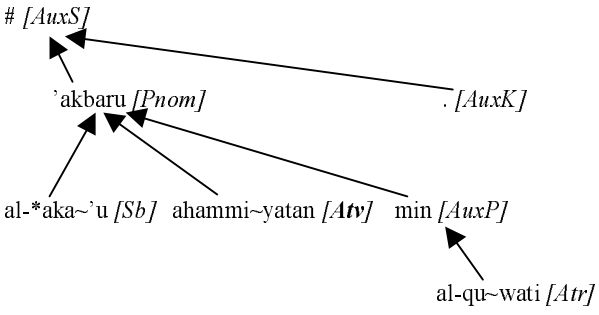
inna al-la~ha nu'minu bi-hi Tu~la Hajawa~ti-na~ – In God we believe throughout our lives.

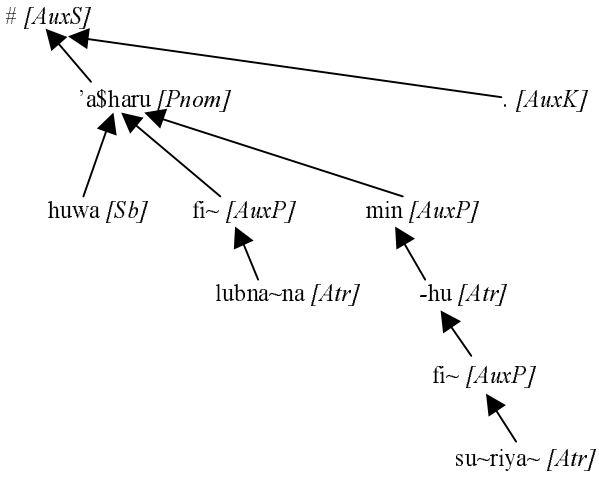


Comments: Note the function suffix used to mark the referencing pronoun.

5.6 Comparison

a) *al-*aka~'u akbaru ahammi~yatan min al-qu~wati.* Intelligence is more important than strength.
b) *huwa a$haru fi~ lubna~na min-hu fi~ su~riya~.* He is more famous in Lebanon then in Syria.

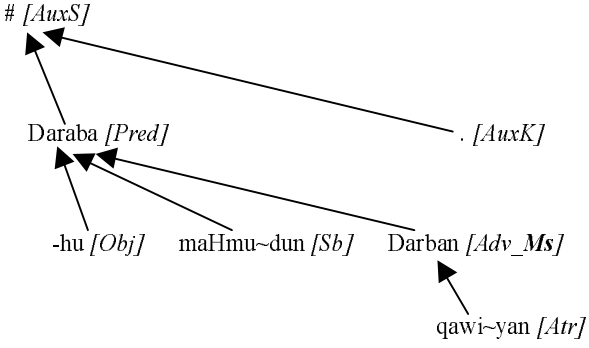


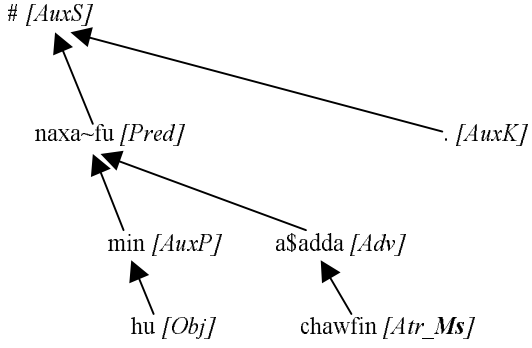


Comments: Look at the composition of comparative structures: The defining complement (*ahammi~yatan*) as well as the expression being compared to (*min…*) are both linked directly to the elative.

5.7 Accusative of the inner object (figura etymologica)

a) *Daraba-hu maHmu~dun Darban qawi~yan* – Mahmud hit him strongly.
b) *naxa~fu minhu a$adda chawfin* – We fear him most.

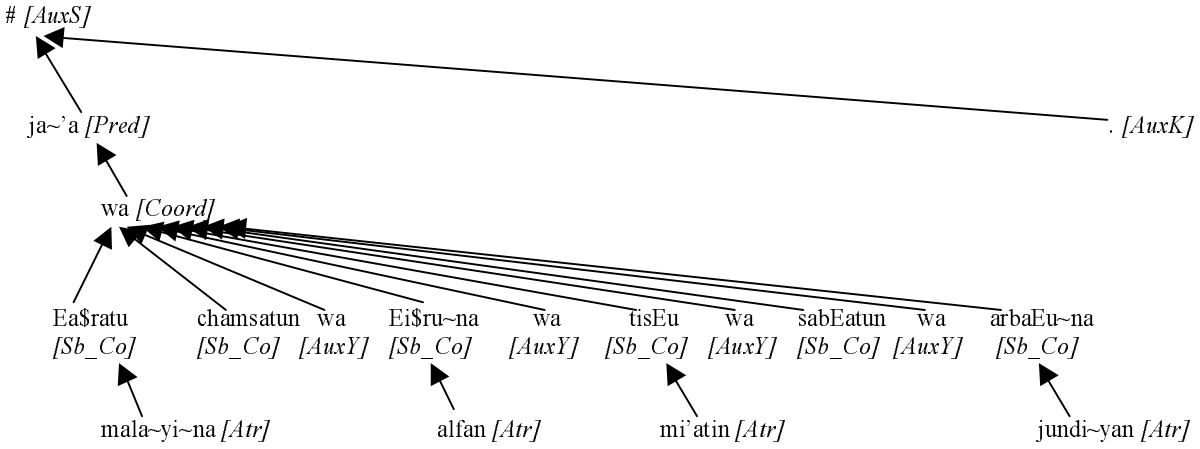




Comments: Note the use of the function suffix to indicate a masdar adverbial form. This so-called inner object in accusative is semantically empty, thus requires a marking in order to recognize its grammatical connection with the verb.

5.8 Numerals

Ea$ratu mala~yi~na wa chamsatun wa Ei$ru~na alfan wa tisEu mi'atin wa sabEatun wa arbaEu~na jundi~yan – 10,025,947 soldiers



Comments: The structure of numerals is easy to understand.

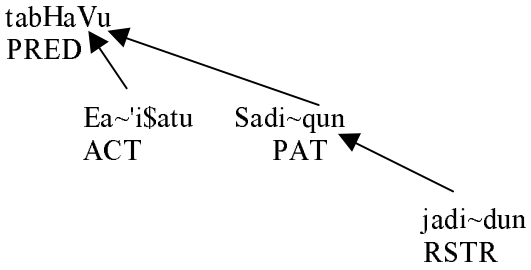
The examples given above are only to illustrate our approach to specific and particular problems of the dependency representation of Arabic. A full and systematic description of our approach will be given in a separate publication.

6 Tectogrammatical level of annotation

This level describes the linguistic meaning of a sentence, derived from the analytical level. The Functional Generative Perspective [3, 4] is applied. It is the highest (underlying) level of annotation. A tectogrammatical sentence representation prototypically has the same structure as the analytical representation; however, in specific cases some nodes may be eliminated, some added, so that the tectogrammatical structure of a sentence can differ from that at the analytical level. Basically, here only autosemantic words have a node of their own, while the correlates of function words are attached as indices to the words to which they "belong" (i.e., auxiliary verbs and subordinating conjunctions to verbs, prepositions to nouns, etc.). Analytic functions are substituted by tectogrammatical functions (such as Actor/Bearer, Patient, Addressee, etc.). This means that this level corresponds to such concepts as "deep syntax", etc.

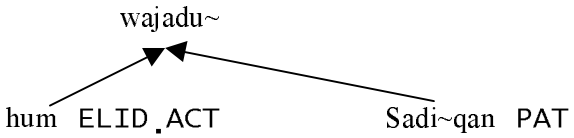
At the time when this contribution was written, this level has not been dealt with in greater detail. Here, we will present only a few examples which should demonstrate the nature of this level.

a) an example of standard tectogrammatical analysis
tabHaVu Ea~'i$atu Ean Sadi~qin jadi~din
Aisha is looking for a new friend.



Comment: note the elision of the preposition, which is in a given example not necessary for understanding the sentence. PRED stands for Predicate, ACT for Actor, PAT for Patient, RSTR for Restriction.

b) addition of a node
wajadu~ Sadi~qan
they found a friend



Comment: A new node has been added, which explicitly expresses the Actor of the sentence. ELID.ACT stands for Elided Actor.

7 Conclusion

This paper described the procedures in construction of a richly annotated corpus of Arabic. While the morphological analysis is a standard one, in the syntactic representation we chose an approach, which is different from commonly used methodologies, but which proved as computationally interesting (cf. the Prague Dependency Treebank for Czech, [1, 2]).

References

[1] Böhmová, Alena – Hajič, Jan – Hajičová, Eva – Hladká, Barbora: The Prague Dependency Treebank: A Three-Level Annotation Scenario. In print.

[2] Prague Dependency Treebank (PDT). http://ufal.ms.mff.cuni.cz/pdt/index.html.

[3] Sgall, Petr – Hajičová, Eva – Panevová, Jarmila: The Meaning of the Sentence and Its Semantic and Pragmatic Aspects. Reidel Publishing Company, Dordrecht, Academia, Prague, 1986.

[4] Sgall, Petr: Underlying structure of sentences and its relations to semantics. In: Festschrift für Viktor Jul´evič Rozencvejg zum 80. Geburtstag. Hrsg. von Tilmann Reuther. Wiener Slawistischer Almanach, Sonderband 33, Wien 1992, p. 273-282.

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

[6] Sawaf, Hassan - Zaplo, Jörg - Ney, Hermann: Statistical Classification Methods for Arabic News Articles. In: Association for Computational Linguistics. 39th Annual Meeting and 10th Conference of the European Chapter. Workshop Proceedings: Arabic Language Processing: Status and Prospects. July 6th 2001. CNRS - Institut de Recherche en Informatique de Toulouse, and Universite des Sciences Sociales, Toulouse, France, pp. 127-132.

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

APPENDIX

Verbal structures on the analytical level (Sisters of ka~na" & compound verbs)

Jan Šnaidauf

1 Introduction

This provisional working paper should be regarded as a groundwork for discussion.

There is a thesis that I will try to promote in this study: the equality of Arabic verbs within one single verbal system. We will observe some vital characteristics of different verb samples, intending to demonstrate that all of them, whether sisters of ka~na or non-sisters, transitive or intransitive, are parts of the same system. It means – in other words – that there is no need defining a special class for 'auxiliary' verbs such as 'sisters of ka~na'.

Note: My brief analysis of this traditional issue is to the greater part based on a mutual comparison of the seven following model sentences. Therefore, mistakes due to inappropriate or insufficient choice of these models could not be excluded. For transliteration and function tables see above.

2 Complements of simple clauses

Model statements:

1. *ra'a~ muHammadun maHmu~dan [ka~tiban].*
   Muhammad saw Mahmud [writing].
2. *i$taGala muHammadun [muhandisan].*
   Muhammad worked [as an engineer].
3. *$aGGala maHmu~dun muHammadan [muhandisan].*
   Mahmud employed Muhammad [as an engineer].
4. *jaEala maHmu~dun muHammadan muhandisan*
   Mahmud made an engineer of Muhammad.
5. *uEtubira muHammadun muhandisan.*
   Muhammad was considered an engineer.
6. *aSbaHa muHammadun muhandisan.*
   Muhammad became an engineer.
7. *ka~na muHammadun muhandisan.*
   Muhammad was an engineer

Note: Brackets indicate where the complement is optional, i.e. where it may be omitted or replaced by a prepositional phrase (*ka-muhandisin*). In those cases, the sentence can be used without the complement, causing no harm to the grammatical correctness.

The following table (Table 1) lists analyses of the given verbs. **'Sisters of ka~na' is description of verbs that have a direct semantic relation to the fact of being (existence) while introducing a special circumstantial aspect** (he became = he had *not been*, then he *was* – the circumstance is the process of achieving being/existence)**.** Thus, the complement is <u>required</u> as the target i.e. the final state of the circumstance.

Other verbs, however, do not require a complement since there is no targeted circumstance. The complement in such cases merely extends the meaning of the verb, referring even here to an existence (he worked and/while he was an engineer). The optional complement may also be in Arabic expressed by a prepositional phrase (*i$taGala ka-muhandisin*).

Note: Surprisingly, the type of the complement in transitive sentences does not depend on whether the verb is a "sister" or a "non-sister", but rather on the source of transitiveness. Where transitiveness results from the substantial meaning (like seeing), it seems to presume a subject complement. On the other hand, where circumstance is the cause, we observe an object complement (like employ, consider, make sb. do etc.).

Table 1: Overview of verbal features

	Semantic core	Aspect (circumstance)	Transitivity	Voice	Complement is obligatory
ra'a~	seeing [1]	none	yes	active	no
i$taGala	working	none	no	active	no
$aGGala	working	factitive [2]	yes	active	no
jaEala	being	factitive [2]	yes	active	yes
uEtubira	being	estimative [2]	yes	passive	yes
aSbaHa	being	resultative	no	active	yes
ka~na	being	none (duration)	no	active	yes

[1] an interactive meaning [2] an interactive circumstance – both resulting in transitivity

Note to Table 1: Verbal transitivity might either be an effect of the verb's interactive meaning or its interactive circumstance. In the first case, the action itself presumes a number of participants (usually two, e.g. someone seeing something), whereas in the other one the transitivity is caused by relating the action towards someone or somebody in a special way (consider someone etc.).

3 Compound verbs

Table 2: Principles of building compound verbs

	Given forms	Semantic core	Aspect (circumstance)	Tense	Meaning
Input	*aSbaHa* + *ja$taGilu*	being + working	resultative + none	finished past + unfinished present	he became + he works
	⇩	⇩	⇩	⇩	⇩
Transformations	*aSbaHa* + *ja$taGilu*	['being' exercises no semantic influence] [1] + working	resultative + [0]	[refers to the aspect/circumstance] + [refers to the aim of the aspect, is not essential] [2]	⇩
	⇩	⇩	⇩	⇩	⇩
Output	***aSbaHa ja$taGilu***	working	resultative	finished past [overall tense]	he began to work

[1] Like in English: being writing ⇔ writing [2] in English the inner verb is represented either by an infinitive or a gerund form (he started reading / to read) which substantially lacks any temporal aspect, while the finite forms of Arabic do lose it

Table 2 shows how individual features of the original verbs participate in defining the characteristics of the final one. The first part of a compound verb is generally one of the "sisters of ka~na", which in the end transfers its circumstantial aspect to the second member of the pair. It is also the carrier of the tense, while the second part appears in the tense-neutral form of imperfect indicative.

On the syntactical level, **the mechanism works in exactly the same way** as it does with nominal complements (e.g. he started working = he was NOT working, then he was – once again, the matter is the process of achieving existence for working).

The next question should then be, whether the Arabic grammar sustains a construction, where the 1st part of a compound verb is **not** a sister of ka~na i.e. is loaded with a full-scope semantic core? If so, is it only the circumstantial aspect which is always transferred onto the second-position verb or is it also the meaning?

The greatest example for such a case are the well-known verbs of perception, which incidentally show some specific behavior in English too.

Table 3: Principles of building compound verbs – verbs of perception

	Given forms	Semantic core	Aspect (circumstance)	Meaning
Input	*ra'a~(ha~)* + *ta$taGilu*	seeing + working	none + none	he saw (her) + she works
	⇩	⇩	⇩	⇩
Output	***ra'a~(ha~) ta$taGilu***	seeing working	[0]	he saw her work

Since the sustainability of the system seems to be sufficient, we may even dare to move one step further and extend the question: Are these constructions possible even apart from the particular case of "sensorial" verbs?

A simple test may be done using two of the sentences from our list:

$aGGalaha~ muHammadun taktubu.
Muhammad charged her with writing.
(Muhammad gave her work + she writes)

Should this construction be accepted by a native speaker of Arabic then it would be the final evidence not only that sisters of ka~na are not principally different then other verbs. Moreover, it would deny any differences at all.

4 Representation on the analytical level

It has been shown that there is a structural homogeneity among all sentence complements, whether they are attached to sisters of *ka~na* or non-sisters, transitive or intransitive. As explained in part II., this also matches the compound verbs, since they differ neither syntactically nor semantically from the usual structure 'verb ⇔ nominal complement'.

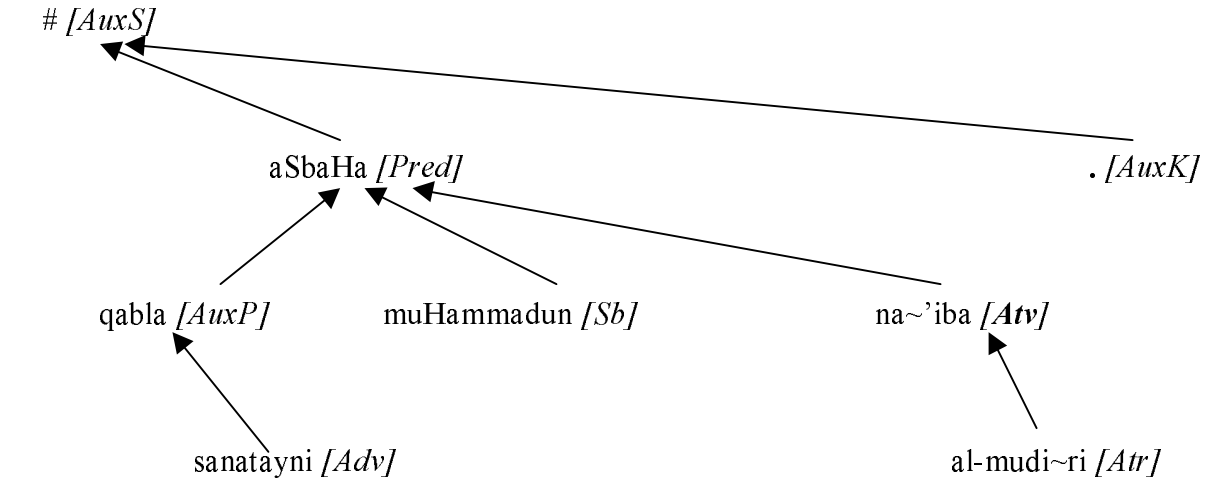
The logical consequence of this approach is that a **compound verb structure corresponds to a hypotactical compound sentence**. So the second part of the compound verb becomes the top node of a new sentence tree which takes place in the superior (main) sentence as the complement.

A general rule concerning paratactical compound sentences says that the top node of a subordinate clause lose its original analytic function (Pred) and be assigned that function, which the whole clause occupies in the frame of the superior sentence. Finally, this is what we needed to achieve, in order to maintain consistency of our model: the 2nd part of a compound verb, which may even be the top node of a large complex subordinate clause, assumes the same analytical function as an ordinary complement.

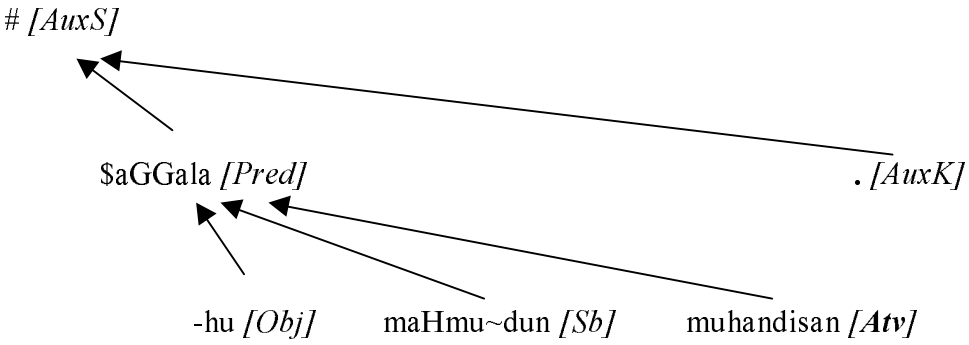
5 Examples

In our project, the description **Atv** has been chosen to indicate complements of any type.
The following figures show a few examples:

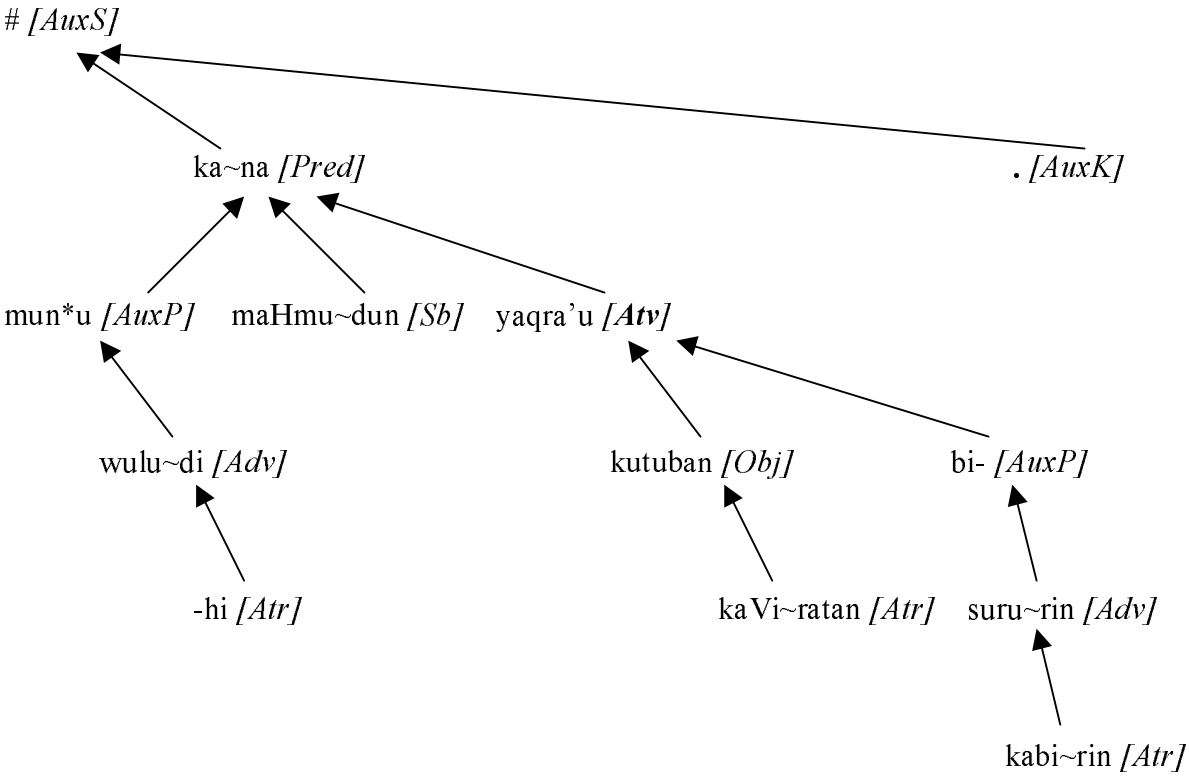
(1) *qabla sanatayni aSbaHa muHammadun na~'iba al-mudi~ri.*
Two years ago, Muhammad became a deputy manager



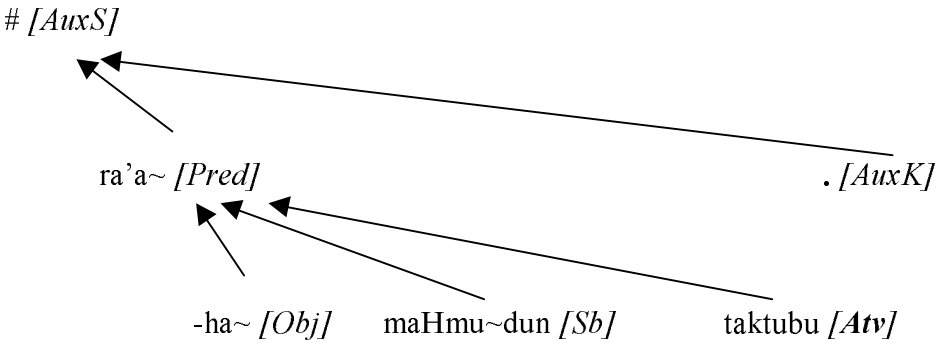
(2) *$aGGalahu maHmu~dun muhandisan.* - Mahmud employed him as an engineer.



(3) *mun*u wulu~dihi ka~na maHmu~dun yaqra'u kutuban kaVi~ratan bi-suru~rin kabi~rin.* Since his childhood Mahmud read many books with a great pleasure.



(4) *ra'a~aha~ maHmu~dun taktubu.* - Mahmud saw her write.



6 Conclusion

This study is a first attempt at an alternative view of structural representation of the so-called auxiliary verbs in Arabic. Its aim is to promote the affirmation that these verbs (the sisters of ka~na as well as other ones usually considered specific) are in fact an integral part of the one overall system.