Prague Arabic Dependency Treebank: Development in Data and Tools

Jan Hajič[♠], Otakar Smrž[◇], Petr Zemánek[♣],
Jan Šnaidauf[◇], Emanuel Beška[◇]

hajic@ufal.mff.cuni.cz, smrz@ckl.mff.cuni.cz, petr.zemanek@ff.cuni.cz,
jan.snaidauf@amo.cz, emanbe@web.de

[♠]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague
[◇]Center for Computational Linguistics, Faculty of Mathematics and Physics, Charles University in Prague
[♣]Institute of Comparative Linguistics, Faculty of Arts, Charles University in Prague

Abstract

Prague Arabic Dependency Treebank not only consists of multi-level linguistic annotations over the language of Modern Standard Arabic, but even provides a variety of unique software implementations designed for general use in Natural Language Processing (NLP). This paper delivers an overview of the recent and most interesting results, findings and innovations within the project.

1 Introduction

Let us refer to other literature (Smrž et al., 2002; Smrž and Zemánek, 2002; Žabokrtský and Smrž, 2003) instead of repeating the initial motivation for and the history of the Prague Arabic Dependency Treebank (PADT) project.

It might be summarized as an open-ended activity of the Center for Computational Linguistics, the Institute of Formal and Applied Linguistics, and the Institute of Comparative Linguistics, Charles University in Prague, resting in multi-level annotation of Arabic language resources in the light of the theory of Functional Generative Description (Sgall et al., 1986; Hajičová and Sgall, 2003). The project is a younger sibling to Prague Dependency Treebank for Czech (Hajič et al., 2001), and is maintained upon co-operation with the Linguistic Data Consortium, University of Pennsylvania, who release non-annotated corpora of Arabic newswire and develop an independent Penn Arabic Treebank (Maamouri and Cieri, 2002).

1.1 Levels of Description

The PADT scenario of annotations employs the upper three levels of the Functional Generative Description (FGD), intending to infer linguistic meaning from the orthographical or phonological realization of the language, and skipping the lower two levels that decompose it down to phonetics.

Morphological annotations identify the textual forms of a discourse lexically and recognize the grammatical categories they assume. Processing on the analytical level describes the superficial syntactic structures present in the discourse, whereas the tectogrammatical level reveals the underlying ones and restores the linguistic meaning.

The **morphological level** of PADT has for long been the same as that available in Penn Arabic Treebank, Part 2 (Maamouri et al., 2004). PADT adopted the way of Tim Buckwalter's morphological analyses (Buckwalter, 2002) and the annotators were using the SelectPOS disambiguation tool written in Python by Kazuaki Maeda.

As reasoned in (Smrž, in prep), the confrontation of this and numerous other implementations of Arabic morphology, which all turned out to give descriptions of morphs rather than morphemes, with the grammatical rules and syntactic behavior of the language (Fischer (2001), inter alia) brought us to reviewing the system and introducing the Functional Arabic Morphology (see Section 2).

The increasing need for the new type of annotations required a different disambiguation tool, and the general idea of MorphoTrees hierarchy came into existence, implemented as an annotation context for TrEd (see Section 5).

Annotations on the **analytical level** have been treated earlier in (Žabokrtský and Smrž, 2003), where the relations between the PADT dependency analytical trees and the phrase-structure trees of the Penn Arabic Treebank were studied. Here, we explain the principles of analytical annotation proper, extending on the types of predicates and discussing their representation. We formulate a hypothesis on using the analytical data to supplement the lexicons of Arabic morphological analyzers with important grammatical categories like humanness, logical gender, etc.

The third, **tectogrammatical level**, has not yet been outlined in Arabic in such a detail that would let PADT annotations commence. The power and success of tectogrammatics in Prague Dependency Treebank for Czech is, however, more than promising and motivating (Čmejrek et al., 2003; Hajič et al., 2003).

1.2 Release of the Data

The corpus of PADT currently consists of morphologically and analytically annotated newswire texts of Modern Standard Arabic, which originate from LDC's resources — Arabic Gigaword (Graff, 2003) and the plain data of Penn Arabic Treebank, Part 1 (Maamouri et al., 2003) and Part 2 (Maamouri et al., 2004). Our annotations are going to be published via LDC by the end of 2004.

The PADT distribution is expected to comprise over 100 000 tokens of data annotated analytically and provided with the disambiguated morphological information. In addition, the release will include complete annotations of MorphoTrees resulting in more than 125 000 tokens, 35 000 of which will have received the analytical processing. The contents will further be divided as indicated in Table 1.

Data Set	[A] Tokens	[M]	T/Para	T/Doc	Original Data Provider	News Period	Related Corpora
AFP	13 000	—	34.6	260	Agence France Presse	July 2000	Penn ATB Part 1
UMH	38 500	—	43.6	290	Al Hayat News Agency	Spring 2002	Penn ATB Part 2
XIN	13 500	—	31.2	160	Xinhua News Agency	May 2003	Arabic Gigaword
ALH	5 000	51 000	51.4	450	Al Hayat News Agency	September 2001	Arabic Gigaword
ANN	10 000	25 500	50.2	630	An Nahar News Agency	November 2002	Arabic Gigaword
XIA	20 000	48 500	25.9	210	Xinhua News Agency	May 2003	Arabic Gigaword

Table 1: Survey of the expected contents of the first release of PADT. Tokens give the number of syntactic units that are annotated [A] analytically [M] within MorphoTrees. Approximate ratios of tokens per paragraph and tokens per document come in the next columns. The sets of selected documents could cover only a couple of days of the specified period of time.

1.3 Annotation Process Research

Manual annotations are the necessary grounding for machine-learning techniques and statistical modeling in NLP. Even during creation of such resources, it is desirable to help the annotators with preliminary models and automation utilities derived from the just-finished data and improving as annotations proceed. In treebanking, this is true not only on the syntactic levels, but even in the tasks of morphological analysis and disambiguation.

The newswire language teems with out-of-vocabulary expressions, mostly proper names and loan-words, which might get included into the lexicon of the morphological analyzer[1]. Providing the annotators with a user interface to perform these updates and re-run the morphological analyzer instantly, is our current programming concern.

In Arabic morphological disambiguation, a prototype feature-based tagger has been developed recently. The morphological data of Penn Arabic Treebank, Part 2 were converted to the quasi-functional representation (Smrž and Pajas, 2004, in this volume) on which the tagger, a modification of (Hajič and Hladká, 1998), was trained. Its core, the weighted features, discriminate individual grammatical categories for up to four tokens present in an input string.

The tagger achieves 3.6 % error rate in assignment of the major part-of-speech (15 possible values), and 10.8 % error rate in choosing from the whole tag set (317 evidenced combinations of the categories). In tokenization, which we evaluate by comparing two series of non-vocalized standard orthographical forms of tokens in terms of the Longest Common Subsequence problem, the tagger excels with the error rate between 0.8 % and 0.6 %, depending on which data — either the testing, or the produced — are the referential sequence.

The process of annotation of analytical structures was, as soon as the AFP data set had been finished, facilitated by a syntactic parser trained on these trees. Since then, the TrEd tree editor has offered the annotators an option to apply automatic parsing and assignment of analytical functions right during the time of annotation.

Looking for linguistic structures in PADT and revising the data is well feasible with Netgraph (Mírovský et al., 2002). This treebank search engine has been, within the mutual co-operation, installed in LDC and customized for the needs of their various projects.

[1] This has been in competence of Tim Buckwalter of LDC.

2 Functional Morphology & MorphoTrees

While highlighting the structure of word forms and the derivational and inflectional processes of the language, even the best computational models of Arabic morphology (Beesley (2001), Buckwalter (2002), Kiraz (2001) and the works referenced therein) never question the information they provide with respect to the linguistic functions the word forms represent.

Supported fully by FGD and (Sproat, 1992), we stress that the task of morphology should be to analyze word forms of a language not only by finding their internal structure, i.e. recognizing morphs, but even by *strictly* discriminating their functions, i.e. providing the true morphemes. This doing in such a way that it should be *completely* sufficient to generate the word form that represents a lexical unit and features all grammatical categories (and structural components) required by context, exclusively from the information comprised in the analyses.

2.1 Functional and Illusory Categories

In morphological description of Arabic, the senses in which grammatical categories are used, get very often confused or not distinguished at all by the computational systems.

Functional Morphology (Smrž, in prep) respects the grammatical requirements of syntactic constructs and needs to model the complete control over word forms. It revives the different senses and fixes them for the categories like number, gender, case, definiteness.

For number and gender, studying the important phenomenon of agreement classifies the senses as follows:

functional category involved in syntactic consideration

- **logical** agreement with numerals and quantifiers
- **formal** other agreement, pronominal reference

illusory category identifying morphs of an expression

The information on these categories commonly returned by Arabic morphological analyzers is functional formal for verbs, which is most relevant, but only illusory for nominal expressions.

It is also common that the oblique case, the mere denotation for homonymous morphs of genitive and accusative in dual, plural and diptotic singular (all meant as illusory), is mistaken for a grammatical category.

صحّحت مصادر مطّلعة رفيعة المستوى تصريحاتهم هذه غير الدّقيقة.

Corrected sources well-informed high-of the-level declarations-of theirs this other-of the-accurate.
Well-informed high-level sources corrected these inaccurate declarations of theirs.

Ṣaḥḥaḥat [**Pred**] (*maṣādiru* [**Sb**] (*muṭṭaliʿatun* [**Atr**]) (*rafīʿatu* [**Atr**] (*al-mustawā* [Atr])))
(*taṣrīḥāti* [**Obj**] (*-him* [Atr]) (*hāḏihi* [**Atr**]) (*ġayra* [**Atr**] (*ad-daqīqati* [Atr]))) .

Tokens	اَلدَّقِيقَةِ	غَيرَ	هٰذِهِ	هِم	تَصرِيحَاتِ	اَلمُستَوَى	رَفِيعَةُ	مُطَّلِعَةٌ	مَصَادِرُ	صَحَّحَت
Logical	`...D`	`FS.D`	`...D`	`...D`	`MP.D`	`MS.D`	`...I`	`...I`	`MP.I`	`....`
Formal	`FS2D`	`..4R`	`FS4.`	`MP2.`	`FS4R`	`MS2D`	`FS1R`	`FS1I`	`FS1I`	`FS..`
Illusory	`FS2D`	`--4-`	`??--`	`MP--`	`FP3-`	`--?D`	`FS1-`	`FS1I`	`--1-`	`FS--`

Figure 1: Functional (logical and formal) and illusory categories versus agreement. Legend: `.` irrelevant, `-` unset, `?` vague, `M` masculine, `F` feminine, `S` singular, `P` plural, `1` nominative, `2` genitive, `3` oblique case, `4` accusative, `D` definite, `I` indefinite, `R` reduced. Bold constituents in brackets agree over each line, but only functional categories reflect this properly.

Considering definiteness, one issue is the logical definiteness of an expression within a sentence, the other is the formal use of morphs and yet the third, illusory presence or absence of the definite or the indefinite article.

Logical definiteness is binary, i.e. an expression is syntactically either definite, or indefinite. It figures in rules of agreement and propagation.

Formal definiteness introduces, in addition to indefinite and definite, the reduced and complex definiteness values describing word formation of *nomen regens* in construct states and logically definite improper annexations, respectively. Let us give examples:

indefinite *ḥulwatun* حلوةٌ nom. *a-sweet*, *Ṣanʿāʾa* صنعاءَ gen. *Sanaa*, *ḥurrayni* حرَّين acc. *two-free*, *tisʿūna* تسعونَ nom. *ninety*, *sanawātin* سنوَاتٍ acc. *years*

definite *al-ḥulwatu* الحلوَةُ nom. *the-sweet*, *al-ḥurrayni* الحرَّين acc. *the-two-free*, *at-tisʿūna* التّسعُونَ nom. *the--ninety*, *as-sanawāti* السّنوَاتِ acc. *the-years*

reduced *ḥulwatu* حلوَةُ nom. *sweet-of*, *wasāʾili* وسائل gen. *means-of*, *ḥurray* حرَّي acc. *two-free-in*, *muḥāmū* محامُو nom. *lawyers-of*, *sanawāti* سنوَاتِ acc. *years-of*

complex *al-ḥulwatu ’l-ibtisāmi* الحلوَةُ الابتسام nom. *the--sweet-of the-smile, the sweet-smiled*, *al-mutaʿaddiday-i ’l-luġāti* المتعدّدَي اللّغاتِ gen. *the-two-multiple--of the-languages, the two multilingual*[2]

Proper names and abstract entities can be logically definite while formally and illusorily indefinite: *fī Kānūna ’t-ṯānī* في كانونَ الثّاني *in January, the second month of ‘Kaanoon’*. There are adjectival construct states that are logically indefinite, but formally not so: *rafīʿu ’l-mustawā* رفيعُ المستوَى *high-level, high of the level.*

Figure 1 exemplifies the principal difference between the functional and the illusory categories and shows the impossibility to restore agreement, and thus to have an *excellent clue for parsing*, if relying on the illusory analyses.

[2]The dropped-ن-plus-ال cases of *al-ʾiḍāfa ġayr al-ḥaqīqīya* الإضافة غير الحقيقيّة *the improper annexation* clearly belong here.

Our hypothesis is that the *analytical* data can provide enough information to refine, using the structures that imply agreement or other grammatical control, the morphological analyzers towards the functional sense. An iterative algorithm would extend the lexicons with static grammatical categories like humanness, logical gender, intrinsic definiteness, which are generally missing in current computational resources, systematize diptotic declension indicators, and bring other improvements.

2.2 MorphoTrees Briefly

The classical concept of morphological analysis is, technically, to take an input substring of a discourse and produce a list of different strings, each of which represents a reading of the input in terms of the underlying lexical units and morphs, and some abstract labels revealing the process of derivation of the input from the lexical units.

The practice has been, at least in Arabic, that the output information is not organized any further. The different analyses are not clustered together according to their common features, and the output strings are linear in structure and need explicit parsing. It is very difficult for a human to interpret the analyses and to discriminate among them. For a machine, it is undefined how to compare the distance of two analyses, as they are naturally all unequal strings.

MorphoTrees is the idea of building effective and intuitive hierarchies over and among the input and output strings of morphological systems. It is especially interesting for Arabic and the Functional Morphology, but it is in no sense limited to either of these.

We continue the description of this technology in an extra paper (Smrž and Pajas, 2004, in this volume).

3 Analytical Dependency Syntax

In FGD, the superficial syntactic structure of a discourse is modeled as a series of dependency trees whose nodes map, one to one, to the tokens resulting from the morphological analysis and tokenization, and whose roots group the nodes according to the division into sentences or paragraphs. These trees are called analytical.

Edges in the trees show that there is a syntactic relation between the governor and its dependent, or rather, the

1. (a) *In ends-of the-month dismissed the-government the-minister.* في أواخر الشّهر أقالت الحكومة الوزير.
   (*Fī* [AuxP] (*ʾawāḫiri* [Adv] (*aš-šahri* [Atr]))) *ʾaqālat* [**Pred**] (*al-ḥukūmatu* [Sb]) (*al-wazīra* [Obj]) .

2. (a) *The-matter clear.* (*Al-ʾamru* [Sb]) *wāḍiḥun* [**Pnom**] . الأمر واضح.

   (b) *Not was the-matter clear.* لم يكن الأمر واضحا.
   (*Lam* [AuxM]) *yakun* [**Pred**] (*al-ʾamru* [Sb]) (*wāḍiḥan* [Pnom]) .

   (c) *Said that the-matter clear.* قال إنّ الأمر واضح.
   Qāla [**Pred**] (*ʾinna* [AuxC] ((*al-ʾamra* [Sb]) *wāḍiḥun* [Obj_**Pnom**])) .

3. (a) *For the-movement six-of representatives.* للحركة ستّة نوّاب.
   Li- [**PredP**] (*al-ḥarakati* [Obj]) (*sittatu* [Sb] (*nūwābin* [Atr])) .

   (b) *From affair-of hers settlement-of this the-dispute.* من شأنها تسوية هذا النّزاع.
   Min [**PredP**] (*šaʾni* [Pnom] (*-hā* [Atr])) (*taswiyatu* [Sb] ((*hāḏā* [Atr]) *an-nizāʿi* [Atr])) .

4. (a) *No influence of his on the-operation.* لا تأثير له على العمليّة.
   Lā [**PredE**] (*taʾṯīra* [Sb] (*la-* [AuxY] (*-hu* [Atr])) (*ʿalā* [AuxP] (*al-ʿamalīyati* [Atr]))) .

Figure 2: Analytical treatment of Arabic predicates. See the running text of Section 3, along with the analogies in Figure 3.

whole subtree under and including the dependent. The nature of the government is expressed by the analytical functions of the nodes being linked.

3.1 Annotation Principles

The concepts of dependency and valency imply the guidelines for determining the hierarchy of the discourse constituents, as well as for resolving the syntactic functions of their units, the tokens. The principle of analysis by reduction (Plátek et al., 2003) is very often the pursued method.

First, where a token requires the presence or form of other syntactic units or where it can be freely complemented by optional constituents, it becomes the governing node and the other ones its subordinates. Conversely, a governor should neither be influenced by its subordinate nodes[3], nor be itself a complement of any of these.

Second, structural consistency and recursiveness of the language must be generally respected. This means that the description of clauses in a sentence treats them equally with non-clausal expressions, and that the internal structure of a clause would not change with its position in the sentence.

3.2 Identifying Root Nodes of Sentences

Let us discuss these issues within the problem of finding and classifying the topmost nodes of sentences and clauses. We will not make much distinction in the terminology further, and will refer to such nodes as the heads or roots of the structures in question.

One would probably consider deciding between the two dominant syntactic elements, the subject and the predicate, to assume the topmost position.

A choice might establish the subject as the root node, which is supported by its obvious independence on other elements. However, in Arabic and many other languages, subject is only optional and need not be explicitly expressed as a token. More seriously, the role of the subject is conditioned by the valency frame of the predicate, and violating it would bring a regular and useless non-projectivity into the trees.

The solution taking the *predicate* as the root node of a clause is therefore preferred. Although its form might be dependent on the subject, the predicate makes up the very core of a sentence and can never be omitted, unless the clause loses its contextual independence as to the meaning.

3.3 Predicate Types and Representation

The following typology is based on the diverse sentence types encountered during the annotations, and fits our theoretical expectations. Please, consult Figure 2 and Figure 3 as you proceed with reading.

The easiest type of structure for analysis is a sentence with a simple verb. Here, the verb acquires the root node position and is assigned the [Pred] analytical function designating it as the predicate. Other nodes are then attached below it. The immediate subordinates of the verb are, if expressed, the subject, objects, adverbials, verbal complements and other modifiers. The declared immediateness is not violated by the possible linking via nodes with coordinating or other auxiliary functions (prepositions [AuxP], conjunctions [AuxC], generic [AuxY], etc.).

The situation becomes more complicated for verbless sentences. There are basically three distinct types of such constructions, which resemble each other as there is no element representing the *is* or *are* in the elementary function of a copula.

One can imagine that the tree is built as if there were a verbal copula. Since empty nodes are not acceptable in the analytical description, we must decide which of the child subtrees of the empty root will move to its position. The new analytical functions reflect this process, as is apparent in $\{\text{Pred}\} \leftarrow [\text{AuxP}] = [\text{PredP}]$, for instance. On the other

[3]Definitely, the influence of the dependents on the governor should be less substantial than that in the reverse direction.

وشكل هذا الإعلان مفاجأة للإسلاميّين في الكويت وعلّق النّائب يوم ٩ / ١١ الحاليّ على تصريحات الشّيخ قائلا: «إنّنا مستمرّون في اقتراحنا العقوبات الشّرعيّة» مشدّدا على أنّ «ممثّلي الشّعب الكويتيّ هم من له الكلمة الأخيرة» .

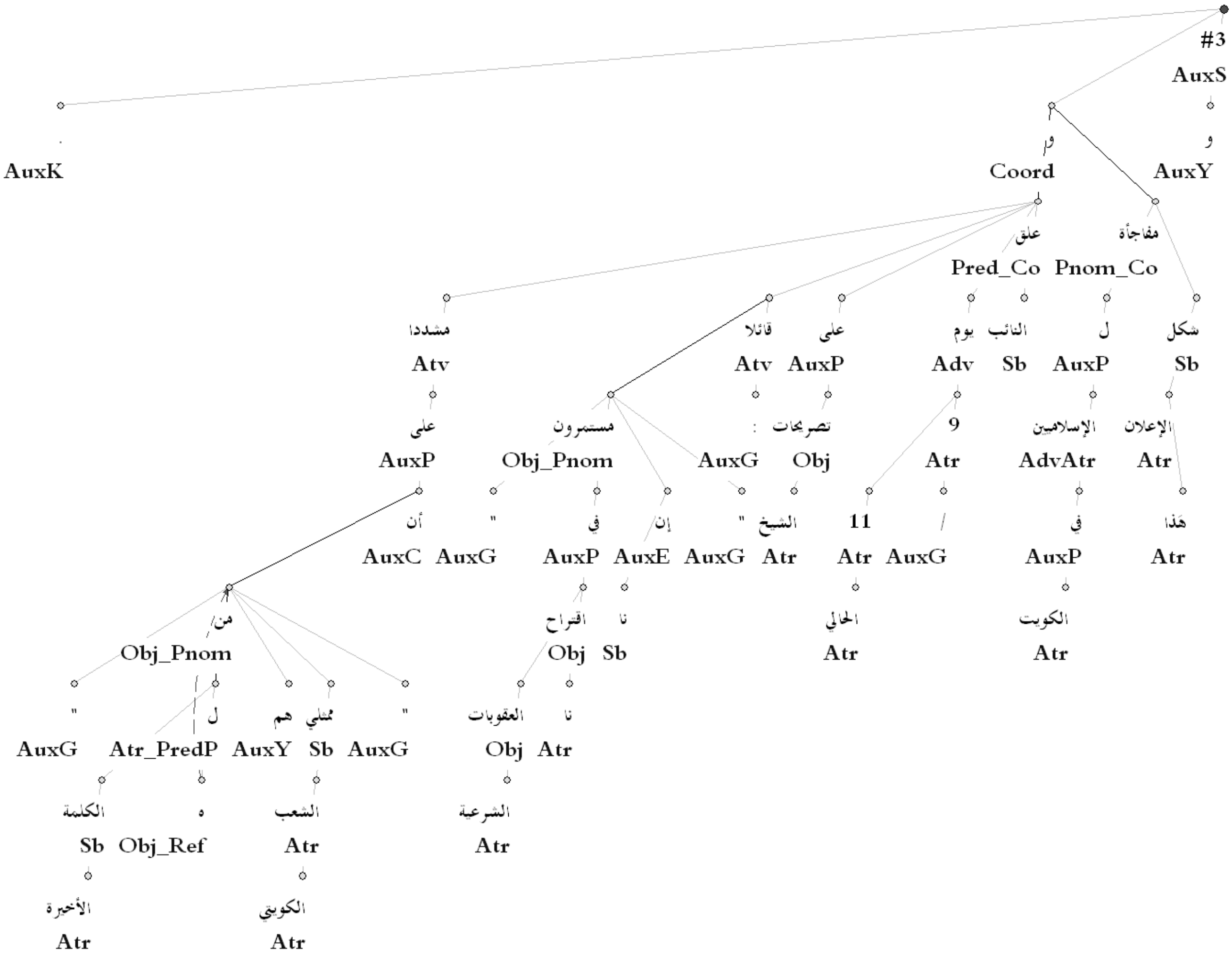

Figure 3: Complex analytical annotation. This abridged sentence comes from the UMH data set and features five predicates (hanging on dark edges), the first two of which are coordinated. The combined analytical function [AdvAtr] is used here to denote the slight homonymy of *sth. is a surprise to sb.*, somebody being partly the adverbial of respect, partly the attribute.

hand, the nominal part of a predicate need not modify its function with the shift, as its status is derivable from the tree. Thus {Pred} ← [Pnom] = [Pnom].

This case, the predicate being a simple nominal expression, can be identified thanks to, e.g., its usual position close after the subject and being either undetermined (مفاجأة and مستمرّون in Figure 3), or determined and immediately preceded by a pronoun (the phrase هم من therein).

The second verbless type covers a wide range of sentences where prepositional phrases are used to express the predicate. They usually designate a locative, temporal or other adverbial relation, or a possessive one. The preposition is assigned the analytical function [PredP] and the root node position, while the head token of the predicate branch receives [Adv] if adverbial, [Obj] if possessive.

There is a particular case of prepositional predication in which the preposition *min* مِن *from* and the following nominal phrase express a part–whole relation (see Figure 2, item 3b). Here, the preposition receives [PredP] and the rest of the phrase depends on it, marked as the nominal part of the predicate [Pnom].

The last type consists of sentences where the existence of things is either asserted, or denied. There is only a very small set of words that can exert such aspect (in fact, only *hunāka* هُنَاكَ, *ṯammata* ثَمَّةَ *there* and *lā* لَا *no*). In analogy to verbal predicates, these elements acquire the root node position, but a distinct denotation [PredE]. In verbal sentences, these expressions are treated as [Adv], adverbial, or [AuxM], auxiliary verbal modification.

Yet, one more type is accounted for — the rare case of verbless sentences where an anticipated nominal predicate is instead substituted by a clause started with a conjunction. The conjunction gets the function [PredC] and becomes the root node of the sentence. Thus, it governs the preceding subject and the following subordinate clause, which forms the [Pnom].

3.4 Predicates in Compound Sentences

Subordinate clauses should fill in the position that otherwise their non-clausal paraphrase would fill in. The analytical function of such a hypothetical expression is then ascribed to the whole clause, whose internal structure must be preserved[4]. The way to do this is to link the clause's head node to the superordinate clause and always assign this head node the analytical function that the whole clause assumes relative to the superordinate clause.

To indicate unambiguously, however, that the substructure is a clause on its own, we mark the head node with its internal function, i.e. the clause's predicate type. The analytical denotation then consists of two functions, the external and the internal one, which are joined with an underscore when rendered.

Another problem to be mentioned here concerns compound sentences whose subordinate part is a verbless prepositional-type clause. Due to the known rule of Arabic saying that relative clauses for undetermined expressions are not started with any relative pronoun, it is sometimes impossible to clearly decide whether the prepositional phrase actually constitutes a subordinate clause, or whether it is a constituent of the sentence. The solution is preferred that better suits the context.

3.5 Other Pieces of the Puzzle

Several particular problems emerged while the data were analyzed, from which only a few examples have been chosen for this article.

One of the characteristics of Arabic is an obligatory use of personal pronouns in certain syntactic constructs. Such pronouns are only grammatical correferents to other entities and are present due to the formal requirements of the language. As a corollary of the definition, the target of a grammatical correference is inferable merely from the analytical structure of the discourse.

The pronouns qualifying as grammatical correferents obtain the marker [_Ref] next to their analytical function. In the TrEd annotation tool, resolution of grammatical correference is implemented, too, and special links from the correferents to their targets get displayed without other human intervention (see the dashed arc in Figure 3).

An independent issue is Arabic compound verbs, i.e. constructs where the predicate is expressed by two verbs, one carrying the meaning and the other one indicating a certain additional feature[5]: temporality, inchoativity, duration, etc. In our approach, the first verb is always marked as the root-node predicate and usually governs the elements[6] preceding the occurrence of the second verb, which in turn is seen (Smrž et al., 2002) as a verbal complement [Atv] of the first verb and governs all the other constituents.

[4]Unless a conjunction occurs that further affects the clause, like *ʾanna* أَنَّ, *ʾinna* إِنَّ *that* requiring the subject's accusative.

[5]Far more rarely, both the verbs retain their own meanings.

[6]Most often, the subject, like in Figure 4, item 2.

4 Tectogrammatical Level

Tectogrammatics, the underlying syntax reflecting the linguistic meaning of a sentence, is the highest level of the PADT annotation. It is driven by FGD and captures dependency and valency with respect to the deep linguistic relations of discourse participants. In its generality, the description also includes topic-focus articulation, correference resolution and other non-dependency relations.

The topology of a tectogrammatical representation of a sentence is similar to that of the analytical level. In contrast to it, nodes in the tree may be deleted, inserted, and even reorganized. We speak of a transfer of structures from analytical to tectogrammatical, which can be automated to some extent.

Basically, on this level of annotation, only autosemantic words have a node of their own, while the correlates of function words are attached as indices to the words which they belong to (i.e., auxiliary verbs and subordinating conjunctions to verbs, prepositions to nouns, etc.). The nodes appear as lexical entries rather than inflected forms.

The participants are labeled with the roles they assume, called here tectogrammatical functions or functors (such as Actor/Bearer, Patient, Addressee, Effect, various types of local and temporal modifications, Cause, Benefactive). The information from the morphological and analytical levels (indispensable grammatical categories like logical number, degree of comparison, modality, tense) can be preserved in gramatemes, another type of attributes of the nodes.

The work on the annotation guidelines for this level has started only recently, and just a few remarks can be made at this point. In the following, an outline of solutions of the transfer of predicate types to the tectogrammatical level is given. Examples are delivered in Table 4.

4.1 Verbal Predication

If the predicate node is occupied by an autosemantic verb[7], the node remains predicative also on the tectogrammatical level. The verb itself, finite verbal form, is substituted by its lemma. Additional grammatical information from the finite verb form is transferred to the gramatemes of the node.

The case of compound verbs occurs when the [Pred] node is occupied by an auxiliary and the autosemantic verb is subordinated as [Atv], its complement (Smrž and Zemánek, 2002). Here, the auxiliary hides in the background as a feature of the former [Atv] node, which is elevated and obtains the PRED label. All the former brother subtrees of the [Atv] node depend on the new PRED, and their functors are assigned based on the conditions of the context[8].

4.2 Non-verbal Predication

All these types of constructions in Arabic are characterized by a zero verbal copula receiving an extra node with a fictitious lemma that represents the expected meaning. Alternatively, the lemma can be chosen such that employing it

[7]This may well be the case of the *kāna* كان *and its sisters* verbs, as long as they are used in their autosemantic senses.

[8][Sb] usually becomes Actor in active sentences, Patient in passive ones, but not always so, as shown in Figure 4, item 4!

1. *Welcomed Greece by the-initiative.* رحّبت اليونان بالمبادرة.
   Raḥḥabat [Pred] (*al-Yūnānu* [Sb]) (*bi-* [AuxP] (*al-mubādarati* [Obj])) .
   Raḥḥab PRED.Ind.Ant (*al-Yūnān* ACT.Def) (*mubādarat* PAT.Def) .

2. *And was Murad pilots the-plane.* وكان مراد يقود طائرة.
   (*Wa* [AuxY]) (*kāna* [Pred] (*Murādun* [Sb]) (*yaqūdu* [Atv] (*ṭā'iratan* [Obj]))) .
   Qād PRED.Ind.Ant.Proc (*Murād* ACT.Def) (*ṭā'irat* PAT.Indef) .

3. *The-climate humid.* (*Al-munāḫu* [Sb]) *ǧāffun* [Pnom] . المناخ جافّ.
   Empty_Copula PRED.Ind.Sim (*munāḫ* ACT.Def) (*ǧāff* PAT.Indef) .

4. *For the-movement representatives.* *Li-* [PredP] (*al-ḥarakati* [Obj]) (*nūwābun* [Sb]) . للحركة نوّاب.
   Empty_Possess PRED.Ind.Sim (*ḥarakat* ACT.Def) (*nūwāb* PAT.Indef) .

Figure 4: Elementary tectogrammatical constructs given in correspondence to the main analytical types of predication. Functors: `PRED` Predicate, `ACT` Actor, `PAT` Patient. Gramatemes: `Ind` indicative/affirmative mood, `Ant` anterior tense, `Sim` simultaneous tense, `Proc` processual aspect, `Def` logically definite, `Indef` logically indefinite.

yields a synonymous sentence. In Figure 4, Empty_Copula could possibly by understood as *kān* كان *to be*, transforming the [Pnom] to the PAT.

For the [PredP], [PredE] and [PredC] types of predicates, the node for the zero copula has to be either inserted, or created from the existing root node (the preposition, conjunction or existential expression, respectively). Then, it is renamed to PRED and labeled with the lemma that delivers the meaning of the predication the best.

5 Programming and Annotation Tools

The indispensable annotation environment for Prague (Arabic) Dependency Treebank is **TrEd**, the tree editor written in Perl by Petr Pajas (Hajič et al., 2001, http://ckl.mff.cuni.cz/pajas/tred/). It is not only a fully programmable and customizable GUI, but also an excellent suite of utilities needed in the every-day project management and data batch processing. Using the TrEd's API to implement the language-specific functions for Arabic analytical annotation, and even to design the complete MorphoTrees annotation context, was extremely quick and comfortable.

Netgraph (Mírovský and Ondruška, 2002, http://quest.ms.mff.cuni.cz/netgraph/) is a client-server application for efficient searching in treebanks. Unlike TrEd, it provides the user with an easy-to-learn query language that does not presume any programming skills.

Next to the other linguistically significant solutions, there is the **Encode::Arabic** module (Smrž, 2003, http://ckl.mff.cuni.cz/smrz/Encode/Arabic/) for Perl offering the non-trivial transducers for turning the ArabTEX transliteration (Lagally, 2004) into the Arabic script or its phonetical transcription. It covers the Buckwalter transliteration, too.

6 Conclusions and Prospects

This paper gives the most complete account on the PADT project ever. Many of its ideas and approaches are original and unprecedented in Arabic NLP.

The morphological description in the terms of syntactic tokens and their relevant, functional grammatical categories seems to be missing in computational literature. The Functional Arabic Morphology is not implemented yet. Its approximation used in the MorphoTrees annotations is quite a close one, though, and will improve.

Our prototype tagger shows that this new approach is promising. Re-training of the system to the truly functional data is expected. The work might constitute a baseline in full Arabic morphological tagging, considering of course the interesting results by Khoja (2001), Kirchhoff et al. (2002), Schafer and Yarowsky (2003), Rogati et al. (2003).

The data and tools of PADT will be released before the end of 2004 by LDC. Morphological and analytical annotations will proceed. The tectogrammatical level will be studied thoroughly, including our intentions to set up a group for building Arabic valency lexicons.

Acknowledgements

The project of Prague Arabic Dependency Treebank is a joint venture of experts in computer science on the one hand, and in linguistics and Arabic studies on the other.

Petr Pajas, Zdeněk Žabokrtský, Jiří Mírovský, Roman Ondruška and Jiří Hana must be credited for the essential contribution in the field of software development and data management.

Ivona Kučerová, Jarmila Panevová and Jan Štěpánek are the consultants who helped design the annotation guidelines and who came up with other important suggestions to the project.

Ondřej Beránek, Viktor Bielický, Kamila Hassanová, Simona Hlaváčová, Markéta Husinecká, Emíra Klementová, Monika Kolbová, Jakub Kráčmar, Alena Pejcharová, Martin Špáta and Pavel Ťupek have been, in different periods of time, in charge of building the treebank and refining the linguistic description critically.

Much of inspiration for the work originates from the enriching co-operation with the colleagues of the LDC, Mohamed Maamouri, Tim Buckwalter, Ann Bies, Wigdan Mekki, Hubert Jin, Mark Liberman and Christopher Cieri in particular.

The research described herein has been supported by the Ministry of Education of the Czech Republic, projects LN00A063 and MSM113200006, and partially by the Grant Agency of the Czech Republic, project 405/02/0823.

Arabic script displays in this paper were typeset using the ArabTEX package (Lagally, 2004) for TEX and LATEX by Prof. Dr. Klaus Lagally of the University of Stuttgart.

References

Kenneth R. Beesley. 2001. Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. In *EACL 2001 Workshop Proceedings on Arabic Language Processing: Status and Prospects*, pages 1–8, Toulouse, France, July 2001.

Tim Buckwalter. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. LDC catalog number LDC2002L49, ISBN 1-58563-257-0.

Martin Čmejrek, Jan Cuřín, and Jiří Havelka. 2003. Czech-English Dependency-based Machine Translation. In *EACL 2003 Proceedings of the Conference*, pages 83–90, Budapest, Hungary, April 2003.

Wolfdietrich Fischer. 2001. *A Grammar of Classical Arabic*. Yale Language Series. Yale University Press, third revised edition. Translated by Jonathan Rodgers.

David Graff. 2003. Arabic Gigaword. LDC catalog number LDC2003T12, ISBN 1-58563-271-6.

Jan Hajič and Barbora Hladká. 1998. Tagging Inflective Languages: Predicting Morphological Categories for a Rich, Structured Tagset. In *Proceedings of COLING--ACL 1998, Montreal, Canada*, pages 483–490. ACL.

Jan Hajič, Barbora Hladká, and Petr Pajas. 2001. The Prague Dependency Treebank: Annotation Structure and Support. In *Proceeding of the IRCS Workshop on Linguistic Databases*, pages 105–114, Philadelphia, December 2001. University of Pennsylvania.

Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. 2003. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, pages 57–68, Växjö, Sweden, November 2003.

Eva Hajičová and Petr Sgall. 2003. Dependency Syntax in Functional Generative Description. In *Dependenz und Valenz – Dependency and Valency*, volume I, pages 570–592. Walter de Gruyter.

Shereen Khoja. 2001. APT: Arabic Part-of-Speech Tagger. In *Proceedings of Student Research Workshop at NAACL 2001*, pages 20–26, Pittsburgh, June 2001.

George Anton Kiraz. 2001. *Computational Nonlinear Morphology with Emphasis on Semitic Languages*. Studies in Natural Language Processing. Cambridge University Press.

Katrin Kirchhoff, Jeff Bilmes, et al. 2002. Novel Speech Recognition Models for Arabic: JHU Summer Workshop 2002 Final Report. Technical report, Johns Hopkins University.

Klaus Lagally. 2004. ArabTEX: Typesetting Arabic and Hebrew, User Manual Version 4.00. Technical Report 2004/03, Fakultät Informatik, Universität Stuttgart.

Mohamed Maamouri and Christopher Cieri. 2002. Resources for Natural Language Processing at the Linguistic Data Consortium. In *Proceedings of the International Symposium on Processing of Arabic*, pages 125–146, Manouba, Tunisia, April 2002.

Mohamed Maamouri, Ann Bies, Hubert Jin, and Tim Buckwalter. 2003. Arabic Treebank: Part 1 v 2.0. LDC catalog number LDC2003T06, ISBN 1-58563-261-9.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Hubert Jin. 2004. Arabic Treebank: Part 2 v 2.0. LDC catalog number LDC2004T02, ISBN 1-58563-282-1.

Jiří Mírovský and Roman Ondruška. 2002. Netgraph System: Searching through the Prague Dependency Treebank. *Prague Bulletin of Mathematical Linguistics*, (77):101–104.

Jiří Mírovský, Roman Ondruška, and Daniel Průša. 2002. Searching through Prague Dependency Treebank. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories*, pages 114–122, Sozopol, Bulgaria, September 2002.

Martin Plátek, Markéta Lopatková, and Karel Oliva. 2003. Restarting Automata: Motivations and Applications. In *Proceedings of the Workshop Petrinetze*, pages 90–96, München, Germany, September 2003.

Monica Rogati, Scott McCarley, and Yiming Yang. 2003. Unsupervised Learning of Arabic Stemming Using a Parallel Corpus. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 391–398, Sapporo, Japan, July 2003.

Charles Schafer and David Yarowsky. 2003. A Two-Level Syntax-Based Approach to Arabic-English Statistical Machine Translation. In *Proceedings of the MT Summit IX Workshop on Machine Translation for Semitic Languages: Issues and Approaches*, pages 45–52, New Orleans, September 2003.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel & Academia, Dordrecht & Prague.

Otakar Smrž and Petr Pajas. 2004. MorphoTrees of Arabic and Their Annotation in the TrEd Environment. In *NEMLAR 2004 Conference Proceedings*.

Otakar Smrž and Petr Zemánek. 2002. Sherds from an Arabic Treebanking Mosaic. *Prague Bulletin of Mathematical Linguistics*, (78):63–76.

Otakar Smrž, Jan Šnaidauf, and Petr Zemánek. 2002. Prague Dependency Treebank for Arabic: Multi-Level Annotation of Arabic Corpus. In *Proceedings of the International Symposium on Processing of Arabic*, pages 147–155, Manouba, Tunisia, April 2002.

Otakar Smrž. 2003. Encode::Arabic. Programming module registered in the Comprehensive Perl Archive Network, http://search.cpan.org/dist/Encode-Arabic/.

Otakar Smrž. in prep. *Functional Arabic Morphology. Formal System and Implementation*. Ph.D. thesis, Charles University in Prague.

Richard Sproat. 1992. *Morphology and Computation*. ACL–MIT Press Series in Natural Language Processing. MIT Press.

Zdeněk Žabokrtský and Otakar Smrž. 2003. Arabic Syntactic Trees: from Constituency to Dependency. In *EACL 2003 Conference Companion*, pages 183–186, Budapest, Hungary, April 2003.