Arabic Computational Linguistics: Current Implementations

Edited by Ali Farghaly

January 17, 2008

CENTER FOR THE STUDY
OF LANGUAGE
AND INFORMATION

1

The Other Arabic Treebank: Prague Dependencies and Functions

Otakar Smrž and Jan Hajič

The words in the title of this chapter seem to like each other to a surprising extent. Not only are the notions of dependency and function central to many modern linguistic theories and 'inherent' to computer science and logic. Their connection to the study of the Arabic language and its meaning is interesting, too, as the traditional literature on these topics, with some works dating back more than a thousand years, actually involved and developed similar concepts.

One of the theories of linguistic meaning and its relation to written or spoken language is Functional Generative Description (FGD). It has become the background for a family of Prague Dependency Treebanks, including Prague Arabic Dependency Treebank (PADT), which represent natural languages by formal means on multiple and mutually inter-operating levels of abstraction: morphological, analytical, and tectogrammatical.

In the current contribution, we would like to discuss the most prominent issues in the description of Arabic that we have encountered during the building of PADT. In particular, we will focus on:

a. the functional model of the morphology–syntax interface in Arabic
b. the morphological hierarchies and their annotation
c. description of surface syntax in the dependency framework
d. tectogrammatics and the representation of information structure

We will try to give enough references that can provide the context for our research as well as inspire to deeper investigations into the problems.

Note on style For the presentation of Arabic, two alternative modes are used next to the original script. Buckwalter transliteration appears in the `typewriter` font, whereas phonetic transcription is typeset sans serif.

Arabic Computational Linguistics: Current Implementations.
Edited by Ali Farghaly.

1 Functional Description of Language

Prague Arabic Dependency Treebank is a project of analyzing large amounts of linguistic data in Modern Written Arabic in terms of the formal representation of language that originates in the Functional Generative Description (Sgall et al., 1986, Sgall, 1967, Panevová, 1980, Hajičová and Sgall, 2003).

Within this theory, the formal representation delivers the linguistic meaning of what is expressed by the surface realization, i.e. the natural language. The description is designed to enable generating the natural language out of the formal representations. By constructing the treebank, we provide a resource for computational learning of the correspondences between both languages, the natural and the formal.

Functional Generative Description stresses the principal difference between the form and the function of a linguistic entity,[1] and defines the kinds of entities that become the building blocks of the respective level of linguistic description—be it underlying or surface syntax, morphemics, phonology or phonetics.

In this theory, a morpheme is the least unit representing some linguistic meaning, and is understood as a function of a morph, i.e. a composition of phonemes in speech or orthographic symbols in writing, which are in contrast the least units capable of distinguishing meanings.

Similarly, morphemes build up the units of syntactic description, and assume values of abstract categories on which the grammar can operate. In FGD, this very proposition implies a complex suite of concepts, introduced with their own terminology and constituting much of the theory. For our purposes here, though, we would only like to reserve the generic term 'token' to denote a syntactic unit, and defer any necessary refinements of the definition to later sections.

The highest abstract level for the description of linguistic meaning in FGD is that of the underlying syntax. It comprises the means to capture all communicative aspects of language, including those affecting the form of an utterance as well as the information structure of the discourse. From this deep representation, one can generate the lower levels of linguistic analysis, in particular the surface syntactic structure of a sentence and its linear sequence of phonemes or graphemes.

In the series of Prague Dependency Treebanks (Hajič et al., 2001, 2006, Cuřín et al., 2004, Hajič et al., 2004a), this generative model of the linguistic process is inverse and annotations are built, with minor modifications to the theory, on the three layers denoted as morphological, analytical and tectogrammatical.

Morphological annotations identify the textual forms of a discourse lexically and recognize the morphosyntactic categories that the forms assume. Processing on the analytical level describes the superficial syntactic relations present in the discourse, whereas the tectogrammatical level reveals the underlying structures and restores the linguistic meaning (cf. Sgall et al., 2004, for what concrete steps that takes).

[1] It seems important to note that the assignment of function to form is arbitrary, i.e. subject to convention—while Kay (2004) would recall *l'arbitraire du signe* in this context, Hodges (2006, section 2) would draw a parallel to waḍʿ وضع *convention.*

2 Functional Arabic Morphology

Arabic is a language of rich morphology, both derivational and inflectional (Holes, 2004). Due to the fact that the Arabic script does usually not encode short vowels and omits some other important phonological distinctions, the degree of morphological ambiguity is very high.

2.1 The Tokenization Problem

In addition to this complexity, Arabic orthography prescribes to concatenate certain word forms with the preceding or the following ones, possibly changing their spelling and not just leaving out the whitespace in between them. This convention makes the boundaries of lexical or syntactic units, which need to be retrieved as tokens for any deeper linguistic processing, obscure, for they may combine into one compact string of letters and be no more the distinct 'words'.

Tokenization is an issue in many languages. Unlike in Chinese or German or Sanskrit (cf. Huet, 2003), in Arabic there are clear limits to the number and the kind of tokens that can collapse in such manner.[2] This idiosyncrasy may have lead to the prevalent interpretation that the clitics, including affixed pronouns or single-letter 'particles', are of the same nature and status as the derivational or inflectional affixes. Cliticized tokens are often considered inferior to some central lexical morpheme of the orthographic string, which yet need not exist if it is only clitics that constitutes the string . . .

We think about the structure of orthographic words differently. In treebanking, it is essential for morphology to determine the tokens of the studied discourse in order to provide the units for the syntactic annotation. Thus, it is *nothing but these units* that must be promoted to tokens and considered equal in this respect, irrelevant of how the tokens are realized in writing.

To decide in general between pure morphological affixes and the critical run-on syntactic units, we use the criterion of substitutability of the latter by its synonym or analogy that can occur isolated. Thus, if hiya هي nom. *she* is a syntactic unit, then the suffixed -hā ها gen. *hers*/acc. *her* is tokenized as a single unit, too. If sawfa سوف *future marker* is a token, then the prefixed sa- سـ, its synonym, will be a token. Definite articles or plural suffixes do not qualify as complete syntactic units, on the other hand.

The leftmost columns in Figure 1 illustrate how input strings are tokenized in PADT, which may in detail contrast to the style of the Penn Arabic Treebank (examples in Maamouri and Bies, 2004).

Discussions can be raised about the subtle choices involved in tokenization proper, or about what orthographic transformations to apply when reconstructing the tokens. Habash and Rambow (2005, section 7) correctly point out the following:

> There is not a single possible or obvious tokenization scheme: a tokenization scheme is an analytical tool devised by the researcher.

Different tokenizations imply different amount of information, and further influence the options for linguistic generalization (cf. Bar-Haim et al., 2005, for the case of Hebrew). We will resume this topic in Section 3 on MorphoTrees.

[2] Even if such rules differ in the standard language and the various dialects.

String	Token Tag	Buckwalter Morph Tags	Token Form	Token Gloss
سيخبرهم	F---------	FUT	sa-	will
	VIIA-3MS--	IV3MS+IV+IVSUFF_MOOD:I	yu-ḫbir-u	he-notify
	S----3MP4-	IVSUFF_DO:3MP	-hum	them
بذلك	P---------	PREP	bi-	about/by
	SD----MS--	DEM_PRON_MS	ḏālika	that
عن	P---------	PREP	ʿan	by/about
طريق	N-------2R	NOUN+CASE_DEF_GEN	ṭarīq-i	way-of
الرسائل	N-------2D	DET+NOUN+CASE_DEF_GEN	ar-rasāʾil-i	the-messages
القصيرة	A-----FS2D	DET+ADJ+NSUFF_FEM_SG+ +CASE_DEF_GEN	al-qaṣīr-at-i	the-short
والإنترنت	C---------	CONJ	wa-	and
	Z-------2D	DET+NOUN_PROP+ +CASE_DEF_GEN	al-ʾinternet-i	the-internet
وغيرها	C---------	CONJ	wa-	and
	FN------2R	NEG_PART+CASE_DEF_GEN	ġayr-i	other/not-of
	S----3FS2-	POSS_PRON_3FS	-hā	them

FIGURE 1 Tokenization of orthographic strings into tokens in *he will notify them about that through SMS messages, the Internet, and other means*, and the disambiguated morphological analyses providing each token with its tag, form and gloss (lemmas are omitted here).

2.2 Functional and Illusory Categories

Once tokens are recognized in the text, the next question comes to mind—while concerned with the token forms, what morphosyntactic properties do they express?

It appears from the literature and the implementations of morphological analyzers (many summarized in Al-Sughaiyer and Al-Kharashi, 2004) that Arabic computational morphology has understood its role in the sense of operations with morphs rather than morphemes (cf. El-Sadany and Hashish, 1989), and has not concerned itself systematically and to the necessary extent with its role for syntax.[3] In other words, the syntax–morphology interface has not been clearly established in most computational models.

The outline of formal grammar in (Ditters, 2001), for example, builds on grammatical categories like number, gender, humanness, definiteness, but many morphological analyzers (eg. Beesley, 2001, Buckwalter, 2002, 2004a, Kiraz, 2001) would not return this information completely right. It is discussed in (Smrž, 2007b, Hajič et al., 2005, 2004b) that these systems misinterpret some morphs for bearing a category, and underspecify lexical morphemes in general as to their intrinsic morphological functions.

In Figure 1, the Buckwalter analysis of the word ar-rasāʾil-i الرّسائل *the messages* says that this token is a noun, in genitive case, and with a definite article. It does not continue, however, that it is also the actual plural of risāl-ah رسالة *a message*, and that this logical plural formally behaves as feminine singular, as is the grammatical rule for every noun not referring to a human. Its congruent attribute al-qaṣīr-at-i القصيرة *the short* is marked as feminine singular due to the presence of the -ah ة morph. Yet, the mere presence of a morph does not guarantee its function, and vice versa.

[3] Versteegh (1997, chapter 6) describes the traditional Arabic understanding of ṣarf صرف *morphology* and naḥw نحو *grammar, syntax*, where morphology studied the derivation of isolated words, while their inflection in the context of a sentence was part of syntax.

What are the genders of ṭarīq طريق *way* and al-ʾinternet الإنترنت *the Internet*? Their tags do not tell, and ṭarīq طريق actually allows either of the genders in the lexicon.

This discrepancy between the implementations and the expected linguistic descriptions compatible with e.g. (Fischer, 2001, Badawi et al., 2004, Holes, 2004) can be seen as an instance of the general disparity between inferential–realizational morphological theories and the lexical or incremental ones. Stump (2001, chapter 1) presents evidence clearly supporting the former methodology, according to which morphology needs to be modeled in terms of lexemes, inflectional paradigms, and a well-defined syntax–morphology interface of the grammar. At least these three of Stump's points of departure deserve remembering in our situation (Stump, 2001, pages 7–11):

> The morphosyntactic properties associated with an inflected word's individual inflectional markings may underdetermine the properties associated with the word as a whole.
>
> There is no theoretically significant difference between concatenative and nonconcatenative inflection.
>
> Exponence is the only association between inflectional markings and morphosyntactic properties.

Many of the computational models of Arabic morphology are lexical in nature, i.e. they associate morphosyntactic properties with individual affixes regardless of the context of other affixes. As these models are not designed in connection with any syntax–morphology interface, their interpretation is destined to be incremental, i.e. the morphosyntactic properties are acquired only as a composition of the explicit inflectional markings. This cannot be appropriate for such a language as Arabic,[4] and leads to the series of problems that we observed in Figure 1.

Functional Arabic Morphology (Smrž, 2007b) is our revised morphological model that endorses the inferential–realizational principles. It re-establishes the system of inflectional and inherent morphosyntactic properties (or grammatical categories or features, in the alternative naming) and discriminates precisely the senses of their use in the grammar. It also deals with syncretism of forms (cf. Baerman et al., 2006) that seems to prevent the resolution of the underlying categories in some morphological analyzers.

The syntactic behavior of ar-rasāʾil-i الرّسائل *the messages* disclosed that we cannot dispense with a single category for number or for gender, but rather, that we should always specify the sense in which we mean it: [5]

functional category is for us the morphosyntactic property that is involved in grammatical considerations; we further divide functional categories into

  logical categories on which agreement with numerals and quantifiers is based

  formal categories controlling other kinds of agreement or pronominal reference

illusory category denotes the value derived merely from the morphs of an expression

[4] Versteegh (1997, chapter 6, page 83) offers a nice example of how the supposed principle of 'one morph one meaning', responsible for a kind of confusion similar to what we are dealing with, complicated some traditional morphological views.

[5] One can recall here the terms maʿnawīy معنويّ *by meaning* and lafẓīy لفظيّ *by expression* distinguished in the Arabic grammar. The logical and formal agreement, or *ad sensum* resp. grammatical, are essential abstractions (Fischer, 2001), yet, to our knowledge, implemented only in El Dada and Ranta (2006).

Does the classification of the senses of categories actually bring new quality to the linguistic description? Let us explore the extent of the differences in the values assigned. It may, of course, happen that the values for a given category coincide in all the senses. However, promoting the illusory values to the functional ones is in principle conflicting:

1. Illusory categories are set only by a presence of some 'characteristic' morph, irrespective of the functional categories of the whole expression. If lexical morphemes are not qualified in the lexicon as to the logical gender nor humanness, then the logical number can be guessed only if the morphological stem of the logical singular is given along with the stem of the word in question. Following this approach implies interpretations that declare illusory feminine singular for e.g. sād-ah سادة *men*, qād-ah قادة *leaders*, quḍ-āh قضاة *judges*, dakātir-ah دكاترة *doctors* (all functional masculine plural), illusory feminine plural for bāṣ-āt باصات *buses* (logical masculine plural, formal feminine singular), illusory masculine dual for ʿayn-āni عينان *two eyes*, biʾr-āni بئران *two wells* (both functional feminine dual), or even rarely illusory masculine plural for sin-ūna سنون *years* (logical feminine plural, formal feminine singular), etc.

2. If no morph 'characteristic' of a value surrounds the word stem and the stem's morpheme does not have the right information in the lexicon, then the illusory category remains unset. It is not apparent that ḥāmil حامل *pregnant* is formal feminine singular while ḥāmil حامل *carrying* is formal masculine singular, or that ǧudud جدد *new* is formal masculine plural while kutub كتب *books* is formal feminine singular. The problem concerns every nominal expression individually and pertains to some verbal forms, too. It is the particular issue about the internal/broken plural in Arabic, for which the illusory analyses do not reveal any values of number nor gender. It would not work easily to set the desired functional values by some heuristic, as this operation could only be conditioned by the pattern of consonants and vowels in the word's stem, and that can easily mislead, as this relation is also arbitrary. Consider the pattern in ʿarab عرب *Arabs* (functional masculine plural) vs. ǧamal جمل *camel* (functional masculine singular) vs. qaṭaʿ قطع *stumps* (logical feminine plural, formal feminine singular), or that in ǧimāl جمال *camels* (logical masculine plural, formal feminine singular) vs. kitāb كتاب *book* (functional masculine singular) vs. ʾināṯ إناث *females* (logical feminine plural, formal feminine singular or plural depending on the referent), etc.

Functional Arabic Morphology enables the functional gender and number information thanks to the lexicon that can stipulate some properties as inherent to some lexemes, and thanks to the paradigm-driven generation that associates the inflected forms with the desired functions directly.

Another inflectional category that we discern for nominals as well as pronouns is case. Its functional values are nominative, genitive, and accusative. Three options are just enough to model all the case distinctions that the syntax–morphology interface of the language requires. The so-called oblique case is not functional, as long as it is the mere denotation for the homonymous forms of genitive and accusative in dual, plural and diptotic singular (all meant in the illusory sense, cf. Fischer, 2001, pages 86–96).

Neither do other instances of reduction of forms due to case syncretism need special treatment in our generative model. In a nutshell—if the grammar asks for an accusative of maʿn-an معنى *meaning*, it does not care that its genitive and nominative forms incidentally look identical. Also note that case is preserved when a noun is replaced by a pronoun in a syntactic structure. Therefore, when we abstract over the category of person, we can consider even ʾanā أنا nom. *I*, -ī/-ya ي gen. *mine*, and -nī ني acc. *me* as members of the pronominal paradigm of inflection in case.

The final category to revise with respect to the functional and illusory interpretations is definiteness. One issue is the logical definiteness of an expression within a sentence, the other is the formal use of morphs within a word, and yet the third, the 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 rules of propagation of definiteness (cf. the comprehensive study by Kremers, 2003).

Formal definiteness, denoted also as state, is independent of logical definiteness. It introduces, in addition to indefinite and definite, the reduced and complex definiteness values describing word formation of *nomen regens* in genitive constructions and logically definite improper annexations, respectively. In (Smrž, 2007a,b), we further formalize this category and refine it with two more values, absolute and lifted. Let us give examples:

indefinite ḥulwatu-n حلوةٌ nom. *a-sweet*, Ṣanʿāʾa صنعاءَ gen./acc. *Sanaa*, ḥurray-ni حرَّين gen./acc. *two-free*, tisʿū-na تسعُونَ nom. *ninety*, sanawāti-n سنوَاتٍ gen./acc. *years*

definite al-ḥulwatu الحلوَةُ nom. *the-sweet*, al-ḥurray-ni الحرَّين gen./acc. *the-two-free*, at-tisʿū-na التّسعُونَ nom. *the-ninety*, as-sanawāti السّنَوَاتِ gen./acc. *the-years*

reduced ḥulwatu حلوَةُ nom. *sweet-of*, wasāʾili وسائلِ gen. *means-of*, wasāʾila وسائلَ acc. *means-of*, ḥurray حرَّي gen./acc. *two-free-in*, muḥāmū محامُو nom. *attorneys-of*, maʿānī معاني nom./gen. *meanings-of*, sanawāti سنوَاتِ gen./acc. *years-of*

complex al-ḥulwatu ʾl-ibtisāmi الحلوَةُ الابتسام nom. *the-sweet-of the-smile, the sweet-smiled*, al-mutaʿaddiday-i ʾl-luġāti المتعدّدَي اللّغَاتِ gen./acc. *the-two-multiple-of the-languages, the two multilingual*[6]

Proper names and abstract entities can be logically definite while formally and illusorily indefinite: fī Kānūna ʾṯ-ṯānī في كانونَ الثّاني *in January, the second month of Kānūn*. Kānūna كانونَ *Kānūn* follows the diptotic inflectional paradigm, which is indicative of formally indefinite words. Yet, this does not prevent its inherent logical definiteness to demand that the congruent attribute aṯ-ṯānī الثّاني *the-second* be also logically definite. Aṯ-ṯānī الثّاني *the-second* as an adjective achieves this by way of its formal definiteness.

From the other end, there are adjectival construct states that are logically indefinite, but formally not so: rafīʿu ʾl-mustawā رفيعُ المستوَى *a high-level, high-of the-level*. Rafīʿu رفيعُ *high-of* has the form that we call reduced, for it is the head of an annexation. If, however, this construct is to modify a logically definite noun, the only way for it to mark its logical definiteness is to change its formal definiteness to complex, such as in al-masʾū-lu ʾr-rafīʿu ʾl-mustawā المسؤولُ الرّفيعُ المستوَى *the-official the-high-of the-level*. We can now

[6] The dropped-ن-plus-ال cases of al-ʾiḍāfah ġayr al-ḥaqīqīyah الإضافة غير الحقيقيّة *the improper annexation* clearly belong here (cf. Smrž et al., 2007, for how to discover more examples of this phenomenon).

inflect the phrase in number. Definiteness will not be affected by the change, and will ensure that the plural definite and complex forms do get distinguished: al-masʾūlū-na 'r-rafīʿū 'l-mustawā المسؤولُونَ الرّفيعُو المستوَى *the-officials the-highs-of the-level.*

In our view, 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, purely from the information comprised in the analyses. Functional Arabic Morphology is a model that suits this purpose.

2.3 ElixirFM Implementation

We first presented the elements of Functional Arabic Morphology in (Hajič et al., 2004b). In PADT 1.0 (Hajič et al., 2004a) and the feature-based morphological tagger that used it (Hajič et al., 2005), this model could not be fully implemented yet. Instead, the functional approximation (Smrž and Pajas, 2004) based on the Buckwalter Arabic Morphological Analyzer (Buckwalter, 2002, 2004a) was developed.

The functional approximation essentially takes the output of the Buckwalter morphology and transforms it in two steps (illustrated in Figure 1):

1. The morphs of the original orthographic strings are re-grouped to form tokens.
2. The corresponding sequences of morph tags are mapped into the fixed-width positional notation in which the two initial positions identify the token's part-of-speech category and its refinement, and the other positions express features like mood, voice,[7] person, (illusory) gender, (illusory) number, case, and formal definiteness.

ElixirFM (Smrž, 2007a,b) is the original implementation of Functional Arabic Morphology, and is being applied as a definitive replacement of the functional approximation for the next versions of the Prague Arabic Dependency Treebank.

ElixirFM is implemented in Haskell, a modern purely functional programming language (cf. eg. Hudak, 2000, Wadler, 1997). ElixirFM extends and reuses the Functional Morphology library and methodology by Forsberg and Ranta (2004).[8]

The lexicon of ElixirFM is derived from the open-source Buckwalter lexicon—it is however redesigned in important respects and extended with functional inherent information learned from the PADT annotations. Thanks to the declarative possibilities of Haskell and the abstraction that it allows, the resulting format of the lexicon resembles the printed human-readable dictionaries. It can be exported or otherwise reused.

The whole morphological model adopts the multi-purpose notation of ArabTEX (Lagally, 2004) as a meta-encoding of both the orthography and phonology. With our Haskell implementation of Encode Arabic (Smrž, 2003–2007) interpreting the notation,

[7] The fifth position is reserved for dialectal features, and is always unset with - in standard data. The complete list of mappings from morph tags to token tags is available from the authors. Similar notations have been used in various projects, most notably the European Multext and Multext-East projects, for languages ranging from English to Czech to Hungarian.

[8] Functional Morphology itself builds on the computational toolkit Zen for Sanskrit (Huet, 2002, 2005). Both elegantly reconcile what put Paradigm Function Morphology (Stump, 2001) and KATR (Finkel and Stump, 2002) under critique by proponents of finite-state methodology (Karttunen, 2003).

```
data Mood = Indicative | Subjunctive | Jussive | Energetic
                                                 deriving (Eq, Enum)
data Gender = Masculine | Feminine               deriving (Eq, Enum)
data Number = Singular | Dual | Plural           deriving (Eq, Enum)

data ParaVerb = VerbP      Voice Person Gender Number
              | VerbI Mood Voice Person Gender Number
              | VerbC                   Gender Number  deriving Eq

paraVerbC :: Morphing a b => Gender -> Number -> [Char] -> a -> Morphs b
paraVerbC g n i = case n of

          Singular  -> case g of  Masculine  -> prefix i . suffix ""
                                  Feminine   -> prefix i . suffix "I"

          Plural    -> case g of  Masculine  -> prefix i . suffix "UW"
                                  Feminine   -> prefix i . suffix "na"

          _         ->                          prefix i . suffix "A"
```

FIGURE 2 Excerpt of the implementation of inflectional features and paradigms in ElixirFM.

ElixirFM can process either the original Arabic script (non-)vocalized to any degree or some kind of transliteration or even transcription thereof (details in Smrž, 2007b).

Morphology is modeled in terms of paradigms, grammatical categories, lexemes and word classes (Figure 2). Inflectional parameters are represented as values of distinct enumerated types (note the three initial `data` declarations). The algebraic data type `ParaVerb` implements the space in which verbs are inflected by defining three Cartesian products of the elementary categories: a verb can have `VerbP` perfect forms inflected in voice, person, gender, number, `VerbI` imperfect forms inflected also in mood, and `VerbC` imperatives inflected in gender and number only (cf. Forsberg and Ranta, 2004).

The paradigm for inflecting imperatives, the one and only such paradigm in ElixirFM, is implemented in `paraVerbC`. It is a function (note its `::` type signature) parametrized by some particular value of gender `g` and number `n`. It further needs the initial auxiliary vowel `i` and the verbal stem (provided by rules or the lexicon) to produce the full form. The definition of `paraVerbC` is very concise due to the chance to compose with `.` the partially applied `prefix` and `suffix` functions and to virtually omit the next argument (cf. the morphology-theoretic views in Spencer, 2004). By evaluating the function for varying parameters in some Haskell interpreter, we get the inflected forms:

`paraVerbC Feminine Plural "u" "ktub"` → `"uktubna"` uktubna اُكْتُبْنَ fem. pl. *write!*

`[ paraVerbC g n "i" "qra'" | g <- values, n <- values ]` →

masc.: `"iqra'"` iqraʾ إِقْرَأْ sg. `"iqra'A"` iqraʾā إِقْرَآ du. `"iqra'UW"` iqraʾū إِقْرَؤُوا pl.

fem.: `"iqra'I"` iqraʾī إِقْرَئِي sg. `"iqra'A"` iqraʾā إِقْرَآ du. `"iqra'na"` iqraʾna إِقْرَأْنَ pl. *read!*

ElixirFM provides a modern computational model of Arabic morphology on which many other applications can be based, cf. (Smrž, 2007a,b). ElixirFM and Encode Arabic are open-source projects available and documented at `http://sourceforge.net/`.

Morphs	Form	Token Tag	Lemma	Morph-Oriented Gloss
`|laY+(null)`	ʾālā	`VP-A-3MS--`	ʾālā	promise/take an oath + he/it
`|liy~+u`	ʾālīy-u	`A-------1R`	ʾālīy	mechanical/automatic + [def.nom.]
`|liy~+i`	ʾālīy-i	`A-------2R`	ʾālīy	mechanical/automatic + [def.gen.]
`|liy~+a`	ʾālīy-a	`A-------4R`	ʾālīy	mechanical/automatic + [def.acc.]
`|liy~+N`	ʾālīy-un	`A-------1I`	ʾālīy	mechanical/automatic + [indef.nom.]
`|liy~+K`	ʾālīy-in	`A-------2I`	ʾālīy	mechanical/automatic + [indef.gen.]
`|l` +	ʾāl	`N--------R`	ʾāl	family/clan
+ `iy`	-ī	`S----1-S2-`	ʾanā	my
`IilaY`	ʾilā	`P---------`	ʾilā	to/towards
`Iilay` +	ʾilay	`P---------`	ʾilā	to/towards
+ `ya`	-ya	`S----1-S2-`	ʾanā	me
`Oa+liy+(null)`	ʾa-lī	`VIIA-1-S--`	waliy	I + follow/come after + [ind.]
`Oa+liy+a`	ʾa-liy-a	`VISA-1-S--`	waliy	I + follow/come after + [sub.]

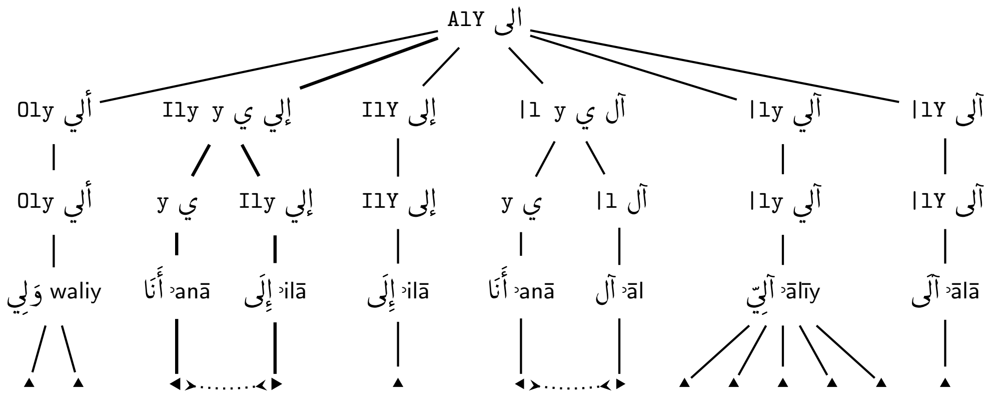


FIGURE 3 Analyses of the orthographic word AlY الى turned into the MorphoTrees hierarchy. The full forms and morphological tags in the leaves are schematized to triangles. The bold lines indicate the annotation, i.e. the choice of the solution Ily y إلي ي ʾilay-ya *to me.*

3 MorphoTrees

The classical concept of morphological analysis is, technically, to take individual subparts of some linear representation of an utterance, such as orthographic words, interpret them regardless of their context, and produce for each of them a list of morphological readings revealing what hypothetical processes of inflection or derivation the given form could be a result of. One example of such a list is seen at the top of Figure 3.

The complication has been, at least with Arabic, that the output information can be rather involved, yet it is linear again while some explicit structuring of it might be preferable. The divergent analyses are not clustered together according to their common characteristics. 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 differences of the analyses, as there is no disparity measure other than unequalness.

MorphoTrees (Smrž and Pajas, 2004) is the idea of building effective and intuitive hierarchies over the information presented by morphological systems (Figure 3). It is especially interesting for Arabic and the Functional Arabic Morphology, yet, it is not limited to the language, nor to the formalism, and various extensions are imaginable.

3.1 The MorphoTrees Hierarchy

As an inspiration for the design of the hierarchies, let us consider the following analyses of the string `fhm` فهم. Some readings will interpret it as just one token related to the notion of *understanding*, but homonymous for several lexical units, each giving many inflected forms, distinct phonologically despite their identical spelling in the ordinary non-vocalized text. Other readings will decompose the string into two co-occurring tokens, the first one, in its non-vocalized form `f` ف, standing for an unambiguous conjunction, and the other one, `hm` هم, analyzed as a verb, noun, or pronoun, each again ambiguous in its functions.

Clearly, this type of concise and 'structured' description does not come ready-made—we have to construct it on top of the overall morphological knowledge. We can take the output solutions of morphological analyzers and process them according to our requirements on tokenization and 'functionality' stated above. Then, we can merge the analyses and their elements into a five-level hierarchy similar to that of Figure 4. The leaves of it are the full forms of the tokens plus their tags as the atomic units. The root of the hierarchy represents the input string, or generally the input entity (some linear or structured subpart of the text). Rising from the leaves up to the root, there is the level of lemmas of the lexical units, the level of non-vocalized canonical forms of the tokens, and the level of decomposition of the entity into a sequence of such forms, which implies the number of tokens and their spelling.

Let us note that the MorphoTrees hierarchy itself might serve as a framework for evaluating morphological taggers, lemmatizers and stemmers of Arabic, since it allows for resolution of their performance on the different levels, which does matter with respect to the variety of applications.

3.2 MorphoTrees Disambiguation

The linguistic structures that get annotated as trees are commonly considered to belong to the domain of syntax. Thanks to the excellent design and programmability of TrEd,[9] the general-purpose tree editor written by Petr Pajas, we could happily implement an extra annotation mode for the disambiguation of MorphoTrees, too. We thus acquired a software environment integrating all the levels of description in PADT.

The annotation of MorphoTrees rests in selecting the applicable sequence of tokens that analyze the entity in the context of the discourse. In a naive setting, an annotator would be left to search the trees by sight, decoding the information for every possible analysis before coming across the right one. If not understood properly, the supplementary levels of the hierarchy would rather tend to be a nuisance...

Instead, MorphoTrees in TrEd take great advantage of the hierarchy and offer the option to restrict one's choice to subtrees and hide those leaves or branches that do

[9] TrEd is open-source and is documented and available at `http://ufal.mff.cuni.cz/~pajas/tred/`.

not conform to the criteria of the annotation. Furthermore, many restrictions may be applied automatically, and the decisions about the tree can be controlled in a very rapid and elegant way.

The MorphoTrees of the entity `fhm` فهم in Figure 4 are in fact annotated already. The annotator was expecting, from the context, the reading involving a conjunction. By pressing the shortcut `c` at the root node, he restricted the tree accordingly, and the only one eligible leaf satisfying the `C---------` tag restriction was selected at that moment. Nonetheless, the `fa-` ف *so* conjunction is part of a two-token entity, and some annotation of the second token must also be performed. Automatically, all inherited restrictions were removed from the `hm` هم subtree (notice the empty tag in the flag over it), and the subtree unfolded again. The annotator moved the node cursor[10] to the lemma for the pronoun, and restricted its readings to the nominative `--------1-` by pressing another mnemonic shortcut `1`, upon which the single conforming leaf `hum` هم *they* was selected automatically. There were no more decisions to make and the annotation proceeded to the next entity of the discourse.

Alternatively, the annotation could be achieved merely by typing `s1`. The restrictions would unambiguously lead to the nominative pronoun, and then, without human intervention, to the other token, the unambiguous conjunction. These automatic decisions need no linguistic model, and yet they are very effective.

Incorporating restrictions or forking preferences sensitive to the surrounding annotations is in principle just as simple, but the concrete rules of interaction may not be easy to find. Morphosyntactic constraints on multi-token word formation are usually hard-wired inside analyzers and apply within an entity—still, certain restrictions might be generalized and imposed automatically even on the adjacent tokens of successive entities, for instance. Eventually, annotation of MorphoTrees might be assisted with real-time tagging predictions provided by some independent computational module.

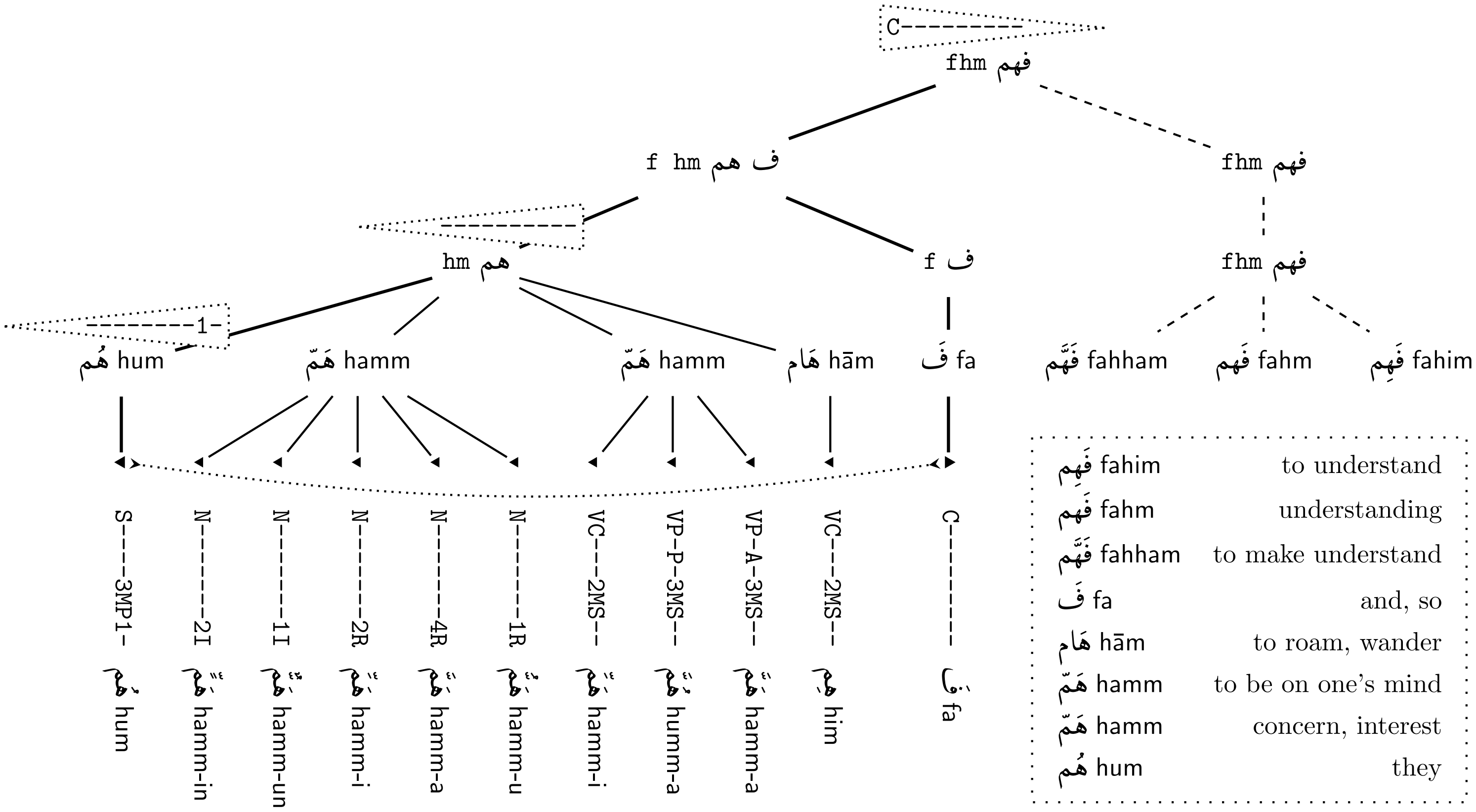
3.3 Further Discussion

Hierarchization of the selection task seems to be the most important contribution of the idea. The suggested meaning of the levels of the hierarchy mirrors the linguistic theory and also one particular strategy for decision-making, neither of which are universal. If we adapt MorphoTrees to other languages or hierarchies, the power of trees remains, though—efficient top-down search or bottom-up restrictions, gradual focusing on the solution, refinement, inheritance and sharing of information, etc.

The levels of MorphoTrees are extensible internally (More decision steps for some languages?) as well as externally in both directions (Analyzed entity becoming a tree of perhaphs discontiguous parts of a possible idiom? Leaves replaced with derivational trees organizing the morphs of the tokens?) and the concept incites new views on some issues encompassed by morphological analysis and disambiguation.

In PADT, whose MorphoTrees average roughly 8–10 leaves per entity depending on the data set while the result of annotation is 1.16–1.18 tokens per entity, restrictions as a means of direct access to the solutions improve the speed of annotation significantly.

[10] Navigating through the tree or selecting a solution is of course possible using the mouse, the cursor arrows, and the many customizable keyboard shortcuts. Restrictions are a convenient option to consider.



فَهِم fahim	to understand
فَهَم fahm	understanding
فَهَّم fahham	to make understand
فَ fa	and, so
هَام hām	to roam, wander
هَمّ hamm	to be on one's mind
هَمّ hamm	concern, interest
هُم hum	they

FIGURE 4 MorphoTrees of the orthographic string `fhm` فهم including annotation with restrictions. The dashed lines indicate that there is no solution suiting the inherited restrictions in the given subtree. The dotted line symbolizes the fact that there might be implicit morphosyntactic constraints between the adjacent tokens in the analyses.

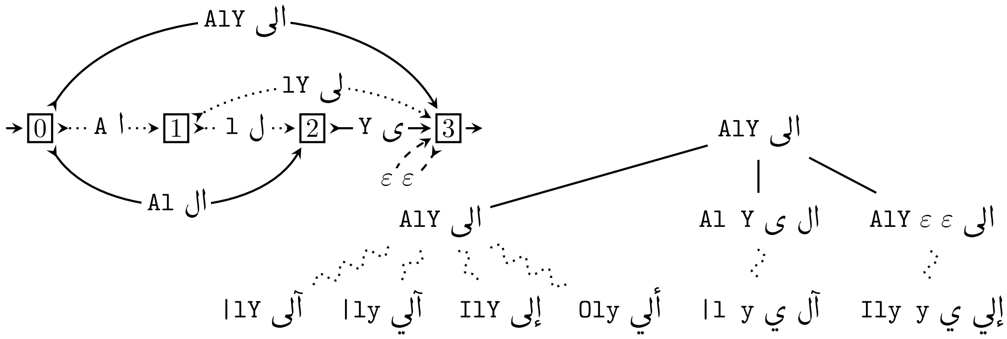


FIGURE 5 Discussion of partitioning and tokenization of input orthographic strings.

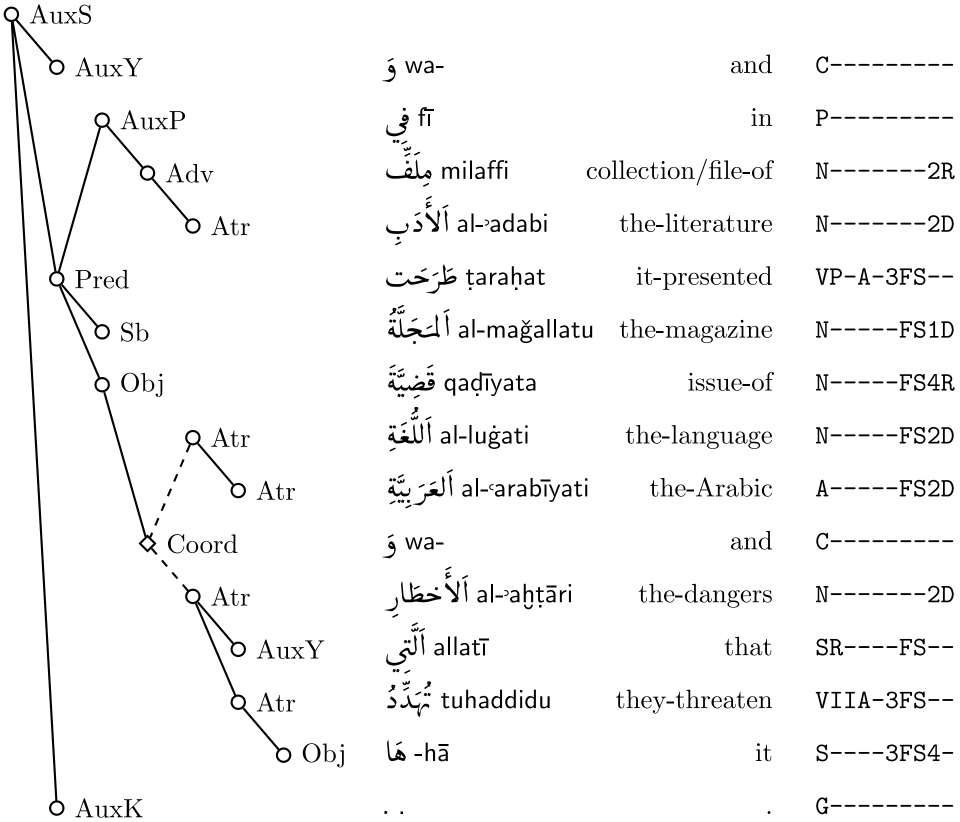
How would the first and the second level below the root in MorphoTrees be defined, if we used a different tokenization scheme? Some researchers do not reconstruct the canonical non-vocalized forms as we do, but only determine token boundaries between the characters of the original string (cf. Diab et al., 2004, Habash and Rambow, 2005). Our point in doing the more difficult job is that (a) we are interested in such level of detail (b) disambiguation operations become more effective if the hierarchy reflects more distictions (i.e. decisions are specific about alternatives).

The relation between these tokenizations is illustrated in Figure 5. The graph on the left depicts the three 'sensible' ways of partitioning the input string AlY الى in the approach of (Diab et al., 2004), where characters are classified to be token-initial or not. In the graph, boundaries between individual characters are represented as the numbered nodes in the graph. Two of the valid tokenizations of the string are obtained by linking the boundaries from 0 to 3 following the solid edges in the directions of the arrows. The third partitioning AlY ε ε الى indicates that there is another fictitious boundary at the end of the string, yielding some 'empty word' ε ε, which together corresponds to leaping over the string at once and then taking the dashed edge in the graph.

Even though conceptually sound, this kind of partitioning may not be as powerful and flexible as what MorphoTrees propose, because it rests in classifying the input characters only, and not actually constructing the canonical forms of tokens as an *arbitrary function* of the input. Therefore, it cannot undo the effects of orthographic variation (Buckwalter, 2004b), nor express other useful distinctions, such as recover the spelling of tāʾ marbūṭah or normalize hamzah carriers.

We can conclude with the tree structure of Figure 5. The boundary-based tokenizations are definitely not as detailed as those of MorphoTrees given in Figure 3, and might be occasionally thought of as another intermediate level in the hierarchy. But as they are not linguistically motivated, we do not establish the level as such.

In any case, we propose to evaluate tokenizations in terms of the Longest Common Subsequence (LCS) problem (Crochemore et al., 2000, Konz and McQueen, 2000–2006). The tokens that are the members of the LCS with some referential tokenization, are considered correctly recognized. Dividing the length of the LCS by the length of one of the sequences, we get recall, doing it for the other of the sequences, we get precision. The harmonic mean of both is $F_{\beta=1}$-measure (cf. e.g. Manning and Schütze, 1999).



AuxS				
AuxY	وَ wa-		and	C---------
AuxP	فِي fī		in	P---------
Adv	مِلَفِّ milaffi		collection/file-of	N-------2R
Atr	اَلأَدَبِ al-ʾadabi		the-literature	N-------2D
Pred	طَرَحَت ṭaraḥat		it-presented	VP-A-3FS--
Sb	اَلمَجَلَّةُ al-maǧallatu		the-magazine	N-----FS1D
Obj	قَضِيَّةَ qaḍīyata		issue-of	N-----FS4R
Atr	اَللُّغَةِ al-luġati		the-language	N-----FS2D
Atr	اَلعَرَبِيَّةِ al-ʿarabīyati		the-Arabic	A-----FS2D
Coord	وَ wa-		and	C---------
Atr	اَلأَخطَارِ al-ʾaḫṭāri		the-dangers	N-------2D
AuxY	اَلَّتِي allatī		that	SR----FS--
Atr	تُهَدِّدُ tuhaddidu		they-threaten	VIIA-3FS--
Obj	هَا -hā		it	S----3FS4-
AuxK	. .		.	G---------

FIGURE 6 Analytical annotation of example (1). Orthographic words are tokenized into lexical words, and grammatical categories are encoded using the positional notation.

4 Syntactic Dependency Description

The tokens with their disambiguated grammatical information enter the annotation of analytical syntax (Žabokrtský and Smrž, 2003, Hajič et al., 2004b), which is itself a precursor to the deep syntactic annotation (Sgall et al., 2004, Mikulová et al., 2006).

In Figures 6 and 7, one can compare both representations given the following sentence from our treebank:

(1) وفي ملف الأدب طرحت المجلة قضية اللغة العربية والأخطار التي تهددها.

Wa-fī milaffi ’l-ʾadabi ṭaraḥati ’l-maǧallatu qaḍīyata ’l-luġati ’l-ʿarabīyati wa-’l-ʾaḫṭāri ’llatī tuhaddiduhā.

‘In the section on literature, the magazine presented the issue of the Arabic language and the dangers that threaten it.’

4.1 Analytical Syntax

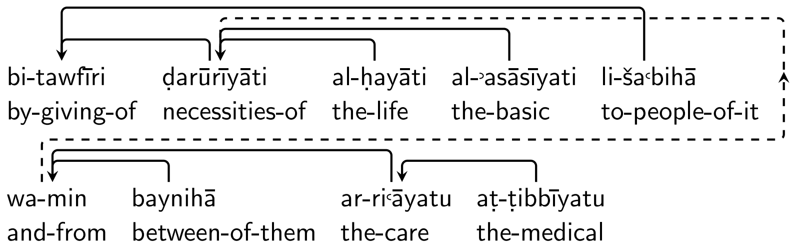
This level is formalized into dependency trees the nodes of which are the tokens. Relations between nodes are classified with analytical syntactic functions. More precisely, it is the whole subtree of a dependent node that fulfills the particular syntactic function with respect to the governing node.

Both clauses and nominal expressions can assume the same analytical functions—the attributive clause in our example is Atr, just like in the case of nominal attributes. Pred denotes the main predicate, Sb is subject, Obj is object, Adv stands for adverbial. AuxP, AuxY and AuxK are auxiliary functions of specific kinds.

The coordination relation is different from the dependency relation. We can, however, depict it in the tree-like manner, too. The coordinative node becomes Coord, and the subtrees that are the members of the coordination are marked as such (cf. dashed edges). Dependents modifying the coordination as a whole would attach directly to the Coord node, yet would not be marked as coordinants—therefrom, the need for distinguishing coordination and pure dependency in the trees.

The immediate-dominance relation that we capture in the annotation is independent of the linear ordering of words in an utterance, i.e. the linear-precedence relation (Debusmann, 2006). Thus, the expressiveness of the dependency grammar is stronger than that of phrase-structure context-free grammar. The dependency trees can become non-projective by featuring crossing dependencies, which reflects the possibility of relaxing word order while preserving the links of grammatical government.

(2) بتوفير ضروريات الحياة الأساسية لشعبها ومن بينها الرعاية الطبية



'by providing the basic necessities of life to its people, including medical care'

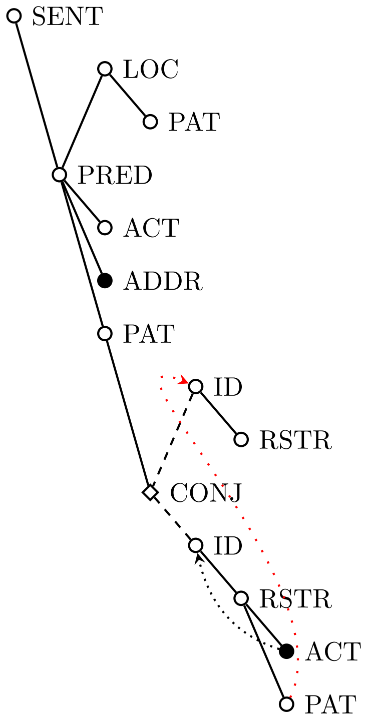
In example (2), a non-projective edge occurs between the word ḍarūrīyāti and its dependent, the relative attributive clause. In between of the two, there is the phrase li-šaʿbihā, which depends directly on bi-tawfīri and is not a descendant of ḍarūrīyāti, as a projective structure would require.

4.2 Tectogrammatics

We can note these characteristics of the representations of the underlying syntax:

deleted nodes only autosemantic lexemes and coordinative nodes are involved in tectogrammatics; synsemantic lexemes, such as prepositions or particles, are deleted from the trees and may instead be reflected in the values of deep grammatical categories, called grammatemes, associated with the relevant autosemantic nodes

inserted nodes autosemantic lexemes that do not appear explicitly in the surface syntax, yet that are demanded as obligatory by valency frames or by other criteria of tectogrammatical well-formedness, are inserted into the deep syntactic structures; the elided lexemes may be copies of other explicit nodes, or may be restored even as generic or unspecified



مِلَفّ milaff	collection	Masc.Sing.Def	*B*
أَدَب ʾadab	literature	Masc.Sing.Def	*C*
طَرَح ṭaraḥ	to present	Ind.Ant.Act	*B*
مَجَلَّة maǧallah	magazine	Fem.Sing.Def	*B*
هُوَ huwa	someone	GenPronoun	*B*
قَضِيَّة qaḍīyah	issue	Fem.Sing.Def	*N*
لُغَة luġah	language	Fem.Sing.Def	*N*
عَرَبِيّ ʿarabīy	Arabic	Adjective	*N*
وَ wa-	and	Coordination	
خَطَر ḫaṭar	danger	Masc.Plur.Def	*N*
هَدَّد haddad	to threaten	Ind.Sim.Act	*N*
هِيَ hiya	it	PersPronoun	*B*
هِيَ hiya	it	PersPronoun	*B*

FIGURE 7 Tectogrammatical annotation of example (1) with resolved coreference (extra arcs) and indicated values of contextual boundness. Lexemes are identified by lemmas, and selected grammatemes are shown in place of morphological grammatical categories.

functors are the tectogrammatical functions describing deep dependency relations; the underlying theory distinguishes *arguments* (inner participants, including: ACTor, PATient, ADDRessee, ORIGin, EFFect) and *adjuncts* (free modifications, such as: LOCation, CAUSe, MANNer, TimeWHEN, ReSTRictive, APPurtenance) and specifies the type of coordination (e.g. CONJunctive, DISJunctive, ADVerSative, ConSeQuential)

grammatemes are the deep grammatical features that are necessary for proper generation of the surface form of an utterance, given the tectogrammatical tree as well (cf. Mikulová et al., 2006, Hajič et al., 2004b)

coreference pronouns are matched with the lexical mentions they refer to; we distinguish *grammatical* coreference (the coreferent is determined by grammar) and *textual* coreference (otherwise); in Figure 7, the black dotted arcs indicate grammatical coreference, the loosely dotted red curves denote textual coreference

contextual boundness is the elementary distinctive feature from which the topic–focus dichotomy in a sentence is derived; as explained below, nodes can be contextually *B*ound, *C*ontrastively bound, or *N*on-bound

4.3 Describing Information Structure

The issue of information structure in language has been studied extensively both in the Prague School of Linguistics (Mathesius, 1929) and in the Functional Generative Description, one of the modern theories of representation of linguistic meaning (cf. Hajičová and Sgall, 2003, 2004).

In the flow of the discourse, the salience of the concepts that the interlocutors entertain changes and develops. Individual underlying components of each proposition differ in their *communicative dynamism*, in accordance with which the surface sentence is organized. The linguistic means for expressing the dynamism can include word order variation with respect to some prototypical systemic ordering, using of marked intonation and stress within an utterance, or employing extra constructs of the grammar.

Each sentence can be divided into two parts that exhibit the relation of aboutness. Topic (theme) is that part of sentence that links the content of the utterance with the context of the discourse. Focus (rheme, comment) is the other part that provides or modifies some information about the topic.

The *topic–focus dichotomy* is recognized, with varying terminology, in most theories of information structure (cf. Kruijff-Korbayová and Steedman, 2003). In the Praguian approach (Sgall et al., 1986, Kruijff-Korbayová, 1998), this distinction is understood as derived from the structural notion of contextual boundness and non-boundness:

context-bound lexical reference to an already *explicitly mentioned* entity, or to an entity *implicitly evoked* in the context of the discourse

non-bound lexical item that is not contextually bound, i.e. *not retrievable* in the interlocutor's mind *as reference*

One can use the so called question test to identify the context-bound and non-bound items. Let us assume that without breaking the felicitousness of the discourse, a question summarizing the preceding context is inserted immediately before the sentence whose boundness we study. Those items in the sentence that are also present in or implied by the question, are considered contextually bound, others are non-bound.

The relation of definiteness and boundness is not trivial and the notions cannot be interchanged (Kruijff-Korbayová, 1998, Brustad, 2000). Contextual boundness can neither be equated to the cognitive given/new opposition, due to the important possibility of implicitness in our definitions.

The topic–focus dichotomy can be determined recursively for a sentence and its clauses, and on every level of nesting, the following rules relating it to boundness apply (cf. Kruijff-Korbayová, 1998, Postolache, 2005):

1. the predicate node belongs to the focus if it is non-bound (value *N*), and to the topic if it is context-bound (values *B* or *C*)
2. the non-bound tectogrammatical nodes that depend directly on the predicate belong to the focus, and so do all their descendants
3. if the predicate and all of its direct dependents are context-bound, the focus is constituted by the more deeply embedded nodes that are non-bound, and all their descendants
4. all other nodes belong to the topic

Thus, based on information in Figure 7, the sentence of example (1) and its relative clause receive this annotation of focus (underlined):

(3) وفي ملف الأدب طرحت المجلة قضية اللغة العربية والأخطار التي تهددها.

Wa-fī milaffi ʼl-ʾadabi ṭaraḥati ʼl-maǧallatu qaḏīyata ʼl-luġati ʼl-ʿarabīyati wa-ʼl-ʾaḫṭāri allatī tuhaddiduhā.

‘In the section on literature, the magazine presented the issue of the Arabic language and the dangers that threaten it.’

The topic–focus articulation is relevant for semantic as well as pragmatic interpretation, as argued by many authors and treated in detail in (Kruijff-Korbayová, 1998). It is the focus of a sentence that becomes the scope of *focalizer* particles, adverbs of quantification or frequency, and prototypically also negation.

4.4 Annotation Examples

(El-Shishiny, 1990, Pedersen et al., 2004, Anoun, 2006) (Kruijff and Duchier, 2002, Nivre, 2005) (Žabokrtský, 2005, Lopatková et al., 2005)

Verbal clauses and coordination

In Figure 6, we see an example of a verbal sentence, including a verbal relative clause and a coordination, in the analytical dependency representation. The adverbial phrase precedes the main predicate due to the requirements of information structure—the contrastive context the sentence was used in. Word order in Arabic is relatively freer than what the classical VSO characterization would suggest—word order does reflect/express information structure and the prototypical ordering differs for verbal vs. nominal clauses as well as for main vs. subordinate clauses.

Figure 7 depicts the deep syntactic relations in the very sentence, i.e. the tectogrammatical structure and functors (in this presentation, we disregard deep word order rearrangements due to information structure). Note the differences in the set of nodes actually represented, esp. the restored ADDRessee which is omitted in the surface form of the sentence, but is obligatory in the valency frame of the semantics of the PREDicate.

Ellipsis and ‘inner objects’

(4) ودمر وفقا لإحصائية أولية عشرة منازل تدميرا كليا إضافة إلى خمسة عشر منزلا تدميرا جزئيا.

Wa-dummira wifqan li-ʾiḥṣāʾīyatin ʾawwalīyatin ʿašratu manāzila tadmīran kullīyan ʾiḍāfatan ʾilā ḫamsata ʿašara manzilan tadmīran ǧuzʾīyan.

‘And according to first statistics, ten houses were destroyed completely and fifteen partially.’

The sentence in Figure 8 exhibits ellipsis of the predicate—the sentence includes two propositions that share the verbal frame, yet each is instantiated with a different set of modifications. On the analytical level, the otherwise coordinative phrase إضافة إلى ʾiḍāfatan ʾilā is classified with ExD to mark the actual ellipsis. The adverbial phrase expressing extent is realized with the ‘inner object’, the تدميرا tadmīran, which is the deverbal noun of the predicate verb دمّر dammar. Note the red dashed arcs that indicate this fact.

On the tectogrammatical level, in Figure 9, 'inner objects' are removed and the EXTent is represented directly with the former dependents of each of these nodes. The elided nodes are restored by copying and linking them together to preserve their identity (cf. the loosely dashed red curves). Note how the passive voice affects the structures, and that quantifiers are represented as dependent RSTR modifiers, contrary to the analytical level.

Non-projectivity and complements

(5) ولم يكن من السهل عليه مواجهة كاميرات التلفزيون وعدسات المصورين وهو يصعد الباص.

Wa-lam yakun min as-sahli ʿalayhi muwāǧahatu kāmīrāti 't-tilfizyūni wa-ʿadasāti 'l-muṣawwirīna wa-huwa yaṣʿadu 'l-bāṣa.

'It was not easy for him to face the television cameras and the lenses of photographers as he was getting on the bus.'

Figure 10 depicts a sentence with a non-projective complement clause Atv expressing state. The subject of the clause is grammatically coreferring with the object of the main clause. The main predicate is the negated verb *to be* in the so called jussive mood, so there is no particular issue about it, unlike clauses without the verbal copula, cf. below.

The tectogrammatical tree in Figure 11 is projective already, as the COMPLement is attached directly to the head of the clause, and the reference to the original parent node is captured with the loosely dashed red curve.

The functors of the arguments of مواجهة muwāǧahah, in either of the Figures, respect the underlying verbal character of this gerund, i.e. the maṣdar as called in the Arabic linguistic terminology. The ACTor of the *facing* is coreferring with the BENefactor.

Non-verbal clauses and topicalization

(6) ويرى القائمون على الملف أن ما تتعرض له اللغة العربية له أهداف محددة منها ... .

Wa-yarā 'l-qāʾimūna ʿalā 'l-milaffi ʾanna mā tataʿarraḍu lahu 'l-luġatu 'l-ʿarabīyatu lahu ʾahdāfun muḥaddada-tun minhā ... .

'The ones in charge of the section are of the opinion that what the Arabic language is exposed to has its specific goals, including . . . .'

Figure 12 presents a rather complex objective clause featuring topicalization and non-verbal predication mediated by the preposition ل li- to express ownership or APPurtenance. The topicalized, or antepositioned, part includes the pronoun ما mā that is further modified by a relative clause. This subordinate clause, as well as the non-verbal clause itself, both include additional resumptive pronouns that are grammatically coreferring with the topicalized ما mā.

On the tectogrammatical level, in Figure 13, the missing verbal predicate is restored with the most generic كان kān. The resumptive pronoun that matches with a non-ancestor is removed, and its functions are transferred to the coreferent. Naturally, resumptive pronouns in relative clauses do not undergo such transformations.

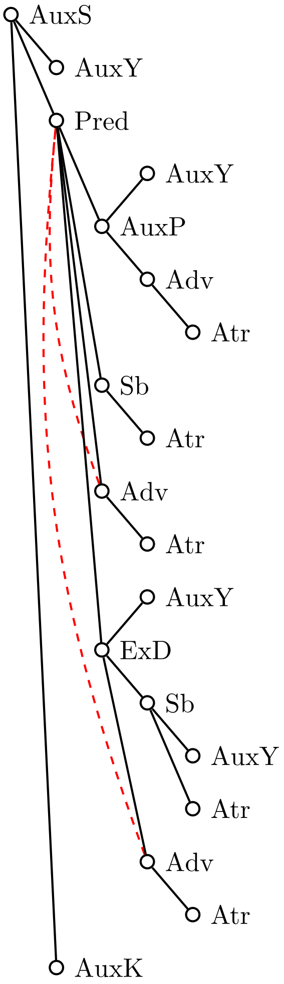
There is another instance of using non-verbal predication in this example. The sentence would continue with *including* . . . , which translates literally as *from them be* . . . . This introduces a new relative clause with another resumptive pronoun and the predicate كان kān inserted into the tectogrammatical tree.

In (Hajič et al., 2004b), we give some more examples of the tectogrammatical treatment of non-verbal predication. Note, however, that we now prefer not to distinguish between the predicative and possessive senses of كان kān by introducing distinct fictitious lexemes with the non-distinctive ACTor PATient valency frame—instead, as presented here, we rather capture the possessive sense by using the ACTor APPurtenance frame.

Wa-fī milaffi ’l-ʾadabi ṭaraḥat-i ’l-maǧallatu qaḍīyata ’l-luġati ’l-ʿarabīyati wa-’l-ʾaḫṭāri ’llatī tuhaddiduhā. wa-yarā ’l-qāʾimūna ʿalā ’l-milaffi ʾanna mā tataʿarraḍu lahu ’l-luġatu ’l-ʿarabīyatu lahu ʾahdāfun muḥaddada-tun minhā ʾibʿādu ’l-ʿarabi ʿan luġatihim wa-muzāḥamatu ’l-luġāti ’l-ġarbīyati lahā wa-huwa mā yaʿnī ḍuʿfa ’ṣ-ṣilati bihā wa-muḥāwalatu ʾizāḥati ’l-luġati ’l-fuṣḥā bi-kulli ’l-wasāʾili wa-ʾiḥlāli ’l-lahaǧāti ’l-muḫtalifati fī ’l-bilādi ’l-ʿarabīyati maḥallahā.

وفي ملف الأدب طرحت المجلة قضية اللغة العربية والأخطار التي تهددها. ويرى القائمون على الملف أن ما تتعرض له اللغة العربية له أهداف محددة منها إبعاد العرب عن لغتهم ومزاحمة اللغات الغربية لها وهو ما يعني ضعف الصلة بها ومحاولة إزاحة اللغة الفصحى بكل الوسائل وإحلال اللهجات المختلفة في البلاد العربية محلها.

In the section on literature, the magazine presented the issue of the Arabic language and the dangers that threaten it. The ones in charge of the section are of the opinion that what the Arabic language is exposed to has its specific goals, including the separation of Arabs from their language and the competition of the Western languages with it, which means weakness of the link to it, and the attempt to remove the literary language by all means and to replace it with the different dialects of the Arab world.



AuxS				
AuxY	وَ	wa-	and	C---------
Pred	دُمِّرَ	dummira	it-was-destroyed	VP-P-3MS--
AuxY	وِفقًا	wifqan	in-accordance	N-----MS4I
AuxP	لِ	li-	to	P---------
Adv	إِحصَائِيَّةٍ	ʾiḥṣāʾīyatin	statistics	N-----FS2I
Atr	أَوَّلِيَّةٍ	ʾawwalīyatin	an-initial	A-----FS2I
Sb	عَشرَةُ	ʿašratu	ten	N-----FS1R
Atr	مَنَازِلَ	manāzila	houses	N-------2I
Adv	تَدمِيرًا	tadmīran	destroying	N-----MS4I
Atr	كُلِّيًّا	kullīyan	a-complete	A-----MS4I
AuxY	إِضَافَةً	ʾiḍāfatan	in-addition	N-----FS4I
ExD	إِلَى	ʾilā	to	P---------
Sb	خَمسَةَ	ḫamsata	five	N-----FS--
AuxY	عَشَرَ	ʿašara	ten	N---------
Atr	مَنزِلًا	manzilan	a-house	N-----MS4I
Adv	تَدمِيرًا	tadmīran	destroying	N-----MS4I
Atr	جُزئِيًّا	ǧuzʾīyan	a-partial	A-----MS4I
AuxK	.	.	.	G---------

FIGURE 8 Analytical annotation of example (4).

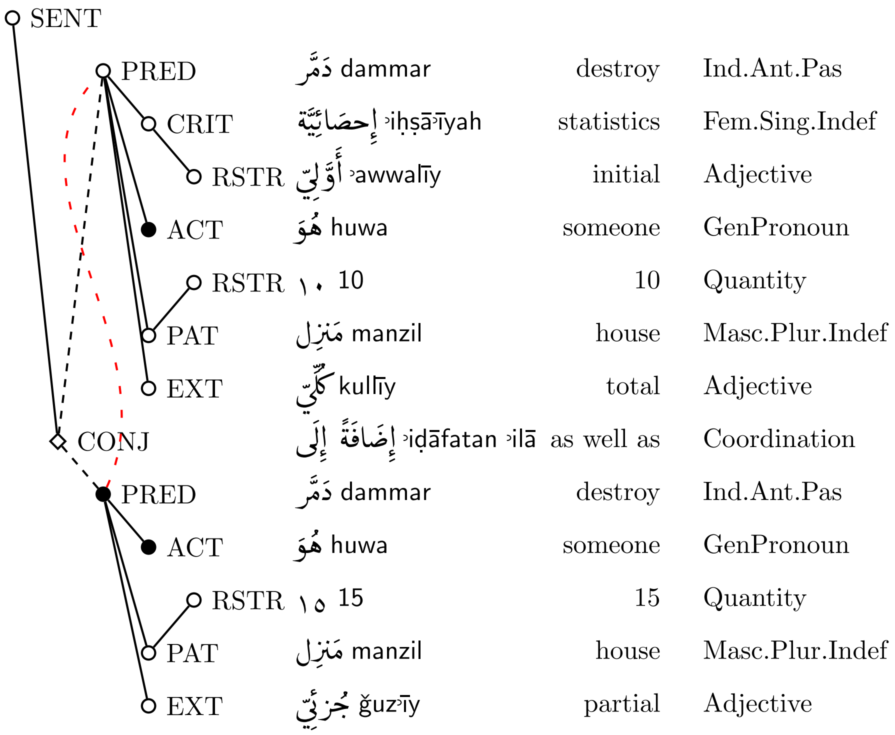
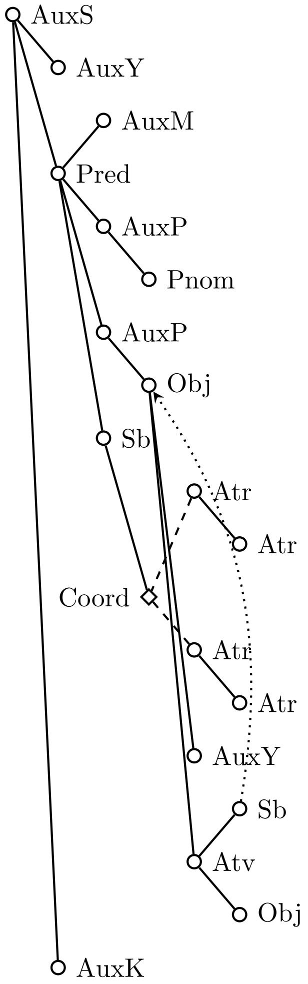
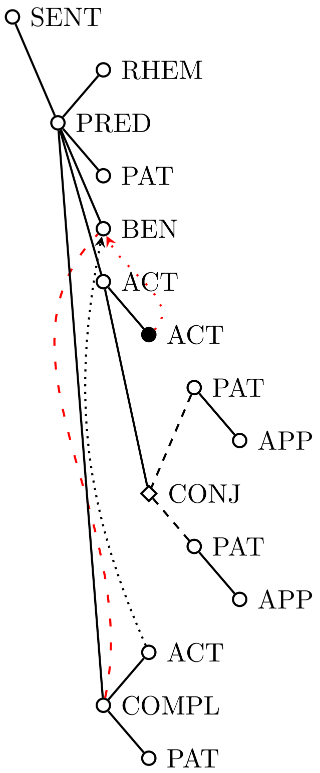

SENT
PRED دَمَّر dammar destroy Ind.Ant.Pas
CRIT إِحصَائِيَّة ʾiḥṣāʾīyah statistics Fem.Sing.Indef
RSTR أَوَّلِيّ ʾawwalīy initial Adjective
ACT هُوَ huwa someone GenPronoun
RSTR ١٠ 10 10 Quantity
PAT مَنزِل manzil house Masc.Plur.Indef
EXT كُلِّيّ kullīy total Adjective
CONJ إِضَافَةً إِلَى ʾiḍāfatan ʾilā as well as Coordination
PRED دَمَّر dammar destroy Ind.Ant.Pas
ACT هُوَ huwa someone GenPronoun
RSTR ١٥ 15 15 Quantity
PAT مَنزِل manzil house Masc.Plur.Indef
EXT جُزئِيّ ǧuzʾīy partial Adjective

FIGURE 9 Tectogrammatical annotation of example (4).



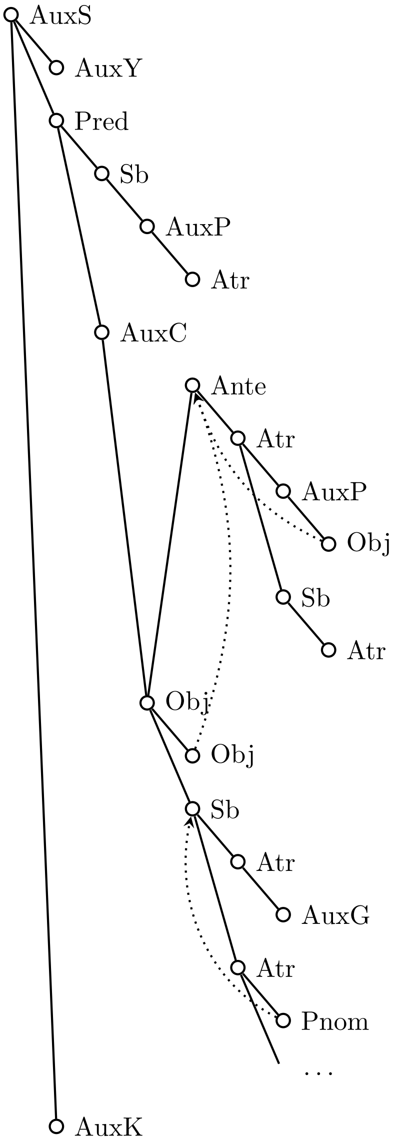
وَ wa-	and	C---------
لَم lam	did-not	FN--------
يَكُن yakun	it-be	VIJA-3MS--
مِن min	from	P---------
اَلسَّهلِ as-sahli	the-easy	A-------2D
عَلَي ʿalay	on	P---------
ـهِ -hi	him	S----3MS2-
مُوَاجَهَةُ muwāǧahatu	facing-of	N-----FS1R
كَامِيرَاتِ kāmīrāti	cameras-of	N-----FP2R
اَلتِّلفِزيُونِ at-tilfizyūni	the-television	N-------2D
وَ wa-	and	C---------
عَدَسَاتِ ʿadasāti	lenses-of	N-----FP2R
اَلمُصَوِّرِينَ al-muṣawwirīna	the-photographers	N-----MP2D
وَ wa-	and	C---------
هُوَ huwa	he	S----3MS1-
يَصعَدُ yaṣʿadu	he-gets-on	VIIA-3MS--
اَلبَاصَ al-bāṣa	the-bus	N-------4D
. .	.	G---------

FIGURE 10 Analytical representation of example (5).



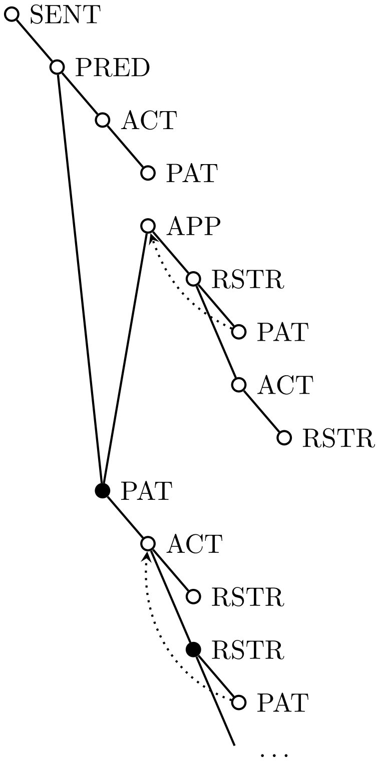
لَا lā	not	Negation
كَان kān	to be	Ind.Ant.Act
سَهل sahl	easy	Adjective
هُوَ huwa	he	PersPronoun
مُوَاجَهَة muwāǧahah	facing	Fem.Sing.Def
هُوَ huwa	he	PersPronoun
كَامِيرَا kāmīrā	camera	Fem.Plur.Def
تِلفِزيُون tilfizyūn	television	Masc.Sing.Def
وَ wa-	and	Coordination
عَدَسَة ʿadasah	lens	Fem.Plur.Def
مُصَوِّر muṣawwir	photographer	Masc.Plur.Def
هُوَ huwa	he	PersPronoun
صَعِد ṣaʿid	to get on	Ind.Sim.Act
بَاص bāṣ	bus	Masc.Sing.Def

FIGURE 11 Tectogrammatical annotation of example (5).



وَ	wa-	and	C---------
يَرَى	yarā	they-see	VIIA-3MS--
اَلقَائِمُونَ	al-qāʾimūna	the-ones-in-charge	N-----MP1D
عَلَى	ʿalā	of	P---------
اَلمِلَفِّ	al-milaffi	the-collection	N-------2D
أَنَّ	ʾanna	that	C---------
مَا	mā	what	SR--------
تَتَعَرَّضُ	tataʿarraḍu	it-is-exposed	VIIA-3FS--
لَ	la-	to	P---------
هُ	-hu	it	S----3MS2-
اَللُّغَةُ	al-luġatu	the-language	N-----FS1D
اَلعَرَبِيَّةُ	al-ʿarabīyatu	the-Arabic	A-----FS1D
لَ	la-	for	P---------
هُ	-hu	it	S----3MS2-
أَهدَافٌ	ʾahdāfun	goals/intentions	N-------1I
مُحَدَّدَةٌ	muḥaddadatun	specific	A-----FS1I
،	,	,	G---------
مِن	min	from	P---------
هَا	-hā	them	S----3FS2-
.	.	.	G---------

FIGURE 12 Analytical annotation of example (6).



رَأَى	raʾā	see	Ind.Sim.Act
قَائِم	qāʾim	being in charge	Masc.Plur.Def
مِلَفّ	milaff	collection	Masc.Sing.Def
مَا	mā	what	RelPronoun
تَعَرَّض	taʿarraḍ	to be exposed	Ind.Sim.Act
هُوَ	huwa	it	PersPronoun
لُغَة	luġah	language	Fem.Sing.Def
عَرَبِيّ	ʿarabīy	Arabic	Adjective
كَان	kān	to be	Ind.Sim.Act
هَدَف	hadaf	goal	Masc.Plur.Indef
مُحَدَّد	muḥaddad	specific	Adjective
كَان	kān	to be	Ind.Sim.Act
هِيَ	hiya	they	PersPronoun

FIGURE 13 Tectogrammatical annotation of example (6).

5 Conclusion

In PADT, which now consists of the morphological and the analytical levels of description of Arabic, the annotation of information structure and tectogrammatics is being established.

In our contribution, we have tried to overview the theoretical concepts we work with and the original implementations we develop, and to present our formal treatment of a number of corpus-based instances of interesting linguistic phenomena.

Acknowledgements

This research has been supported by the Ministry of Education of the Czech Republic, projects MSM0021620838 and MSM0021620823, by the Grant Agency of Charles University in Prague, project UK 373/2005, and by the Grant Agency of the Czech Academy of Sciences, project 1ET101120413.

References

Al-Sughaiyer, Imad A. and Ibrahim A. Al-Kharashi. 2004. Arabic Morphological Analysis Techniques: A Comprehensive Survey. *Journal of the American Society for Information Science and Technology* 55(3):189–213.

Anoun, Houda. 2006. Towards a Logical Approach to Nominal Sentences Analysis in Standard Arabic. In *Proceedings of the Eleventh ESSLLI Student Session*.

Badawi, Elsaid, Mike G. Carter, and Adrian Gully. 2004. *Modern Written Arabic: A Comprehensive Grammar*. Routledge.

Baerman, Matthew, Dunstan Brown, and Greville G. Corbett. 2006. *The Syntax-Morphology Interface. A Study of Syncretism*. Cambridge Studies in Linguistics. Cambridge University Press.

Bar-Haim, Roy, Khalil Sima'an, and Yoad Winter. 2005. Choosing an Optimal Architecture for Segmentation and POS-Tagging of Modern Hebrew. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 39–46. Ann Arbor, Michigan: Association for Computational Linguistics.

Beesley, Kenneth R. 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.

Brustad, Kristen E. 2000. *The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects*. Georgetown University Press.

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

Buckwalter, Tim. 2004a. Buckwalter Arabic Morphological Analyzer Version 2.0. LDC catalog number LDC2004L02, ISBN 1-58563-324-0.

Buckwalter, Tim. 2004b. Issues in Arabic Orthography and Morphology Analysis. In *Proceedings of the COLING 2004 Workshop on Computational Approaches to Arabic Script-based Languages*, pages 31–34.

Crochemore, Maxime, Costas S. Iliopoulos, Yoan J. Pinzon, and James F. Reid. 2000. A Fast and Practical Bit-Vector Algorithm for the Longest Common Subsequence Problem. In *Proceedings of the 11th Australasian Workshop On Combinatorial Algorithms*. Hunter Valley, Australia.

Cuřín, Jan, Martin Čmejrek, Jiří Havelka, Jan Hajič, Vladislav Kuboň, and Zdeněk Žabokrtský. 2004. Prague Czech-English Dependency Treebank 1.0. LDC catalog number LDC2004T25, ISBN 1-58563-321-6.

Debusmann, Ralph. 2006. *Extensible Dependency Grammar: A Modular Grammar Formalism Based On Multigraph Description*. Ph.D. thesis, Saarland University.

Diab, Mona, Kadri Hacioglu, and Daniel Jurafsky. 2004. Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. In *HLT-NAACL 2004: Short Papers*, pages 149–152. Association for Computational Linguistics.

Ditters, Everhard. 2001. A Formal Grammar for the Description of Sentence Structure in Modern Standard Arabic. In *EACL 2001 Workshop Proceedings on Arabic Language Processing: Status and Prospects*, pages 31–37. Toulouse, France.

El Dada, Ali and Aarne Ranta. 2006. Open Source Arabic Grammars in Grammatical Framework. In *Proceedings of the Arabic Language Processing Conference (JETALA)*. Rabat, Morocco: IERA.

El-Sadany, Tarek A. and Mohamed A. Hashish. 1989. An Arabic morphological system. *IBM Systems Journal* 28(4):600–612.

El-Shishiny, Hisham. 1990. A Formal Description of Arabic Syntax in Definite Clause Grammar. In *Proceedings of the 13th Conference on Computational Linguistics*, pages 345–347. Association for Computational Linguistics.

Finkel, Raphael and Gregory Stump. 2002. Generating Hebrew Verb Morphology by Default Inheritance Hierarchies. In *Proc. of the Workshop on Computational Approaches to Semitic Languages*, pages 9–18. Association for Computational Linguistics.

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

Forsberg, Markus and Aarne Ranta. 2004. Functional Morphology. In *Proceedings of the Ninth ACM SIGPLAN International Conference on Functional Programming, ICFP 2004*, pages 213–223. ACM Press.

Habash, Nizar and Owen Rambow. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 573–580. Ann Arbor, Michigan: Association for Computational Linguistics.

Hajič, Jan, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, and Barbora Vidová-Hladká. 2001. Prague Dependency Treebank 1.0. LDC catalog number LDC2001T10, ISBN 1-58563-212-0.

Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, and Marie Mikulová. 2006. Prague Dependency Treebank 2.0. LDC catalog number LDC2006T01, ISBN 1-58563-370-4.

Hajič, Jan, Otakar Smrž, Tim Buckwalter, and Hubert Jin. 2005. Feature-Based Tagger of Approximations of Functional Arabic Morphology. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, pages 53–64. Barcelona, Spain.

Hajič, Jan, Otakar Smrž, Petr Zemánek, Petr Pajas, Jan Šnaidauf, Emanuel Beška, Jakub Kráčmar, and Kamila Hassanová. 2004a. Prague Arabic Dependency Treebank 1.0. LDC catalog number LDC2004T23, ISBN 1-58563-319-4.

Hajič, Jan, Otakar Smrž, Petr Zemánek, Jan Šnaidauf, and Emanuel Beška. 2004b. Prague Arabic Dependency Treebank: Development in Data and Tools. In *NEMLAR International Conference on Arabic Language Resources and Tools*, pages 110–117. ELDA.

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

Hajičová, Eva and Petr Sgall. 2004. Degrees of Contrast and the Topic–Focus Articulation. In *Information Structure: Theoretical and Empirical Aspects*, vol. 1 of *Language, Context & Cognition*, pages 1–13. Berlin: Walter de Gruyter.

Hodges, Wilfrid. 2006. Two doors to open. In *Mathematical Problems from Applied Logic I: Logics for the XXIst century*, pages 277–316. New York: Springer.

Holes, Clive. 2004. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown Classics in Arabic Language and Linguistics. Georgetown University Press.

Hudak, Paul. 2000. *The Haskell School of Expression: Learning Functional Programming through Multimedia*. Cambridge University Press.

Huet, Gérard. 2002. The Zen Computational Linguistics Toolkit. ESSLLI Course Notes, FoLLI, the Association of Logic, Language and Information.

Huet, Gérard. 2003. Lexicon-directed Segmentation and Tagging of Sanskrit. In *XIIth World Sanskrit Conference*, pages 307–325. Helsinki, Finland.

Huet, Gérard. 2005. A Functional Toolkit for Morphological and Phonological Processing, Application to a Sanskrit Tagger. *Journal of Functional Programming* 15(4):573–614.

Karttunen, Lauri. 2003. Computing with Realizational Morphology. In *CICLing Conference on Intelligent Text Processing and Computational Linguistics*, pages 205–216. Springer Verlag.

Kay, Martin. 2004. Arabic Script-Based Languages Deserve to Be Studied Linguistically. In *COLING 2004 Computational Approaches to Arabic Script-based Languages*, page 42. Geneva, Switzerland.

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

Konz, Ned and Tye McQueen. 2000–2006. Algorithm::Diff. Programming module registered in the Comprehensive Perl Archive Network, `http://search.cpan.org/dist/Algorithm-Diff/`.

Kremers, Joost. 2003. *The Arabic Noun Phrase. A Minimalist Approach*. Ph.D. thesis, University of Nijmegen. LOT Dissertation Series 79.

Kruijff, Geert-Jan M. and Denys Duchier. 2002. Formal and Computational Aspects of Dependency Grammar. ESSLLI Course Notes, FoLLI, the Association of Logic, Language and Information.

Kruijff-Korbayová, Ivana. 1998. *The Dynamic Potential of Topic and Focus: A Praguian Approach to Discourse Representation Theory*. Ph.D. thesis, Charles University in Prague.

Kruijff-Korbayová, Ivana and Mark Steedman. 2003. Discourse and Information Structure. *Journal of Logic, Language and Information* 12(3).

Lagally, Klaus. 2004. ArabTeX: Typesetting Arabic and Hebrew, User Manual Version 4.00. Tech. Rep. 2004/03, Fakultät Informatik, Universität Stuttgart.

Lopatková, Markéta, Martin Plátek, and Vladislav Kuboň. 2005. Modeling Syntax of Free Word-Order Languages: Dependency Analysis by Reduction. In *Proceedings of Text, Speech and Dialogue*, LNCS/LNAI, pages 140–147. Springer Verlag.

Maamouri, Mohamed and Ann Bies. 2004. Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. In *Proceedings of the COLING 2004 Workshop on Computational Approaches to Arabic Script-based Languages*, pages 2–9.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Mathesius, Vilém. 1929. Functional Linguistics. In *Praguiana: Some Basic and Less Known Aspects of the Prague Linguistic School*, pages 121–142. Amsterdam: John Benjamins.

Mikulová, Marie et al. 2006. A Manual for Tectogrammatical Layer Annotation of the Prague Dependency Treebank. Tech. rep., Charles University in Prague.

Nivre, Joakim. 2005. Dependency Grammar and Dependency Parsing. Tech. Rep. MSI Report 05133, Växjö University, School of Mathematics and Systems Engineering.

Panevová, Jarmila. 1980. *Formy a funkce ve stavbě české věty [Forms and Functions in the Structure of the Czech Sentence]*. Academia.

Pedersen, Mark, Domenyk Eades, Samir K. Amin, and Lakshmi Prakash. 2004. Relative Clauses in Hindi and Arabic: A Paninian Dependency Grammar Analysis. In *COLING 2004 Recent Advances in Dependency Grammar*, pages 9–16. Geneva, Switzerland.

Postolache, Oana. 2005. Learning Information Structure in the Prague Treebank. In *Proceedings of the ACL Student Research Workshop*, pages 115–120. Ann Arbor, Michigan: Association for Computational Linguistics.

Sgall, Petr. 1967. *Generativní popis jazyka a česká deklinace [Generative Description of Language and Czech Declension]*. Academia.

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

Sgall, Petr, Jarmila Panevová, and Eva Hajičová. 2004. Deep Syntactic Annotation: Tectogrammatical Representation and Beyond. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 32–38. Association for Computational Linguistics.

Smrž, Otakar. 2003–2007. Encode::Arabic. Programming module registered in the Comprehensive Perl Archive Network, `http://search.cpan.org/dist/Encode-Arabic/`.

Smrž, Otakar. 2007a. ElixirFM — Implementation of Functional Arabic Morphology. In *ACL 2007 Proceedings of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 1–8. Prague, Czech Republic: ACL.

Smrž, Otakar. 2007b. *Functional Arabic Morphology. Formal System and Implementation*. Ph.D. thesis, Charles University in Prague.

Smrž, Otakar and Petr Pajas. 2004. MorphoTrees of Arabic and Their Annotation in the TrEd Environment. In *NEMLAR International Conference on Arabic Language Resources and Tools*, pages 38–41. ELDA.

Smrž, Otakar, Petr Pajas, Zdeněk Žabokrtský, Jan Hajič, Jiří Mírovský, and Petr Němec. 2007. Learning to Use the Prague Arabic Dependency Treebank. In E. Benmamoun, ed., *Perspectives on Arabic Linguistics*, vol. XIX. John Benjamins.

Spencer, Andrew. 2004. Generalized Paradigm Function Morphology. `http://privatewww.essex.ac.uk/~spena/papers/GPFM.pdf`.

Stump, Gregory T. 2001. *Inflectional Morphology. A Theory of Paradigm Structure*. Cambridge Studies in Linguistics. Cambridge University Press.

Versteegh, Kees. 1997. *The Arabic Language*. Edinburgh University Press.

Wadler, Philip. 1997. How to Declare an Imperative. *ACM Computing Surveys* 29(3):240–263.

Žabokrtský, Zdeněk. 2005. *Valency Lexicon of Czech Verbs*. Ph.D. thesis, Charles University in Prague.

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