Information Structure with the Prague Arabic Dependency Treebank

Otakar Smrž, Petr Zemánek, Jakub Kráčmar, Viktor Bielický

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague
padt@ufal.mff.cuni.cz

1 Introduction

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 (FGD), one of the modern theories of representation of linguistic meaning (Sgall, 1967; Sgall et al., 1986; Hajičová and Sgall, 2003, 2004).

In its entirety, FGD constitutes the framework for a family of projects in computational linguistics concerned with explicit multi-level annotation of linguistic resources, which include the Prague Dependency Treebank (PDT) for Czech (Hajič et al., 2001, 2006) as well as the Prague Arabic Dependency Treebank (PADT) (Hajič et al., 2004a; Smrž et al., 2006).

Information structure—the question of "the given" and "the new" in an utterance and how it is expressed—is recognized as an important component of the communicative function of language and is considered to influence the meaning of a discourse. Its annotation in PDT is part of the third, the most detailed and abstract level of linguistic description, called *tectogrammatical*. Next to determining which elements in a sentence are context-bound and which are non-bound (the elementary distinctive feature from which the topic–focus dichotomy is derived), attention is also paid to capturing the

communicative dynamism of a proposition by introducing ordering on the participants of its deep syntactic structures (cf. Mikulová et al., 2006).

In PADT, which now consists of the *morphological* and the *analytical* levels of description of Modern Written Arabic, a similar annotation of information structure is being established. In this contribution, we would like to overview the theoretical concepts we work with, and present our formal treatment of several prototypical, yet corpus-based, instances of linguistic phenomena that have their role in the study of the structure of information in Arabic (cf. Brustad, 2000; Holes, 2004).

The applicability of the general approach to written as well as spoken Arabic will be the main point of our account. Theoretical works on information structure, including those by the Prague School, incorporate also the notions of intonation center and sentence prosody, contrast, subjective word order, or potential ellipsis, which are considered as manifestations of the deeper formal model of information structure.

In this document, the following conventions are used: *italics* for phonetic transcription of Arabic in the ZDMG style, `typewriter` for Buckwalter transliteration of the script, sans serif for linguistic glosses and *slanted* for translations.

2 Prague Dependencies and Functions

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.

The formal representation delivers the linguistic meaning of what is expressed by the surface realization, i.e. the natural language. The description is also designed to enable synthesizing 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.

The linguistic analysis takes place in three stages: the morphological level (inflection of lexemes), the analytical level (surface syntax), and the tec-

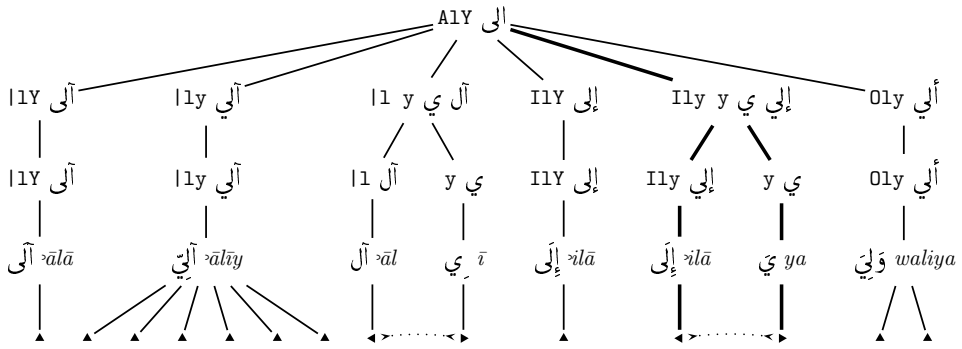


Figure 1: MorphoTrees of analyses of the orthographic word AlY الى and its spelling variants. The morphological tags in the leaves are schematized to triangles. The bold lines in the hierarchy indicate the human annotation, i.e. the choice of the morphological solution Ily y إلي ي *>ilay-ya 'to me'*.

togrammatical level (underlying syntax). Annotation of information structure is best associated with the tectogrammatical structures.

2.1 Morphological Annotation

The first step in our formal analysis of written (or even, transcribed spoken) language is the recovery of the grammatical categories that the word forms carry in the context, and of the subsuming lexemes of these forms.

Thus, from a non-vocalized Arabic text, we obtain the abstract information that is relevant for further processing of the discourse, and for syntactic analysis in particular. Moreover, morphological analysis can be reversed into generation in most computational morphological models. Due to that, we can produce the phonologically qualified, fully vocalized version of the text as another result.

Morphologically annotated data are used as training examples for taggers, which are systems that can do automatic morphological analysis and its context-aware disambiguation. There is a number of taggers already developed for Arabic on the basis of treebanks (Habash and Rambow, 2005; Smith et al., 2005; Hajič et al., 2005).

Morphological analysis in PADT is pioneering the MorphoTrees technique (Smrž and Pajas, 2004; Smrž, in prep.). For every word form found in a text, MorphoTrees organize the list of its possible morphological readings into a hierarchy, and allow the annotator to systematize and speed up his/her selecting of the one analysis that is appropriate in the context.

Figure 1 illustrates this further. The analyzed orthographic word constitutes the root of the hierarchy, the full forms and morphological tags of the analyzing syntactic tokens project into its leaves. Lexemes occupy the first level above the leaves, then there is the level of canonical non-vocalized spelling of the tokens, and the level of partitioning of the original word into such token forms.

2.2 Analytical Syntax

The tokens with their disambiguated grammatical information enter the annotation of analytical syntax (Žabokrtský and Smrž, 2003; Hajič et al., 2004b).

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.

In Figure 2, we analyze 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.'

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.

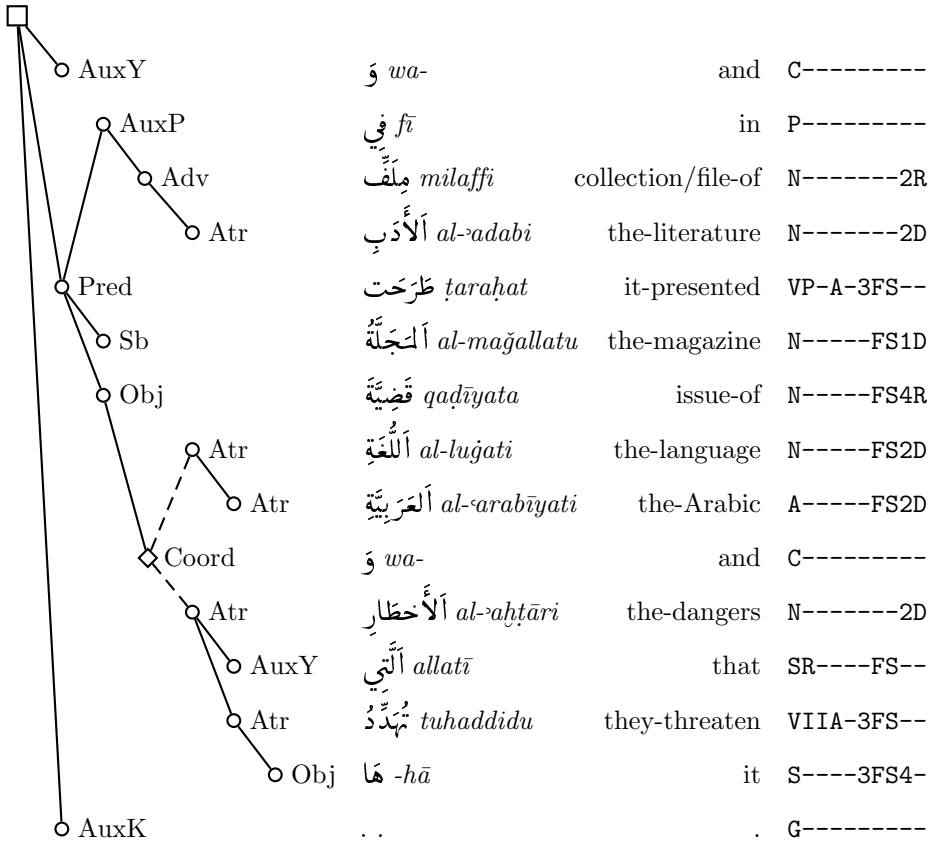


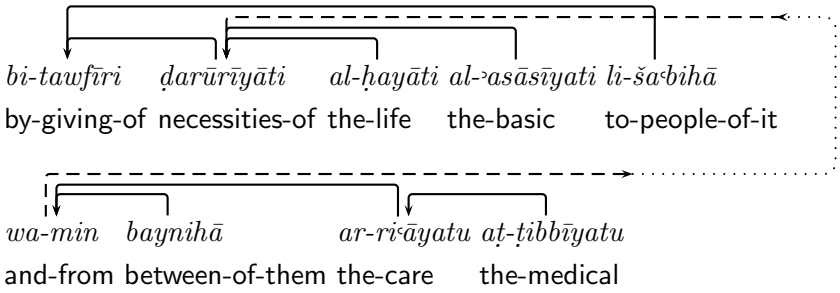
Figure 2: Analytical annotation of example (1). Grammatical categories are encoded using the positional notation explained in (Hajič et al., 2005).

The coordination relation is different from the dependency relation, however, we can 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 et al., 2005). 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.

2.3 Tectogrammatics

The analytical syntax is yet a precursor to the deep syntactic annotation (Hajičová and Sgall, 2003; Sgall et al., 2004; Mikulová et al., 2006). We can note these characteristics of the tectogrammatical level, and compare the representations of example (1) in Figure 2 and Figure 3:

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 reflect in the values of deep grammatical categories, called *grammatemes*, that are associated with the relevant autosemantic nodes

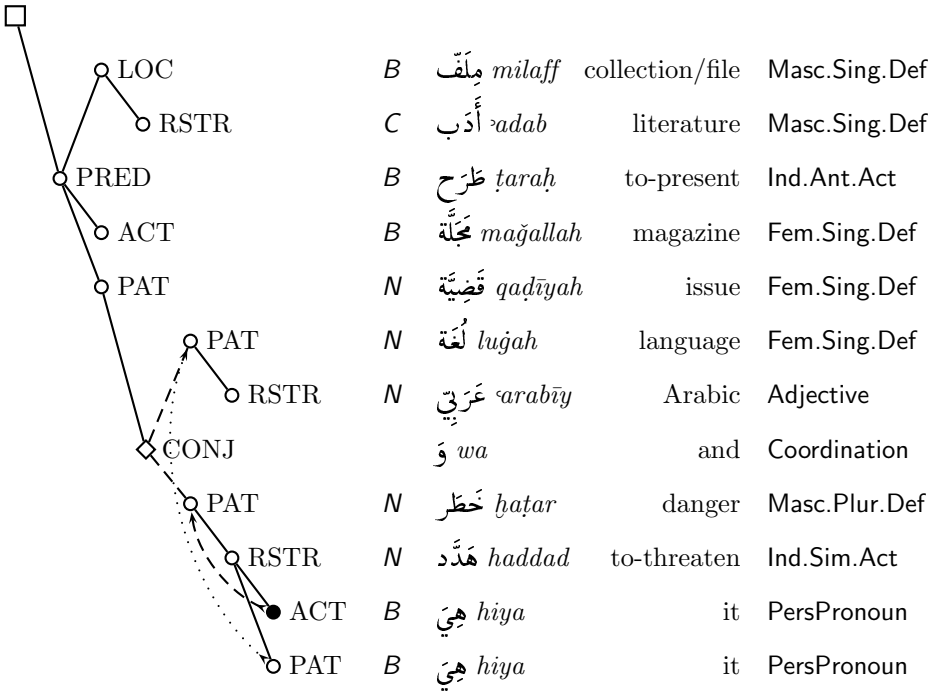


Figure 3: Tectogrammatical annotation of example (1) with resolved coreference (extra arcs) and indicated values of contextual boundness. Lexemes are identified by lemmas, and selected grammatemes are shown in place of morphological grammatical categories (compare with tags in Figure 2).

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

functors are the tectogrammatical functions describing deep dependency relations; the underlying theory distinguishes *arguments* (inner participants: ACTor, PATient, ADDRessee, ORIGin, EFFect) and *adjuncts* (free modifications, e.g.: 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. Hajič et al., 2004b; Smrž, in prep.)

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 3, the pairs are rendered using dashed and dotted arcs for each respective type

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

3 Describing Information Structure

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 in the syntax or morphology.

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 (for an overview, cf. e.g. Kruijff-Korbayová and Steedman, 2003). Yet 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 3, 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 '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.'

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.

3.1 Systemic Ordering

The systemic ordering as such can be viewed as a standard, unmarked ordering of the predicate and its participants in a sentence. Such an ordering yields a normal flow of information, unless a particular context interferes with it. In this section, we will deal with this issue more extensively, since for Arabic, only a little has been published about it (cf. e.g. Holes (2004, esp. p. 250 ff.), Mohammad (2000) and Shlonsky (1997)).

Intuitively for an Arabist, under the label *systemic ordering*, the first thing to come to mind is the standard order of sentence constituents given by the syntactic typology. Arabic is generally viewed as the VSO type of language, at least for the verbal sentences, and one should add that the usual word order for nominal (non-verbal, non-copular) sentence is Subject and Nominal predicate.

It is, however, clear, that such a view in case of Arabic holds especially for Modern Standard Arabic. However, as the definition of systemic ordering is language-dependent, or, in case of Arabic, also dialect-dependent, it has to be set for every dialect or dialect group individually—in other words, the dialects of Arabic and the MSA do not share the same systemic ordering. This has been stated in many studies concerning both the MSA and the dialects of Arabic. But even for the MSA, there are many sentence types that at least from the statistical point of view are almost as common as the two types mentioned above.

Most of the studies that deal with the word order in Arabic concentrate on the order of the basic constituents (verb, subject, object), quite a lot

of them use speculative examples, continuing the tradition of Arabic grammarians. Such examples are in many cases relevant, but on the other hand, corpus-based examples may bring up different sentence structures that appear in the current usage.

It should be further noted that the above mentioned structures (esp. the VSO word order) can be considered labels that give a general structure, but as such give quite a little information on more complex sentences, which are quite common in Arabic, both written and spoken. Such general labels do not meet the requirements of a more minute description of the syntax of Arabic—they do not cover other types of substructures that can be present in a sentence, such as various types of attributive or adverbial clauses, to mention the most common ones. These can be called *free modifications*. A typical ordering of such items within a sentence should also be studied.

As an example of the interplay between information structure and systemic ordering, consider the following sentence that is different from what we can find in most of the treatises on the word order in Arabic, yet a sentence quite typical in Arabic newspaper texts. The sentence is divided into chunks reflecting the arguments and free modifications in the main clause patterned by the verb توقّع *tawaqqaʿ 'to expect'*, as well as in the object clause patterned by انخفض *inḫafaḍ 'to diminish'*. The chunks' overall analytical functions and tectogrammatical functors are given, and bracketing indicates the dependency nesting of the chunks.

(4) a. في غضون ذلك توقع اتحاد الغرف التجارية

Fī ġuḍūni ḏālika tawaqqaʿa ittiḥādu 'l-ġurafi 't-tiǧārīyati
(Adv / Time PARallel) PREDicate (Sb / ACTor)
'In the meantime the Union of the Chambers of Trade expected'

b. أن تنخفض قيمة الصادرات إلى الولايات المتحدة

ʾan tanḫafiḍa qīmatu 's-ṣādirāti ʾilā 'l-wilāyāti 'l-muttaḥidati
(Obj / PATient (Sb / ACTor)
'that the value of exports to the United States will diminish'

c. في النصف الثاني من السنة الجارية

fī ’n-niṣfi ’l-ṯānī min-a ’s-sanati ’l-ǧāriyati
(Adv / Time WHEN)
‘in the second half of the current year’

d. إلى ٤٠٠ مليون دولار من ٥٩٣ مليون دولار حاليا

›ilā 400 milyūni dūlārin min 593 milyūna dūlārin ḥālīyan
(Obj / PATient ← EFFect) (Obj / ORIGin)
‘to 400 million dollars from 593 million dollars at present’

e. بسبب تداعيات الأحداث.

bi-sababi tadāʿiyāti ’l-›aḥdāṯi .
(Adv / CAUSe))
‘because of associations of the events.’

The functor PATient ← EFFect means that a participant that semantically would be understood as the EFFect of the action expressed by the predicate, fulfills linguistically the role of the PATient participant. This is known as the *actant-shifting principle* of the valency theory of FGD.

The questions that arise with example (4) might include: Is the ordering of the contents of (4c) and (4d) significant for the message that is delivered? What is the most communicatively dynamic participant of the object clause? Why does PATient ← EFFect precede ORIGin in (4d)? If the two phrases were swapped, would their functors be PATient ← ORIGin and EFFect, or would they remain unchanged, or what would they be?

For some more insight, let us have a look at the behavior of prepositional phrases introduced by من *min ‘from’* and إلى *›ilā ‘to’*. Such prepositional phrases are intuitively perceived as naturally forming a sequence starting with the *min* phrase and continuing with the *›ilā* phrase. Reversing this order is usually perceived as a signal of a change in a standard flow of information, a change in the ordering of the deep-syntactic participants in the sentence, and thus a change of the ordering of functions fulfilled by these participants.

The analysis of our corpus shows that from the statistical point of view, the above mentioned ordering of the two prepositional phrases works well. In majority of cases found in the corpus, the order of *min* before *ʾilā* is the one that was found, cf. example (5) below. E.g., in cases of such phrases as من حين إلا حين *min ḥīnin ʾilā ḥīnin 'from time to time'* in example (6), one cannot think of reversing the order of the items; reversing the order is very unusual with time reference (a period from ... to ...); in case of a reference to a place, several examples of reversed order can be found, cf. example (7).

(5) إن قرارات نقل المعلمين من التدريس إلى الأعمال الإدارية ...

ʾinna qarārāti naqli 'l-muʿallimīna min-a 't-tadrīsi ʾilā 'l-ʾaʿmāli 'l-ʾidārīyati

verily decisions-of moving-of the-teachers from the-teaching to the-work the-administrative ...

'the decisions to move the teachers from teaching to administrative work ...'

(6) يتم القبض عليهم من حين إلى حين

yatimmu 'l-qabḍu ʿalayhim min ḥīnin ʾilā ḥīnin

is-performed the-seizure on-them from some-time to some-time

'they are arrested from time to time'

(7) عودة الشيخ إلى الكويت من لندن

ʿawdatu 'š-šayḫi ʾilā 'l-kuwayti min lundun

return-of the-sheikh to Kuwait from London

'the return of the sheikh to Kuwait from London'

However, the situation can also considerably change with different lexemes. The data from the corpus show that in case of verbs and verbal nouns derived from the root خفض *ḫfḍ*, such as خفض *ḫafaḍ 'to decrease'*, خفض *ḫafḍ*

'lowering, decrease', انخفض *inḫafaḍ 'to decrease'*, انخفاض *inḫifāḍ 'lowering, decrease'* is usually reversed (see examples below). This means that the ordering of such elements can be dependent also on the valency frame of the particular verb. Or, to make this statement even more general, the ordering may depend on the valency characteristics of the lexical unit.

3.2 Expressing Dynamism

Even the textbooks of Arabic say that this language can easily change its word order, which has its impact on the structure of information yielded by the changed sentence. There are several types of syntactic construction that can be viewed as signaling a change in the flow of information.

In most of such structures, we find also words (usually function ones) that are generally called topicalizers or rhematizers/focalizers that help to introduce the respective piece of information.

The most common topicalizers are: إنّ *ʾinna 'verily, truly'*, أنّ *ʾanna 'that'*, أمّا *ʾammā 'as to, as for, as far as'*, the most common rhematizers are: ف *fa- 'then, and then, and so, so that'*, بل *bal 'rather, even'*, فقط *faqaṭ 'only'*, etc., but we could also add some phrases on this list, such as بما فيه *bi-mā fīhi 'including'*, بعبارة أخرى *bi-ʿibāratin ʾuḫrā 'in other words'*, بعبارة أحسن *bi-ʿibāratin ʾaḥsana 'better said'*, بعبارة أدقّ *bi-ʿibāratin ʾadaqqa 'more precisely'*, etc. It should be also noted that negation usually serves as rhematizer, too.

As an example, the prototypical rhematizer in Arabic can be considered. The particle ف *fa- 'so, then'* functioning as a conjunction is interesting also in connection with its function as a "subject switcher" in medieval texts written in Classical Arabic. In a way, such a function can be viewed as a substitution for punctuation. The *fa-* retained its function of introducing new, contextually unbound information also in the MSA. A prototypical example is the usage of *fa-* in the structure … ف … أمّا *ʾammā … fa- …*, where أمّا *ʾammā* is used for introducing the topic (topicalizer) and ف as introducing the focus (rhematizer). Other uses of *fa-* can be also viewed as typically introducing new information, too—cf. examples below.

(8) *ʾammā ʾīrānu ... fa-tuʿāriḍu ʾayya ziyādatin lil-ʾintāǧi*
as-to Iran ... then-opposes any-of increase to-the-production
'as for Iran ..., it opposes any increase in the production'

(9) *ʾiḏā ḥaṣala ḏālika fa-sa-yakūnu ʿamalan ẖaṭiran*
if happened that then-will-be act dangerous
'if that happens, (then) it will be a dangerous act'

(10) *rafaḍa 'l-ʾadillata fa-lam yuṣdir ʾaḥkāman mušaddadatan*
refused the-evidences then-not issued judgments severes
'he refused the evidence and did not pass severe judgments'

The conjunction و *wa-* *'and'* on the other hand, renders linkage to the before mentioned information. It is often used to show continuity with the previous information flow. Also its usage at the beginning of a new sentence, which is very common in Arabic, can be viewed as an expression of a continuity of the information flow from previous sentence (cf. also its function marking the sentence boundaries in Classical Arabic). It can also stand in opposition as a topicalizer to the rhematizer ف *fa-*.

(11) وتم تدوير الهواء كل تلاث دقائق لذا فإن فرصة التقاط الفيروس شديد الانخفاض

wa-tamma tadwīru 'l-hawāʾi kulla talāṯi daqāʾiqa li-ḏā fa-ʾinna furṣata 'ltiqāṭi 'l-fīrūs šadīdu 'l-inẖifāḍi

and-finished circulation the-air(acc) every three minutes for-that then-verily opportunity picking-up(gen) the-virus strong the-lowering

'and the circulation of air has been performed every three minutes which significantly diminished the opportunity of being infected by the virus'

Another example of a rhematizer in Arabic is the particle بل *bal* *'even'*, which is opposed to preceding affirmative or negative proposition, a command or a prohibition.

(12) إن المبادرة ليست عملا فرديا ... بل هي عمل يسهم في تأسيس دفاع مشترك

ʾinna ʾl-mubādarata laysat ʿamalan fardīyan ... bal hiya ʿamalun yashumu fī taʾsīsi difāʿin muštarakin

verily the-initiative not-is action individual ... even she action participates in founding defence collective

'and the initiative is not an individual act ... moreover, it is an act which helps in founding collective defence'

The particles mentioned above function as rhematizers mainly when used as conjunctions. It should be mentioned that in Arabic, after these conjunctions, the standard structure of the sentence is retained, which means that even after these conjunctions in a vast majority of cases at least a formal pointer (semantically empty function word) to the topic of the previous sentence (such as هي *hiya 'she'* in example (12) referring to the subject of the previous sentence) is also present in the sentence (or clause) introduced by these conjunctions.

Most of the rhematizers are rather function words, but content words functioning as rhematizers can be found, too, although in such a use its semantic independence may be seen as somewhat restricted. As an example of such a word, cf. the following examples of the usage of the verb يعني *yaʿnī* (ما يعني *mā yaʿnī*):

(13) وتراكمت فوائد الديون ... ما يعني أن ذلك سيؤثر في أي أرباح تحققها الشركة

wa-tarākamat fawāʾidu ʾd-duyūni ... mā yaʿnī ʾanna d̲ālika sa-yuʾat̲tiru fī ʾayyi ʾarbāḥin tuḥaqqiqu-hā ʾš-šarikatu

and-accumulated interests the-debts ... which means that this FUT-will-influence in any profits realize-her the-company

'and the interest of the debts accumulated ... which means that this will influence any profit that the company will make'

(14) إن ذلك يعني استمرار ارتفاع كلفة الاستمرار المباشر

ʾinna ḏālika yaʿnī ʾstimrāra ʾrtifāʿi kulfati ʾl-istimrāri ʾl-mubāširi

verily that means continuing rising costs the-continuing the-direct

'that means the continuation of the increase in the costs of the direct continuation'

The scope of various rhematizers in a sentence or clause is limited and to a great extent depends on the position of the rhematizer in a particular sentence. As an example, we have chosen the particle فقط *faqaṭ 'only'*. This particle as such can appear in several positions in a sentence. Somewhat outside the frame of rhematizing functions is its function with numbers, especially with financial operations, such as

(15) *dafaʿa ʾalfa dūlārin faqaṭ*
paid.he thousand dollars only
'he paid one thousand dollars only/exactly'

Such a meaning is, however, limited to the domain of financial operations and most probably it will not appear in spoken language. Other instances include the appearance of *faqaṭ* bound to the predicate or appearing after the phrase it limits.

(16) ليس فقط يدفع تنفيذ الاستراتيجية ... بل يوفر أيضا المعلومات لتنفيذ المشاريع

laysa faqaṭ yadfaʿu tanfīḏa ʾl-istrātīǧīyati ... bal yuwaffiru ʾayḍan al-maʿlūmāti li-tanfīḏi ʾl-mašārīʿi

not-be only pays.he realization the-strategy ... but will-provide.he also the-informations for-realization the-projects

'not only will he pay the implementation of the strategy ... but he will also provide information for the project implementation'

(17) يسعا فقط إلى وقف العنف من دون النظر إلى حقوق الشعب

yasʿā faqaṭ ʾilā waqfi ʾl-ʿunfi min dūni ʾn-naẓari ʾilā ḥuqūqi ʾš-šaʿbi

attempts.he only to stopping the-violence from without the-look to rights the-people

'he is only trying to stop the violence with no respect to the rights of the people'

(18) *yuḥaqqiqu ṣāliḥa ʾl-mustaṯmirīna wa-riǧāli ʾl-ʾaʿmāli faqaṭ*
realize.he benefit the-investors and-people the-works only
'he acts only in the interests of the investors and businessmen'

(19) *kāna taqtaṣiru fī ʾl-māḍī ʿalā ʾl-ḥizbi ʾl-ḥākimi faqaṭ*
was confines.she in the-past on the-party the-ruling only
'it was usually confined in the past to the ruling party only'

4 Conclusion and Future Work

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.

Annotated corpora for written and spoken Arabic are becoming available for quantitative evaluation of linguistic theories, large-scale analysis of linguistic material, computational processing and modeling.

In our contribution, we have tried to overview the theoretical concepts we work with, and present our formal treatment of a number of corpus-based instances of linguistic phenomena that have a principal impact on the structure of information in Arabic.

Rich linguistic literature and interesting computational systems are available (cf. e.g. Hajičová et al., 1995; Kruijff-Korbayová, 1998; Hajičová and Sgall, 2004; Debusmann et al., 2005; Mikulová et al., 2006).

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

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

Ralph Debusmann, Oana Postolache, and Maarika Traat. A Modular Account of Information Structure in Extensible Dependency Grammar. In *Proceedings of the CICLING 2005 Conference*, 2005.

Nizar Habash and Owen Rambow. 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, 2005. Association for Computational Linguistics.

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

Jan Hajič, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, and Barbora Vidová-Hladká. Prague Dependency Treebank 2.0. LDC catalog number LDC2006T01, ISBN 1-58563-370-4, 2006.

Jan Hajič, Otakar Smrž, Tim Buckwalter, and Hubert Jin. 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, 2005.

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

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

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

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

Eva Hajičová, Petr Sgall, and Hana Skoumalová. An Automatic Procedure for Topic–Focus Identification. *Computational Linguistics*, 21(1):81–94, 1995.

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

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

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

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

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

Mohammad A. Mohammad. *Word Order, Agreement and Pronominalization in Standard and Palestinian Arabic*. John Benjamins, 2000.

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

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

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

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

Ur Shlonsky. *Clause Structure and Word Order in Hebrew and Arabic. An Essay in Comparative Semitic Syntax*. Oxford University Press, 1997.

Noah A. Smith, David A. Smith, and Roy W. Tromble. Context-Based Morphological Disambiguation with Random Fields. In *Proceedings of HLT/EMNLP 2005*, pages 475–482, Vancouver, 2005. Association for Computational Linguistics.

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

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

Otakar Smrž, Petr Pajas, Zdeněk Žabokrtský, Jan Hajič, Jiří Mírovský, and Petr Němec. Learning to Use the Prague Arabic Dependency Treebank. In *Perspectives on Arabic Linguistics*, volume XIX. John Benjamins, 2006.

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