Building the Valency Lexicon of Arabic Verbs

Viktor Bielický **Otakar Smrž**

Institute of Formal and Applied Linguistics, Charles University in Prague
Malostranské náměstí 25, Prague 1, 118 00, Czech Republic
padt@ufal.mff.cuni.cz

Abstract

This paper describes the building of a valency lexicon of Arabic verbs using a morphologically and syntactically annotated corpus, the Prague Arabic Dependency Treebank, as its primary source. We present the theoretical account on valency developed within the Functional Generative Description theory. We apply the framework to Arabic and discuss various valency-related phenomena with respect to examples from the corpus. We then outline the methodology and the linguistic and technical resources used in the building of the lexicon. Valency lexicons can find application in automatic parsing as well as in language generation.

1. Introduction

Valency of a lexical unit, in particular a verb, is a set of its obligatory and/or optional arguments potentially or actually realized in an utterance. Valency information is useful in restoring the syntactic structure of an utterance, and has consequences for the study of the meaning.

The goal of this paper is to prepare the theoretical (Sections 2 and 3) and methodological (Sections 4 and 5) background for creating the valency lexicon of the most frequent Arabic verbs, exploiting various resources of information. Our approach is inspired by the VALLEX lexicon of Czech verbs (Lopatková et al., 2006, 2008) and its treebank-oriented twin project, the PDT-VALLEX (Hajič et al., 2003, 2006).

In our case, we focus on Modern Standard Arabic (MSA) and take as reference the **Prague Arabic Dependency Treebank** (PADT). It provides refined linguistic annotations whose multi-level description scheme discerns functional morphology, analytical dependency syntax, and tectogrammatical representation of linguistic meaning. The current, largely extended version of PADT (cf. Hajič et al., 2004; Smrž, 2007) covers over one million words of text.

2. Theory of Valency in FGD

Before we focus on some issues concerning verbal valency in Arabic and our proposed methodology for creating the valency lexicon, let us briefly outline the theoretical framework we have adopted. The **Functional Generative Description** (FGD) theory, which has been elaborated since the sixties of the last century (in particular in Sgall et al., 1986; Hajičová and Sgall, 2003), is a multi-stratal dependency-oriented description of language. The valency theory of verbs has been thoroughly researched within the framework of FGD since the seventies (Panevová, 1974, 1975, 1994; Lopatková and Panevová, 2005). The question of valency is closely associated with the underlying tectogrammatical level of language description representing the meaning of the discourse.

According to the valency theory of FGD, valency information of the given verb is defined by the valency frame—the sequence of frame slots—which is filled by a specific number of various valency complements, i.e. a variety of either required or specifically permitted syntactic units dependent on a verb. Each verb has at least one valency frame. The exact number of valency frames depends on the number of different meanings of the particular verb. For expressing relations between a verb and its complements, FGD uses various functors. These functors are divided into actants (inner participants, arguments) and free (adverbial) modifications (adjuncts). The entire number of actants is five (for examples in English, see Table 1):

Actant	Meaning	Example
ACT	Actor	*Peter* read a letter.
ADDR	Addressee	Peter gave *Mary* a book.
PAT	Patient	I saw *him*.
EFF	Effect	We made her *the secretary*.
ORIG	Origin	She made a cake *from apples*.

Table 1: Types of actants (inner participants) illustrated on English sentences (Lopatková et al., 2006: xvi).

Adjunct	Meaning
MANN	Manner
MEANS	Means
LOC	Location

DIR1	Direction from
DIR3	Direction to
TWHEN	Time when
THO	Time how often

Table 2: Types of adjunct (free modifications) appearing in this paper. For the complete list, cf. (Mikulová et al., 2006).

ACTor – usually the agent (the surface subject) or the bearer of some property/quality; **PATient** – the goal/target or the object affected by the action with consequences for its morphemic representation (the case in inflectional languages) brought about by verbal government (usually the direct object of transitive verbs); **ADDRessee** – usually the indirect object on the surface; **ORIGin** – this participant is probably never obligatory; **EFFect** – usually the second (inanimate) object, the predicative complement or the adverbial of result.

As regards the actants, they have to fulfill two conditions. The first condition is that the set of certain actants is characteristic for a particular verb—in other words, not every actant can depend on every verb. The second is that every actant can occur only once as a complement of the given verb, disregarding coordination or apposition.

On the contrary, there are different kinds of free modifications denoting various types of adverbial complementation (e.g. time, location, direction, manner, aim, cause, regard, accompaniment). These free modifications can appear more than once with a single verb and theoretically can modify any verb. It means that they are actually not restricted to a certain group of verbs, as is the case with actants. For examples of free modifications, see Section 3. The verbal valency frame in its narrow sense consists of both obligatory and optional inner participants and obligatory free modifications (see Table 3). The criterion of obligatoriness or optionality of verbal complements was introduced in the dialogue test by (Panevová, 1974, 1975) with respect to possibility to intentionally omit a contextually bound obligatory complement on the surface morphemic level of representation through the ellipsis or, for instance, as a general ("dummy") subject or object, etc.

	obligatory	optional
inner participants (actants)	+	+
free modifications (adjuncts)	+	−

Table 3: Members included in the valency frames.

It is to be stated that the approach adopted by the FGD takes into account both syntactic and semantic criteria for assigning functors to verbal complements (contrary to other more semantically-based approaches). Within this approach the concept of "shifting of cognitive roles" was adopted (Panevová, 1974, 1975, 1994). This "shifting" denotes application of primarily syntactic criteria for identifying the first two actants (Actor and Patient). Due to this fact, the first actant of the given verb is always identified as Actor and the second one as Patient regardless of their actual semantics. On the contrary, semantic criteria are applied when assigning functors to other actants as well as to all free (adverbial) modifications of a verb. For the concept of "shifting" see Figure 1. Some examples of "shifting" will be illustrated on Arabic in Section 3.

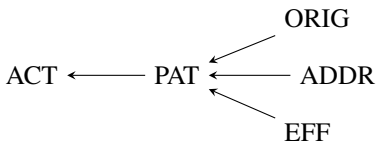


Figure 1: Shifting of cognitive roles as a criterion for assigning functors to actants (inner participants) of a verbal frame (Panevová, 1994: 234).

The valency frames as appeared in the valency lexicon of Czech verbs VALLEX are enriched with two other sets of complements, namely quasi-valency complements and typical complements (Lopatková and Panevová, 2005; Lopatková, 2003). The former quasi-valency type (consisting of newly introduced Obstacle and Difference and of revised previously existing complements Intention and Mediator) is the kind of complement lying somewhere in between the free modifications and actants, while the latter typical type denotes optional free modification usually co-occurring with a particular verb. Those complements will not be taken into consideration in our present study of Arabic and will be the subject of our further research. In this preliminary phase of our research, we adopt the valency frames in their narrow sense, i.e. including obligatory and optional actants and only obligatory free modifications as has been pointed at in Table 3.

3. Valency in Arabic: Preliminary Overview

In this section, we will adapt some aspects of the above mentioned theoretical approach of FGD for Arabic in order to make our preliminary observations about the valency behavior of Arabic verbs and their verbonominal derivatives. The only elaborate work on valency in Arabic which has come to our knowledge is (Al-Qahtani, 2004). Contrary to FGD, al-Qahtani has adopted predominantly semantic approach, since he deals with verbal valency in terms of Case Grammar theory. He applies the Matrix Model of (Cook, 1979) to the semantic classification of Arabic verbs (state, action, and process verbs). To each class a specific set of required semantic complements ("deep cases") is assigned—namely Agent, Experiencer, Benefactive, Object, and Locative. The obligatory Object is omnipresent with every verb (in contrast to Actor in FGD) and can occur more than once in a case frame. Experiencer, Benefactive, and Locative are mutually exclusive. Sometimes, a particular case is not realized on the surface ("covert case role"), i.e. it is either partially covert ("deletable") or totally covert ("coreferential" or "lexicalized"). Those deletable case roles can be omitted on the surface (optional or elided complements in terms of FGD, see (X) and Table 4), whereas the so-called coreferential and lexicalized case roles are always absent from the surface. The former coreferential roles denote instances where a single noun cumulates two case roles simultaneously (not permitted in FGD, see (Y)), while the latter lexicalized roles include instances where a certain case role (usually Object) is incorporated in the semantics of the verb (see (Z)). No shift of case roles takes place in this approach. (Al-Qahtani, 2004: 148, 178)

(X) *qāla Zaydun maqūlata-hu* he-said Zayd said-of-him
Zayd said what he had to say
qāl AEO/E-del (Experiencer is deleted)

(Y) *darasa Zaydun al-kitāba* he-studied Zayd the-book
Zayd studied the book
daras AEO/A=E (Agent equals Experiencer)

(Z) *ʿamila Zaydun* he-worked Zayd
Zayd worked = Zayd did some work
ʿamil AO/O-lex (Object is lexicalized)

(subject)	*li-* (prep.)	*ʿan* (prep.)	4–*ʾinna* (conj.)
ACT^{obl}	ADDR^{opt}	PAT^{opt}	EFF^{obl}
someone	to someone	about sth.	something/that

Table 4: Valency frame of the verb *qāl* قَالَ 'to say'.

3.1. Verbal valency

First, let us demonstrate some basic issues postulated by the FGD approach. In all the following examples in this section, the complements highlighted in bold are considered

to be obligatory, the others are optional. Some examples derived from available corpora had to be abridged.

In case that the verbal valency frame consists of only one inner participant, it is always Actor, whatever the semantics of that complement would be. Here, the syntactic criteria play the major role in assigning the functor to a complement. Those verbs are typically intransitive stative (1) or passive/reflexive (2).

(1) *kāna yanāmu ʿādatan fī šāriʿin ṣaġīrin*
he-was ⌜he$^{\mathbf{ACT}}$-sleeps ⌜usually$^{\text{THO}}$ ⌜in a-street a-small$^{\text{LOC}}$
he usually slept in a small street

(2) *intaḥara bi-taṣwībi-hi ʾl-musaddasa ʾilā raʾsi-hi*
⌜he$^{\mathbf{ACT}}$-commited-a-suicide ⌜by-aiming-of-him the-gun at head-of-him$^{\text{MEANS}}$
he committed a suicide by aiming the gun at his head

If the valency frame includes two actants, the first actant is considered to be Actor and the other Patient. Some verbs are directly transitive (3), whereas others are transitive indirectly through a preposition (4).

(3) *ʿaqadat-i ʾl-laǧnatu ʾl-munaẓẓimatu muʾtamaran ṣiḥāfīyan ʾawwala min ʾamsi*
it-held ⌜the-committee the-organizational$^{\mathbf{ACT}}$ ⌜a-conference a-press$^{\mathbf{PAT}}$ ⌜a-first-day from yesterday$^{\text{TWHEN}}$
the organizational committee held a press conference the day before yesterday

(4) *naǧaḥa ʿulamāʾu faransīyūna fī ʾstinsāḫi ʾarāniba*
he-succeeded ⌜scientists French$^{\mathbf{ACT}}$ ⌜in cloning-of rabbits$^{\mathbf{PAT}}$
French scientists succeeded in cloning rabbits

In case of verbs with three or more actants (no matter if obligatory or optional) where Patient is from the semantic viewpoint not realized in the valency frame, the above mentioned preference of syntactic criteria is applied. This means that the other actant (EFF, ORIG, or ADDR) undergoes the "shift" (see Section 2., esp. Figure 1) to occupy the unfilled slot of Patient. To the remaining actants, functors are assigned according to the semantic criteria. In example (5), Effect undergoes the shift to Patient.

(5) *taḥawwalat-i ʾl-munaẓẓamatu min ʾadāti muwāǧahatin ʾilā ʾadātin li-ʾl-baḥṯi*
it-changed ⌜the-organization$^{\mathbf{ACT}}$ ⌜from instrument-of a-confrontation$^{\text{ORIG}}$ ⌜to an-instrument for-the-research$^{\mathbf{PAT}}$ (EFF → PAT)
the organization changed from an instrument of confrontation to an instrument for research

In the examples below, some verbs with three and more actants are illustrated. Valency frames with ACT, ADDR, PAT:

(6) *samaḥa la-hu bi-ʾd-duḫūli ʾilā ʾl-bayti*
⌜he$^{\mathbf{ACT}}$-permitted ⌜to-him$^{\mathbf{ADDR}}$ ⌜by-the-entering into the-house$^{\mathbf{PAT}}$
he permited him to enter the house

(7) *šārakat zawǧa-hā fī ʾl-ḥukmi*
⌜she$^{\mathbf{ACT}}$-shared ⌜husband-of-her$^{\mathbf{ADDR}}$ ⌜in the-reign$^{\mathbf{PAT}}$
she shared the reign with her husband

It is worth mentioning that the usual word order of double transitive verbs as *ʿaṭā* IV 'to give' (=both objects (ADDR, PAT) are in accusative (8)) can be used regularly also in the reversed position (PAT, ADDR). In that case, the indirect object (i.e., ADDR) appears with the preposition *li-* (9).

(8) *ʿaṭā-hu ʾl-furṣata*
⌜he$^{\mathbf{ACT}}$-gave-⌜him$^{\mathbf{ADDR}}$ ⌜the-opportunity$^{\mathbf{PAT}}$
he gave him the opportunity

(9) *ʿaṭat-i ʾs-sayṭarata li-ʾl-bunūki*
⌜it$^{\mathbf{ACT}}$-gave ⌜the-power$^{\mathbf{PAT}}$ ⌜to-the-banks$^{\mathbf{ADDR}}$
it gave the power to the banks

Valency frames with ACT, PAT, EFF. The following verbs are also double transitive:

(10) *ʿayyana-hu ḥākiman li-ʾl-Kuwayti*
⌜he$^{\mathbf{ACT}}$-appointed-⌜him$^{\mathbf{PAT}}$ ⌜a-ruler for-Kuwait$^{\mathbf{EFF}}$
he appointed him as a ruler of Kuwait

(11) *iʿtabara ʾAdūnīs ʿūrūbīyan*
⌜he-considered$^{\mathbf{ACT}}$ ⌜Adonis$^{\mathbf{PAT}}$ ⌜a-European$^{\mathbf{EFF}}$
he considered Adonis to be European

Valency frames with four actants ACT, PAT, ORIG, EFF:

(12) *tarǧama ʾakṯara min ḫamsīna kitāban min-a ʾl-fārisīyati ʾilā ʾl-ʿarabīyati*
⌜he$^{\mathbf{ACT}}$-translated ⌜a-more than fifty a-book$^{\mathbf{PAT}}$ ⌜from Persian$^{\text{ORIG}}$ ⌜into Arabic$^{\text{EFF}}$
he translated more than fifty books from Persian into Arabic

(13) *ġayyarat [aš-šarikātu] našāṭa-hā min ʾintāǧi ʾl-qamḥi ʾilā ʾintāǧi ʾl-buḏūri*
⌜it$^{\mathbf{ACT}}$-changed [the companies] ⌜activity-of-it$^{\mathbf{PAT}}$ ⌜from production-of the-wheat$^{\mathbf{ORIG}}$ ⌜to production-of the-seeds$^{\mathbf{EFF}}$
[the companies] changed their activity from the production of grain to the production of seeds

In the following examples, let us mention some verbal valency frames that comprise some type of free (adverbial) modifications.

(14) *badʾu ʾl-ḥarbi waḍaʿa-hu ʾamāma ʾamrin wāqiʿin*
⌜beginning-of the-war$^{\mathbf{ACT}}$ it-put-⌜him$^{\mathbf{PAT}}$ ⌜in-front-of a-thing a-real$^{\mathbf{DIR3}}$
the beginning of the war put him in front of the reality

(15) *ʿādat min-a ʾl-Qāhirati ʾilā Bayrūta*
⌜she$^{\mathbf{ACT}}$-returned ⌜from Cairo$^{\text{DIR1}}$ ⌜to Beirut$^{\mathbf{DIR3}}$
she returned from Cairo to Beirut

It should be pointed out that we make a difference in verbal frames when assigning a functor to a verbal complement that could be semantically regarded as a free modification (e.g. some directional meaning), but on the surface level this complement is the direct object in accusative. In this case, the syntactic (or morpho-syntactic) viewpoint (verbal government affects the morphemic form of a complement, i.e. the criterion of direct transitivity) is preferred, and consequently a functor of Patient is assigned as in sentence (16). If there are two (or more) different morphemic realizations on the surface (i.e. prepositional phrase versus direct verbal government), although the meaning of that verb is in both cases the very same, two (or more) different valency frames are distinguished ((17) with DIR3 and (18) with PAT).

(16) *ġādarat-i ʾl-Qāhirata ʾilā Tall ʾAbīb*

⌜she$^{\textbf{ACT}}$-left ⌜Cairo$^{\textbf{PAT}}$ ⌜to Tel Aviv$^{\text{DIR3}}$
she left Cairo for Tel Aviv

(17) *waṣala 'l-muntaḫabu ʔilā madīnati Sālīrnū 'l-ʔiṭālīyati*
it-arrived ⌜the-representation$^{\textbf{ACT}}$ ⌜to town-of Salerno the-Italian$^{\textbf{DIR3}}$
the representation arrived to the Italian town Salerno

(18) *waṣaltu-hā [Dimašq] min-a 'd-Dawḥati*
⌜I$^{\textbf{ACT}}$-arrived-to-⌜it$^{\textbf{PAT}}$ [Damascus] ⌜from ad-Dawha$^{\text{DIR1}}$
I arrived there [to Damascus] from ad-Dawha

On the contrary, when a more abstract meaning of a particular verb occurs, the complement is no longer considered to be a free modification (directional meaning) and both (or more) variants—that with a prepositional phrase and the other with a direct object—are regarded as morphemic variants of the same actant (Patient in this case) within one single valency frame (19) and (20).

(19) *waṣalat qīmatu-hā ʔilā ḫamsati yūrūhātin*
it-reached ⌜value-of-it$^{\textbf{ACT}}$ ⌜to five euros$^{\textbf{PAT}}$
its value reached 5 euros

(20) *waṣalat qīmatu 'ṣ-ṣādirāti 625 milyūna dūlārin*
it-reached ⌜value-of the-exports$^{\textbf{ACT}}$ ⌜625 million-of a-dollar$^{\textbf{PAT}}$
the value of export reached 625 million dollars

When dealing with verbal valency in MSA, some issues concerning diathesis should be briefly discussed as well. MSA (contrary to Arabic dialects), as the successor of Classical Arabic, has preserved one of its characteristic features—regularly formed passive by changing the vowel pattern of active verb (so-called inflectional, internal or apophonic passive)—which is usually used when the agent of an action is not known or is preferred not to be mentioned. With some rare exceptions, only transitive verbs[1] undergo passivisation, no matter if they are transitive directly or indirectly through the preposition. In the passive, the position of the underlying Actor is reduced and the understood surface object (usually PAT or ADDR) of the active verb becomes a subject (Agameya, 2008: 558). However, besides this type of diathetical transformation, another type of passive exists in Arabic, namely "a derivational passive, where a derivational verb form (typically V, VII, or VIII) is used to convey a passive, reflexive or mediopassive sense of the action involved in the verb" (Ryding, 2005: 657). Those cases are then, as a result of derivation through verbal morpho-semantic patterns, autonomous lexicalized passive or passive-related verbs with their own valency frames with more or less probable word-formational relation to some active verb (causative, factitive, etc.) they are derived from, cf. Figure 2.

With directly transitive verb, Actor is reduced and the direct object (Patient) becomes a grammatical subject (compare to the active voice in (3)).

(21) *wa-lam yuʿqad ʾayyu muʾtamarin ṣiḥāfīyin muštarakin*
and-not it-was-held ⌜any a-conference a-press a-joint$^{\textbf{PAT}}$
and no joint press conference was held

With double transitive verbs, those with complements Patient and Effect, the first object (PAT) substitutes the grammatical (surface) subject while the second object (EFF) remains in accusative (compare to the active voice (10)).

(22) *ʿuyyina 'd-duktūru Mawṣilī wakīlan li-kullīyati 'ṭ-ṭibbi*
he-was-appointed ⌜the-doctor Mawsili$^{\textbf{PAT}}$ ⌜an-assistant-dean to-faculty-of the-medicine$^{\textbf{EFF}}$
doctor Mawsili was appointed as an assistant dean of the faculty of medicine

It is to be pointed out that those double transitive verbs with complements ADDR and PAT (verbs as *ʿaṭā* 'to give') might be passivized in two ways, either the former object usually referred to as indirect (23) or the latter direct object (24) can substitute the grammatical (surface) subject (Agameya, 2008: 559) (compare the active voice (8) and (9)).

(23) *ʿuṭiyat furṣatan ṯāniyatan*
⌜she$^{\textbf{ADDR}}$-was-given ⌜a-chance a-second$^{\textbf{PAT}}$
she was given the second chance

(24) *ʿuṭiya 'ḍ-ḍawʾu 'l-ʾaḫḍaru li-'l-malikati*
it-was-given ⌜the-light the-green$^{\textbf{PAT}}$ ⌜to-the-queen$^{\textbf{ADDR}}$
the queen was given the green light

In case of indirectly transitive verbs through prepositions when passivized, the verb itself always remains in the 3rd person masculine singular in the passive while the surface subject (the previous object of the active verb) goes after the preposition (Badawi et al., 2004: 387–388; Ryding, 2005: 666–667).

(25) *yuḥkamu ʿalay-hi bi-'s-siǧni*
it-is-sentenced ⌜upon-him$^{\textbf{PAT}}$ ⌜by-the-jail$^{\textbf{EFF}}$
he is sentenced to jail

Sometimes, the agent (Actor on the underlying level of representation) of some verb in the passive voice is expressed periphrastically after particular prepositional phrases like *min qibali* 'by, on part of', etc. (Badawi et al., 2004: 385–386; Retsö, 2006: 624–625).

(26) *šarikātun tudāru min qibali mudarāʾa mutaḫaṣṣiṣīna*
⌜companies$^{\textbf{PAT}}$ is-managed ⌜from directions-of managers specialist$^{\textbf{ACT}}$
companies are managed by professional managers

It is worth mentioning that several Arabic verbs in passive voice have undergone some kind of semantic shift and are used figuratively. Due to this fact they have to be considered as idiomatic, since they are no longer real semantic counterparts of their active forms. We can mention at this point two very frequent verbs (27) *tuwuffiy* V 'to die, to pass away' and (28) *ustušhid* X 'to die as a martyr' reflecting some degree of euphemism in connection with religious feeling. Those verbs would be treated in our proposed lexicon as separate word entries.

(27) *tuwuffiya wālidu-hu fī ḥādiṯi sayyāratin*
he-died ⌜father-of-him$^{\textbf{ACT}}$ ⌜in accident-of a-car$^{\text{MANN}}$
his father died in a car accident

(28) *ustušhida ʿāma 1991 fī ḥādiṯi 'ǧtiyālin*
⌜he$^{\textbf{ACT}}$-died-as-a-martyr ⌜year-of 1991$^{\text{TWHEN}}$ ⌜in

[1] Some intransitive verbs, esp. those of movement, become transitive secondarily when taking a preposition, e.g. *ǧāʾ* 'to come' → *ǧāʾ bi-* 'to bring sth.'; *qām* 'to stand up' → *qām bi-* 'to carry out sth.' (Badawi et al., 2004: 382–383; Drozdík, 2001).

event-of an-assassination$^{\text{MANN}}$
he died as a martyr in 1991 in an assassination

Reflexivity is expressed in MSA either lexically by derivational morpho-semantic classes (especially reflexive/passive forms V, VII, VIII and some reflexive meanings of X) or periphrastically using the nouns *nafs* 'a self, a soul' or *ḏāt* 'a self' in the position of all possible verbal complements. Some verbs with reflexive meaning have been already illustrated in examples (2) and (5). Another closely related phenomena to reflexivity, i.e. reciprocity, is expressed in MSA either by derived morpho-semantic form VI (29) (then lexicalized) or by an anaphoric means using the noun *baʿḍ* 'part, some' (Badawi et al., 2004: 391–394).

(29) *taʿāqadat maʿa 'l-banki 'l-ʾislāmīyi bi-Dubay ʿalā ʾidārati funduqi ʾAbū Ẓabī*
⌜it$^{\textbf{ACT}}$-came-to-the-agreement ⌜with the-bank the-islamic in-Dubay$^{\textbf{ADDR}}$ ⌜on managing-of hotel-of Abu Dhabi$^{\text{PAT}}$
it came to the agreement with the Islamic Bank in Dubay on managing the Abu Dhabi Hotel

The last issue related to diathesis to some degree concerns a certain type of verbs that are sometimes regarded as impersonal owing to the fact that they frequently seem to be neutralized in gender (3rd person singular masc.), since they disagree with their grammatical subject (modal verbs of necessity (*wağab* I, etc.) and possibility (*ʾamkan* IV)) (McCarus, 1976: 17; Al-Qahtani, 2004: 114–118; Badawi et al., 2004: 395–398, 418–419). These verbs have a verbal noun (30) or subordinate clause after the conjunction *ʾan* (31) as a surface subject which corresponds according to our approach to Actor on the underlying level. They have their Patient, if realized on the surface, after the preposition (usually *ʿalā*) or as a direct object (when pronoun) or after preposition *li-* (when a noun) in case of *ʾamkan* IV.

(30) *yumkinu-hu 'd-duẖūlu fī 'l-ḥarbi*
it-enables-⌜him$^{\textbf{PAT}}$ ⌜the-entering in the-war$^{\textbf{ACT}}$
it is possible for him to enter the war

(31) *yağibu ʿalā qūwāti 't-taḥālufi ʾan tuḥīṭū bi-hāḏā 'l-baladi*
⌜it-is-necessary on forces-of the-alliance$^{\textbf{PAT}}$ ⌜that they-encircle by-this the-country$^{\textbf{ACT}}$
the ally forces have to encircle that country

According to the FGD approach, auxiliary verbs do not have their own valency frames, as they only modify the main verb. These verbs (e.g. "*kān* and her sisters") are beyond the scope of this paper.

3.2. Valency of verbonominal derivatives

Valency—a potential of a particular lexical unit to bind some other syntactic element(s)—is not limited only to verbs but it is the property of all autosemantic words, i.e. including nouns and adjectives.

In case of Arabic, studying the verbal valency should be very fruitful with respect to its verbonominal derivatives—a participle (active and passive) and a verbal noun (*maṣdar*). Not only can these bear similar syntactic function as the verb (e.g. active participle as a predicate), but in many cases they preserve the same or almost the same valency frame as the verb they are derived from. In this respect, we have to emphasize that almost all verbonominal derivatives (except for verbal nouns of form I) are in Arabic derived regularly. A great number of verbonominal derivatives can also be used as substantives and adjectives. In case of many verbal nouns, they are no longer used in their verbal meaning of action and they are capable of forming plural. As regards the participles, many of them have been substantivized and also form plurals or are used as adjectives. Despite the loss of the verbal character of both types of verbonominals, not only do they preserve some of their valency complements (esp. in case of participles), but those complements are usually expressed on the surface level by the same morphemic representation, as it is with the verb (by preposition or with accusative). However, only those cases where the verbonominals preserve their verbal character (action) are in the scope of our interest in this paper.

As concerns the verbal noun, it is capable of preserving all slots of the original verbal valency frame when Actor is annexed to it in the position of nomen rectum (*muḍāf ʾilay-hi*). It is worth pointing out that those verbal nouns derived from directly transitive verbs can emphasize their verbal character through affecting the accusative case in their object (see (33)) (Badawi et al., 2004: 238).

As regards the participle, namely the active participle, in cases where it substitutes for verbal predicate, it behaves the same way as the verb and keeps all the verbal valency slots (see (34)). Actor is realized as a surface subject or is absorbed by the subject of the governing sentence in such frequent cases where the active participle as a verbal attribute (*ḥāl*) develops the preceding sentence.

As an example of preserving the valency frame (Table 5), we can mention here verb *ṭālab* 'to demand' and its active participle *muṭālib* 'demanding' and verbal noun *muṭālabah* 'demanding, a demand'. This verb is double transitive; one object is direct, the other is indirect through the preposition.

(subject)	*bi-* (prep.)/*ʾan* (conj.)	4– (object)
ACT^{obl}	PAT^{obl}	ORIG^{opt}
someone	something/that	of someone

Table 5: Valency frame of the verb *ṭālab* طَالَب 'to demand'.

(32) *ṭālabat-i 'l-wikālatu 'l-ittiḥādāti bi-taḥlīli ...*
it-demanded ⌜the-agency$^{\textbf{ACT}}$ ⌜the-unions$^{\text{ORIG}}$ ⌜by-analysis-of ...$^{\textbf{PAT}}$
the agency demanded of the unions an analysis of ...

(33) *wa-muṭālabati 'l-wafdi 'l-ğamāʿata bi-ḥalli nafsi-hā*
and-demanding-of ⌜the-Wafd$^{\textbf{ACT}}$ ⌜the-group$^{\text{ORIG}}$ ⌜by-dissolution-of self-of-it$^{\textbf{PAT}}$
and al-Wafd's demand that the group dissolves itself

(34) *muṭāliban qiyādata 'l-ğayši bi-ʾiqāmati ḥiwārin*
[⌜he$^{\textbf{ACT}}$-is]-demanding ⌜leadership-of the-army$^{\text{ORIG}}$ ⌜by-establishing-of a-dialog$^{\textbf{PAT}}$
[he is] demanding of the military leadership to enter into a dialog

4. Methodology and Resources

The linguistic work on exploring the language data, extracting the valency information, and recording it in form of a

compact lexicon, needs to be supported with reusable resources, effective tools, and scalable procedures.

The key concept in our scenario is that of PDT-VALLEX (Hajič et al., 2003; Hajič and Urešová, 2003). The lexicon is gradually developed by linking its conceivable entries with their instances in the treebank. Conversely, the treebank's annotations are linked to the lexicon, so that the validity of annotations can be verified, and the mutual consistency of the treebank and the lexicon can be established. Due to Petr Pajas (p.c., 2008), the TrEd[2] annotation environment can be extended for the new linking and editing tasks, as the data of the valency lexicon can conveniently be formatted in terms of trees (note the valency frame treelets and the capturing of alternatives in the trees of Figure 2).

While a comparable functionality is readily in place for the Czech treebank projects, our own contribution is to design and implement this model thoroughly for Arabic and the PADT data. The new valency lexicon will be interlinked with the morphological lexicon of ElixirFM, which itself is linked with the morphological annotations.

ElixirFM[3] (Smrž, 2007) is an open-source implementation of inflectional and derivational morphological processes in Arabic allowing, among other things, to automatically compute the conversions between verbs, participles, and verbal nouns required for the study of valency. Entries of the ElixirFM lexicon are encoded using morphophonemic templates (shown in Figure 2). Most of the lexical information is adapted from the Buckwalter lexicon (Buckwalter, 2002). Converting the ElixirFM lexicon into a data structure accessible and editable with TrEd leaves open the possibility to expand the annotation of valency also to general nouns and adjectives (cf. Hajič et al., 2003). The work on the valency lexicon will however first concern the most frequent verbs and verbonominal derivatives, as discussed previously.

Using the PADT data for the extraction of Arabic valency information brings significant advantages. In the development version of the treebank (to be published as PADT 2.0 in 2008), over one million tokens are annotated into surface syntactic trees, and more than thirty thousand tokens are equipped with tectogrammatical features. The underlying morphological annotations provide disambiguated lexical and inflectional (functional and structural) morphological abstractions compatible with the ElixirFM system.

The contents of the arising valency lexicon will not be limited to evidence from PADT. The Arabic Gigaword (Graff, 2007) supplies even more raw data of the newswire domain, while the CLARA[4] corpus offers documents from literature and other types of texts. The printed dictionaries to be consulted include (e.g. Wehr, 1979; Baalbaki, 2000; Hoogland et al., 2003; Zemánek et al., 2006).

We believe that the valency frames gathered in VALLEX[5] (Žabokrtský, 2005; Lopatková et al., 2006, 2008) and PDT-VALLEX can serve as a useful source of inspiration for describing valency in other languages as well.

Based on our previous experience with exploring and managing large linguistic corpora, the search tools used in the project will include TrEd, Netgraph, and Xaira. TrEd and Netgraph[6] allow structural queries into the trees. Xaira[7] searches in linear text, but the underlying data for it can comprise node-related annotations of various kinds, in particular some automatically disambiguated morphological information (cf. Hajič et al., 2005).

The complex representation of the valency lexicon based on the principle of links will yet enable miscellaneous options for the reuse, processing, and exporting of the data. One of the possible linearizations of the lexicon is depicted in Figure 2. Other output formats can follow, for instance, the online, printed, or interactive interfaces of VALLEX.

5. Proposed Structure of the Lexicon

With respect to specific needs of Arabic, a slightly modified structure of a lexical entry used in VALLEX will be adopted (Lopatková et al., 2006). Entries will be organized alphabetically according to the Arabic three-/four-consonantal root, as is usual in most lexicographic works.[8] The proposed lexical entry will provide the following information:

Word Root an abstract consonantal morpheme that can include more than one lexical entry (lexeme);

Lexeme the most abstract notion represented by the lemma (citation form); one lexeme encompasses all inflectional morphological forms and discerns one or more lexical units corresponding to different meanings;

Morphological Information identification of the derivational class (I–X) and the corresponding verbal noun;

Lexical Unit each lexeme would contain at least one lexical unit, i.e. a particular meaning of the given verb with its syntactic properties described by the valency frame; a potential idiomatic usage of a verb with its restricted collocability will be taken into account—those cases will be treated as separate lexical units;

Valency Frame a sequence of frame slots—each of them is assigned to a different valency complementation of the given verb (lexical unit) bringing the valency information in terms of the FGD functors as well as the surface morphemic representation of such complementation (e.g. `4-` (accusative case), `-I` (indefinite state), the form of a particular bound lexeme, conjunction, or preposition, like *li-*, *ʾan*, etc.);

Obligatoriness/Optionality valency frames will consist of both obligatory and optional inner participants (actants) and only of obligatory free modifications;

Diathesis possible passivisation of the given verb by changing the internal vowel pattern, in case it is semantically permitted, should be mentioned and illustrated with an example taken from available corpora;

Additional information examples from the corpus as well as the exact frequency of occurrence and rank of a lexeme in PADT will be provided; syntactic-semantic verb class will be the subject of our future research.[9]

[2] `http://ufal.mff.cuni.cz/~pajas/tred/`

[3] `http://sf.net/projects/elixir-fm/`

[4] `http://enlil.ff.cuni.cz/`

[5] `http://ufal.mff.cuni.cz/vallex/2.5/`

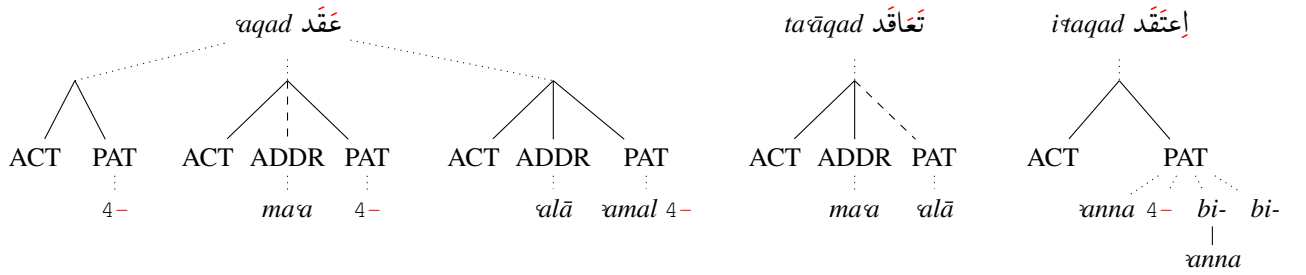
[6] `http://quest.ms.mff.cuni.cz/netgraph/`

[7] `http://sf.net/projects/xaira/`

[8] Weak verbs will be treated as triconsonantal with *w* or *y* included among the root consonants.

[9] In this respect, works of (McCarus, 1976) and (Al-Qahtani, 2004) as well as verb classes proposed in VALLEX are of high inspiration for us.

```
|> "ʿ q d" <| [                                                                    ع ق د
    FaCaL     'verb'  [ "hold", "convene", "conclude" ]  'imperf'  FCiL,   ʿaqad   I (i)   عَقَد
    FaCCaL    'verb'  [ "complicate" ],                                    ʿaqqad   II      عَقَّد
    TaFaCCaL  'verb'  [ "be complicated" ],                                taʿaqqad V       تَعَقَّد
    TaFACaL   'verb'  [ "contract", "convene" ],                           taʿāqad  VI      تَعَاقَد
    InFaCaL   'verb'  [ "be held", "be gathered", "be convened" ],         inʿaqad  VII     اِنعَقَد
    IFtaCaL   'verb'  [ "believe" ]                                    ]   iʾtaqad  VIII    اِعتَقَد
```



I *ʿaqad yaʿqid ʿaqd* عَقَد ـ يَعقِد ـ عَقد ~ 526

ACT PAT (4–) to call, to hold (a meeting, a conference, etc.) |^{act} *ʿaqadat-i 'l-laǧnatu 'l-wizārīyatu 's-sūrīyatu 'l-kuwaytīyatu 'ǧtimāʿa-hā 'd-dawrīya fī Dimašqa* the committee of Syrian and Kuwaiti ministers held its regular meeting in Damascus |^{pas} *wa-lam yuʿqad ʾayyu muʾtamarin ṣiḥāfīyin muštarakin* and no joint press conference was held

ACT ADDR (*maʿa*) **PAT** (4–) to conclude (a contract, a treaty, etc.) |^{act} *al-muṭaqqafu 'llaḏī yaʿqidu ṣafqatan maʿa 'l-ḥukūmati* the intelectual that is concluding a bargain with the government |^{pas} *ʿuqidat-i 'ttifāqīyatu 'l-Ǧazāʾiri ʿāma 1975 bayna ʾIrāna wa-'l-ʾIrāqi* the Algerian agreement between Iran and Iraq was concluded in 1975

ACT ADDR (*ʿalā*) **PAT** (*ʾamal* 4–) to set one's hope(s) to (idiom) |^{act} *yaʿqidu 'n-nādī 'l-ʾisbānīyu ʿāmālan ʿalā ʾan tusāʿida šaʿbīyatu Bīkhām al-ǧārifatu fī 'ḫtirāqi 's-sūqi* the Spanish club is setting its hopes to that Beckham's great popularity will help to penetrate into the market |^{pas} *allaḏīna yuʿqadu ʿalay-him al-ʾamalu fī taʿlīmi 'ṣ-ṣiġāri wa-'l-kibāri li-luġati-nā 'l-ḥayyati* those to whom the hope is set that they will teach both the young and the old our live language

II *ʿaqqad* عَقَّد ~ 9

ACT PAT to complicate |^{act} *ḏālika ʾamrun yuʿaqqidu 'l-muškilata wa-lā yaḥullu-hā* this is a thing that only complicates the problem and does not solve it

V *taʿaqqad* تَعَقَّد ~ 3

ACT to be/become complicated |^{act} *al-waḍu taʿaqqada lāḥiqan bi-mā ḥadaṯa li-ʾAlmāniyā min taqsīmin* the situation became complicated afterwards as a consequence of the division of Germany

VI *taʿāqad* تَعَاقَد ~ 20

ACT ADDR (*maʿa*) PAT (*ʿalā*) to come to an agreement, to make a contract |^{act} *kamā taʿāqadat [aš-šarikatu] maʿa 'l-banki 'l-ʾislāmīyi bi-Dubay ʿalā ʾidārati funduqi ʾAbū Ẓabī 't-tābiʿi la-hu* it [the company] also came to the agreement with the Islamic Bank in Dubai on managing its Abu Dhabi Hotel

VII *inʿaqad* اِنعَقَد ~ 42

ACT to assemble, to meet, to be held, to take place (a meeting, a conference, etc.) |^{act} *wa-fī tišrīna 't-ṯānī 1995, inʿaqada fī Baršilūnah al-muʾtamaru 'l-wizārīyu li-'š-širākati 'l-ʾūrūbīyati 'l-mutawassiṭīyati* in November 1995, the Ministerial Conference of the Euro-Mediterranean Partnership was held in Barcelona

VIII *iʾtaqad* اِعتَقَد ~ 285

ACT PAT (4–, *bi-*, *ʾanna*, *bi-ʾanna*) to believe, to think |^{act} *kamā ʾanna-nī ʾaʿtaqidu ʾanna 'r-raǧula 'š-šarqīya lā yuḥibbu ʾan yakūna murattabu zawǧati-hi ʾaʿlā min murattabi-hi* I also think that the man from the Orient does not like when his wife's salary is higher than his own | *fa-'l-ʾIrāqu laysa milkan li-qūwāti 't-taḥālufi, wa-ʾin iʾtaqada baʿḍu 'l-muḥāfiẓīna 'l-ǧududi ḏālika* Iraq is not property of the ally forces, even if some neo-conservatives have that opinion |^{pas} *baʿda našri ḫarīṭati 'ṭ-ṭarīqi 'llatī yuʾtaqadu ʾanna-hā sa-takūnu ḥayawīyatan fī ḥalli 'l-qaḍīyati 'l-filasṭīnīyati* after the publication of the Road Map, which is believed to be vital for the solution of the Palestinian Question

Figure 2: *Top left:* Verb entries of the ElixirFM lexicon nested under the ʿ*q d* عقد root. *Top right:* Possible layout of these entries including the explicit derivational class, showing that various pieces of information can be inferred directly from this lexicon's representation. *Middle:* Valency frame treelets and the constraints on the surface realization of the functors, organized into trees. Optional slots are marked with dashed edges. Multiple options with frames or constraints are rendered as dotted links. *Bottom:* Linearized valency information augmented with the derivational class number, frequency, glosses, and corpus-based examples. Obligatory functors are highlited in bold.

6. Conclusion

In our contribution, we have overviewed the theoretical concept of valency developed in FGD and have adapted its essential part for Modern Standard Arabic. We have illustrated various valency-related phenomena on a number of instances, extracted mainly from the PADT treebank.

We have presented the methodology, tools, and resources used for creating the valency lexicon of Arabic verbs, which is intended for applications in computational parsing or language generation, and for use by human researchers.

The proposed valency lexicon will be exploited in particular during further tectogrammatical annotations of PADT, and might possibly serve for enriching the expected second edition of the corpus-based Arabic-Czech Dictionary (Zemánek et al., 2006).

Acknowledgements

We would like to thank Petr Zemánek, Markéta Lopatková, Petr Pajas, and Zdeněk Žabokrtský for their valuable comments and inspiring ideas. This work has been funded by the Ministry of Education of the Czech Republic (MSM00-21620823 and MSM0021620838) and by the Grant Agency of the Czech Academy of Sciences (1ET101120413).

References

Amira Agameya. Passive (syntax). In Kees Versteegh, editor, *Encyclopaedia of Arabic Language*, volume 3, pages 558–564. Brill, Leiden – Boston, 2008.

Duleim Masoud Al-Qahtani. *Semantic Valence of Arabic Verbs*. Libraire du Liban, Beirut, 2004.

Rohi Baalbaki. *Al-Mawrid: A Modern Arabic-English Dictionary*. Dar El-Ilm Lilmalayin, Beirut, 2000.

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

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

Walter A. Cook. *Case Grammar: Development of the Matrix Model (1970–1978)*. Georgetown University Press, 1979.

Ladislav Drozdík. Grammatical, Derivational and Lexical Dimensions of Transitivity in Arabic. In *Modern Written Arabic: Studies in Grammar, Lexicon and Prestigious Oral Communication*, pages 87–103. Veda, Bratislava, 2001.

David Graff. Arabic Gigaword Third Edition. LDC2007T40, 1-58563-460-3, 2007.

Jan Hajič and Zdeňka Urešová. Linguistic Annotation: from Links to Cross-Layer Lexicons. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, pages 69–80, Växjö, Sweden, 2003.

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

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

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

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.

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

Jan Hoogland, Kees Versteegh, and Manfred Woidich. *Woordenboek Arabisch-Nederlands*. Bulaaq, Amsterdam, 2003.

Markéta Lopatková. Valency in the Prague Dependency Treebank: Building the Valency Lexicon. *Prague Bulletin of Mathematical Linguistics*, 78–80:37–60, 2003.

Markéta Lopatková and Jarmila Panevová. Recent developments of the theory of valency in the light of the Prague Dependency Treebank. In Mária Šimková, editor, *Insight into Slovak and Czech Corpus Linguistic*, pages 83–92. Veda, Bratislava, 2005.

Markéta Lopatková, Zdeněk Žabokrtský, and Václava Benešová. Valency Lexicon of Czech Verbs VALLEX 2.0. Technical Report 34, UFAL MFF UK, Charles University in Prague, 2006.

Markéta Lopatková, Zdeněk Žabokrtský, and Václava Kettnerová. *Valenční slovník českých sloves [Valency Lexicon of Czech Verbs]*. Karolinum, Praha, 2008.

Ernest N. McCarus. A Semantic Analysis of Arabic Verbs. In *Michigan Oriental Studies in Honour of George G. Cameron*, pages 3–28. Department of Near Eastern Studies, University of Michigan, Ann Arbor, 1976.

Marie Mikulová et al. A Manual for Tectogrammatical Layer Annotation of the Prague Dependency Treebank. Technical Report 30, UFAL MFF UK, Charles University in Prague, 2006.

Jarmila Panevová. On Verbal Frames in Functional Generative Description, Part I. *Prague Bulletin of Mathematical Linguistics*, 22:3–40, 1974.

Jarmila Panevová. On Verbal Frames in Functional Generative Description, Part II. *Prague Bulletin of Mathematical Linguistics*, 23:17–52, 1975.

Jarmila Panevová. Valency Frames and the Meaning of the Sentence. In *The Prague School of Functional and Structural Linguistics*, pages 223–243. John Benjamins, Amsterdam–Philadelphia, 1994.

Jan Retsö. Diathesis. In Kees Versteegh, editor, *Encyclopaedia of Arabic Language*, volume 1, pages 622–626. Brill, Leiden – Boston, 2006.

Karin C. Ryding. *A Reference Grammar of Modern Standard Arabic*. Cambridge University Press, 2005.

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

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

Hans Wehr. *Arabic-English Dictionary: The Hans Wehr Dictionary of Modern Written Arabic*. Spoken Language Services, 1979. Edited by J. Milton Cowan.

William Wright. *A Grammar of the Arabic Language*. Cambridge University Press, 1991.

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

Petr Zemánek, Andrea Moustafa, Naděžda Obadalová, and František Ondráš. *Arabsko-český slovník [Arabic-Czech Dictionary]*. Set Out, Praha, 2006.