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 representation 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):

- **ACTor** – usually the agent (the surface subject) 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 (Paneková, 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.

Table 3: Members included in the valency frames.

<table>
<thead>
<tr>
<th>Inner participants (actants)</th>
<th>Obligatory</th>
<th>Optional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free modifications (adjuncts)</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>

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 (Paneková, 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 (Paneková, 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 Paneková, 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) q¯ała Zaydun maq¯ulatu-hu he-said Zayd said-of-him Zayd said what he had to say q¯al AEO/E-del (Experiencer is deleted)
(Y) darasa Zaydun al-kit¯aba he-studied Zayd the-book Zayd studied the book daras AEO/A=E (Agent equals Experiencer)
(Z) amila Zaydun he-worked Zayd Zayd worked = Zayd did some work amil AO/O-lex (Object is lexicalized)

(subject)  | li- (prep.)  | un (prep.)  | 4-/1inma (conj.)
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>ADDR</td>
<td>PAT</td>
<td>EFF</td>
</tr>
<tr>
<td>someone</td>
<td>to someone</td>
<td>about sth.</td>
<td>something/that</td>
</tr>
</tbody>
</table>

Table 4: Valency frame of the verb q¯al َعَقَلَ “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 transitive reflexive (2).

(1) kana yanama adatan fi sarvin saqirin
he was himACT-sleeps usuallyTHO in a-street a-smallLOC
he usually slept in a small street

(2) intahara bi-ta’swibihi ’l-muwasatta ila rasih
he committed a suicide himACT-by-aiming-of-him the-gut at head-of-himMEANS

If the valency frame includes two actants, the first actant is considered to be Actor and the other Patient. Some verbs go 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) tahawwalat-i ’l-munazzamatu min udatin muwaddati ila udatin li-’l-bahtsi
it-changed the-organizationACT from-instrument of a-confrontationORG to an-instrument for-the-researchEFF (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) samaha la-hu bi-’d-dahbati ila ’l-bayti
heACT-permitted to-himADDR by-the-entering into the-housePAT
he permitted him to enter the house

(7) sharakat zawqa-ha fr ’l-hukmi
sheACT-shared husband-of-herADDR in-the-reignPAT
she shared the reign with her husband

It is worth mentioning that the usual word order of double transitive verbs as urch 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-

(8) urch-hu ’l-fursata
heACT-gave-himADDR the-opportunityPAT
he gave him the opportunity

(9) urch-i ’s-sayyaratara li-’l-bunaki
itACT-gave the-power PAT to-the-banks ADDR
it gave the power to the banks

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

(10) wyyana-hu hikiman li-’l-Kuwayti
heACT-appointed-himPAT a-ruler for-KuwaitEFF
he appointed him as a ruler of Kuwait

(11) intahara Addinis urufiyam
he-consideredACT Adonis PAT a-EuropeanEFF he considered Adonis to be European

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

(12) tagarma uk’ara min hamsina kitabun min-a ’l-farsi-yati ila ’l-arabi-yati
he-consideredACT translated a-more than fifty a-bookPAT from PersianORG into ArabicEFF
he translated more than fifty books from Persian into Arabic

(13) gayyarat [aš-šarikatu] nashta-hu min t mantiqi ’l-qamhi ila t mantiqi ’l-budari
itACT-changed [the companies] activity-of-itPAT from production-of the-wheatORG to production-of the-seedsEFF [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) baddu ’l-harbi wadada-hu umama umrini waq’in
beginning-of-the-war ACT it-put-himPAT in-front-of a-thing a-realDIR3
the beginning of the war put him in front of the reality

(15) ’adat min-a ’l-Qahiratila Bayriita
sheACT-returned from CairoDIR1 to BeirutDIR3 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) gadarat-i ’l-Qahiratila Tall Abib
1. Some intransitive verbs, esp. those of movement, become transitive secondarily when taking a preposition, e.g. ǧā’ ‘to come’ → ǧā- bi- ‘to bring sth.’; qām ‘to stand up’ → qām bi- ‘to carry out sth.’ (Badawi et al., 2004: 382–383; Drozdik, 2001).

17. wašala ‘l-muntaḥabu ūla madinati Sālinrū ‘l-ţāliyyat it-arrived ‘the-representation to-town-of Salerno the-Italian

the representation arrived to the Italian town Salerno

18. wasaltu-hā [Dimašiq] min-a ‘d-Dawhaṭi

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. wašalat qimatu-hā ūla ḥamsati yārīḥatān

it-reached ‘value-of-it to five euros its value reached 5 euros

20. wašalat qimatu ‘s-sādiratū 625 milyūna dålarin

it-reached ‘value-of-the-exports to 625 million-of a-dollar 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 verbal 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 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-لام waqad wa-yuu mu’tamarin siḥāfīyin muṣṭarakin

and-not it-was-held ‘any a-conference a-press a-

joint 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. ‘uṣūna ‘d-dukťārū Mawsili wakilān li-kulliyatī ‘l-ţībī

he-was-appointed ‘the-doctor Mawsili PAT ‘an- assistant-dean to-faculty-of-the-medicine 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 ʿurtā ‘to give’) might be passivised 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. ʿuṭyāt fursatān tāniyatān

she ADDR-was-given ‘a-chance a-second she was given the second chance

24. ʿuṭiya ‘d-ḍawu ‘l-ʿabdū hu fī ‘l-šalākiṭī

it-was-given ‘the-light the-green ADDR ‘to-the- queen the queen was given the green light

In case of indirectly transitive verbs through prepositions when passivised, 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. ʿuḥkamu alay-hi bi-ṣ-sīgīnī

it-is-sentenced ‘upon-him PAT ‘by-the-jail 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. ʿašrıkāttūn tudāru min qibalā muddarī a mutahāssisīna

‘companies PAT is-managed ‘from-directions-of managers specialist 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) tuwuffiya V ‘to die, to pass away’ and (28) ustūšid X ‘to die as a martyr’ reflecting some degree of euphemism in connection with religious being. Those verbs would be treated in our proposed lexicon as separate word entries.

27. tuwuffiya wǎlīdu-hu fī ḥādīti sāyyārātīn

he-died ‘father-of-him ACT ‘in accident-of a-car MANN

his father died in a car accident

28. ustūšidā ðumā 1991 fī ḥādīti ʿgītīyālin

‘he ACT-died-as-a-martyr ‘year-of 1991 WHEN ‘in
event-of an-assassination

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 dāt ‘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) or by an anaphoric means using the noun bād ‘part, some’ (Badawi et al., 2004: 391–394).

(29) taqādat mā yl-banki l-islāmiyti bi-Dubay ulā idārati funduqī Abū Zabi

‘it-CAME-TO-THE-AGREEMENT with-the-bank the-islamic in-Dubay on-managing-OF hotel-of Abu Dhabi

it came to the agreement with the Islamic Bank in Dubai 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 (waqab I, etc.) and possibility (umkân 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 un (31) as a subject surface 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 ulā) or as a direct object (when pronoun) or after preposition li- (when a noun) in case of umkân IV.

(30) yunjina-hu d-dahāla fi l-barbi

it-enables-him PAT the-entering in-the-war ACT

it is possible for him to enter the war

(31) yaqību ulā gūwāti l-tahlūfī un tuḥṭa bi-hādā l-baladi

‘it-IS-NEEDED on-forces-of the-alliance PAT that they-encircle by-this the-country ACT

the ally forces to encircle that country

According to the FGD approach, auxiliary verbs do not have their own valence 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 autonomous words, i.e. including nouns and adjectives.

In case of Arabic, studying the verbal valency should be very fruitful with regard to its verbonominal derivatives—a participle (active and passive) and a verbal noun (masdar).

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 valence 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 (mudaf ‘slay-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 tālāb ‘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.

<table>
<thead>
<tr>
<th>(subject)</th>
<th>bi- (prep.)/un (conj.)</th>
<th>4 – (object)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT&lt;sup&gt;OPT&lt;/sup&gt;</td>
<td>PAT&lt;sup&gt;OPT&lt;/sup&gt;</td>
<td>ORIG&lt;sup&gt;OPT&lt;/sup&gt;</td>
</tr>
<tr>
<td>someone</td>
<td>something/that</td>
<td>of someone</td>
</tr>
</tbody>
</table>

Table 5: Valency frame of the verb tālāb طلاب ‘to demand’.

(32) tālābat-i ‘l-wikālatu ‘l-ittihādāti bi-taḥlīlī ... it-demanded [the-agency<sup>ACT</sup> of-the-unions<sup>ORIG</sup> of-by] analysis-of ... PAT

the agency demanded of the unions an analysis of...

(33) wa-muṭālabat-i ‘l-wafī ‘l-ǧamā’ata bi-hallī nafaṣi-hā and-demanding-of the-Wafī<sup>ACT</sup> the-group<sup>ORIG</sup> by-dissolution-of self-of it-PAT

and the-Wafī’s demand that the group dissolves itself

(34) muṭālaban qiyādatu ‘l-ṭayyib bi-taqāmati hiwārin [he<sup>ACT</sup>-is]-demanding ‘leadership-of the-army<sup>ORIG</sup> of-by-establishing-of a-dialog<sup>PAT</sup> [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 Pajás (p.c., 2008), the TrEd2 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. ElixirFM3 (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 verbomonominal 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 CLARA4 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 VALLEX5 (Zabokrtský, 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 Netgraph6 allow structural queries into the trees. Xaira7 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 discurs 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), –1 (indefinite state), the form of a particular bound lexeme, conjunction, or preposition, like li-, un, etc.);
- **Obligatoriness/Optionality** valency frames will consist of both obligatory and optional inner participants (arguments) 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

---

2http://ufal.mff.cuni.cz/~pajas/tred/
3http://sf.net/projects/elixir-fm/
4http://enlil.ff.cuni.cz/
5http://ufal.mff.cuni.cz/vallex/2.5/
6http://quest.ms.mff.cuni.cz/netgraph/
7http://sf.net/projects/xaira/
8Weak verbs will be treated as triconsonantal with w or y included among the root consonants.
9In this respect, works of (McCarus, 1976) and (Al-Qahtani, 2004) as well as verb classes proposed in VALLEX are of high inspiration for us.
FaCAL 'verb' [ "hold", "convene", "conclude" ] 'imperf' FCIL, vqad I (i)
FaCCal 'verb' [ "complicate" ], tvaqad II
TaFaCal 'verb' [ "be complicated" ], tvaqad V
TaFACal 'verb' [ "contract", "convene" ], tvqad VI
InfACal 'verb' [ "be held", "be gathered", "be convened" ], ivqad VII
IftaCAL 'verb' [ "believe" ] itqad VIII

Figure 2: Top left: Verb entries of the ElixirFM lexicon nested under the 'q d` 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 highlighted 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 (MSM0021620823 and MSM0021620838) and by the Grant Agency of the Czech Academy of Sciences (1ET10120413).

References


