Enhancing the ElixirFM Lexicon with Verbal Valency Frames

Viktor Bielický
Institute of Near Eastern and African Studies
Charles University in Prague
viktor.bielicky@ff.cuni.cz

Otakar Smrž
Institute of Formal and Applied Linguistics
Charles University in Prague
otakar.smr@ff.cuni.cz

Abstract
This paper reports on the status of our ongoing project to extend, refine, and maintain the lexicon of the ElixirFM system. The latest release of ElixirFM has been enriched with valency frames for selected verbal lexemes. We briefly outline some other similar projects for Modern Standard Arabic and present crucial axioms of the approach we have adopted, which is Functional Generative Description. Our contribution focuses mainly on describing the annotation procedure, resources and tools, as well as on the representation of the valency phenomena in form of dependency trees in the lexicon. We also point at some intricate issues concerning reciprocal meaning associated with some verbs. Such a valency lexicon for Arabic can find broad use in various NLP tasks and applications, in addition to its reusability by the users of the language.

1. Introduction
Valency of a lexical unit, in particular a verb, represents its potential to bind specific numbers and types of other lexical units. Valency properties are lexicalized and cannot be inferred using grammatical rules only. Creating a valency lexicon of the most frequent verbs for Modern Standard Arabic (MSA) that will be reusable by both humans and computational systems is a desirable task. This kind of lexicon can find use in a wide variety of NLP problems including word-sense disambiguation, tagging, syntactic and semantic analysis, machine translation, and language generation.

The aim of this paper is to report on the status of our ongoing project to extend, refine, and maintain the lexicon of the ElixirFM system, which has been recently enhanced with valency frames for some chosen verbal lexemes. This is the result of the initial phase of our effort to build a verbal valency lexicon for MSA, the basic concept of which was presented in (Bielický and Smrž, 2008).

First, we describe the ElixirFM Lexicon and its web application ElixirFM Online Interface (Sec. 2), and we briefly outline some projects and resources closely related to ours (Sec. 3). Then we touch upon some crucial axioms of the theoretical framework we adopted (Sec. 4), discuss several topics concerning verbal valency frames and their annotation procedure, and examine closely reciprocal meaning associated with some verbs (Sec. 5).

2. ElixirFM
ElixirFM (Smrž, 2007a, b) is a high-level open-source implementation of Functional Arabic Morphology reusable for various NLP applications. Most of its lexical material was originally derived from the open-source Buckwalter lexicon (Buckwalter, 2002), however the latest release of ElixirFM (Smrž and Bielický, 2009) has undergone considerable computational and lexicographical corrections, refinements, and extensions.

The online implementation ElixirFM Online Interface (see Fig. 1 at the end) can serve as an example of the lexicon’s computational potential.1 This user interface works in four different modes: 1) Resolve mode provides tokenization and morphological analysis of the inserted text (optionally in Unicode, Buckwalter or ArabTeX notations); 2) Inflect mode offers complete or partial paradigms of inflected forms for a given lexeme; 3) Derive mode gives verbal nominal derivatives (verb, verbal noun, active and passive participles) for the particular class of lexemes; 4) Lookup mode enables to search for a desired lexeme and its possibly existing root-related derivatives, i.e. it serves as an online dictionary. It is even possible to browse the dictionary in a reverse direction by inserting expressions in English.

3. Similar Projects for Arabic
To the best of our knowledge, the only existing projects that are related to a systematic description of valency phenomena or semantic-syntactic properties of a considerable number of Arabic verbs are the Arabic Propbank (Palmer et al., 2008), the work by Loukil et al. (2008), and the printed study by al-Qahtani (2005). Let us shortly review their main characteristic features.

3.1. Arabic Propbank
The Pilot Arabic Propbank project, which is based on corpus data derived from the Penn Arabic Treebank and which is similar to the English and Chinese Propbanks, aims at adding a level of semantic annotation into a large Arabic corpus.

First, framesets of particular verbs are created by which individual senses for the given verbal lexeme are distinguished. A frameset contains all semantically required arguments for each sense of the verb (lexical unit). The arguments are labeled as ARG0-4 and are assigned specific semantic roles (experiencer, organizer, event, agent, theme, etc.). Consequently, the ARG0-4 labels must be seen as variables containing different semantic roles according to and depending on every single verb sense they are assigned to.

Second, the framesets are used as guidelines by annotators who match them to corpus instances of verbal predicates and their arguments. In addition to the above-mentioned 5 numbered arguments, 19 fixed adjunctive arguments are distinguished (time, location, direction, manner, etc.).

1 http://quest.ms.mff.cuni.cz/elixir/
3.2. Arabic Syntactic Lexicon

As concerns the creation of a syntactic lexicon of Arabic verbs, the project announced by Loukil et al. (2008) focuses on building a machine-readable resource distinguishing syntactic (though surface) properties for each sense of a verbal lexeme, which are encoded into syntactic frames using the extensional Lexical Markup Framework formalism. Such a syntactic frame brings information on transitivity/intransitivity of a verb and a number and type of its syntactic complementations (its realization through the accusative case or after a preposition – if it is NP/PP, identification of its syntactic function (subject, object, etc.)). Each sense is illustrated with an example sentence. Unfortunately, the authors’ standpoint seems to be affected by the traditional Arabic approach to syntactic structures, where classifying VSO and SVO word orders as different types of sentences approach to syntactic structures, where classifying VSO and SVO word orders as different types of sentences.

3.3. Semantic Valence of Arabic Verbs

The printed study on semantic valence of Arabic verbs by al-Qahtani (2005), which has been briefly compared to our approach in (Bielicíký and Smrž, 2008), presents thorough examination of Arabic verbs in terms of the Case Grammar theory based on Cook’s Matrix Model of verb classification (1979 and later improvements). It distinguishes three semantic types of verbs, i.e. state, process and action verbs, each of which is further subclassified into basic, experiential, benefactive and locative verbs according to a particular case role (semantic complementation) which it is characteristic for. These case roles are Agent, Experiencer, Benefactive, Object, Locative/Time. The obligatory Object is omnipresent with every verb and can occur more than once in a case frame. On the other hand, Experiencer, Benefactive and Locative are mutually exclusive. With some verbs, a particular case role is not realized on the surface (“covert case role”), i.e. it is either partially covert (“deletable”) or totally covert (“coreferential” or “lexicalized”). The deletable case roles are simply those complementations which can be omitted on the surface as a result of the ellipsis. On the contrary, the coreferential and lexicalized case roles are always absent from the surface. The former coreferential roles denote instances where a single noun on the surface structure cumulates two case roles simultaneously, while the latter lexicalized roles include instances where a certain case role (usually Object) is incorporated in the semantics of a given verb. Let us look at some al-Qahtani’s examples:

(1) deleted case role from the surface

\[ \text{wašafa Zaydun al-manzar} \]
Zayd described the scene

\[ \text{wašaf ÆOE/E-del} \] (Experiencer is deleted)

(2) coreferential case roles

\[ \text{gādāra Zaydun al-bayt} \]
Zayd left the house

\[ \text{gādar AOL/A=O} \] (Agent equals Object)

(3) lexicalized case role

\[ \text{āmilə Zaydun} \]
Zayd Worked = Zayd did some work

\[ \text{āmil AO/O-lex} \] (Object is lexicalized)

To conclude, all the aforementioned approaches suggest some interesting ideas which are worth noticing and can serve us as a source of inspiration.

4. Theoretical Framework

The theoretical framework that we adopted is the valency theory elaborated within the Functional Generative Description (FGD), a multi-stratal dependency-oriented language representation formalism (Panevová, 1974, 1975, 1994; Sgall et al., 1986). This approach has been successfully applied on two valency lexicons for Czech, namely VALLEX (Lopatková et al., 2006, 2008; Žabokrtský, 2005) and PDT-Vallex (Haji et al., 2003).

Although the study of valency phenomena pertains mainly to the highest (underlying) tectogrammatical layer representing the linguistic meaning, it is also closely associated with the surface morphemic layer where deep-level entities emerge and are realized as morphemes. According to FGD, it is possible to describe the semantic-syntactic properties of a verb or any other autosemantic lexical unit by its valency frame – a sequence of frame slots – where constraints on the specific type, number and possible morphemic realizations of the dependents of the lexical unit are encoded.

The valency frame of a verb is filled with a particular set of functors denoting a type of relation between the governor (a verb) and its dependent (verbal complementation/argument) as units of the tectogrammatical layer of language representation. Two kinds of verbal complementations are distinguished, i.e. actants (or inner participants – ACTor, PATient, ADDRessee, ORGin, EFFECT) and free (adverbial) modifications (or adjuncts, e.g. adverbials of direction, location, means, time, result, respect). Both actants and free modifications are considered to be either obligatory or optional for each lexical unit (each sense of a verbal lexeme). The so-called “dialogue test” is introduced as the criterion of obligatoriness for both types of verbal complementations (e.g. Panevová, 1974, 1975). The valency frame in its narrow sense consists of both obligatory and optional actants and only obligatory free modifications (see Table 1).

<table>
<thead>
<tr>
<th>obligatory</th>
<th>optional</th>
</tr>
</thead>
<tbody>
<tr>
<td>actants (inner participants)</td>
<td>+</td>
</tr>
<tr>
<td>free modifications (adjuncts)</td>
<td>+</td>
</tr>
</tbody>
</table>

Table 1: Members included in the valency frame

Thus, for instance, the valency frame of the basic meaning of the quadrilateral verb tarğım ("to translate") consists of two obligatory (ACT, PAT) and two optional (ORIG, EFF) actants (Fig. 2). If we intend to transform this sentence into the form of a dependency tree structure, its possible representation can be depicted as is shown in the following (Fig. 3).

\[ \text{84} \]

For a full list of verbal complementations and their description, detailed annotation guidelines for the tectogrammatical layer of the Prague Dependency Treebank for Czech (Mikulová et al., 2006) should be consulted.
5. Valency frames in ElixirFM

5.1. Present coverage, tools and resources

The latest release of ElixirFM (Smrž and Bielický, 2009) has been enhanced with nearly 3,500 valency frames for selected verbal lexemes. Roughly 2,000 frames were created automatically exploiting the information on intransitivity from the Buckwalter Arabic Morphological Analyzer (Buckwalter, 2002). In addition to this, around 1,500 non-trivial valency frames have been created manually. All these frames are being subject to further additions, corrections, and refinements. In this initial phase of the annotation process, mostly basic meanings of verbal lexemes have been covered. In the following phases, we will focus more on figural meanings and on phraseology.

The ElixirFM lexicon is maintained and annotated in the environment of the open-source tree editor TrEd. The main resource for extracting the valency frames is the Prague Arabic Dependency Treebank (PADT) (Hajič et al., 2004a; Smrž et al., 2008) the second release of which containing a million word tokens is in preparation. This linguistically annotated corpus of Arabic newswire texts consists of three layers of annotations, i.e. functional morphology, surface analytical syntax, and tectogrammatics. The ElixirFM lexicon is already used in the MorphoTrees annotations of PADT, and the valency frames are being linked to the nodes in the tectogrammatical as well as analytical syntactic structures. PADT also serves us as a source for the study of frequency of verbs. Besides PADT, CLARA (Zemánek, 2001) and Arabic Gigaword are also exploited. We further consult major contemporary printed dictionaries (Baalbaki, 2000; Hoogland et al. 2003; Wehr, 1979).

5.2. Dependency-tree representation

Each annotated verbal lexeme, which is nested under its triliteral or quadriliteral root, is represented by a citation form, a lemma. The lexeme associates all its inflected forms and may include links to its deverbal nouns. Apart from these formal properties, a verbal lexeme comprises a set of lexical units, i.e. its individual senses (at least one), the semantic-syntactic properties of which are defined by means of their valency frame.

The valency frame, which is represented by a tree structure of alternative dependency subtrees, not only encodes the valency properties of a verb in terms of the FGD functors, but in most cases it also provides constraints on the possible surface morphemic realizations of a particular verbal complementation.

All valency frames in the ElixirFM lexicon include both obligatory and optional actants and only obligatory free modifications (see Table 1 above), but in some cases even optional free modifications are added, since they can be regarded as typical complementations.

To demonstrate the representation of all these phenomena in ElixirFM, the previously quoted sentence with verb ṭargam (Fig. 2 and 3) can serve as a suitable example (see Fig. 4 at the end).

---

3 http://ufal.mff.cuni.cz/~pajas/tred/
5.3. Morphemic constraints
As is the case in VALLEX (Zabokrtský, 2005; pp. 63-66), two types of morphemic constraints can be distinguished and modified for our Arabic valency lexicon as well. These are:
Explicit morphemic constraints – one or more possible surface realizations for the given valency slot (alternantive subtree) is explicitly declared in the frame.
- Inflectional case ending (section of the morphological tag): 1- (nominative), 2- (genitive), 4- (accusative);
- Indefiniteness/definiteness (section of the morphological tag): 0/- D;
- Constructions with adjectival (section of the morphological tag): A-;
  - ii) Prepositionals (yet incomplete list): bayna, dānā, ḥawa, tabhā;
- Subordinate content clauses with a conjunction (yet incomplete list): ‘inna, ‘anna, ‘an, mā ‘idā;
- Dependent part of a phrase: If the lexical choice of a particular verbal complement is very limited due to a very restricted collocability of such a lexical unit, the lexical choice should be listed in the lexicon, such as for example (6). 5

(6) inšaraḥa la-hu ṣadr-ī
it-was-delighted because-of-him.CAUS chest-of-me.ACT
I was delighted by him valency frame: ACT60 (sadrqalbhiṭṭir) CAUS95 (li-)

Implicit morphemic constraints – the surface realization of the valency slot results from the semantics of the particular functor (usually a free modification), and thus it does not have to be specified in the valency frame, since it is registered in the list of possible forms (yet incomplete) gathered during the annotation process (e.g. LOC (locative) = fi, bi-, ḥawa, bayna, ‘alā, tabhā, favqa, ‘amūna, warā`, ḥulfa, ḥāliha, ḥaṛgīza . . . ; DIR1 (direction from) = min, min ‘alā, min tabhī, min warā`, min favqa . . . ; CAUS (cause) = li-, bi-sababī, bi-faḍlī, min ‘aġlī, nātīgṭaṭan li-, bi-ḥukmi, li-’anna . . . ).

5.4. Reciprocity
When we try to capture the semantic-syntactic properties of a verb in its valency frame, we might face some difficulties, if the inherent reciprocal meaning of form VI verbs (and even of some form VIII verbs) is involved. This reciprocal character originates from the composition of the associative meaning of the form III verb with the anaphoric prefix tu- (LeTourneau, 1998).
In order to emphasize the reciprocal meaning of such verbs, we propose that a special attribute “Rcp” can be attached to the valency frame where particular complementations, which enter the relation of reciprocity, will be listed. In this case, we follow the practice in Vallex (Lopatková et al. 2006, 2008).
For example, if the surface subject of the verb tabāḥaṭ VI (“to dicuss, to confer with each other”) is in plural or dual (or as two or more coordinated elements), in fact, two different actants (ACT, ADDR) are mixed up into a single surface realization (7). On the other hand, the very same verbal complementations can be realized as two surface entities, if the preposition mēa is used. Thus, the valency frame created by us for this verb contains ACT60 ADDR60 (mēa) PAT95 (filḥawa) and will receive the attribute Rcp: ACT-ADDR depicting that ACT and ADDR are in reciprocal relation (see (8) and Fig. 5 at the end).

(7) tabāḥaṭa ʾz-aʿīmīnī fi ʾaḥīrī l-mustuqaddāṭī
he-discussed the-two-leaders.ACT-ADDR last-the-news.PAT
The two leaders discussed latest news

(8) tabāḥaṭa Netanyahu mēa wuzarā`ī tīh
he-discussed Netanyahu.ACT with-ministers-of-him.ADDR
Netanyahu discussed with his ministers

With some other verbs than those of form VI another type of reciprocal relation can be found, if some other actants than Actor, e.g. PAT and EFF, enter the reciprocal usage. It is the case when the verb gāran III (“to compare”) is used with preposition bayn (10). Its standard valency frame can be determined as ACT60 PAT60 (4.) EFF60 (bi-) and is illustrated on (10). For the reciprocal usage, this verb receives the attribute Rcp: PAT-EFF.

(9) sa-yuqārinu ʾAristū bayna l-ʾaġnāṣī
he-will-compare Aristotle.ACT between-the-genes.PAT-EFF
Aristotle will compare the genes

(10) wā-yuqārinu ʾl-ʾadāba bi-ʾl-fumūnī
and-he.ACT-compares the-literature.PAT with-the-arts.EFF
and he compares the literature with arts

6. Conclusions and Future Work
In our contribution, we have overviewed the current status of our ongoing work on the valency lexicon of Arabic verbs, which is an integral part of the ElixirFM lexicon. We have sketched the main characteristics of the latest release of ElixirFM including its web interface and we have reviewed the projects and sources for MSA that are related to our work. We have also outlined our theoretical background (Functional Generative Description), the present coverage of the valency frames of ElixirFM, and the representation of various valency phenomena in form of dependency trees. Last but not least, we have discussed issues about depicting the reciprocal meaning of verbs.
In the forthcoming phase of our work, we would like to focus on annotating less usual, figurative and idiomatic meanings of verbs as well as on solving some more intricate issues, such as the valency status of some

---

4 This distinction between prepositions and prepositionals is made according to Badawi et al. (2004).
5 It might be possible to include among the explicit morphemic constraints a possibility that one of valency slots is realized by the imperfect form of a verb (e.g. for verbs as ġa‘al oratak), but it is not clear yet if this element can be regarded as an actant or circumstantial qualifier (Badawi et al., 2004). This issue must be subject to further examination.
auxiliary, modal, impersonal and other defective verbs. We also intend to enrich the valency frames with additional information on potential passivization, correlation between secondary derived verb forms with their original counterparts, or semantic-syntactic class. We would like to link the entries with the instances of occurrence in PADT, as well as with other corpus examples and Czech glosses. We would like to continue building this complex language resource and make it most reusable for various NLP tasks and applications.

Acknowledgements
We would like to thank Petr Zemánek for his valuable comments and suggestions. This work has been funded by the Ministry of Education of the Czech Republic, projects MSM0021620823 and MSM0021620838.

References
Figure 1: Screenshot of the Derive mode of the ElixirFM Online Interface showing verbonominal derivatives (verbal noun, active and passive participles) for the verb *tarğam* ("to translate").

Figure 4: Screenshot of part of the lexical entries nested under the quadriliteral root *ṭrğm*. The ElixirFM lexicon is displayed in the TrEd environment. The verbal lexeme *ṭarğam* contains two lexical units, since two different senses are distinguished (1. "to translate, interpret"; 2. "to write someone’s biography"). The first level of the valency frame (dependency tree) represents slots for various valency complements; those marked with dashed edges are considered to be optional. Their children represent all alternative morphemic representations of such complementations, which are expressed by their subtrees. Entries are encoded in form of morphophonemic patterns – the pattern *KaRDaS* corresponds to the quadriliteral strong verb that is traditionally denoted by the pattern *fa’lal*. 
Figure 5: Screenshot of the valency frame for the reciprocal verb *tabāḥat* ("to discuss, to confer") with obligatory Actor, and Addressee. Optional Patient allows two possible surface morphemic realizations with prepositions *fī* and *hawla*. The reciprocal attribute Rcp: ACT-ADDR should be attached to this valency frame in order to emphasize the implicit reciprocal meaning of this verb.