ISBN 978-83-60810-27-9
ISSN 1896-7094

# PerGram: A TRALE Implementation of an HPSG Fragment of Persian

Stefan Müller
German Grammar Group
Freie Universität Berlin
14195 Berlin
Stefan.Mueller@fu-berlin.de

Masood Ghayoomi
German Grammar Group
Freie Universität Berlin
14195 Berlin
Masood.Ghayoomi@fu-berlin.de

***Abstract*—In this paper, we discuss an HPSG grammar of Persian (PerGram) that is implemented in the TRALE system. We describe some of the phenomena which are currently covered. While working on the grammar, we developed a test suite with positive and negative examples from the linguistic literature. To be able to test the coverage of the grammar with respect to naturally occurring sentences, we use a subcorpus of a big corpus of Persian.**

## I. Introduction

IN THE past two decades, Head-driven Phrase Structure Grammar (HPSG, [1], [2]) was used successfully to formalize the phonology, morphology, syntax, semantics, and information structure of various languages. Apart from the role HPSG plays in theoretical linguistics, there is also an active community developing large scale implemented grammar fragments that can be used for parsing and generation. While there are large scale grammar fragments available for major languages like English [3] and German [4], [5], [6], [7], other languages such as Persian are understudied.

In this paper, we describe a fraction of the phenomena that are covered in an implemented fragment of Persian (PerGram). The focus is on the implementation of the grammar.

The structure of the paper is as follows: we will say a few words about Persian in general and also briefly describe some of its specific syntactic properties in Section II. Since the grammar is implemented in the TRALE system, we will introduce the general features of TRALE in Section III. In Section IV, some of the phenomena implemented in PerGram are discussed. Section V describes the test suites we are using. In Section VI, we summarize the paper.

## II. Persian

Persian is a member of the Indo-European language family and it has many features in common with the other languages of this family in terms of phonology, morphology, syntax, and lexicon. Persian uses a modified version of the Arabic script and is written right to left. However, the two languages differ from one another in many respects. Persian belongs to the subject-drop languages with an SOV constituent order in unmarked structures. The constituent order is relatively free. Verbs are inflected for tense and aspect, and they agree with the subject in person and number. The particle /rā/ is used as an object marker. The language does not make use of gender [8]. Persian has simple and compound prepositions [9], [10]. There exists a closed list of simplex verbs and an open list of compound verbs.

## III. TRALE

The TRALE system [11], [12], [13] is an extension of the Attribute Logic Engine (ALE) [14], [15]. TRALE has a parser and a generator and can be used to implement theoretical HPSG proposals rather directly. The only specification that is needed in addition to the constraints that are known from theoretical papers is a phrase structure backbone that has to be defined to guide the parsing process. The system is run with a graphical user interface called Grale that can be used to visualize lexical items, lexical rules, types, and macros as an Attribute Value Matrix (AVM) including relational constraints that might be attached to the respective objects. The GUI does unfilling, that is, features and types that are not more specific than the types in the type definition (the signature, see below) are not displayed. This is a very important feature for debugging large scale grammars. The most recent extension of TRALE features a Java-based graphical debugger, which makes it possible to debug the unification operations stepwise and to debug relational constraints during parsing [16].

To declare a grammar in TRALE, a *signature* file is required which defines the types and the features that are appropriate for them. In addition, a TRALE grammar has one or several files containing the definitions of lexical items, lexical rules, phrase structure rules, relational constraints, principles, etc.

To make the development of large scale grammars feasible, macros can be used in the definition of linguistic objects. Like types, macros can be organized in hierarchies. While the type hierarchy is stated explicitly in the signature file, the macro hierarchy is defined rather implicitly by calling macros in the definition of other macros. In contrast to types, macros can be parametrized. (1) shows an example of a parametrized macro in the lexical entry for /man/ ('I'):

(1) man ~~> @pers_pronoun(first, sg, human).

The macro for personal pronouns takes three parameters: one for person, one for number, and one for the semantic type of the pronoun. Macros are very useful for the development

978-83-60810-27-9/09/$25.00 © 2010 IEEE

of the lexicon since they hide the complexity of the grammar and therefore make it possible for inexperienced users to write lexical entries.

While HPSG in general deals with linguistic phenomena belonging to all main dimensions of grammar (phonology, morphology, syntax, semantics, pragmatics), the implementation currently considers morphology, syntax, and semantics. The syntactic analysis uses the feature geometry and makes the basic assumptions that were worked out in [7] for German. See [17] for the details on Persian. The semantic analysis is based on Minimal Recursion Semantics (MRS) introduced by Copestake et al. [18]. MRSes can be displayed using 'utool' [19]. 'utool' also provides a scope resolver for MRS.

The TRALE system supports Unicode and it is therefore possible to parse Persian text that is written in the Arabic script. However, since in written Persian short vowels are omitted, a lot of information such as Ezafe, which we will describe in the following section, is left implicit. In order to be able to formalize and test constraints regarding the distribution of linguistically important short vowels, we transcribe Persian words with Latin characters. We already have a version of the lexicon in Arabic script and plan to extend the grammar in a way that makes it possible to use it with or without the transcription.

The implemented fragment of Persian shares a common core with fragments of German, Mandarin Chinese [20], Danish [21], and Maltese [22]. For more information on this core and for downloading the grammars see http://hpsg.fu-berlin.de/Projects/core.html.

## IV. The Covered Linguistic Phenomena

### A. Principles and Schemata

The grammar uses several immediate dominance schemata and principles that are similar to the ones that were originally suggested by P&S94 [2]. The Persian grammar uses a Head-Adjunct-Schema, a Head-Complement-Schema, a Head-Specifier-Schema, and a Head-Filler-Schema. In addition to these schemata, a Head-Cluster-Schema is used for the formation of complex predicates (see [23], [24], [17] for analyses of predicate complexes in German and Persian). In addition to these more general ID schemata, the grammar uses language specific schemata for the combination of nominal elements with their possessor and for noun compounding.

Principles that hold for all of the mentioned ID schemata are factored out of the schemata and are represented as constraints on an appropriate type, for instance *phrase* or *headed-phrase*. Examples of such principles are the 'head feature principle', the 'semantics principle', the 'specifier principle', and the 'nonlocal feature principle'.

### B. Morphological Rules

The grammar contains morphological rules both for derivation and inflection. The morphological rules are modeled as lexical rules. For instance, for inflectional morphology, we use lexical rules that map roots or stems to fully inflected words. The following lexical rule (LR) is responsible for noun inflection. It maps a nominal stem onto a word with exactly the same syntactic and semantic properties.

$$\begin{bmatrix} \text{PHON}\ \boxed{1} \\ \text{SS}\ \boxed{2}\big[\text{LOC} \mid \text{CAT} \mid \text{HEAD}\ \textit{noun}\big] \\ \text{AFFIX} \begin{bmatrix} \text{PHON} & \boxed{3} \\ \text{NUM} & \boxed{4} \\ \text{SORT} & \boxed{5} \\ \textit{noun-i-affix} & \end{bmatrix} \\ \textit{stem} \end{bmatrix} \mapsto \begin{bmatrix} \text{PHON}\ \boxed{1} \oplus \boxed{3} \\ \text{SS}\ \boxed{2} \left[\text{LOC} \mid \text{CONT} \mid \text{IND} \begin{bmatrix} \text{NUM}\ \boxed{4} \\ \text{SORT}\ \boxed{5} \end{bmatrix}\right] \\ \textit{word} \end{bmatrix}$$

The input of the rule has a special feature the value of which contains the information about the affix. For nominal inflection, the affix has to have the type *noun-i-affix*. There is a constraint on this type that disjunctively specifies the inflectional paradigm.

$$\textit{noun-i-affix} \Rightarrow \begin{bmatrix} \text{PHON} & \langle\rangle \\ \text{NUM} & \textit{sg} \end{bmatrix} \vee \begin{bmatrix} \text{PHON} & \langle \textit{ān} \rangle \\ \text{NUM} & \textit{pl} \\ \text{SORT} & \textit{human} \end{bmatrix} \vee \begin{bmatrix} \text{PHON} & \langle \textit{hā} \rangle \\ \text{NUM} & \textit{pl} \end{bmatrix}$$

The respective PHON values provide information about the phonological contribution of the stem and the affix. These values are lists of phonemes and they are concatenated in the output of the lexical rule. ⊕ stands for the *append* relation. In addition to the concatenation of the PHON values, the values of NUM and SORT of the output of the rule are instantiated with the features provided by the affix. SORT is a feature that is used to enforce selectional restrictions. The values are based on a semantic ontology which will be described in the subsection IV-E.

In the lexical rule given above, the SYNSEM value of the input is identified with the SYNSEM value of the output. This is different in LRs for derivational morphology. For instance, in the LR that derives adjectives from verbs by appending the suffix *-i* ('-able'), the part-of-speech and the valence specification changes. In addition to these syntactic changes, the semantic contribution of the verb is embedded under a modal operator.

Apart from this derivational LR, we have lexical rules for participle to adjective conversion and for agentive nominalizations. All these morphological rules interact properly with the formation of complex predicates. See [17] for details.

### C. Ezafe

The so-called Ezafe is a short vowel /e/ which functions to link the elements of a noun phrase (see for instance [25]). Ezafe appears on: a noun before another noun (attributive); a noun before an adjective; a noun before a possessor (noun or pronoun); an adjective before another adjective; a pronoun before an adjective; the first names before the last names; a combination of the above [26]. Ezafe is realized as /e/ after consonants and /i/ and as /ye/ after vowels other than /i/. Ezafe does not appear on a bare noun or adjective and its appearance indicates that the end of the syntactic phrase is not reached.

We defined an LR which adds the Ezafe at the end of a word. To distinguish Ezafe-marked words from unmarked ones, we employ a binary valued feature EZAFE. The lexical rule applies to words that have the value '−' and licenses words with the value '+'.

EZAFE is an edge feature, that is, a complex phrase is Ezafe-marked if an Ezafe is present at its right periphery. (This is similar to the possessive *'s* in English.) The Ezafe marking of phrases is taken care of by the following constraint on phrases:

$$\textit{phrase} \Rightarrow \begin{bmatrix} \text{SS} \mid \text{EZAFE}\ \boxed{1} \\ \text{DTRS}\ \boxed{2} \end{bmatrix} \wedge \textit{last}(\boxed{2}, \left[\text{SS} \mid \text{EZAFE}\ \boxed{1}\right])$$

The relational constraint *last* succeeds if the second argument ($[\text{SS} \mid \text{EZAFE}\ \boxed{1}]$ in the example above) is the last element of the list that is provided as the first argument. The DTRS list is a list of daughters that is ordered according to the surface order of constituents, so *last* returns the rightmost daughter. The EZAFE value of this daughter is shared with the EZAFE value of the mother. Since there is no reference to the number of daughters in the constraint above, it applies to unary and binary branching phrases alike. Currently we only have unary and binary branching rules in the grammar, but of course the constraint would apply to structures with three or more daughters as well.

The distribution of the Ezafe is constrained by implicational statements like the following:

$$\begin{bmatrix} \text{HEAD-DTR} \mid \text{SS} \mid \text{LOC} \mid \text{CAT} \mid \text{HEAD}\ \textit{noun} \\ \textit{head-argument-phrase} \end{bmatrix} \Rightarrow \left[\text{HEAD-DTR} \mid \text{SS} \mid \text{EZAFE}\ -\right]$$

This constraint applies to combinations of nouns with their arguments. Since prepositional arguments have to be realized outside of the Ezafe domain, the head daughter is required to have the EZAFE value '−'. The schema that is used for the combination of a noun with a possessor requires the nominal constituent to have the EZAFE value '+'.

### D. Negation

A verb or an auxiliary can be negated by attaching the prefix /na-/. This is implemented by a lexical rule that adds the phonological material and embeds the content of the verb under the negation relation. The syntactic properties of the verb are not affected by the negation and are carried over from the input of the rule to the output. Persian differs from languages like German in that it is impossible to negate a non-finite verb that is embedded under a modal. This is captured by a constraint that requires that the input to the lexical rule is a finite verb.

The LR applies to auxiliaries, simplex verbs, and the verbal element of complex predicates. In the latter case, the negation scopes over the whole complex predicate even though the negation attaches to the verb before the other part of complex predicate is combined with the negated verb. For details see [17].

### E. Nouns

Several kinds of nouns are modeled. We implemented common nouns with and without arguments. The arguments are always optional and we have subclasses for nouns that take CPs and for nouns that take PPs as complement. In addition to common nouns, the grammar contains lexical items for proper nouns. Apart from common nouns and proper nouns, we have lexical entries for nouns that play a special role in complex predicate formation (process nouns and verbal nouns). See [17] for details on these nouns.

All non-expletive linguistic objects are classified with respect to an ontology. The ontology contains types like *human*, *agentive*, *substance*, and *geo-location*. This ontology is an extended version of the ontology that was developed in Verb*mobil* [27]. It can be used to specify sortal restrictions of governing verbs with respect to their arguments. Apart from this, it can be used to enforce certain syntactic constraints. For instance, one allomorph of the plural affix /-ān/ is only used with nouns that refer to humans.

### F. Verbs

In Persian, there are two classes of verbs: a closed list of simplex verbs; and an open list of compound verbs. The latter group is composed of a preverbal and a verbal element. The verbal elements which belong to a subclass of the simplex verbs are called 'light verbs' and the whole predicate formation process is called 'light verb construction'. The implementation of this phenomenon in our grammar is based on [17] and is not discussed here.

Currently, the grammar has lexical entries for the following kinds of verbs: mono-valent verbs with one NP argument, bi-valent verbs with two NP arguments, bi-valent verbs with an NP and a PP argument, ditransitive verbs with two NPs and a PP, verbs with an NP and a clausal argument, copula verbs, modal verbs, mono-valent unaccusative verbs, and several types of light verbs.

As said, Persian is a language with a relatively free constituent order. This is captured by allowing the combination of an arbitrary element of the SUBCAT list with a head. For instance the head may be combined with the subject and then with the object or the other way round. Languages with strict constituent order like English do not allow this but require the combination of heads with the arguments in order of their obliqueness. See [28] for further discussion of this difference.

We follow [2] in assuming that there is a special representation of valence information, nowadays called Argument Structure (ARG-ST). For all heads there is a mapping from the ARG-ST list to other valence features like COMPS and SPR. For SVO languages like English and Danish the least oblique argument of a verb is mapped to SPR and all the other arguments are mapped to COMPS[1] [29]. The verb forms a VP together with the arguments that are selected via COMPS. This VP is combined with the subject to form a sentence. In contrast, in languages like German and Persian, all arguments of finite verbs are mapped to COMPS. This makes it possible to account for orders like OSV, in which the object and the verb are not adjacent.

While the arguments can be realized in any order with respect to each other, the order with respect to their head

[1] For historical reasons COMPS is still called SUBCAT in the implementation.

is rather fixed: Persian is an SOV language, that is, the verb follows its arguments.[2] On the other hand, Persian has prepositions, that is, the adposition precedes its complement. To capture this, we assume two phrase structure rules that are instances of the general Head-Argument-Schema: one head initial and another one head final. We use a head feature INITIAL, which has the value '+' for heads that are serialized initially and '−' for heads that follow their arguments.

All tense and aspect forms of the verbal paradigm are covered. The progressive and subjunctive marking is done by inflectional lexical rules. The auxiliaries that are used for periphrastic forms are described in [17] and the description will not be repeated here. Three types of complex predicate formation are also discussed in [17] and covered in the grammar.

### G. Agreement

Subject verb agreement is handled by the lexical rule that licenses finite verbs. The rule requires that the least oblique NP in the ARG-ST list that bears structural case shares its person and number values with the person and number features of the inflectional affix. This treatment of subject verb agreement is also used for the grammars of German, Maltese, and Hindi. For a similar analysis of agreement in Spanish see [30]. In comparison to the other languages, Persian allows for an additional case: Plural NPs referring to non-agentive entities may also occur with verbs inflected for third person singular.

### H. Prepositions

As mentioned above, prepositions differ from verbs in governing their complement to the right.[3] This is enforced by their INITIAL value, which is '+'. The fragment currently contains prepositions that form PPs that can be used as arguments, prepositions that can be used as modifiers, and a preposition similar to the English preposition *by* that can be used in passive constructions.

### I. Clitics

Pronominal clitics can be used as possessives (2). As in noun phrases with a full NP as possessor phrase, the clitic is the rightmost element in the NP. Therefore there are two possible hosts for clitic attachment: a noun as in (2a) or an adjective as in (2b):

(2) a. ketāb=aš
book=3SG
'his/her book'

b. ketāb-e jadid=aš
book-EZ new=3SG
his/her new book

Clitics can also fill the slot of the direct object of a verb. In this case, the clitic attaches to the verb as in (3a) or to the preverbal element in a complex predicate construction (3b,c), or to the future auxiliary (3d):

(3) a. did-am=aš.
saw-1SG=3SG
'I saw him/her.'

b. bāz=aš kard-am.
open=3SG did-1SG
'I opened it.'

c. dust=aš dār-am.
friend=3SG have-1SG
'I love him/her.'

d. dust xāh-am=aš dāšt.
friend FUT-1SG=3SG have.SG
'I will love him/her.'

Currently these clitics are treated as postlexical clitics, that is, clitics are treated in the syntactic component. For clitics that fill the object slot of verbs, there is a special grammar rule in the phrase structure backbone used by TRALE. This rule is necessary since the order of clitic and verb is different from the usual order. Apart from these order differences, this grammar rule is an instance of the general Head-Argument-Schema. The noun possessor construction has the same structure for possessors that are realized as clitics and for possessors realized as full NPs. The only difference is the impossibility of the Ezafe when the possessor is realized as a clitic.

However, Samvelian argued for a treatment of clitics as phrasal affixes [25], and the grammar will be adapted in order to cover lexical and morphological idiosyncrasies.

### J. Determiners

Persian makes use of prenominal and postnominal determiners: demonstratives like /in/ ('this') and /ān/ ('that') and the indefinite element /yek/ ('a/an') are realized prenominally and the indefinite marker /*-i*/ occurs postnominally.

The prenominal determiners are treated as dependents of the noun and are selected via the SPR feature as suggested by [2]. To establish the semantic relation between the quantifier and the noun, the determiner is able to access the variable of the nominal projection via the SPEC feature.

The indefinite marker /*-i*/ is currently not covered but the implementation of a lexical rule that has the desired phonological, syntactic, and semantic effects is straight forward.

### K. Direct Object Marker

The particle /rā/ is used as a marker of a noun in object position. Since object marking is optional, we treat /rā/ as an adjunct. The following is the lexical entry of the particle /rā/:

$$\begin{bmatrix}
\text{PHON} & \langle \textit{r}\bar{\textit{a}} \rangle \\
\text{SS} \mid \text{L} & \begin{bmatrix} \text{CAT} \mid \text{HEAD} \mid \text{MOD} \mid \text{L} & \begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \text{CASE} \mid \text{MORPH-CASE}\ \textit{acc} \\ \textit{noun} \end{bmatrix} \\ \text{SPR} & \langle\rangle \end{bmatrix} \\ \text{CONT} & \boxed{1} \begin{bmatrix} \text{LTOP} & \boxed{2} \\ \text{IND} & \boxed{3} \end{bmatrix} \end{bmatrix} \\ \text{CONT} & \boxed{1} \end{bmatrix} \\
\text{RELS} & \left\langle \begin{bmatrix} \text{LTOP} & \boxed{2} \\ \text{ARG1} & \boxed{3} \\ \textit{specific} & \end{bmatrix} \right\rangle
\end{bmatrix}$$

The modified linguistic object is required to have accusative case. This ensures that /*rā*/ attaches to an object. In addition, it is required to have an empty SPR list, i.e. it has to be an NP. We assume that the list of relations that is contributed by signs is not part of CONT, but rather represented at the top level of

[2] We are aware of the fact that arguments may be realized postverbally. Currently, only the postverbal realization of clausal arguments is implemented. We leave the other serialization options to further research.

[3] This shows that the assumption of a headedness parameter that has the same value for verbs and prepositions makes the wrong predictions for Persian. Therefore such a parameter should not be part of an innately specified UG, if there is such a thing at all. See [31] on a detailed discussion of language acquisition including a discussion of Principle & Parameter approaches.

the feature description. This yields a more restrictive theory as far as the locality of selection is concerned [32]. Since RELS is not part of CONT, the CONT values of the modifier and the modified element can be shared. */rā/* is treated as an intersective modifier, and hence the LTOP of the modified noun is identified with the LTOP of the specificity relation contributed by */rā/*. Of course, the argument of the specificity relation is identical to the referential index of the NP ($\boxed{3}$).

### *L. Complementizers for Subordinated Clauses and Relative Clauses*

Embedded clauses are introduced by the complementizer */ke/* ('that'). We treat the complementizer as a head that selects a finite clause. In addition to this complementizer, we have another lexical entry for */ke/* which is used in relative clauses [33]. The grammar covers relative clauses with and without resumptive pronouns. Free relatives will be implemented in the near future. The analysis is described in detail in Taghvaipour's thesis about relative clauses.

### *M. Pro Drop*

Initially pro drop was handled by a unary branching rule that discharges the subject from a valence representation after all other arguments of a finite verb have been saturated. While this rule works correctly, it is inefficient since it contributes edges to the bottom up chart parser even if an overt subject is present in a clause. We, therefore, implemented an underspecification analysis that originally was used for imperatives in BerliGram, an implementation of a grammar of German [7]. The grammars use a valence representation that comes with an additional feature REALIZED [34], [35]. Arguments that are not realized have the value '−' and once they are realized, the value is changed to '+' in the valence representation at the mother node. This representation can be used to represent optional arguments by underspecification (see [36] for a different solution using a binary feature): For the subject of finite verbs, the REALIZED value can remain unspecified. The unspecified value is compatible with '+' in the pro drop case and '−' in the case when a verbal projection is combined with an overt subject. During a parse of the 165 test sentences from the PerGram Test Suite (see Section V), the grammar version with unspecified REALIZED value licensed 12.7 % less passive edges in comparison to the one that uses the unary branching rule. This resulted in a reduction of parse times of 30.6 % in average.

### *N. Coordination*

The grammar handles symmetric coordinations. The coordination of two or more NPs with */va/* ('and') results in a plural NP, so that the agreement facts are captured correctly. The analysis of coordination is basically the one suggested by [2], that is, the CAT and NONLOCAL values of the conjuncts are shared. However, there is a slight complication: since we use a non-cancellation approach to valence, examples like the one in (4) are problematic.

(4) Ali [[mard rā did] va [xandid]].
Ali man RA saw.3SG and laughed.3SG
'Ali saw the man and laughed.'

In (4), a VP with a transitive verb and one with an intransitive verb is coordinated. The valence representations of the respective VPs is shown in (5):

(5) a. mard rā did: SUBCAT ⟨ NP, ~~NP~~ ⟩
b. xandid: SUBCAT ⟨ NP ⟩

The valence representation of the VP with a transitive verb contains an NP that is marked as realized. Meurers [34] called realized arguments *spirit*. In comparison, the VP with the intransitive verb does not contain such an NP. The consequence is that the SUBCAT values of the conjuncts cannot be unified since the lengths of the lists are different. This problem is solved in the implemented grammar by using a relational constraint that returns all unrealized elements in a valence list. This constraint is applied to both conjuncts and the respective result lists are unified and represented at the mother node. As a result, the valence representation of [[mard rā did] va [xandid]] is ⟨ NP ⟩. The spirit NP (~~NP~~) is not represented at the mother node. The subject NPs of */didan/* ('see') and */xandan/* ('laugh') are unified and hence it is explained why *Ali* fills the respective slots of both verbs.

This analysis captures a lot of complicated coordination phenomena like Across the Board Extraction in unbounded dependencies (questions and relative clauses) and also interacts nicely with resumptive pronouns in relative clauses. However, there is one problem left: the analysis does not extend to German fronting data and case assignment for which it was introduced originally. Meurers [34] and Przepiórkowski [35] suggested representing the saturated complements at the VP level. Auxiliary verbs attract the arguments of the verb they embed. Case is assigned to arguments that are not raised by higher verbs. In the analysis of the sentences in (6), the auxiliary attracts the arguments of *gelesen* ('read') and assigns nominative to the subject and accusative to the object.

(6) a. [Einen Aufsatz gelesen] hat er nicht.
an essay.ACC read has he.NOM not
'He did not read an essay.'
b. [Ein Aufsatz gelesen] wurde nicht.
an essay.NOM read was not
'An essay was not read.'

The important point is that the case is not determined locally in the fronted VP, but assigned by the finite verb. In order to assign case, the finite verb has to raise the realized argument (the spirit) and hence it has to be accessible at the VP node.

With this background, the problem of the coordination analysis is obvious: We can coordinate two VPs and front them. According to the coordination analysis sketched above, realized arguments are not contained in the valence lists of the mother nodes of coordinated structures. Since these spirits are needed for case assignment, we have conflicting demands: coordination requires VPs with verbs of different arity to be syntactically parallel and case assignment (in German) requires all arguments to be present at the VP node.

(7) a. [[Einen Aufsatz gelesen] und [einen Report
an essay.ACC read and a report.ACC
geschrieben]] hat er nicht.
written has he.NOM not
'He neither read an essay nor did he write a report.'

b. [[Ein Aufsatz gelesen] und [ein Report
an essay.NOM read and a report.NOM
geschrieben]] wurde nicht.
written was not
'An essay was not read and a report was not written either.'

It remains to be seen if it is possible to develop a consistent analysis of coordination and non-cancellation approaches to valence.

### *O. Empty Elements*

Currently, two types of analyses of unbounded dependency phenomena are entertained in the HPSG framework: one assumes an empty element for the introduction of a nonlocal dependency [2] and the other one introduces nonlocal dependencies lexically [37]. See [38] for an extended discussion. In our implementation we adopt a trace-based approach.

In addition to the empty element that is used in nonlocal dependencies, we also use an empty determiner for the analysis of NPs that do not have an overt determiner. The empty determiner introduces the quantifier relation that is needed for the interpretation of the NP. The alternative to this treatment would be a unary branching ID schema that discharges the valence requirement represented under SPR and introduces the appropriate semantics. By adopting this solution, one would miss the generalization about determiners. In our approach, the type definitions for overt determiners can be used for the covert determiner as well. No idiosyncratic ID rule is needed. There is just one place in the grammar where it is said that the phonology of a determiner may be empty.

We agree that empty elements should be avoided wherever possible and that they should not be stipulated on a cross-linguistic basis but rather be motivated by evidence from within the language under consideration. That is, a topic morpheme in Japanese should not be seen as evidence for an empty topic head in German, English, and Danish. If empty elements are assumed based on the evidence from the language under consideration, the language acquisition model does not have to assume a rich UG, but is compatible with data-driven approaches like the one that is entertained by Tomasello [39].

As is known from research on formal grammars [40], phrase structure grammars with empty elements can be converted into grammars without empty elements. This result does not transfer directly to grammars with typed feature structures, but most of the grammars that are currently suggested can be converted into grammars without empty elements by applying the techniques developed for PSGs (See [24], [31] for examples). TRALE does this kind of grammar conversion for the relevant cases automatically and transparently for the user and hence the grammars can be specified in a more compact way. For example, in the case of the empty determiner, a special variant of the Head-Specifier-Schema is created that has no daughter for the determiner. That is, TRALE compiles the grammar into one that has the unary branching rule that was mentioned above. See [41] for further discussion.

## V. The Test Suites

During the development of the Persian grammar, we put together two test suites that are used for systematic testing and grammar profiling [42]. The first one contains examples from the linguistic literature that are relevant for the phenomena that are covered by the grammar. In addition, it contains ungrammatical examples that were constructed in the development process in order to rule out overgeneration of the grammar which was detected by systematic testing. This test suite consists of 165 sentences including both positive (132 sentences) and negative (33 sentences) examples.

To be able to test the coverage of the grammar, we randomly selected 130 sentences from a Persian corpus called 'Peykare'. Peykare [44] is a big Persian corpus provided by the University of Tehran and the Higher Council for Informatics of Iran. It contains texts from various data sources, both written and spoken.The part-of-speech of each word is annotated according to the EAGLES guidelines. The sentences were selected randomly in such a way that the balancedness of the original corpus is kept in the subcorpus; as a result, the 130 sentences have the variability of the existing registers in the Peykare corpus. Currently, most of the sentences do not parse because of missing lexical entries. We added lexical items for proper nouns, common nouns, and adjectives to the lexicon, but there are other missing lexical items (verbs, clitic forms of the copula, numerals, adverbs, adjectives derived from nouns) that affect 98 of the 130 sentences.

## VI. Summary and Conclusion

In this paper, we briefly described some of the phenomena that are part of an implemented HPSG grammar of Persian. A full description of the phenomena and the analyses will be provided in [43]. The grammar covers the core aspects of Persian syntax and morphology and provides semantic representations in the form of MRS. The grammar is evaluated with respect to the example sentences that were collected from the linguistic literature and ungrammatical examples that were constructed during the development process. In addition, we started experiments with naturally occurring data that was selected from the Peykare corpus.

## Acknowledgment

We thank Pollet Samvelian, Olivier Bonami, and three anonymous reviewers for comments on an earlier version of this paper. Of course all shortcomings of the paper and the implementation are entirely our fault.

The first author thanks Mina Esmaili, Pollet Samvelian, Yasser Shakeri, and Mehran A. Taghvaipour for help with the

Persian data. We thank Emily M. Bender, Felix Bildhauer, Olivier Bonami, Bob Borsley, Karine Megerdoomian, Ivan Sag, Pollet Samvelian, and several anonymous reviewers, for comments on and discussion of the complex predicate paper and/or Persian grammar in general.

This work was supported by a grant from Agence Nationale de la Recherche and Deutsche Forschungsgemeinschaft to the Franco-German project *PER-GRAM: Theory and Implementation of a Head-driven Phrase Structure Grammar for Persian* (DFG MU 2822/3-1).

## References

[1] C. J. Pollard and I. A. Sag, *Information-Based Syntax and Semantics*. Stanford: CSLI Publications, 1987.

[2] ——, *Head-Driven Phrase Structure Grammar*. University of Chicago Press, 1994.

[3] D. P. Flickinger, A. Copestake, and I. A. Sag, "HPSG analysis of English," in *Verbmobil: Foundations of Speech-to-Speech Translation*, W. Wahlster, Ed. Berlin: Springer Verlag, 2000, pp. 254–263.

[4] S. Müller, "The Babel-System—an HPSG Prolog implementation," in *Proceedings of the Fourth International Conference on the Practical Application of Prolog*, London, 1996, pp. 263–277.

[5] S. Müller and W. Kasper, "HPSG analysis of German," in *Verbmobil: Foundations of Speech-to-Speech Translation*, W. Wahlster, Ed. Berlin: Springer Verlag, 2000, pp. 238–253.

[6] B. Crysmann, "On the efficient implementation of German verb placement in HPSG," in *Proceedings of RANLP 2003*, Borovets, Bulgaria, 2003, pp. 112–116.

[7] S. Müller, *Head-Driven Phrase Structure Grammar: Eine Einführung*, 2nd ed. Tübingen: Stauffenburg Verlag, 2008.

[8] S. Mahootiyan, *Persian*. Routledge, 1997.

[9] Z. A. Chime, "An account for compound preposition in Farsi," in *Proceedings of the COLING/ACL 2006*, 2006, pp. 113–119.

[10] Z. A. Chime and M. Ghayoomi, "Incorporation: Word production of persian prepositions and its application in computational linguistics," in *Proceedings of the 2nd Workshop on the Persian Language and Computer*, 2006, pp. 16–24.

[11] W. D. Meurers, G. Penn, and F. Richter, "A web-based instructional platform for constraint-based grammar formalisms and parsing," in *Proceedings of the Effective Tools and Methodologies for Teaching NLP and CL*, 2002, pp. 18–25.

[12] G. Penn, "Balancing clarity and efficiency in typed feature logic through delaying," in *Proceedings of ACL 2004*, 2004, pp. 239–246.

[13] S. Müller, "The Grammix CD Rom. a software collection for developing typed feature structure grammars," in *Proceedings of the Grammar Engineering Across Frameworks Workshop 2007*, ser. Studies in Computational Linguistics ONLINE, T. H. King and E. M. Bender, Eds. Stanford: CSLI Publications, 2007.

[14] B. Carpenter and G. Penn, "Efficient parsing of compiled typed attribute value logic grammars," in *Recent Advances in Parsing Technology*, H. Bunt and M. Tomita, Eds., no. 1. Dordrecht: Kluwer Academic Publishers, 1996, pp. 145–168.

[15] G. Penn and B. Carpenter, "ALE for speech: a translation prototype," in *Proceedings of the 6th Conference on Speech Communication and Technology (EUROSPEECH)*, G. Gordos, Ed., Budapest, Hungary, 1999.

[16] J. Dellert, K. Evang, and F. Richter, "Kahina, a debugging framework for logic programs and TRALE," 2010, presentation at the HPSG 2010 Conference.

[17] S. Müller, "Persian complex predicates and the limits of inheritance-based analyses," *Journal of Linguistics*, vol. 46, no. 3, To Appear 2010, http://hpsg.fu-berlin.de/~stefan/Pub/persian-cp.html.

[18] A. Copestake, D. Flickinger, C. Pollard, and I. A. Sag, "Minimal recursion semantics: An introduction," *Research on Language and Computation*, vol. 4, no. 3, pp. 281–332, 2006.

[19] A. Koller and S. Thater, "Efficient solving and exploration of scope ambiguities," in *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. Ann Arbor: ACL, 2005, pp. 9–12.

[20] S. Müller and J. Lipenkova, "Serial verb constructions in Mandarin Chinese," in *Proceedings of the 16th International Conference on Head-Driven Phrase Structure Grammar*, S. Müller, Ed. Stanford: CSLI Publications, 2009, pp. 234–254.

[21] B. Ørsnes, "Preposed sentential negation in Danish," in *Proc. of the 16th International Conference on Head-Driven Phrase Structure Grammar*, S. Müller, Ed.Stanford: CSLI Publications, 2009, pp. 255–275.

[22] S. Müller, "Towards an HPSG analysis of Maltese," in *Introducing Maltese Linguistics*, B. Comrie, R. Fabri, B. Hume, M. Mifsud, T. Stolz, and M. Vanhove, Eds. Amsterdam, Philadelphia: John Benjamins Publishing Co., 2009, pp. 83–112.

[23] E. W. Hinrichs and T. Nakazawa, "Linearizing AUXs in German verbal complexes," in *German in Head-Driven Phrase Structure Grammar*, J. Nerbonne, K. Netter, and C. J. Pollard, Eds. Stanford: CSLI Publications, 1994, pp. 11–38.

[24] S. Müller, *Complex Predicates: Verbal Complexes, Resultative Constructions, and Particle Verbs in German*. Stanford: CSLI Publications, 2002.

[25] P. Samvelian, "A (phrasal) affix analysis of the Persian Ezafe," *Journal of Linguistics*, vol. 43, pp. 605–645, 2007.

[26] A. Kahnemuyipour, "Persian ezafe construction revisited: Evidence for modifier phrase," in *Proceedings of the 21st Conference of the Canadian Linguistic Association*, 2000, pp. 173–185.

[27] W. Wahlster, Ed., *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin: Springer Verlag, 2000.

[28] S. Müller, "Head-Driven Phrase Structure Grammar," in *Syntax – Ein internationales Handbuch zeitgenössischer Forschung*, 2nd ed., A. Alexiadou and T. Kiss, Eds. Berlin: Walter de Gruyter Verlag, in Preparation, http://hpsg.fu-berlin.de/~stefan/Pub/hpsg-hsk.html.

[29] ——, "On predication," in *Proceedings of the 16th International Conference on Head-Driven Phrase Structure Grammar*, S. Müller, Ed. Stanford: CSLI Publications, 2009, pp. 213–233, http://hpsg.fu-berlin.de/~stefan/Pub/predication.html.

[30] C. Vogel and B. Villada, "Spanish psychological predicates," in *Grammatical Interfaces in HPSG*, R. Cann, C. Grover, and P. Miller, Eds. Stanford: CSLI Publications, 2000, pp. 251–266.

[31] S. Müller, *Grammatiktheorie: Von der Transformationsgrammatik zur beschränkungsbasierten Grammatik*. Tübingen: Stauffenburg Verlag, To Appear, http://hpsg.fu-berlin.de/~stefan/Pub/grammatiktheorie.html. English translation in preparation.

[32] M. Sailer, "Local semantics in Head-Driven Phrase Structure Grammar," in *Empirical Issues in Formal Syntax and Semantics*, O. Bonami and P. C. Hofherr, Eds. Online, 2004, vol. 5, pp. 197–214. [Online]. Available: http://www.cssp.cnrs.fr/eiss5/sailer/index_en.html

[33] M. A. Taghvaipour, "Persian relative clauses in Head-driven Phrase Structure Grammar," Ph.D. dissertation, Department of Language and Linguistics, University of Essex, 2005.

[34] W. D. Meurers, "Raising spirits (and assigning them case)," *Groninger Arbeiten zur Germanistischen Linguistik (GAGL)*, vol. 43, pp. 173–226, 1999, http://www.sfs.uni-tuebingen.de/~dm/papers/gagl99.html.

[35] A. Przepiórkowski, "On case assignment and "adjuncts as complements"," in *Lexical and Constructional Aspects of Linguistic Explanation*, G. Webelhuth, J.-P. Koenig, and A. Kathol, Eds. Stanford: CSLI Publications, 1999, pp. 231–245.

[36] D. P. Flickinger, "On building a more efficient grammar by exploiting types," *Natural Language Engineering*, vol. 6, no. 1, pp. 15–28, 2000.

[37] G. Bouma, R. Malouf, and I. A. Sag, "Satisfying constraints on extraction and adjunction," *Natural Language and Linguistic Theory*, vol. 19, no. 1, pp. 1–65, 2001.

[38] R. D. Levine and T. E. Hukari, *The Unity of Unbounded Dependency Constructions*. Stanford: CSLI Publications, 2006.

[39] M. Tomasello, *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge: Harvard University Press, 2003.

[40] Y. Bar-Hillel, M. A. Perles, and E. Shamir, "On formal properties of simple phrase-structure grammars," *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, vol. 14, no. 2, pp. 143–172, 1961.

[41] S. Müller, "Elliptical constructions, multiple frontings, and surface-based syntax," in *Proceedings of Formal Grammar 2004, Nancy*, G. Jäger, P. Monachesi, G. Penn, and S. Wintner, Eds. Stanford: CSLI Publications, To Appear. [Online]. Available: http://hpsg.fu-berlin.de/~stefan/Pub/surface.html

[42] S. Oepen and D. P. Flickinger, "Towards systematic grammar profiling. Test suite technology ten years after," *Journal of Computer Speech and Language*, vol. 12, no. 4, pp. 411–436, 1998.

[43] O. Bonami, S. Müller, and P. Samvelian, *Persian in Head-Driven Phrase Structure Grammar*, In Preparation.

[44] M. Bijankhan, "The role of corpora in writing grammar," *Journal of Linguistics,* vol. 19, no. 2, pp. 48-67, 2004, Tehran: Iran University Press.