Romanian Valence Dictionary in XML Format

Ana-Maria Barbu[*], Emil Ionescu[§], Verginica Barbu Mititelu[♣][*]

[*]Romanian Academy Institute of Linguistics “Iorgu Iordan-Al. Rosetti”
13, Calea 13 Septembrie, Bucharest, 050711, Romania
[§]University of Bucharest, Faculty of Letters
14, Edgar Quinet St, Bucharest, 1, 010018, Romania
[♣]Romanian Academy Research Institute for Artificial Intelligence
13, Calea 13 Septembrie, Bucharest, 050711, Romania
anabarbu@unibuc.ro; emilionescu@unibuc.ro; vergi@racai.ro

Abstract

Valence dictionaries are dictionaries in which logical predicates (most of the times verbs) are inventoried alongside with the semantic and syntactic information regarding the role of the arguments with which they combine, as well as the syntactic restrictions these arguments have to obey. In this article we present the incipient stage of the project “Syntactic and semantic database in XML format: an HPSG representation of verb valences in Romanian”. Its aim is the development of a valence dictionary in XML format for a set of 3000 Romanian verbs. Valences are specified for each sense of each verb, alongside with an illustrative example, possible argument alternations and a set of multiword expressions in which the respective verb occurs with the respective sense. The grammatical formalism we make use of is Head-driven Phrase Structure Grammar, which offers one of the most comprehensive frames of encoding various types of linguistic information for lexical items. XML is the most appropriate mark-up language for describing information structured in HPSG framework. The project can be further on extended so that to cover all Romanian verbs (around 7000) and also other predicates (nouns, adjectives, prepositions).

1. Introduction

The project we present below is called “Syntactic and semantic database in XML format: an HPSG representation of verb valences in Romanian”. It is financially supported by the National University Research Counsel of Romania and represents the joint efforts of linguists, computer scientists and students working at the University of Bucharest, at the Romanian Academy, as well as in the industrial field.

The main aim of our project is to fill a gap in the field of resources and tools for Romanian. At the moment, Romanian lacks a valence dictionary, both in paper version, and in electronic format.

Žabokrtský (2005) distinguishes between two methodologies in building valence dictionaries: *verb-wise* (in which the verb entries are completed individually; this is the case of most paper valence dictionaries) and *frame-wise* (in which verbs with given senses belonging to a certain frame are processed together at a time; this is the case of FrameNet, Fillmore, 2002). We adopt here the former methodology: for each sense of each verb we encode, in XML format, its corresponding arguments with the necessary morphological and syntactic restrictions.

The electronic dictionary we are building will contain about 3000 verbal lemmas and will have the following characteristics:

(1) For every given verb, the dictionary will provide the set of meanings associated with that verb.

(2) For every meaning of a given verb, the dictionary will provide:

(i) an illustrative context

(ii) the corresponding grammatical/inflected forms of the verb

(iii) the corresponding argument structure (ARG-ST) of the verb sense, by indicating the syntactic categories, semantic characteristics and semantic roles of the arguments

(iv) the set of argument alternations (if any), which may be obtained from the conjunction of a given meaning with a given ARG-ST.

(3) For every given verb, the dictionary will provide the set of multiword expressions in which the verb occurs.

While (1) and (2) capture what is (usually) regular in a verb representation, (3) is intended to deal with pure idiosyncratic aspects of verb meaning and grammar.

In order to attain the objectives mentioned above, we use the following tools:

(i) an electronic corpus of Romanian (morphologically and syntactically annotated at the word level)

(ii) an electronic dictionary of the verb morphology in Romanian (about 7000 lemmas)

(iii) a concordance program working on corpora in text or XML format, able to look not only for words but also for XML elements or attributes

(iv) printed dictionaries of Romanian

(v) the framework of the HPSG verb subcategorization.

Given that at this phase of the project the linguistic part is better drawn out than the computational part, we will put emphasis on the linguistic aspects of the research. The article is organized as follows: section 2 outlines related work; section 3 presents the devices involved in the linguistic formalism we commit ourselves to; afterwards we exemplify the encoding of verbal valences at the current phase of the project (section 4), and we motivate the choice for the XML encoding of verbal valences. The conclusions and further work section closes the article.

2. Related Work

Valence dictionaries are dictionaries in which logical predicates (most of the times verbs) are inventoried alongside with the semantic and syntactic information regarding the role of the arguments with which they combine, as well as the syntactic restrictions these arguments have to obey. Valence dictionaries in electronic form have proven their importance in many NLP applications, such as (deep and shallow) parsers' development, Question-Answering, Machine Translation, Information Extraction.

Valence dictionaries are either in paper form (for German, Polish, Slovak, Bulgarian, Russian), or in electronic form (for English, German, Japanese, Bulgarian, French and Dutch, Czech, Polish, Russian, Armenian, Turkish, Arabic, Chinese, Indonesian) (for a presentation of these projects see Žabokrtský, 2005).

The procedure for creating valence dictionaries differs from one project to another: some are created entirely manually (especially the paper ones, but also VALLEX, Žabokrtský, 2005, which was created for the most frequent verbs in Prague Dependency Treebank (PDT) and is XML-encoded; the valence frames are created for each sense of the target verbs relying on the annotations in PDT), others in a semiautomatic way (the Japanese-English valence dictionary).

Both the Polish syntactico-semantic lexicon (Przepiórkowski, 2004) and the Bulgarian valence dictionary in electronic form (Balabanova & Ivanova, 2002) use Head-driven Phrase Structure Grammar (HPSG) (Pollard & Sag, 1987, 1994; Sag & Wasow, 1999) for representing grammatical knowledge.

There have been attempts to extract subcategorization frames (SF) for verbs from corpora, using machine learning techniques: Manning (1993), Briscoe & Carroll (1997), Carroll et al. (1998), Sarkar & Zeman (2000), Maragoudakis et al. (2001). Except for the last paper, the other ones make use of syntactically parsed corpora. Maragoudakis et al. use Bayesian Belief Networks and support vector machines to learn SF from corpora. All these approaches do not distinguish between various senses of verbs. Korhonen & Preiss (2003) suggest a way to improve the automatic acquisition of SF from corpora using a word sense disambiguation system. Bazili et al. (1999) use conceptual clustering for learning verb SFs. Salgueiro et al. (2005) present an unsupervised method for learning verb argument structures from corpora, making use of POS and named entity tagging.

3. Devices Involved in the Linguistic Representation in HPSG

HPSG offers one of the most comprehensive frames of encoding various types of linguistic information for lexical items. Apart from phonological information – which does not play a particularly important role in our lexical representations – one may easily represent morphological, syntactic and semantic properties of words.

HPSG makes use of the language of feature structures as a general device for encoding linguistic information. Minimally, a feature structure is an attribute-value pair, in which the attribute is expressed by capitals. The attribute has a value which is symbolized by italics. The value is called *a type* (or a sort). For instance, given the attribute CASE, one of its possible values is the type nominative (*nom*). The statement that a certain linguistic item bears case nominative is therefore expressed as follows: [CASE: *nom*]. This is a representation of a feature structure and it is called an attribute-value matrix (AVM). Representing syntactic, semantic and morphological properties of lexical items amounts to use AVMs in which relevant information is displayed.

3.1. Elements Involved in the HPSG Lexical Representation

Attributes involved in representing subcategorization properties of verbs are HEAD, ARG-ST (argument structure) and a part of the content representation, namely the one dealing with semantic roles.

The HEAD feature indicates the part of speech to which a given lexical item belongs, as well as certain morphological properties. We are interested in those HEAD properties that are relevant for subcategorization and argument alternations (if any). These properties may be captured by means of the features VFORM (verb form) AGR (agreement) and CLTS (clitics).

In Romanian, as in other languages, impersonal meaning of verbs is expressed by finite forms in the third person singular: *Eu dorm/ El doarme/ Se doarme* ("I sleep/ He sleeps/ One sleeps"). This information is to be encoded in the feature AGR:

$$\left[\text{HEAD}:\textit{verb}\left[\text{VFORM}:\textit{fin}\left[\text{AGR}:\begin{bmatrix}\text{PERS}:3\textit{rd}\\ \text{NB}:\textit{sg}\end{bmatrix}\right]\right]\right]$$

On the other hand, the same impersonal meaning is sometimes pointed out by means of the so-called clitic morphology. That is, the verb incorporates an element which is not a word, but an affix-like item. It is the case of the pronoun *se* in Romanian, in the example above, where the impersonal meaning is also marked through this clitic morphology: *Se doarme* ("One sleeps"). To encode it one employs the feature CLTS:

$$\left[\text{HEAD}:\textit{verb}\left[\text{VFORM}:\textit{fin}\begin{bmatrix}\text{AGR}:\begin{bmatrix}\text{PERS}:3\textit{rd}\\ \text{NB}:\textit{sg}\end{bmatrix}\\ \text{CLTS}:\textit{se}\end{bmatrix}\right]\right]$$

The feature ARG-ST accounts for those elements which are selected by the verb. Its value is therefore a list of dependents.

We make a commonplace distinction between *arguments* and *adjuncts*. The distinction may be easily illustrated in the case of the verb *to sleep*: a noun phrase (NP) in nominative like *John* is a required dependent and hence an argument. A prepositional phrase (PP) like *in the afternoon*, on the other hand, is a non-required dependent, hence it is an adjunct: *John sleeps in the afternoon*. According to this distinction, we keep the NP in nominative in the list of the elements defining the arguments of the verb *to sleep*, but we ignore dependents like the PP above. In the AVM below, angle brackets denote a list, that is, a set on which a certain relation of order has been defined. In the case of arguments, the order relation is obliqueness. This order somewhat reflects a certain prominence (or importance) of arguments.

$$\left[\text{ARG}-\text{ST}:\left\langle\text{NP}\left[\text{CASE}:\textit{nom}\right]\right\rangle\right]$$

Semantic roles express an important part of the meaning of a lexical item in its relation to its arguments. Semantic roles denote types of participants specific to the situation denoted by the verb. If the verb denotes for instance the state of sleeping, the situation denoted by it will necessarily make reference to someone who 'experiences' this state. This will therefore be the proper semantic role associated with the situation of sleeping.

In HPSG, semantic roles are currently represented by attributes such as AGENT, PATIENT, THEME, and so on. Recent works (Davis, 2001) minimize the set of these attributes. Our representations slightly depart from this procedure in the following respect: we use particular attributes for each semantic role associated with a given verb. For instance, in the case of the verb *to arrive at*, we indicate one of the semantic roles (the ‘first one’) by means of the attribute HE WHO/WHAT ARRIVES AT and not by means of the well-known attribute AGENT. Likewise, in the case of the other semantic role, we prefer the attribute THE PLACE TO ARRIVE AT and not the standardized semantic role PATH.

From a theoretical point of view, nothing special hinges on this option. In fact, we do not reject the hypothesis that the set of semantic roles might be reduced to a conveniently small number of members. What we get by adopting this way of representing semantic roles is just a more intuitive representation of the meaning of each verb. At the same time, this notation allows us to avoid current puzzles specific to the ‘general’ representation of semantic roles. For instance, we avoid saying whether a certain participant is the theme or the patient.

With these elements, the semantic part in the lexical representation of a verb like for example *to sleep* looks as follows:

$$\left[\mathrm{CONT}: cont\left[\mathrm{REL}: sleep - rel\left[\mathrm{SLEEPER}: \left|1\right| cont\right]\right]\right]$$

In the representation above, CONT stands for ‘content’. This attribute has as value the type *cont*. This type labels a feature structure which is richer than the representation above (other details have been ignored). The central element in the feature structure labeled *cont* is the relation (REL) *sleep* (which resembles a predicate in logic). This relation is defined by the fact that it necessarily requires a participant called SLEEPER. The sleeper is identified with the referent (in the representation above, the content) of one of the dependents in the ARG-ST list of the verb *to sleep*.

4. First Steps

With the elements presented above, the first step in our research was to describe a minimal corpus of ten verbs. The selection of the verbs has been determined by several criteria: membership to the core vocabulary of Romanian, frequency of occurrence in communication, and ambiguity. We chose the following ten verbs: *a ajunge* (to arrive), *a avea* (to have), *a creşte* (to grow), *a da* (to give), *a face* (to do), *a fi* (to be), *a lua* (to take), *a pune* (to put), *a sta* (to stay), *a veni* (to come).

We agreed that the format of the description for each verb in the dictionary must be as follows:

(a) A relevant context for a given meaning
(b) A description of the argument structure (ARG-ST)
(c) A description of the HEAD features (if necessary)
(d) A description of the semantic roles (SEM ROLES)
(e) A description of the argument alternation (if any)
(f) A list of multiword expressions incorporating the word in question.

We give below a sample of description for three meanings (out of seven) of the verb *a ajunge* (to arrive):

A AJUNGE $_1$

- CONTEXTS:

Trenul a ajuns (în gară/acolo) “The train arrived (at the station/there).”
Cine pleacă de dimineaţă ajunge departe. “He who leaves in the morning arrives far away.”
Am ajuns unde am vrut. “I arrived where I wanted to.”

- ARG-ST:

NP[CASE: *nom*] ∨ S[REL-DTR: **cine**];
(AvP[CONT: *loc*]) ∨ (PP[PForm: **la, pe, în**]) ∨ S[REL-DTR: **unde**]

- SEM-ROLES:

He who arrives at
The place to arrive at
ARGUMENT ALTERNANCE of *A AJUNGE$_1$* (Impersonal lexical rule): impersonal reflexive

- CONTEXTS:

Se ajunge greu (pe creastă). “One reaches hard (the ridge).”

- HEAD features

VFORM | AGR: 3rd person singular
[CLTS: |1| *reflexive* [CASE: *acc*]]

- ARG-ST:

NP[CLTS:| 1| *reflexive* [CASE: *acc*]]
(AvP[CONT: *loc*]) ∨ (PP[PForm: **la, pe, în**]) ∨S[REL-DTR: **unde**]

A AJUNGE $_2$

- CONTEXTS:

Ion s-a ajuns. “John got rich.”
Cine a fost strângător s-a ajuns. “He who saved got rich.”

- HEAD features

[CLTS: *reflexive* [CASE: *acc*]]

- ARG-ST:

NP$\begin{bmatrix} CASE: nom \\ CONT: human \end{bmatrix}$ ∨ S[REL-DTR: **cine**]

- SEM-ROLE:

He who becomes rich

A AJUNGE$_3$

- CONTEXTS:

(Îmi) ajunge ce văd/cît cîştig/ cine mă iubeşte/cum mi se vorbeşte/că sunt consolat. “It’s enough (for me) what I see/what I earn/who loves me/how they speak to me/that I am comforted.”
Nu (i)-au mai ajuns banii. “The money was not enough for him.”

- HEAD features

VFORM | [AGR: 3rd person singular]

- ARG-ST:

NP[CASE: *nom*] ∨ S[REL-DTR: **cine, ce, cît, cum**] ∨ S[MARKER-DTR: **că**];
(NP[CASE: *dat*])

- SEM-ROLES:

What is enough (what does suffice)
The person for whom something is enough

MULTIWORD EXPRESSIONS

A ajunge departe, a ajunge bine (a reuşi) (to succeed)
A ajunge rău (a decădea) (to decay)
A ajunge la mal (a finaliza cu bine o acțiune dificilă) (to end well a difficult action)
A-i ajunge cuiva cuțitul la os (a fi într-o situație disperată) (to be in a desperate situation)
A-l ajunge zilele (a îmbătrâni) (to get old)
Etc.

5. XML Format

We have chosen the XML format for encoding our dictionary because XML is the most appropriate mark-up language for describing information structured in HPSG framework. It is very popular due to its wide spread technology. It also supports Unicode encoding that is necessary for Romanian diacritics.

In XML the structuring of information is done by means of elements, which can be sequential (for defining one-level information) or imbricate (for multi-level information). An element usually consists of a pair of tags surrounding the content of the element. Feature structures in HPSG can be easily represented in terms of XML elements. Tags can be assimilated with HPSG attributes and the content with values of attributes. For instance, the argument structure indicating a subject expressed by a noun phrase, has the following representations in HPSG: [ARG-ST: <NP[CASE: *nom*]>] and, the corresponding one, in XML: <ARG-ST><NP><CASE>nom</CASE> </NP></ARG-ST>.

Besides the full compatibility between XML and HPSG representations, the XML technology offers the possibility of implementing different constraints related either to the linguistic theory adopted or to the proper work of building the electronic dictionary by humans.

For building the electronic form of the dictionary, we intend to follow the XML representation used by the team of the BulTreeBank Project with CLaRK System (Balabanova & Ivanova, 2002).

6. Conclusions and Future Work

Up to this stage, the valences descriptions of the verbs were done by hand, especially relying on the information in printed dictionaries of Romanian. The lexical richness of these verbs has offered us a large inventory of semantic roles, grammatical categories and other useful information about arguments. We intend to filter out the most general aspects from this inventory in order to apply the result for automatically extracting the argument structures of the other verbs. For instance, after getting a comprehensive list of possible grammatical categories concerning arguments, for each verb, one can search for the words corresponding to one of the inventoried categories, in the concordance frames of that verb, automatically extracted from our annotated corpus. The most frequent categories found are likely to express arguments of the verb. Further on, human experts will distinguish between true and false arguments and verb meanings. Besides, because concordance frames give access both to grammatical categories and words, it is not difficult for humans to see whether words have some common semantic characteristics.

The methodology presented here can be further on used to develop valence frames for the other Romanian verbs and (with possible amendments) for other grammatical categories expressing predicates (nouns, adjectives, prepositions).

7. Acknowledgements

The work presented here is funded by CNCSIS research grant A-1156.

8. References

Balabanova, E. & Ivanova, K. (2002). Creating a Machine-readable Version of Bulgarian Valence Dictionary. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria.

Basili, R., Pazienza, M. T., Vindigni, M. (1999). Lexical Learning for Improving Syntactic Analysis. In *Proceedings of the ACAI'99*, Workshop on Machine Learning in Human Language Technology, Chanai, Greece.

Briscoe, E. & Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing,* Washington, DC., pp. 356-363.

Carroll, J., Briscoe, E., Sanfilippo, A. (1998). Parser evaluation: A survey and a new proposal. In *Proceedings of the International Conference on Language, Resources and Evaluation*, pp. 447-454.

Davis, A. R. (2001). *Linking by Types in the Hierarchical Lexicon*, CSLI Publications, Stanford.

Fillmore. C.J. (2002). FrameNet and the Linking between Syntactic and Semantic Relations. In Tseng, S.-C. (Ed.) *Proceeding of COLING 2002*, pages xxviii-xxxvi. Howard International House.

Korhonen, A. & Preiss. J. (2003). Improving Subcategorization Acquisition using Word Sense Disambiguation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, pp. 48-55.

Manning, C. D. (1993). Automatic Acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st ACL*, pp. 235-242.

Maragoudakis, M., Kermanidis, K., Fakotakis, N., Gokkinakis, G. (2001). Learning Automatic Acquisition of Subcategorization Frames using Bayesian Inference and Support Vector Machines. ICDM '01, The 2001 IEEE International Conference on Data Mining, San Jose, pp. 623-625.

Pollard, C & Sag, I.A. (1987). *Information-based Syntax and Semantics*, CSLI, Stanford, California.

Pollard, C & Sag, I.A. (1994). *Head-driven Phrase Structure Grammar*, Chicago University Press/CSLI Publications, Chicago, IL.

Przepiórkowski, A. (2004). Towards the Design of a Syntactico-Semantic Lexicon for Polish. In *Proceedings of New Trends in Intelligent Information Processing and Web Mining, Zakopane*, Springer Verlag.

Sag, I. A & Wasow, Th. (1999). *Syntactic Theory: A Formal Introduction*, CSLI, Stanford California.

Salgueiro Pardo, T. A., Marcu, D., das Graças Volpe Nunes, M. (2005). Unsupervised Learning of Verb Argument Structures. In *Proceedings of CICLing 2006*.

Sarkar, A. & Zeman, D. (2000). Automatic Extraction of Subcategorization Frames for Czech. In *19th International Conference on Computational Linguistics*, pp. 691-697.

Žabokrtský, Z. (2005). *Valency Lexicon of Czech Verbs*. PhD Thesis, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague.