Using the Constructionist Approach when Building a Multilingual Valency Lexicon

J. Šindlerová

Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic.

Abstract. The paper presents a preliminary research in the area of verbal valency and argument structure theory. With the perspective of building a multilingual archive of valency characteristics of verbs, the question is raised whether the structure of such a linguistic resource should be straight and simple, or to some extent hierarchical and capturing more relation types, including those among individual frames within a single language.

Introduction

The issue of verbal argument structure has been in the center of linguistic interest for decades. Very soon, it became clear that this verbal property is to a great extent language universal. As *Goldberg* (1995) states it: "It is quite possible that there is a universal inventory of possible argument structure constructions relating form and meaning, and that particular languages make use of a particular subset of this inventory." According to her, the crosslinguistic research is crucial for the understanding of the universal nature of the phenomenon.

The research presented in this paper is strongly connected to a larger project of building a multilingual valency lexicon for the purposes of machine translation experiments to be carried on (in the first step) on the Prague Czech-English Dependency Treebank (see e.g. *Šindlerová et al.*, 2007). The multilingual valency lexicon to be built within the next few years should attest the influence of the valency information on the MT results in contrast to the experiments lacking this kind of information (see e.g. *Bojar and Hajič*, 2008 or *Bojar and Prokopová*, 2006).

Resources

As the lexical resources for the first stage of the project we use two already existing valency lexicons, the PDT-VALLEX (Czech verbs) and EngValLex (English verbs).

PDT-VALLEX (see e.g. *Hajič et al.*, 2003) has been created as a supplemental resource for the manual annotation of the Prague Dependency Treebank (PDT). By now, it contains 9191 valency frames for 5510 verbs, besides storing also the valency characteristics of some verbal nouns, adjectives and adverbs. The verbs and frames come almost exclusively from the data appearing in the PDT, version 2.0.

EngValLex (see e.g. *Cinková et al.*, 2006) is a sister resource to PDT-VALLEX, created by a manual conversion of the PropBank lexicon (see e.g. *Palmer et al.*, 2005) into the Functional Generative Description valency format (see e.g. *Lopatková and Panevová*, 2006). This conversion basically included re-labeling of the arguments, obligatoriness marking and unification of the resulting frames with identical meaning, sometimes even the introduction of new verbal lexemes and frames. It only contains verbs so far. Currently, it covers 6006 valency frames for 3576 verbs.

Both the above mentioned "vallexes" contain the same straight and simple xml structure:

```
<word>
  <valency_frames>
    <frame>
    </frame>
  </valency_frames>
</word>
```

Additonally, the structure of Engvallex already contains linkage to the PropBank lexicon. In the planned multilingual edition of the vallexes, this linkage is to be replaced by a similar linkage pattern between the Czech and English lexicons.

The Conception of a Multilingual Valency Lexicon

The intended multilingual valency lexicon is meant to represent a valuable resource which could be used both for a theoretical linguistic research and for computer science applications. Therefore the question of its structuring must be considered carefully, so that the rich linguistic information and linking it is supposed to contain would, at the end, stay fully available to the user.

The overall idea is that the new enriched and interlinked vallex should contain:

1. cross-language translational links between the corresponding frames;
2. cross-language links between the corresponding participants (frame elements);
3. information on an alternational status of the frame and
4. information on the verb class of the given verb sense/frame.

To improve the lucidity of the structure, we plan to implement a more complex hierarchical structure of the entry. So far, the entries in both PDT-VALLEX and EngValLex are basically unstructured, displaying a mere list of possible valency frames (there is just a slight tendency to organize the list linearly according to the complexity of the frame in PDT-VALLEX). This makes the reading and sorting of the linguistic information slightly difficult for a human user. As it was discussed by *Goldberg* (1995) or *Levin* (1993) previously, it is not always the case that each separate frame denotes a separate verb meaning. The tighter relations between some of the frames available, basically on the basis of a simple semantic derivation, are not indentifiable in the existing vallexes yet. It can be also assumed that capturing these relations will have a positive impact on the search for translational equivalents.

There have already been many attempts to approach the argument structure phenomena in a hierarchical manner, the most inspiring for our project being the Construction Grammar treatment of argument structure (see above all *Goldberg*, 1995) and the FrameNet hierarchies for the so called *semantic frames*, representing the hierarchical solution for the treatment of semantic verb classes (e.g. *Baker*, 2008).

Basic Concepts in the Theory of Argument Structure

The theory of Argument Structure is an alternative approach to valency developed in the generative framework. During the last twenty years several interesting observations were made by the theoreticians of this approach which we would like to make use of when restructuring and reorganizing the available resources into a single, interconnected valency lexicon.

The basic assumption shared by all approaches to valency structure of a verb is that each verb of a language has a limited list of possible valency frames from which a speaker must choose when producing a particular utterance. These frames can be successfully and (perhaps) exhaustively listed in a lexicon.

Individual valency frames have been first thought to identify individual senses of the verb. Later it became clear that there can be metaphorical or synonymical ties between some of them. That means that frames do not create just an unordered list, but rather a sort of net. The notion of *(diathesis) alternation* has been introduced and developed especially by *Levin* (1993) to account for the specific semantic type of relation of the tied frames. Thus, in cases like Substance/Source Alternation or Spray/Load Alternation, where the situation described by the verb involves an identical number and character of participants and only the position of

the participants in the frames is reorganized according to a different point of view, we do not perceive a complete change in the verb sense, but only in the organization of the perspective, resulting in the frame divergence (and possibly influencing some of the minor truth-conditions of the sentence uttered).

An interesting novelty in the theory is the notion of *constructions*. According to *Goldberg* (1995) (citing works of Fillmore et al.), constructions are basic units of language and they are perceived as form-meaning correspondences. It is believed that "simple clause constructions are associated directly with semantic structures which reflect scenes basic to human experience" (*Goldberg*, 1995, p.5). In short, it has been observed that there is a limited number of formal appearances of the frames available, that is a limited number of syntactic patterns the individual semantic frames are mapped to, such as *a ditransitive pattern*, *a pattern with a prepositional phrase* etc. As a consequence, we can basically generalize over the frames mapping to the same syntactic pattern to get a more abstract linguistic notion, the construction. As an example of what a construction looks like we can take the Ditrasitive Construction, which takes the form of Subj V Obj Obj_2 and covers the very general semantic concept of X CAUSES Y TO RECEIVE Z.

An interesting result of such an abstraction is that verbs mapping their frames to the same constructions usually form semantically close clusters. Thus, we can gain an alternative approach to verb classes as well.

In short, constructions are a highly abstract syntactic concept generalizing over a set of frames expressing their meanings by similar syntactic patterns. The verb classes then are formed as clusters of those verbs which belong to the same semantic field and whose valency frames are licensed by the same construction.

Figure 1 shows how approximately the constructionist system of argument structure works. We can see the links from the individual constructions to the individual verb frames, which capture the mapping relations. A given construction assigns a particular syntactic form to the frame. But the constructions themselves are not just scattered in the space, they rather form a separate hierarchy. They are linked by several types of inheritance relations, they take various properties from one another and they can also be tied by means of metaphorical extensions. The individual frames can be linked to one another as well, e.g. by means of metaphorical transfer (as in case of the verb *kick*) or by the relation of a diathesis alternation (as in case of the verb *send*). Verbs which inherit their frames from the same constructions and whose meanings come from the same semantic field can be classed together into the same verb class (*push* and *kick*).

The interesting thing about this scheme is that the licensing effect of the construction does not necessarily affect the meaning of the verb in the particular frame use. Though the verbs in sentences like *He kicked the dog*, *He kicked the ball into the goal* and *He kicked the door open* fall each into a different frame, the rough meaning of the verb itself, *to hit something with your foot*, has not changed dramatically.

Taking into account the mentioned facts, a preliminary hierarchical scheme of a valency lexicon entry can be proposed:

```
<verb>
  <verb_sense>
    <construction>
      <frame></frame>
      <metaphor></metaphor>
      <metaphor></metaphor>
    </construction>
  </verb_sense>
</verb>
```

Both the verb_sense and the construction tags are repeatable in the entry. The frame tag represents the most central and basic frame representation of the given construction, and other

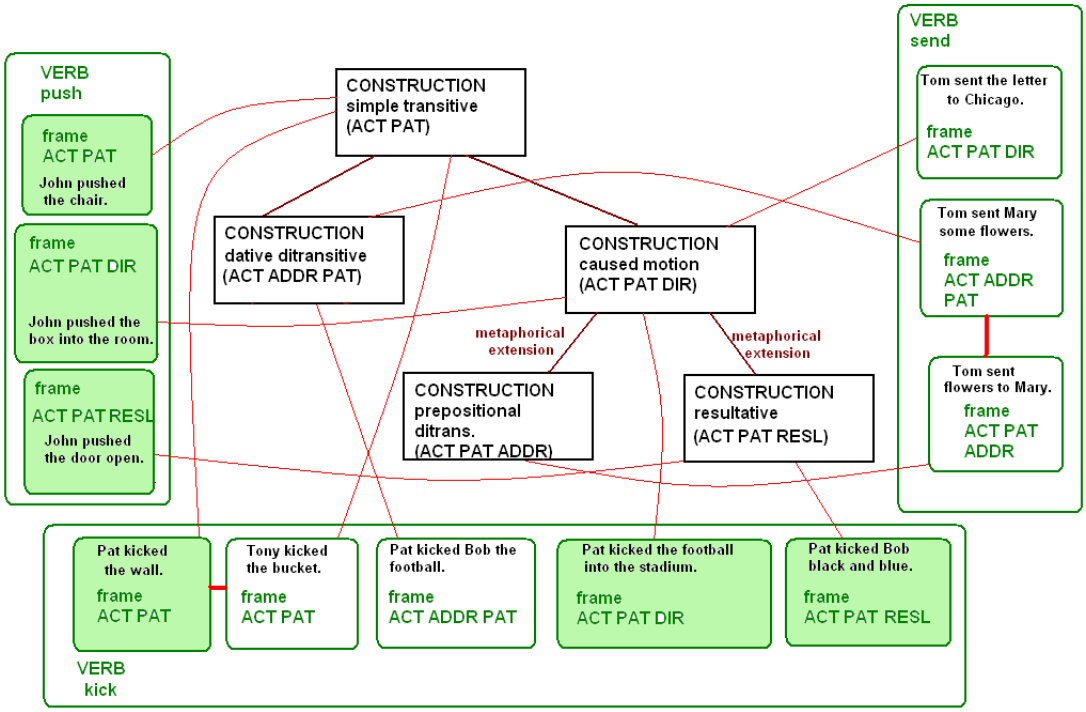

Figure 1. An illustration of the constructionst scheme of argument structure links and hierarchies.

frames within the same construction are marked as its metaphorical derivations.

Within such a scheme, we will be able to store frames related to the same verb sense together. Let us now show how the scheme works on the example of the verb DROP (mind that the further proposed structure of (a part of) the entry is only preliminary and still very rough, invented for the purposes of this paper). The structure proposed here contains three basic verb senses, one depicting the act of *something falling* including the meaning of a non-vertical movement, one depicting the act of *something stopping* and one depicting the act of *something disappearing*. Note that in contrast to a simple list of possible valency frames we have stored here more semantic information: first we have saved the semantic relatedness between the causative and non-causative frames and second we have saved the ties between the metaphorical derivations of the physical and abstract falling.

DROP
verb sense: falling
  construction: causative (X makes Y to fall (from Z to W))
    frame: ACT PAT ?DIR1 ?DIR3
      example: The children were dropping stones off the bridge.
    metaphor: ACT PAT ?DIR3
      example: He dropped his mouth/eyes/head into his hands etc.
    metaphor: ACT PAT ?DIFF ?ORIG ?EFF
      example: The company dropped the prices to $12 per bottle.
  construction: intransitive (Y falls (from Z to W))
    frame: ACT ?DIR1 ?DIR3
      example: The coins dropped into my hand.
    metaphor: ACT PAT DIR3
      example: He dropped into the chair.

metaphor: ACT ?DIFF ?ORIG ?PAT
example: The prices dropped to $12 per bottle.
construction: caused motion (X causes the motion of Y to Z)
frame: ACT PAT DIR3
example: Can I drop you somewhere?
verb sense: stopping
construction: transitive (X stops (doing) Y)
frame: ACT PAT
example: He has dropped all the actions. She dropped history in 1998.
verb sense: disappearing
construction: causative (X makes Y to disappear (from Z))
frame: ACT PAT ?DIR1
example: They dropped the verb from the lexicon. She is dropping her h's.
etc.

There are some theoretical points which can turn out problematic in the future application of the scheme to the real data. First of all it is the question of alternations, that is where to store this kind of information in the structure of the entry. It appears that it will be lower than the verb_sense, nevertheless, it is not clear yet, whether it will be possible to store the information by means of a separate tag, or whether some indexing will be more suitable. We let this question open for the time being.

Moreover, two other things will have to be decided before the annotation process begins. First, we will have to determine (at least very roughly) the boundaries of the metaphor, i.e. the point where a mere metaphorical extension turns into a different verb sense. Second, we will have to revise the traditional Levin alternations and decide which of them to employ in our scheme.

Approaches to Verb Class Assignment

There are two influential approaches to verb semantic classes in the current linguistics we would like to draw on in our project.

Beth Levin (see *Levin*, 1993) has based her observations concerning the syntactic and semantic behavior of verbs on the likelihood of a verb appearing in a particular alternation. Her classification is roughly hierarchized (there are three levels od specificity at the maximum, e.g. Verbs of Motion - Manner of Motion Verbs - Roll Verbs).

FrameNet (see e.g. *Baker*, 2008) hierarchy of *semantic frames* is based on the character of frame elements. It is more fine-grained and captures relations between the individual classes.

Standing at the beginning of the project, we have two reasons why to prefer the FrameNet as the primary source of methodology. As the constructionist approach mentions, Levin's assumption that verbs of the same class will have the same frames/alternations is basically not fully correct. Goldberg shows that at least some of the constructions do not operate on the whole of a semantic class, they can be only partially productive. And second, there is currently a project trying to map Framenet categories to a valency lexicon made within the Praguian formalism, namely to VALLEX 2.5, having already very interesting and inspiring results on the subclass of communication verbs (*Benešová et al.*, 2008).

Conclusion

In our future work, we would like to build a multilingual, enriched and interlinked lexicon of valency characteristics. We plan to start with the Czech and English verbs, making use of the available valency lexicons PDT-VALLEX and EngValLex.

In contrast to both the already existing sources, the multilingual vallex should contain

more information on the relations between the individual frames, such as information about their alternations and metaphorical extensions, and its entries should be finer-structured.

The multilingual valency lexicon should be of use both to a human reader – linguist – and in MT applications. We expect to gain not only a valuable source of linguistic information about the crosslinguistic similarities and differences in verb behavior, particularly about the distribution of verbs into classes, about the number and character of participants and about the structure of the frame hierarchy, but also a resource for attesting the idea that a lexicon information on the verbal valency structure could improve the results of MT experiments carried on the tectogrammatical layer of Prague dependency corpora.

On the basis of the Construction Grammar approach findings about the nature of valency frames relations, we have proposed here a simple hierarchy for the lexicon entry. The usability of the hierarchy must be still proven on the data.

In the following process we will face several decisions about the extent of the information captured within the entry and about the way of its classification. Though it is necessary to make these theoretical decisions in advance, we are aware of the fact that it is always the real data which are to confirm the propriety of the theoretical approach, which stand at the very core of the project and which will finally bring the most interesting results.

Acknowledgments. The present work was supported by the Charles University Grant Agency under Contracts 19008/2008 and 52408/2008.

References

Baker, C., FrameNet, Present and Future, in: *Proceedings of The First International Conference on Global Interoperability for Language Resources*, Hong Kong, 12-17, 2008.

Bojar, O. and Prokopová, M., Czech-English Word Alignment, in: *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 1236-1239, 2006.

Bojar, O. and Hajič, J., Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation, in: *Proceedings of the Third Workshop on Statistical Machine Translation*, 143-146, 2008.

Benešová, V., Lopatková, M. and Hrstková, K., Enhancing Czech Valency Lexicon with Semantic Information from Framenet: The Case of Communication Verbs, in: *Proceedings of The First International Conference on Global Interoperability for Language Resources*, Hong Kong, 18-25, 2008.

Cinková, S. and Semecký, J., Constructing an English Valency Lexicon, in: *Proceedings of Frontiers in Linguistically Annotated Corpora*, Sydney, 111-113, 2006.

Goldberg, A., *A Construction Grammar Approach to Argument Structure*, The University of Chicago Press, Chicago, 1995.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., and Pajas, P. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation, in: *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, Vaxjo, Sweden, 57-68, 2003.

Levin, B., *English Verb Classes and Alternations: A Preliminary Investigation*, The University of Chicago Press, Chicago, 1993.

Lopatková, M. and Panevová, J., Recent Developments in the Theory of Valency in the Light of the Prague Dependency Treebank, in: *Insight into Slovak and Czech Corpus Linguistic*, Bratislava, Slovakia, 83-92, 2006.

Palmer, M., Gildea, D. and Kingsbury, P., The Proposition Bank: An Annotated Corpus of Semantic Roles in: *Computational Linguistics*, 31(1), 71-106, 2005.

Šindlerová, J., Mladová, L., Toman, J. and Cinková, S., An Application of the PDT-scheme to a Parallel Treebank, in: *Proceedings of the 6th International Workshop on Treebanks and Linguistic Theories (TLT 2007)*, Bergen, Norway, 163-174, 2007.