# Towards a Semantically Motivated Organization of a Valency Lexicon for NLP: The GREG-Proposal

**Leo Wanner, Stefan Klatt, Oleg Kapanadze, Nunu Kapanadze**
Computer Science Department
University of Stuttgart, Breitwiesenstr. 20-22
70565 Stuttgart, Germany
{wanner | klatt}@informatik.uni-stuttgart.de | ok@iatp.org.ge

**Abstract**
The regularity of valency information across entries for different lexical units calls for its generalization. In this paper, we present the results of our research on the generalization of syntactic and semantic valency patterns in German for a mid-size multilingual verbal valency NLP lexicon. We suggest semantic field-oriented multiple inheritance hierarchies. These hierarchies capture fine-grained differentiations between subcategorization frames and allow for a direct relation between case frames and corresponding subcategorization frames. For the implementation, we use the DATR-formalism.

## 1 Introduction

The regularity of valency patterns across different (semantically related) lexical units (LUs) calls for their non-redundant representation in the lexicon. The research on this issue reported on in this paper has been carried out within the GREG Project. The goal of GREG was to develop a non-redundant and thus efficient lexical representation for multi-lingual valency lexica for NLP and to demonstrate the adequacy of this representation by implementing a mid-size verb lexicon for Georgian, Russian, English, and German. Non-redundant representation of lexical representation means, as a rule, construction of an inheritance hierarchy in which information common to several items is extracted and placed higher in the hierarchy so as to be inherited to all items that possess this information. If an item inherits information that is not compatible with its patterns, this information is overriden; if an item possesses local information, this information is added to the information inherited. However, as is well known, detailed valency patterns are highly language-specific. This suggests that an optimal *multilingual* representation will not necessarily be optimal with respect to the individual languages involved - especially when these languages belong to different language families as Georgian, Russian and English/German do. Therefore, we decided to pursue two strands of research in GREG. The first strand was on the lexical representation formalism suitable for the representation of both common valency patterns across languages and language-specific valency patterns. The second strand was on the representation suitable for a to a maximal extent non-redundant encoding of valency patterns. Due to the above-mentioned observation that valency patterns tend to be language-specific, the first phase of the second strand of our research was dedicated to the investigation to what extent *monolingual* valency information can be generalized. Because for German a large number of fine-grained subcategorization information was available, German was chosen for this investigation. In the second phase of

the second strand of our research, the findings of the first phase were applied to the multilingual representation.
In this paper, we present the results of the first phase of the second strand of research in GREG. For the description of the first strand, see [Evans et al. forthcoming].

## 2 The German Material in GREG

The material in GREG has been compiled starting from the list of the 1000 most common verbs in Georgian. The English, German and Russian parts of the lexicon have been obtained by translating the Georgian originals and adding the most common 300 verbs of the respective language.

The valency information in GREG covers both syntactic valency (subcategorization frames) and semantic valency (case frames [Fillmore 1982], thematic [Jackendoff 1990] or functional roles [Halliday 1985; Chafe 1970]).
The German part of GREG contains about 1200 German verbs. The major part of the German subcategorization frames stem from the lexicon provided to us by the IMS, University of Stuttgart [Lezius et al. 2000]. Additional frames have been extracted from corpora using a valency extraction program [Wauschkuhn 1999] and some others were added manually. The case frames have been compiled manually. Consider, for illustration, a part of the valency information for BEZAUBERN '[to] charm' and LIEFERN '[to] deliver' in Figure 1.

BEZAUBERN '[to] charm'

ACTOR SENSER
ACT SENSER
$NP_{nom}$ $NP_{acc}$
ACTOR SENSER MEANS
$NP_{nom}$ $NP_{acc}$ PP [*mit* $NP_{dat}$]
$NP_{nom}$ $NP_{acc}$ PP [*durch* $NP_{acc}$]
ACT SENSER
*dass*-CLAUSE $NP_{acc}$
*zu*-$VP_{inf}$ $NP_{acc}$

LIEFERN '[to] deliver'

ACTOR OBJECT
$NP_{nom}$ $NP_{acc}$
ACTOR OBJECT SOURCE
$NP_{nom}$ $NP_{acc}$ PP [*aus* $NP_{dat}$]
$NP_{nom}$ $NP_{acc}$ PP [*von* $NP_{dat}$]
ACTOR OBJECT RECEIVER
$NP_{nom}$ $NP_{acc}$ PP [*zu* $NP_{dat}$]
$NP_{nom}$ $NP_{acc}$ PP [*an* $NP_{acc}$]
$NP_{nom}$ $NP_{acc}$ $NP_{dat}$
ACTOR OBJECT DESTINATION
$NP_{nom}$ $NP_{acc}$ PP [*nach* $NP_{dat}$]
$NP_{nom}$ $NP_{acc}$ PP [*in* $NP_{acc}$]

Figure 1: Partial specification of the valency information for BEZAUBERN and LIEFERN

The entry for BEZAUBERN indicates that BEZAUBERN possesses the subcategorizat-ion frame $NP_{nom}$ $NP_{acc}$, which corresponds to the case frames ACTOR SENSER; cf. (1a) below) and ACT SENSER (1b):

1. (a) *Maria hat Hans bezaubert*, lit. 'Maria has John charmed',
   (b) *Der Gesang der Kinder bezauberte Hans*
   lit. 'The singing of the children charmed John';

the subcategorization frames $NP_{nom}$ $NP_{acc}$ PP [*mit* $NP_{dat}$] and $NP_{nom}$ $NP_{acc}$ PP [*durch* $NP_{dat}$], which correspond to the case frame ACTOR SENSER MEANS; cf. (2a,b):

2. (a) *Maria bezauberte Hans mit ihrem Gesang.*
   lit.'Maria charmed John with her singing',
   (b) *Maria bezauberte Hans durch ihren Gesang*
   lit. 'Maria charmed Hans by her singing';

and the subcategorization frames *dass*-CLAUSE $NP_{acc}$ and *zu*-$VP_{inf}$ $NP_{acc}$, which correspond to the case frame ACT SENSER; cf. (3a,b):

3. (a) *Dass Hans Maria Blumen geschenkt hat, bezauberte sie*
   lit. 'That Hans Maria flowers gave as a present', charmed her',
   (b) *Ihren Sohn in dieser Position zu sehen, bezauberte Maria*
   lit. 'Her son in this position to see charmed Maria'.

The entry for LIEFERN is somewhat more complex; consider sentential examples for each of its subcategorization frames:

4. ACTOR OBJECT:
   (a) *Hans* [$NP_{nom}$] *lieferte die Waffen* [$NP_{acc}$] 'John delivered the arms',
5. ACTOR OBJECT SOURCE:
   (a) *Hans* [$NP_{nom}$] *lieferte die Waffen* [$NP_{acc}$] *aus Deutschland* [*aus* $NP_{dat}$]
   lit. 'John delivered the arms from Germany',
   (b) *Hans* [$NP_{nom}$] *lieferte die Waffen*[$NP_{acc}$] *von den Malediven* [*von* $NP_{dat}$],
   lit. 'John delivered the arms from the Maldives',
6. ACTOR OBJECT RECEIVER:
   (a) *Hans* [$NP_{nom}$] *lieferte die Waffen* [$NP_{acc}$] *zu den Rebellen* [*zu* $NP_{dat}$]
   lit. 'John delivered the arms to the rebels',
   (b) *Hans* [$NP_{nom}$] *lieferte die Waffen* [$NP_{acc}$] *an die Rebellen* [*an* $NP_{acc}$]
   lit. 'John delivered the arms at the rebels',
7. ACTOR OBJECT DESTINATION:
   (a) *Hans* [$NP_{nom}$] *lieferte die Waffen* [$NP_{acc}$] *nach Taschkent* [*nach* $NP_{dat}$]
   lit. 'John delivered the arms to Tashkent',
   (b) *Hans* [$NP_{nom}$] *lieferte die Waffen* [$NP_{acc}$] *in die Türkei* [*in* $NP_{acc}$]
   lit. 'John delivered the arms in Turkey'.

## 3 Representation of Valency Information

The problem of the representation of valency information must be considered from two angles: (i) the way LUs that share some or all valency patterns can be grouped together and (ii) the way valency information can be inherited.

### 3.1 Classification of Lexical Units with Common Valency Information

Syntactic valency tends to be dominant in valency lexica. This might suggest a syntactic classification in which all LUs that possess (a) common subcategorization frame(s) form one class. In German, for instance, $NP_{nom}$ $NP_{acc}$ would subsume BEZAUBERN '[to] charm',

GLAUBEN '[to] believe', ERREICHEN '[to] reach', FAHREN '[to] drive', and many others. However, this approach would lead to very flat hierarchies and, furthermore, not allow for a unified (and simultaneous) classification with respect to syntactic and semantic valency. This also applies to mixed classifications (see, e.g., the organization of the grammar in systemic linguistics [Matthiessen 1996] and the classification in [Kilgarriff 1993]) in which syntactic criteria such as transitivity provide a coarse-grained classification, which is then made more delicate by semantic criteria.
A case-frame based classification (similar to the Frame Semantics approach [Baker et al. 1998]), in which a case frame hierarchy forms the backbone of the classification seems more appropriate. However, this classification does not allow, e.g., for a grouping of several different case frames that are realized by a single subcat frame; cf. BEZAUBERN above where ACTOR SENSER and ACT SENSER both correspond to $NP_{nom}$ $NP_{acc}$. Therefore, we adopt yet another approach. In this approach, the verbal material is first grouped with respect to a number of semantic fields. Then, a syntactico-semantic classification is carried out for each of the fields. What this classification looks like is illustrated in the next section.

### 3.2 Generalization of Valency Information

Two approaches are possible for a generalized representation of valency information. In the first approach, individual valency patterns are inherited as a whole. For instance, in German, the subcategorization frame $NP_{nom}$ $NP_{acc}$ is inherited by FEIERN '[to] celebrate', KALKULIEREN '[to] calculate', VERWIRKLICHEN '[to] realize', etc., which belong to the same class in the mental field and to all other classes in the same field whose members possess this pattern.
In the second approach, parts of valency patterns rather than whole patterns are inherited. The inherited parts are then concatenated to a complete pattern. Thus, $NP_{nom}$ for the first actant is inherited by all classes of a field in question. The class of transitive verbs adds then the specification of the subcategorization information of the second actant, i.e., $NP_{acc}$. This approach is adopted, e.g., by Kilgarriff [1993]. However, while seemingly attractive because it reduces the redundancy of information, it turns out to be problematic in the case of a relatively large number of detailed valency patterns. For instance, Germ. VORAUSSEHEN '[to] foresee' inherits the subcategorization frame $NP_{nom}$ $NP_{acc}$: *Hans sah dieses Unglück voraus* lit. 'John foresaw this disaster'. But it also inherits $NP_{nom}$ *dass* CLAUSE: *Hans sah voraus, dass dieses Unglück passieren wird* lit. 'John foresaw that this disaster will happen'. In this case, an overriding of inherited information is required. Such conflicts occur on a regular basis. Therefore, we adopted the first approach.

## 4 German GREG-Hierarchies

For the formal representation of the valency information hierarchies, we use the DATR-formalism [Evans & Gazdar 1996]. However, due to the lack of space, we provide here merely the general picture of what our hierarchies look like. Figure 2 shows a fragment of the communication hierarchy. Except the root of the hierarchy, which is named, the classes in the hierarchy are, as a rule, numbered. This is done to avoid arbitrary and, given the degree of delicateness of our hierarchies, unavoidably opaque names. The valency patterns that belong to a given class are specified in brackets below the number of the class. As might be intuitively clear, $NP_{nom}$ stands for "NP in the nominative", PP $von_{dat}$ stands for "PP with the preposition *von*, which requires the governed NP to be in the dative", etc. PAV as in

PAV *von*, PAV *durch*$_{acc}$, PAV *über*$_{acc}$, ... stands for "pronominal adverb" obtained by a concatenation of *da-* with the preposition cited (cf. *davon*, *dadurch*, *darüber*, etc.). Note that the notation of subcategorization frame information used in the German part of GREG stems from the IMSLex.

```
1 communication
          1.0
             [SAYER]
              NP_nom
              1.0.1 ^1.1
              grüssen
          1.1
               NP_nom NP_acc
               aussprechen
               befehligen
               1.1.1
                   NP_nom NP_acc PP_von_dat
                   abberufen
                   1.1.1.1 ^1.1.6.1 ^1.11.1.1.1 ^1.1.7
                   NP_nom NP_acc PAV_von_dat C_dass
                        NP_nom NP_acc PAV_über_acc C_dass
                        NP_nom NP_acc PP_durch_acc PP_bezüglich_gen
                        NP_nom NP_acc PP_durch_acc PP_über_gen
                   benachrichtigen
             ...
             1.1.6
                   NP_nom NP_acc PP_über_acc
                   1.1.6.1
                         NP_nom NP_acc PAV_über_acc C_ob
                         NP_nom NP_acc PAV_über_acc C_wh
                         verhören
                   1.1.6.2
                         NP_nom NP_acc PAV_über_acc C_dass
             1.1.7
                   NP_nom NP_acc PP_bezüglich_gen
                   1.1.7.1 ^1.1.6.1
                         1.1.7.1.1 ^1.16
                               ausfragen
                         1.1.7.1.1 ^1.7 ^1.8
                               beraten
                ...
```

Figure 2: An excerpt of the German communication hierarchy in the GREG lexicon

If the valency patterns of one or several verbs are fully covered by a class, these verbs are listed below the patterns of the class. Number labels of classes which are listed after the first label and which are preceded by the '^'-sign indicate the mother classes of the first class. That is, `1.1.1.1 ^1.1.6.1 ^1.1.7 ^1.11.1.1.1` signals that the class `1.1.1.1` inherits the valency patterns from its immediate predecessor (i.e., `1.1.1`) and the classes `1.1.6.1`, `1.1.7` and `1.11.1.1.1`. As Figure 2 illustrates, the communcation valency hierarchy is a multiple inheritance hierarchy: a class can inherit valency patterns from several mother classes.

Figure 2 also illustrates the depth of the hierarchy and, thus, the potential of the generalization of valency information.

## 5 Conclusions and Future Work

The work in the GREG Project demonstrated that a semantic field bound generalization of valency information is advantageous in that it allows for a more detailed hierarchization of valency patterns with less conflicting cases in which an LU would inherit patterns that it does not possess. However, Figure 2 also shows that a considerable redundancy is encountered as far as case frames are concerned. Future work will address this critical aspect of the current structure of the lexicon – before the same schema will be adapted for the multilingual environment. With 1200 lemmata the German part of the GREG lexicon is still relatively small. For broad coverage NLP, a considerably larger lexicon is needed. Therefore, another important part of future work will consist in enlarging the lexicon.

## Acknowledgements

O. and N. Kapanadze's address is: Department of Applied Informatics, University of Tbilisi, Charchavadse av. 1, 380028 Tbilisi, Georgia. The work described in this paper has been funded by the EC under the contract number INTAS Georgia '97, 1921. The partners involved in the GREG project were: University of Stuttgart, University of Brighton, University of Tbilisi and Georgian Academy of Sciences.
We would like to thank Ulrich Heid for making the IMSLex available to us.

## References

[Baker et al. 1998] Baker, C.F., C.J. Fillmore & J.B. Lowe, 1998. The Berkeley FrameNet Project, in: *Proceedings of COLING/ACL 1998,* Montreal.

[Chafe 1970] Chafe, W.L., 1970. *Meaning and the structure of language.* University of Chicago Press, Chicago.

[Evans & Gazdar 1996] Evans, R. & G. Gazdar, 1996. DATR: A Language for Lexical Knowledge Representation, in: *Computational Linguistics*, 22(2), pp. 167-216.

[Evans et al. forthcoming] Evans, R. et al., forthcoming. *The GREG Framework for Multilingual Valency Lexicons*, Technical Report, ITRI, University of Brighton, Brighton, U.K.

[Fillmore 1982] Fillmore, C.J., 1982. Frame Semantics, in *Proceedings of the Conference Linguistics in the Morning Calm*, pp. 111-137. Hanshin Publishing Co., Seoul.

[Halliday 1985] Halliday, M.A.K., 1985. *Introduction to Functional Grammar.* Edward Arnold, London.

[Jackendoff 1990] Jackendoff, R., 1990. *Semantic Structures.* The MIT Press, Cambridge, MA.

[Kilgarriff 1993] Kilgarriff, A., 1993. Inheriting verb alternations, in: *Proceedings of the 6th European ACL Meeting*, pp. 213-221.

[Lezius et al. 2000] Lezius, W., S. Dipper & A. Fitschen, 2000. IMSLex – Representing morphological and syntactical information in a relational database, in: *Proceedings of the 9th EURALEX International Congress*, pp. 133-139. Stuttgart, Germany.

[Matthiessen 1996] Matthiessen, C., 1996. *Lexicogrammatical Cartography: English Systems.* International Language Sciences Publishers, Singapour.

[Wauschkuhn 1999] Wauschkuhn O., 1999. *Automatische Extraktion von Verbvalenzen aus deutschen Textkorpora.* Shaker Verlag, Aachen, Germany.