Using High-Quality Resources in NLP: The *Valency Dictionary of English* as a Resource for Left-Associative Grammars

Thomas Proisl, Besim Kabashi

Friedrich-Alexander-Universität Erlangen-Nürnberg
Department Germanistik und Komparatistik
Professur für Computerlinguistik
Bismarckstr. 6, 91054 Erlangen
{tsproisl,kabashi}@linguistik.uni-erlangen.de

Abstract

In Natural Language Processing (NLP), the quality of a system depends to a great extent on the quality of the linguistic resources it uses. Due to the unpredictable character of valency properties, a reliable source for information about valency is important for syntactic and semantic analysis. With this in mind, we discuss how the *Valency Dictionary of English* in machine-readable form can be used as a resource for NLP. We will show that the valency data can be integrated into a Left-Associative Grammar and thus can be used for accurately parsing natural language with a rule-based approach.

1. Introduction

Valency as a model for the description of language goes back to Tesnière (1959) and his dependency grammar. The main idea of valency theory is that certain words, most prominently verbs but also members of other word classes such as adjectives or nouns, have a determining influence on the syntactic structure of a sentence.[1] The verb *admire*, for example, requires a noun phrase or a wh-clause expressing the thing or person admired to form a syntactically and semantically well-formed sentence:[2]

(1) a. He was not one for *admiring* the sunset. [BNC: J10 1822]
b. I know that Jack is a master chairmaker and I *admire* what he does. [BNC: A0X 292]

Valency dictionaries give detailed descriptions of these phenomena for every single word – this is necessary because even semantically related or similar words can differ in their valency properties.[3] It is these detailed descriptions that make valency dictionaries such valuable resources for NLP purposes.

Among the most well-known valency resources are COMLEX (Grishman et al., 1994), VALEX (Korhonen et al., 2006) and FrameNet (Fillmore et al., 2003). In this paper we want to show how valency information from the *Valency Dictionary of English* (Herbst et al., 2004) can be integrated into a Left-Associative Grammar.

[1]The notion of valency is also called complementation or subcategorization in other theoretical frameworks.

[2]Examples marked with the subscript 'BNC' have been extracted from the British National Corpus (2007), distributed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved.

[3]The nouns *amateur* and *master*, for example, require specific prepositions to express the theme of amateur- or mastership. There is, however, no way to exactly predict which noun requires which preposition (*master at sth.* vs. *amateur in sth.*).

2. The *Valency Dictionary of English*

2.1. The Printed Dictionary

The *Valency Dictionary of English* (VDE) is "a highly specialized dictionary that attempts to provide a detailed description of valency" (Herbst et al., 2004, vii). It contains valency descriptions of 511 verbs, 274 nouns and 544 adjectives.[4] Each entry presents the syntactic valency of the lemma in the form of formal patterns illustrated by authentic examples drawn from the *Bank of English*. The verb *discuss*, for example, possesses a valency pattern + wh-$CL_{P(it)}$ that indicates that *discuss* can take a wh-clause as complement (2-a). The subscript P(it) indicates that the wh-clause can occur as subject of a finite passive clause (2-b) and that when occurring as subject, extraposition with a dummy subject *it* is possible (2-c).

(2) a. Then they *discussed* what they had in mind for the publicity. [BNC: CHE 1968]
b. Why this should be so will be *discussed* in a moment [...]. [BNC: A75 922]
c. It has been *discussed* whether this increase may also be related to the proliferative capacity of the cells [...]. [BNC: FTE 726]

VDE entries also indicate the senses of the valent uses of a lemma, provide a list of all complements and the patterns in which they occur, group semantically similar complements together and informally characterize these groups of semantically similar complements. Let us give an example to illustrate the point.

(3) If you are living in council property you must *discuss* adapting your house with the housing department. [BNC: A0J 1891]

Example (3) is an instance of another valency pattern of *discuss*, namely + N_P/V-ing_P + with N. The VDE

[4]The VDE was created at the universities of Erlangen, Reading and Augsburg. The descriptions are based on the *Bank of English* and have been checked and complemented by native speaker informants (Herbst et al., 2004, xxxix).

tells us that the subject (which is not indicated in the pattern) belongs to the group of complements labelled I. The passivizable noun phrase (N_P) and its alternative, the passivizable clause introduced by the ing-form of a verb (V-ing_P), belong to the group of complements labelled II. The prepositional phrase introduced by *with* followed by a noun phrase (with N) belongs to the group of complements labelled III. The VDE gives the following informal characterization of these groups of complements: "**A person^{I} can discuss a matter^{II} with another person^{III}**, i. e. talk about it."
The VDE contains only 1 329 entries, but as the majority of lemmata are high-frequent, they have a surprisingly high token coverage. The VDE covers 12.36% of all noun tokens in the BNC, 19.30% of all adjective tokens and 77.36% of all verb tokens. If the verbs *be, do, have* and *go* are excluded because of their grammatical functions which are not covered by the VDE entries, the VDE still covers 65.32% of all verb tokens in the BNC.

2.2. Derived Electronic Resources

The VDE "shows considerable affinity with NLP, even though it was not conceived with the use by automatic systems in mind" (Heid, 2007, 378). Consequently, efforts have been made to turn the VDE into an electronic resource that can also serve NLP purposes.
Heid (2007, 374–375) reports a positive effect of VDE data extracted by Spohr (2004) on syntactic analysis coverage of a Lexical Functional Grammar for English and concludes that the VDE can "indeed serve as valency dictionary for a formal grammar."
Electronic versions of the VDE commissioned by Mouton de Gruyter and worked on by Proisl (2008) are the basis for the latest electronic version, which is available online via the *Erlangen Valency Pattern Bank*[5] (Herbst and Uhrig, 2009). The *Pattern Bank* is aimed at linguists working in the area of valency and argument structure constructions. Due to licensing issues, it contains mainly the syntactic valency patterns and omits most of the other information from the dictionary.

3. Left-Associative Grammar

Left-Associative Grammar (LAG) is the formalism of the SLIM theory of language (Hausser, 2001).[6] In contrast to other grammar formalisms, LAG models natural language not by a hierarchy of substitutions but by a sequence of continuations.[7] To illustrate this in a simplified form, consider example sentence (4-a) which will be processed as indicated in (4-b).

(4) a. He opened the envelope. [BNC: GUF 2069]

b. ((((He + opened) + the) + envelope) + .)

In the first step, the first two word forms are concatenated by combination rules that make use of the lexical and grammatical categories of the word forms. The result contains the concatenated word forms and a so called rule package that indicates the possible continuations. Processing continues word form by word form until the last word form is reached. Formally, an LAG rule i can be represented as follows:

$$r_i\colon\ cat_1\ cat_2 \Rightarrow cat_3\ rp_i$$

The formalism of LAG is currently being used within Database Semantics (DBS) (Hausser, 2006) which is based on the SLIM theory of language.
LAG has proven adequate for modelling even complex phenomena of natural languages, e. g. the doubling or substitution of objects by pronominal clitics in Albanian (Kabashi, 2007).

4. Integration

We will illustrate the integration of the valency data available via the *Erlangen Valency Pattern Bank* into a Left-Associative Grammar by analyzing the following example:[8]

(5) The following analogy may prove helpful in understanding these statements. [BNC: HSE 176]

First, consider the relevant lexical information:

```
[sur: the
 cat: {(sn' snp) (pn' pnp) ...}]

[sur: following
 cat: {(adj)}  ]

[sur: analogy
 cat: {(sn) (for_npo' sn)
       (with_npo' sn) ...}]

[sur: may
 cat: {(nps' inf' v) ...}]

[sur: prove
 cat: {(adjp' inf)
       (npo' to_npo' inf) ...}]

[sur: helpful
 cat: {(to_npo' adj)
       (in_v-ing' adj) ...}]

[sur: in
 cat: {(npo' in_npo)
       (v-ing' in_v-ing) ...}]

[sur: understanding
 cat: {(npo' v-ing)
       (npo' as_npo' v-ing) ...}]

[sur: this
 cat: {(sn' snp) ...}]
```

[5] http://www.patternbank.uni-erlangen.de

[6] The letters of the acronym SLIM indicate the main principles of this theory of language: Surface Compositional Linear Internal Matching.

[7] According to the CoNSyx hypothesis put forward by Hausser (2001, 236), "[t]he natural languages are contained in the class of C1-languages and parse in linear time." The formalism of LAG is therefore regarded as particularly well-suited for NLP purposes.

[8] The use of resources in Natural Language Generation (NLG) is part of an ongoing project by B. Kabashi.

```
[sur: statement
 cat: {(sn) (of_npo' sn)
       (about_npo' sn) ...}]

[sur: .
 cat: {(ip)}]
```

The shortened[9] entries displayed here are feature structures containing the surface of a word form as a string and its grammatical categories as a set of tuples. The categories consist of segments indicating valency slots (marked with an apostrophe) and a segment expressing formal properties of the word form. The valency properties of *analogy, prove, helpful, understanding* and *statement* are based on information from the *Pattern Bank*,[10] the other entries were created by the authors.
In the first step, the determiner *the* (categorized, inter alia, as a singular noun phrase still needing a singular noun:[11] `(sn' snp)`) is combined with the adjective *following* by means of general linguistic rules. In the next step, the singular noun *analogy* fills the empty `sn'` valency slot of the determiner, thus eliminating the other category sequences of *the*. The next word form, the modal verb *may*, is categorized as needing both an NP in subjective case (`nps'`) and an infinitive (`inf'`). The category value `snp` is (by definition of variable restrictions) able to fill the valency slot `nps'` of the modal verb. The combination with the modal verb also rules out all the categorizations of *analogy* except the avalent one, `(sn)`.
In the next combination step, the lexical verb *prove* is able to fill the `inf'` slot of *may*. *Prove* possesses various valency patterns, inter alia one which demands an adjective phrase (`adjp'`). The variable restrictions allow that the next word, the adjective *helpful*, can fill this valency slot. All other category sequences of *prove* are ruled out by this combination step.
Helpful itself also possesses valency patterns. Indicated in the listing above are patterns demanding a prepositional phrase (PP) with *to* followed by an NP in objective case (`to_npo'`) or a PP with *in* followed by the ing-form of a verb (`in_v-ing'`). The next word form, the preposition *in*, illustrates how we are dealing with PPs. One of its category sequences is `(v-ing' in_v-ing)`. This means that *in* is able to fill the `in_v-ing'` valency slot of *helpful* but still has an open `v-ing'` slot.
By now, the mechanism should have become clear. In the next step, *understanding* fills the empty `v-ing'` slot of *in* and opens other slots. The demonstrative *this* fills the `npo'` slots of understanding but still needs a singular noun. *Statement* is able to fill this slot. The next word is a full stop. At this point, there are analyses that still have unfilled valency slots, e. g. the `as_npo'` slot of the second pattern of *understanding* or the PP slots of some patterns of *statement*. As the full stop ends the current sentence, these analyses are considered ungrammatical. However, there is also one analysis which has no open valency slots (using the first patterns of *understanding* and *statement*). The dependency structure of this analysis is visualized in the stemma in Fig. 1.

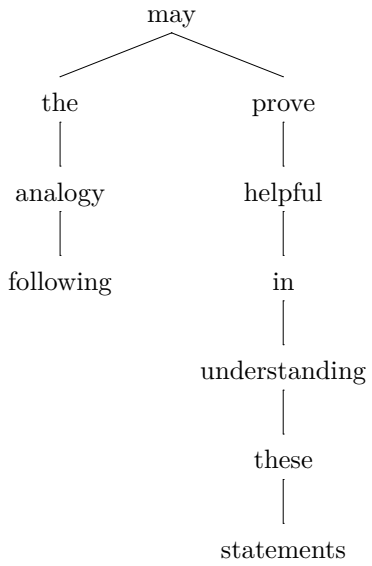

Figure 1: Stemma for example (5)

During the analysis, we have the possibility of fusing certain function words with content words (Hausser, 2006, 87–90). The resulting more compact dependency structure is shown in Fig. 2.

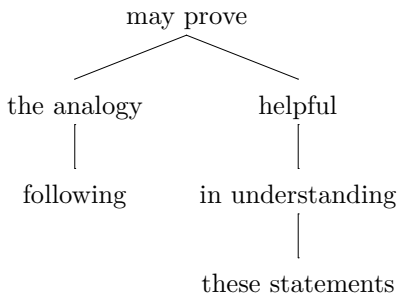

Figure 2: Stemma for example (5), illustrating the fusion of function words with content words

5. Problems

Here we will highlight only some of the more prominent difficulties, for a more detailed discussion cf. Spohr (2004, 34–40) and Proisl (2008, 107–110).
There are some pattern elements in the VDE that are not determined in their form by the valency carrier. The symbol ADV, for example, represents a valency slot that can be filled by an adverb phrase, a noun phrase, a prepositional phrase or an adverbial clause. This kind of underspecified category is difficult to deal with in an NLP system.
Sometimes, the VDE provides information on collocations or contextual lexical tendencies, e. g. + wh-CL

[9] We display only the attributes discussed in the text.
[10] Changes include the addition of valency slots for subjects of finite verb forms and explicit case marking for NPs.
[11] The alert reader will have noticed that the determiner is categorized as needing a noun, not vice versa. So, strictly speaking, we are dealing here with determiner phrases (DPs) instead of noun phrases (NPs).

(often: how or what) or [cannot] + N. Such information, although very valuable, is also difficult to integrate.
A further issue, which is not a "problem" as such, is coverage. While we pointed out above that token coverage is surprisingly high, the absolute number of entries is still quite small. It would therefore be desirable to increase the size of the resource.

6. Comparison

The *Valency Dictionary of English* and the treatment of valency in Left-Associative Grammar have been compared to other resources and formalisms (Proisl, 2008, 33–42, 50–68). We will give a very brief summary of the most important findings here.
A comparison with COMLEX, VALEX and FrameNet suggests that the VDE entries provide a more detailed (and in some cases also more accurate) description of the syntactic and semantic valency properties of their lemmata. As it is a highly specialized dictionary, the VDE can of course not compete with the wealth of non-valency information contained for example in COMLEX.
From a valency point of view, Left-Associative Grammar has two main advantages over other systems such as Head-Driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994) or various flavors of Categorial Grammar, e. g. Dependency Categorial Grammar (Pickering and Barry, 1993). First, long distance dependencies can be treated in an uncontroversial way. Second, in contrast to other formalisms, LAG is closer in spirit to most versions of valency theory in that it treats valency exclusively as a property of words and not also as a property of constituents or phrases.

7. Conclusion

In this paper we tried to show three things. First, that the *Valency Dictionary of English* in electronic form is very well suited for being used as a resource in Natural Language Processing. Second, that the integration of its main data into the formalism of Left-Associative Grammar can be accomplished without problems. Third, that LAG provides a simple way of handling natural language phenomena.
A resource of high quality, such as the VDE, simplifies the grammar development because the developer need not be concerned with its correctness but can use it as a reliable source. Therefore, the use of such a resource makes it possible to derive automatic analyses of equally high quality.

8. References

The British National Corpus. 2007. Version 3 (BNC XML edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. `http://www.natcorp.ox.ac.uk`.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R. L. Petruck. 2003. Background to framenet. *International Journal of Lexicography*, 16(3):235–250.

Ralph Grishman, Catherine Macleod, and Adam Meyers. 1994. Comlex syntax: Building a computational lexicon. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, volume 1, pages 268–272.

Roland Hausser. 2001. *Foundations of Computational Linguistics: Human-Computer Communication in Natural Language*. Springer, Berlin, New York, 2nd edition.

Roland Hausser. 2006. *A Computational Model of Natural Language Communication: Interpretation, Inference, and Production in Database Semantics*. Springer, Berlin, Heidelberg, New York.

Ulrich Heid. 2007. Valency data for natural language processing: What can the *Valency Dictionary of English* provide? In Thomas Herbst and Katrin Götz-Votteler, editors, *Valency. Theoretical, Descriptive and Cognitive Issues*, pages 365–382. Mouton de Gruyter, Berlin, New York.

Thomas Herbst and Peter Uhrig. 2009. Erlangen Valency Pattern Bank – a corpus-based research tool for work on valency and argument structure constructions. Website. `http://www.patternbank.uni-erlangen.de`.

Thomas Herbst, David Heath, Ian F. Roe, and Dieter Götz. 2004. *A Valency Dictionary of English: A Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives*. Mouton de Gruyter, Berlin, New York.

Besim Kabashi. 2007. Pronominal clitics and valency in Albanian: A computational linguistics perspective and modelling within the LAG-framework. In Thomas Herbst and Katrin Götz-Votteler, editors, *Valency. Theoretical, Descriptive and Cognitive Issues*, pages 339–352. Mouton de Gruyter, Berlin, New York.

Anna Korhonen, Yuval Krymolowski, and Ted Briscoe. 2006. A large subcategorization lexicon for natural language processing applications. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1015–1020, Genua.

Martin Pickering and Guy Barry. 1993. Dependency categorial grammar and coordination. *Linguistics*, 31:855–902.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

Thomas Proisl. 2008. Integration von Valenzdaten in die grammatische Analyse unter Verwendung des *Valency Dictionary of English*. Master's thesis, Philosophische Fakultät und Fachbereich Theologie, Friedrich-Alexander-Universität Erlangen-Nürnberg.

Dennis Spohr. 2004. Using *A Valency Dictionary of English* to enhance the lexicon of an English LFG grammar. M. A. study. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris.