Text and Language

Structures · Functions · Interrelations
Quantitative Perspectives

Edited by
Peter Grzybek
Emmerich Kelih
Ján Mačutek

Language

prae
sens

Peter Grzybek
Emmerich Kelih
Ján Mačutek
(eds.)

Advisory Editor
Eric S. Wheeler

Text and Language

Structures · Functions · Interrelations.
Quantitative Perspectives

SONDERDRUCK

prae
sens

On the quantitative analysis of verb valency in Czech

Radek Čech, Ján Mačutek

1 Introduction

It is a matter of common knowledge in linguistics that verb valency is a verbal property which governs the other parts of a sentence. Although valency has been analysed in detail for more than fifty years (cf. Agel et al. 2004), some fundamental problems have not been resolved so far. For instance, no common criteria or tests for the distinguishing complements and adjuncts have been found, despite the fact that a distinction between them plays a crucial role in any valency approach (see Section 2). Since the absence of these criteria seriously undercuts the whole conception of valency, the question about the validity or the suitability of the valency approach emerged.

The goal of the present study is not to solve any of the fundamental problems of valency. We just decided to test empirically whether valency, in spite of the mentioned problems, reflects some important language property or mechanism. The only attempt, to our knowledge, to analyse valency empirically was presented in Köhler (2005a), where some properties of verb valency in German were observed: specifically the distribution of valency frames of each verb, the distribution of unique valency patterns, and the distribution of complement variants (a variant being the possibility to express a given complement of the verb by different semantic roles). Also the relationship between the number of complements of each verb and the number of complement variants was observed. In all cases regular distributions were detected which means that the distribution of observed entities could be viewed as a result of a diversification process (cf. Altmann 2005). In the present study we follow Köhler's methodological approach; we examine the distribution of valency frames in Czech and test the hypothesis concerning a relationship between the number of valency frames and word length.

The article is organized as follows: a very short overview of the main valency properties, in the "traditional" sense, is given in Section 2; valency hypotheses which were tested are presented in Section 3; Section 5 is focused on a methodology and language material used for the hypotheses testing; the results are presented in Section 4; and the article is closed by further research proposals.

2 Valency properties

Valency is usually viewed as a kind of a lexico-syntactic property which "involves the relationship between, on the one hand, the different subclasses of a word-class (such a verb) and, on the other, the different structural environments required by the subclasses, these environments varying both in the number and in the type of elements. Valency is thus seen as the capacity a verb has for combining with particular patterns of other sentence constituents" (Allerton 2005: 4878). In other words, valency "denotes the property of the verb to claim or to admit, respectively, particular kinds and forms of complements. The verb opens up slots, in which the complements enter as arguments" (Heringer 1993: 303). More concretely, valency determines

(1) the number of complements, compare monovalent verb *sleep*:
   a. Baby – sleeps
      versus bivalent verb *write*
   b. Mary – writes – the letter
      versus trivalent verb *give*
   c. Peter – gave – Mary – the book

(2) the form of the complements, compare verb *look* claiming adverbial complementation:
   a. Mary looks nice
      NOUN VERB ADVERB
      versus verb *bring* claiming nominal complementation
   b. Peter brought the book,
      NOUN VERB NOUN

(3) the meaning of the complements, compare the subject of the verb *see* which is assigned as the experiencer:
   a. Mary saw the house
      EXPERIENCER PATIENT
      versus the subject of the verb *kick* which is assigned as the agent
   b. Peter kicked the ball
      AGENT PATIENT

As we noted in Section 1, in any valency theory, a distinction between obligatory complements and facultative (optional) complements (they are usually called adjuncts) of the verb plays a crucial role. However, despite a huge endeavour (for more details see Buysschaert 1982, Herbst 2007, Panevová 1974, Storrer 1992, Van Valin & LaPolla 1997) to find common criteria or tests for distinguishing complements and adjuncts, a satisfying outcome has not been reached yet (Comrie 1993: 906ff.). So, some authors admit that "[t]he state of distinction into *C* [complement] and *A* [adjunct] and the position of valency

theory suggests that an intuitively substantiated basis (…) has not yet been sufficiently justified by theory. The different relational criteria – as far as they are methodically applicable in a controlled way – yield similar results in the majority of cases but also opposite ones. There are no adequate criteria to evaluate the quality of the results. (…) *It seems likely, however, that valency is a semantic phenomenon of which we have not yet found a clear view or which we perhaps have not even understood properly*" (Heringer 1993: 307; emphasis added by the authors).

It is clear that this fact seriously undermines the conception of valency in general. In other words, how can one seriously talk about "valency theory" without clear criteria for determining one of the most important properties of verb valency? Consequently, is not valency the notion which, although it fits one's intuition, does not reflect any important language mechanism? Or even, is it not just a matter of tradition?

Of course, the fact that the criteria have not been found yet does not necessarily mean that valency is an "empty" or senseless notion. However, if valency indeed reflects some important language property or mechanism, it is necessary, according to us, to prove the validity of this notion empirically. Therefore we tested two hypotheses concerned with (1) a regular distribution of valency frames in a language and (2) the relationship between the number of valency frames and the word length (several hypotheses on valency can be found in Köhler and Altmann 2009: 16ff.). So, if these hypotheses are not rejected, it seems reasonable to consider valency as a linguistically meaningful notion. Moreover, it will be possible to integrate valency to the synergetic linguistic framework (Köhler 2005b).

3 Valency hypotheses

3.1 Regular distribution of verb valency

Let us assume that valency, contrary to all problems related to the notion, reflects some important language mechanism and it could be considered as a verb classification enabling hypotheses testing and the exploration of relationships between valency and other language properties. One of the ways of evaluation of any classification scheme is an observation of rank-frequency distribution. It has been shown that "linguistic classification is 'good', 'useful' or 'theoretically prolific' if the taxa follow a 'decent rank-frequency distribution' " (Altmann 2005: 647). The regular distribution is viewed as a consequence of a diversification process and there is an assumption which says "that if an entity diversifies on one direction, the frequencies of the resulting classes are not equal but can be ordered according to decreasing frequency" (Altmann 2005: 646). So, if valency represents a "theoretically prolific" class, it should have a regular distribution.

3.2 The shorter the verb, the more verb valency frames

A relationship between the length of the verb and the number of valency frames of the given verb should be a consequence of the relationship between frequency and length. In other words, the shorter the verb, the more frequent the verb, and so the more frequent the verb occurs in more contexts, i.e. in more valency frames.

3.3 Language material and methodology

The crucial aspect of the testing of the hypotheses lies in both the choice of language material and the clear definition of valency. As for language data, we have used the Czech valency lexicon Vallex 1.0 (Lopatková et al. 2003) which contains about the 1400 most frequent Czech verbs.[1] Vallex 1.0 is based on Sgall's theoretical approach known as the Functional Generative Description (Sgall et al. 1986, Hajičová et al. 1998) and is closely related to the Prague Dependency Treebank project (Hajič et al. 2006).

As for definition of valency, we follow the Prague Dependency Treebank approach and we use the Vallex 1.0 annotation. In this study, we take into account only those verb modifications assigned as obligatory. The obligatoriness of a verb modification is determined by means of a so-called dialogue test in Vallex 1.0. The main principle of the dialogue test is defined as follows: "If [speaker] *A* uses a sentence *S* and [speaker] *B* asks him *wh*-question concerning the participant *P*, *A*'s answer might be "I don't know" (without disturbing the dialogue) if and only if the participant *P* is not semantically obligatory in *S*" (Panevová 1974: 15). More concretely, in the dialogue (4) the answer "I don't know" is unacceptable, so the verb *come* has assigned obligatory complementation "direction-to" and it is taken as bivalent in Vallex 1.0, although it is properly used as monovalent in the "surface" sentence structure.

(4) A: My friends have come.
B: Where to?
A: *I don't know.

On the contrary, in the dialogue (4) the answer "I don't know" is acceptable, so the complementation "direction-from" is optional.

1. Concretely, verbs were selected as follows: the 1000 most frequent Czech verbs, according to their number of occurrences in a part of the Czech National Corpus, were taken at the beginning and then their perfective or imperfective aspectual counterparts were added, if they were missing. For more details, see Vallex's 1.0 official web pages: `http://ufal.mff.cuni.cz/vallex/1.0/` and the technical report (Lopatková et al. 2006).

(5) A: My friends have come.
B: Where from?
A: I don't know.

For the hypotheses testing we counted verb valency frames which consist just of obligatory complementation (Vallex 1.0 comprises also other types of complementation; these ones we omit in this study). It is necessary also to note that we just counted formally unique valency frames; this means that if the verb has, for instance, two identical valency frames (as a consequence of a semantic shift), we count only one.

4 Results

4.1 Distribution of valency frames

As it can be seen in Table 1, the distribution of valency frames is indeed regular – in fact, so regular that there are many distributions with a very good fit.

Table 1: Distribution of valency frames

x – Number of valency frames	Number of verbs with x valency frames
1	815
2	319
3	152
4	73
5	38
6	17
7	7
8	7
9	4
10	2
11	1
14	1
17	1

Tentatively, we present the fit of the Good distribution (cf. Wimmer and Altmann 1999: 219ff.),

$$P_x = C\frac{p^x}{x^a} \tag{1}$$

where a, p are parameters and C is a normalization constant. We obtain an excellent fit (in terms of the chi square goodness of fit test, with $P = 0.9693$,

$a = 0.6562$, $p = 0.6034$). We do not claim that the Good distribution should be a general model; here only the 'smoothness' or 'regularity' of the distribution is demonstrated. Most probably the model would have to be modified or generalized when data from more languages are available.

4.2 Relationship between verb length and number of valency frames

The hypothesis "The shorter the verb, the more valency frames" is also corroborated, see Table 2. We note that the verb length was measured in syllables and the infinitive form of verbs was considered.

Table 2: Mean length of valency frames

x – Number of valency frames	Mean length (in syllables) of verbs with x valency frames
1	3.40
2	3.14
3	2.97
4	2.71
5	2.45
6	2.41
7	2.00
8	2.57
9	1.50
10	1.50
11	2.00
14	1.00
17	1.00

Again only tentatively, we suggest the function $y = Cx^a e^{-bx}$ as a model. The suggested model is a special case of a very general scheme derived by Wimmer and Altmann (2005). The goodness of fit, although not so excellent as for the distribution of valency frames, is still satisfying ($R^2 = 0.8959$, with $C = 3.6675$, $a = 0.0308$, $b = 0.0834$). Some discrepancies (the observed values are not decreasing) can be caused by relatively small numbers of verbs with many valency frames (e.g., we have only one verb with 11 valency frames, which is one of two problematic cases).

5 Further research

The corroboration of the hypotheses presented in this study allows us to consider valency as an important property of the language, despite many obscurities associated with this notion in linguistics. Nevertheless, further analyses

should be done: first, it is necessary to observe valency properties in other languages; next, hypotheses predicting relationships between valency and synonymy, polysemy, frequency and the other language characteristics should be tested. A fresh view to valency could be achieved by analyses focused on valency "in use", meaning that the distribution of valency frames given by both obligatory and optional complements in actual language usage are the subject of the analysis.

Acknowledgments. Radek Čech was supported by the Czech Science Foundation GAČR (grant no. 405/08/P157: "Components of transitivity analysis of Czech sentences (emergent grammar approach)". Ján Mačutek was supported by the Austrian FWF Lise Meitner Program.

References

Agel, V.; Eichinger, L.M.; Eroms, H.-W.; Hellwig, P.; Heringer, H.J.; Lobin, H.
2004 *Dependenz und Valenz / Dependency and Valency: Ein Internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research.* Berlin, New York: de Gruyter.

Allerton, D.J.
2005 "Valency Grammar." In: Brown, E.K. (ed.), *The Encyclopedia of Language and Linguistics.* Amsterdam: Elsevier Science Ltd., 4878–4886.

Altmann, G.
2005 "Diversification processes." In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook.* Berlin, New York: de Gruyter, 646–659.

Buysschaert, J.
1982 *Criteria for the Classification of English Adverbials.* Brussels: Koninklijke Academie.

Comrie, B.
1993 "Argument Structure". In: Jacobs, J.; Stechow, A.; Sternefeld, W.; Vennemann, T. (eds.), *Syntax. An International Handbook of Contemporary Research.* Berlin, New York: de Gruyter, 905–914.

Hajič, J.; Panevová, J.; Hajičová, E.; Pajas, P.; Štěpánek, J.; Havelka, J.; Mikulová, M.
2006 *Prague Dependency Treebank 2.0.* Philadelphia: Linguistic Data Consortium.

Hajičová, E.; Partee, B.H.; Sgall, P.
1998 *Topic-Focus Articulation, Tripartite Structures, and Semantic Content.* Dordrecht: Kluwer.

Herbst, T.
2007 "Valency complements or valency patterns?" In: Herbst, T.; Götz-Votteler, K. (eds.), *Valency: Theoretical, Descriptive and Cognitive Issues.* Berlin, New York: de Gruyter, 15–36.

Heringer, H.J.
1993 "Basic Ideas and the Classical Model?" In: Jacobs, J.; Stechow, A.; Sternefeld, W.; Vennemann, T. (eds.), *Syntax. An International Handbook of Contemporary Research.* Berlin, New York: de Gruyter, 297–316.

Köhler, R.
2005a "Quantitative Untersuchungen zur Valenz deutscher Verben", in: *Glottometrics*, 9; 13–20.
2005b "Synergetic Linguistics." In: Köhler, R.; Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook.* Berlin, New York: de Gruyter, 760–775.

Köhler, R.; Altmann, G.
2009 *Problems in Quantitative Linguistics 2.* Lüdenscheid: RAM-Verlag.

Lopatková, M.; Žabokrtský, Z.; Skwarska, K.; Benešová, V.
2003 "Vallex 1.0. Valency lexicon of Czech Verbs." Prague: Center of Computational Linguistics. `http://ufal.mff.cuni.cz/vallex/1.0/`

Lopatková, M.; Žabokrtský, Z.; Benešová, V.
2006 "Valency lexicon of Czech verbs VALLEX 2.0. Technical Report 34." Prague: ÚFAL MFF UK. `http://ufal.mff.cuni.cz/vallex/2.0/publ/06-techrep.pdf`

Panevová, J.
1974 "On verbal frames in functional generative description I.", in: *Prague Bulletin of Mathematical Linguistics*, 22; 3–40.

Sgall, P.; Hajičová, E.; Panevová, J.
1986 *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* Dordrecht: Reidel Publishing Company.

Storrer, A.
1992 *Verbvalenz. Theoretische und methodische Grundlagen ihrer Beschreibung in Grammatikographie und Lexikographie.* Tübingen: Niemeyer.

Van Valin, R.D.; LaPolla, R.J.
1997 *Syntax: Structure, Meaning, and Function.* Cambridge: Cambridge University Press.

Wimmer, G.; Altmann, G.
1999 *Thesaurus of univariate discrete probability distributions.* Essen: Stamm.
2005 "Unified derivation of some linguistic laws." In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook.* Berlin, New York: de Gruyter, 791–807.

Contents