On the relation between verb full valency and synonymy

Radek Čech
University of Ostrava
Faculty of Arts
Department of Czech Language
Czech Republic
cechradek@gmail.com

Ján Mačutek and **Michaela Koščová**
Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics
Department of Applied Mathematic and Statistics
Slovakia
jmacutek@yahoo.com
michaela.koscova@fmph.uniba.sk

Abstract

This paper investigates the relation between the number of full valency frames (we do not distinguish between complements and optional adjuncts, both are taken into account) of a verb and the number of its synonyms. It is shown that for Czech verbs from the Prague Dependency Treebank it holds *"the greater the full valency of a verb, the more synonyms the verb has"*.

1 Introduction

Verb valency has been studied for more than fifty years in linguistics and the study of this phenomenon has enhanced knowledge about sentence functioning substantially. Although there still remain some problems (even fundamental ones) which need to be solved in this research area (see Section 2), verb valency is considered to have a decisive impact on the sentence structure. Consequently, it has become a standard part of the majority of grammar books, verb valency lexicons have appeared for many languages, and plenty of articles focused on it have been published so far. These analyses are mostly descriptive; usually valency patterns, relationship between syntax and semantics, classification criteria etc. are investigated, see, e.g., Mukherjee (2005), Herbst and Götz-Votteler (2007), and Faulhaber (2011). However, in linguistics there are also attempts to overcome the descriptive character of research and to ground the discipline on empirically testable hypotheses, see, e.g., Zipf (1935), Sampson (2001), Sampson (2005), Gries (2009), and Köhler and Altmann (2011). The goal of such a methodology is not only to describe phenomena under study but also to interpret them, i.e., to find their relations to other language properties, and, in the ideal case, to explain them within a theory of language. It is to be emphasized that, within this approach, all conclusions are based on statistically testable hypotheses, and the aim is to build a theory, i.e., a system of hypotheses and scientific laws (which are statements theoretically derived and empirically tested), see Bunge (1967) in general and Altmann (1993) more specifically for linguistics. As for verb valency, results achieved by this methodology were presented by Köhler (2005a), Liu (2009), Čech and Mačutek (2010), Čech et al. (2010), Liu (2011), Köhler (2012), Gao et al. (2014), and Vincze (2014). The authors tested hypotheses on relations between the number of valency frames and the frequency, length of verb and its polysemy; further, it was shown that the distribution of valency frames is a special case of a very general distribution which is used very often as a mathematical model in linguistics (Wimmer and Altmann, 2005).

All these studies are somewhat connected to a synergetic theory of language, see Köhler (1986) and Köhler (2005b), and they represent first steps in the endeavor to implement verb valency (or valency in general) to a synergetic model of syntax (Köhler, 2012). The paper by Gao et al. (2014) deserves a special mention, as it contains an explicit synergetic scheme of interrelations. The scheme includes the verb valency and some other verb properties (frequency, length, polysemy, polytextuality, and, in addition, two properties which are specific for the Chinese language, namely the number of strokes and the number of pinyin letters). The present study follows the same direction. Our goal is to analyse the relationship between verb valency (to be exact, its variant which is called full valency, see Section 2) and another important language property – synonymy. Specifically, we test a hypothesis on the relationship between the number of full valency frames of verb and its synonymy, namely, we suppose that it holds "the more full valency frames of a verb, the

Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), pages 68–73,
Uppsala, Sweden, August 24–26 2015.

more synonyms the verb has". The validity of this statement will be tested on data from the Czech language.

2 Full valency

The concept of full valency was introduced by Čech et al. (2010). It can be viewed as a reaction to the absence of reliable criteria for distinguishing obligatory arguments (complements) and non-obligatory arguments (optional adjuncts), see Rickheit and Sichelschmidt (2007) and Faulhaber (2011). Full valency does not distinguish between these two types of arguments; it takes into account all arguments of a verb which occur in the actual language usage (i.e., all nodes in a syntactic tree which depend directly on the verb represent its full valency frame). Following the paper by Čech et al. (2010), only formally unique full valency frames are considered. This means that if the verb occurs in two or more identical full valency frames in the corpus, only one of them is counted.

Čech et al. (2010) assumed that the distribution of the number of full valency frames is not chaotic or accidental but it is governed by fundamental principles which have an impact also on other language characteristics (such as the distribution of word frequencies, word lengths, morphological categories, etc.). Further, according to the authors, full valency of verbs should be systematically related to other language properties (e.g., to the frequency of verb, to its length, etc.) as a result of the synergetic character of language, see Köhler (2005b) and Köhler (2012).

First results – Čech et al. (2010), Gao et al. (2014) and Vincze (2014) – corroborated the reasonability of the approach. They revealed, for instance, that the distribution of full valency frames can be modelled by the same model as the distribution of valency frames based on the traditional argument classification, see Čech and Mačutek (2010) for Czech, Liu (2011) for English, Gao et al. (2014) for Chinese, and Vincze (2014) for Hungarian. Given these results, "traditional" valency and full valency seem to be governed by the same mechanism, and traditional valency can be interpreted, tentatively at least, as a special case of full valency.

3 Verb full valency a synonymy

Every hypothesis should be based on some theoretical assumption(s). Without it, one can find even strong correlation (e.g., inductively) between observed phenomena, however, it does not have to mean anything. Therefore, a crucial question is why one should expect the existence of a relationship between verb valency and synonymy. To find an answer, let us start from a wider perspective. At least since Zipf (1935), it is known that semantic properties of language are systematically related to other language characteristics (e.g., relative frequency, degree of intensity of accent, etc.). These systematic relationships can be interpreted as a consequence of the dynamic evolution of language caused by language usage (Bybee and Hopper, 2001). For an illustration, assume a development of usage of any word. Initially, it was used in a unique sense and in a specific context. Next usages of the word led both to a strengthening of the sense and to an increase of the number of contexts in which the word occurs. More generally, the word properties were formed by two opposite forces: a unification and a diversification (Zipf, 1935). As a result, fundamental characteristics of the word were established (for instance, the length of the word is a consequence of its frequency as well as the number of its derivatives, compounds in which it occurs etc.). As for the meaning of the word, a high frequency of its usage increases a chance that the word is used in different contexts. Different contexts usually modify slightly the word meaning, which leads (sometimes) to a "codification" of a new meaning of the word. Therefore, a relationship between frequency and polysemy emerges. Further, the more meanings the word has, the more semantic domains exist in which the word can occur. Obviously, different semantic domains are represented by different sets of words. Consequently, a word which occurs in more semantic domains increases its chance of having more synonyms.

As for verb valency, there is, as can be seen from any valency dictionary, a clear relationship between polysemy of the verb and its valency. Specifically, different meanings of the verb are often represented by different valency frames, see Liu (2011) for an analysis of the relation between the two properties. Consequently, it seems reasonable to hypothesize the relationship between verb valency and synonymy; to be precise, we expect that the number of synonyms of a verb tends to increase with the increasing number of its full valency frames. We thus have a deductive hypothe-

sis which will be tested empirically in Section 5. A quantification (which necessarily precedes tests) not only enables the application of statistical methods, it also opens a way towards a mathematical model (which, in turn, makes possible more objective comparisons of different languages, language typology based on values of its parameters, etc.).

4 Language material

For the counting of full valency verb frames, the Prague Dependency Treebank 2.0 was used (Hajič et al., 2006); specifically, the data annotated on an analytical layer, which consists of 4264 documents, 68,495 sentences and 1.2 million tokens. For the determination of synonyms of a verb, we use the Czech WordNet from the EuroWordNet project (Vossen, 1997); it contains 32,116 words and collocations, 28,448 synsets, 43,958 literals, see Horák and Smrž (2004) and Hlaváčková et al. (2006).

The term "full valency" means that all verb directly dependent words (arguments) which occur in the sentence are taken into account. To determine a full valency frame of a verb, we use argument characteristics as follows: analytical functions (e.g., subject, object), morphological cases (e.g., nominative, genitive), and lemmas (only in the case of prepositions). Particular characteristics are assigned to arguments in accordance with the PDT 2.0 annotation. Specifically, from the sentence *John gave four books to Mary yesterday*, we obtain the following full valency frame of the verb *give*: GIVE [subject/nominative; object/accusative; AuxP/dative/lemma TO; Adv], see Figure 1.

This procedure is used for all predicate verbs in the corpus and, finally, we get list of verbs (lemmas) with assigned full valency frames.

The number of synonyms of a verb is determined from the database CzechWordNet which is organized as a network of basic entities called synsets, i.e., synonym sets. Each synset corresponds to one meaning of a word or a collocation. In this paper, synonymy of each verb is defined as the number of lemmas which appear with the verb in particular synsets. For instance, the verb *intend* has four synsets in English Wordnet:

1. intend: 1, mean: 4, think: 7;
2. intend: 2, destine:2 , designate: 4, specify: 6;
3. mean: 1, intend: 3;
4. mean: 3, intend: 4, signify: 1, stand for: 2;

in which nine different lemmas appear (in order to avoid confusion, it should be emphasized that, e.g., "mean: 1" and "mean: 4" express two different meanings, and hence they also represent two different lemmas) – i.e., the verb *intend* has nine synonyms. Hereby we do not claim that other possibilities of determining the number of synonyms (e.g., distinguishing among different senses of the verb) are worse; quite on the contrary, using several of them (while keeping in mind what they have in common and in what they differ) and comparing results can lead to a deeper understanding of mechanisms "behind" synonymy (and language in general).

Altogether, we work with 2120 verbs in this study.

5 Methodology and results

The validity of our hypothesis for Czech data was checked in two different (albeit related) ways.

First, one can compute the correlation coefficient between full verb valency and synonymy. There is no a priori reason to suppose the linearity of the relation; therefore, the Kendall correlation coefficient – see, e.g., Hollander and Wolfe (1999) – was used (similarly as the well-known Pearson correlation coefficient, it takes values from the interval [-1,1]; value 1 means that the relation "the greater one variable, the greater the other" is valid for all data without an exception). It is a measure of a monotonous relation (without specifying the type of a functional relation, like, e.g., linearity) between two variables (full valency and synonymy in our case). Thus it is a more general and more robust characteristic of the relation than the Pearson correlation coefficient (which is a measure of linearity of the relation).

The Kendall correlation coefficient evaluates to 0.18 for our data. It is, quite clearly, a non-zero value (if we test the hypothesis of zero value of the coefficient, we obtain the p-value lesser than 0.0001, hence, the hypothesis is rejected for all reasonable significance levels). There are, however, several minor problems associated with the test.

First, it is well-known that practically all hypotheses are rejected if sufficiently high amount of data are used. This fact was discussed specifically with respect to linguistic data by Mačutek and Wimmer (2013). Our sample size (2120 verbs)

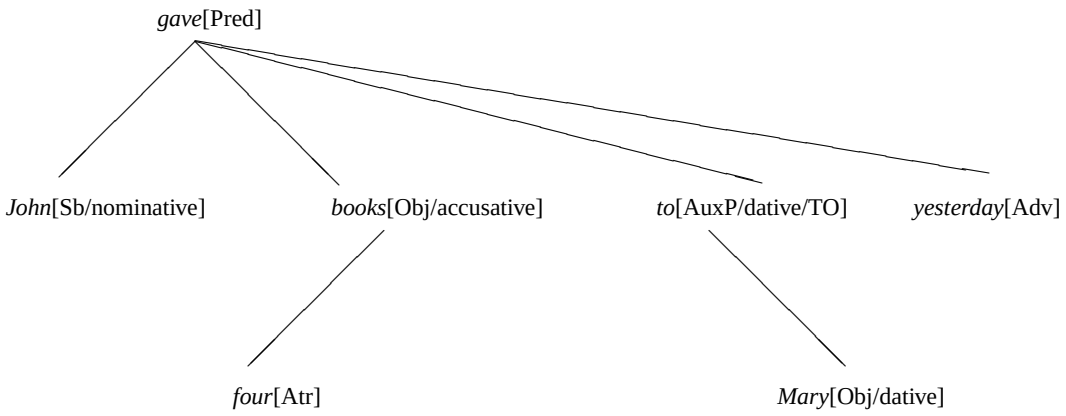

Figure 1: Syntactic tree of the sentence *John gave four books to Mary yesterday*.

is not too high yet, but studies using higher volumes of language material can appear in future (see also comments in Section 6), for which (almost) any hypothesis would be rejected in terms of the p-value. Thus, a need of a unified approach to checking the validity of the hypothesis arises.

Anyway, the p-value should be read cautiously. It can serve as a decision rule whether to reject a hypothesis or not, but p-values resulting from different tests are not directly comparable (Grendár, 2012). Applied to our problem, based on the p-value we reject the hypothesis that full valency and synonymy are (monotonously) independent, however, from the p-value we cannot deduce a strength (or a type) of their relationship.

Next, the test for the Kendall correlation coefficient supposes no ties in the data, but there are many verbs with the same full valency (especially the low values of full valency frames occur very often – which is true also for the "traditional" valency.

Finally, if an "optical criterion" is taken into account, the data fluctuate quite strongly, as can be seen in Figure 2, and the increasing trend indicated by the positive value of the Kendall correlation coefficient is not too obvious.

Therefore, in order to be able to see a clearer picture and to provide a tool applicable also to higher sample sizes, we performed also the analysis of pooled data. Groups of at least 20 verbs were created as follows. Starting from the verbs with the highest number of full valency frames, a group of the first 20 verbs was taken. Then, it

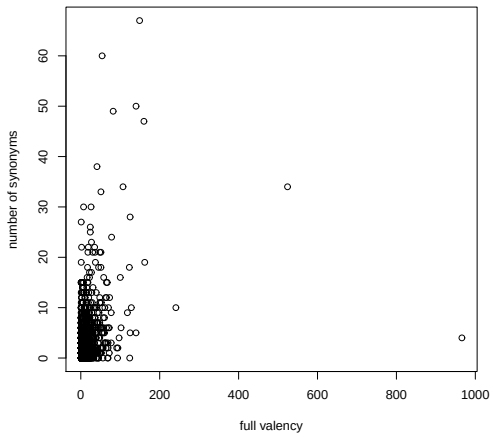

Figure 2: Number of full valency frames and number of synonyms for all verbs under study.

was checked whether the last verb in this groups has more full valency frames than the first verb in the next group – if the respective numbers of full valency frames were equal, the group was enlarged so that all verbs with the same full valency belonged to the same group. This approach was repeteadly applied, until all verbs were divided into groups. Resulting groups do not contain the same numbers of verbs, however, we prefer to keep verbs with the same number of full valency frames in one group, as there is no reasonable ordering of verbs (ones with the same full valency are either ordered alphabetically, or they appear in the chronological order as they were entered into treebanks, etc.). Then, the mean number of full

valency frames and the mean number of synonyms per verb were calculated in each group. The pooling process results in much smoother data, see Figure 3. Obviously, the mean number of synonyms per group tends to increase with the increasing mean full valency.

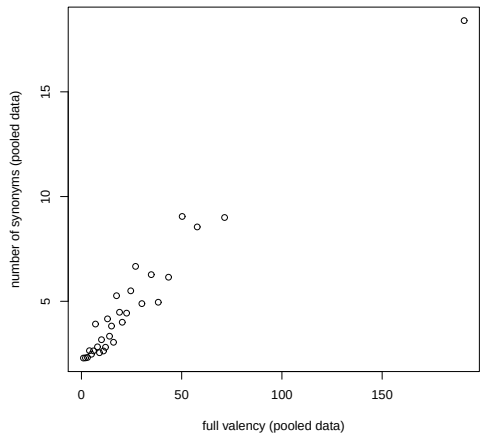

Figure 3: Number of full valency frames and number of synonyms (pooled data).

Admittedly, the minimal size of the group used (i.e., 20 in our case) is purely heuristic; however, other choices lead to very similar pooled data behaviour (an increasing, seemingly even a linear trend is observed). As we consider this paper to be a kind of a pilot study, we postpone a deeper analysis of the full valency – synonymy relation (is there really a linear dependence, or, what we see in Figure 3 is a part of a flat power law curve? are parameters of the line/curve language specific? if yes, do they correspond to an established syntax-based language typology? etc.) until results for more languages are available.

6 Conclusion

The results presented in this study can be seen as the first step in the empirical research of the relation between the number of full valency frames of verbs and the number of synonyms. It goes without saying that an analysis based on a single language cannot be interpreted as an "honest", general enough corroboration of the respective hypothesis. However, tentatively the results allow to expect that synonymy can be related to verb (full) valency, i.e., to one of fundamental syntax properties.

This paper, we hope, will serve also as an impetus for future research in this field. Some questions were already asked at the end of Section 5; in addition, our results call for substantial generalizations in (at least) two directions. First, the same phenomenon (the relation between verb valency and synonymy) should be investigated in several typologically different languages. Second, we suppose that valency of other parts of speech, see, e.g., Spevak (2014), is also related to synonymy; this topic waits for empirical approaches as well. Given the lack of a clear distinction between obligatory and non-obligatory arguments, full valency (of other parts of speech) can again be of help.

Finally, if the hypothesis on a systematic relation between (full) valency and synonymy is more generally corroborated, it should be integrated into the network of (inter)relations among linguistic units and their properties, see Köhler (2005b) and Gao et al. (2014).

Acknowledgement

Supported by the grant VEGA 2/0047/15 (J. Mačutek and M. Koščová) and by Slovak Literary Fund (J. Mačutek).

References

Gabriel Altmann. 1993. Science and linguistics. In Reinhard Köhler and Burghard B. Rieger, editors, *Contributions to Quantitative Linguistics*, pages 3–10. Kluwer, Dordrecht.

Mario Bunge. 1967. *Scientific Research I.* Springer.

Joan Bybee and Paul Hopper. 2001. *Frequency and the Emergence of Linguistic Structure*. John Benjamins, Amsterdam/Philadelphia.

Radek Čech and Ján Mačutek. 2010. On the quantitative analysis of verb valency in Czech. In Peter Grzybek, Emmerich Kelih, and Ján Mačutek, editors, *Text and Language. Structures, Functions, Interrelations, Quantitative Perspectives*, pages 21–29. Praesens, Wien.

Radek Čech, Petr Pajas, and Ján Mačutek. 2010. Full valency. verb valency without distinguishing complements and adjuncts. *Journal of Quantitative Linguistics*, 17(4):291–302.

Susen Faulhaber. 2011. *Verb Valency Patterns. A Challenge for Semantics-Based Accounts.* De Gruyter.

Song Gao, Hongxin Zhang, and Haitao Liu. 2014. Synergetic properties of Chinese verb valency. *Journal of Quantitative Linguistics*, 21(1):1–21.

Marian Grendár. 2012. Is the p-value a good measure of evidence? Asymptotic consistency criteria. *Statistics & Probability Letters*, 82(6):1116–1119.

Stefan T. Gries. 2009. *Statistics for Linguistics with R: A Practical Introduction*. De Gruyter.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková-Razimová, and Zdenka Uresová. 2006. *Prague Dependency Treebank 2.0*. Linguistic Data Consortium, Philadelphia.

Thomas Herbst and Katrin Götz-Votteler. 2007. *Valency: Theoretical, Descriptive, and Cognitive Issues*. De Gruyter.

Dana Hlaváčková, Aleš Horák, and Vladimír Kadlec. 2006. Exploitation of the VerbaLex verb valency lexicon in the syntactic analysis of Czech. In *Proceedings of 9th International Conference on Text, Speech, and Dialogue*, pages 79–85. Springer.

Myles Hollander and Douglas A. Wolfe. 1999. *Nonparametric Statistical Methods*. Wiley, second edition.

Aleš Horák and Pavel Smrž. 2004. VisDic - WordNet browsing and editing tool. In *Proceedings of the Second International WordNet Conference - GWC 2004*, pages 136–141. Masaryk University, Brno.

Reinhard Köhler and Gabriel Altmann. 2011. Quantitative linguistics. In Patrick Colm Hogan, editor, *The Cambridge Encyclopedia of the Language Sciences*, pages 695–697. Cambridge University Press.

Reinhard Köhler. 1986. *Zur linguistische Synergetik. Struktur und Dynamik der Lexik*. Brockmeyer, Bochum.

Reinhard Köhler. 2005a. Quantitative Untersuchungen zur Valenz deutscher Verben. *Glottometrics*, 9:13–20.

Reinhard Köhler. 2005b. Synergetic linguistics. In Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski, editors, *Quantitative Linguistics. An International Handbook*, pages 760–774. De Gruyter.

Reinhard Köhler. 2012. *Quantitative Syntax Analysis*. De Gruyter.

Haitao Liu. 2009. Probability distribution of dependencies basen on a Chinese dependency treebank. *Journal of Quantitative Linguistics*, 16(3):256–273.

Haitao Liu. 2011. Quantitative properties of English verb valency. *Journal of Quantitative Linguistics*, 18(3):207–233.

Ján Mačutek and Gejza Wimmer. 2013. Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics*, 20(3):227–240.

Joybrato Mukherjee. 2005. *English Ditransitive Verbs: Aspects of Theory, Description and a Usage-Based Model*. Rodopi, Amsterdam/New York.

Gert Rickheit and Lorenz Sichelschmidt. 2007. Valency and cognition – a notion in transition. In Thomas Herbst and Katrin Götz-Votteler, editors, *Valency: Theoretical, Descriptive, and Cognitive Issues*, pages 163–182. De Gruyter.

Geoffrey Sampson. 2001. *Empirical Linguistics*. Continuum, London/New York.

Geoffrey Sampson. 2005. Quantifying the shift towards empirical methods. *International Journal of Corpus Linguistics*, 10(1):15–36.

Olga Spevak. 2014. *Noun Valency*. John Benjamins, Amsterdam/Philadelphia.

Veronika Vincze. 2014. Valency frames in a Hungarian corpus. *Journal of Quantitative Linguistics*, 21(2):153–176.

Piek Vossen. 1997. EuroWordNet: a multilingual database for information retrieval. In *Proceedings of the DELOS Workshop on Cross-language Information Retrieval*.

Gejza Wimmer and Gabriel Altmann. 2005. Unified derivation of some linguistic laws. In Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski, editors, *Quantitative Linguistics. An International Handbook*, pages 791–807. De Gruyter.

George K. Zipf. 1935. *The Psychobiology of Language*. Houghton-Mifflin, Boston.