WDS'11 Proceedings of Contributed Papers, Part I, 171–176, 2011. ISBN 978-80-7378-184-2 © MATFYZPRESS

Nominal Valency in Lexicons

A. Vernerová

Abstract. The term valency refers to the number, type and form of arguments that are bound to a word. Valency is specific to any given lexical unit and therefore is covered by lexicons. This is a preliminary survey conducted with the creation of a valency lexicon of Czech nouns in mind. The authors of such a lexicon have to decide who will be the intended users, how the material will be presented and which aspects of valency behaviour will be covered; we present the choices made by the authors of several Czech, English and German resources that cover the valency of nouns, both machine readable [FrameNet 1.5, 2010; NomBank 1.0, 2008; PDT-Vallex, 2006] and printed [Herbst et al., 2004; Sommerfeldt and Schreiber, 1996; Svozilová et al., 2005].

Introduction

Valency plays a crucial role in the Czech linguistic tradition [Panevová, 1980; Daneš et al., 1987; Karlík, 2000; Sgall, 2006]. Lexicographic description of valency has been most extensive in the case of verbs: *Valenční slovník českých sloves* [Lopatková et al., 2008] gives rich linguistic information including valency patterns, division of verbs into semantic classes, information on control, reflexivity and reciprocity. In the last two years, two monographs concerning valency of Czech nouns were published [Čermáková, 2009; Kolářová, 2010]. However, although valency behaviour of nouns is covered by two existing lexical resources [PDT-Vallex, 2006; Svozilová et al., 2005], neither of them offers rich linguistic information. The aim of this survey is to present and compare existing English, Czech and German lexicons[1] which cover nominal valency and to identify some of the crucial decisions (both concerning the material and its presentation) that have to be made before any lexicographic work is begun.

In the first section, we discuss different kinds of uses and users of valency lexicons; in the second section, we touch upon the alternatives to the alphabetical ordering of the items; then we compare the presentation of the syntactic and semantic aspects of valency patterns in some of the available lexicons; finally, we mention the role of corpus evidence and the choice of example sentences.

Intended uses and users. Choice of entries

The creation of a valency lexicon is a lengthy and expensive process; therefore it should ideally produce a resource useful for a wide range of users. In this section, we have a look at how the intended group of users influences the choice of entries in the lexicon.

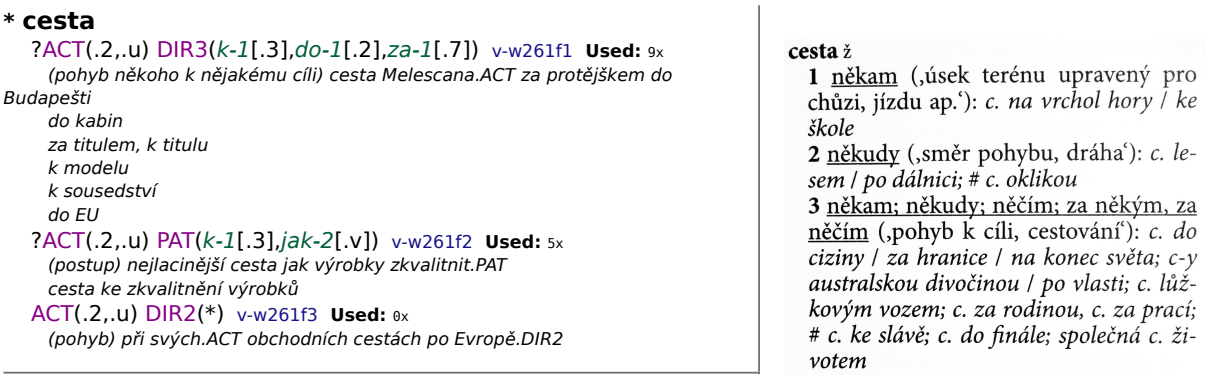
Second language learners

The term "valency," coined by Lucien Tesnière in his 1959 book *Eléments de syntaxe structurale*, was quickly adopted by researchers working in the area of foreign language education. Many valency lexicons are therefore primarily intended for non-native speakers, whether it is Helbig and Schenkel's lexicon of German verbs (published as early as 1969), its adjectival and nominal extensions [Sommerfeldt and Schreiber, 1974, 1977] or newer lexicons covering all three parts-of-speech in one volume [Sommerfeldt and Schreiber, 1996; Herbst et al., 2004].

These lexicons cover around 1500 words (counting the series of German lexicons from the 60's and 70's as one lexicon), of which 250–750 are nouns. The wordlist is based upon two criteria: frequency (only frequent words are important for learners) and the complexity of patterns (learners are more likely to struggle with such words).

However, research such as Bräunling's 1989 survey among the teachers of European Goethe Institutes shows that only very few teachers use valency lexicons in their classes. The rest either don't know valency lexicons at all or find them too theoretical, complex and specialized. Thus it seems that students are best served when valency information is included directly in learner's dictionaries.

[1] For lexical resources which are particularly concerned with valency information, we use the term *lexicon*; the term *dictionary* is reserved for general dictionaries. This may not coincide with the actual titles of the works discussed.



*** cesta**

?ACT(.2,.u) DIR3(*k-1*[.3],*do-1*[.2],*za-1*[.7]) v-w261f1 **Used:** 9x
(pohyb někoho k nějakému cíli) cesta Melescana.ACT za protějškem do Budapešti
do kabin
za titulem, k titulu
k modelu
k sousedství
do EU

?ACT(.2,.u) PAT(*k-1*[.3],*jak-2*[.v]) v-w261f2 **Used:** 5x
(postup) nejlacinější cesta jak výrobky zkvalitnit.PAT
cesta ke zkvalitnění výrobků

ACT(.2,.u) DIR2(*) v-w261f3 **Used:** 0x
(pohyb) při svých.ACT obchodních cestách po Evropě.DIR2

cesta ž

1 někam (‚úsek terénu upravený pro chůzi, jízdu ap.'): *c. na vrchol hory / ke škole*
2 někudy (‚směr pohybu, dráha'): *c. lesem / po dálnici; # c. oklikou*
3 někam; někudy; něčím; za někým, za něčím (‚pohyb k cíli, cestování'): *c. do ciziny / za hranice / na konec světa; c-y australskou divočinou / po vlasti; c. lůžkovým vozem; c. za rodinou, c. za prací; # c. ke slávě; c. do finále; společná c. životem*

Figure 1. Entries for the word *cesta* in PDT-Vallex and in *Slovník vazeb a spojení*

Native speakers and linguists

Dictionaries for native speakers differ from those previously mentioned mainly in size, as it is expected that native speakers tend to look up less frequent words. *Slovník vazeb a spojení* [Svozilová et al., 2005] has 16 000 entries, most of which are verbs. The entries are selected from *Slovník spisovné češtiny* [Filipec et al., 1994], the authoritative dictionary of current standard Czech. The selection includes all verbs which take valency arguments, but only a limited number of adjectives and nouns. Nouns are included only if they are deverbal or have similar valency patterns as deverbals (*hovor o Praze* "a talk about Prague" → *kniha o Praze* "a book about Prague"). Moreover, deverbal and deadjectival nouns are sometimes omitted if their patterns can be inferred from the patterns of the corresponding verb/adjective. Thus, the authors assume that the user is capable of making the inference from *potopit loď* "to sink a ship" to *potopení lodi* "the sinking of a ship," from *odolný vůči/proti/k něčemu* "resistant to something" to *odolnost vůči/proti/k něčemu* "resistance to something." In our opinion it would be necessary to conduct research among users to justify this assumption.

When lexicons are primarily intended as a resource for linguistic research, they may contain richer linguistic information, more complex notation and more elaborate search tools than what would be appropriate for general users. For example, in [PDT-Vallex, 2006] (Figure 1), the arguments are named by their tectogrammatical functors of the Functional Generative Description. A linguist may think that the meaning of these functors is fairly intuitive—*ACT* and *PAT* are the first and the second argument and as such are syntactically determined; the names of the other arguments are then determined by semantic criteria (in our example we have *DIR2* "which way" and *DIR3* "where to").[2] However, it is rather unlikely that general users would make the effort to understand this formalism.

Natural language processing

In Natural Language Processing, valency lexicons play two complementary roles: 1. during the creation of annotated data, valency lexicons enable consistency of annotation which could otherwise not be reached; this is important especially if the data is not large enough for statistical methods to filter out the "noise" [Hajič and Honetschläger, 2003]; 2. they are indispensable to many NLP applications that rely on accurate description of language phenomena, e.g. word sense disambiguation, data mining, language data visualisation and machine translation.

Lexicons created with NLP applications in mind include FrameNet 1.5 [2010]; NomBank 1.0 [2008] and PDT-Vallex [2006]. All three are connected with a project of corpus annotation: FrameNet is based on the *British National Corpus*[3]; NomBank uses the the Wall Street Journal Corpus of the *Penn Treebank*[4]; and PDT-Vallex was created in order to bring consistency into the tectogrammatical annotation of the *Prague Dependency Treebank*[5]. For the latter two, the aim of the project was to cover all nouns, resp. all words with valency behaviour in the corpus. On the other hand, FrameNet annotation does not progress by words but by semantic frames. Some semantic frame is declared to be

[2] For each argument, a list of possible surface forms is given in the brackets.

[3] `http://www.natcorp.ox.ac.uk/`

[4] `http://www.cis.upenn.edu/~treebank/`

[5] `http://ufal.mff.cuni.cz/pdt2.0/`

Travel

Definition:

In this frame a Traveler goes on a journey, an activity, generally planned in advance, in which the Traveler moves from a Source location to a Goal along a Path or within an Area. The journey can be accompanied by Co_participants and Baggage. The Duration or Distance of the journey, both generally long, may also be described as may be the Mode_of_transportation. Words in this frame emphasize the whole process of getting from one place to another, rather than profiling merely the beginning or the end of the journey.

FEs:

Area [Area]
Semantic Type: Location
This is the Area in which the traveling takes place. This frame element describes the enclosed area inside which travelling, of unspecified Source, Path or Goal takes place.
We TRAVELLED in Europe.

Direction [dir]
Excludes: Area
The direction in which the Traveler goes.
They began their ODYSSEY north.

Lexical Units:

commute.v, excursion.n, expedition.n, getaway.n, jaunt.n, journey.n, journey.v, junket.n, odyssey.n, peregrination.n, pilgrimage.n, safari.n, tour.n, tour.v, traveler.n, travel.n, travel.v, trip.n, voyage.n, voyage.v

Figure 2. Parts of the Travel frame in FrameNet: definition, first two core FEs, list of lexical units

finished if all lexical units that the lexicographers have assigned to it have been created and annotated. However, other senses of the same words may be left unannotated; moreover, only a small number of corpus occurrences of each lexical unit are annotated.

Multi-purpose lexicons

Sometimes, electronically available data of a lexicon originally intended for human users can be turned into a valuable resource for NLP applications [Boguraev et al., 1987; Herbst and Uhrig, 2009]. The availability of the data in clearly structured data formats is crucial for NLP usage.

On the other hand, human users benefit from tools that convert machine readable data into browsable form.[6] In an ideal world, flexible visualisation and search tools would serve various kinds of human users, each according to their needs. In particular, we believe that the presentation to general users should be so self-contained that no prior knowledge would be necessary.

Organisation of the lexicon: semantic frames, word fields and derived words

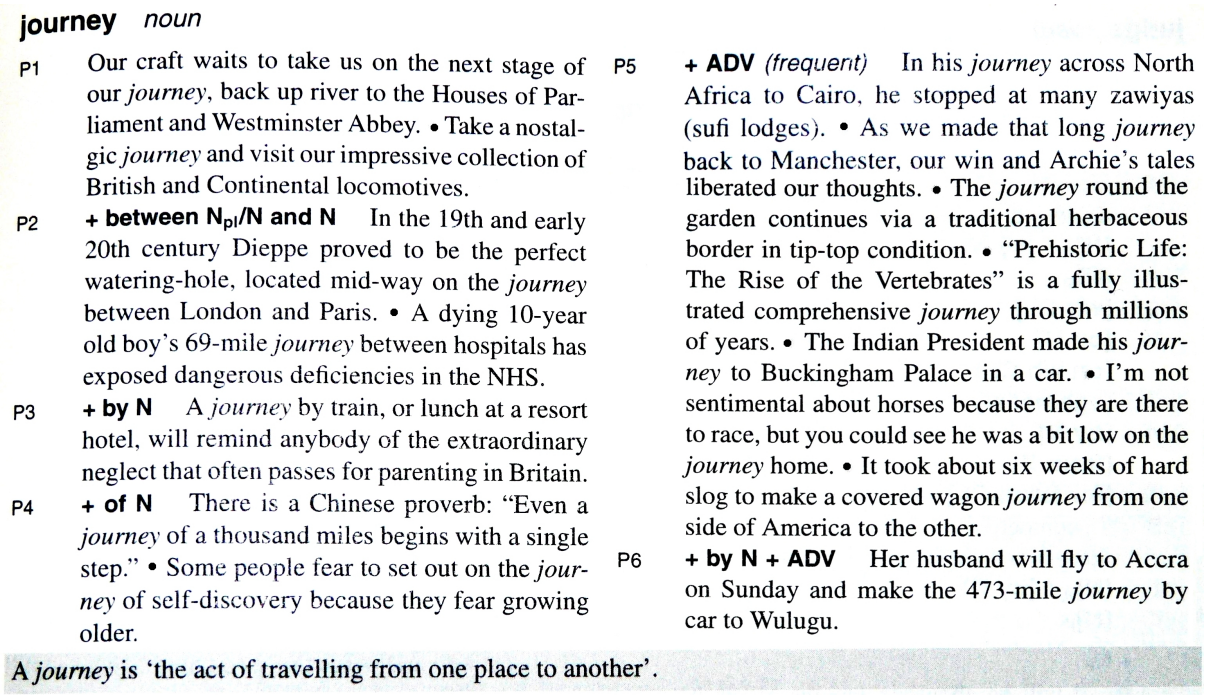
Most dictionaries, and valency lexicons are no exception, are organised so that entries are marked by their headwords and subdivided into "senses." The headwords are usually ordered alphabetically. However, valency is a syntacto-semantic phenomenon, and some regularities stand out more vividly when entries are grouped according to their semantic, or syntacto-semantic characteristics. In this section, we discuss examples of such groupings.

We have already mentioned that the creation of FrameNet proceeds from semantic frames to lexical units. A semantic frame is a schematic representation of a situation type (eating, spying, removing, classifying, etc.) together with a list of participants, propositions, and other conceptual roles that are seen as components of such situation. These participants are called frame elements. As we can see in Figure 2, the entry for each semantic frame contains a definition that characterizes the given semantic situation and the relationships between the most important frame elements, a list of frame elements with more detailed definition of each, and their relationships (e.g. if Direction is expressed, then Area is not expressed), and finally a list of lexical units that belong to this frame.[7]

Another lexicon which organises the entries into semantically based groups is the *Wörterbuch der Valenz etymologisch Vervandter Wörter* [Sommerfeldt and Schreiber, 1996]. The entries are divided into

[6] For example, FrameNet data is stored in XML files on the server; each XML file contains a link to its associated XSLT stylesheet which allows the client's browser to convert the data into a visually friendly report.

[7] There is one lexical entry for each lexical unit. We will discuss the structure of the lexical entries later.



journey *noun*

P1 Our craft waits to take us on the next stage of our *journey*, back up river to the Houses of Parliament and Westminster Abbey. • Take a nostalgic *journey* and visit our impressive collection of British and Continental locomotives.

P2 **+ between N_{pl}/N and N** In the 19th and early 20th century Dieppe proved to be the perfect watering-hole, located mid-way on the *journey* between London and Paris. • A dying 10-year old boy's 69-mile *journey* between hospitals has exposed dangerous deficiencies in the NHS.

P3 **+ by N** A *journey* by train, or lunch at a resort hotel, will remind anybody of the extraordinary neglect that often passes for parenting in Britain.

P4 **+ of N** There is a Chinese proverb: "Even a *journey* of a thousand miles begins with a single step." • Some people fear to set out on the *journey* of self-discovery because they fear growing older.

P5 **+ ADV** *(frequent)* In his *journey* across North Africa to Cairo, he stopped at many zawiyas (sufi lodges). • As we made that long *journey* back to Manchester, our win and Archie's tales liberated our thoughts. • The *journey* round the garden continues via a traditional herbaceous border in tip-top condition. • "Prehistoric Life: The Rise of the Vertebrates" is a fully illustrated comprehensive *journey* through millions of years. • The Indian President made his *journey* to Buckingham Palace in a car. • I'm not sentimental about horses because they are there to race, but you could see he was a bit low on the *journey* home. • It took about six weeks of hard slog to make a covered wagon *journey* from one side of America to the other.

P6 **+ by N + ADV** Her husband will fly to Accra on Sunday and make the 473-mile *journey* by car to Wulugu.

A *journey* is 'the act of travelling from one place to another'.

Figure 3. Valency Dictionary of English: entry for the noun *journey*

thirteen "word fields" such as "locomotion" (the mover and the moved thing are identical), "transport" (someone or something is causing someone or something else to move), "change of ownership," "feelings" etc. In a short introductory passage about each field, the common characteristics such as the prevalent number of arguments or most common syntactic structures are described, then the words are classified into smaller groups according to further semantic criteria (eg. locomotion is divided into "general," "without auxiliary means: slow/quick," "with auxiliary means," "through water," "through air"). Finally, detailed entries of all the words in the word field are listed alphabetically; each entry comprises several etymologically related words. See Figure 4 for the entry of the words *reisen / einreisen / verreisen - Reisen / Reise* "to travel / to enter / to go on a journey - (the) travelling / (a) journey."

We consider this combined approach particularly fruitful: the division of words into word fields or semantic frames brings attention to the differences between the surface form of elements with the same or similar semantic role. On the other hand, the simultaneous presentation of etymologically related words shows the changes in argument structure (both as to the number, form and semantics) that take place during derivation.

Valency patterns and arguments

Obviously the most important part of a valency lexicon are the valency patterns.

Sometimes, the patterns are characterized purely by their surface form, as in the *Valency Dictionary of English* [Herbst et al., 2004] (see Figure 3). In this case, the different surface forms in patterns 2–5 in fact imply different semantic roles (the argument with preposition *by* is the means, with preposition *of* is the attribute, and the arguments with the preposition *between* as well as the adverbial expressions denote the direction or location in which the journey takes place). However, the user is expected to infer the information about the semantic roles of the arguments from the examples.

On the other hand, the NomBank lexicon presents the patterns as rolesets, which means they are purely semantically defined. For example, the roleset for the noun *journey* consists of two roles, the "traveller" and the "destination or path." How the roles are expressed in the surface structure of the sentences can only be seen from the annotated data.

We have already seen that in FrameNet, the arguments (here called the frame elements) are characterized by their semantic roles. The information about the surface form is, similarly as in NomBank, a result of the annotation process: the user may look up all combinations of frame elements that were found within the span of a single sentence during the annotation, together with their syntactic realizations such as "a prepositional phrase with preposition *in*," "definite null instantiation" (the argument

reisen / einreisen / verreisen – Reisen / Reise

Die Familie (a) reist an die Ostsee (b). Viele Polen (a) reisen nach Deutschland (b) ein. In diesem Jahr verreisen wir (a) ins Gebirge (b). Das Reisen in ferne Länder (b) ist meine liebste Freizeitbeschäftigung. Die Reise der Expedition (a) zum Basislager (b) verlief ohne Störungen.

1. 'allgemeine Fortbewegung auf ein Ziel', 'über eine größere Entfernung hinweg', 'mit einem Instrument (Verkehrsmittel)', 'von einem Ort an einen anderen', 'für eine längere Zeit'
2. a – Täter / Mensch /
   V: Sn; S: Sg/Sp (von)
   b – Richtung / Ding /
   V: Sp (von – über – nach, zu, in. . .); S: Sp (von – über – nach, zu. . .)
3. Die Verwandten / Nachbarn reisen / verreisen ans Meer. Sie reisen in den Süden. Immer mehr Touristen reisen nach Deutschland ein. Das Reisen mit dem Flugzeug nimmt zu. Die Reise zum Nordkap war ein einmaliges Erlebnis.

Figure 4. [Sommerfeldt and Schreiber, 1996], entry for *reisen* "to travel" and its etymologically related words

did not appear in the sentence, but its value was clear from the context).

However, there is more to the semantics of the arguments than just semantic roles. This can be seen in part 2 of the entry in the *Wörterbuch der Valenz etymologisch Vervandter Wörter* [Sommerfeldt and Schreiber, 1996] (Figure 4). In this case, the semantic roles are *Täter* "actor" and *Richtung* "direction," which corresponds to the frame elements Traveller and Direction/Goal/Source/Area in FrameNet or to the roles of traveller and destination-or-path in the NomBank roleset. Besides that, there is the semantic requirement on the argument: the actor has to be a human, the direction is an object. Another example of a lexicon which lists the semantic requirements is the *Slovník vazeb a spojení* [Svozilová et al., 2005] (Figure 1), where the indefinite pronouns *někdo* "someone" and *něco* "something" mark the difference between animate and inanimate nominal arguments.

Corpus evidence for patterns. Examples

In FrameNet as well as in the *Valency Dictionary of English* [Herbst et al., 2004], only patterns that were actually found during corpus annotation are listed. This has the disadvantage that some more complex patterns may be left out not because they are ungrammatical, but because of lack of corpus evidence.

Example sentences and sentence fragments play an integral part of any valency lexicon. The *Wörterbuch der Valenz etymologisch Vervandter Wörter* [Sommerfeldt and Schreiber, 1996] (Figure 4) offers an interesting solution of the dilemma between illustrative examples made up by the lexicographers and corpus evidence: the first set of examples directly under the headword are made up so that each word appears with its full valency potential (all arguments are expressed in the same sentence). Section 3 of the entry then gives natural examples.

The use of made up examples may also reflect the findings of Opavská [2002] that two thirds of general users prefer examples in the form of short phrases to full sentences taken from the corpus.

Conclusion

Among the approaches to the creation of valency lexicons, we find the following ideas and strategies particularly useful:

- the availability of the data in an electronic form, with tools which can be adjusted to the needs of various kinds of users;
- the organisation of the entries into semantically and linguistically motivated groups,
- the inclusion of both semantic roles and semantic requirements,
- the listing of the surface forms that the arguments may take,
- the availability of real-life examples to end users and of simplified or made up examples for the needs of foreign learners and users who prefer short, compact entries.

References

Boguraev, B., Briscoe, T., Carroll, J., Carter, D., and Grover, C., The derivation of a grammatically indexed lexicon from the Longman Dictionary of Contemporary English, in *Proceedings of the 25th annual meeting on Association for Computational Linguistics*, ACL '87, pp. 193–200, Association for Computational Linguistics, Stroudsburg, PA, USA, 1987.

Bräunling, P., Umfrage zum Thema Valenzwörterbücher, *Lexikographica*, *5*, 168–177, 1989.

Daneš, F., Hlavsa, Z., Jirsová, A., Macháčková, E., Prouzová, H., and Svozilová, N., *Větné vzorce v češtině, 2. opravené vydání*, no. 23 in Studie a práce lingvistické, Academia, 1987.

Filipec, J., Daneš, F., Machač, J., and Mejstřík, V., *Slovník spisovné češtiny pro školu a veřejnost, 2. upravené a doplněné vydání*, Academia, Praha, 1994.

FrameNet 1.5, URL `http://framenet.icsi.berkeley.edu/`, 2010.

Hajič, J. and Honetschläger, V., Annotation lexicons: Using the valency lexicon for tectogrammatical annotation, *Prague Bulletin of Mathematical Linguistics (PBML)*, 2003.

Helbig, G. and Schenkel, W., *Wörterbuch zur Valenz und Distribution deutscher Verben*, VEB Bibliographisches Institut, Leipzig, 1969.

Herbst, T. and Uhrig, P., Erlangen Valency Patternbank, URL `http://www.patternbank.uni-erlangen.de/cgi-bin/patternbank.cgi`, 2009.

Herbst, T., Heath, D., Roe, I. F., and Götz, D., *A valency dictionary of English: a corpus-based analysis of the complementation patterns of English verbs, nouns, and adjectives*, vol. 40 of *Topics in English Linguistics*, Walter de Gruyter, Berlin, New York, URL `http://books.google.com/books?id=HC6wUJeq6MUC&printsec=frontcover#v=onepage&q&f=false`, 2004.

Karlík, P., Hypotéza modifikované valenční teorie, *Slovo a slovesnost*, *61*, 170–189, 2000.

Kolářová, V., *Valence deverbativních substantiv v češtině (na materiálu substantiv s dativní valencí)*, Karolinum, Praha, 2010.

Lopatková, M., Žabokrtský, Z., and Kettnerová, V., *Valenční slovník českých sloves*, Karolinum, Praha, 2008.

NomBank 1.0, URL `http://nlp.cs.nyu.edu/meyers/NomBank.html`, 2008.

Opavská, Z., Postoje a preference uživatelů slovníku. K jednomu aspektu dotazníkovho průzkumu, in *Varia IX. Zborník materiálov zo IX. kolokvia mladých jazykovedcov*, pp. 87–96, Slovenská jazykovedná spoločnost pri SAV, Bratislava, URL `http://lexiko.ujc.cas.cz/index.php?page=14&idStudie=8`, 2002.

Panevová, J., Formy a funkce ve stavbě české věty, *Academia, Praha*, 1980.

PDT-Vallex, URL `http://ufal.mff.cuni.cz/pdt2.0/browse/visual-data/pdt-vallex/`, 2006.

Sgall, P., Valence jako jádro jazykového systému, *Slovo a slovesnost*, *67*, 163–179, 2006.

Sommerfeldt, K. E. and Schreiber, H., *Wörterbuch zur Valenz und Distribution deutscher Adjektive*, VEB Bibliographisches Institut, Leipzig, 1974.

Sommerfeldt, K. E. and Schreiber, H., *Wörterbuch zur Valenz und Distribution der Substantive*, VEB Bibliographisches Institut, Leipzig, 1977.

Sommerfeldt, K. E. and Schreiber, H., *Wörterbuch der Valenz etymologisch Vervandter Wörter: Verben, Adjektive, Substantive*, Max Niemeyer Verlag, Tübingen, 1996.

Svozilová, N., Prouzová, H., and Jirsová, A., *Slovník slovesných, substantivních a adjektivních vazeb a spojení*, Academia, Praha, 2005.

Tesnière, L., *Eléments de syntaxe structurale*, Libraire C. Klincksieck, 1959.

Čermáková, A., *Valence českých substantiv*, no. 9 in Studie z korpusové lingvistiky, Nakladatelství Lidové noviny, Praha, 2009.