Research in Language, 2025, vol. 23 (130-143) DOI: 10.18778/1731-7533.23.08

HOW TO COLLECT CROSS-LINGUISTIC DATA ON VALENCY ONLINE

JERZY GASZEWSKI
University of Lodz
jerzy.gaszewski@uni.lodz.pl

Abstract
The paper presents a method of online collection of linguistic data from informants, developed in a cross-linguistic project on verbal valency in the languages of Europe. The method is essentially an online questionnaire. The rules of the creation of its contents as well as specific problems related to data provision and data annotation are discussed in detail. The proposed method ensures the appropriate level of detail needed, enables the collection of a database sizeable enough and secures the comparability of the data.

Keywords: areal linguistics, linguistic methodology, linguistic typology, online questionnaire, valency

1. Introduction

The paper is meant to present a method that allows collecting comparable data on valency for multiple languages. The method was developed within a project[1] focused on areal patterns of marking of valency positions in European languages and could be used in other cross-linguistic studies.

The central concept of valency is that some words (*valency carriers*) open slots in their textual surrounding and require (both semantically and formally) lexical material to fill them (cf. Tesnière 1976: 239, Ágel & Fischer 2009: 230). For example, the verbs *break* and *wait* both open two valency positions. On a more abstract and language-neutral ground we can say that the verbal meanings

[1] The current work was supported by the National Science Centre, Poland (grant no, 2016/20/S/HS2/00285 under the guidance of the late Joanna Błaszczak, University of Wrocław). I am indebted to Tomasz Naskręt (Wrocław University of Science and Technology), who programmed the questionnaire app, and to Elaine Horyza (University of Wrocław), Martin Hinton (University of Lodz) and John Hodgetts (University of Lodz), who checked the input sentences.

BREAK and WAIT define two elements each (two *semantic microroles*, i.e. semantic roles defined with respect to a single predicate, cf. Van Valin 2005: 53-54, Hartmann et al. 2014: 464) they need to be complete semantically.[2] The formal requirements of the two verbs differ, *break* is a typical transitive one (cf. Blasi 2015), as shown in (1).

(1) English *break*: 1-NOM > V.subj[1] > 2-ACC[3]

‘sb/sth(arg 1) breaks sth(arg 2)’

With the verb *wait*, the second argument has to be marked with the preposition *for*, as shown in (2).

(2) English *wait*: 1-NOM > V.subj[1] > for 2-ACC

‘sb(arg 1) waits for sth/sb(arg 2)’

Unlike the English ones, the corresponding verbs in Croatian, *lomiti* ‘break’ and *čekati* ‘wait’ have the same pattern of marking, as shown in (3-4). The examples illustrate the kind of difference between languages that comparative studies on formal valency concentrate on, a difference in the assignment of verbal meanings to valency classes in the given language.

(3) Croatian *lomiti*: 1-NOM V.subj[1] 2-ACC

‘sb/sth(arg 1) breaks sth(arg 2)’

(4) Croatian *čekati*: 1-NOM V.subj[1] 2-ACC

‘sb(arg 1) waits for sth/sb(arg 2)’

Let us add that there can also be internal variation within languages. Hungarian and Ukrainian have two different markers for the second argument of the verbs meaning WAIT. In other words, they are like both English in (2) and Croatian in (4).

Valency is thus bound to individual words and the analysis must be fine-grained and start at the level of individual verbal meanings and microroles. This bottom-up approach provides a solid basis for generalizations within and across languages, and is common in present-day research on valency (e.g. Bickel et al. 2014, Malchukov & Comrie 2015, Gaszewski 2020, Say 2020, Luraghi et al. 2024).

[2] Verbal meanings and microroles are language-neutral comparative concepts in the sense of Haspelmath (2010), throughout the text they are marked with capital letters.

[3] The format of (1-4) follows the conventions of the ValPaL database (Hartmann et al. 2013), also used in the ongoing PaVeDa project (Luraghi et al. 2024). (1) is meant to cover all instances of English bivalent *break*. Order (shown by arrows) is the primary means of distinguishing arguments, case marking and subject indexing appear with some forms only.

To have an informative and “representative” (Hartmann et al. 2014, Say 2014) picture of valency, many valency carriers must be considered.

The paper presents a method of data collection for research of that kind, its rationale (section 2) and consecutive steps (section 3). We do not discuss any research questions or analyse any data as these are issues separate from data collection. However, references are made to the actual application of the method in the author’s project (the main source of examples) as well as the tool, an online questionnaire app developed as part of the CLARIN infrastructure and used in the project. Due to space limitations no thorough description of either the project or the app is included.

2. The choice of the method

Several features of data are desirable in a cross-linguistic study on valency. Due to the lexical specificity of valency, the database should be sizeable and detailed. The data should also be comparable, respective data points from different languages being as close semantically as possible. The method presented here meets all these criteria.

Information on valency in particular languages can be obtained from various types of sources: general dictionaries, valency dictionaries, corpora or via elicitation from speakers. If possible, one can combine different types of sources. This was to an extent the case in my project, but it was necessary to pick one kind of source as primary and the choice fell upon using native speaker informants. This to an extent follows from the organisational limitations of my project, which excluded a research team. Many studies on valency (e.g. Malchukov & Comrie 2015, Say 2020, Luraghi et al. 2024) rely on linguists as contributors, each providing data from one language. Data collection and work with the sources is then ‘decentralised’. The proposed method is a way to proceed with uniform data collection in multiple languages.

The use of ready language resources involves certain problems. They are not intractable, but become greatly exacerbated when dealing with a number of languages at once. First, the availability of the resources varies considerably from language to language, which impacts the quality and comparability of the data. Furthermore, any dictionary lacks some information and when working with a gallery of dictionaries for particular languages that necessarily differ in their size, coverage, lexicographic approach and time of publication, gaps related to relevant data points are a given. Even when dictionaries do provide the relevant valency frames, there remains doubt (sometimes considerable) if the verbs found separately for each language are indeed equivalent in meaning. As for corpora, establishing (rather than confirming) valency frames is not straightforward at all. Lexemes usually have multiple meanings and occur in numerous structures, only some of which may be relevant. Working through the noise in the results to recover the specific patterns is practically tantamount to creating valency dictionary entries on the basis of corpus data, a task very different to the one at hand.

By resorting to informants, we avoid such problems. The access to languages is on a par as native speakers are employed in all cases. Data collected in this way can also have gaps or dubious data points, but it is possible to consult an informant on the data they provided. There is thus a clear way to fill the gaps using the same source of data. As for comparability, if informants are faced with the same questionnaire content, response sentences as such are equivalent, careful analysis still has to identify those data points where the actual verbs are not, cf. the discussion of (7), (13) and (16) below.

There are still practical difficulties with relying on informants as the primary source of data. It is rather obvious that they should be native speakers, ideally raised and educated monolingually in their language. As a result, a number of individuals of very different backgrounds need to be found. In my project one good pool of potential collaborators were exchange students. However, the form of online questionnaire enables contact across borders and this opportunity was also used to reach out to informants. As for the level of involvement, the proposed method relies on intensive cooperation, but with a limited number of non-anonymous informants.

In sum, the proposed method is an alternative to more common practices in research on valency, developed out of necessity. Let us, however, point out that the database compiled in my project is of a size comparable to that of the other existing databases.

3. Data collection

3.1. Overview

The overall plan of action in the proposed method is as follows. First, the contents of the questionnaire needs to be prepared on the basis of a relevant selection of verbal meanings. Informants should provide sentences rather than abstract valency frames because this is how valency manifests itself in real usage. The preparation of the input is of course the task of the linguist(s) collecting the data. For practical reasons, the actual contents of the questionnaire used in my project was in English, the present-day lingua franca. It is also necessary to distinguish here between the whole valency frame (represented by a clause in the questionnaire) and its parts, microroles (argument phrases within the clause). Collected data are whole sentences, but the cross-linguistic comparison relates first of all to the level of microroles.

Second, the questionnaire is made available to informants who provide raw data for their respective languages. As has been said, all informants in my project were native speakers. The method relies mostly on native speaker competence so theoretically there could even be just one informant per language, but this was avoided. As a rule, there were at least two people covering each language so as not to rely on one person as a representative of a whole language community.

Third, the raw data returns to the linguist and undergoes annotation. While informants provide the language material, the assumption is that they do not analyse it in any way. It is the linguist who abstracts the patterns from the provided material. The questionnaire is designed so as to facilitate the annotation. In case of interpretative problems with particular data points, other sources may be consulted, including possible follow-up questions to informants.

The stages of the process described in the following subsections are simple actions, but they actually need to be simple in the face of the extent of the enterprise. The questionnaire in my project had 360 cues for 150 verbal meanings (cf. 70 in Malchukov et al. 2015) and yielded more than thirteen thousand response sentences in almost 20 languages. One of the goals in the project was to capture variation within languages,[4] of the kind mentioned for Hungarian and Ukrainian verbs meaning WAIT, cf. the discussion of (1-4) above. This results in more structures and contributes to the overall size of the database.

3.2. Cue preparation

Valency frames are templates for sentences. Thus, once verbal meanings are selected for the study, what cue sentences to include in the questionnaire seems straightforward. What matters for the research agenda are the grammatical elements, after all. Theoretically then, the actual words that fill the valency positions in the sentences used are of little importance. However, poor choices here might yield odd or irrelevant data points so one should try to predict and preclude such problems.

Most verbal meanings have semantic preferences, some (classes of) nouns combine with them naturally, creating meaningful sentences that could easily occur in natural language. It is advisable to pick such typical words for argument phrases. Thus, even though both (5) and (6) are realisations of the frame (2), their usefulness as cues differs.

(5) I'm waiting for my mum.

(6) The cucumber will have to wait for a better moment.

The verbal meaning WAIT typically has persons as the first argument, so that the given participant can be understood as experiencing the process of waiting. Impersonal nouns are possible, but less typical and less frequent. Thus, (5) is much more plausible in real communication than (6) and makes a better cue sentence. (7-8) show a less obvious example of problematic input.

[4] For the sake of simpler structure of the data, some valency databases (e.g. Say 2020) ignore such variation and allow only one grammatical marker for one microrole in one language.

(7) Have you added salt to the soup?.

(8) The government wants to add two new articles to the constitution.

The second and third argument of the predicate ADD should be of the same general kind and allow for a part-whole relationship between them. These conditions are successfully met in both (7) and (8). In terms of direct experience of informants, (7), referring to everyday life, is a better choice. However, many languages have a special verb that describes the action of adding salt to dishes, e.g. German *salzen*, Hungarian *sóz* or Polish *solić*. If such a verb surfaces in the results, it does match the meaning of the cue sentence, but not the valency frame wanted. Thus, a natural cue sentence may still hinder data collection.

It is unfortunately impossible to predict all such situations and some number of irrelevant datapoints is inevitable. A measure that may help is to use several cues for one verbal meaning. It is less likely that all of them will produce irrelevant results in any given language. Ideally, the sentences should differ considerably, but within the limits of real-life plausibility set by the verbal meaning, cf. (9-10).

(9) He mistook my mum for my sister.

(10) Some people confuse social policy with socialism.

The first argument of the predicate CONFUSE is typically human. Variation is possible still: (9) has an individuated subject, (10) an indefinite one. The further arguments have few semantic restrictions, but they need to be of the same broad class, in (9) they are human nouns, in (10) abstractions. Note also the different tenses used. Lastly, the verbal meaning can even be expressed by different synonymous verbs.

Of the very similar cues below, (12) may be preferable.

(11) The shop assistant accused the man of theft..

(12) The shop assistant accused the woman of theft.

Slavic languages (several of which were covered in my project) typically have a single form for the accusative and genitive with masculine personal nouns, but distinct ones for feminines. Since both cases may mark arguments, with the former group it is impossible to establish the marking just on the basis of the provided sentence. Thus, using feminine words for arguments that are likely to surface as objects may help.[5]

Measures such as the use of feminine nouns involve predicting actual categories in the responses before any data are even collected. The syncretism described above is irrelevant for non-Slavic languages, which have their own

[5] Dealing with syncretism is further discussed in Section 3.4.

peculiarities. Predicting and precluding all such problems in advance is impossible and such measures were in fact used only sparingly in my project.

If the input is provided in English, one should still bear in mind the accessibility and accuracy of the language. To ensure the former, one should obviously avoid too difficult words in the cues. Accuracy is best secured by consulting the input with native speakers of English, which was done in my project, cf. fn. 1.

We have discussed various aspects of the exact contents of cue sentences. The last but crucial element of preparation of the input is to describe the valency frames exemplified so that this information is present in the database. Table 1. shows this for (5). The format of description reflects the part-whole relationship between the valency frame (cue sentence) and microroles (phrases in the cue), inherent in the phenomenon of valency.

Table 1: An example of a ready cue with microroles assigned to phrases
Note: elements presented to informants are in bold

cue sentence	**I'm waiting for my mum.**
parts of the sentence	
microrole	part of cue sentence
valency carrier WAIT	**'m waiting**
WAIT-1	**I**
WAIT-2	**for my mum**

To be described in the database, individual microroles need labels, for example ones with numbers like WAIT-1 in Table 1. Other possible versions like "the experiencer of WAIT" or "waiting person" are more informative. In my project the succinct numerical format was used, but it necessarily came with a list explaining the labels.

3.2. Data provision

At this stage informants are presented with the contents of the questionnaire. They are asked to provide the versions of the cue sentences in their native languages. Then, they need to describe selected parts of each cue representing elements of the valency frame in question. Table 1. above shows such a division of a cue sentence, Table 2. presents another cue with Croatian data provided. Apart from the data itself, it is also advisable to allow the informants to provide comments on their sentences (a dedicated field in not shown in Table 2.).

Table 2: An example of provided Croatian data
Note: questionnaire input presented to informants in bold, data provided by the informant in bold italic

cue sentence	**He probably hits his victims with a baseball bat.**		
provided sentence	***On vjerojatno udara svoje žrtve s bejzbolskom palicom.***		
parts of the sentence			
microrole	part of cue sentence	part of provided sentence	basic form
valency carrier HIT	**hits**	***udara***	***udarati***
HIT-1	**he**	***on***	***on***
HIT-2	**his victims**	***svoje žrtve***	***svoj, žrtva***
HIT-3	**with a baseball bat**	***s bejzbolskom palicom***	***bejzbolska palica***

Providing exact sentence fragments is generally easy. It becomes problematic when arguments are dropped, which happens most often with pronoun subjects (not the case in Table 2). The information about the given argument is typically present on the verb so informants can give the verb as the relevant part even though it also represents the predicate. Confusing as it may feel to the informant, the annotation of such cases is not really problematic, cf. the discussion of Table 3. below.

Argument phrases include grammatical marking that identifies microroles. The markers are easiest to identify if one can compare the actual forms in the sentence with citation forms of the words used. Informants are thus asked to provide the 'basic form' of sentence parts. This is the only moment when the informant has to use conscious metalinguistic knowledge about their native language. Informants are told to use basic forms like the nominative for nouns or infinitives for verbs.[6] When a sentence part has several words, the basic form is less obvious. In Table 2., the phrase *s bejzbolskom palicom* 'with a baseball bat' is turned into the nominative as a whole. For the phrase *svoje žrtve* 'his (own) victims' both words are given separately in their citation forms. The adjective is then masculine: *svoj*, while the noun *žrtva* happens to be feminine and so they do not make a well-formed phrase. In both cases the comparison of the actual phrase to the citation form is possible and so the annotation is facilitated.

Ideal response sentences should be natural in the given language, but also comparable within the project. These two features need not always agree very well, as shown by (13-14).

[6] The informants are to follow general practice here, which may differ for individual languages, e.g. the citation form of the verb is 3SG in Hungarian, and 1SG in Greek.

(13) Ez az anyámé.

it DEF mum.1SG.POSS

'It is my mum's.'

(14) Ez az anyámhoz tartozik.

it DEF mum.1SG.ALL belong.3SG

'It belongs to my mum.' (= cue for 13-14)

My informant saw (13) as the most basic way to express the meaning of the cue in Hungarian. However, it does not have a verb meaning BELONG, the central element of the valency frame wanted. The less obvious (14) does include the right structural elements. Let us add that the verb *tartozik* exists in the language and it is not unnatural by itself, it just happens that an alternative structure is even more common and feels even more natural. In the light of the above, it is not clear to what extent naturalness of the responses should be emphasised in the instructions. We want to avoid slavish copying of the English structure, but also free paraphrases. Practice in my project shows that informants are generally aware of the structural parallelism (or lack of it) between their sentences and the cues. There are numerous comments notifying about departures from the structure of the English sentence for the sake of naturalness.

(13) is an example of how data provision may go astray. As has been said, it is inevitable that some portion of the collected data will be irrelevant. Another interesting examples is the cue (15), with the response (16), back-translated from Czech. Here the meaning of the predicate ENABLE was expressed by the preposition *díky* 'thanks to', totally reworking the structure of the sentence.

(15) In fact the accident enabled him to make a positive change in life.

(16) In the end his life changed thanks to the accident.

(13-16) reflect variation in language use. Here it was unwanted, but languages do have synonymous verbs, all of which may qualify as realisations of a given verbal meaning, cf. the verbs in (9-10) above. In my project, informants could provide several responses to one cue. With a diligent informant, one could get both sentences in (13-14) or (15-16) thanks to this option. Obligatory multiple responses would be counterproductive, though. For many cues informants feel there is one best way to render the meaning and extra versions would run the risk of being really forced, at the same time inflating the database.

Having several cues for one verbal meaning, illustrated above with (9-10), may yield variation in the data. (17-18), showing Finnish responses to (9-10), have two different verbs with different marking for the last argument.

(17) Hän luuli äitiäni siskokseni.

3SG confuse.PST.3SG mum.1SG sister.TRANS.1SG

‘He took my mum for my sister.’

(18) Jotkut sekoittavat sosiaalipolitiikan sosialismiin.

some.PL confuse-3PL social.policy.GEN socialism.ILL

‘Some people confuse social policy with socialism.’

Alternatively, the results from separate cue sentences may produce the same structure in the output, as in (19-20), Romanian results for (9-10). Such a situation is also acceptable on principle, we simply get stronger support for one structure.

(19) A confundat-o pe mama cu sora mea.

AUX.3SG confuse.PTCP-3SG.F.ACC DOM mum.DEF with sister my.F.DEF

‘He took my mum for my sister.’

(20) Unii oameni confundă politicile sociale cu socialismul.

some.PL human.PL confuse.3PL policy.PL.DEF social.PL with socialism.DEF

‘Some people confuse social policy with socialism.’

3.4. Data annotation

The purpose of this stage is to link grammatical markers with particular microroles in the database. Table 3. gives an example of annotated Hungarian data.

Table 3: An example of annotated Hungarian data.
Note: questionnaire input presented to informants in bold, data provided by the informant in bold italic, annotation in italic.

cue sentence	**I'll teach your dog a funny trick.**			
response	***Megtanítok a kutyádnak egy vicces trükköt.***			
parts of the sentence				
microrole	part of cue	part of response	basic form	grammatical marker
valency carrier TEACH	**'ll teach**	***megtanítok***	***megtanít***	
TEACH-1	**I**	***megtanítok***	***én***	*NOM*
TEACH-2	**your dog**	***a kutyádnak***	***kutya***	*DAT*
TEACH-3	**a funny trick**	***egy vicces trükköt***	***vicces, trükk***	*ACC*

Comparing the actual phrase in the sentence and the basic form in Table 3., we can notice that the phrase *a kutyádnak* has a number of morphological elements added to the plain *kutya* 'dog': the article *a*, the possessive suffix -*d* and the case suffix -*nak*. It is the case, the dative, that marks the syntactic function and the associated microrole. Likewise, we identify the accusative marker -*t* as the key morpheme in *egy vicces trükköt*. Both case labels are entered into the database. The situation is only slightly more complex with the first argument. There is no overt corresponding phrase, the argument is indexed on the verb. The suffix -*ok* unambiguously marks the 1SG subject and subjects in Hungarian have the nominative form. Any overt phrase in this position would have this case.

Table 4. shows that the collected responses are not ideal and annotation must be carried out carefully.

Table 4: An example of annotated Lithuanian data.
Note: presentation conventions as in Table 3.

cue sentence	**It concerned my parents directly.**			
response	***Tai tiesiogiai susiję su mano tėvais.***			
parts of the sentence				
microrole	part of cue	part of response	basic form	grammatical marker
valency carrier CONCERN	**concerned**	***susiję***	***susieti***	
CONCERN-1	**it**	***tai***	***tai***	*NOM*
CONCERN-2	**my parents**	***mano tėvais***	***mano, tėvas***	*su + INS*

The English sentence in Table 4. has a plain object phrase (with no overt marking) for the microrole CONCERN-2. Consequently, the questionnaire asks about the phrase *my parents*. The obvious equivalent fragment is the phrase *mano tėvais* 'my parents, INS' and this is what the informant gave. Alas, this is only a part of the argument phrase in Lithuanian as the response also includes the preposition *su*. The actual marker is *su* (and the instrumental). The problem in Table 4. can hardly be blamed on the informant, it results from the difference between English and Lithuanian. Let us only add that more avoidable mistakes also surface in the data.

Annotation shown so far relies solely on the response sentence as provided and basic grammar rules of the languages. Most phrases in the database can be dealt with in this way, but there are complications related to syncretism and differential argument marking. Both phenomena are involved in (19-20) above. The phrase *pe mama* in (19) by itself is difficult to classify, it could be the special differential marker of definite personal direct objects or the preposition *pe*, which some Romanian verbs govern. Both are expressed in the same way, this being syncretism of form, cf. also section 3.2. Yet, in the sentence (20), the parallel argument with an impersonal noun has no preposition. We can then easily conclude that the second argument of the verb *a confunda* is the direct object, marked by *pe* (and the accusative) or the plain accusative in line with the rules of differential object marking in the language. In sum, what counts for the research agenda are the abstract valency frames, which should cover all types of phrases that can appear in the given position, cf. fn. 3. The discussed phrases in (19-20) should then be annotated the same. The annotation departs here from actual surface marking but does so in order to reflect the overall marker variation across the realisations of the argument in question.

Syncretism is present in a number of languages. Typically, syncretic forms appear in some paradigms, but not in others. Since there are at least two sentences collected for any valency frame, the comparison of different sentences in the database usually suffices to resolve the ambiguity, as with the phrases in (19-20). There are naturally bound to be less fortunate cases of syncretism that cannot be clarified on the basis of the gathered data alone. It is only then that we need to turn to other sources like online searches, corpora, dictionaries, or follow-up questions to informants.

4. Final remarks

The presented procedure allows online collection of sizeable sets of comparable data at a necessary level of fine-grained detail. Apart from the method itself, software incorporating it, the questionnaire app, was also developed and all the stages of data collection in my project as described in Section 3 were performed with the use of the app. Various features postulated above as advisable practical details of the execution of the method are built into it, e.g. several cues for one verbal meaning, many responses to a single cue, or comments by informants.

The method could obviously be applied in other studies on valency. The method's potential is even greater as it essentially relies on two abstract relations. One is that between the whole and parts in the constructions analysed, the other is that between the language-neutral concept and language-specific instantiations. Table 5. illustrates these relations for the data in the last line in Table 4. In principle, any linguistic study that can be cast in such terms could make use of the method to collect data.

Table 5: Exemplification of relations between various levels in the data in the proposed method.

	language-neutral comparative concept	language-specific instantiation
verbal meaning (licensing whole valency frame)	CONCERN	susieti
microrole in the valency frame	CONCERN-2	su + INS

References

Ágel, Vilmos & Klaus Fischer. 2009. Dependency Grammar and Valency Theory. In Bernd Heine & Heiko Narrog (eds.), *The Oxford Handbook of Linguistic Analysis.* 223-256. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199544004.013.0010

Blasi, Damian E. 2015. Assessing transitivity prominence from a statistical perspective: A commentary on Martin Haspelmath's "Transitivity prominence". In Andrej Malchukov & Bernard Comrie (eds.), *Valency Classes in the World's Languages.* Vol. 1. 149-153. Berlin, Boston: de Gruyter Mouton. https://doi.org/10.1515/9783110338812-009

Bickel, Balthasar, Zakharko, Taras, Bierkandt, Lennart & Alena Witzlack-Makarevich. 2014. Semantic role clustering: An empirical assessment of semantic role types in non-default case assignment. *Studies in Language* 38 (3). 485–511. https://doi.org/10.1075/sl.38.3.03bic

Gaszewski, Jerzy. 2020. Does Verb Valency Pattern Areally in Central Europe? A First Look. In Luka Szucsich, Agnes Kim & Uliana Yazhinova (eds.), *Areal Convergence in Eastern Central European Languages and Beyond.* 13-53. Wien, Berlin: Peter Lang.

Hartmann, Iren, Haspelmath, Martin & Michael Cysouw. 2014. Identifying semantic role clusters and alignment types via microrole coexpression tendencies. *Studies in Language* 38 (3). 463-484. https://doi.org/10.1075/sl.38.3.02har

Hartmann, Iren, Haspelmath, Martin & Bradley Taylor (eds.) 2013. *Valency Patterns Leipzig*. Leipzig: Max Planck Institute for Evolutionary Anthropology. [Online] Available from: https://valpal.info, [Accessed on 7th May 2025].

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in cross-linguistic studies. *Language* 86(3). 663-687. https://doi.org/10.1353/lan.2010.0021

Luraghi, Silvia, Palmero Aprosio, Alessio, Zanchi, Chiara & Martina Giuliani. 2024. Introducing PaVeDa – Pavia Verbs Database: Valency Patterns and Pattern Comparison in Ancient Indo-European Languages. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*. 79–88. Torino: ELRA and ICCL.

Malchukov, Andrej & Bernard Comrie (eds.) 2015. *Valency Classes in the World's Languages*, 2 Vols. Berlin, Boston: de Gruyter Mouton.

Malchukov, Andrej & the Leipzig Valency Classes Project team. 2015. Leipzig Questionnaire on valency classes. In Andrej Malchukov & Bernard Comrie (eds.), *Valency Classes in the World's Languages.* Vol. 1. 27-39. Berlin, Boston: de Gruyter Mouton. https://doi.org/10.1515/9783110338812-005

Say, Sergey. 2014. Bivalent Verb Classes in the Languages of Europe. A Quantitative Typological Study. *Language Dynamics and Change* 4(1). 116–166. https://doi.org/10.1163/22105832-00401003

Say, Sergey (ed.) 2020. *BivalTyp: Typological database of bivalent verbs and their encoding frames*. St. Petersburg: Institute for Linguistic Studies. [Online] Available from: https://www.bivaltyp.info. [Accessed on 13th May 2025].

Tesnière, Lucien. 1959/1976. *Éléments de syntaxe structurale.* Paris: Klincksieck.

Van Valin, Robert D., Jr. 2005. *Exploring the syntax-semantics interface*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511610578.001