Proposal for a Master's thesis

Typology of verbal valency systems: A quantitative study based on Universal Dependencies

Siyu Tao

Draft as of December 16, 2022

Abstract

TODO

1 Introduction

Universal Dependencies (UD) treebanks, a multilingual collection of dependency treebanks based on a shared, cross-lingually consistent annotation scheme (Nivre et al., 2020) and covering 138 languages with 243 treebanks in its most recent `v2.11` release (Zeman, Nivre, et al., 2022), has enabled significant advances in the development of multilingual dependency parsers and other NLP technologies (Zeman, Hajič, et al., 2018; Zeman, Popel, et al., 2017). This proposed thesis will explore their potential in typology research through a cross-lingual quantitative study of verbal valency systems.

The starting point of this study is the assumption, consistent with those behind Levin (1993) and other work on verb classes, that the syntactic behavior of verbs are at least in part determined by their lexical semantics, and that, as such, verb classes based on their syntactic distribution should be semantically coherent as well. This study will test this assumption computationally by performing clustering experiments on a subset of UD treebanks in order to explore whether the UD annotations support an automated induction of the valency frames in a language and whether

verb classes can be further inducted based on the distribution of verbs across the valency frames. In the process of the experiments, factors that influence the outcome of clustering, particularly with respect to data quantity and quality, as well as typological features of languages, will be examined. The results of these clustering experiments will then, in combination with a computationally derived cross-lingual lexicon, support typological investigations into possible universals in the organization of verbal lexicon.

This proposal itself is organized as follows: relevant literature on valency and dependency grammar is surveyed in §2 to provide the theoretical background; §4 lays out the proposed data sources and methodology to be used in the study; §3 presents the aims of the study and the key research questions; §5 reports the preliminary results of a pilot study; §6 provides a work plan and tentative timeline for completing the thesis; §7 concludes the proposal.

2 Background and related work

2.1 Valency and valency phenomena

In chemistry, *valency*, or *valence*, refers to the combining power of an atom or radical. The valency of any atom can be measured by the number of hydrogen atoms that it can combine with or displace in a chemical compound (Law and Rennie, 2020). This same term has been used in linguistics to similar effect and refers to the combining power of a word, primarily a verb or other predicate, with other words or elements of the sentence.

Lucien Tesnière is generally credited with introducing the term valency to linguistics with his syntactic theory of valency and dependence, as presented in the posthumously published *Éléments de syntaxe structurale* (1959; English translation 2015).[1] In another of Tesnière's metaphors, each verbal node, being the center of sentence structure, is not unlike a "theatrical performance" with the verb expressing the process and the nouns being the *actants* (what we would now call *arguments*) in this performance. Just like how atoms of different elements allow for a greater or lesser number of bonds, different verbs can combine with a greater or lesser number of actants, i.e., their valency.

While the term valency is borrowed into linguistics from chemistry, the study of the phenomena which are covered by or otherwise overlap with valency has a much longer tradition, dating to the early beginnings of linguistics from the kāraka concept

[1] It should be noted that while Tesnière is rightly credited with the introduction of a theory of linguistic valency, the metaphor of valency itself has made appearances as early as in Peirce (1897), among others (Przepiórkowski, 2018).

of semantic relation between verb and noun (Ganeri, 2011) in Pāṇinian grammar to modern case grammar (Fillmore, 1968).

Most linguistic theories assert the centrality of the verb in determining either or both the syntactic and semantic structure of a sentence, corroborated also by psycholinguistic evidence (Healy and Miller, 1970). This places valency and the issues of *argument structure* squarely at the center of the inquiry into the interface between lexical semantics and syntax.

In generative grammar, the syntactic valency of a verb is treated under a similar notion of *subcategorization* (Chomsky, 1965). As an example, a transitive verb must be followed by a direct object, whereas intransitive verb cannot, as such transitive and intransitive verbs form subcategories of the category verb. Verbs are therefore assigned to *subcategorization frames* which specifies the number and type of complements, i.e., objects and obliques, as well as of the subject in later theories, that the verb can be subcategorized for. In addition to being syntactically driven, it is worth noting that the subcategorization frames are considered as part of the lexical entry of the verb. Later work in generative grammar, in particular Jackendoff (1972, 1987, 1992), following Katz and Fodor (1963) and Gruber (1962), further develops a theory of thematic relations and posits that argument structure serves as the interface between syntactic and thematic structures.

As compared to broader distinctions such as those made between transitive and intransitive verbs, the verb classes in Levin (1993) provide a drastically more fine-grained categorization of verbs based on their syntactic behavior. Guided by the assumption that the syntactic behavior of verbs are determined semantically, Levin reasons that patterning together classes of verbs based on their diathesis alternations should result in semantically coherent verb classes. Levin's work has been highly influential both in the development of valency theory and in computational approaches to lexical semantics, the VerbNet (Kipper et al., 2006, 2008; Kipper-Schuler, 2005) being a prominent example of such projects, extending the Levin verb classes into a computational lexicon that links with other resources such as WordNet (Fellbaum, 1998; Miller, 1995), PropBank (Kingsbury and Palmer, 2002) and FrameNet (C. F. Baker, Fillmore, et al., 1998; Fillmore and C. Baker, 2015).

In terms of their theoretical foundations, however, FrameNet differs from VerbNet in that it derives from a different line of research that stems from Charles Fillmore's frame semantics (Fillmore, 1977a,b, 1982), which in turn traces to his earlier work on case grammar (Fillmore, 1968, 1970). While they are computationally interoperable to some extent, there remains a key conceptual distinction made in frame semantics Fillmore (1968), namely the *frames*-driven analysis of argument encoding. While the verbal lexicon continues to play a role in placing selectional restrictions on the frames in which a given verb can be found in, the frames are themselves said to have semantics through their grouping of frame elements, which are similar to thematic roles

but local to their specific frames. The frame semantics approach is consolidated by further development in construction grammar where the frame, or *argument structure constructions* (Goldberg, 1992, 1995), is viewed as a construction. According to construction grammar theories, a construction is considered distinct or autonomous if one or more of its properties are not predictable from other grammatical units.

q: I am still considering how the conceptual distinction of valency information as part of the lexical entry of the verb vs. valency information as part of a frame construction affects their predictions in terms of typological evidence. Would, say, a hypothetical result where cross-lingual patterns in the organization of verbal lexicon can be observed, but the frames themselves are more language-specific, necessarily be evidence against frame as a distinct / autonomous construction? need to think more about frame semantics and typology

2.2 Typological perspectives on valency and dependency

In terms of their mathematical foundations, dependency grammars, based on the notion of dependencies, can be considered in contrast with constituency grammars, based on the notion of substitution (Stabler, 2019). However, even most iterations of generative grammar theories, which are primarily constituency-based, adopt some notion of head-dependency (such as X-bar theory). de Marneffe and Nivre (2019) cites the easiness of generalization across languages, its operationalization of human sentence processing facts, and the transparency and simplicity of representation as reasons why dependency-based representations have become increasingly widely adopted in linguistic theory and even more so in NLP. Here, however, the focus will be on why dependency representations lend themselves to cross-linguistic contrastive studies.

It is perhaps not surprising that, in addition to introducing a first valency theory, Tesnière (1959) also introduced the notion of dependency to modern linguistic theory. As Tesnière (1959) introduces his theory of valency and dependency, the cross-lingual differences in the structure were already in focus. Tesnière describes the process of *metataxis*, by which syntactic structures of one language is "translated" to those of another. This implies, as this study argues explicitly, that the primary typological interest in valency is in comparing the mismatches.

What would a universal look like: Tsunoda (1981, 1985, 2015) proposes a hierarchy of verbs

For example, Say (2014) rejects the equating of minor valency classes cross-lingually and study how the individual verbs care grouped into valency.

Computational work on semantic frame induction / verb classes: Abend et al.

(2009), Basili et al. (1993), Bickel et al. (2014), Dowty (1991), Fellbaum (1998), Fillmore (1968), Fürstenau and Rambow (2012), Kipper et al. (2008), Kipper-Schuler (2005), Korhonen et al. (2006), Levin (2015), Majewska, Collins, et al. (2021), Majewska, McCarthy, et al. (2018), Majewska, Vulić, et al. (2020), Miller (1995), Miller et al. (1990), Navarretta (2000), Palmer et al. (2005), Say (2014), Sayeed et al. (2018), Schulte im Walde (2003, 2006), Schulte im Walde and Brew (2002), Snider and Diab (2006), Sun and Korhonen (2009), Sun, Korhonen, and Krymolowski (2008), Sun, McCarthy, et al. (2013), Titov and Klementiev (2012), Watanabe et al. (2010), and Yamada et al. (2021)

pick a few examples to introduce, e.g., one each for semantics- and synatx-based approaches and omit the rest / relegate to citations only

C. F. Baker and Lorenzi (2020) and Ellsworth et al. (2021) addresses FrameNet and typology

Croft et al. (2017) address UD specifically and propose more typologically-informed modifications to the dependency annotations of UD.

3 Research questions

The aim of this thesis study is twofold: the first is exploratory and computational, namely whether the existing computational resources based on dependency grammar can be effectively utilized in quantitative typology; the second is investigative and typological, whether a corpus-based study of valency features reveals patterns or universals in how languages organize their valency systems.

Levin (1993) observes in her study of English verb classes that

> Distinctions induced by diathesis alternations help to provide insights into verb meaning, and more generally into the organization of the English verb lexicon, that might not otherwise be apparent, bringing out unexpected similarities and differences between verb. (p.15)

A typological study then aims to examine these linguistic universals.

Will we see cross-lingual patterns or universals in how verb classes aggregate and within each cross-lingual clusters, different strategies being used to encode the verb classes. If a semantic universal exists for different levels of transitivity for example, this should show up in the verb classes.

Different strategies for the same construction, e.g. adpositions and case markings

While the valency frames can themselves be considered a component of the syntactic structure of sentences, it is nevertheless clear that they are primarily a feature of the verbal lexicon. While there are certainly exceptions to the rule (such as the non-canonical use of verbs), it is generally possible to determine the possible valency frames given the verb. Cross-lingually, this means the comparison of the dis-

tribution of verbs across different verb classes and valency frames allows us to test possible universals regarding the organization of the verbal lexicon. The object of cross-lingual comparison therefore is crucially *not* the valency frames or verb classes themselves, but the organization of the frames and classes.

The difficulty in a finite categorical classification of valency class systems can thus be overcome through statistical, information theoretic methods.

4 Data and methodology

This section presents the proposed data sources and methodology of the thesis. §4.1 introduces the Universal Dependencies treebanks as well as additional resources that will be used as reference and validation in this study. The rest of the section, §4.2-4.5, presents the main computational methods to be used in the thesis.

4.1 Data sources

Universal Dependencies (UD) is designed to be a cross-linguistically consistent system for annotating morphosyntactic information within a dependency grammar framework (de Marneffe, Manning, et al., 2021). The UD treebanks (Zeman, Nivre, et al., 2022) is the collection of cross-lingual treebanks annotated in the UD framework by an open community of more than 300 contributors.

See 1 for a table of languages available in UD v2.5 (to be updated for v2.11)

A subset of the UD treebanks, the Parallel Universal Dependencies (PUD) treebanks were originally developed for the CoNLL-2017 Shared Task (Zeman, Popel, et al., 2017) and include 1000 sentences in 18 languages that were randomly picked from newswire and Wikipedia and annotated according to UD v2 guidelines. The 18 languages are English, German, French, Italian, Spanish, Arabic, Hindi, Chinese, Indonesian, Japanese, Korean, Portuguese, Russian, Thai, Turkish, Czech, Finnish and Swedish. Of the sentences, 750 were originally English, while the remaining 250 sentences come from German, French, Italian and Spanish texts and translated to other languages through English. While facing obvious limitation in terms of language coverage, corpus size, and possible artifacts due to the so-called "translationese", parallel corpora allow for cross-lingual comparison with a smaller data size and will also be considered in this thesis.

In additional to the main data source of UD treebanks, additional resources will be used in the study as reference and to perform validation and evaluation of the intermediate results. As an example, the valency frames and verb classes as induced from the UD treebanks will be validated, where possible, against the expert-annotated data from **ValPaL**(Hartmann et al., 2013).

Language	#	Sents	Words	Language	#	Sents	Words	Language	#	Sents	Words
Afrikaans	1	1,934	49,276	German	4	208,440	3,753,947	Old Russian	2	17,548	168,522
Akkadian	1	101	1,852	Gothic	1	5,401	55,336	Persian	1	5,997	152,920
Amharic	1	1,074	10,010	Greek	1	2,521	63,441	Polish	3	40,398	499,392
Ancient Greek	2	30,999	416,988	Hebrew	1	6,216	161,417	Portuguese	3	22,443	570,543
Arabic	3	28,402	1,042,024	Hindi	2	17,647	375,533	Romanian	3	25,858	551,932
Armenian	1	2502	52630	Hindi English	1	1,898	26,909	Russian	4	71,183	1,262,206
Assyrian	1	57	453	Hungarian	1	1,800	42,032	Sanskrit	1	230	1,843
Bambara	1	1,026	13,823	Indonesian	2	6,593	141,823	Scottish Gaelic	1	2,193	42,848
Basque	1	8,993	121,443	Irish	1	1,763	40,572	Serbian	1	4,384	97,673
Belarusian	1	637	13,325	Italian	6	35,481	811,522	Skolt Sámi	1	36	321
Bhojpuri	1	254	4,881	Japanese	4	67,117	1,498,560	Slovak	1	10,604	106,043
Breton	1	888	10,054	Karelian	1	228	3,094	Slovenian	2	11,188	170,158
Bulgarian	1	11,138	156,149	Kazakh	1	1,078	10,536	Spanish	3	34,693	1,004,443
Buryat	1	927	10,185	Komi Permyak	1	49	399	Swedish	3	12,269	206,855
Cantonese	1	1,004	13,918	Komi Zyrian	2	327	3,463	Swedish Sign Language	1	203	1,610
Catalan	1	16,678	531,971	Korean	3	34,702	446,996	Swiss German	1	100	1,444
Chinese	5	12,449	285,127	Kurmanji	1	754	1,0260	Tagalog	1	55	292
Classical Chinese	1	15,115	74,770	Latin	3	41,695	582,336	Tamil	1	600	9,581
Coptic	1	1,575	40,034	Latvian	1	13,643	219,955	Telugu	1	1,328	6,465
Croatian	1	9,010	199,409	Lithuanian	2	3,905	75,403	Thai	1	1,000	22,322
Czech	5	127,507	2,222,163	Livvi	1	125	1,632	Turkish	3	9,437	91,626
Danish	1	5,512	100,733	Maltese	1	2,074	44,162	Ukrainian	1	7,060	122,091
Dutch	2	20,916	306,503	Marathi	1	466	3,849	Upper Sorbian	1	646	11,196
English	7	35,791	620,509	Mbyá Guaraní	2	1,144	13,089	Urdu	1	5,130	138,077
Erzya	1	1,550	15,790	Moksha	1	65	561	Uyghur	1	3,456	40,236
Estonian	2	32,634	465,015	Naija	1	948	12,863	Vietnamese	1	3,000	43,754
Faroese	1	1,208	10,002	North Sámi	1	3,122	26,845	Warlpiri	1	55	314
Finnish	3	34,859	377,619	Norwegian	3	42,869	666,984	Welsh	1	956	16,989
French	7	45,074	1,157,171	Old Church Slavonic	1	6,338	57,563	Wolof	1	2,107	44,258
Galician	2	4,993	164,385	Old French	1	17,678	170,741	Yoruba	1	100	2,664

Table 1: Languages in UD v2.5 with number of treebanks (#), sentences (Sents) and words (Words).

more details on valpal and other possible data

4.2 Verb valency features

A list of binary slot features

4.3 Clustering

The verb class induction from UD data can be broken down into a three-step process.

Step 1: Coding Feature Selection In the first step, the specific uses of verbs are abstracted through a feature selection process, where only features that are relevant to valency frame encoding are included. A verb can therefore be represented by a list of its features. This is in order to focus on whether semantically coherent verb

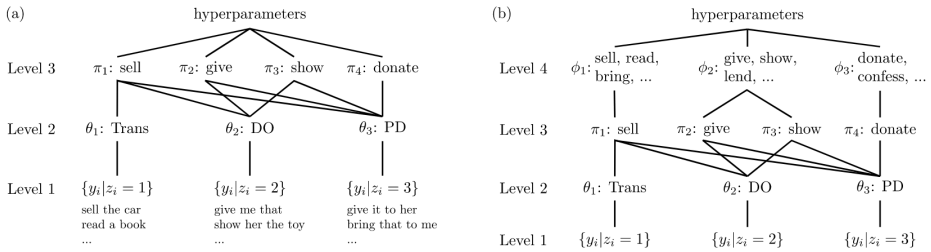

Figure 1: (a) Model 1, a Hierarchical Dirichlet Process applied to learning verb argument structure constructions. (b) Model 2, an extension of Model 1 to learn verb alternation classes.

classes can be induced on valency frame information. In selecting the features, cross-lingual differences in valency frame coding will be taken into account, e.g. whether a language uses cases or word order to encode valency frame information.

Examples from EN, DE, ZH

Step 2: Valency Frame Induction Given the selected features, the valency frame are then derived using unsupervised clustering algorithms such as k-means (Macqueen, 1967), which iteratively updates the center of cluster which is represented by the center of data points, until the criteria for convergence is met. Other clustering algorithms should also be investigated (Xu and Tian, 2015).

what distance measure to use?

Step 3: Verb Class Induction Approach similar to Parisien and Stevenson (2010), where a Hierarchical Dirichlet process is extended to account for diathesis alternations. Each verb will be represented by its distribution over the valency frames of the language, which are then clustered in a similar process as step 2.

4.4 Cross-lingual verb sense alignment

A cross-lingual aligned list of counterpart verbs will be needed to compare the verb classes and valency frames. The easiest way to do this is likely through existing cross-lingual word lists such as LanguageNet, part of the PanLex project. http://uakari.ling.washington.edu/languagenet/available/

Alternatively, lexicon induction from cross-lingual word embeddings and other NLP methods may also be considered.

4.5 Information theory

Complexity and point-wise mutual information, like in Say (2014). Complexity measure:

Point-wise mutual information (PMI)

5 Preliminary results

This section will present preliminary results of a small pilot study where verb clustering is done on English and German treebanks from the Parallel UD dataset and the results are compared against the ValPaL database and manually inspected. (work-in-progress)

6 Work plan

This section presents a work plan and the tentative timeline for the thesis project.

Work plan: (1) language selection (2) feature selection (3) clustering (4) verb alignment list (5) information theoretical metrics (6) linguistic analysis

1->2->3;3->1;4->1;3+4->5->6

This thesis study is intended to be completed within roughly three months after the submission of this proposal even though the maximum time available for completing it remains six months.

Given time constraints, an iterative process is envisioned and priority will be put on completing a functional pipeline of the computational part already in the first month of the work, i.e. January, allowing for more flexibility later in the project. Iterative improvements will then be made upon the code and methodological modifications tested. The primary experimental parts should conclude by the end of second month to allow for time needed for the write-up and revisions in the final month.

7 Conclusion

This thesis aims to contribute to the study of valency by using corpus linguistic approaches. The aim of study is not dissimilar to that of projects such as ValPaL or studies by Say (2014) but instead of focusing on a limited set of samples, the aim is to make the best use of the available cross-lingual corpora while still basing the study in a consistent theoretical framework. Exploring linguistic universals regarding the organization of verb lexicon and valency systems contributes to the overall project of typological studies, as well as shedding new light on the development of

valency theories, which have thus far been driven by introspective studies of single languages or contrastive studies of two languages.

In the process of doing so, computational methods and their applicability will also be explored.

Bibliography

Abend, Omri, Roi Reichart, and Ari Rappoport (Aug. 2009). "Unsupervised Argument Identification for Semantic Role Labeling". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. ACL-IJCNLP 2009. Suntec, Singapore: Association for Computational Linguistics, pp. 28–36.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe (Aug. 1998). "The Berkeley FrameNet Project". In: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*. ACL 1998. Montreal, Quebec, Canada: Association for Computational Linguistics, pp. 86–90. DOI: 10.3115/980845.980860. URL: https://aclanthology.org/P98-1013 (visited on 12/16/2022).

Baker, Collin F. and Arthur Lorenzi (May 2020). "Exploring Crosslinguistic Frame Alignment". In: *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*. Marseille, France: European Language Resources Association, pp. 77–84. ISBN: 979-10-95546-58-0. URL: https://aclanthology.org/2020.framenet-1.11 (visited on 09/25/2022).

Basili, Roberto, Maria Pazienza, and Paola Velardi (1993). "Hierarchical Clustering of Verbs". In: *Acquisition of Lexical Knowledge from Text*. URL: https://aclanthology.org/W93-0107 (visited on 11/23/2022).

Bickel, Balthasar, Taras Zakharko, Lennart Bierkandt, and Alena Witzlack-Makarevich (Jan. 1, 2014). "Semantic Role Clustering: An Empirical Assessment of Semantic Role Types in Non-Default Case Assignment". In: *Studies in Language* 38.3, pp. 485–511. ISSN: 0378-4177, 1569-9978. DOI: 10.1075/sl.38.3.03bic. URL: https://www.jbe-platform.com/content/journals/10.1075/sl.38.3.03bic (visited on 09/25/2022).

Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: The MIT Press.

Croft, William, Dawn Nordquist, Katherine Looney, and Michael Regan (2017). "Linguistic Typology Meets Universal Dependencies". In: *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15), Bloomington, IN, USA, January 20-21, 2017*. Ed. by Markus Dickinson, Jan Hajic, Sandra Kübler, and Adam Przepiórkowski. Vol. 1779. CEUR Workshop Proceedings. CEUR-WS.org, pp. 63–75. URL: http://ceur-ws.org/Vol-1779/05croft.pdf (visited on 08/22/2022).

De Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman (July 13, 2021). "Universal Dependencies". In: *Computational Linguistics* 47.2, pp. 255–308. ISSN: 0891-2017. DOI: 10.1162/coli_a_00402. URL: https://doi.org/10.1162/coli_a_00402 (visited on 08/22/2022).

De Marneffe, Marie-Catherine and Joakim Nivre (2019). "Dependency Grammar". In: *Annual Review of Linguistics* 5.1, pp. 197–218. DOI: 10.1146/annurev-linguistics-011718-011842. eprint: https://doi.org/10.1146/annurev-linguistics-011718-011842. URL: https://doi.org/10.1146/annurev-linguistics-011718-011842.

Dowty, David (1991). "Thematic Proto-Roles and Argument Selection". In: *Language* 67.3, pp. 547–619. ISSN: 0097-8507. DOI: 10.2307/415037. JSTOR: 415037.

Ellsworth, Michael, Collin Baker, and Miriam R. L. Petruck (2021). "FrameNet and Typology". In: *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*. Proceedings of the Third Workshop on Computational Typology and Multilingual NLP. Online: Association for Computational Linguistics, pp. 61–66. DOI: 10.18653/v1/2021.sigtyp-1.6. URL: https://www.aclweb.org/anthology/2021.sigtyp-1.6 (visited on 11/01/2021).

Fellbaum, Christiane, ed. (1998). *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. Cambridge, Mass: MIT Press. 423 pp. ISBN: 978-0-262-06197-1.

Fillmore, Charles J. (1968). "The Case for Case". In: *Universals in Linguistic Theory*. Ed. by Emmon Bach and Robert T. Harms. New York: Rinehart and Winston, pp. 21–119.

– (1970). "The Grammar of HITTING and BREAKING". In: URL: https://www1.icsi.berkeley.edu/pubs/ai/ICSI_grammarofhitting12.pdf (visited on 09/18/2022).

– (1977a). "Scenes-and-Frames Semantics". In: *Linguistic Structure Processing*. Ed. by Antonio Zampolli. Fundamental Studies in Computer Science. Amsterdam: North Holland Publishing Company, pp. 55–82. ISBN: 978-0-444-85017-1.

– (Dec. 20, 1977b). "The Case for Case Reopened". In: *Grammatical Relations*. Ed. by Peter Cole and Jerrold M. Sadock. BRILL, pp. 59–81. ISBN: 978-90-04-36886-6 978-90-04-36852-1. DOI: 10.1163/9789004368866_005. URL: https://brill.com/view/book/edcoll/9789004368866/BP000005.xml (visited on 11/24/2022).

– (1982). "Frame Semantics". In: *Linguistics in the Morning Calm*. Ed. by Linguistic Society of Korea. Seoul, Korea: Hanshin Publishing Company, pp. 111–137.

Fillmore, Charles J. and Collin Baker (Jan. 1, 2015). "A Frames Approach to Semantic Analysis". In: *The Oxford Handbook of Linguistic Analysis*. Ed. by Bernd Heine and Heiko Narrog. Oxford University Press. ISBN: 978-0-19-967707-8. DOI: 10.1093/oxfordhb/9780199677078.013.0013. URL: https://academic.oup.com/edited-volume/28050/chapter/211991006 (visited on 11/23/2022).

Fürstenau, Hagen and Owen Rambow (2012). "Unsupervised Induction of a Syntax-Semantics Lexicon Using Iterative Refinement". In: **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the Main*

Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). *SEM/SemEval 2012. Montréal, Canada: Association for Computational Linguistics, pp. 180–188. URL: https://www.aclweb.org/anthology/S12-1026 (visited on 09/18/2019).

Ganeri, Jonardon (July 21, 2011). "Kāraka: Meanings in Composition". In: *Artha: Meaning*. Oxford University Press. ISBN: 978-0-19-807413-7. DOI: 10.1093/acprof:oso/9780198074137.001.0001. URL: http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198074137.001.0001/acprof-9780198074137 (visited on 11/22/2022).

Goldberg, Adele E. (Jan. 1, 1992). "The Inherent Semantics of Argument Structure: The Case of the English Ditransitive Construction". In: 3.1, pp. 37–74. ISSN: 1613-3641. DOI: 10.1515/cogl.1992.3.1.37. URL: https://www.degruyter.com/document/doi/10.1515/cogl.1992.3.1.37/html (visited on 11/24/2022).

– (Mar. 1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Cognitive Theory of Language and Culture Series. Chicago, IL: University of Chicago Press. 271 pp. ISBN: 978-0-226-30086-3. URL: https://press.uchicago.edu/ucp/books/book/chicago/C/bo3683810.html (visited on 11/23/2022).

Gruber, Jeffrey S. (1962). "Studies in Lexical Relations". PhD thesis. Massachusetts Institute of Technology.

Hartmann, Iren, Martin Haspelmath, and Bradley Taylor (2013). *The Valency Patterns Leipzig Online Database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://valpal.info/.

Healy, Alice F. and George A. Miller (June 1, 1970). "The Verb as the Main Determinant of Sentence Meaning". In: *Psychonomic Science* 20.6, pp. 372–372. ISSN: 2197-9952. DOI: 10.3758/BF03335697. URL: https://doi.org/10.3758/BF03335697 (visited on 11/17/2022).

Jackendoff, Ray S. (1972). *Semantic Interpretation in Generative Grammar*. Studies in Linguistic Series. Cambridge, MA: The MIT Press. 400 pp. ISBN: 978-0-262-10013-7. URL: https://bac-lac.on.worldcat.org/oclc/323868 (visited on 11/22/2022).

– (1987). "The Status of Thematic Relations in Linguistic Theory". In: *Linguistic Inquiry* 18.3, pp. 369–411. ISSN: 0024-3892. JSTOR: 4178548.

– (Apr. 22, 1992). *Semantic Structures*. MIT Press. 340 pp. ISBN: 978-0-262-60020-0. Google Books: 7wbYlHis6OEC.

Katz, Jerrold J. and Jerry A. Fodor (1963). "The Structure of a Semantic Theory". In: *Language* 39.2, pp. 170–210. ISSN: 0097-8507. DOI: 10.2307/411200. JSTOR: 411200.

Kingsbury, Paul and Martha Palmer (May 2002). "From TreeBank to PropBank". In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. LREC 2002. Las Palmas, Canary Islands - Spain: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2002/pdf/283.pdf (visited on 12/16/2022).

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer (May 2006). "Extending VerbNet with Novel Verb Classes". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. LREC 2006. Genoa, Italy: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/468_pdf.pdf (visited on 11/23/2022).

– (Mar. 1, 2008). "A Large-Scale Classification of English Verbs". In: *Language Resources and Evaluation* 42.1, pp. 21–40. ISSN: 1572-8412. DOI: 10.1007/s10579-007-9048-2. URL: https://doi.org/10.1007/s10579-007-9048-2 (visited on 09/18/2019).

Kipper-Schuler, Karin (2005). "VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon". PhD thesis. USA: University of Pennsylvania.

Korhonen, Anna, Yuval Krymolowski, and Ted Briscoe (May 2006). "A Large Subcategorization Lexicon for Natural Language Processing Applications". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. LREC 2006. Genoa, Italy: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/558_pdf.pdf (visited on 11/23/2022).

Valency (Valence) (Mar. 19, 2020). In: *A Dictionary of Chemistry*. Ed. by Jonathan Law and Richard Rennie. Oxford University Press. ISBN: 978-0-19-884122-7. URL: https://www.oxfordreference.com/view/10.1093/acref/9780198841227.001.0001/acref-9780198841227 (visited on 11/22/2022).

Levin, Beth (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, Ill.: Univ. of Chicago Press. ISBN: 978-0-226-47532-5 978-0-226-47533-2.

– (2015). "Semantics and Pragmatics of Argument Alternations". In: *Annual Review of Linguistics* 1.1, pp. 63–83. DOI: 10.1146/annurev-linguist-030514-125141. URL: https://doi.org/10.1146/annurev-linguist-030514-125141 (visited on 09/12/2019).

Macqueen, J (1967). "Some Methods for Classification and Analysis of Multivariate Observations". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

Majewska, Olga, Charlotte Collins, Simon Baker, Jari Björne, Susan Windisch Brown, Anna Korhonen, and Martha Palmer (July 15, 2021). "BioVerbNet: A Large Semantic-Syntactic Classification of Verbs in Biomedicine". In: *Journal of Biomedical Semantics* 12.1, p. 12. ISSN: 2041-1480. DOI: 10.1186/s13326-021-

00247-z. URL: https://doi.org/10.1186/s13326-021-00247-z (visited on 11/23/2022).

Majewska, Olga, Diana McCarthy, Ivan Vulić, and Anna Korhonen (May 2018). "Acquiring Verb Classes Through Bottom-Up Semantic Verb Clustering". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1153 (visited on 11/23/2022).

Majewska, Olga, Ivan Vulić, Diana McCarthy, and Anna Korhonen (Dec. 2020). "Manual Clustering and Spatial Arrangement of Verbs for Multilingual Evaluation and Typology Analysis". In: *Proceedings of the 28th International Conference on Computational Linguistics*. COLING 2020. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 4810–4824. DOI: 10.18653/v1/2020.coling-main.423. URL: https://aclanthology.org/2020.coling-main.423 (visited on 11/23/2022).

Miller, George A. (Nov. 1, 1995). "WordNet: A Lexical Database for English". In: *Communications of the ACM* 38.11, pp. 39–41. ISSN: 0001-0782. DOI: 10.1145/219717.219748. URL: https://doi.org/10.1145/219717.219748 (visited on 11/23/2022).

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller (Dec. 1, 1990). "Introduction to WordNet: An On-line Lexical Database". In: *International Journal of Lexicography* 3.4, pp. 235–244. ISSN: 0950-3846. DOI: 10.1093/ijl/3.4.235. URL: https://doi.org/10.1093/ijl/3.4.235 (visited on 11/23/2022).

Navarretta, Costanza (Dec. 2000). "Semantic Clustering of Adjectives and Verbs Based on Syntactic Patterns". In: *Proceedings of the 12th Nordic Conference of Computational Linguistics (NODALIDA 1999)*. NoDaLiDa 2000. Trondheim, Norway: Department of Linguistics, Norwegian University of Science and Technology, Norway, pp. 124–132. URL: https://aclanthology.org/W99-1013 (visited on 11/23/2022).

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman (May 2020). "Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection". In: *Proceedings of the 12th Language Resources and Evaluation Conference*. LREC 2020. Marseille, France: European Language Resources Association, pp. 4034–4043. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.497 (visited on 08/22/2022).

Palmer, Martha, Daniel Gildea, and Paul Kingsbury (2005). "The Proposition Bank: An Annotated Corpus of Semantic Roles". In: *Computational Linguistics* 31.1, pp. 71–

106. DOI: 10.1162/089120105363026 4. URL: https://www.aclweb.org/anthology/J05-1004 (visited on 09/18/2019).
Parisien, Christopher and Suzanne Stevenson (2010). "Learning Verb Alternations in a Usage-Based Bayesian Model". In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 32.32. URL: https://escholarship.org/uc/item/186313ch (visited on 12/09/2022).
Peirce, Charles S. (1897). "The Logic of Relatives". In: *The Monist* 7.2, pp. 161–217. ISSN: 0026-9662. URL: https://doi.org/10.5840/monist18977231 (visited on 09/18/2022).
Przepiórkowski, Adam (Jan. 1, 2018). "The Origin of the Valency Metaphor in Linguistics". In: *Lingvisticæ Investigationes* 41.1, pp. 152–159. ISSN: 0378-4169, 1569-9927. DOI: 10.1075/li.00017.prz. URL: https://www.jbe-platform.com/content/journals/10.1075/li.00017.prz (visited on 09/18/2022).
Say, Sergey (Jan. 1, 2014). "Bivalent Verb Classes in the Languages of Europe: A Quantitative Typological Study". In: *Language Dynamics and Change* 4.1, pp. 116–166. ISSN: 2210-5832, 2210-5824. DOI: 10.1163/22105832-00401003. URL: https://brill.com/view/journals/ldc/4/1/article-p116_4.xml (visited on 06/20/2022).
Sayeed, Asad, Pavel Shkadzko, and Vera Demberg (May 2018). "Rollenwechsel-English: A Large-Scale Semantic Role Corpus". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1488 (visited on 06/07/2022).
Schulte im Walde, Sabine (Apr. 2003). "Experiments on the Choice of Features for Learning Verb Classes". In: *10th Conference of the European Chapter of the Association for Computational Linguistics*. EACL 2003. Budapest, Hungary: Association for Computational Linguistics. URL: https://aclanthology.org/E03-1037 (visited on 11/23/2022).
– (2006). "Experiments on the Automatic Induction of German Semantic Verb Classes". In: *Computational Linguistics* 32.2, pp. 159–194. DOI: 10.1162/coli.2006.32.2.159. URL: https://aclanthology.org/J06-2001 (visited on 11/16/2022).
Schulte im Walde, Sabine and Chris Brew (July 2002). "Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. ACL 2002. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 223–230. DOI: 10.3115/1073083.1073121. URL: https://www.aclweb.org/anthology/P02-1029 (visited on 09/18/2019).
Snider, Neal and Mona Diab (July 2006). "Unsupervised Induction of Modern Standard Arabic Verb Classes Using Syntactic Frames and LSA". In: *Proceedings of the*

COLING/ACL 2006 Main Conference Poster Sessions. COLING-ACL 2006. Sydney, Australia: Association for Computational Linguistics, pp. 795–802. URL: https://aclanthology.org/P06-2102 (visited on 11/23/2022).

Stabler, Edward P. (2019). "Three Mathematical Foundations for Syntax". In: *Annual Review of Linguistics* 5.1, pp. 243–260. DOI: 10.1146/annurev-linguistics-011415-040658. URL: https://doi.org/10.1146/annurev-linguistics-011415-040658 (visited on 08/16/2019).

Sun, Lin and Anna Korhonen (Aug. 2009). "Improving Verb Clustering with Automatically Acquired Selectional Preferences". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2009. Singapore: Association for Computational Linguistics, pp. 638–647. URL: https://aclanthology.org/D09-1067 (visited on 11/23/2022).

Sun, Lin, Anna Korhonen, and Yuval Krymolowski (2008). "Verb Class Discovery from Rich Syntactic Data". In: *Computational Linguistics and Intelligent Text Processing*. Ed. by Alexander Gelbukh. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 16–27. ISBN: 978-3-540-78135-6. DOI: 10.1007/978-3-540-78135-6_2.

Sun, Lin, Diana McCarthy, and Anna Korhonen (Aug. 2013). "Diathesis Alternation Approximation for Verb Clustering". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL 2013. Sofia, Bulgaria: Association for Computational Linguistics, pp. 736–741. URL: https://aclanthology.org/P13-2129 (visited on 11/23/2022).

Tesnière, Lucien (1959). *Éléments de syntaxe structurale*. Paris: C. Klincksieck.

– (2015). *Elements of Structural Syntax*. John Benjamins Publishing Company. ISBN: 978-90-272-1212-2. DOI: 10.1075/z.185. URL: https://library.oapen.org/handle/20.500.12657/30722 (visited on 08/22/2022).

Titov, Ivan and Alexandre Klementiev (Apr. 2012). "A Bayesian Approach to Unsupervised Semantic Role Induction". In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. EACL 2012. Avignon, France: Association for Computational Linguistics, pp. 12–22. URL: https://aclanthology.org/E12-1003 (visited on 09/18/2022).

Tsunoda, Tasaku (Jan. 1, 1981). "Split Case-Marking Patterns in Verb-Types and Tense/Aspect/Mood". In: 19.5-6, pp. 389–438. ISSN: 1613-396X. DOI: 10.1515/ling.1981.19.5-6.389. URL: https://www.degruyter.com/document/doi/10.1515/ling.1981.19.5-6.389/html?lang=en (visited on 09/25/2022).

– (1985). "Remarks on Transitivity". In: *Journal of Linguistics* 21.2, pp. 385–396. ISSN: 0022-2267. JSTOR: 4175793.

– (2015). "The Hierarchy of Two-Place Predicates: Its Limitations and Uses". In: Malchukov, Andrej and Bernard Comrie. *Valency Classes in the World's Languages:*

Volume 2, Case Studies from Austronesia and the Pacific, the Americas, and Theoretical Outlook. Vol. 2. 2 vols. Berlin; Boston: De Gruyter Mouton, pp. 1597–1626. ISBN: 978-3-11-043844-4.

Watanabe, Yotaro, Masayuki Asahara, and Yuji Matsumoto (July 2010). "A Structured Model for Joint Learning of Argument Roles and Predicate Senses". In: *Proceedings of the ACL 2010 Conference Short Papers*. ACL 2010. Uppsala, Sweden: Association for Computational Linguistics, pp. 98–102. URL: https://aclanthology.org/P10-2018 (visited on 09/25/2022).

Xu, Dongkuan and Yingjie Tian (June 2015). "A Comprehensive Survey of Clustering Algorithms". In: *Annals of Data Science* 2.2, pp. 165–193. ISSN: 2198-5804, 2198-5812. DOI: 10.1007/s40745-015-0040-1. URL: http://link.springer.com/10.1007/s40745-015-0040-1 (visited on 12/09/2022).

Yamada, Kosuke, Ryohei Sasano, and Koichi Takeda (Aug. 2021). "Semantic Frame Induction Using Masked Word Embeddings and Two-Step Clustering". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. ACL-IJCNLP 2021. Online: Association for Computational Linguistics, pp. 811–816. DOI: 10.18653/v1/2021.acl-short.102. URL: https://aclanthology.org/2021.acl-short.102 (visited on 11/23/2022).

Zeman, Daniel, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov (Oct. 2018). "CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies". In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. CoNLL 2018. Brussels, Belgium: Association for Computational Linguistics, pp. 1–21. DOI: 10.18653/v1/K18-2001. URL: https://aclanthology.org/K18-2001 (visited on 12/14/2022).

Zeman, Daniel, Joakim Nivre, et al. (Nov. 15, 2022). *Universal Dependencies 2.11*. Universal Dependencies Consortium. URL: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-4923 (visited on 11/25/2022).

Zeman, Daniel, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez

Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li (2017). "CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies". In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Vancouver, Canada: Association for Computational Linguistics, pp. 1–19. DOI: 10.18653/v1/K17-3001. URL: http://aclweb.org/anthology/K17-3001 (visited on 12/09/2022).