Proposal for a Master's thesis

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

Siyu Tao

December 23, 2022

Abstract

Issues of verbal valency has long occupied a central place in the study of argument structure and the interface between lexicon and grammar. However, the cross-lingual comparison of valency systems and their general typology have proved challenging due to both disagreements on the linguistic basis of comparison and the difficulty in arriving at categorical types. For this thesis, I propose clustering experiments on the automatic induction of valency frames and of verb classes using morphosyntactically annotated data from the Universal Dependencies (UD) treebanks and the interpretation and analysis of the results through information-theoretic metrics of complexity and similarity. The aim of the proposed quantitative typological study is to explore the utility and limits of unsupervised clustering methods and dependency-annotated corpora in the cross-lingual comparison of verbal valency systems, in hopes that it may also yield quantitative insights on the complex issue.

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), have 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 have an impact on 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 theory and dependency grammar is surveyed in §2 to provide the theoretical background for the proposed study; §3 formulates the key research questions the study seeks to address; §4 lays out the data sources and potential methodology to be used in the study; §5 provides a work plan and suggests a tentative timeline for completing the thesis; §6 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 was introduced to linguistics by analogy and refers to the combining power of a word, primarily a verb or 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 analogies, each verbal node, being the center of sen-

[1] It should be noted that while Tesnière is rightly credited with the introduction of a theory of linguistic

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

Implicit in the focus on verbal valency is the assumption, shared by most linguistic theories, of the centrality of the verb in determining either or both the syntactic and semantic structure of a sentence. This assumption has also been corroborated by psycholinguistic evidence (Healy and Miller, 1970) and places valency and the issues of *argument structure* squarely at the center of the inquiry into the interface between syntax and lexical semantics.

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 an intransitive verb cannot. As such, transitive and intransitive verbs form subcategories of the category of verb. Verbs are thus further assigned to *subcategorization frames* which specify the number and type of complements, i.e., objects and obliques, (and of subjects as well in later theories), that the verb can be subcategorized for. In addition to being syntactically driven, a notable feature of generative theories' treatment of valency is 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 developed a theory of thematic relations and posited 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, Levin (1993) categorized verbs in a much more fine-grained manner based on their syntactic behavior into different verb classes. Starting from 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, where it spurred further work on verb classes, and in computational approaches to lexical semantics, where the

valency, the metaphor of valency itself has made appearances as early as in Peirce (1897), among others (Przepiórkowski, 2018).

VerbNet (Kipper et al., 2006, 2008; Kipper-Schuler, 2005) is a prominent example of 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). Further work on verb class induction based on syntactic patterns includes Basili et al. (1993), Korhonen et al. (2006), Navarretta (2000), Sun and Korhonen (2009), Sun, Korhonen, and Krymolowski (2008), and Sun, McCarthy, et al. (2013) in English, schulteimwalde2002, schulteimwalde2003, schulteimwalde2006 in German, Snider and Diab (2006) in Arabic. Sun, McCarthy, et al. (2013) in particular included diathesis alternation as input feature. Other work focused instead on the induction of semantic verb classes such as Fürstenau and Rambow (2012), Majewska, McCarthy, et al. (2018), and Majewska, Vulić, et al. (2020). And work such as Abend et al. (2009), Bickel et al. (2014), Dowty (1991), Sayeed et al. (2018), Titov and Klementiev (2012), Watanabe et al. (2010), and Yamada et al. (2021), among others, work on the induction of semantic roles, a topic arguably tightly related to the induction of the verb classes.

Another computational project focused on verbal valency, FrameNet (C. F. Baker, Fillmore, et al., 1998; Fillmore and C. Baker, 2015) differs from VerbNet in terms of their theoretical foundations, in that it derives from a divergent line of research that stemmed from Charles Fillmore's frame semantics (Fillmore, 1977a,b, 1982), which in turn has its roots in his earlier work on case grammar (Fillmore, 1968, 1970). While they are often 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 frames are viewed as a level of constructions on their own, cf. e.g., Goldberg (1992, 1995)'s *argument structure constructions*. Furthermore, construction grammar theories often argue for frames to be considered distinct or autonomous constructions, as it is not strictly predictable from other constructions.

2.2 Typological perspectives on valency and dependency

It is perhaps not surprising that, besides introducing the analogy of valency, Tesnière (1959) also introduced the notion of dependency into modern linguistics. In terms of their mathematical foundations, dependency grammar, based on the notion of dependencies, can be viewed in contrast with constituency grammars which are based on the notion of substitution instead (Stabler, 2019). However, even most iterations of generative grammar theories, which are primarily constituency-based,

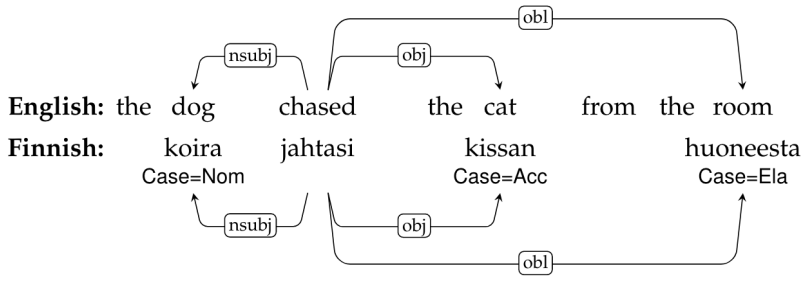

Figure 1: Simplified UD annotation for equivalent sentences from English (top) and Finnish (bottom) (de Marneffe, Manning, et al., 2021).

incorporate some version of a head-dependent relationship (cf. X-bar theory). de Marneffe and Nivre (2019) cited 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.

The usefulness of dependency grammar in allowing for cross-lingual generalizations and comparisons of linguistic structures should not be understated. Universal Dependencies (UD) (de Marneffe, Manning, et al., 2021; Nivre, 2015) in particular is an initiative that aims to develop a uniform grammatical annotation system that are cross-lingually consistent. The basic structure of the UD annotation is to segment *sentences* into *syntactic words* which are annotated with their *morphological properties* and linked together by *syntactic relations*. A comparison of UD annotations of equivalent sentences in two languages shows how they can show both the structural parallel and differences between how two languages encode the same sentence, as seen in Fig. 1, where both the similarities between how English and Finnish encoded semantically equivalent sentences (same syntactic relationships between the arguments and the verb) and the differences (case markings in Finnish, preposition in English) are easily discernable. And further enhancements have also been proposed that would make the UD annotation scheme more compatible with contemporary typological theory (Croft et al., 2017).

Specifically on verbal valency, already Tesnière (1959) was paying attention to the cross-lingual differences in the argument structure of semantically equivalent sentences while describing his dependency grammar. Tesnière described the process of *metataxis*, by which syntactic structures of one language are "translated" to those of another. Such a process points to the clear typological interest in valency systems, namely the mismatch between how languages encode their argument structure.

In terms of possible universals that can be observed, Tsunoda (1981, 1985) pro-

posed a transitivity hierarchy of verbs:

Effective action » Perception » Pursuit » Knowledge » Feeling » Relation

The idea is that languages that encode verbs that are lower in this hierarchy as transitive verbs will encode all those above it as well, with the effective action being the most prototypical transitive verb hence most likely to be transitive in a language. This approach is further extended by Malchukov, 2005 who used the semantic map method and proposed a two-dimensional transitivity hierarchy with the semantic map method.

There has been some recent work from advocates of both the lexeme- and frames-based approaches on the cross-lingual alignment of their respective units of linguistic analysis. On the frames-based side, C. F. Baker and Lorenzi (2020) and Ellsworth et al. (2021) explored the cross-lingual alignment of frames based on FrameNet; in contrast, Say (2014) rejected the equating of minor valency classes cross-lingually and studied how verb classes compare cross-lingually instead, seeing that as a more valid method of measuring how languages organize their verbal lexicon differently.

3 Research questions

The proposed thesis has both computational and theoretical goals in mind, aiming to not only experiment whether dependency-based datasets can support a quantitative study of verbal valency, but also explore whether the quantitative results will support certain theoretical models of verbal valency over others. This section formulates in more detail the key research questions it seeks to address.

The first set of research questions relate primarily to the experiments proposed for the study.

- Can semantically coherent verb classes be derived from UD-style morphosyntactic annotations? More specifically:
  - Do UD annotations support an automated induction of the valency frames in a language given that it only encodes morphosyntactic information?
  - If so, can verb classes can be further induced based on the distribution of verbs across the valency frames?

To answer these questions, I propose in this thesis clustering experiments on automatically inducing valency frames based on UD treebanks and afterwards verb classes based on the induced valency frames. Successful results will confirm the assumption that verb valency works at the interface between lexical semantics and

syntax and offer support to the hypothesis that syntactic behavior of verbs are semantically determined. Note however that they cannot provide support either for or against lexeme- or frame-based view of verbal valency, as both predict interactions between syntax and semantics with the disagreement primarily on whether the valency information is stored in the verbal lexicon or the frames.

The second set of research questions relate primarily to the metrics proposed for the study.

- Can the automatically induced valency frames and/or the verb classes reveal cross-linguistic patterns in how languages organize their verbal lexicon? More specifically:
  - Are information-theoretic metrics a good measure of the complexity of verbal valency systems and cross-linguistic (dis)similarity?
  - And could they be evidence for or against a lexeme- or frames-based view of valency?

To answer these questions, I propose in this thesis the use of information-theoretic metrics to measure the internal complexity and cross-lingual similarity of valency systems as characterized by the results of the clustering experiments. The metrics will be interpreted in combination with manual analysis of the results to assess whether they should be taken as revealing typological patterns of valency systems as well as whether they consequently support certain theoretical models of valency over others.

For example, if verb classes can be shown to be a more consistent basis of typological comparison, such results may be viewed as in support of a lexeme-based view of valency systems. Whereas if languages are shown to employ different strategies to encode cross-lingually consistent frames, this may instead be evidence favoring a construction-based view. This has further consequences for the search of linguistic universals regarding verbal valency systems as well: if we consider valency frames primarily a feature of the verbal lexicon, then cross-lingually, this means the comparison of the distribution of verbs across different verb classes allows us to test possible universals regarding the organization of the verbal lexicon; the opposite is true if we consider valency frames a construction on their own.

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.

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) (Nivre et al., 2020).

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, as an example.

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 **the Valency Patterns Leipzig Online Database (ValPaL)**(Hartmann et al., 2013). Other datasets will be introduced as necessary.

4.2 Representing verb instances / feature selection

In the first step, the specific uses of verbs are abstracted through a feature selection process. Each instance of verb use will be represented by the morphosyntactic features of the sentence, namely only features that are considered part of valency frame encoding are included. This is in order to focus on whether semantically coherent verb classes can be induced from valency frame information alone. In selecting the features, cross-lingual differences in valency frame coding will be taken into account, e.g., whether a language uses morphological cases or word order to encode valency frame information. Word order information, although not explicitly specified in UD, should also be extracted from the dataset. An alternative approach considered is to keep manual feature selection at a minimum and to allow the clustering algorithms to weigh the features as needed.

4.3 Clustering

The clustering process after feature selection consists of two steps, but the clustering algorithms used need not be the same. The first is the automatic induction of valency frames in a language given the selected features and the second is the clustering of verbs, represented by their distribution over the valency frames, into verb classes. Since we intend to perform unsupervised clustering, the number of valency frames and the number of verb classes cannot be assumed *a priori*. This requires either using of algorithms that do not require a predefined number of clusters (e.g. Ward clustering), or experimenting with cluster sizes with each language (cf. Schulte im Walde, 2006, which used the k-means algorithm with a predefined the gold standard number). Due to the lack the gold standard for many of the languages to be experimented on, the former seems preferable. A bottom-up agglomerative clustering method will

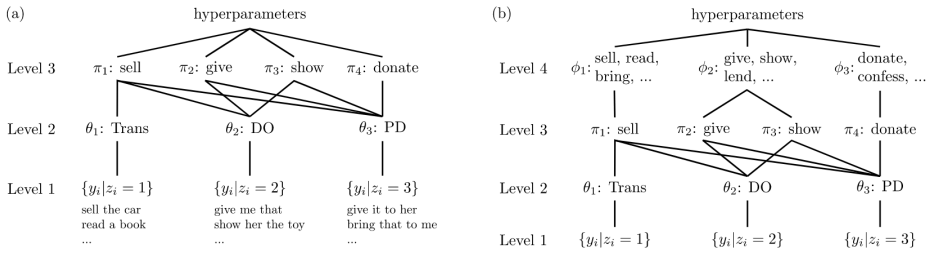

Figure 2: (a) Model 1, a Hierarchical Dirichlet Process applied to learning verb argument structure constructions. (b) Model 2, an extension of Model 1 to learn verb alternation classes.

also be favored over top-down methods.

Given the relative low dimensionality of hand-selected features, complex clustering algorithms are not anticipated to be necessary. Nevertheless, more modern clustering algorithms should also be investigated (Xu and Tian, 2015). Given the two levels of clustering, one method to be considered for the verb class induction is the Hierarchical Dirichlet process, which is particularly suited for clustering grouped data (cf. Parisien and Stevenson (2010), Fig. 2, where a Hierarchical Dirichlet process was extended to account for diathesis alternations).

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 (Kamholz et al., 2014). However, multilingual word embeddings and induction of a cross-lingual verbal lexicon could be considered as another option is the existing resources prove insufficient.

4.5 Information theory metrics

The specific metrics to be used on the results will have to be determined in combination with the decisions to be made in the experiments such as the number of the language studied and whether parallel or non-parallel datasets are used. Preliminarily, however, information theory metrics modeled on those used by Say (2014) are considered, specifically an internal complexity metric measuring the entropy of the distribution of verbs among valency classes and a similarity metric based on mutual

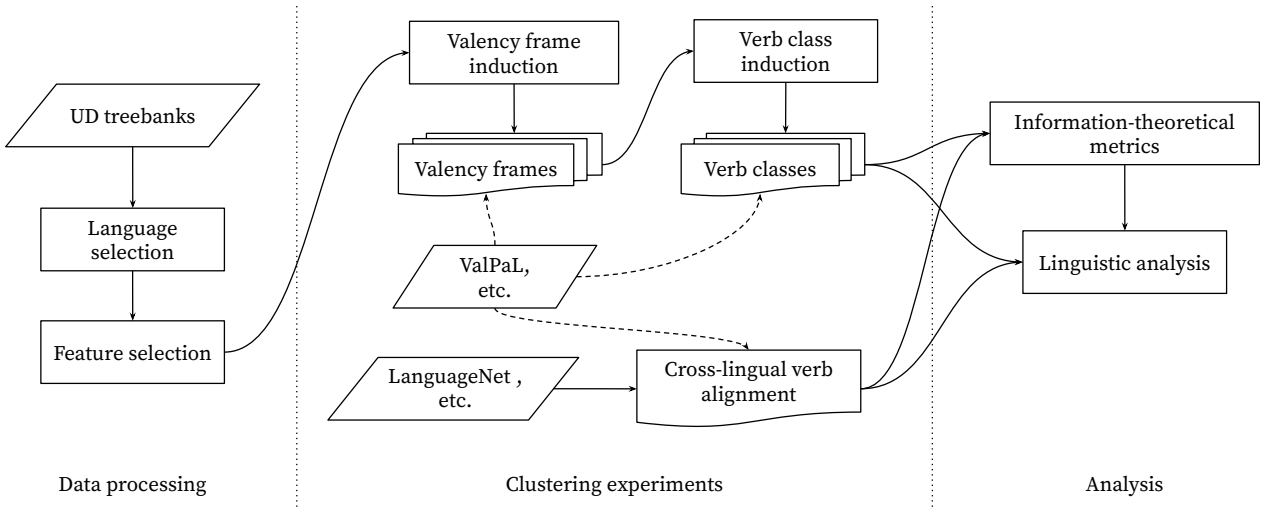

Figure 3: Flowchart depicting the data sources and processes as described in the proposal. Dotted lines depict validation steps.

information measuring the similarity / dissimilarity between valency systems of two or more languages.

5 Work plan

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

The work plan for the thesis is shown in the flowchart in Fig. 3. The general process with most of the specific steps already discussed in §4 are divided into the data processing stage, the clustering experiments stage, and the analysis stage. In terms of the timeline, this thesis study is intended to be completed within roughly three months after the submission of this proposal, with a statutory maximum of six months. Given the preference for a more speedy completion, 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 alternate methods 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.

6 Conclusion

The proposed thesis would like to contribute to the study of verbal valency systems by including the typological and quantitative perspectives. It will first and foremost attempt novel combinations of computational methods and corpora in service of a quantitative typological investigation, therefore exploring the utility and limits of dependency-based datasets and automatic clustering methods. In doing so, it also aims to contribute to the typological study and theoretical debate on verbal valency. It is hoped that the difficulty in working out a finite categorical typology of valency class systems may be overcome through statistical and information-theoretic methods and modern computational corpora; and that furthermore the empirical study can bring out patterns and observations that may not have been obvious in an introspective study, which may help settle or intensify theoretical debates on valency.

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.

Kamholz, David, Jonathan Pool, and Susan Colowick (May 2014). "PanLex: Building a Resource for Panlingual Lexical Translation". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. LREC 2014. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 3145–

3150. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1029_Paper.pdf (visited on 12/23/2022).

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.

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).

Malchukov, Andrej L. (Jan. 1, 2005). "Chapter 4 - Case Pattern Splits, Verb Types and Construction Competition". In: *Competition and Variation in Natural Languages*. Ed. by Mengistu Amberber and Helen De Hoop. Perspectives on Cognitive Science. Oxford: Elsevier, pp. 73–117. DOI: 10.1016/B978-008044651-6/50006-9. URL: https://www.sciencedirect.com/science/article/pii/B9780080446516500069 (visited on 12/20/2022).

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).

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 (2015). "Towards a Universal Grammar for Natural Language Processing". In: *Computational Linguistics and Intelligent Text Processing*. Ed. by Alexander Gelbukh. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 3–16. ISBN: 978-3-319-18111-0. DOI: 10.1007/978-3-319-18111-0_1.

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).

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 (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).

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.

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).