Saarland University
Department of Language Science and Technology
Faculty of Humanities

Master's Thesis

Typological investigations of verbal valency systems

A quantitative study based on Universal Dependencies

Siyu Tao

June 14, 2023

Advisors: Prof. Dr. Michael Hahn
Dr. Lucia Donatelli

Supervisors: Jun.-Prof. Dr. Annemarie Verkerk
Dr. Lucia Donatelli

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Ich versichere, dass die gedruckte und die elektronische Version der Masterarbeit inhaltlich übereinstimmen.

Statutory Declaration

I hereby declare that the thesis presented here is my own work and that no other sources or aids, other than those listed, have been used. I affirm that the electronic version is identical in content to the printed version of the Master's thesis.

Ort, Datum / *Place, date*: ____________________

Unterschrift / *Signature*: ____________________

Acknowledgments

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

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. In this thesis, I propose information-theoretic metrics to characterize the complexity of verbal valency at a lexical level and experiment using morphosyntactically annotated data from the Universal Dependencies (UD) treebanks. The aim of the proposed quantitative typological study is to explore the utility and limits of quantitative 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.

Contents

Abstract vii

1. Introduction 1

2. Background and theoretical framework 3
   2.1. Valency and valency phenomena . . . 3
   2.2. Typological perspectives on valency and dependency . . . 5

3. Data selection 9
   3.1. Data sources . . . 9
      3.1.1. Universal Dependencies . . . 9
   3.2. Representing verb instances . . . 11
   3.3. Clustering . . . 11
   3.4. Cross-lingual verb sense alignment . . . 12
   3.5. Information theory metrics . . . 12

4. Experiments and analysis 13
   4.1. Experiment 1: Transitivity ratios . . . 13
      4.1.1. From transitivity categories to transitivity ratios . . . 13
      4.1.2. Transitivity ratios and the lexicon . . . 16
      4.1.3. Genetic and areal patterns in transitivity . . . 17
   4.2. Experiment 2: Valency frame entropy . . . 19
      4.2.1. Valency frame and the efficient organization of the lexicon . . . 19
      4.2.2. Measuring valency frame entropy in UD . . . 19
      4.2.3. Core and non-core dependents . . . 22
   4.3. Experiment 3: Word order and case information . . . 22
      4.3.1. Motivation . . . 22
      4.3.2. Experiment design: ablation studies . . . 22
      4.3.3. Results . . . 22
   4.4. Experiment 4: Verb entropy . . . 22
      4.4.1. Motivation . . . 22
      4.4.2. Experiment design . . . 22
      4.4.3. Results . . . 22

4.5. Experiment 5: Verb-finalness 22
4.5.1. Motivation 22
4.5.2. Experiment design 22
4.5.3. Results 22

5. Conclusion 27

A. Experiment 1 Additional Results 29

Bibliography 33

Chapter 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.

We develop a quantitative methodology. The methodology is described along with experiment design in the respective sections.

Chapter 2. Background and theoretical framework

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 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 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

[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).

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 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, Schulte im Walde (2003, 2006) and Schulte im Walde and Brew (2002) 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, worked 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, 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 annota-

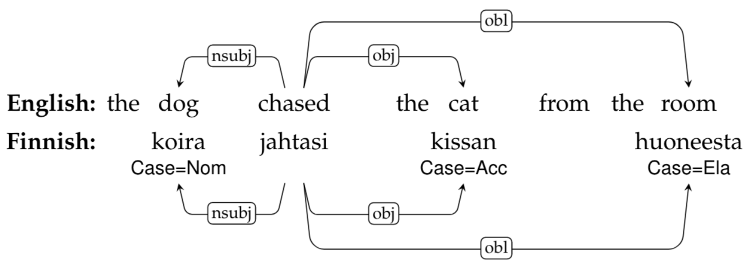

Figure 2.1.: Simplified UD annotation for equivalent sentences from English (top) and Finnish (bottom) (de Marneffe, Manning, et al., 2021).

tion 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. 2.1, 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.

transitivity

In terms of possible universals that can be observed, Tsunoda (1981, 1985) proposed 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 them too as transitive, 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.

Chapter 3.

Data selection

3.1. Data sources

3.1.1. Universal Dependencies

Universal Dependencies (UD) is the main source of primary data used for the present study. It 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 v2 update to the UD annotation guidelines also introduced changes that intend to decrease the reliance on language-specific categories (Nivre et al., 2020). Inevitably, these efforts had to be balanced against the practicality of computational efficiency but nevertheless converged in many cases with proposals by typologists, as the core principles converged with a functional typology approach. Croft et al. (2017)

1. UD needs to be satisfactory on linguistic analysis grounds for individual languages.

2. UD needs to be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.

3. UD must be suitable for rapid, consistent annotation by a human annotator.

4. UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. We refer to this as seeking a *habitable* design, and it leads us to favor traditional grammar notions and terminology.

5. UD must be suitable for computer parsing with high accuracy.

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 3.1.: Languages in UD v2.5 with number of treebanks (#), sentences (Sents) and words (Words) (Nivre et al., 2020).

6. UD must support well downstream language understanding tasks (relation extraction, reading comprehension, machine translation, ...).

The **Universal Dependencies** 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 3.1 for a table of languages available in UD v2.5, as an example.

The thesis uses the UD v2.11 release.

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.

3.2. Representing verb instances

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.

3.3. 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 unsupervised clustering will be used, the number of valency frames and the number of verb classes cannot be assumed *a priori*. This requires either using 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 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. 3.1, where

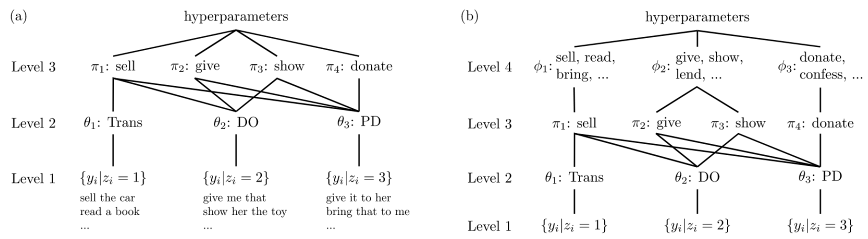

Figure 3.1.: (a) Model 1, a Hierarchical Dirichlet Process applied to learning verb argument structure constructions. (b) Model 2, an extension of Model 1 to learn verb alternation classes.

a Hierarchical Dirichlet process was extended to account for diathesis alternations). ·

3.4. Cross-lingual verb sense alignment

A cross-lingual aligned list of counterpart verbs will be needed to compare the verb classes. 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 should the existing resources prove insufficient.

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

Chapter 4.

Experiments and analysis

4.1. Experiment 1: Transitivity ratios

4.1.1. From transitivity categories to transitivity ratios

If we were to approach the task of verb classification, i.e., to categorize verbs of a language into verb classes according to their syntacto-semantic properties and behavior, we could well imagine ourselves with fine-grained verb classes à la Levin (1993) in the end, but will likely have to start with more basic distinctions and, first among them, that of verb *transitivity*.

In contrast with more fine-grained metrics of valency to be used in further experiments, the notion of *transitivity*, i.e., whether a verb can take one or more objects, is probably familiar to any of us who has ever been in a foreign language classroom. Traditionally, a binary distinction is made between *intransitive* verbs, which take only a subject and no objects and *transitive* verbs, which take one or more objects. Further categories, some overlapping, deal with finer distinctions, such as *ditransitive* verbs (those taking two objects), ambitransitive verbs (those that can be used both transitively and intransitively), etc.

While transitivity categories can and are already used in verb classification, as well as the cross-lingual comparison thereof, a functional and qualitative approach demands a better picture of transitivity *use* in language. An example illustrating what transitivity categories do *not* capture comes from the study of near-synonyms: (Biber et al., 1998) compares the English verbs *begin* and *start* in a corpus-based study. At first glance, English appears to have provided us with two verbs that are not only semantically synonymous but share valency properties as well, as they can both be used in transitive and intransitive constructions:

(1) a. I had better issue a survival kit before we *start*/*begin*.
intransitive

b. Then they *started/began* the quota system.
transitive with noun phrase

c. They'd *started/begun* leaving before I arrived.
transitive with *-ing* clause

d. One of the wheels had *started/begun* to wobble.
transitive with *to* clause

This however belies the different usage patterns exhibited by these verbs, as Biber et al. (1998, p. 95) demonstrate with statistics from the British National Corpus (BNC): while both uses are clearly grammatical for both verbs, *begin* is used more often in a transitive frame than *start* across different genres in the BNC: in fiction, 78% of *begin* occurrences (196/250) are with various transitive patterns vs. only 60% for *start* (149/250); transitive uses are less frequent in academic texts in general but the observation of relatively higher transitivity for *begin* still holds (57% vs. 36%, or 110/192 vs. 51/142).

To capture such differences in degrees of transitivity, we propose measuring the **transitivity ratio** based on language corpora, defined as the percentage of verb use that are transitive. Necessarily then, and in contrast with transitivity categories, any values are calculated on an *ad hoc* basis in a given corpus. Unless we expect the corpus to be a representative sample of all language use in that language, which is certainly not the case for the UD corpora we are using, the absolute values of these ratios at a lexeme-level cannot be directly interpreted. Instead, intralinguistic analysis will take the form of analyzing the distribution of verbs according to their observed transitivity ratios. We take on an additional assumption, that the corpus would reflect the general tendency towards (in-)transitivity of the language, if too piecemeal for individual verbs, which then allows for cross-lingual comparison of transitivity ratios at a token-level.

The UD annotation scheme provides us with good facility to investigate transitivity at lexeme- and token-levels, as the relevant dependency relations NSUBJ and OBJ mark respectively the first and second core arguments of a verb with their typical syntactic roles as subject and object. This is defined without respect to specific cases (even though typically the accusative in languages with a case system) or semantic roles (even though they would typically be the proto-agent and proto-patient) in an effort to avoid *a priori* categories to the extent possible. The renaming of the DOBJ relation to OBJ, among the changes introduced by UD v2 (Nivre et al., 2020), reflects the same effort.

Given the clear and typologically sound UD dependency relation annotations, one can be forgiven for already jumping at the task. On closer examination, however, we see that arriving at a clear definition of transitivity ratio is not trivial. We consider four different definitions of a quantitative transitivity ratio within the UD annotation scheme here:

1. the ratio of verb instances with both NSUBJ and OBJ dependents, as compared to verb instances with an NSUBJ dependent

2. the ratio of verb instances with an OBJ dependent, as compared to verb instances with an NSUBJ dependent

3. the ratio of verb instances with an OBJ dependent, as compared to all verb instances

4. the ratio of verb instances with an OBJ dependent, as compared to verb instances with either an NSUBJ or an OBJ dependent

Def. 1 is an attempt at enforcing a definition of the transitive object as the *second* core argument of the verb by excluding from calculation instances where the first core argument (i.e., subject) is not realized. This turns out counterproductive for two reasons. Firstly, this does not sit well with the core definition on transitivity, as instances of verb use where the subject is not expressed should not count against the fact that the verb is taking a transitive object; secondly, this is undesirable in practice when accounting for typological variations, as the metric would be biased against pro-drop languages that drops subject pronouns more often than objects such as Spanish.

Revising def. 1 and dropping the requirement in the numerator for verbs to have an NSUBJ dependent gives us def. 2. However, this is not sufficient, as we are now concerned with the opposite problem where subject-dropping languages are likely to have a smaller denominator, resulting in a high transitivity ratio that is not representative. Def. 3 is where we drop the NSUBJ requirement from the denominator as well. This still faces problems, as verb instances where both subject and object are dropped would affect the denominator, and such usage, e.g., non-predicative usage of verbs, is unlikely to be equally frequent in different languages and would therefore interfere with the cross-lingual comparability of our transitivity ratio focusing on argument structure of verb predicates. Taking all these potential drawbacks into consideration, we arrive at def. 4 with the number of verb instances with either an NSUBJ or an OBJ dependent in the denominator.

While there is a strong case for Def. 4 being the most principled definition, we nevertheless implement all four definitions in this experiment to empirically verify the intuitions. They are also represented with feature matrices in Tab. 4.1 for quick reference.

#	**Definition**
1	[+NSUBJ, +OBJ]/[+SUBJ]
2	[+OBJ]/[+NSUBJ]
3	[+OBJ]/[±NSUBJ, ±OBJ]
4	[+OBJ]/[+NSUBJ] or [+OBJ]

Table 4.1.: Potential definitions of transitivity ratio considered in §4.1, represented with feature matrices

4.1.2. Transitivity ratios and the lexicon

We go through all eligible UD corpora and first compile transitivity ratio statistics for each verb lexeme based on the definitions. From there, we compute the per-language statistics that will become the basis for our cross-lingual comparison: we calculate the lexeme-level and token-level transitivity ratios for each language, respectively the arithmetic mean of the lexeme transitivity ratios and the mean of lexeme transitivity ratios weighted by the frequency of the lexeme. In addition to our transitivity ratio metrics, we will also calculate an additional metric, percentage of transitive verbs, i.e., the percentage of verbs in the observed lexicon that are not strictly intransitive (defined as never observed to take an OBJ), for comparison purposes, as it should correspond better with the traditional binary distinction between transitive and intransitive verbs.

We perform the experiment on the selected subset of UD data as described in §3.1.1. For the analysis, we include only languages with at least 100 observed verb lexemes (56 out of 79 languages); the full results from the experiments can be found in the accompanying data and appendices.

def.	lexeme tr., token tr.	tr. verb %, lexeme tr.	tr. verb %, token tr.
1	$\rho(54) = .83, p = .000$	$\rho(54) = .61, p = .000$	$\rho(54) = .70, p = .000$
2	$\rho(54) = .81, p = .000$	$\rho(54) = -.17, p = .221$	$\rho(54) = -.03, p = .817$
3	$\rho(54) = .77, p = .000$	$\rho(54) = .64, p = .000$	$\rho(54) = .73, p = .000$
4	$\rho(54) = .87, p = .000$	$\rho(54) = .61, p = .000$	$\rho(54) = .61, p = .000$

Table 4.2.: Spearman's rank correlation between the transitivity metrics

To compare between the different definitions of transitivity, we compute Spearman's rank correlations between the lexeme- and token-level means of transitivity ratios according to each of our four definitions, as well as between the transitive verb percentage and each of them. The correlation statistics are listed in Tab. 4.2. We observe overall strong correlations between the mean transitivity ratios at lexeme- and token-levels for all four definitions, with the highest observed for definition 4 ($\rho(54) = .87, p = .000$) and lowest observed for definition 3 ($\rho(54) = .77, p = .000$). The strong correlation is not surprising as we have no reason to expect the more frequent verbs to behave differently from the less frequent verbs with regard to transitivity ratios. This can also be confirmed by correlation tests between verb frequency and verb transitivity ratios for each language, which show no strong correlation.

The correlation statistics between the transitive verb percentages and the transitivity ratios are slightly more revealing. If nothing else, they help us eliminate definition 2 from the competition as it shows no statistically significant correlation ($\rho(54) = -.17, p = .221$ for lexeme-level transitivity ratios and $\rho(54) = -.03, p = .817$ for token-level) while strong correlations are observed for all three other definitions (see Tab. 4.2).

In the absence of a practical reason based on correlation statistics to select a definition among the remaining three over the others, we proceed with the rest of the analysis with results using def. 4, which we have considered *a priori* to be the most principled. We proceed with the intralinguistic analyses by plotting histograms of the distributions of verbs among the transitivity ratios in different languages, as shown in Fig. 4.1.

Most distributions are bimodal with peak at both ends, which supports the overall cross-lingual validity of a binary conception of transitivity. But exceptions abound as well, among others Indonesian (unimodal with peak in the middle), Catalan, Galician, Spanish (unimodal with peak on right). And even between the bimodal distributions, we note that they are rarely symmetric, with differing levels of skew towards either end.

4.1.3. Genetic and areal patterns in transitivity

Tab. 4.3a and 4.3b list the most and least 'transitive' languages in our study, respectively according to the transitive verb percentage and the token-level transitivity ratio. Recall that we favor token-level transitivity ratios over lexeme-level ones for cross-lingual comparison.

Among the most transitive languages are Romance (Catalan, Spanish, French, Galician) and Germanic (German, Norwegian, English, Danish) languages of Europe, Sinitic languages (Chinese, Classical Chinese, particularly when measured by token-level transitivity ratio), Indonesian, Hindi. On the opposite end of the spectrum are Hebrew, Irish, Japanese, as well as Baltic (Lithuanian, Latvian) and Slavic (Slovak, Russian, Polish) languages, which have the lowest transitivity ratios.

Language	# Verbs	Tr. verb %	Language	# Verbs	Tr. verb %
Catalan	627	98.2%	Hebrew	531	52.0%
Spanish	941	92.9%	O.C. Slavonic	211	54.5%
Indonesian	330	91.2%	Slovak	279	55.9%
French	733	90.9%	Lithuanian	212	58.0%
German	1998	89.7%	Gothic	195	58.5%
Norwegian	765	89.7%	Russian	2558	59.3%
Galician	339	88.8%	Japanese	381	61.2%
English	771	88.7%	Polish	1059	61.9%
Danish	212	88.7%	Latvian	787	63.8%
Naija	200	88.5%	Coptic	127	65.4%
...	...	...	...	...	...

(a) by transitive verb percentage

Language	# Verbs	Token tr.	Language	# Verbs	Token tr.
Catalan	627	75.9%	Irish	107	30.9%
Galician	339	70.4%	Faroese	114	32.5%
Afrikaans	117	65.1%	Hebrew	531	33.5%
Vietnamese	150	64.2%	Japanese	381	33.5%
Cl. Chinese	1166	63.8%	Polish	1059	34.1%
Chinese	369	62.9%	Russian	2558	37.4%
Pomak	229	61.8%	Dutch	497	37.5%
Spanish	941	61.1%	Latvian	787	37.7%
Hindi	202	60.7%	Serbian	214	38.4%
Anc. Greek	1111	58.5%	Arabic	407	39.7%
...	...	...	...	...	...

(b) by token-level transitivity ratio

Table 4.3.: Most and least transitive languages by different metrics

To look at any potential areal patterns in transitivity, we also map our token-level transitivity ratio results for European languages in Fig. 4.2. (A less Euro-centric study of areal patterns is unfortunately difficult for the lack of enough language samples in the UD.) We find a particularly high transitivity area in the Iberian peninsular as well as another relatively high transitivity area in the Balkans, in contrast to eastern and northern Europe with lower transitivity.

Where the languages overlap, these observations match well with those from Say (2014)'s survey of transitivity in European languages (as measured by the percentage of verbs that are transitive from a fixed list), who observed high transitivity areas in western Europe except Irish and south-western Balkans, and a corresponding low transitivity area in eastern Europe.

4.2. Experiment 2: Valency frame entropy

4.2.1. Valency frame and the efficient organization of the lexicon

Cognitive linguists and typologists have increasingly sought to integrate the two approaches in the same functionalist research paradigm (Croft, 2016). Among others is research that seeks to examine and explain language-internal and cross-linguistic features of human languages through the lens of communicative efficiency, at various levels including the lexicon, syntax and morphology (see Gibson et al., 2019 for a survey).

An early example where efficiency is used to explain phenomena in human languages is the work of George Kingsley G. Zipf (1935) and G. K. Zipf (1949). He first studied what is now known as Zipf's law, namely the empirical observation that "the magnitude of words tends, on the whole, [stands] in an inverse (not necessarily proportionate) relationship to the number of occurrences", and sought to explain it through the principle of least effort. A more recent body of work focuses on using information theory to show how the structure of human language reflects the need for efficient communication.

citations re lexicon, dependency

Psycholinguistic studies consistently show the effect of semantic and syntactic attributes of the verb on online sentence processing (Collina et al., 2001; Shapiro et al., 1987). More specifically, studies suggest that such attributes are organized in frames as opposed to parameters such as transitivity: e.g., Shetreet et al. (2007), an fMRI study in Hebrew, shows that the number of options in terms of subcategorization and thematic frames is better correlated to activity in the cortical areas that are associated with linguistic processing, as opposed to the number of complements.

4.2.2. Measuring valency frame entropy in UD

Besides the information about transitivity, UD annotations make available further information that are related to the morphosyntactic encoding of valency frame, including information about further dependents, word order, as well as case and adpositional markings. In this experiment, we make use of such information for a more fine-grained characterization of the variation in valency frame encoding for different verbs using entropy-based measures.

To compute the entropy measures, we need to first extract the **valency frame encoding** information from the UD annotations at a token level for each verb instance. We start by determining the scope of our investigation

UD label	Dependent
NSUBJ	nominal subject
OBJ	object
OBL	oblique nominal
IOBJ	indirect object
ADVMOD	adverbial modifier
CCOMP	clausal complement
XCOMP	open clausal complement
ADVCL	adverbial clause modifier

Table 4.4.: UD dependent labels included in valency frame extraction

and select a list of dependency relations in Tab. 4.4 which categorizes different syntactic dependencies that can be attached to the verb. These include core dependents, namely NSUBJ and OBJ, but also so-called non-core dependents as well.

justification for that: syntactic classification of verb using non-core information, cite previous work

It is important to note here that it is the cross-lingual taxonomy of the dependencies that serves as the cross-linguistic basis of comparison and that the name of the labels, while linguistically informed, do not affect the entropy calculation in any way.

We extract information about (1) the presence or absence of dependency relations attached to a verb token, (2) the word order information, whether specific dependents precede or follow the head verb, and (3) morphological case marking on the dependents, if any.

example for what is to be extracted

We propose **valency frame entropy**, conditioned on the verb, as a measure of the variation of valency frame encoding for a verb. This is formalized as the joint entropy of the relevant UD dependency relations, i.e., number of bits needed to encode the entire valency frame.

Let $X_1, X_2, \ldots, X_n$ be discrete random variables where each X_k represents a UD dependency relation (e.g., NSUBJ, OBJ, ...). These variables have corresponding sample spaces (images? support?) $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$, which represent the possible levels or outcomes of each variable with regard to its presence, linearized order, and any case information.

Additionally, let there be a variable V that represents the choice of a verb from the lexicon $\mathscr{V}$. We are first interested in calculating the entropy of each dependency relation given a specific verb $v \in \mathscr{V}$. This is defined as:

$$H(X_k|V = v) = -\sum_{x_k \in \mathscr{X}_k} P(x_k|V = v) \log_2 P(x_k|V = v)$$

Here, $P(x_k|V = v)$ represents the conditional probability of the outcome x_k of the dependency relation X_k given that the verb variable V takes on the value v. The entropy of X_k is calculated by summing the products of these probabilities with their logarithms (base 2) taken, each corresponding to a possible outcome x_k.

The joint entropy of all dependent relations, denoted as $H_{\text{joint}}(X_1, X_2, \ldots, X_n|V = v)$, quantifies the uncertainty associated with the combined set of random variables $X_1, X_2, \ldots, X_n$, again given a specific value v for the verb variable V. This is defined as:

$$\begin{aligned} &H_{\text{joint}}(X_1, X_2, \ldots, X_n|V = v) \\ &= -\sum_{x_1 \in \mathscr{X}_1} \cdots \sum_{x_n \in \mathscr{X}_n} P(x_1, x_2, \ldots, x_n|V = v) \log_2 P(x_1, x_2, \ldots, x_n|V = v) \end{aligned}$$

Here, $P(x_1, x_2, \ldots, x_n|V = v)$ represents the joint probability distribution of the outcomes $x_1, x_2, \ldots, x_n$ of the random variables $X_1, X_2, \ldots, X_n$, given that the verb variable V takes on the value v. The joint entropy is calculated by summing the products of these joint probabilities with their logarithms (base 2) taken, each corresponding to a specific combination of outcomes $x_1, x_2, \ldots, x_n$.

4.2.3. Core and non-core dependents

4.3. Experiment 3: Word order and case information

4.3.1. Motivation

4.3.2. Experiment design: ablation studies

4.3.3. Results

4.4. Experiment 4: Verb entropy

4.4.1. Motivation

4.4.2. Experiment design

4.4.3. Results

4.5. Experiment 5: Verb-finalness

4.5.1. Motivation

4.5.2. Experiment design

4.5.3. Results

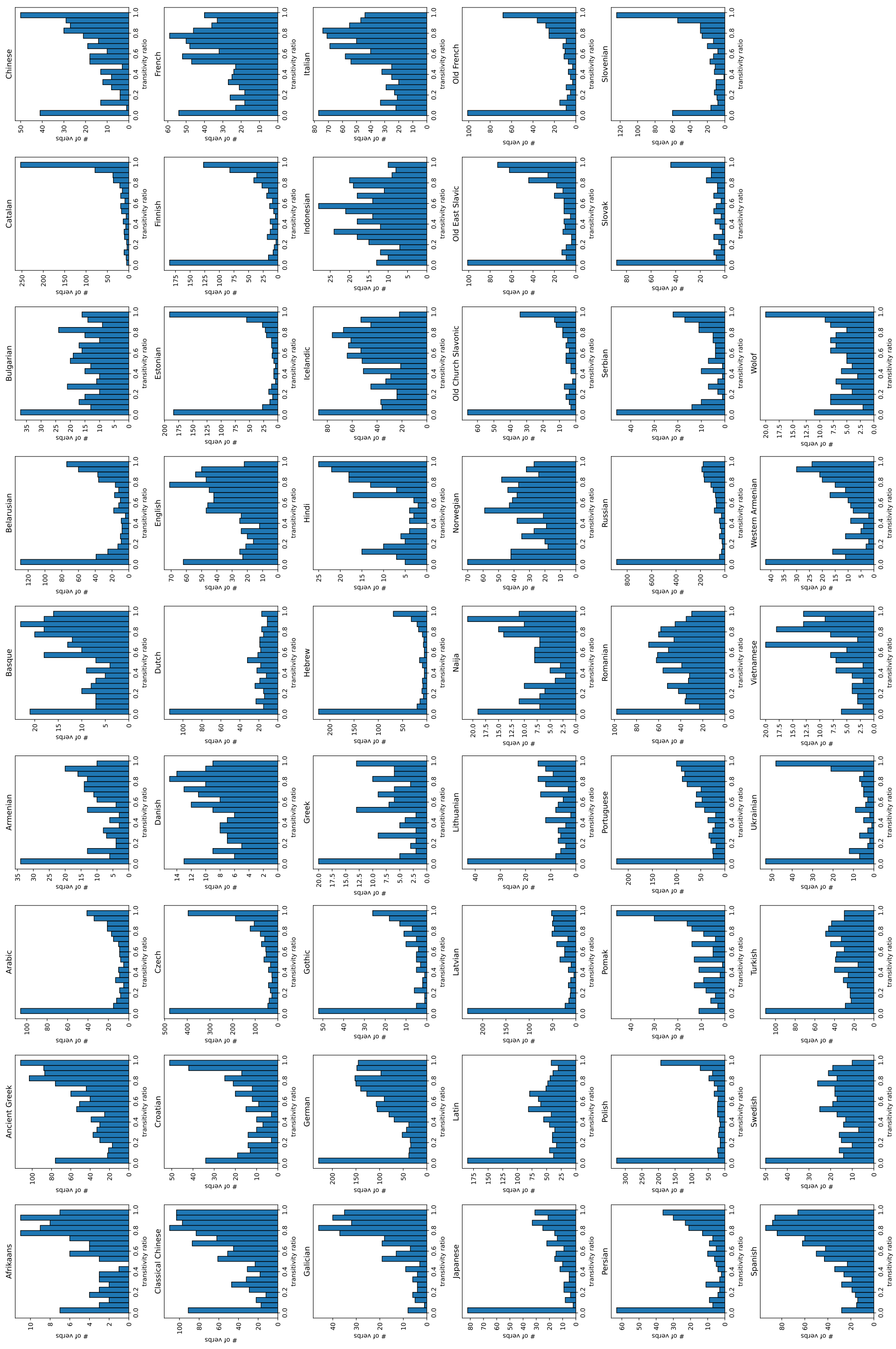

Figure 4.1.: Histograms showing the binned distributions of verbs according to their transitivity ratio in different languages

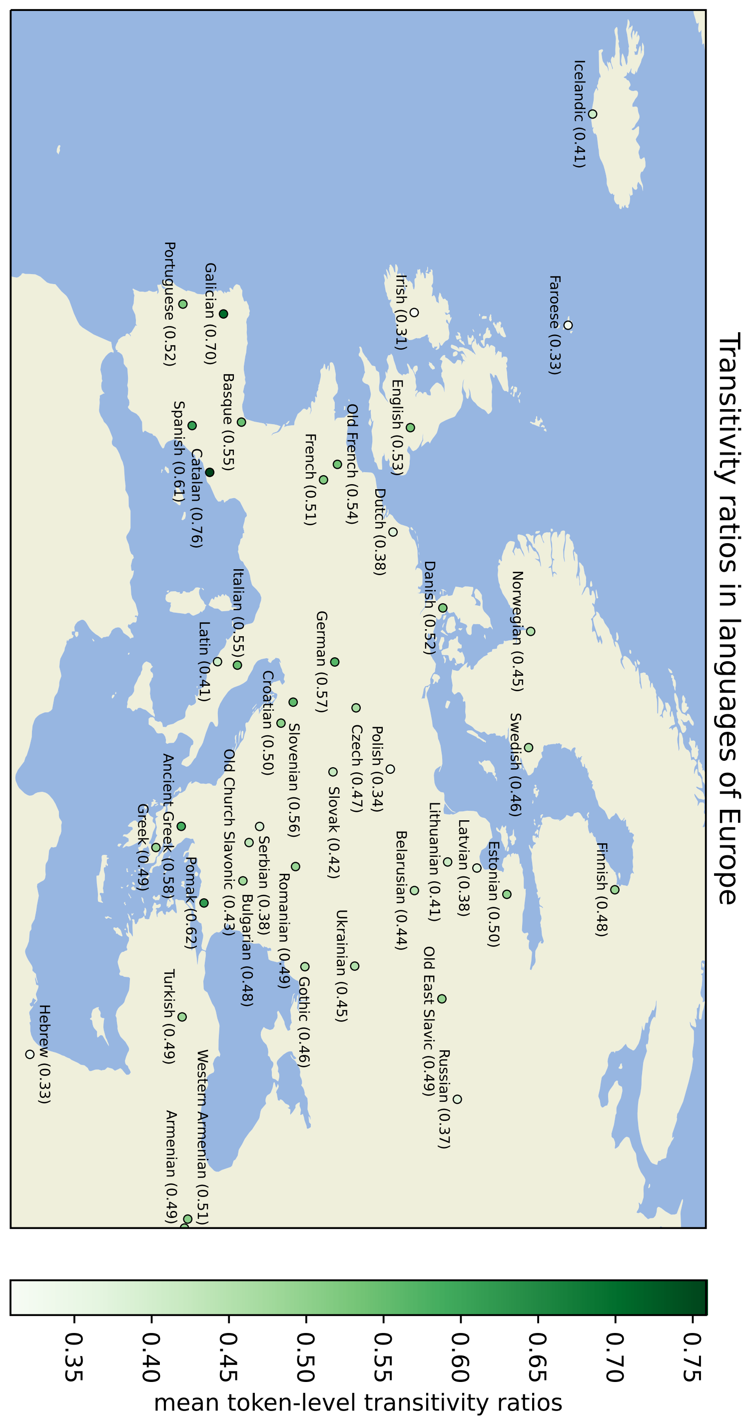

Figure 4.2.: Mean transitivity ratios in languages of Europe

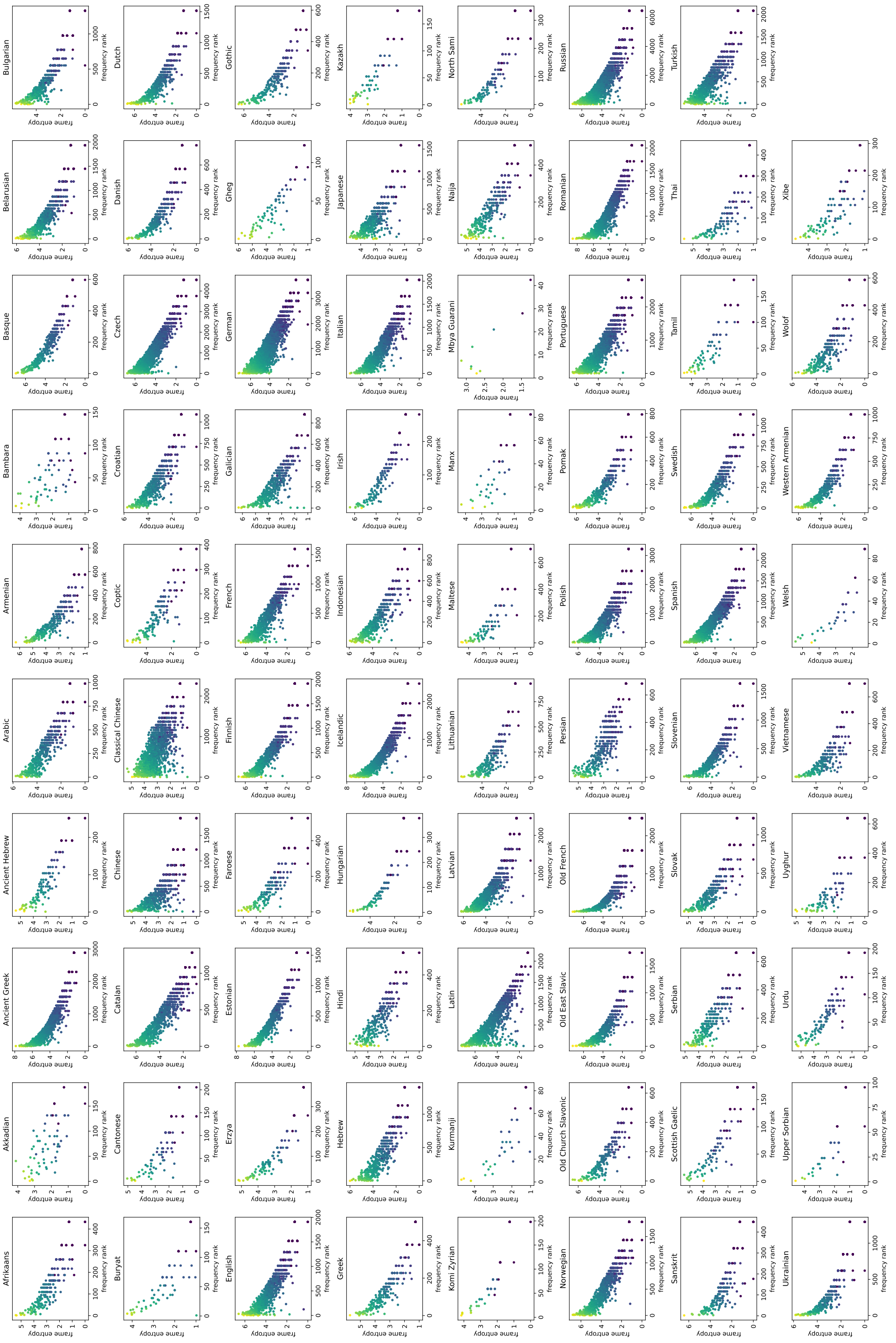

Figure 4.3.: Scatter plots showing relationship between valency frame entropy and frequency rank of verbs, color shows number of frames associated with the verb

Chapter 5.

Conclusion

Appendix A.

Experiment 1 Additional Results

language name	total	tr verb percent	4 mean	4 weighted mean
Akkadian	45	62.22%	73.38%	80.32%
Armenian	227	70.48%	51.53%	48.98%
Welsh	17	70.59%	35.53%	9.45%
Gheg	50	86.00%	65.95%	64.82%
Norwegian	765	89.67%	48.30%	44.80%
Old East Slavic	500	65.80%	55.43%	48.87%
English	771	88.72%	55.51%	52.90%
French	733	90.86%	55.15%	51.29%
Slovenian	514	79.57%	64.30%	55.55%
Hebrew	531	51.98%	35.94%	33.50%
Kurmanji	21	52.38%	27.94%	38.02%
Italian	921	82.41%	55.70%	54.50%
Turkish	773	73.74%	47.20%	48.54%
Finnish	682	69.94%	53.94%	48.41%
Indonesian	330	91.21%	49.74%	51.21%
Ukrainian	228	66.23%	50.64%	45.33%
Dutch	497	75.65%	38.10%	37.51%
Polish	1059	61.95%	47.67%	34.07%
Portuguese	1203	77.47%	52.46%	52.45%
Kazakh	30	63.33%	45.43%	39.91%
Latin	1143	72.35%	44.50%	40.85%
Old French	422	70.14%	52.96%	53.50%
Spanish	941	92.88%	65.36%	61.08%
Buryat	27	51.85%	35.67%	36.83%
Icelandic	997	86.56%	51.81%	41.05%
Estonian	672	67.71%	53.48%	49.69%

Continued on next page

language name	total	tr verb percent	4 mean	4 weighted mean
Croatian	386	84.97%	58.34%	50.27%
Gothic	195	58.46%	50.92%	45.80%
North Sami	75	76.00%	49.35%	37.91%
Naija	200	88.50%	54.92%	45.43%
German	1998	89.69%	57.18%	56.84%
Latvian	787	63.79%	45.18%	37.71%
Chinese	369	83.74%	60.56%	62.90%
Irish	107	77.57%	54.22%	30.87%
Xibe	62	70.97%	65.22%	58.37%
Bambara	50	62.00%	54.14%	43.08%
Lithuanian	212	58.02%	44.38%	41.48%
Galician	339	88.79%	71.14%	70.44%
Vietnamese	150	84.00%	62.00%	64.25%
Greek	145	77.24%	50.65%	48.68%
Catalan	627	98.25%	81.12%	75.86%
Swedish	391	86.45%	47.79%	46.49%
Russian	2558	59.34%	43.43%	37.35%
Czech	2083	72.88%	54.72%	46.76%
Erzya	87	54.02%	40.36%	36.01%
Thai	69	69.57%	51.99%	63.16%
Basque	249	78.71%	56.54%	54.51%
Slovak	279	55.91%	44.26%	42.32%
Tamil	41	80.49%	50.32%	42.73%
Maltese	67	43.28%	30.35%	27.34%
Ancient Greek	1111	81.55%	60.97%	58.47%
Ancient Hebrew	89	73.03%	53.43%	48.40%
Mbya Guarani	5	60.00%	57.33%	51.46%
Urdu	69	85.51%	62.27%	65.23%
Romanian	1002	79.84%	48.65%	48.86%
Persian	308	72.73%	54.05%	50.27%
Japanese	381	61.15%	50.06%	33.54%
Hungarian	73	60.27%	44.39%	40.16%
Hindi	202	83.66%	61.24%	60.74%
Classical Chinese	1166	79.42%	60.76%	63.84%
Faroese	114	71.93%	40.13%	32.51%
Sanskrit	106	74.53%	57.49%	51.74%

Continued on next page

language name	total	tr verb percent	4 mean	4 weighted mean
Arabic	407	72.48%	46.74%	39.65%
Wolof	155	87.10%	55.11%	53.52%
Bulgarian	348	81.32%	47.10%	47.64%
Cantonese	45	73.33%	55.92%	50.99%
Pomak	229	72.49%	63.20%	61.76%
Old Church Slavonic	211	54.50%	46.60%	43.49%
Upper Sorbian	10	70.00%	40.82%	38.59%
Danish	212	88.68%	54.47%	51.92%
Komi Zyrian	25	32.00%	28.64%	25.52%
Manx	25	72.00%	47.98%	22.92%
Afrikaans	117	82.91%	60.90%	65.13%
Belarusian	587	68.14%	49.27%	44.41%
Coptic	127	65.35%	41.30%	43.55%
Serbian	214	73.36%	47.94%	38.45%
Western Armenian	299	75.92%	54.87%	50.66%
Scottish Gaelic	53	77.36%	37.74%	16.17%
Uyghur	80	53.75%	37.84%	32.95%

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).

"Lexico-Grammar" (1998). In: *Corpus Linguistics: Investigating Language Structure and Use*. Ed. by Douglas Biber, Randi Reppen, and Susan Conrad. Cambridge Approaches to Linguistics. Cambridge: Cambridge University Press, pp. 84–105. ISBN: 978-0-521-49957-6. DOI: 10.1017/CBO9780511804489.005. URL: https://www.cambridge.org/core/books/corpus-linguistics/lexicogrammar/8849B881ECF55E7890631A75C0515B59 (visited on 04/04/2023).

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.

Collina, S., P. Marangolo, and P. Tabossi (2001). "The Role of Argument Structure in the Production of Nouns and Verbs". In: *Neuropsychologia* 39.11, pp. 1125–1137. ISSN: 0028-3932. DOI: 10.1016/s0028-3932(01)00058-6. pmid: 11527549.

Croft, William (Nov. 1, 2016). "Typology and the Future of Cognitive Linguistics". In: *Cognitive Linguistics* 27.4, pp. 587–602. ISSN: 1613-3641. DOI: 10.1515/cog-2016-0056. URL: https://www.degruyter.com/document/doi/10.1515/cog-2016-0056/html?lang=en (visited on 06/14/2023).

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. URL: https://www.jstor.org/stable/415037 (visited on 09/18/2019).

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).

Gibson, Edward, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy (May 1, 2019). "How Efficiency Shapes Human Language". In: *Trends in Cognitive Sciences* 23.5, pp. 389–407. ISSN: 1364-6613. DOI: `10.1016/j.tics.2019.02.003`. URL: `https://www.sciencedirect.com/science/article/pii/S1364661319300580` (visited on 07/12/2022).

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`. URL: `https://www.jstor.org/stable/4178548` (visited on 11/22/2022).

– (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. URL: https://www.jstor.org/stable/411200 (visited on 11/22/2022).

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 (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).

Shapiro, L. P., E. Zurif, and J. Grimshaw (Dec. 1987). "Sentence Processing and the Mental Representation of Verbs". In: *Cognition* 27.3, pp. 219–246. ISSN: 0010-0277. DOI: 10.1016/s0010-0277(87)80010-0. pmid: 3691026.

Shetreet, Einat, Dafna Palti, Naama Friedmann, and Uri Hadar (Aug. 1, 2007). "Cortical Representation of Verb Processing in Sentence Comprehension: Number of Complements, Subcategorization, and Thematic Frames". In: *Cerebral Cortex* 17.8, pp. 1958–1969. ISSN: 1047-3211. DOI: 10.1093/cercor/bhl105. URL: https://doi.org/10.1093/cercor/bhl105 (visited on 06/11/2023).

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 Confer-*

ence 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. URL: https://www.jstor.org/stable/4175793 (visited on 09/25/2022).

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).

Zipf, G.K. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Philology.* The Psycho-Biology of Language: An Introduction to Dynamic Philology. Oxford, England: Houghton Mifflin.

Zipf, George Kingsley (1949). *Human Behavior and the Principle of Least Effort.* Human Behavior and the Principle of Least Effort. Oxford, England: Addison-Wesley Press, pp. xi, 573. xi, 573.