Saarland University
Department of Language Science and Technology
Faculty of Humanities

Master's Thesis

Exploring Cross-Linguistic Patterns in Verbal Valency

A Quantitative Typological Study

Siyu Tao

July 31, 2023

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

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

UNIVERSITÄT
DES
SAARLANDES

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 have long occupied a central place in the study of language structure, as it stands at the interface between syntax and semantics, as well as between grammar and lexicon. However, the cross-lingual comparison of valency systems and their general typology have proved challenging due to both disagreements on the basis of comparison and the difficulty in arriving at categorical types. This thesis advocates a quantitative and functionalist approach to tackling this challenge. Using morphosyntactically annotated data from the Universal Dependencies (UD) treebanks, it experiments with corpus-based transitivity ratio metrics and novel entropy-based measures for valency system analysis. These experiments reveal areal and genetic patterns of transitivity among languages and suggest universal effects of learnability and efficiency on valency systems. In doing so, it hopes to contribute to a better understanding of verb valency typology and its underlying drivers.

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 9
   3.1. Universal Dependencies . . . 9
   3.2. Data selection . . . 10

4. Comparing transitivity across languages 11
   4.1. From transitivity categories to transitivity ratios . . . 11
   4.2. Experiment 1: Transitivity at token and lexeme levels . . . 12
      4.2.1. Introduction . . . 12
      4.2.2. Methodology . . . 13
      4.2.3. Results and discussion . . . 14
      4.2.4. Genetic and areal patterns in transitivity . . . 16

5. Entropy-based measures of verbal valency 21
   5.1. Valency frame, frequency and the efficient organization of the lexicon . . . 22
   5.2. Extracting valency frame encoding from UD . . . 23
      5.2.1. Dependency relations . . . 24
      5.2.2. Feature extraction . . . 24
   5.3. Experiment 2: Frequency correlation with valency frame entropy 25
      5.3.1. Introduction . . . 25
      5.3.2. Methodology . . . 26
      5.3.3. Results and discussion . . . 27
   5.4. Experiment 3: Word order or case? Cross-lingual variation in valency encoding strategies . . . 30
      5.4.1. Introduction . . . 30
      5.4.2. Methodology . . . 30

5.4.3. Results and analysis 31
5.5. Experiment 4: Frequency correlation with verb entropy 31
5.5.1. Introduction 31
5.5.2. Methodology 32
5.5.3. Results and analysis 32

6. Conclusion and outlook 35
6.1. Summary 35
6.2. Limitations and future work 35
6.3. Reproducibility 35

A. Experiment 1 Results 37
A.1. Per-language transitivity statistics 37

B. Experiment 2 Results 41
B.1. Per-language Spearman's rank correlation between verb frequency and valency frame entropy 41

Bibliography 45

Chapter 1.

Introduction

Verbal valency deals with the relationship between verbs and their arguments, such as subjects and objects. It occupies a central place in investigations into language structure, as it lies at the interface between syntax and semantics, as well as between grammar and lexicon. It also exhibits considerable cross-lingual variations. A typological study of verb valency is therefore crucial to getting a better picture of the range and pattern of variations as well as to understanding the underlying causes of the similarities and differences. However, the study of cross-lingual differences and similarities in valency systems is not without difficulties. This reflects a diversity of both theoretical and methodological challenges that result from the same central role valency plays in language structure.

On the theoretical side, valency frames is alternately viewed either as syntactic expressions of lexical semantics, where the argument structure is determined by verbal meaning, or constructions in their own right that interact with verbs and their arguments to produce sentence meaning. The corollary of this debate extends to the relationship between valency structures and the lexicon: whereas the former view favors a lexeme-based approach, where valency information is an inherent part of a verb's lexical entry, the latter advocates a frame-based approach, where the valency frames are separate constructions with its own semantic contribution associated with but not determined by the verb.

The theoretical debates inevitably have implications for typological research on valency and, of greater interest to typology, valency systems. Most prominently, this makes agreeing on a sound *tertium comparationis* (Latin for "the third part of comparison"), or shared point of comparison, difficult. A lexeme-based approach would argue for verb classes to be the basis of comparison between languages, while a frame-based approach would attempt to find cross-lingually valid valency frames and see how they are distributed in different languages. However, regardless of the view subscribed, features of valency systems on their own pose methodological challenges. This is namely due to

the fact that neither verb classes nor valency frames lend themselves to easy categorical types of how a language organize them, at least not in the same way a language can be said to have a SOV, SVO or OSV word order.

Such theoretical and methodological complexity also translates to challenges in formulating any expected cross-lingual similarities and differences and what may have caused them. This present thesis makes an attempt to address them. The goal of the study is twofold: (1) Firstly, it advocates for a quantitative and functionalist approach to the study of verb valency. By employing empirical and corpus-based methods, this approach remains relatively theory-agnostic and enables the characterization of valency systems without relying on rigid categorical types, which in turn facilitates cross-lingual comparison; (2) Secondly, it seeks to be explanatory as well and uses the methods developed to identify typological differences across languages as well the test hypotheses for possible universals motivated by view of language as having evolved for communication and thus language structure as reflecting pressures in the communication context.

The remainder of this thesis is structured as follows: §2 provides more background and theoretical framework on functional typology, dependency grammar, and valency grammar. The main data source, Universal Dependencies treebanks, is introduced in §3, where its suitability to typological research in general and compatibility with this study in particular are highlighted. §4 presents a first experiment using corpus-based versions of the transitivity ratio metric and shows that they reveal areal and genetic patterns of transitivity among languages. §5 introduces a further set of experiments using novel entropy-based metrics to measure the average surprisal of valency frames given the verb and vice versa. Hypotheses regarding the correlation between these metrics and verb or valency frame frequency are posited and tested, suggesting that learnability and efficiency effects shape valency structures of languages in a universal way. Lastly, §6 discusses the implications as well as limitations of the studies, suggests future directions, and concludes the thesis.

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-

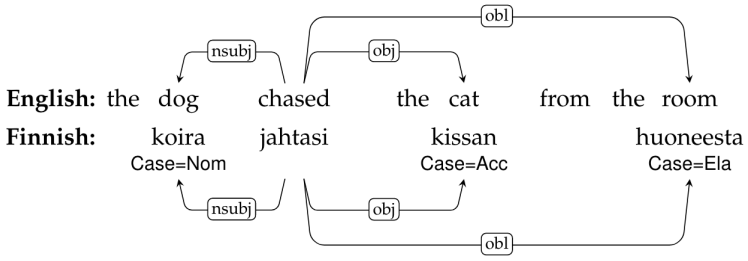

Figure 2.1.: Simplified UD annotation for equivalent sentences from English (top) and Finnish (bottom) (de Marneffe, Manning, et al., 2021).

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 is one aspect of valency that has received particular attention for cross-linguistic comparison. 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

3.1. Universal Dependencies

Universal Dependencies (UD) is the main source of primary data used for the present study. It is designed as 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.

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

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).

See 3.1 for a table of languages available in UD v2.5, as an example. The thesis uses the UD v2.11 release, which covers 138 languages with 243 treebanks.

3.2. Data selection

Explain the choices made in data selection, exclusions of certain treebanks e.g. L2 speakers, code-switch, lemmatization issues, unavailable texts

Chapter 4.

Comparing transitivity across languages

4.1. From transitivity categories to transitivity ratios

It is logical when doing cross-linguistic comparison to start with simpler metrics and features before developing to more complex ones. On the one hand, it is true that regardless of how we approach the task of verb classification, i.e., to categorize the verbs of a language into verb classes according to their syntacto-semantic properties and behavior, we would expect to arrive at fine-grained verb classes à la Levin (1993) in the end. On the other hand, such an expectation does not render obsolete the more basic distinctions like that of verb *transitivity*. In part, this is due to their utility as convenient starting points of comparison for verbs within a language, but their simplicity also translates to being more cross-lingually valid bases of typological comparison.

This first experiment deals with metrics of *transitivity*, i.e., the ability of a verb to take one or more objects. It is surely a more familiar and intuitive concept as compared to the finer-grained metrics of valency to follow in the experiments in the next chapter. In traditional grammars, a basic binary distinction is made between *intransitive* verbs, which take only a subject and no objects and *transitive* verbs, which take one or more objects. Additional categories, some overlapping, make finer distinctions, such as *ditransitive* verbs (those taking two objects), *ambitransitive* verbs (those that can be used both transitively and intransitively), etc.

I focus here on the simple distinction between transitive and intransitive verbs, but, hewing to a functional and quantitative outlook, find that binary categories do not sufficiently capture the nuances of verb use. An example illustrating why comes from the study of near-synonyms: (Biber et al., 1998), an early corpus linguistics study, compares the English verbs *begin* and *start* in the British National Corpus (BNC). 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 their different usage patterns as observed in the 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 levels of transitivity, I propose measuring **transitivity ratios** based on language corpora, defined as the percentage of verb instances that are transitive. The verbs of a language can then be thought of as on a spectrum of transitivity, with strictly intransitive verbs on the one end, strictly transitive verbs on the opposite end, and all other verbs[1] somewhere in between.

4.2. Experiment 1: Transitivity at token and lexeme levels

4.2.1. Introduction

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 used here, the absolute values of these

[1]Note that not all of them are ambitransitive verbs, as the approach here is only concerned with the surface realization of transitivity. In particular, this means instances of pro-drop of objects are counted as intransitive, where some would argument for a null object analysis instead.

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. I 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 justifies the cross-lingual comparison of transitivity ratios at a token-level as well.

4.2.2. Methodology

The UD annotation scheme facilitates the investigation of 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 nominative and the accusative in languages with a case system) or semantic roles (even though they would typically be the proto-agent and the 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 laudable effort.

Despite the typologically sound UD dependency relation annotations, arriving at a clear definition of transitivity ratio in the UD context is still not trivial upon close examination. In fact, I 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 and Catalan.

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 the opposite problem surfaces, where subject-dropping languages are likely to have a smaller denominator, resulting in a high transitivity ratio that is not representative. Def. 3 drops 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.

#	**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 §4.2, represented with feature matrices

While there is a strong case for Def. 4 being the most principled definition, I 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.

4.2.3. Results and discussion

Transitivity ratio statistics for each verb lexeme based on the definitions are first compiled. From there, per-language statistics that will become the basis for our cross-lingual comparison are computed: I 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 the transitivity ratio metrics, 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), is calculated for comparison purposes, as it should correspond better with the traditional binary distinction between transitive and intransitive verbs.

I perform the experiment on the selected subset of UD data as described in §3.2. For the analysis, I include only languages with at least 50 observed verb lexemes (69 out of 79 languages); the full results from the experiments can be found in the accompanying data.

def.	lexeme tr., token tr.	tr. verb %, lexeme tr.	tr. verb %, token tr.
1	$\rho(67) = .76, p = .000$	$\rho(67) = .50, p = .000$	$\rho(67) = .65, p = .000$
2	$\rho(67) = .85, p = .000$	$\rho(67) = .20, p = .106$	$\rho(67) = .27, p = .025$
3	$\rho(67) = .78, p = .000$	$\rho(67) = .56, p = .000$	$\rho(67) = .67, p = .000$
4	$\rho(67) = .88, p = .000$	$\rho(67) = .70, p = .000$	$\rho(67) = .72, p = .000$

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 ($\rho(67) = .88, p = .000$) and lowest observed for definition 1 ($\rho(67) = .76, p = .000$). 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 (Spearman's ρ between all languages).

The correlation statistics between the transitive verb percentages and the transitivity ratios are more revealing. Definition 2 is eliminated as it does not show statistically significant correlation at lexeme level ($\rho(67) = .20, p = .106$) and weak correlation at token level ($\rho(67) = .27, p = .025$). Of the other three, only definition 4, our *a priori* favorite, show strong correlations at both token and lexeme levels ($\rho(67) = .70, p = .000$ and $\rho(67) = .72, p = .000$, respectively).

While this comparison is not conclusive, it indicates that definition 4 provides more robust results across the levels of measurement, providing circum-

stantial support to our *a priori* determination. For the sake of simplicity, the rest of the experiment are reported relying on the transitivity ratios as defined in definition 4.

Further intralinguistic analyses are done 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, they are notably rarely symmetric, with differing levels of skew towards either end.

4.2.4. 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 token-level transitivity ratios are favored over lexeme-level ones for cross-lingual comparison. Appendix A lists the full results.

Among the most transitive languages are Romance (Catalan, Spanish, French, Galician) and most Germanic (German, Norwegian, English, Danish, with the exception of Dutch) 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.

To look at any potential areal patterns in transitivity, the token-level transitivity ratio results for European languages are also mapped in Fig. 4.2. A less Eurocentric study of areal patterns is unfortunately difficult due to corpus size constraints of the annotated UD. A particularly high transitivity area can be observed 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.

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

Language	# Verbs	Tr. verb %	Language	# Verbs	Tr. verb %
Catalan	628	99.2%	Maltese	78	55.1%
Galician	341	97.7%	Hebrew	536	58.4%
Urdu	69	97.1%	Hungarian	73	64.4%
Hindi	207	97.1%	Russian	2583	65.0%
Spanish	948	96.9%	Slovak	284	66.5%
Indonesian	330	96.7%	Uyghur	93	66.7%
Vietnamese	156	95.5%	Coptic	128	68.0%
French	735	94.8%	Old Church Slavonic	223	68.6%
Gheg	50	94.0%	Polish	1080	68.8%
Danish	212	93.9%	Bambara	50	70.0%
English	773	93.4%	Erzya	87	70.1%
Afrikaans	119	93.3%	Latvian	794	71.2%
Norwegian	765	92.9%	Gothic	201	73.1%
Basque	255	92.9%	Old French	444	74.1%
Ancient Greek	1127	92.7%	Faroese	117	74.4%
…	…	…	…	…	…

(a) by transitive verb percentage

Language	# Verbs	Token tr.	Language	# Verbs	Token tr.
Akkadian	76	75.9%	Scottish Gaelic	53	16.2%
Catalan	628	75.9%	Irish	108	30.8%
Galician	341	70.5%	Maltese	78	31.7%
Afrikaans	119	65.3%	Faroese	117	32.6%
Urdu	69	65.2%	Japanese	395	33.4%
Gheg	50	64.8%	Hebrew	536	33.5%
Vietnamese	156	64.8%	Polish	1080	34.0%
Thai	77	64.7%	Uyghur	93	34.1%
Classical Chinese	1192	63.8%	Erzya	87	36.0%
Pomak	244	62.1%	Russian	2583	37.4%
Spanish	948	61.1%	Dutch	497	37.5%
Hindi	207	60.8%	Latvian	794	37.8%
Chinese	380	60.7%	North Sami	76	38.2%
Xibe	74	58.8%	Serbian	214	38.4%
Ancient Greek	1127	58.5%	Arabic	407	39.7%
…	…	…	…	…	…

(b) by token-level transitivity ratio

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

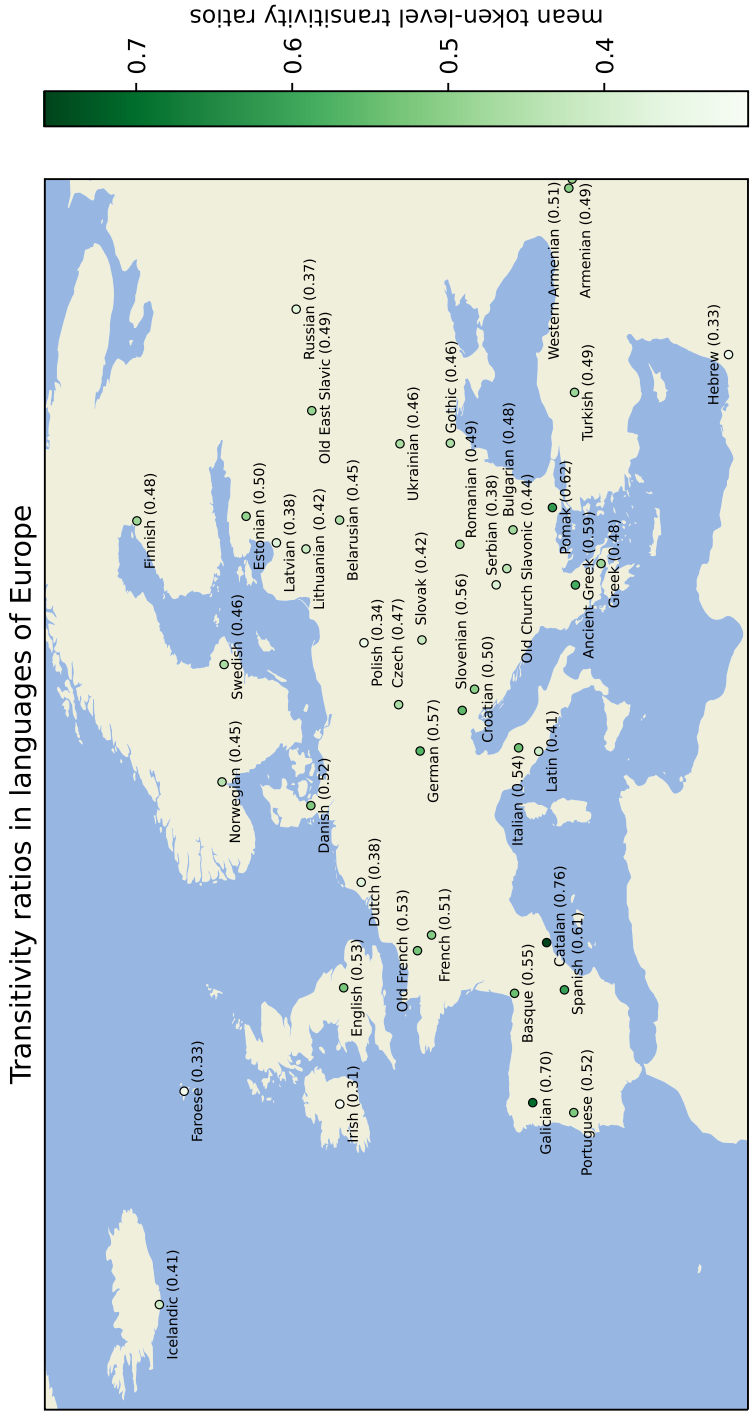

Figure 4.2.: Mean transitivity ratios in languages of Europe

Chapter 5.

Entropy-based measures of verbal valency

While the previous chapter has demonstrated that corpus-based transitivity ratios can serve as a useful basis for typological comparison, it is still is a relatively simple metric that captures one aspect of verbal valency. Put another way, transitivity ratio can be seen as a subset of the scope needed for a more holistic investigation into verbal valency, as it concerns only one type of verb dependent, i.e., the object, and one feature of this dependent, i.e., its presence or absence.

This chapter undertakes to expand the scope of investigation accordingly, to cover both more verb dependent types and more features of them. Instead of performing exhaustive comparisons on each individual feature of each dependent, the focus is on the variability of *valency frame alternations* (also called diathesis alternations), i.e., the range of valency frames in which a verb appears and how likely it is to appear with each frame. New entropy-based measures of verbal valency are proposed and tested in conjunction with hypotheses about the structure of valency systems.

The sections are organized as follows: §5.1 provides a brief background on cognitive and psycholinguistic perspectives on language structure, in particular relating to the frequency effect and valency frame alternation, which motivate the hypotheses tested in the experiments; in §5.2, the feature selection and extraction procedures of valency frame encoding from UD are described, as they form the basis for the experiments in this chapter.

Four experiments are then presented: Experiment 2 (§5.3) uses joint entropy to define the valency frame entropy measure and tests its correlation with verb frequency; Experiment 3 (§5.4) examines cross-lingual variations in how languages encode the valency frames, by means of an ablation study using conditional entropy to assess the contribution of word order and case marking to the overall valency frame entropy; Experiment 4 (§5.5) takes a frame-based approach and, in symmetry to Experiment 2, calculates the correlation between

verb entropy and frame frequency.

5.1. Valency frame, frequency 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 Zipf (1935, 1949). He first studied what is now known as Zipf's law, the empirical observation of the negative correlation between word length and frequency, 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.

Frequency is a particularly frequent lens through which the lexicon is examined and the correlation between frequency and other features often subjects of hypotheses. Its importance has been further underlined by psycholinguistic studies which show a consistent "frequency effect" for both open and closed class word (Marslen-Wilson, 1990; Segui et al., 1982), where more frequently occurring words have a higher resting activation, making their lexical retrieval easier. That the most frequent lexical items are also more likely to be associated with irregularity is not surprising. This correlation between frequency and irregularity has most often been hypothesized and studied for morphology Wu et al. (2019). Bybee (1998) considers lexicon from a learnability perspective and postulates a trade-off in the lexical memory that "being easier to access, they are less likely to be replaced by regular formations".

When it comes to verbal valency, psycholinguistic studies have also consistently shown the effect of semantic and syntactic attributes of the verb on online sentence processing (Collina et al., 2001; Shapiro et al., 1987). Results differ on whether subcategorization (syntactic), thematic frames (semantic) or both have an effect on lexical processing. Shapiro et al. (1987) reported that RTs for lexical decisions increased as the function of the number of thematic options instead of subcategorization options. In contrast, Shetreet et al. (2007) reported on an fMRI study on Hebrew speakers, shows that the number of options in terms of subcategorization and thematic frames is better correlated to

UD label	Dependent
core arguments	
NSUBJ	nominal subject
OBJ	object
CSUBJ	clausal subject
CCOMP	clausal complement
XCOMP	open clausal complement
IOBJ	indirect object
non-core dependents	
OBL	oblique nominal
EXPL	expletive
ADVMOD	adverbial modifier
ADVCL	adverbial clause modifier

Table 5.1.: UD dependency relation labels included in valency frame extraction

activity in the cortical areas that are associated with linguistic processing, as opposed to the number of complements or thematic frames.

In the following experiments, I will test a few hypotheses between frequency metrics on the one hand and entropy metrics on the other hand and show that, regardless of the theoretical stance, how a language structures its valency system reflects trade-offs predicted by considerations of efficiency and learnability.

5.2. Extracting valency frame encoding from UD

As briefly discussed in the data chapter, UD annotations make available a range of information related to the morphosyntactic encoding of valency frame, beyond the transitivity information that we have already used. The following experiments in this chapter make use of this for a more fine-grained characterization of the variation in valency frame encoding for different verbs using entropy-based measures. To do so, **valency frame encoding**, i.e., the scope of the morphosyntactic features that compose a valency frame, must first be extracted from the UD annotations so that variation patterns can be consistently captured. The extraction procedures are described in this section.

5.2.1. Dependency relations

From UD annotations, I start by determining which dependency relations of the verb to include as part of the valency frame. The distinction between argument (complement) and adjunct is a well-established one in linguistics, the former being obligatory and the latter optional. UD annotations schema (de Marneffe, Dozat, et al., 2014), including in the up-to-date v2 guidelines[1], makes the distinction between *core arguments* (i.e. subject and object) and everything else (called *non-core dependents*) instead. All core arguments as classified by UD are included in the analysis. This includes nominal dependents (*nominal subject*, *object*, *indirect object*) as well as clausal dependents (*clausal subject*, *clausal complement*, *clausal complement*).

As the non-core dependents still include arguments which complete the verbal meaning (in particular, *obliques*), and, as a priori distinctions between arguments and adjuncts are unnecessary for this study, possibly even counterproductive, given the quantitative experiment design, a subset of non-core dependents are included as well, namely oblique nominal, expletive, adverbial modifier and adverbial clause modifier. Other non-core dependents from UD are excluded for various reasons, either due to relatively low cross-lingual uniformity in interpretation (e.g., *dislocated element*), or due to being suprasentential elements (e.g. *vocative*, *discourse element*). Tab. 5.1 shows the list of dependents included in the study.

It is important to note that the basis of cross-linguistic comparison will be the taxonomy of the dependencies and the valency frames they compose. The validity of the study is therefore predicated on the cross-lingual validity of the UD relations, which, while certainly not perfect, is as good as one can do, given that UD is designed with it in mind, but otherwise agnostic as far as linguistic theory is concerned. In other words, it does not matter whether the UD category *indirect object* corresponds to the traditional grammatical category of indirect object, in so far as the cross-linguistically dependents serving equivalent functions are consistently annotated.

5.2.2. Feature extraction

For each of dependent relations in Tab. 5.1, I extract three features: (1) the presence or absence of dependency relations attached to a verb token, (2) the relative word order information, whether specific dependents precede or fol-

[1] https://universaldependencies.org/u/dep/, archived on 30-07-2023 at https://web.archive.org/web/20230730071650/https://universaldependencies.org/u/dep/

low the head verb, and (3) morphological case marking on the dependents, if any. In terms of implementation, the valency frame of each verb token is represented in a feature array encoding the three types of feature with the size of 3× the number of dependents. Any verbs that share the same feature array are said to have the same valency frame.

5.3. Experiment 2: Frequency correlation with valency frame entropy

5.3.1. Introduction

This experiment introduces a valency frame entropy metric, conditioned on the verb, that measures the average amount of surprisal, i.e., uncertainty, associated with the valency frame alternation for a verb, and hypothesize a positive correlation between a verb's frequency and the valency frame entropy conditioned on it that should hold across languages. In other words, the more frequent a verb is, the more information content one can expect on average from its valency frames.

A learnability perspective on the lexicon provides one motivation behind the hypothesis. Taking the view that the lexicon is acquired from linguistic experience, more exposure to and access of the more frequent words leads to higher resting activation (cf. Bybee, 1998), therefore allowing for more complexity or uncertainty being retained in the lexicon. This is analogous to the correlation between word frequency and irregularity in morphological patterns. However, whereas morphological irregularity is a purely formal feature, the choice of valency frame often entails a semantic choice as well.

If viewed from a production and comprehension perspective, the hypothesis is also potentially relevant to the Uniform Information Density (UID) hypothesis (Fenk and Fenk-Oczlon, 1980; Levy and Jaeger, 2006), which posits that language speakers prefer a more even distribution of surprisal values across utterances in order to maximize but not overload the capacity of the communication channel. Less frequent verbs would already have high surprisal, having high entropy in the valency frames associated with it is undesirable due to channel capacity constraints. I note, however, that UID hypothesis has mostly focused on surprisal at a token/phoneme-level. This is the case for the verb in question, but the valency entropy measures aspects of the morphosyntax of the sentence instead and requires a clearer formulation of the relationship between the frames and the tokens that compose it. Nevertheless, at a high level the hypothesis here is congruent what is expected given the UID hypothesis.

5.3.2. Methodology

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 $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$, which represent the possible 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 $\mathcal{V}$.

The entropy of each dependency relation given a specific verb $v \in \mathcal{V}$ quantifies the average surprisal associated with that dependency relation. The **dependency relation entropy** for X_k is defined as:

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

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 **valency frame entropy** is formalized as the joint entropy of the relevant UD dependency relations, i.e., number of bits needed to encode the entire valency frame. Denoted as $H_{\text{joint}}(X_1, X_2, \ldots, X_n|V = v)$, it 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:

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

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 $x_1, x_2, \ldots, x_n$.

In practice, for this study, the valency frame entropy is calculated as cross-entropy, following other studies using entropy measures (Hahn et al., 2021), in an effort to reduce artifacts introduced by data sparsity for rare frames. Treebanks of the same language are combined and then split randomly into two halves, resulting in two sets of distributions $X_1, \ldots, X_n$ and $X'_1, \ldots, X'_n$. It follows that the joint cross-entropy between them is:

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

The difference from the standard entropy measure is that the probabilities and their logarithms are estimated from the two different distributions with the same image. Laplace smoothing is used to prevent fringe cases where a frame is observed in only one of the two distributions.

The correlation between verb frequency and valency frame entropy as conditioned on it is assessed using Spearman's rank correlation coefficient (Spearman, 1904), which measures the correlation between two rank variables.

As a simple, related metric that direct measures the range of valency frame alternations, but does not take into the relative frequency of different frames into account, I calculate also the number of valency frames each verb is associated with for comparison and perform Spearman's rank correlation coefficient between verb frequency and number of frames.

At first glance, there may be concerns about circularity with the correlation: the verb frequency is the first variable, but it is also the number of observations made for estimating the second variable. This is in part mitigated by the use of cross-entropy, as the probabilities of one frame appearing and the average surprisal of that frame are estimated from two separate splits of the corpora. To further address such concerns, a subsampling experiment is performed where I take subsamples of a fixed size (the subsampling threshold) for all verbs with frequency above the threshold and verbs with frequency below the subsampling threshold are not included in the analysis. As corpus size varies dramatically between languages, the subsampling threshold also cannot be one-size-fits-all. I use a heuristically determined subsampling ratio of 0.1 but capped at a maximum of 25 samples. In this way, a lower-resource language such as Greek will see a threshold of 18, as determined by 0.1 × the frequency of the most frequent verb *μπορώ* (186), whereas a higher-resource language such as English will see a fixed threshold of 25 instead of 313, as it would have been by 0.1 × the frequency of the most common verb *have* (3134).

5.3.3. Results and discussion

The results of the valency frame entropy calculations with the full test set. are plotted into scatter plots in Fig. 5.1, where each dot represents a single lexeme with frequency rank on the x-axis and valency frame entropy on the y-axis for each of the languages. The dots are additionally colored by the number

of valency frames associated with the verb. To improve plot legibility, only verbs with frequency rank below 1000 are plotted. A visual inspection suggests a clear negative relationship across languages between frequency rank and valency frame as well a negative relationship between frequency rank and number of frames, i.e. positive relationship between frequency and the respective variables. Notably, however, the plots for Japanese and to a slightly lesser extent Chinese and Classical Chinese show a significant number of outliers where high frequency verbs nevertheless have lower valency frame entropy values.

Spearman's rank correlation results show robust correlation between frequency and valency frame entropy. I include again only languages with at least 50 verbs for which valid entropy measures can be calculated. Strong to very strong correlations (defined as $\rho \geq .70$, following Schober et al., 2018) are observed in 48 out of 59 languages and moderate correlations ($\rho \geq .40$) are observed in 58 out of 59 languages, with Japanese being the only exception. The mean ρ value is 0.78, with a standard deviation of 0.14.

Subsampling results help guard against possible circularity by using the same sample size for each verb, at the expense of possibly underestimating entropy for more frequent verbs in particular. One further language (Mbya Guarani) is excluded from subsampling results due to data sparsity. As expected, subsampling results in decreased correlation strength, but 50 out of 59 languages still show moderate correlation strength. The mean ρ value is 0.53, with a standard deviation of 0.14. That the standard deviation remains almost the same indicates that subsampling has a relatively uniform impact across the languages.

Full correlation results are shown in Appendix B.1. Scatter plots using subsampling results per verb are shown in Fig. 5.2. For some languages, the leveling effect of the subsampling is more visible for higher frequency verbs (e.g. Czech, Romanian), whereas for others, the effect is more uniform across frequencies (e.g. Russian, German).

As can be observed from the scatter plots, the number of frames also shows a strong correlation with verb frequency. In fact, in all the 59 languages studied, the number of frames is more strongly correlated with frequency than valency frame entropy does. This is, however, not to be taken as evidence that the number of frames is a better metric for valency than the valency entropy, because it does not take into account how often the verb appears with a certain valency frame. The high correlation between number of frames and verb frequency may also be taken to indicate that the number of frames are more prone to being affected by the number of observations made hence being a mere proxy of the frequency. While a correlation is indeed expected between frequency and

metrics of the valency frame alternation, it stands to reason that more factors, in particular semantics, may affect it as well. A future study with a careful consideration of other factors and using mixed effect models may be better able to explain the results.

Overall, the results strongly support my hypothesis of a positive correlation between a verb's frequency and the valency frame entropy as conditioned on it. It is evidence in support of features of language and lexical structure being shaped by learnability, efficiency and other pressures derived from the communicative function, further bolstered by the cross-lingual applicability of the results.

However, this does not absolve the need for further investigations, in particular due to the variation in correlation strength between different languages. Several reasons may cause this, all of which require further investigative work either into the UD annotation practices for specific languages or into other aspects of grammar and lexicon.

One possibility is that the UD annotation schema does not sufficiently account for cross-linguistic differences in argument encoding, or the grammatical categories as designed by UD are not well suited to that languages. This is not unrelated to the second possibility, that because the valency structures of a language interacts with other aspects of grammar, taking further factors into account may be necessary.

An example of this can be seen in the case of Chinese: many of the verbs that are the outliers in terms of being high frequency but having relatively lower entropy are from the category often termed coverbs, e.g., 自 'from', 隨 'follow / with', as their role in the lexicon straddles prepositions and verbs in English. They often serve semantic functions similar to prepositions while behave syntactically as a verb. As their semantic function are relatively narrow, however, so do they only appear much fewer valency frames than other verbs of similar frequency does. Such examples bring into question how strict the boundaries of word categories should be and a fuller picture of valency may well need to better situate the verb category within the overall lexicon.

5.4. Experiment 3: Word order or case? Cross-lingual variation in valency encoding strategies

5.4.1. Introduction

Typological differences regarding word order and case marking necessarily means that different languages will have to use different strategies for encoding valency frames. On the one hand, it is a straightforward matter that a language cannot use case marking to encode a valency frame if it does not have case marking and a language with more flexible word order is less likely to use word order to encode its valency frame. On the other hand, the trade-off between word order and case marking in languages has been a well-studied topic in typology, which naturally begets the question if such a trade-off can be observed in how languages encode their valency frames.

This experiment examines the cross-lingual variation in valency encoding strategies by testing two hypotheses. The first a relatively simple hypothesis, simply that languages use word order and case marking to encode valency information. The operationalized prediction is that the correlation between valency frame entropy and verb frequency confirmed in Experiment 2 should be weaker, if the valency frame encoding leaves out either word order information, case marking, or both. An ablation study will seek to confirm it.

The second hypothesis uses conditional entropy to quantify the contribution of word order and case marking information to entropy. If the trade-off does

5.4.2. Methodology

The valency encoding information is split into three sets of variables, with one set each storing information about dependency relation presence $X_1, \dots, X_n$, relative word order $Y_1, \dots, Y_n$, and case marking $Z_1, \dots, Z_n$, with the subscript number representing a UD argument slot (e.g. *nsubj*, *obj*, ...) and with images $\mathcal{X}_1, \dots, \mathcal{X}_n$, $\mathcal{Y}_1, \dots, \mathcal{Y}_n$, $\mathcal{Z}_1, \dots, \mathcal{Z}_n$ and variable V representing a verb in the lexicon $\mathcal{V}$. The full valency frame entropy as calculated in Experiment 2 is now represented as

$$
\begin{aligned}
&H_{joint}(X_1, \dots, X_n, Y_1, \dots, Y_n, Z_1, \dots, Z_n | V = v) \\
&= - \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_n \in \mathcal{X}_n} \sum_{y_1 \in \mathcal{Y}_1} \cdots \sum_{y_n \in \mathcal{Y}_n} \sum_{z_1 \in \mathcal{Z}_1} \cdots \sum_{z_n \in \mathcal{Z}_n} \\
&P(x_1, \dots, x_n, y_1, \dots, y_n, z_1, \dots, z_n | V = v) \log_2 P(x_1, \dots, x_n, y_1, \dots, y_n, z_1, \dots, z_n | V = v)
\end{aligned}
$$

For the correlation strength comparison, we calculate (1) valency frame entropy without word order information; (2) valency frame entropy without case marking information; and (3) valency frame entropy without word order and without case marking information. Spearman's rank correlation is then calculated between verb frequency and each of the three post-ablation valency measures.

The conditional entropy is calculated using the chain rule by subtracting the entropy of all other variables from the full entropy. For example, the conditional entropy of word order information will be calculated by subtracting the entropy of presence and case marking variables from the full valency entropy:

$$\begin{aligned} &H_{joint}(Y_1, \dots, Y_n | V = v, X_1, \dots, X_n, Z_1, \dots, Z_n) \\ =&H_{joint}(X_1, \dots, X_n, Y_1, \dots, Y_n, Z_1, \dots, Z_n | V = v) \\ &- H_{joint}(X_1, \dots, X_n, Z_1, \dots, Z_n | V = v) \end{aligned}$$

and likewise for the conditional entropy of case marking. The conditional entropy can be intuitively understood as the unique contribution of the variables towards the full entropy, not predictable from other variables. It can also be understood as the entropy of this variable minus mutual information shared with other variable. Conditional measures are then aggregated at a language level for cross-lingual comparison.

5.4.3. Results and analysis

Lower correlation strength are observed

Delta between them

5.5. Experiment 4: Frequency correlation with verb entropy

5.5.1. Introduction

So far the experiments have focused on calculating the valency frame entropy conditioned on verb choice. While this does not necessitate the lexeme-based view of valency, lexeme is still the level at which experiments are performed and analysis made. As the hypotheses themselves are motivated by constraints on language structure that derive from its communicative function, they should nevertheless be neutral with respect to which theory of valency one subscribes to.

Consequently, if one were to adopt the frame-based view of valency and insist on using the frame as the level of analysis, a symmetric hypothesis can be made: namely that given a new metric of verb entropy conditioned on valency frame, we would see a similar frequency effect where the entropy of lexical choice would correlate with the frequency of the valency frame.

This experiment undertakes to examine exactly that hypothesis and verify the expected symmetry.

5.5.2. Methodology

Given the same variable definitions as in experiment 2, the entropy value for the verb given a single slot is defined as

$$\begin{aligned} &H(V|X_1 = x_1, \dots, X_n = x_n) \\ &= -\sum_{v \in \mathscr{V}} P(V = v|X_1 = x_1, \dots, X_n = x_n) \log_2 P(V = v|X_1 = x_1, \dots, X_n = x_n) \end{aligned}$$

Here, $P(V = v|X_1 = x_1, \dots, X_n = x_n)$ represents the conditional probability of the outcome that the verb v from the vocabulary $\mathscr{V}$ is selected for the verb slot V given an already determined valency frame $X_1 = x_1, \dots, X_n = x_n$. The entropy of V is calculated by summing the products of these probabilities with their logarithms (base 2) taken, each corresponding to a possible outcome v_k.

The rest of the experiment is similarly done in symmetry to experiment 2. Spearman's rank correlation will be calculated with valency frame frequency on the one hand and verb entropy conditioned on the valency frame on the other.

5.5.3. Results and analysis

The results are as expected.

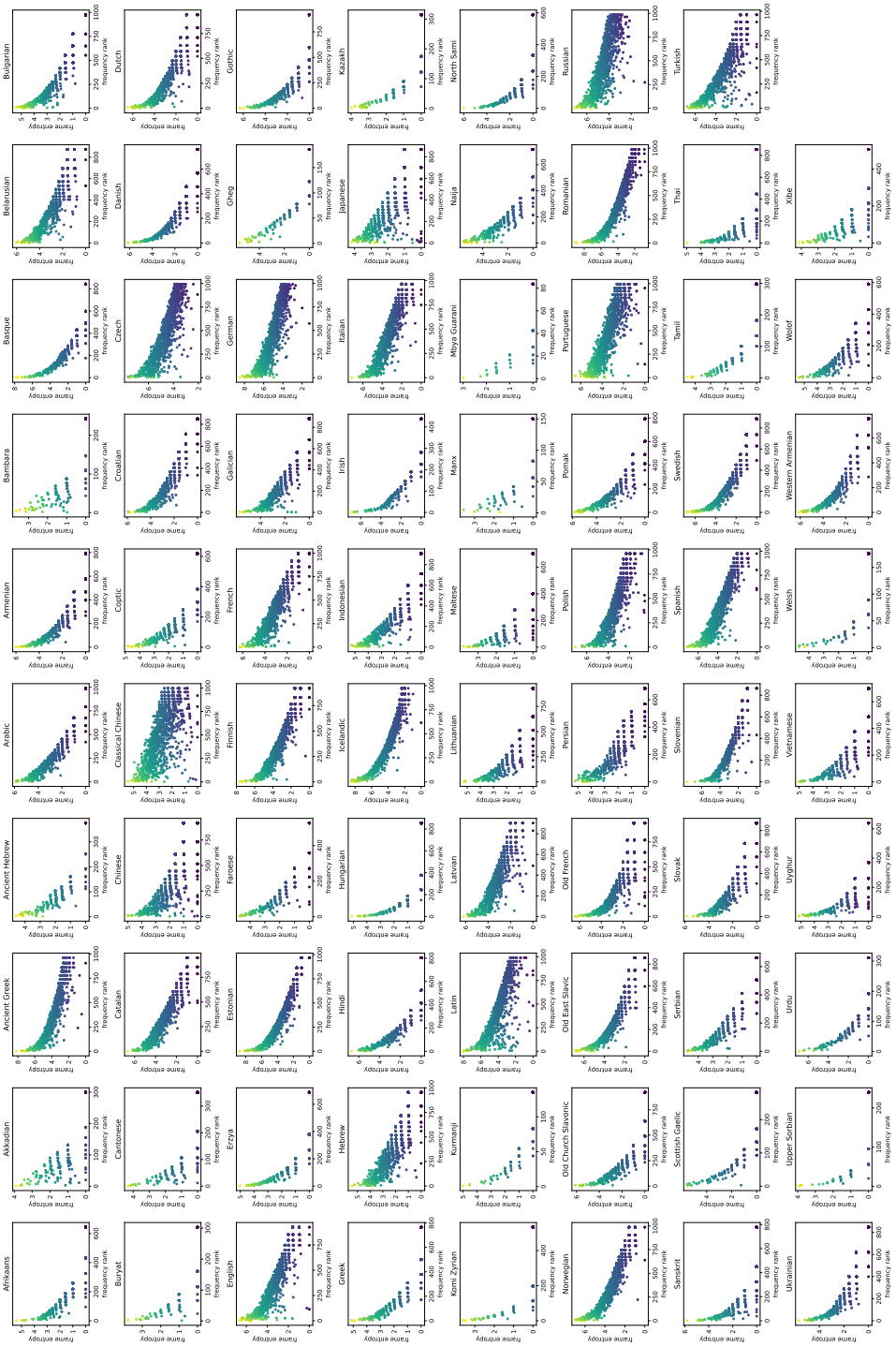

Figure 5.1.: Scatter plots showing relationship between valency frame entropy and frequency rank of verbs, color shows number of frames associated with the verb (yellow = higher, blue = lower)

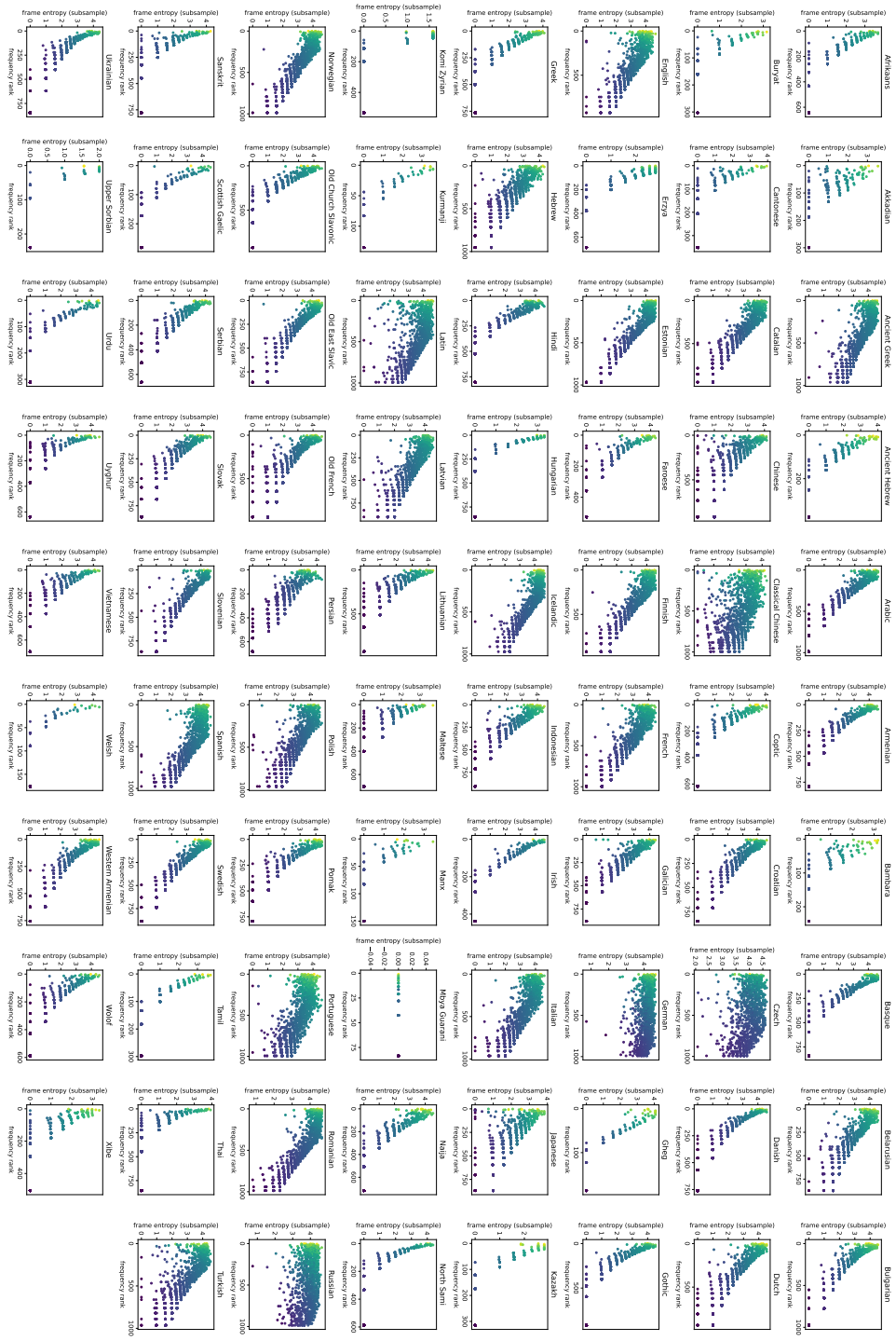

Figure 5.2.: Scatter plots showing relationship between valency frame entropy as estimated from the subsample and frequency rank of verbs, color shows number of frames associated with the verb

Chapter 6.

Conclusion and outlook

6.1. Summary

6.2. Limitations and future work

6.3. Reproducibility

Code and results for the thesis will be made available at the following repository: `https://github.com/siyutao/verbal-valency-ud`

Appendix A.

Experiment 1 Results

A.1. Per-language transitivity statistics

Language	# Verbs	Tr. verb %	Token tr.	Lexeme tr.
Afrikaans	119	93.3%	65.3%	61.2%
Akkadian	76	85.5%	75.9%	82.9%
Ancient Greek	1127	92.7%	58.5%	61.4%
Ancient Hebrew	90	84.4%	48.6%	53.9%
Arabic	407	76.9%	39.7%	46.7%
Armenian	228	86.0%	49.1%	51.7%
Bambara	50	70.0%	43.1%	54.1%
Basque	255	92.9%	54.5%	57.2%
Belarusian	597	80.1%	45.4%	50.0%
Bulgarian	352	88.6%	47.6%	47.4%
Catalan	628	99.2%	75.9%	81.2%
Chinese	380	88.2%	60.7%	61.0%
Classical Chinese	1192	91.3%	63.8%	61.4%
Coptic	128	68.0%	43.4%	41.3%
Croatian	388	90.7%	50.3%	58.4%
Czech	2091	78.7%	46.8%	54.8%
Danish	212	93.9%	51.9%	54.5%
Dutch	497	78.9%	37.5%	38.1%
English	773	93.4%	52.8%	55.6%
Erzya	87	70.1%	36.0%	40.4%
Estonian	676	75.1%	49.7%	53.7%
Faroese	117	74.4%	32.6%	40.7%
Finnish	690	76.4%	48.5%	54.4%
French	735	94.8%	51.3%	55.3%

Continued on next page

Language	# Verbs	Tr. verb %	Token tr.	Lexeme tr.
Galician	341	97.7%	70.5%	71.3%
German	2004	90.4%	56.8%	57.2%
Gheg	50	94.0%	64.8%	66.0%
Gothic	201	73.1%	46.1%	52.1%
Greek	146	85.6%	48.3%	50.7%
Hebrew	536	58.4%	33.5%	36.2%
Hindi	207	97.1%	60.8%	62.2%
Hungarian	73	64.4%	40.2%	44.4%
Icelandic	1000	91.3%	41.1%	51.9%
Indonesian	330	96.7%	51.2%	49.7%
Irish	108	88.9%	30.8%	54.2%
Italian	934	91.9%	54.4%	55.9%
Japanese	395	75.2%	33.4%	50.8%
Latin	1157	88.0%	40.9%	45.0%
Latvian	794	71.2%	37.8%	45.5%
Lithuanian	224	75.9%	41.9%	46.2%
Maltese	78	55.1%	31.7%	37.8%
Naija	203	91.6%	45.3%	54.9%
North Sami	76	77.6%	38.2%	50.0%
Norwegian	765	92.9%	44.8%	48.3%
Old Church Slavonic	223	68.6%	44.0%	48.6%
Old East Slavic	518	80.5%	49.3%	56.7%
Old French	444	74.1%	53.2%	53.8%
Persian	310	81.9%	50.3%	54.2%
Polish	1080	68.8%	34.0%	48.4%
Pomak	244	92.6%	62.1%	65.0%
Portuguese	1211	81.5%	52.4%	52.6%
Romanian	1007	91.1%	48.9%	48.8%
Russian	2583	65.0%	37.4%	43.7%
Sanskrit	108	88.0%	52.1%	58.3%
Scottish Gaelic	53	84.9%	16.2%	37.7%
Serbian	214	78.5%	38.4%	47.9%
Slovak	284	66.5%	42.2%	44.5%
Slovenian	517	88.6%	55.7%	64.5%
Spanish	948	96.9%	61.1%	65.5%
Swedish	391	90.5%	46.5%	47.8%

Continued on next page

Language	# Verbs	Tr. verb %	Token tr.	Lexeme tr.
Thai	77	87.0%	64.7%	56.4%
Turkish	787	86.3%	48.6%	47.8%
Ukrainian	233	75.1%	46.3%	51.5%
Urdu	69	97.1%	65.2%	62.3%
Uyghur	93	66.7%	34.1%	42.2%
Vietnamese	156	95.5%	64.8%	63.5%
Western Armenian	302	84.8%	50.7%	55.2%
Wolof	156	92.3%	53.6%	55.4%
Xibe	74	89.2%	58.8%	69.2%

Appendix B.

Experiment 2 Results

B.1. Per-language Spearman's rank correlation between verb frequency and valency frame entropy

Language	# Verbs	ρ	p-value	ρ (subsample)	p-value
Afrikaans	81	.73	.000	.58	.000
Ancient Greek	536	.87	.000	.59	.000
Arabic	206	.81	.000	.51	.000
Armenian	97	.84	.000	.66	.000
Basque	126	.95	.000	.66	.000
Belarusian	264	.79	.000	.57	.000
Bulgarian	121	.73	.000	.48	.000
Catalan	340	.85	.000	.54	.000
Chinese	125	.44	.000	.31	.000
Classical Chinese	683	.61	.000	.36	.000
Coptic	51	.52	.000	.28	.046
Croatian	137	.80	.000	.51	.000
Czech	1164	.89	.000	.55	.000
Danish	72	.91	.000	.77	.000
Dutch	198	.77	.000	.54	.000
English	419	.75	.000	.46	.000
Erzya	123	.90	.000	.72	.000
Estonian	390	.92	.000	.65	.000
Faroese	54	.72	.000	.55	.000
Finnish	395	.87	.000	.61	.000
French	411	.76	.000	.46	.000

Continued on next page

Language	# Verbs	ρ	p-value	ρ (subsample)	p-value
Galician	130	.54	.000	.42	.000
German	1250	.92	.000	.57	.000
Gothic	82	.82	.000	.55	.000
Greek	76	.73	.000	.57	.000
Hebrew	252	.65	.000	.43	.000
Hindi	116	.88	.000	.51	.000
Hungarian	73	.95	.000	.86	.000
Icelandic	547	.95	.000	.64	.000
Indonesian	159	.69	.000	.46	.000
Italian	535	.84	.000	.50	.000
Japanese	135	.21	.016	.06	.493
Kazakh	57	.87	.000	.69	.000
Komi Zyrian	137	.97	.000	.90	.000
Latin	693	.72	.000	.35	.000
Latvian	366	.79	.000	.54	.000
Lithuanian	75	.60	.000	.43	.000
Mbya Guarani	113	.71	.000	nan	nan
Naija	106	.66	.000	.32	.001
Norwegian	394	.91	.000	.60	.000
Old Church Slavonic	96	.76	.000	.51	.000
Old East Slavic	218	.87	.000	.64	.000
Old French	154	.78	.000	.59	.000
Persian	130	.75	.000	.40	.000
Polish	483	.78	.000	.53	.000
Pomak	114	.84	.000	.56	.000
Portuguese	675	.81	.000	.48	.000
Romanian	589	.93	.000	.65	.000
Russian	1327	.85	.000	.54	.000
Serbian	63	.73	.000	.53	.000
Slovak	99	.74	.000	.52	.000
Slovenian	220	.84	.000	.63	.000
Spanish	564	.85	.000	.45	.000
Swedish	172	.89	.000	.69	.000
Turkish	421	.70	.000	.33	.000
Ukrainian	56	.85	.000	.59	.000
Western Armenian	115	.86	.000	.62	.000

Continued on next page

B.1. Per-language Spearman's rank correlation between verb frequency and valency frame entropy

Language	# Verbs	ρ	p-value	ρ (subsample)	p-value
Wolof	60	.53	.000	.35	.006
Xibe	74	.69	.000	.51	.000

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).

Bybee, Joan (1998). "The Emergent Lexicon". In.

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, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning (May 2014). "Universal Stanford Dependencies: A Cross-Linguistic Typology". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. LREC 2014. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 4585–4592. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1062_Paper.pdf (visited on 11/24/2022).

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.

Fenk, August and Gertraud Fenk-Oczlon (Jan. 1, 1980). "Konstanz Im Kurzzeitgedächtnis - Konstanz Im Sprachlichen Informationsfluß?" In: *Zeitschrift für experimentelle und angewandte Psychologie* 27, pp. 400–414.

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.

Hahn, Michael, Judith Degen, and Richard Futrell (2021). "Modeling Word and Morpheme Order in Natural Language as an Efficient Trade-off of Memory and Surprisal". In: *Psychological Review* 128.4, pp. 726–756. ISSN: 1939-1471. DOI: 10.1037/rev0000269.

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.

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.

Levy, Roger and T. Florian Jaeger (Dec. 4, 2006). "Speakers Optimize Information Density through Syntactic Reduction". In: *Proceedings of the 19th International Conference on Neural Information Processing Systems*. NIPS'06. Cambridge, MA, USA: MIT Press, pp. 849–856.

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).

Marslen-Wilson, William (1990). "Activation, Competition, and Frequency in Lexical Access". In: *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*. ACL–MIT Press Series in Natural Language Processing. Cambridge, MA, US: The MIT Press, pp. 148–172. ISBN: 978-0-262-01117-4.

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).

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).

Schober, Patrick, Christa Boer, and Lothar A. Schwarte (May 2018). "Correlation Coefficients: Appropriate Use and Interpretation". In: *Anesthesia & Analgesia* 126.5, p. 1763. ISSN: 0003-2999. DOI: 10.1213/ANE.0000000000002864. URL: https://journals.lww.com/anesthesia-analgesia/Fulltext/

2018/05000/Correlation_Coefficients__Appropriate_Use_and.50.aspx (visited on 07/23/2023).

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).

Segui, Juan, Jacques Mehler, Uli Frauenfelder, and John Morton (Jan. 1982). "The Word Frequency Effect and Lexical Access". In: *Neuropsychologia* 20.6, pp. 615–627. ISSN: 00283932. DOI: 10.1016/0028-3932(82)90061-6. URL: https://linkinghub.elsevier.com/retrieve/pii/0028393282900616 (visited on 07/01/2023).

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).

Spearman, C. (1904). "The Proof and Measurement of Association between Two Things". In: *The American Journal of Psychology* 15.1, pp. 72–101. ISSN: 0002-

9556. DOI: 10.2307/1412159. JSTOR: 1412159. URL: https://www.jstor.org/stable/1412159 (visited on 07/23/2023).

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. 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).

Wu, Shijie, Ryan Cotterell, and Timothy O'Donnell (July 2019). "Morphological Irregularity Correlates with Frequency". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019. Florence, Italy: Association for Computational Linguistics, pp. 5117–5126. DOI: 10.18653/v1/P19-1505. URL: https://aclanthology.org/P19-1505 (visited on 07/20/2023).

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).

Zipf, George Kingsley (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.

– (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.