Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

Evaluation of Complex Typological Universals with Language Vectors and Real-Valued Logics

Tillmann Dönicke
Master Thesis

Examiner: Prof. Dr. Jonas Kuhn
Supervisor: Xiang Yu, M.Sc.

April 24, 2020

Erklärung (Statement of Authorship)

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet. Die eingereichte Arbeit ist weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen. Sie ist weder vollständig noch in Teilen bereits veröffentlicht. Die beigefügte elektronische Version stimmt mit dem Druckexemplar überein.[1]

Erklärung bei elektronischer Abgabe einer Abschlussarbeit

Hiermit bestätige ich, dass

- ich die Druckversion schnellstmöglich nachreiche.
- die Druckversion mit dem elektronischen Exemplar übereinstimmt.
- mir bekannt ist, dass Abweichungen zwischen dem elektronischen Exemplar und der Druckversion als Täuschungsversuch angesehen und entsprechend geahndet werden.

(Tillmann Dönicke)

[1] Non-binding translation for convenience: I declare that I have developed and written the enclosed thesis completely by myself and that I have not used sources or means without declaration in the text. Any thoughts from others or literal quotations are clearly marked. The thesis was not used in the same or in a similar version to achieve an academic grading or is being published elsewhere. The enclosed electronic version is identical to the printed version.

Abstract

Language representations and typological universals have received increasing attention in computational linguistics over the past few years. Most approaches make use of binary language vectors from typological databases and/or focus on the correlation of only two typological variables at the same time. This thesis shows that real-valued logics can be used to evaluate even more complex formulae, as they are formulated by typologists, on continuous vectors from existing corpora. Syntactic language vectors are extracted from the Universal Dependencies treebanks (Nivre et al., 2016) and serve as the basis for the evaluation of word-order universals.

Contents

List of Tables

List of Figures

List of Symbols

List of Syntactic Categories

Adj	adjective
AdjP	adjective phrase
Adp	adposition
Adv	adverb
Art	article
Aux	auxiliary
Comp	complementiser
Conj	conjunction
Conjunct	conjunct
Cop	copula
Dem	demonstrative
DemP	demonstrative phrase
DP	determiner phrase
Gen	genitive
iO	indirect object
N	noun
N'	noun projection
NP	noun phrase
Num	numeral
NumP	numeral phrase
O	object
PP	adpositional phrase
Pred	predicate
Rel	relative clause
S	subject
Sent	sentence
Sent'	complement clause
V	verb
VP	verb phrase
X	oblique

1 Introduction

In linguistic typology, language universals are statements that appear to hold for all natural languages. A simple example consisting of only one variable is given in (1).

(1) a. All languages have verbs.

b. $\forall \ell : \mathrm{V}(\ell)$

Universals are typically proposed by typologists and evaluated on as many languages as possible. Using logical connectives like conjunction ($\wedge$) or implication ($\Rightarrow$) allows to create an infinite number of potential universals, although they are usually kept simple (Occam's law of parsimony). The following examples are sorted by increasing complexity.

(2) a. All languages have verbs and nouns.

b. $\forall \ell : \mathrm{V}(\ell) \wedge \mathrm{N}(\ell)$

(3) a. "Languages with dominant VSO order are always prepositional." (Greenberg, 1963, #3)

b. $\forall \ell : \text{V-S-O}(\ell) \Rightarrow \text{Adp-NP}(\ell)$

(4) a. "In languages with prepositions, the genitive almost always follows the governing noun, while in languages with postpositions it almost always precedes." (Greenberg, 1963, #2)

b. $\forall \ell : \text{Adp-NP}(\ell) \Leftrightarrow \text{N-Gen}(\ell)$

(5) a. "When the descriptive adjective precedes the noun, the demonstrative and the numeral, with overwhelmingly more than chance frequency, do likewise." (Greenberg, 1963, #18)

b. $\forall \ell : \text{Adj-N}(\ell) \Rightarrow (\text{Dem-N}(\ell) \wedge \text{Num-N}(\ell))$

Although the formulation of such universals is quite simple, the evaluation is rather difficult. If one were to use Boolean values for the variables and a strict interpretation of the universal quantifier then most proposed universals would evaluate to false; for most candidates there will be at least one language which is an "exception to the rule". Typologists certainly do not have this interpretation in mind, as formulations like "dominant VSO order" in (3) or "more than chance frequency" in (5) suggest. Today's corpora make it possible to extract real statistics for the variables and calculate pairwise correlations and implications. However, a comprehensive framework for the evaluation of logically complex statements does not exist, so finding such a framework shall be the object of investigation in this work.

The thesis is composed of three parts. After introducing the reader to various types of language representations and typological variables in section 2, language vectors are put at the heart of section 3. That section describes how the language vectors used in this work are extracted from dependency treebanks and thereafter attends to the question of how well phylogenetic relatedness between languages can be modelled by different sorts of language vectors. The investigations mainly follow the recent works of Rabinovich et al. (2017), Östling and Tiedemann (2017), and Bjerva et al. (2019c). The final paragraphs of section 3 show how language vectors can be used to detect inconsistent annotations in treebanks and includes a manual analysis of the detected inconsistencies.

Section 4 reiterates real-valued logics and motivates their use for the evaluation of typological statements. The question of how one can deal with universal quantification over languages shall be answered under several aspects. These theoretical considerations on logical calculi mainly take up ideas from the early works of Łukasiewicz and Tarski (1930), Zadeh (1965), and Hempel (1945a). Concluding section 4, universals from a typological database are evaluated with the new framework and the results are compared to those from typological literature.[2]

In section 5, implication relations between two variables are evaluated in comparison to the findings of the similar works by Daumé and Campbell (2007), and Bjerva et al. (2019b). After that, logical formulae are mapped to a search space to seek new potential universals. The final paragraphs of section 5 provide a linguistic analysis of the high scored formulae found, inspired by the head-dependent analysis of Lochbihler (2017) but more sympathetic to the branching-direction theory of Dryer (2009a).

[2]The term "universal" is very vague. While "statement" and "formula" only refer to a syntactic expression of a proposition, "universal" further entails that the proposition can be evaluated as "(very) true". Neither is there a common definition of what "very true" means nor is there a unique way to assign truth values to logical formulae. Numerous ways to do so are presented in this thesis. Nevertheless, the term "universal" is used in this work for formulae that are claimed to be true in the typological literature.

2 Previous Work

2.1 Typological universals

The study of typological universals is usually traced back to the work of Greenberg (1963) and Comrie (1981). Greenberg (1963) manually compares over 30 languages and lists 45 universals—6 for general typology, 19 for word order, and 20 for morphology—which is still the most prominent set of universals, despite there being other collections of universals, such as the Universals Archive with over 2,000 entries (Plank and Filimonova, 2000; Plank and Mayer, 2006). The discovery and testing of potential universals has typically been performed "by painstaking hand analysis over a small sample of languages" (Daumé and Campbell, 2007), before large databases, such as the Word Atlas of Language Structures (WALS; Haspelmath et al., 2005; Dryer and Haspelmath, 2013), and annotated corpora, such as the Universal Dependencies treebanks (UD treebanks; Nivre et al., 2016), gave rise to computer-assisted methods. Daumé and Campbell (2007), and Bjerva et al. (2019a;b) use the WALS to test Greenberg's universals and discover new universals, and Lochbihler (2017), Gerdes et al. (2019), and Sharma et al. (2019) use the UD treebanks to inspect selected universals by Greenberg through the medium of corpus statistics. Recent work has also focused on universal tendencies concerning the optimality of natural languages, such as information locality (Futrell, 2019) and dependency length minimisation (Yu et al., 2019b; Kahane and Yan, 2019; Chen and Gerdes, 2019). This is, however, not relevant for typological universals in the strict sense.

2.2 Language vectors

Vector-like representations of languages are well-established in comparative linguistics and typology where they are used for language comparison and clustering. Namely, given a set of language vectors and a distance measure, languages can be clustered hierarchically. Many different forms of information have been used as features of the language vectors. Serva and Petroni (2008) use basic concepts as features with concrete words as values, and a modification of the Levenshtein distance as distance measure. Bjerva et al. (2019c) use dependency links (from dependency treebanks) as features with normalised counts as values, and the cosine distance as distance measure. The two methods aim at different aspects of distance between languages. Serva and Petroni (2008) intend to measure genetic distance between languages, whereas Bjerva et al. (2019c) intend to measure structural distance. The corresponding dendrograms in Figure 1 look accordingly different. For example, the Germanic languages form a cluster in Serva and Petroni's (2008) dendogram; but they are split in North Germanic languages (Danish, Swedish) and West Germanic languages (German, Dutch) in Bjerva et al.'s (2019c) dendrogram, with English being grouped to the North Germanic languages despite being West Germanic.

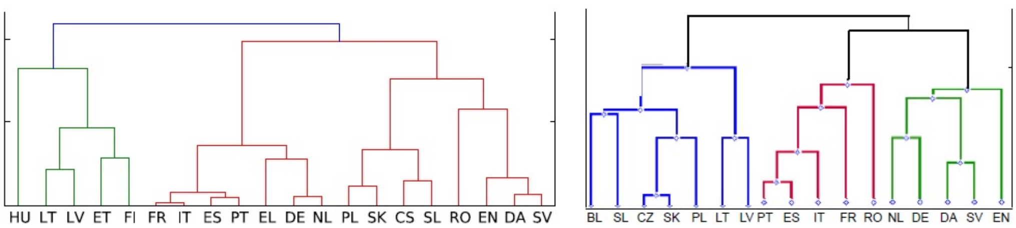


Figure 1: Dendrograms based on dependency link statistics on the left, and words on the right (taken from Bjerva et al., 2019c).

Bjerva et al.'s (2019c) kind of language vectors are so-called typological vectors since they encode typological features, e.g. concerning word order such as "subject-verb-object" or "noun-adjective". It is possible to calculate the corresponding statistics directly from corpora like the UD treebanks, as in Liu (2010), Chen and Gerdes (2017) or Bjerva et al. (2019c), or to use (mostly) binary/binarised information from large databases like the WALS, as in Itoh and Ueda (2004) or Littell et al. (2017). Littell et al. (2017) also append non-typological information to their language vectors, namely geographical and phylogenetic vectors. An example for the resulting language vectors is given in Table 1.

	Typological features				Phylogenetic features				
	SOV	SVO	Pref.	Suff.	ine	gem	gmw	gmq	roa
English	0	1	0	0	1	1	1	0	0
German	1	1	0	1	1	1	1	0	0
Vietnamese	0	1	0	0	0	0	0	0	0

Table 1: Binary language vectors for English, German and Vietnamese with typological and phylogenetic features: SOV, SVO, case prefixes, case suffixes; Indo-European, Germanic, West Germanic, North Germanic, Romance.

The theoretical background for typological features resonates with the "principles and parameters" framework (Chomsky, 1993; Chomsky and Lasnik, 1995). The idea is that there are universal properties shared by all languages (principles), e.g. a universal grammar, and that there are properties that are "on" or "off" for a specific language, e.g. whether the head should be placed left or right of the dependent. Modern typology, however, has shifted away from this theoretical side, and aims to describe and explain the distribution of typological parameters rather than seeking universal principles such as a universal theory of grammar (Bickel, 2007).

A completely different approach to language vectors was developed by Rabinovich et al. (2017). They look at English texts which have been translated from another source language and extract features like part-of-speech (POS) trigrams and function words with counts as values. The hierarchical clustering of their language vectors shows surprising similarity to the dendrogram of Serva and Petroni (2008) which they, as others, use as gold tree (see section 3.4).

In the recent past, distributed language representations, commonly known as language embeddings, gained some attention. Östling and Tiedemann

(2017) learn language embeddings from a multilingual language model and show that they can be used for hierarchical clustering. Malaviya et al. (2017) learn language embeddings from a many-to-one translation model and use them to predict syntactic, phonological and phonetic typological features. Bjerva and Augenstein (2018) fine-tune the language embeddings from Östling and Tiedemann (2017) on four different tasks for three linguistic levels: grapheme-to-phoneme prediction and phoneme reconstruction (phonology), morphological inflection (morphology), and part-of-speech tagging (syntax). The resulting embeddings mostly outperform the original embeddings in predicting typological features of the respective linguistic level. The hierarchical language clustering on the syntactically fine-tuned embeddings is also slightly better than with the original embeddings. Bjerva et al. (2019c) follow Rabinovich et al. (2017) and obtain language embeddings from training an English language model on texts in various languages and their translation into English. They compare the clustering results for different language embeddings trained on characters, POS sequences, dependency links etc. and find out that dependency-based language embeddings yield the best clustering results.

Table 2 summarises the language vectors discussed above.

Type of language vector	Value range	Reference
word list	$\{A, \ldots, Z\}^+$	Serva and Petroni (2008)
typological vector, binary	$\{0, 1\}$	Littell et al. (2017)
typological vector, continuous	$[0, 1]$	Liu (2010)
typological vector, unnormalised	$\mathbb{Q}$	Chen and Gerdes (2017)
phylogenetic vector	$\{0, 1\}$	Littell et al. (2017)
geographical vector	$[0, 1]$	Littell et al. (2017)
translational vector	$\mathbb{N}^0$	Rabinovich et al. (2017)
language embedding	$\mathbb{Q}$	Östling and Tiedemann (2017)

Table 2: Different types of language vectors, their value range and a reference for further reading.

2.3 Correlation, implication and causation

The search for typological universals, be it computationally driven or not, usually selects two or more typological variables and investigates their relationship. In the most basic case, these variables are vectors, containing values for the same languages in the same order, as the vectors for S-O-V ordering[3] and NP-Adp ordering (postpositions) in Table 3.

	S-O-V	NP-Adp
English	0	0
German	1	0
Hindi	1	1
Japanese	1	1
Vietnamese	0	0

Table 3: Typological variables: S-O-V order and postpositions.

The simplest case of a relationship might be an (undirected) correlation. Lochbihler (2017) applies Spearman's rank correlation coefficient to detect the strength of the correlation; an alternative are vector distance measures such as cosine or Euclidean distance. These measures, however, do not tell anything about the direction of the relationship. The values in Table 3 signify that NP-Adp ⇒ S-O-V, in the sense of Boolean algebra, since S-O-V is true whenever Adp-NP is true. But the implication does not hold vice versa. Lochbihler (2017) uses an additional measure, the difference in conditional probability $P(\text{S-O-V}|\text{NP-Adp}) - P(\text{NP-Adp}|\text{S-O-V})$, to determine the direction of the correlation. A value close to 0 indicates that the relationship holds in both directions (equivalence), whereas a value close to 1 or -1 indicates that the relationship is only a one-way implication. Daumé and Campbell

[3]In this thesis, the hyphenation in, e.g., "S-O-V" and "SOV" indicates a small difference: "S-O-V" is used when referring to a syntactic construction, whereas "SOV" is used when referring to some classes of languages. In an SOV language, the dominant order of object, subject and verb is S-O-V.

(2007) and Bjerva et al. (2019b), on the other hand, directly start off with the conditional probability $P(\text{S-O-V}|\text{NP-Adp})$, which measures the strength of an implication. To check whether the relationship is an equivalence, the measure must also be applied to swapped arguments.

Neither correlation nor implication entails causation. For the example above, it might be true that S-O-V order and postpositions frequently cooccur but this does not mean that the presence of postpositions causes S-O-V order. In Dryer (2009b), the author shows that a high correlation between tone system and the order of object and verb can be found in the WALS, namely Complex tone $\Rightarrow$ V-O, and traces this implication back to a geographical coincidence.

3 Construction and Clustering of Language Vectors

This section introduces continuous typological language vectors as they are used in this work. After describing their background, two sets of experiments follow:

1) Language vectors are used for hierarchical clustering and the results are compared to those from previous work.
2) Treebank vectors are used to compare and cluster treebanks and to detect annotation inconsistencies.

3.1 Definitions and requirements

Given a set of languages L and a set of properties P, a language–property matrix V has the size $|L| \times |P|$.[4] Rows are $|P|$-dimensional language vectors whereas columns are $|L|$-dimensional property vectors. The vector of the property $p \in P$ is denoted as $\vec{v}_p$ and its value for the language $\ell \in L$ as v_p^ℓ. For example, $v_{\text{HEAD-amod}}^{\text{English}} = 0.03$ states that the property "HEAD-amod" (see 3.2 below) has the value 0.03 for English.

This work especially requires vectors with 1) interpretable dimensions and 2) real values in the interval $[0, 1]$, since the framework developed in section 4 shall interpret these as real truth values. Typological vectors fulfil both requirements whereas e.g. language embeddings fulfil none of them.

[4] The terms "property" and "feature" both denote a measurable characteristic of an entity (here: a language) but have different connotations: "feature" is more associated with experimental characteristics in machine learning and "property" is used rather for theoretically grounded characteristics. A distinction is not necessary in this work and the term "property" is used for all kinds of characteristics.

3.2 Vectors from the Universal Dependencies treebanks

The Universal Dependencies (UD)[5] treebanks version 2.5 (Zeman et al., 2019) consist of 157 treebanks for 90 languages. A list of the languages is provided in appendix A.1. The sentence-level annotations in the treebanks include dependency relations and parts-of-speech (POS).

Three groups of properties are extracted from the UD treebanks: single-link properties, double-link properties, and chain-link properties. For single-link properties, each dependency relation is used as property where the value expresses the order of head and dependent. For example, (1) expresses the degree to which a language ℓ places an adjectival modifier after its governor.

$$v^{\ell}_{\textsc{head}\text{-amod}} = \frac{\#\left[\overset{\text{amod}}{\curvearrowright}\right](\ell)}{\#\left[\overset{\text{amod}}{\curvearrowleft}\right](\ell) + \#\left[\overset{\text{amod}}{\curvearrowright}\right](\ell)} \tag{1}$$

Here, an expression of the form $\#[*](\ell)$ returns how often the construction $*$ appears for ℓ. The value ranges between 0 and 1; 0 indicates that adjectives never succeed (i.e. always precede) their governor, 1 indicates that adjectives always succeed their governor, and 0.5 indicates that both orderings are equally common.

Table 4 shows a language–property matrix of 17 Indo-European languages and the 10 most frequent dependency relations (the matrix with all languages is provided in appendix A.2). The matrix already shows some well-known typological features. For example, the Romance languages have a strong tendency to place an adjective after its governing noun (cf. Gulordava and Merlo, 2015) and thus have higher values for the property "HEAD-amod", whereas the other languages have low values.

[5]https://universaldependencies.org/

	HEAD-punct	HEAD-nsubj	HEAD-case	HEAD-nmod	HEAD-obl	HEAD-obj	HEAD-advmod	HEAD-amod	HEAD-det	HEAD-cc
IE, Baltic, Latvian	0.51	0.24	0.02	0.16	0.45	0.68	0.13	0.00	0.01	0.00
IE, Baltic, Lithuanian	0.51	0.28	0.01	0.17	0.60	0.65	0.11	0.03	0.02	0.02
IE, Germanic, Danish	0.79	0.16	0.03	0.77	0.82	0.95	0.36	0.13	0.00	0.07
IE, Germanic, Dutch	0.62	0.21	0.02	0.81	0.46	0.41	0.24	0.04	0.00	0.01
IE, Germanic, English	0.71	0.04	0.04	0.63	0.85	0.96	0.24	0.03	0.00	0.00
IE, Germanic, German	0.60	0.22	0.01	0.92	0.31	0.38	0.21	0.00	0.00	0.00
IE, Germanic, Swedish	0.78	0.17	0.01	0.56	0.82	0.94	0.34	0.03	0.00	0.02
IE, Romance, French	0.67	0.06	0.00	0.98	0.83	0.85	0.34	0.70	0.00	0.00
IE, Romance, Italian	0.77	0.21	0.00	0.97	0.76	0.87	0.23	0.68	0.00	0.01
IE, Romance, Portuguese	0.62	0.10	0.01	0.93	0.79	0.91	0.24	0.69	0.01	0.01
IE, Romance, Romanian	0.54	0.27	0.01	0.93	0.76	0.66	0.20	0.88	0.19	0.01
IE, Romance, Spanish	0.59	0.17	0.01	0.97	0.77	0.79	0.28	0.73	0.01	0.01
IE, Slavic, Bulgarian	0.62	0.24	0.00	0.98	0.66	0.81	0.19	0.03	0.42	0.00
IE, Slavic, Czech	0.56	0.33	0.00	0.96	0.58	0.69	0.18	0.08	0.02	0.03
IE, Slavic, Polish	0.60	0.27	0.00	0.93	0.73	0.82	0.19	0.35	0.06	0.01
IE, Slavic, Slovak	0.69	0.32	0.00	0.85	0.61	0.67	0.17	0.04	0.02	0.01
IE, Slavic, Slovenian	0.55	0.27	0.00	0.95	0.55	0.59	0.23	0.03	0.00	0.00

Table 4: Relative frequencies for the 10 most frequent single-link properties, excluding the “root” relation, for 17 languages in the Universal Dependencies treebanks.

For double-link properties, every combination of two relations with the same head is used as property. For example, (2) indicates to which degree the language uses the word order head–subject–object.

$$v^{\ell}_{\text{HEAD-nsubj-obj}} = \frac{\#\left[\;\overset{\text{nsubj}}{\curvearrowright}\;\overset{\text{obj}}{\curvearrowright}\;\right](\ell)}{\underbrace{\#\left[\;\overset{\text{nsubj}}{\curvearrowright}\;\overset{\text{obj}}{\curvearrowright}\;\right](\ell) + \ldots + \#\left[\;\overset{\text{obj}}{\curvearrowleft}\;\overset{\text{nsubj}}{\curvearrowleft}\;\right](\ell)}_{\text{six possibilities}}} \tag{2}$$

Chain-link properties are similar to double-link properties. Here, every combination of two chained relations is used as property. All single-, double- and chain-link properties are extracted in two versions, once without POS information (as in the examples above) and once with POS information. For example, (3) is a chain-link property with POS information and indicates to which degree prepositional phrases follow a governing noun in the language.

$$v^{\ell}_{\mathrm{HEAD}_1\text{:NOUN-case:ADP-nmod:}\mathrm{HEAD}_2\text{:NOUN}} = \frac{\#\left[\overset{\text{nmod}}{\overset{\text{case}}{\text{NOUN}\quad\text{ADP}\quad\text{NOUN}}}\right](\ell)}{\underbrace{\#\left[\overset{\text{nmod}}{\overset{\text{case}}{\text{NOUN}\quad\text{ADP}\quad\text{NOUN}}}\right](\ell) + \ldots + \#\left[\overset{\text{nmod}}{\overset{\text{case}}{\text{NOUN}\quad\text{ADP}\quad\text{NOUN}}}\right](\ell)}_{\text{six possibilities}}} \tag{3}$$

Treebanks for the same language are merged, except for the analysis in section 3.5. Subtypes of relations are mapped to their base type (e.g. “compound:prt” $\mapsto$ “compound”).

3.3 Vectors from the URIEL database

The URIEL typological database[6] (Littell et al., 2017) provides binary typological, geographical and phylogenetic information about languages. Corresponding language vectors can be created with the Python library `lang2vec`.[7] Note that binary vectors fulfil the requirements in section 3.1.

[6] http://www.cs.cmu.edu/~dmortens/uriel.html

[7] https://github.com/antonisa/lang2vec

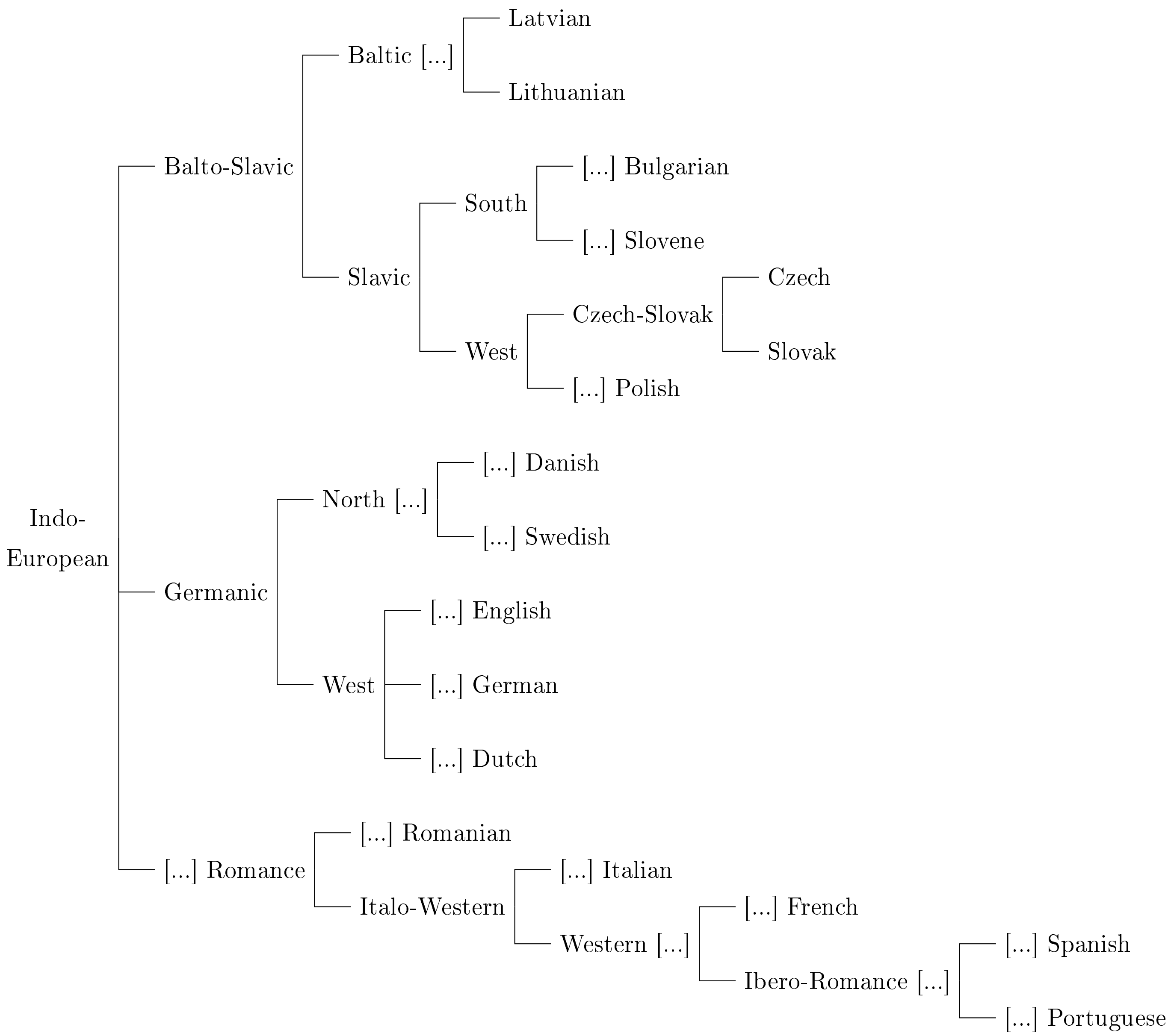


Figure 2: Example tree of language families from Ethnologue (pruned).[8]

[8] www.ethnologue.com/subgroups/indo-european

3.4 Hierarchical clustering

Given a set of language vectors, a distance measure and a linkage method[9], languages can be clustered hierarchically. The output is a weighted binary tree with languages as leaf nodes. This method is scientifically valuable for it gives an impression of how well the vectors represent the languages. However, sensible evaluation is difficult, because a "perfect" gold tree is lacking.

3.4.1 Gold trees

A textbook tree of languages and language families might look like the tree in Figure 2. The missing edge weights can be assumed to be 1. Such trees, however, are non-binary which makes it impossible to use them as gold tree.

Rabinovich et al. (2017) and Bjerva et al. (2019c) use the tree of Serva and Petroni (2008) as gold tree after pruning it to contain 17 languages. Serva and Petroni (2008) use word lists as word vectors and a temporal distance as distance measure which they define as

$$TD(\vec{x}, \vec{y}) = -1750 \cdot \ln(1 - 1.09 \cdot GD(\vec{x}, \vec{y})) \tag{4}$$

where

$$GD(\vec{x}, \vec{y}) = \frac{1}{|\vec{x}|} \cdot \sum_{i=1}^{|\vec{x}|} \frac{LD(x_i, y_i)}{\max\{|x_i|, |y_i|\}} \tag{5}$$

is the genetic distance and LD is the Levenshtein distance. The genetic distance is the average normalised Levenshtein distance. As linkage method, Serva and Petroni (2008) use the unweighted pair group method with arithmetic mean (UPGMA) algorithm by Sokal and Michener (1958) to calculate the distance between two clusters X and Y.

$$d_{average}(X, Y) = \frac{1}{|X| \cdot |Y|} \cdot \sum_{\vec{x} \in X} \sum_{\vec{y} \in Y} D(\vec{x}, \vec{y}) \tag{6}$$

[9] A distance measure calculates the distance of two vectors; a linkage method calculates the distance of two clusters.

D is a vector distance measure—the temporal distance in Serva and Petroni (2008). Figure 3 shows the recreated tree using their method and the original data for the 17 languages used in Rabinovich et al. (2017). The latter list three major advantages for using this tree as gold tree:

1) It is more or less similar to well accepted trees (cf. Gray and Atkinson, 2003).
2) It is a binary tree.
3) Its branches are weighted with the approximate number of years after the splitting occurred which allows a more precise evaluation.

However, there are also some disadvantages:

1) The genetic distance is based on orthographic similarity rather than phonemic, morphological or syntactic similarity.
2) Serva and Petroni's (2008) data is only accessible through an internet archive service[10].
3) It does not contain the same languages as in the UD treebanks which makes it impossible to compare clusterings for all of those languages.

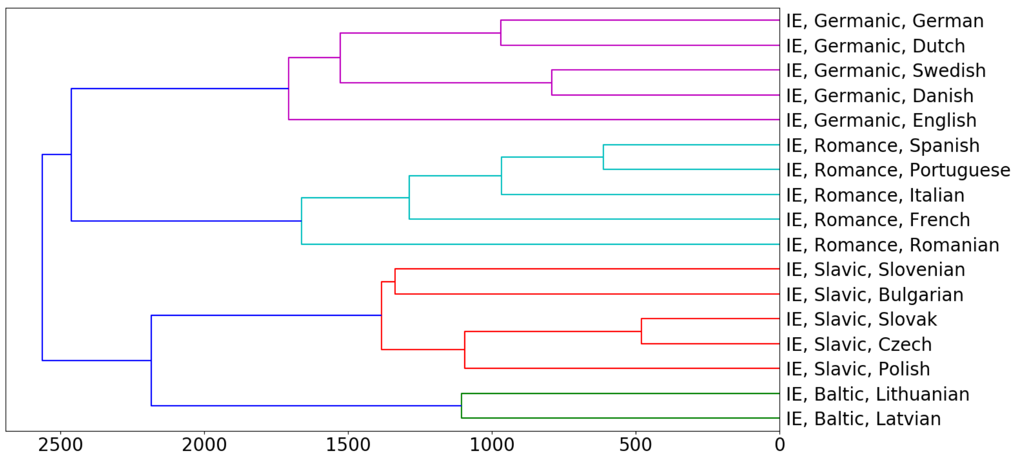


Figure 3: Gold tree following Serva and Petroni (2008).

[10]https://web.archive.org/web/20091203105416/http://univaq.it/~serva/languages/languages.html

3.4.2 Euclidean distance and tree distance

All numeric vectors can be clustered using the Ward variance minimisation algorithm (Ward, 1963) as linkage method and the Euclidean distance ED as distance measure.

$$ED(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{|\vec{x}|}(x_i - y_i)^2} \tag{7}$$

However, the vectors obtained from UD and URIEL are not necessarily complete, i.e. they might contain missing values. Therefore, the Euclidean distance has to be adapted for vectors with possibly missing values.

$$ED(\vec{x},\vec{y}) = \sqrt{\frac{|\vec{x}|}{|\{1 \leq i \leq |\vec{x}| : x_i \neq \text{NULL} \land y_i \neq \text{NULL}\}|}} \cdot \sqrt{\sum_{\substack{i=1 \\ x_i \neq \text{NULL} \\ y_i \neq \text{NULL}}}^{|\vec{x}|} (x_i - y_i)^2} \tag{8}$$

Only properties with values in both vectors are taken into account and the result is normalised. This solution is suitable if the vast majority of properties have values in both vectors—(almost) disjoint vectors produce pointless results.[11]

Following Rabinovich et al. (2017), the Euclidean distance is also used to calculate the distance between two trees. To do so, the trees are vectorised as follows. Given a tree τ, its leaf-distance vector $\vec{l}_\tau$ is defined as

$$\begin{array}{rccccl} \vec{l}_\tau = (& 0, & l_\tau(2,1), & \dots, & l_\tau(|L|,1), & \\ & l_\tau(1,2), & 0, & \dots, & l_\tau(|L|,2), & \\ & \vdots & \vdots & \ddots & \vdots & \\ & l_\tau(1,|L|), & l_\tau(2,|L|), & \dots, & 0 &) \end{array} \tag{9}$$

where $l_\tau(i,j)$ denotes the sum of the weights of all edges on the shortest path between the leaves (i.e. languages) ℓ_i and ℓ_j in τ (when calculating unweighted distances, all edges have the weight 1). The tree distance or

[11] Replacing missing values by 0.5 led to generally equal or worse results compared to the solution in (8).

squared path-difference distance PD (Mir and Rosselló, 2010; Bogdanowicz et al., 2012) between two trees is the squared Euclidean distance of the leaf-distance vectors.

$$PD(\tau, g) = ED(\vec{l}_\tau, \vec{l}_g)^2 \tag{10}$$

This measure has two issues. First, when $\vec{l}_\tau$ contains greater numbers compared to $\vec{l}_g$ or vice versa, then the tree distance becomes accordingly large, independently of whether the languages are similarly clustered. To counteract this, the leaf-distance vectors are standardised as in (12, 13) and the tree distance becomes

$$PD(\tau, g) = \begin{cases} ED(\vec{l}_\tau, \vec{l}_g)^2 & \text{unweighted} \\ ED(z(\vec{l}_\tau), z(\vec{l}_g))^2 & \text{weighted} \end{cases} \tag{11}$$

with

$$z(x_i) = \frac{x_i - \mu(\vec{x})}{\sigma(\vec{x})} \tag{12}$$

$$z(\vec{x}) = (z(x_1), \ldots, z(x_{|\vec{x}|})) \tag{13}$$

where μ is the mean and σ is the standard deviation.

The second issue is that the tree distance is not normalised according to the number of leaves. As a consequence of that, tree distances are not comparable if the number of languages is not fixed. Following Bogdanowicz et al. (2012), the normalised tree distance is obtained by dividing through the average distance between two random trees with $|L|$ leaves:

$$PD^*(\tau, g) = \frac{PD(\tau, g)}{\overline{PD}_{|L|,weighting}} \tag{14}$$

For 10,000 runs with randomly generated trees, $\overline{PD}_{17,unweighted} \approx 1473.2$ and $\overline{PD}_{17,weighted} \approx 490.4$.

3.4.3 Evaluation

Table 5 shows the clustering performance of different vector types from the UD treebanks, URIEL and previous work. The corresponding dendrograms can be found in appendix A.3.1.

				Weighted	
Language vectors	Source	$\|P\|$	Null	No	Yes
phylogenetic features	URIEL	3,718	0.00	0.23	0.15
chain links + POS	UD	682,830	0.91	0.25	0.12
double links	UD	3,810	0.34	0.26	0.15
chain links	UD	7,284	0.53	0.29	0.14
WALS features	URIEL	103	0.30	0.29	0.23
double links + POS	UD	599,196	0.90	0.37	0.20
single links + POS	UD	5,890	0.79	0.41	0.23
single links	UD	37	0.09	0.51	0.35
geographic features	URIEL	299	0.00	0.64	0.59
single links + POS	Bjerva et al. (2019c)			0.32	–
POS	Rabinovich et al. (2017)			0.35	0.30
POS + function words	Rabinovich et al. (2017)			0.36	0.28
phrase structure	Bjerva et al. (2019c)			0.36	–
POS	Bjerva et al. (2019c)			0.52	–
raw text	Bjerva et al. (2019c)			0.53	–
POS + function words	Bjerva et al. (2019c)			0.56	–
random		–	–	0.86	0.85

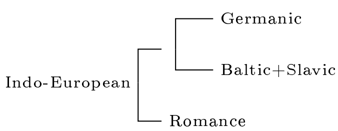
Table 5: Normalised tree distances between different language vectors and the 17-languages gold tree following Serva and Petroni (2008), separated into current and previous work and sorted by unweighted tree distance; $|P|$ ... number of properties (i.e. dimensionality of the vectors), Null ... percentage of missing values in the 17 language vectors.

The numbers of Bjerva et al. (2019c) have to be treated with caution since they do not specify how they normalised their data or whether they calculated weighted or unweighted tree distances. Since they compare their scores to the unweighted tree distances of Rabinovich et al. (2017), it is carefully assumed that the numbers of Bjerva et al. (2019c) and Rabinovich et al. (2017) really are comparable.

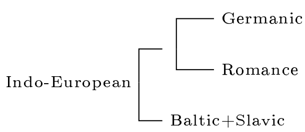
Rabinovich et al.'s (2017) number for the "POS" condition (0.35) was recalculated as 0.37. Since the difference is rather small, the results by Rabinovich et al. (2017) can be considered comparable to the results of this work, although they do not specify how they normalised their tree distances.[12] Rabinovich et al. (2017) also report different average distances between a random tree and the gold tree (last line of Table 5). For unweighted and weighted tree distances, they report 0.72 and 0.68, respectively. The differences might be caused by differences in the algorithm for generating random trees (although both works sample random leaf-distance vectors from the uniform distribution).

Because of the many obscurities in Rabinovich et al.'s (2017) and Bjerva et al.'s (2019c) notions of the tree distance, a comparison between this work and previous work is omitted and the upper and the lower part of Table 5 are discussed separately in the following. It is further noteworthy that the UD treebanks comprise several fictional and non-fictional text genres whereas Rabinovich et al. (2017) and Bjerva et al. (2019c) use the Europarl parallel corpus, which only contains political debates.

It is no surprise that the phylogenetic vectors produce the best phylogenetic tree. The crucial difference is the clustering of the main sub-branches



, which is



in the gold tree (in the Ethnologue tree on page 13, these branches are same-level siblings).

[12]Rabinovich et al. (2017), Bjerva et al. (2019c), and Oncevay et al. (2019) all use normalised tree distances as evaluation measure but neither of them specifies the normalisation constants they used.

The WALS vectors put Slovenian, Bulgarian and Polish to the Romance languages and the chain-link and double-link vectors separate German and Dutch from the other Germanic languages. Comparing the different UD vectors, chain links and double links perform better than single links; presumably because they describe the syntactic nature of a language more precisely. Those vectors also perform better than WALS features (double links with POS only in the weighted evaluation), which illustrates the accuracy of finer-grained, continuous values over coarser-grained, binary values.

The performance is generally higher for the UD vectors with POS information, except for double links with POS, presumably because the large number of constructions which are very rare or even missing in a lot of the languages make a comparison of two language vectors unstable. If this assumption is true, then (the 17 Indo-European) languages have more chain constructions in common than constructions involving one head with two dependents. Geographical features, as one might expect, lead to the most distant clustering compared to the gold standard.

Rabinovich et al. (2017) and Bjerva et al. (2019c) both work with English texts that have been translated from other source languages. Rabinovich et al. (2017) use counts of POS trigrams and function words as language vectors. Bjerva et al. (2019c) train a multilingual neural language model by using the text representation and a language representation as input. The language representations are updated during training and result in the final language vectors. The text is represented on different levels of (syntactic) abstraction: the raw text, POS tags and function words, only POS tags, a phrase structure representation as well as dependency links and POS tags. The latter is similar to “single links + POS” of this work. The general tendency is that more abstract vectors yield more accurate clustering results.

For more than 17 languages, only a qualitative analysis is possible since a gold tree is lacking. Constructed trees for 66 languages can be found in appendix A.3.2. These are exactly those languages in both URIEL and the UD treebanks for which at least two other languages of the same family are included in the UD treebanks. Phylogenetic features, single links and double links recognise most of the families and main sub-branches correctly.[13] The WALS vectors perform worse, probably due to a lot of missing values.

3.5 Inconsistencies among treebanks

The hierarchical clustering of the UD vectors showed that the extracted properties and the underlying data are expressive enough to build approximate phylogenetic trees. However, the construction of a language vector from different treebanks for the same language in a sort makes the assumption that all involved treebanks are annotated consistently. For that reason, the extracted properties are now used to detect inconsistencies in the UD treebanks. This is primarily important because annotation inconsistencies might influence subsequent processing and analysis steps.

3.5.1 Dependency directions

Chen and Gerdes (2017) introduce language vectors whose entries are directional dependency distances. The directional dependency distance DDD for a dependency relation R and a language ℓ is the average dependency distance of R's instances in the treebanks of ℓ, where left-directed dependencies have a negative distance and right-directed dependencies have a positive distance. Figure 4 shows an example for dependency distances; (15) shows the formula to compute the directional dependency distance.

[13]Only Afro-Asiatic languages are scattered through the tree for the UD vectors. A look at the dependency links, e.g. in appendix A.2, confirms that these languages have some statistical differences. One possible explanation (apart from an inconsistent annotation) is that Afro-Asiatic is a comparatively old language family. Almost all Afro-Asiatic languages in the UD treebanks belong to different subgroups of Semitic (ca. 3,750—2,500 BCE; cf. Kitchen et al., 2009; Huehnergard and Pat-El, 2019) which itself is as old as, e.g., Indo-European (ca. 4,500—2,500 BCE; cf. Bomhard, 2019).

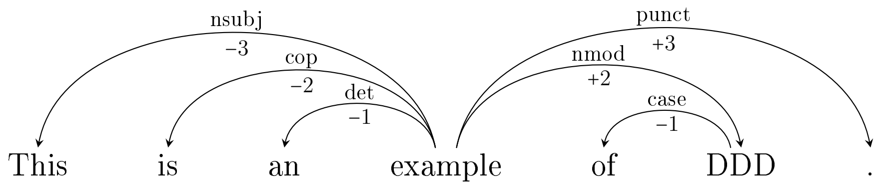


Figure 4: Positive and negative dependency distances.

$$DDD(R,\ell) = \frac{\sum\limits_{r \in \text{instances}(R,\ell)} \text{distance}(r)}{|\text{instances}(R,\ell)|} \tag{15}$$

The authors use the language vectors for clustering but briefly mention that treebank vectors could be used to discover inconsistencies among treebanks for the same language. In the following, this idea shall be pursued.

The DDD is similar to the single-link properties from section 3.2. To be more precise, $\frac{DDD(R,\ell)+1}{2} = v^{\ell}_{\text{HEAD-}R}$ if every dependency distance was set to -1 or $+1$, respectively. This means that the single-link properties only quantify the average direction of a relation whereas the DDD quantifies the average direction and distance.

Table 6 shows languages and relations which have high differences in the corresponding single-link property, i.e. dependency direction, between treebanks. The dependency relations “dep”, “discourse”, “dislocated”, “fixed”, “goeswith”, “list”, “orphan”, “parataxis”, “reparandum” and “vocative” are ignored because their occurrences are usually very infrequent and dissimilar. Which of these differences result from inconsistent annotation is evaluated individually and discussed below. The cases follow in the same order as in Table 6.

ℓ	R	max – min	All treebank values	Incon.
Chinese	clf	1.00	$\overset{\text{GSD}}{0.00}, \overset{\text{GSDSimp}}{0.00}, \overset{\text{PUD}}{0.02}, \overset{\text{CFL}}{0.99}, \overset{\text{HK}}{1.00}$	T
Korean	aux	1.00	$\overset{\text{PUD}}{0.00}, \overset{\text{GSD}}{0.83}, \overset{\text{Kaist}}{1.00}$	T
Korean	mark	1.00	$\overset{\text{Kaist}}{0.00}, \overset{\text{GSD}}{1.00}$	?
Arabic	compound	0.98	$\overset{\text{PUD}}{0.02}, \overset{\text{NYUAD}}{1.00}$	T
Galician	cc	0.94	$\overset{\text{TreeGal}}{0.00}, \overset{\text{CTG}}{0.94}$	T
Turkish	compound	0.94	$\overset{\text{PUD}}{0.01}, \overset{\text{GB}}{0.05}, \overset{\text{IMST}}{0.95}$	?
English	csubj	0.90	$\overset{\text{GUM}}{0.10}, \overset{\text{LinES}}{0.56}, \overset{\text{EWT}}{0.62}, \overset{\text{ESL}}{0.73}, \overset{\text{ParTUT}}{0.73}, \overset{\text{PUD}}{0.73}, \overset{\text{Pronouns}}{1.00}$	F
Old Russian	expl	0.88	$\overset{\text{RNC}}{0.06}, \overset{\text{TOROT}}{0.93}$	F
Spanish	compound	0.88	$\overset{\text{PUD}}{0.12}, \overset{\text{GSD}}{0.36}, \overset{\text{AnCora}}{1.00}$	T
Chinese	compound	0.80	$\overset{\text{GSD}}{0.00}, \overset{\text{GSDSimp}}{0.00}, \overset{\text{PUD}}{0.01}, \overset{\text{HK}}{0.43}, \overset{\text{CFL}}{0.80}$	T
French	csubj	0.80	$\overset{\text{FTB}}{0.20}, \overset{\text{Spoken}}{0.20}, \overset{\text{PUD}}{0.75}, \overset{\text{GSD}}{0.81}, \overset{\text{ParTUT}}{0.97}, \overset{\text{Sequoia}}{1.0}$	F
Galician	nummod	0.80	$\overset{\text{TreeGal}}{0.18}, \overset{\text{CTG}}{0.98}$	T
Arabic	nummod	0.75	$\overset{\text{PUD}}{0.24}, \overset{\text{NYUAD}}{0.47}, \overset{\text{PADT}}{0.99}$	?
French	cop	0.73	$\overset{\text{GSD}}{0.02}, \overset{\text{FTB}}{0.03}, \overset{\text{PUD}}{0.04}, \overset{\text{Sequoia}}{0.05}, \overset{\text{Spoken}}{0.05}, \overset{\text{ParTUT}}{0.08}, \overset{\text{FQB}}{0.75}$	F
Latin	cc	0.72	$\overset{\text{ITTB}}{0.04}, \overset{\text{PROIEL}}{0.71}, \overset{\text{Perseus}}{0.76}$	T
Komi Zyrian	cop	0.69	$\overset{\text{Lattice}}{0.12}, \overset{\text{IKDP}}{0.81}$	?
French	compound	0.67	$\overset{\text{PUD}}{0.33}, \overset{\text{Spoken}}{0.78}, \overset{\text{ParTUT}}{1.00}$	(F)
Portuguese	compound	0.63	$\overset{\text{PUD}}{0.37}, \overset{\text{Bosque}}{1.00}$	T
German	compound	0.62	$\overset{\text{PUD}}{0.34}, \overset{\text{GSD}}{0.53}, \overset{\text{HDT}}{0.94}, \overset{\text{LIT}}{0.96}$	F
English	compound	0.59	$\overset{\text{EWT}}{0.09}, \overset{\text{PUD}}{0.09}, \overset{\text{ESL}}{0.13}, \overset{\text{GUM}}{0.13}, \overset{\text{LinES}}{0.18}, \overset{\text{ParTUT}}{0.67}$	F

Table 6: Top 20 relations with largest differences in dependency direction within a language. The abbreviations written above numbers are the treebank shortcuts. The last column shows whether the difference is a true (T) or false (F) inconsistency or whether the case in unclear (?).

Inconsistent Chinese classifiers (clf) Figure 5 shows that the annotation for classifiers is categorically different among the Chinese treebanks.

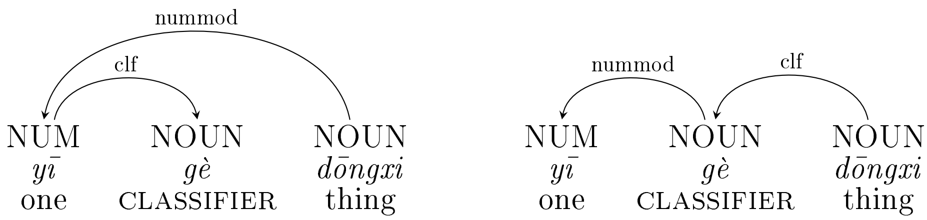


Figure 5: Chinese annotation scheme for noun phrases with a classifier in CFL/HK (left) and GSD/GSDSimp/PUD (right).

Inconsistent Korean auxiliaries (aux) The great inconsistency of the AUX tag and the "aux" relation among the Korean treebanks is already documented by Noh et al. (2018):

> "Of the 37 universal syntactic relations labels [...], [GSD], PUD, and [Kaist] show the biggest difference in the `aux` (Auxiliary) and labels related to MEW [*sic*] (multi-word-expression), `compound` (Compound), `fixed` (Fixed Multiword Expression), and `flat` (Flat Multiword Expression). [GSD] demonstrates quite a low frequency of `aux`, which is because auxiliary predicates are not classified separately in the POS annotation process but are processed with VERB. Since an auxiliary predicate is not used alone but appears next to the main predicate, AUX [...] and `aux`, [...], should be proportional to each other. As such, it is natural for `aux` to show high frequency, as is the case with [Kaist]. Instead, a `flat` label appears in [GSD] with high frequency whereas it appears with low frequency in other corpus [*sic*] as the relationship between the auxiliary predicate and the main predicate is processed as `flat`."

Inconsistent Arabic compounds (compound) In the NYUAD treebank, the “compound” relation is only used to connect numerals (NUM tags) from left to right (in the PUD treebank and the PADT, words tagged as NUM are commonly connected by “nummod”). Whilst the “compound” relation is not used in the PADT at all, there are two usages in the PUD treebank:

1) “compound” to connect nouns, also from left to right, but rarely used.
2) “compound:prt” to connect words, mainly verbs (VERB), with their particles (PART). Here, the particle almost always precedes its governor.

The labelling of particles is, obviously, handled differently in the PADT and the NYUAD treebank.

Inconsistent Galician coordinating conjunctions (cc) Whilst the TreeGal treebank combines the “cc” and the “conj” relation, the CTG treebank does not use “conj” and uses “cc” as shown in Figure 6. Conjuncts which are not nouns are connected similarly.

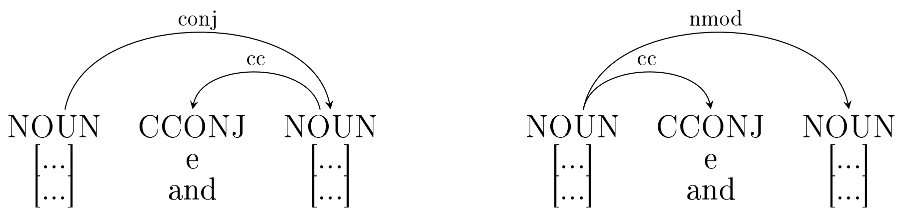


Figure 6: Galician annotation scheme for coordinating conjunctions in TreeGal (left) and CTG (right).

Consistent English clausal subjects (csubj) The presence of an expletive pronoun changes the direction of the “csubj” relation, see Figure 7. The GUM treebank simply contains more sentences with expletives.

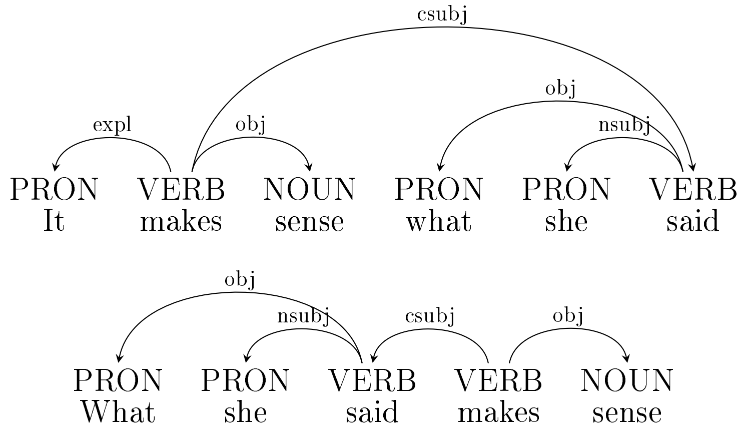


Figure 7: English annotation scheme for clausal subjects with expletive (above) and without expletive (below).

Consistent Old Russian expletives (expl) The only expletives in the Old Russian treebanks are forms of *sebe/sja* 'oneself'. The TOROT contains over 1,000 examples and in 93% of the cases the expletive follows the verb; the RNC treebank addresses those forms with the subtype relation "expl:rv" and contains only nine examples, in all but one of them the expletive precedes the verb. Furthermore, in the RNC treebank the base "expl" relation is used to connect particles (PART) with their governors and the particle usually precedes its governor.

Inconsistent Spanish compounds (compound) First of all, the PUD treebank only contains "compound:prt" whereas the GSD treebank and the AnCora treebank only contain "compound". The usage varies in several points, for example with temporal words and reflexive words:

Figure 8: Spanish annotation scheme for temporal words and *pasado* 'last' in AnCora (left) and GSD/PUD (right).

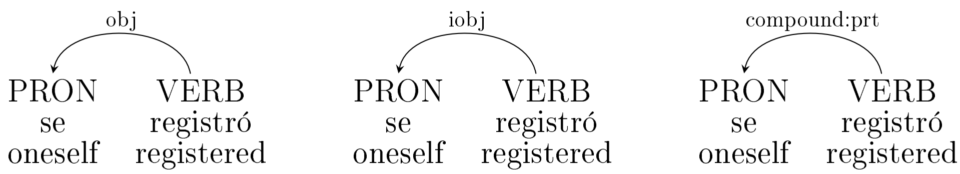


Figure 9: Spanish annotation scheme for reflexive verbs in AnCora (left), GSD (middle) and PUD (right).

Inconsistent Chinese compounds (compound) All Chinese treebanks use the base "compound" relation and the direction is generally from right to left. However, the HK treebank and the CFL treebank also use "compound:dir", "compound:ext", "compound:vo" and "compound:vv" where the direction is generally from left to right. The extended use of "compound" (by subtypes) in some but not all Chinese treebanks is an inconsistency.

Consistent French clausal subjects (csubj) In French, the clausal subjects show the same expletive alternation as in English (see above).

Inconsistent Galician numeral modifiers (nummod) The CTG treebank attaches numerals to the determiner instead of the noun:

Figure 10: Galician annotation scheme for coordinating conjunctions in TreeGal (left) and CTG (right).

Consistent French copulae (cop) The FQB contains a lot of questions where fronted forms of *quel* 'which' are the governor of the copula, i.e. the fronted forms change the average direction of the "cop" relation.

Inconsistent Latin coordinating conjunctions (cc) The conjunction is categorically attached to different conjuncts in the Latin treebanks, analogous to Galician (see above).

(In)Consistent French compounds (compound) The "compound" relation is consistently used for common-noun compounds in the PUD treebank, the Spoken treebank and the ParTUT. An inconsistency, however, which cannot be detected by dependency direction difference, is that the other French treebanks mark compounds by other relations, e.g. "nmod".

Inconsistent Portuguese compounds (compound) The Portuguese treebanks show a different use of "compound" and "compound:prt", similar to the Spanish compounds.

Consistent German compounds (compound) The German treebanks differentiate between "compound" and "compound:prt". The former is used for (hyphenated) nominal compounds and usually spans from right to left; the latter is used for verbal particles and mainly goes the opposite direction, from left to right.

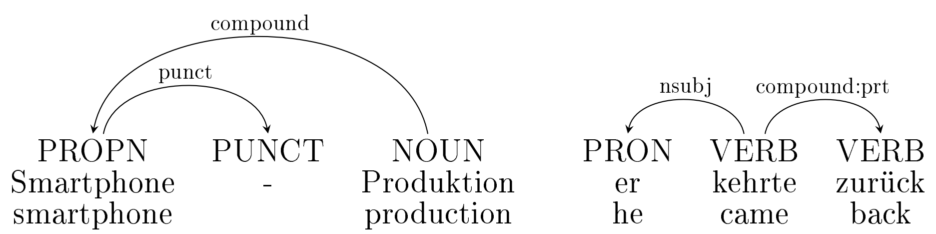


Figure 11: German annotation scheme for nominal compounds (left) and particle verbs (right).

Consistent English compounds (compound) Proper-name compounds are annotated differently from common-noun compounds in English, see Figure 12, and the ParTUT treebank almost exclusively contains proper-name compounds which is in contrast to the other English treebanks.

Figure 12: English annotation scheme for compounds of common nouns (left) and proper names (right).

3.5.2 Treebank clustering

Looking at treebank vectors raises the question of whether treebanks for the same language form a pure "language" cluster, i.e. whether treebanks for the same language are more similiar to each other than to any treebank from another language. Given a gold and a predicted clustering characterised by the functions $gold : L \mapsto \mathfrak{P}(L)$ and $pred : L \mapsto \mathfrak{P}(L)$, where each function maps a treebank $\ell \in L$ to the cluster containing ℓ, the purity is calculated as

$$\mathrm{purity}(L, gold, pred) = \frac{1}{|L|} \cdot \sum_{Y \in \{pred(\ell) : \ell \in L\}} \max_{X \in \{gold(\ell) : \ell \in L\}} |X \cap Y| \qquad (16)$$

and ranges between 0 (bad clustering) and 1 (perfect clustering) (cf. Manning et al., 2008, p. 357).

Table 7 shows the purities for different language vectors using all properties or without properties involving the relations "dep", "discourse", "dislocated", "fixed", "goeswith", "list", "orphan", "parataxis", "reparandum" and "vocative" (the same which are excluded in the previous subsection, p. 22). The purity ranges between 70% and 78% and the overall tendency is that language vectors with fewer dimensions perform better. Figure 13 shows a hierarchical treebank clustering.

	Purity	
Language vectors	37 relations	27 relations
single links	0.72	0.78
single links + POS	0.74	0.76
double links	0.70	0.74
double links + POS	0.74	0.72
chain links	0.70	0.72
chain links + POS	0.72	0.70

Table 7: Purity of the treebank clusterings for the 17 languages in Figure 13 after cutting the constructed tree at 17 clusters.

An automated analysis to find inconsistencies among treebanks of closely related languages is difficult. Too uncertain are the influences of language-internal inconsistencies (due to underspecification in the UD annotation guidelines, conversion errors etc.) and genre differences (fictional vs. non-fictional, spoken vs. written etc.). Again, individual inspections of impure clusters might give an idea of the annotation differences. In Figure 13, for example, Slovak-SNK is in the same cluster as all Czech treebanks except Czech-PUD. Looking into the documentation reveals that Czech-PUD is the only Czech treebank which was natively annotated in the UD format (Zeman et al., 2017); the other Czech treebanks (cf. Zeman, 2015; Jelínek, 2017; Kríž et al., 2016) were built at the Charles University in Prague and later converted into the UD format. Slovak-SNK, on the other side, was annotated following the annotation style of the Prague Dependency Treebank before being converted to UD format. This annotation style seems to make Slovak-SNK and the Czech treebanks more similar than Czech-PUD and the other Czech treebanks.

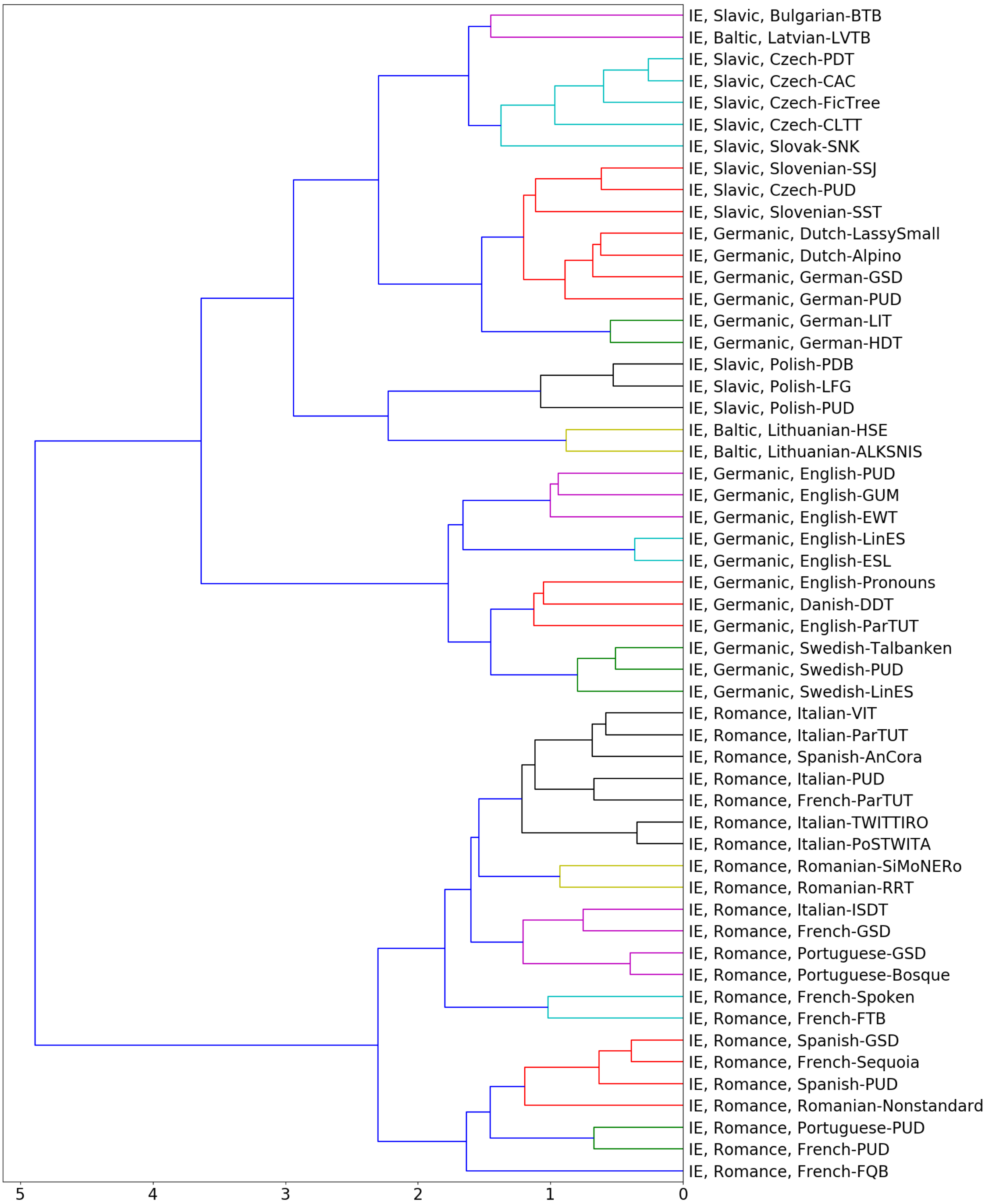


Figure 13: Clustering of 54 treebanks for 17 languages using single-link vectors. The colours mark the clusters after cutting the dendrogram at 17 clusters.

4 A Framework for Property Vectors

This section provides the theoretical basis for the evaluation of formulae based on continuous typological language vectors. After introducing real-valued logics, two concrete use cases follow, in which the framework is used to assess the agreement between (manually collected) typological databases and the statistical data extracted from the Universal Dependencies treebanks.

4.1 Valuation functions

The goal of this work is to evaluate formulae, which are possibly logically complex, on a data basis. Hereby, "logically complex" refers to any formula which connects more than two variables. An example is given in (17), which is the 70^{th} universal in the Universals Archive[14].

$$\phi_{70} \coloneqq (\neg\vec{v}_{\text{HEAD-case}} \wedge \neg\vec{v}_{\text{nsubj-HEAD-obj}}) \Rightarrow (\vec{v}_{\text{HEAD-amod}} \Rightarrow \vec{v}_{\text{HEAD-nmod}}) \tag{17}$$

It is important to differentiate between the syntactic form of a formula, as in (17), and the value of a formula evaluated under certain conditions. The value of a formula $\phi \in \Phi$, where Φ denotes the set of well-formed formulae, can be calculated only with respect to a language ℓ, a universe U and a valuation function $\mathcal{V}$. The universe is the set of possible values which the variables and their connections can represent, i.e. $V \in U^{|L|\times|P|}$; the definition of the connectives, in turn, depends on the valuation $\mathcal{V}$ which assigns a truth value to a formula for a language, i.e. $\mathcal{V} : \Phi^L \to U$:

$$\mathcal{V}(\phi_{70}^{\ell}) \coloneqq \mathcal{V}((\neg v^{\ell}_{\text{HEAD-case}} \wedge \neg v^{\ell}_{\text{nsubj-HEAD-obj}}) \Rightarrow (v^{\ell}_{\text{HEAD-amod}} \Rightarrow v^{\ell}_{\text{HEAD-nmod}})) \tag{18}$$

In Boolean logic, $U = \{0, 1\}$ and the logical connectives $\neg$ (negation), $\wedge$ (conjunction), $\vee$ (disjunction), $\Rightarrow$ (implication) and $\Leftrightarrow$ (equivalence) are defined by $\mathcal{V}_B$ as

[14] http://typo.uni-konstanz.de/archive/

$$\mathcal{V}_B(v_p^\ell) := v_p^\ell \tag{19}$$

$$\mathcal{V}_B(\neg\phi^\ell) := \begin{cases} 1 & \mathcal{V}_B(\phi^\ell) = 0 \\ 0 & \mathcal{V}_B(\phi^\ell) = 1 \end{cases} \tag{20}$$

$$\mathcal{V}_B(\phi_1^\ell \wedge \ldots \wedge \phi_n^\ell) := \begin{cases} 1 & \mathcal{V}_B(\phi_1^\ell) = \ldots = \mathcal{V}_B(\phi_n^\ell) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

$$\mathcal{V}_B(\phi_1^\ell \vee \ldots \vee \phi_n^\ell) := \begin{cases} 0 & \mathcal{V}_B(\phi_1^\ell) = \ldots = \mathcal{V}_B(\phi_n^\ell) = 0 \\ 1 & \text{otherwise} \end{cases} \tag{22}$$

$$\mathcal{V}_B(\phi_1^\ell \Rightarrow \phi_2^\ell) := \mathcal{V}_B(\neg\phi_1^\ell \vee \phi_2^\ell) \tag{23}$$

$$\mathcal{V}_B(\phi_1^\ell \Leftrightarrow \phi_2^\ell) := \mathcal{V}_B((\phi_1^\ell \Rightarrow \phi_2^\ell) \wedge (\phi_2^\ell \Rightarrow \phi_1^\ell)) \tag{24}$$

for any variable $\vec{v}_p \in U^{|L|}$ and formulae $\phi, \phi_1, \ldots, \phi_n$.

Setting $v^\ell_{\text{HEAD-case}} = 1$, $v^\ell_{\text{HEAD-amod}} = 0$, $v^\ell_{\text{HEAD-nmod}} = 1$ and $v^\ell_{\text{nsubj-HEAD-obj}} = 1$ yields $\mathcal{V}_B(\phi^\ell_{70}) = 1$. However, these definitions are not sufficient since they are only applicable for Boolean values, whereas the property values as described in section 3.1 are real values.

4.2 Real-valued logics

In real-valued logics, the logical connectives have to be redefined, so that not only 0 (false) and 1 (true) but also every other value in $[0, 1] = U$ can be a truth value. Product logic and fuzzy logic are two real-valued logics whose commonalities and differences shall be discussed below.

4.2.1 Addition

Addition is not a logical connective but listed here because it is used later on in this work to calculate the value of properties which have not been directly extracted from the UD treebanks. For example, the extent to which a language is "V-final" is the percentage occupied by S-O-V and O-S-V constructions. Thus, the value $v^{\ell}_{\text{V-final}}$ can be calculated by addition:

$$v^{\ell}_{\text{V-final}} = v^{\ell}_{\text{nsubj-obj-}\textsc{head}} + v^{\ell}_{\text{obj-nsubj-}\textsc{head}} \tag{25}$$

Note that $v^{\ell}_{\text{nsubj-obj-}\textsc{head}}$ and $v^{\ell}_{\text{obj-nsubj-}\textsc{head}}$ are calculated with the same denominator (see section 3.2, p. 11) which is required for their addition.

4.2.2 Negation

Negation is a unary operation which is always defined as below.

$$\mathcal{V}(\neg\phi^{\ell}) := 1 - \mathcal{V}(\phi^{\ell}) \tag{26}$$

This general definition has its origin in probability theory: Given a probability function $P : \mathfrak{P}(\Omega) \to [0, 1]$, the probability of any event $A \in \mathfrak{P}(\Omega)$ is $P(A)$ and the probability of its complement $\neg A$ is $1 - P(A)$. It is further a generalisation from Boolean logic.

4.2.3 Conjunction and disjunction

Conjunction and disjunction are n-ary connectives whose definitions vary between different logics.[15] In product logic (e.g. Smith, 2012) (also referred to as probabilistic logic (Mizraji, 1992); the same valuation is used in subjective logic (Jøsang and McAnally, 2005) and others) those connectives are defined as

[15]It is possible to formulate two intuitive constraints which seem to hold across real-valued logics:

$$\mathcal{V}(\phi^{\ell}_1 \wedge \ldots \wedge \phi^{\ell}_n) \overset{!}{\leq} \max\{\mathcal{V}(\phi^{\ell}_1), \ldots, \mathcal{V}(\phi^{\ell}_n)\} \tag{27}$$

$$\mathcal{V}(\phi^{\ell}_1 \vee \ldots \vee \phi^{\ell}_n) \overset{!}{\geq} \min\{\mathcal{V}(\phi^{\ell}_1), \ldots, \mathcal{V}(\phi^{\ell}_n)\} \tag{28}$$

$$\mathcal{V}_P(\phi_1^\ell \wedge \ldots \wedge \phi_n^\ell) \coloneqq \mathcal{V}_P(\phi_1^\ell) \cdot \ldots \cdot \mathcal{V}_P(\phi_n^\ell) \tag{29}$$

$$\mathcal{V}_P(\phi_1^\ell \vee \ldots \vee \phi_n^\ell) \coloneqq \vartheta(\phi_n^\ell) \tag{30}$$

$$\text{with } \vartheta(\phi_i^\ell) = \begin{cases} \vartheta(\phi_{i-1}^\ell) + \mathcal{V}_P(\phi_i^\ell) - \vartheta(\phi_{i-1}^\ell) \cdot \mathcal{V}_P(\phi_i^\ell) & i > 0 \\ 0 & \text{otherwise}^{16} \end{cases}.$$

These definitions also have parallels in probability theory. The probability that two events $A, B \in \mathfrak{P}(\Omega)$, occur together is $P(A) \cdot P(B)$ as long as the events are independent from each other (i.e. $A \cap B = \varnothing$). Similarly, the probability that at least one of the events occurs is $P(A) + P(B) - P(A) \cdot P(B)$. For example, consider any language $\ell \in L$ and let $P(A) = v_{\text{HEAD-case}}^\ell$ the probability that ℓ is postpositional and $P(B) = v_{\text{V-final}}^\ell$ be the probability that ℓ is verb-final. The probability that ℓ is both postpositional and verb-final is $P(A) \cdot P(B)$, only(!) if the postpositionality of ℓ and the verb-finality of ℓ are independent. Under a linguistic point of view, this is certainly not the case (both properties are a subtype of head-finality); however, independence assumptions are known to perform well in computational-linguistic applications.

Another valuation is used in fuzzy logic (Zadeh, 1965), where the connectives are defined by the minimum and maximum of the inputs, respectively:[17]

$$\mathcal{V}_F(\phi_1^\ell \wedge \ldots \wedge \phi_n^\ell) \coloneqq \min\{\mathcal{V}_F(\phi_1^\ell), \ldots, \mathcal{V}_F(\phi_n^\ell)\} \tag{31}$$

$$\mathcal{V}_F(\phi_1^\ell \vee \ldots \vee \phi_n^\ell) \coloneqq \max\{\mathcal{V}_F(\phi_1^\ell), \ldots, \mathcal{V}_F(\phi_n^\ell)\} \tag{32}$$

The underlying theory builds on fuzzy sets. A fuzzy set A is characterised by a membership function $m_A : \Upsilon \to [0, 1]$ (Υ is the element space). For

[16] Since it is not obvious: This definition of $\vee$ is commutative.

[17] These definitions are already used in the earlier Łukasiewicz logic (Łukasiewicz and Tarski, 1930; Łukasiewicz, 1930; see Bergmann, 2008, p. 179) for weak conjunction and weak disjunction.

any element $\upsilon \in \Upsilon$, $m_A(\upsilon)$ is the truth value of υ being in A. Similarly, we define the fuzzy set B and its membership function $m_B : \Upsilon \to [0,1]$. Then the membership function for the intersection is defined as $m_{A\cap B}(\upsilon) := \min\{m_A(\upsilon), m_B(\upsilon)\}$. Now, let $\Upsilon = L$, HEAD-case be the set of languages that are postpositional and V-final be the set of languages that are verb-final. For any language $\ell \in L$, the truth value for ℓ being both postpositional and verb-final is $m_{\text{HEAD-case}\cap\text{V-final}}(\ell) = \min\{m_{\text{HEAD-case}}(\ell), m_{\text{V-final}}(\ell)\}$. In the slightly different notation of this work, the set V-final corresponds to the vector $\vec{\upsilon}_{\text{V-final}}$ and the membership $m_{\text{V-final}}(\ell)$ corresponds to the value $\upsilon^{\ell}_{\text{V-final}}$. Hence the understanding of conjunction and disjunction in fuzzy logic is detached from independence of the languages' properties. Note that the product valuations set lower and upper bounds for the fuzzy valuations, i.e.

$$\mathcal{V}_P(\phi_1^\ell \wedge \ldots \wedge \phi_n^\ell) \le \mathcal{V}_F(\phi_1^\ell \wedge \ldots \wedge \phi_n^\ell) \le \mathcal{V}_F(\phi_1^\ell \vee \ldots \vee \phi_n^\ell) \le \mathcal{V}_P(\phi_1^\ell \vee \ldots \vee \phi_n^\ell) \quad (33)$$

holds for all formulae $\phi_1, \ldots, \phi_n \in \Phi$.

4.2.4 Implication and equivalence

Implication and equivalence are binary connectives which can be defined on basis of the definitions for $\neg$, $\wedge$ and $\vee$ (as in Boolean logic).

$$\mathcal{V}(\phi_1^\ell \Rightarrow \phi_2^\ell) := \mathcal{V}(\neg\phi_1^\ell \vee \phi_2^\ell) \quad (34)$$

$$\mathcal{V}(\phi_1^\ell \Leftrightarrow \phi_2^\ell) := \mathcal{V}((\phi_1^\ell \Rightarrow \phi_2^\ell) \wedge (\phi_2^\ell \Rightarrow \phi_1^\ell)) \quad (35)$$

4.3 Averaging

$|L|$ truth values can be calculated for $\mathcal{V}(\phi^\ell)$, one for every language in L. As mentioned in section 1, a universal is a statement that is supposed to hold for "all" languages, but a universal quantification (i.e. a conjunction of the values) would be misleading since it is not robust to outliers. The score for a formula is hence calculated as the (weighted) average

$$s(\phi) = \frac{\sum\limits_{\ell \in L} w(\phi, \ell) \cdot \mathcal{V}(\phi^\ell)}{\sum\limits_{\ell \in L} w(\phi, \ell)} \tag{36}$$

with w being the weight function. An unweighted average corresponds to a weighted average with all weights being 1. In addition to that, two further weighting methods are tested, which are motivated below.

4.3.1 Antecedent weighting

Hempel (1945a;b) discusses a paradox of confirmation, concerning implicational universals. Consider the hypothesised universal in (39) which states that "if a language is verb-final then it is postpositional". Let $\ell_1, \ell_2, \ell_3, \ell_4$ be four languages so that ℓ_1 is verb-final and postpositional, ℓ_2 is verb-final and prepositional, ℓ_3 is non-verb-final and postpositional, and ℓ_4 is non-verb-final and prepositional. ℓ_1 clearly confirms the hypothesis and ℓ_2 clearly disconfirms it. ℓ_3 and ℓ_4 should have no relevance for a hypothesis concerning verb-final languages since they are not verb-final in the first place.[18] (40) is logically equivalent to (39), hence every language which confirms (40) automatically confirms (39). (40), however, is a hypothesis concerning prepositional languages, making ℓ_4 relevant, which confirms (40) and thus (39). It is counter-intuitive that a non-verb-final language confirms a universal concerning verb-final languages.[19]

$$\phi^\ell_{489} := v^\ell_{\text{V-final}} \Rightarrow v^\ell_{\textsc{head}\text{-case}} \tag{39}$$

$$\phi^{\ell}_{489}{}' := \neg v^\ell_{\textsc{head}\text{-case}} \Rightarrow \neg v^\ell_{\text{V-final}} \tag{40}$$

[18]This selective constraint is called Nicod's criterion.

[19]Hempel's paradox is also known as Raven paradox because of the original example:

$$S_1(x) := \textsc{raven}(x) \Rightarrow \textsc{black}(x) \tag{37}$$

$$S_2(x) := \neg\textsc{black}(x) \Rightarrow \neg\textsc{raven}(x) \tag{38}$$

Any red pencil, green leaf, yellow cow etc. confirms S_2 but could be considered irrelevant for S_1 (cf. Hempel, 1945a, p. 14).

The paradox raises the question of which languages should be considered for evaluating a formula: All or only those which are relevant? Hempel (1945a, p. 11) rightly notices that using a selective constraint "makes confirmation depend not only on the content of the hypothesis, but also on its formulation". And in the case of non-binary logics, it is not even clear what "relevant" means.

Therefore, antecedent weighting makes a generalisation from the binary case. The weight $w(\phi, \ell)$ should be dependent on how relevant ℓ is for ϕ. This can be achieved by setting $w(\phi, \ell) = \mathcal{V}(\phi_1^\ell)$ with $\phi^\ell = (\phi_1^\ell \Rightarrow \phi_2^\ell)$. In the current example:

$$s(\phi_{489}) = \frac{\sum_{\ell \in L} \mathcal{V}\left(v_{\text{V-final}}^\ell\right) \cdot \mathcal{V}\left(\phi_{489}^\ell\right)}{\sum_{\ell \in L} \mathcal{V}\left(v_{\text{V-final}}^\ell\right)} \neq \frac{\sum_{\ell \in L} \mathcal{V}\left(\neg v_{\textsc{head-case}}^\ell\right) \cdot \mathcal{V}\left({\phi_{489}^\ell}'\right)}{\sum_{\ell \in L} \mathcal{V}\left(\neg v_{\textsc{head-case}}^\ell\right)} \tag{41}$$

Since it is possible that several implications occur in ϕ, the definition is extended to $w(\phi, \ell) = w_a(\phi^\ell)$ with

$$w_a(\phi_1^\ell \Rightarrow \phi_2^\ell) = w_a(\phi_1^\ell) \cdot w_a(\phi_2^\ell) \cdot \mathcal{V}(\phi_1^\ell) \tag{42}$$

$$w_a(\phi_1^\ell \Leftrightarrow \phi_2^\ell) = w_a((\phi_1^\ell \Rightarrow \phi_2^\ell) \wedge (\phi_2^\ell \Rightarrow \phi_1^\ell)) \tag{43}$$

$$w_a(\phi_1^\ell \wedge \ldots \wedge \phi_n^\ell) = \begin{cases} \mathcal{V}(w_a(\phi_1^\ell) \wedge \ldots \wedge w_a(\phi_n^\ell)) & \mathcal{V} = \mathcal{V}_P \\ w_a\left(\underset{\phi^\ell \in \{\phi_1^\ell, \ldots, \phi_n^\ell\}}{\arg\min} \mathcal{V}(\phi^\ell)\right) & \mathcal{V} = \mathcal{V}_F \end{cases} \tag{44}$$

$$w_a(\phi_1^\ell \vee \ldots \vee \phi_n^\ell) = \begin{cases} \mathcal{V}(w_a(\phi_1^\ell) \vee \ldots \vee w_a(\phi_n^\ell)) & \mathcal{V} = \mathcal{V}_P \\ w_a\left(\underset{\phi^\ell \in \{\phi_1^\ell, \ldots, \phi_n^\ell\}}{\arg\max} \mathcal{V}(\phi^\ell)\right) & \mathcal{V} = \mathcal{V}_F \end{cases} \tag{45}$$

$$w_a(\neg\phi^\ell) = w_a(\phi^\ell) \tag{46}$$

$$w_a(v_p^\ell) = 1 \tag{47}$$

(48) shows the calculation of the antecedent weight for an example formula. The result is a product of three factors. The first factor is the valuation

of the antecedent of the second (i.e. the inner) implication in ϕ_{70}^{ℓ}, the last two factors are the valuation of the antecedent of the first (i.e. the outer) implication.

$$\begin{aligned}
w_a(\phi_{70}^{\ell}) &= w_a((\neg v_{\text{HEAD-case}}^{\ell} \wedge \neg v_{\text{nsubj-HEAD-obj}}^{\ell}) \Rightarrow (v_{\text{HEAD-amod}}^{\ell} \Rightarrow v_{\text{HEAD-nmod}}^{\ell})) \\
&= w_a(\neg v_{\text{HEAD-case}}^{\ell} \wedge \neg v_{\text{nsubj-HEAD-obj}}^{\ell}) \cdot w_a(v_{\text{HEAD-amod}}^{\ell} \Rightarrow v_{\text{HEAD-nmod}}^{\ell}) \\
&\quad \cdot \mathcal{V}_P(\neg v_{\text{HEAD-case}}^{\ell} \wedge \neg v_{\text{nsubj-HEAD-obj}}^{\ell}) \\
&= \mathcal{V}(w_a(\neg v_{\text{HEAD-case}}^{\ell}) \wedge w_a(\neg v_{\text{nsubj-HEAD-obj}}^{\ell})) \cdot w_a(v_{\text{HEAD-amod}}^{\ell}) \cdot w_a(v_{\text{HEAD-nmod}}^{\ell}) \\
&\quad \cdot \mathcal{V}_P(v_{\text{HEAD-amod}}^{\ell}) \cdot \mathcal{V}_P(\neg v_{\text{HEAD-case}}^{\ell}) \cdot \mathcal{V}_P(\neg v_{\text{nsubj-HEAD-obj}}^{\ell}) \\
&= w_a(\neg v_{\text{HEAD-case}}^{\ell}) \cdot w_a(\neg v_{\text{nsubj-HEAD-obj}}^{\ell}) \cdot 1 \cdot 1 \\
&\quad \cdot v_{\text{HEAD-amod}}^{\ell} \cdot (1 - \mathcal{V}_P(v_{\text{HEAD-case}}^{\ell})) \cdot (1 - \mathcal{V}_P(v_{\text{nsubj-HEAD-obj}}^{\ell})) \\
&= w_a(v_{\text{HEAD-case}}^{\ell}) \cdot w_a(v_{\text{nsubj-HEAD-obj}}^{\ell}) \cdot v_{\text{HEAD-amod}}^{\ell} \cdot (1 - v_{\text{HEAD-case}}^{\ell}) \cdot (1 - v_{\text{nsubj-HEAD-obj}}^{\ell}) \\
&= 1 \cdot 1 \cdot v_{\text{HEAD-amod}}^{\ell} \cdot (1 - v_{\text{HEAD-case}}^{\ell}) \cdot (1 - v_{\text{nsubj-HEAD-obj}}^{\ell}) \\
&= v_{\text{HEAD-amod}}^{\ell} \cdot (1 - v_{\text{HEAD-case}}^{\ell}) \cdot (1 - v_{\text{nsubj-HEAD-obj}}^{\ell})
\end{aligned} \tag{48}$$

4.3.2 Phylogenetic weighting

To confirm a typological universal, it is not sufficient to validate it on as many languages as possible, it is also necessary to validate it on languages from many different language families. If a candidate for a universal is tested on English, German, French, Italian, Spanish and Japanese and the universal holds for all of them but Japanese, then Japanese is not simply an outlier, it is also possible that the "universal" only holds within the Indo-European languages. Typically, "typologists seek to ensure representativity [...] by exhaustive sampling of known families, stratified by area" Bickel (2011). However, giving each language family equal importance can also be achieved by setting $w(\phi, \ell) = w_p(\ell)$ with

$$w_p(\ell) = \begin{cases} \dfrac{w_p(\text{parent}(\ell))}{|\text{siblings}(\ell)| + 1} & \text{if } \ell \text{ has a parent} \\ 1 & \text{otherwise} \end{cases}. \tag{49}$$

The weights for the current example are shown in Figure 14. The weights at each level sum up to 1. This approach does not undersample languages, instead all available data is used. As Bickel (2011) and Dryer (2018) correctly point out, geography is also an important factor, i.e. neighbouring language families are more likely to have common properties than distant language families. However, as Dryer (2018) also admits, measuring geographic distance of languages brings its own complications and therefore it is not taken up in this thesis.

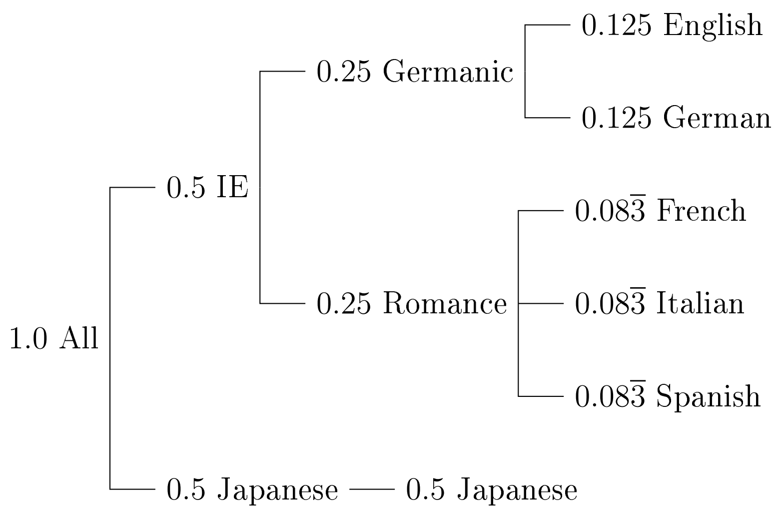


Figure 14: Example language family tree with weights.

4.3.3 Hybrid weighting

Hybrid weighting combines antecedent weighting and phylogenetic weighting by setting $w(\phi, \ell) = w_a(\phi^\ell) \cdot w_p(\ell)$.

4.4 Full example

Now, that the theoretical basis for valuations and averaging are set, the functioning of the complete framework shall be demonstrated by means of an example. Greenberg's fourth universal is widely beyond debate, has been attested in detailed typological studies (eg. Dryer, 1992) and also appears in the top-lists of several computer-driven approaches (Daumé and Campbell, 2007; Bjerva et al., 2019b; Lochbihler, 2017). (6) a. shows the original formulation, (6) b. a formula with typological variables, (6) c. the same formula after replacing the variables with UD variables, and (6) d. the same formula without POS information.

(6) a. "With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional." (Greenberg, 1963, #4)

b. S-O-V ⇒ NP-Adp

c. $\phi_{489} \coloneqq \vec{v}_{\text{nsubj:NOUN-obj:NOUN}\textsc{head}\text{:VERB}} \Rightarrow \vec{v}_{\textsc{head}\text{:NOUN-case:ADP}}$

d. $\phi_{489}{}' \coloneqq \vec{v}_{\text{nsubj-obj-}\textsc{head}} \Rightarrow \vec{v}_{\textsc{head}\text{-case}}$

The formula in d. is used. When evaluating with fuzzy logic, the score for a language ℓ computes as follows.

$$
\begin{aligned}
\mathcal{V}_F(\phi^{\ell}_{489}{}') &= \mathcal{V}_F(v^{\ell}_{\text{nsubj-obj-}\textsc{head}} \Rightarrow v^{\ell}_{\textsc{head}\text{-case}}) \\
&= \mathcal{V}_F(\neg v^{\ell}_{\text{nsubj-obj-}\textsc{head}} \vee v^{\ell}_{\textsc{head}\text{-case}}) \\
&= \max\{\mathcal{V}_F(\neg v^{\ell}_{\text{nsubj-obj-}\textsc{head}}), \mathcal{V}_F(v^{\ell}_{\textsc{head}\text{-case}})\} \\
&= \max\{1 - \mathcal{V}_F(v^{\ell}_{\text{nsubj-obj-}\textsc{head}}), \mathcal{V}_F(v^{\ell}_{\textsc{head}\text{-case}})\} \\
&= \max\{1 - v^{\ell}_{\text{nsubj-obj-}\textsc{head}}, v^{\ell}_{\textsc{head}\text{-case}}\}
\end{aligned}
\tag{50}
$$

The formula should be evaluated on the language–property matrix shown in Table 8 (which is a very reduced version of the actual matrix). The fourth column shows the computed values according to equation (50), their sum and their (unweighted) average. The fifth column contains the weights for hybrid weighting; i.e. the product of the second column (antecedent weight) and the value from Figure 14 (phylogenetic weight) for every language. The sixth column combines the fourth and the fifth by multiplication. Finally, the average with hybrid weighting is the sum of the products divided by the sum of the weights.

Language–property matrix V				Hybrid weighting	
ℓ	$v^{\ell}_{\text{nsubj-obj-HEAD}}$	$v^{\ell}_{\text{HEAD-case}}$	$\mathcal{V}_F({\phi^{\ell}_{489}}')$	$w({\phi_{489}}', \ell)$	$\cdot$
IE, Germanic, English	0.00	0.04	1.00	0.000	0.000
IE, Germanic, German	0.46	0.01	0.54	0.058	0.031
IE, Romance, French	0.09	0.00	0.91	0.008	0.007
IE, Romance, Italian	0.07	0.00	0.93	0.006	0.005
IE, Romance, Spanish	0.16	0.01	0.84	0.013	0.011
Japanese, Japanese	0.96	1.00	1.00	0.480	0.480
$\sum$			5.22	0.564	0.534
$s({\phi_{489}}')$			0.87	0.95	

Table 8: Full example calculation table.

In the case of unweighted averaging, German causes the final score to be low; in the case of hybrid averaging, Japanese dominates the average since it is supported by both its antecedent weight and its phylogenetic weight.

Having two results for the same formula, it is natural to ask which weighting function is most appropriate when evaluating potential universals. No clear answer can be given at this point; a quantitative analysis of the framework paramters, including the weighting functions, follows in section 4.6.2.

4.5 Missing values

The actual language–property matrices can contain missing values, i.e. $V \in (U \cup \{\text{NULL}\})^{|L| \times |P|}$, and the definitions above do not consider this. In the implementation of the framework, missing values are handled by defining:

$$\text{If } \phi^{\ell} \text{ contains any } v^{\ell}_p = \text{NULL then } \mathcal{V}(\phi^{\ell}) = \text{NULL and } w(\phi, \ell) = 0. \tag{51}$$

This means that languages for which a formula is not evaluable are excluded from the final score.

4.6 Comparison between treebank statistics and typological databases

The framework for property vectors can be used to assign a real truth value to every formula. But before applying the framework to test new potential universals, it is crucial to see how it evaluates existing ones, i.e. high truth values should be assigned to established universals. A low score on established universals, on the other side, could be caused by errors in 1) the framework, 2) the UD treebanks, or 3) the typological comparison data. Section 3.5 revealed several inconsistencies in the UD treebanks. However, the following analyses assume that the overwhelming majority of UD annotations is correct, and that only the framework or the typological data can cause discrepancies. The first of the two following subsections focusses on atomic formulae (i.e. variables), the second on longer formulae.

4.6.1 World Atlas of Language Structures

The values in the World Atlas of Language Structures (WALS; Dryer and Haspelmath, 2013) which are extracted through the URIEL database are binary, but can be mapped to similar UD properties. For example, the WALS property vector $\vec{v}_{\text{SVO}}$ and the UD property vector $\vec{v}_{\text{nsubj:NOUN-HEAD:VERB-obj:NOUN}}$ should ideally be equivalent, i.e. $\mathcal{V}(v^\ell_{\text{SVO}} \Leftrightarrow v^\ell_{\text{nsubj:NOUN-HEAD:VERB-obj:NOUN}}) \overset{!}{\approx} 1$ or, equally, $\mathcal{V}(\neg(v^\ell_{\text{SVO}} \Leftrightarrow v^\ell_{\text{nsubj:NOUN-HEAD:VERB-obj:NOUN}})) \overset{!}{\approx} 0$ for every language ℓ. Generally speaking, for a WALS property u and a UD property p, the negated equivalence is equal to the absolute error and calculated as

$$\begin{aligned} e^\ell_{u,p} &= \mathcal{V}(\neg(v^\ell_u \Leftrightarrow v^\ell_p)) \\ &= 1 - \mathcal{V}(v^\ell_u \Leftrightarrow v^\ell_p) \\ &= |v^\ell_u - v^\ell_p|.^{20} \end{aligned} \tag{52}$$

[20] $1 - \mathcal{V}(v^\ell_u \Leftrightarrow v^\ell_p) = |v^\ell_u - v^\ell_p|$ is true for $\mathcal{V}_P$ and $\mathcal{V}_F$, but only if at least one of the two values v^ℓ_u and v^ℓ_p is exactly 0 or 1, which is the case for the binary WALS properties.

This makes it possible to calculate the mean absolute error (MAE) between the two property vectors by

$$MAE(\vec{v}_u, \vec{v}_p) = \frac{1}{|L|} \cdot \sum_{\ell \in L} e^{\ell}_{u,p} \,. \tag{53}$$

Since the vectors obtained from UD and URIEL can contain missing values, the MAE definition has to be adapted for vectors with possibly missing values (in parallel to the Euclidean distance on page 16).

$$MAE(\vec{v}_u, \vec{v}_p) = \frac{1}{\left|\{\ell \in L : v^{\ell}_u \neq \text{NULL} \wedge v^{\ell}_p \neq \text{NULL}\}\right|} \cdot \sum_{\substack{\ell \in L \\ v^{\ell}_u \neq \text{NULL} \\ v^{\ell}_p \neq \text{NULL}}} e^{\ell}_{u,p} \tag{54}$$

Table 23 shows the MAE for all WALS properties from URIEL that can be properly mapped to UD properties. The MAE ranges between 0 (no difference in any language) and 1 (maximum difference in every language).

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.41	0.23	$\vec{v}_{\text{OBLIQUE\_AFTER\_OBJECT}}$	$\vec{v}_{\text{HEAD:VERB-obj:NOUN-obl:NOUN}}$ $+\vec{v}_{\text{obj:NOUN-HEAD:VERB-obl:NOUN}}$ $+\vec{v}_{\text{obj:NOUN-obl:NOUN-HEAD:VERB}}$
0.37	0.21	$\vec{v}_{\text{OBLIQUE\_BEFORE\_OBJECT}}$	$\vec{v}_{\text{HEAD:VERB-obl:NOUN-obj:NOUN}}$ $+\vec{v}_{\text{obl:NOUN-HEAD:VERB-obj:NOUN}}$ $+\vec{v}_{\text{obl:NOUN-obj:NOUN-HEAD:VERB}}$
0.33	0.22	$\vec{v}_{\text{VOX}}$	$\vec{v}_{\text{HEAD:VERB-obj:NOUN-obl:NOUN}}$
0.24	0.19	$\vec{v}_{\text{XVO}}$	$\vec{v}_{\text{obl:NOUN-HEAD:VERB-obj:NOUN}}$
0.23	0.25	$\vec{v}_{\text{SUBJECT\_AFTER\_VERB}}$	$\vec{v}_{\text{HEAD:VERB-nsubj:NOUN}}$
0.19	0.11	$\vec{v}_{\text{OBLIQUE\_AFTER\_VERB}}$	$\vec{v}_{\text{HEAD:VERB-obl:NOUN}}$
0.19	0.11	$\vec{v}_{\text{OBLIQUE\_BEFORE\_VERB}}$	$\neg\vec{v}_{\text{HEAD:VERB-obl:NOUN}}$
0.17	0.16	$\vec{v}_{\text{SUBJECT\_BEFORE\_VERB}}$	$\neg\vec{v}_{\text{HEAD:VERB-nsubj:NOUN}}$
0.14	0.17	$\vec{v}_{\text{SVO}}$	$\vec{v}_{\text{nsubj:NOUN-HEAD:VERB-obj:NOUN}}$
0.13	0.31	$\vec{v}_{\text{OXV}}$	$\vec{v}_{\text{obj:NOUN-obl:NOUN-HEAD:VERB}}$
0.12	0.19	$\vec{v}_{\text{OBJECT\_BEFORE\_VERB}}$	$\neg\vec{v}_{\text{HEAD:VERB-obj:NOUN}}$
0.12	0.15	$\vec{v}_{\text{NUMERAL\_BEFORE\_NOUN}}$	$\neg\vec{v}_{\text{HEAD:NOUN-nummod:NUM}}$

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.12	0.15	$\vec{v}_{\text{NUMERAL\_AFTER\_NOUN}}$	$\vec{v}_{\text{HEAD:NOUN-nummod:NUM}}$
0.11	0.18	$\vec{v}_{\text{OBJECT\_AFTER\_VERB}}$	$\vec{v}_{\text{HEAD:VERB-obj:NOUN}}$
0.09	0.21	$\vec{v}_{\text{RELATIVE\_BEFORE\_NOUN}}$	$\neg\vec{v}_{\text{HEAD:NOUN-acl:VERB}}$
0.09	0.14	$\vec{v}_{\text{SUBJECT\_AFTER\_OBJECT}}$	$\vec{v}_{\text{HEAD:VERB-obj:NOUN-nsubj:NOUN}}$ $+\vec{v}_{\text{obj:NOUN-HEAD:VERB-nsubj:NOUN}}$ $+\vec{v}_{\text{obj:NOUN-nsubj:NOUN-HEAD:VERB}}$
0.09	0.14	$\vec{v}_{\text{SUBJECT\_BEFORE\_OBJECT}}$	$\vec{v}_{\text{HEAD:VERB-nsubj:NOUN-obj:NOUN}}$ $+\vec{v}_{\text{nsubj:NOUN-HEAD:VERB-obj:NOUN}}$ $+\vec{v}_{\text{nsubj:NOUN-obj:NOUN-HEAD:VERB}}$
0.09	0.17	$\vec{v}_{\text{SOV}}$	$\vec{v}_{\text{nsubj:NOUN-obj:NOUN-HEAD:VERB}}$
0.08	0.18	$\vec{v}_{\text{RELATIVE\_AFTER\_NOUN}}$	$\vec{v}_{\text{HEAD:NOUN-acl:VERB}}$
0.07	0.19	$\vec{v}_{\text{VSO}}$	$\vec{v}_{\text{HEAD:VERB-nsubj:NOUN-obj:NOUN}}$
0.07	0.20	$\vec{v}_{\text{ADPOSITION\_AFTER\_NOUN}}$	$\vec{v}_{\text{HEAD:NOUN-case:ADP}}$
0.07	0.11	$\vec{v}_{\text{ADJECTIVE\_BEFORE\_NOUN}}$	$\neg\vec{v}_{\text{HEAD:NOUN-amod:ADJ}}$
0.07	0.11	$\vec{v}_{\text{ADJECTIVE\_AFTER\_NOUN}}$	$\vec{v}_{\text{HEAD:NOUN-amod:ADJ}}$
0.05	0.19	$\vec{v}_{\text{SUBORDINATOR\_WORD\_BEFORE\_CLAUSE}}$	$\neg\vec{v}_{\text{HEAD:VERB-mark:SCONJ}}$
0.05	0.15	$\vec{v}_{\text{XOV}}$	$\vec{v}_{\text{obl:NOUN-obj:NOUN-HEAD:VERB}}$
0.04	0.05	$\vec{v}_{\text{OVS}}$	$\vec{v}_{\text{obj:NOUN-HEAD:VERB-nsubj:NOUN}}$
0.04	0.06	$\vec{v}_{\text{OVX}}$	$\vec{v}_{\text{obj:NOUN-HEAD:VERB-obl:NOUN}}$
0.03	0.15	$\vec{v}_{\text{SUBORDINATOR\_WORD\_AFTER\_CLAUSE}}$	$\vec{v}_{\text{HEAD:VERB-mark:SCONJ}}$
0.03	0.06	$\vec{v}_{\text{ADPOSITION\_BEFORE\_NOUN}}$	$\neg\vec{v}_{\text{HEAD:NOUN-case:ADP}}$
0.03	0.12	$\vec{v}_{\text{VOS}}$	$\vec{v}_{\text{HEAD:VERB-obj:NOUN-nsubj:NOUN}}$
0.02	0.04	$\vec{v}_{\text{OSV}}$	$\vec{v}_{\text{obj:NOUN-nsubj:NOUN-HEAD:VERB}}$

Table 9: Mean absolute error (MAE) between WALS and UD properties and standard deviation over languages, sorted by MAE.

Appendix B.1 provides an extended version of the table which also shows major discrepancies for individual languages. We will only look at two examples more precisely. The overall MAE[21] is 0.10 ($\sigma = 0.17$) which indicates—bearing in mind that the WALS values are restricted to the extremes 0 and 1—a high conformity between the manually collected WALS information and the computationally extracted UD information.

Oblique after/before object The properties with the highest MAE concern the order of oblique and object. Table 10 shows all languages with an absolute error greater than or equal to 0.5: According to the WALS, Cantonese, Amharic, Estonian, Vietnamese and Komi Zyrian are OX-only languages (for all of them, there is a 1 in the second column and a 0 in the fifth

	O-X			X-O		
$e^\ell_{u,p}$	v^ℓ_u	v^ℓ_p	$e^\ell_{u,p}$	v^ℓ_u	v^ℓ_p	ℓ
0.99	1.00	0.01	(0.01	1.00	0.99)	Sino-Tibetan, Chinese
0.90	1.00	0.10	0.90	0.00	0.90	Sino-Tibetan, Cantonese
0.70	1.00	0.30	0.70	0.00	0.70	Afro-Asiatic, Semitic, Amharic
0.57	1.00	0.43	0.57	0.00	0.57	Uralic, Finnic, Estonian
0.52	1.00	0.48	0.52	0.00	0.52	Austro-Asiatic, Viet-Muong, Vietnamese
0.51	1.00	0.49	0.51	0.00	0.51	Uralic, Permic, Komi Zyrian

Table 10: Absolute error, WALS value and UD value of the properties "object before oblique" (O-X) and "object after oblique" (X-O) for all languages with an absolute error greater than or equal to 0.5.

[21] The overall MAE is calculated as

$$MAE_{overall} = \frac{1}{\sum\limits_{(u,p)} |\{\ell \in L : v^\ell_u \neq \text{NULL} \wedge v^\ell_p \neq \text{NULL}\}|} \cdot \sum_{(u,p)} \sum_{\substack{\ell \in L \\ v^\ell_u \neq \text{NULL} \\ v^\ell_p \neq \text{NULL}}} e^\ell_{u,p} \,. \tag{55}$$

If there were no missing values, the overall MAE would be identical to the average MAE over property mappings and languages,

$$MAE_{average} = \frac{1}{|\{(u,p)\}| \cdot |L|} \cdot \sum_{(u,p)} \sum_{\ell \in L} e^\ell_{u,p} \,. \tag{56}$$

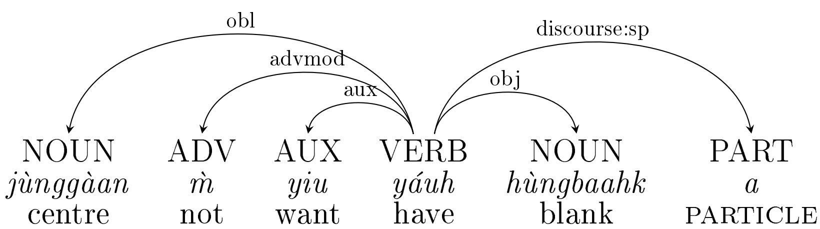
column), but the most frequent order in the UD treebanks is the opposite (for all of them, the value in the third column is smaller than the value in the sixth column).

For the classification of Cantonese as OX language, the WALS refers to Matthews and Yip (1994, p. 115–127)[22], although one can find clear examples with X-O order in their work. For example the sentence in Figure 15 which is syntactically similar to the sentence in Figure 16.

kéuih	*lóuhgūng*	*hái*	*ngoihmihn*	*yáuh*	*go*	*léuihyán*
her	husband	at	outside	have	CLASSIFIER	woman

‘Her husband has a mistress’

Figure 15: Cantonese example sentence from Matthews and Yip (1994, p. 117) with object and oblique.



‘There should be no blank in the middle’

Figure 16: Cantonese example sentence from the HK treebank with one object and one oblique.

Chinese is an XO and an OX language according to the WALS, but almost exclusively shows X-O ordering in the UD treebanks. Figure 17 shows an example sentence with X-O ordering. The WALS refers to Li and Thompson (1981, p. 24)[23] who also give examples for both X-V(-O) and V(-O)-X

[22]Dryer and Gensler (2013), Datapoint Cantonese. Available online at https://wals.info/valuesets/84A-cnt, accessed on 2020-04-16.

[23]Idem, Datapoint Mandarin. Available online at https://wals.info/valuesets/84A-mnd, accessed on 2020-04-16.

constructions.[24] Both constructions are possible and also appear in the Chinese treebanks, as Figure 18 shows. However, the ordering X-O is much more frequent.[25]

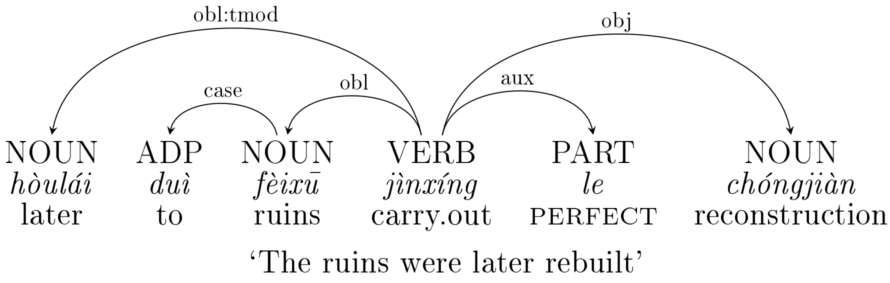


Figure 17: Chinese example sentence from the PUD treebank with one object and two preceding obliques.

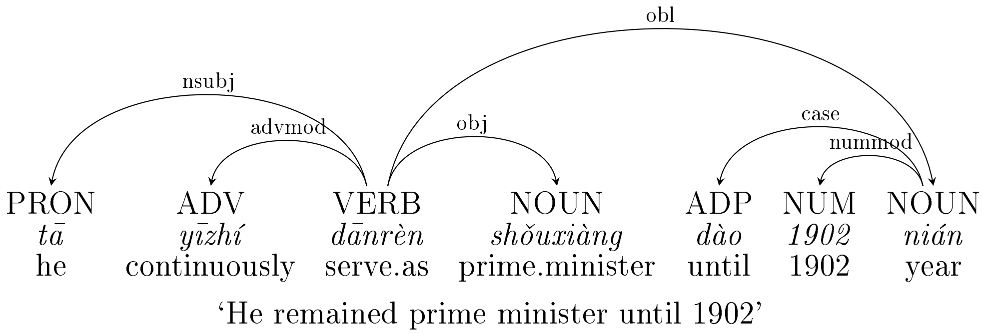


Figure 18: Chinese example sentence from the GSD treebank with one object and one succeeding oblique.

[24] There is a semantic difference between X-V and V-X constructions. A temporal phrase X expresses the time point of the action in X-V and the duration of the action in V-X constructions; a locative phrase X expresses the location of the action in X-V and the location of the result in V-X constructions (Li and Thompson, 1981, p. 21–23).

[25] This draws attention to another issue: The WALS values actually just indicate whether a construction is possible or not, whereas the UD values indicate how frequent a construction is. Imagine X-O and O-X constructions were equally frequent in the Chinese treebanks. Then the MAE would be 0.5 for X-O and 0.5 for O-X which is, on average, the same as the calculated 0.01 for X-O and 0.99 for O-X. This issue could be tackled by normalising the WALS properties which have a "common denominator" before comparing them to UD properties. Then the normalised WALS value for "oblique after object" would be $\frac{v^{\ell}_{\text{OBLIQUE\_AFTER\_OBJECT}}}{v^{\ell}_{\text{OBLIQUE\_AFTER\_OBJECT}} + v^{\ell}_{\text{OBLIQUE\_BEFORE\_OBJECT}}}$. The normalisation would affect only about 2% of the WALS values and is not performed because the simple equation (52) requires binary values (see footnote 20).

VOS in Tagalog With an MAE of 0.03, WALS and UD generally agree on the property “VOS”, i.e. the dominant order being verb–subject–object. The only significant exception is Tagalog. The WALS classifies it as a VSO language and refers to Kroeger (1993, p. 111).[26] In the UD treebank for Tagalog, however, the order V-O-S appears in 94% of the cases. So, Tagalog seems to be either a VSO or a VOS language. It is not possible to fully choose one of these options here; e.g. because the Tagalog treebank contains less than 1,000 tokens. It is, however, possible to say that the information in the WALS regarding the dominant word order(s) of Tagalog is not fully represented in the literature. Even Kroeger (1993, p. 112) assumes that both are “unmarked” word orders after summarising previous works. Figure 19 shows an example sentence from the Tagalog treebank.

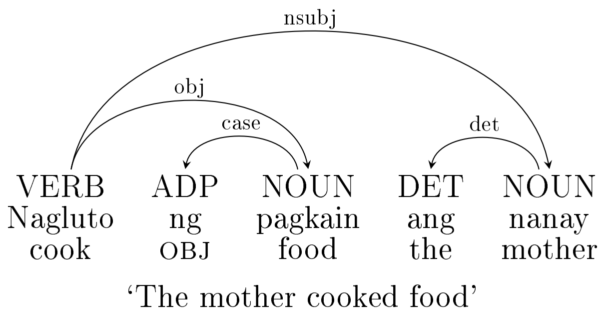
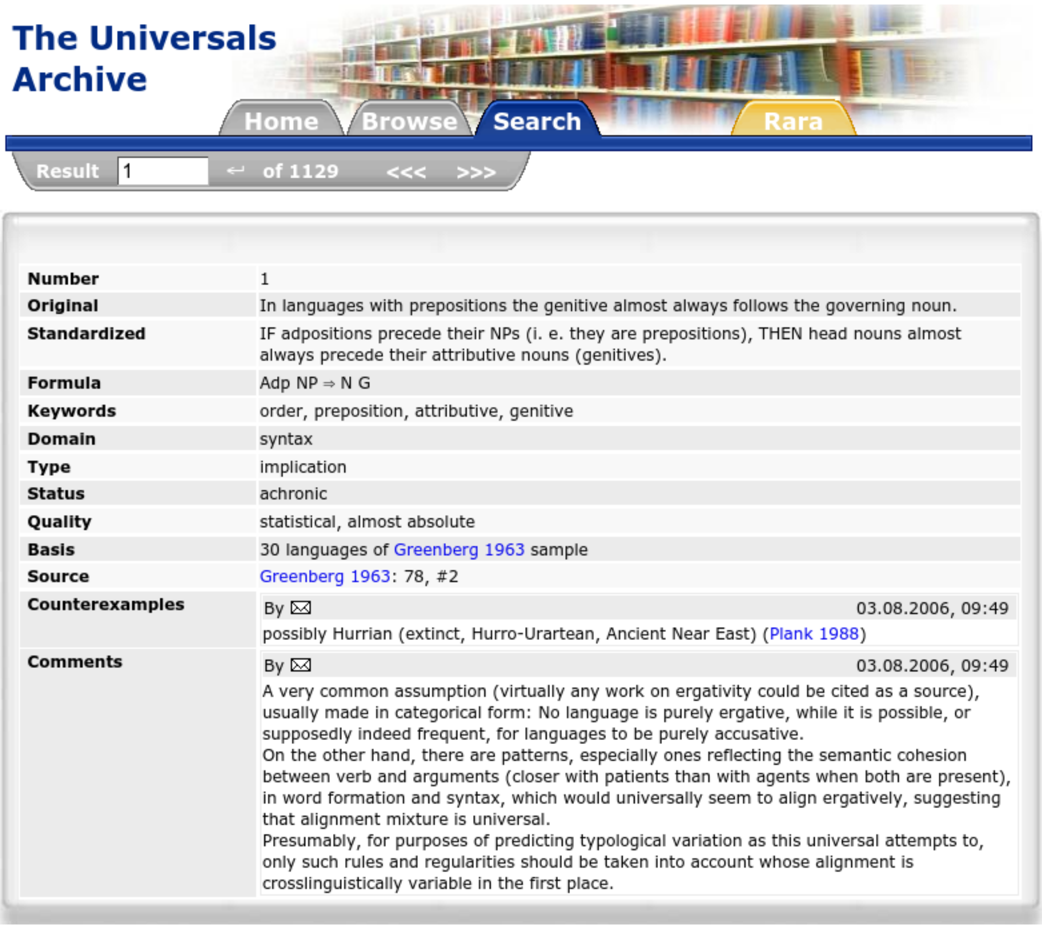


Figure 19: Tagalog example V-O-S sentence from the TRG treebank.

4.6.2 Universals Archive

The Universals Archive (UA; Plank and Filimonova, 2000) contains over 2,000 universals from different sources, including Greenberg’s universals. 1,129 universals are categorised as syntactic universals. From those, 482 universals have a non-empty “Formula” field, e.g. the universal in Figure 20. After crawling the archive, the variables of the formulae were mapped to (terms

[26] Dryer (2013), Datapoint Tagalog. Available online at https://wals.info/valuesets/81A-tag, accessed on 2020-04-16.



Number	1
Original	In languages with prepositions the genitive almost always follows the governing noun.
Standardized	IF adpositions precede their NPs (i. e. they are prepositions), THEN head nouns almost always precede their attributive nouns (genitives).
Formula	Adp NP ⇒ N G
Keywords	order, preposition, attributive, genitive
Domain	syntax
Type	implication
Status	achronic
Quality	statistical, almost absolute
Basis	30 languages of Greenberg 1963 sample
Source	Greenberg 1963: 78, #2
Counterexamples	By ✉ 03.08.2006, 09:49 possibly Hurrian (extinct, Hurro-Urartean, Ancient Near East) (Plank 1988)
Comments	By ✉ 03.08.2006, 09:49 A very common assumption (virtually any work on ergativity could be cited as a source), usually made in categorical form: No language is purely ergative, while it is possible, or supposedly indeed frequent, for languages to be purely accusative. On the other hand, there are patterns, especially ones reflecting the semantic cohesion between verb and arguments (closer with patients than with agents when both are present), in word formation and syntax, which would universally seem to align ergatively, suggesting that alignment mixture is universal. Presumably, for purposes of predicting typological variation as this universal attempts to, only such rules and regularities should be taken into account whose alignment is crosslinguistically variable in the first place.

Figure 20: First universal in the UA.

of) UD properties. For example, "Adp NP" is mapped to $\neg v^{\ell}_{\text{HEAD:NOUN-case:ADP}}$. Only in 66 cases, all variables of the formula could be mapped (variables like "agreement in gender", "head-final order", "rising contour" and many others cannot be expressed with UD properties).[27] All 66 universals are listed in appendix B.3.

Each universal has a "Basis" field which contains either a list of languages

[27]Sometimes the mapping is not bijective. For example, "N-Rel" is mapped to $v^{\ell}_{\text{HEAD:NOUN-acl:VERB}}$ which includes not only relative clauses but also other clausal modifiers, and "N-Gen" is mapped to $v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}}$ which includes not only genitives but also other attributes.

on which the formula was tested or a reference to a typological study in which one can find a list of languages. There is also a “Counterexamples” field with a list of languages for which the universal does not hold. To get an impression of how well a universal is supported in the literature, we can divide the number of tested languages for which the universal holds by the number of all tested languages:

$$\hat{s}(\phi_i) = \frac{|L_{basis}(i) \setminus L_{counter}(i)|}{|L_{basis}(i) \cup L_{counter}(i)|} \tag{57}$$

Hereby, i denotes the index of the universal. Universal 1 (in Figure 20) was tested on the 30 languages in Greenberg (1963): Basque, Berber, Burmese, Burushaski, Chibcha, Finnish, Fulani, Greek, Guarani, Hebrew, Hindi, Indonesian, Italian, Japanese, Kannada, Loritja, Maori, Masai, Maya, Norwegian, Nubian, Quechua, Serbian, Songhai, Swahili, Thai, Turkish, Welsh, Yoruba, and Zapotec. Hurrian is listed as a counterexample. Thus, $\hat{s}(\phi_1) = 30/31 \approx 97\%$ of the examined languages support the universal. However, the examined languages often differ greatly from the laguages which are included in the UD treebanks and the score might change if one looks only at the UD languages L_{UD}:

$$\hat{s}(\phi_i) = \begin{cases} \dfrac{|(L_{basis}(i) \setminus L_{counter}(i)) \cap L_{UD}|}{|(L_{basis}(i) \cup L_{counter}(i)) \cap L_{UD}|} & \text{UD languages} \\ \dfrac{|L_{basis}(i) \setminus L_{counter}(i)|}{|L_{basis}(i) \cup L_{counter}(i)|} & \text{all languages} \end{cases} \tag{58}$$

From the languages above, only 14 are contained in the UD treebanks: Basque, Finnish, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Norwegian, Serbian, Thai, Turkish, Welsh, and Yoruba. Thus, $\hat{s}(\phi_1) = 14/14 = 100\%$ of the examined languages which also appear in the UD treebanks support the universal.

The mean absolute error can be used to compare the UA scores and the UD scores of a set of universals (I denotes the set of the universals' indices):

$$MAE(I) = \frac{1}{|I|} \cdot \sum_{i \in I} e_i \tag{59}$$

with

$$e_i = |\hat{s}(\phi_i) - s(\phi_i)|. \tag{60}$$

For the calculation of s (see section 4.3), L is set as

$$L = \begin{cases} (L_{basis}(i) \cup L_{counter}(i)) \cap L_{UD} & \text{UD languages} \\ L_{UD} & \text{all languages} \end{cases}. \tag{61}$$

In the "UD languages" condition, $\hat{s}$ and s are calculated on the exact same set of languages which makes the comparison more meaningful. In the "all languages" condition, $\hat{s}$ is calculated on all languages in the UA entry whereas s is calculated on all languages in the UD treebanks. Usually, there are much more UA languages which also represent a large number of language families whereas the UD treebanks contain a lot of Indo-European and other widely spoken languages. It is therefore especially interesting to compare the scores from these different sets of languages.

	MAE (σ)	*MAE* (σ)	*MAE* (σ)	*MAE* (σ)		
	Product		Fuzzy			
UD	0.08 (0.11)	0.17 (0.21)	0.09 (0.10)	0.15 (0.16)	N	phylogenetic
	0.07 (0.10)	0.16 (0.21)	0.07 (0.10)	0.13 (0.17)	Y	
all	0.11 (0.15)	0.19 (0.19)	0.11 (0.14)	0.17 (0.17)	N	phylogenetic
	0.10 (0.14)	0.18 (0.21)	0.10 (0.14)	0.15 (0.17)	Y	
	N	Y	N	Y		
	antecedent		antecedent			

Table 11: Mean absolute errors and standard deviations over formulae when comparing UA and UD.

Table 11 shows the MAEs and standard deviations for all 16 combinations of valuation (product logic / fuzzy logic), antecedent weighting (yes/no), phylogenetic weighting (yes/no) and languages (UD/all). Several observations can be made:

1) Product logic and fuzzy logic yield very similar results which indicates that the choice of the valuation function is not crucial.

2) Only phylogenetic weighting could slightly improve the congruence between UA and UD (phylogenetic > unweighted > hybrid > antecedent).

3) Unsurprisingly, the congruence is larger when using only UD languages instead of all languages.[28]

The relatively high standard deviations further indicate that the agreement of UA and UD is very different for the individual universals.

Table 12 ranks all universals which have been examined by absolute error in descending order ("UD languages" condition; no weighting). Errors yielded by fuzzy logic and product logic show only minor relative differences. Note that the first two columns of Table 12 do not show the average truth value of the formula but the difference between the average truth value computed by the framework (s) and the average truth value calculated from the Universals Archive ($\hat{s}$) (hence the values do not indicate how true a formula is but only how much UD and UA agree on the truth value). Appendix B.2 provides an extended version of the table which also shows major discrepancies for individual languages. Most universals from the UA are pretty well reflected in the UD treebanks. Only 13/51 universals[29] have an absolute difference greater than or equal to 0.1. Apparently, the majority of those universals are an equivalence, or could be reformulated as an equivalence, which are harder to fulfil than one-way implications.

e_i			
$\mathcal{V}_F$	$\mathcal{V}_P$	i	Formula
0.39	0.55	1346	(PP-V $\Leftrightarrow$ Adv-V) $\wedge$ (V-PP $\Leftrightarrow$ V-Adv)
0.34	0.34	110	(Rel-N $\Rightarrow$ Adj-N) $\wedge$ (N-Rel $\Rightarrow$ N-Adj)
0.34	0.34	964	(V-O $\Rightarrow$ Art-N) $\wedge$ (O-V $\Rightarrow$ N-Art)

[28] Furthermore, all numbers are lower when not using POS information.

[29] 15 of the original 66 universals have been excluded from this analysis because the original study is undiscoverable and hence no $\hat{s}$ values can be calculated.

e_i			
$\mathcal{V}_F$	$\mathcal{V}_P$	i	Formula
0.33	0.33	10	Common-Proper ⇒ Gen-N
0.27	0.28	107	(V-O ⇒ (N-Rel ∧ N-Adj ∧ N-Gen)) ∧ (O-V ⇒ (Rel-N ∧ Adj-N ∧ Gen-N))
0.25	0.26	1343	(O-V ⇒ PP-V) ∧ (V-O ⇒ V-PP)
0.25	0.25	9	(Proper-Common ⇒ N-Gen) ∧ (Common-Proper ⇒ Gen-N)
0.19	0.16	112	((V-O ∧ Adv-V) ⇒ Adj-N) ∧ ((O-V ∧ V-Adv) ⇒ N-Adj)
0.19	0.15	1153	Rel-N ⇒ N-Rel
0.14	0.12	111	V-Adv ⇒ N-Adj
0.14	0.15	1419	(O-V ⇒ Pred-Cop) ∧ (V-O ⇒ Cop-Pred)
0.14	0.12	501	(V-S-O ⇒ Aux-V) ∧ (S-O-V ⇒ V-Aux)
0.10	0.08	92	Adp-NP ⇒ (((N-Dem ∨ N-Num) ⇒ N-Adj) ∧ (N-Adj ⇒ N-Gen) ∧ (N-Gen ⇒ N-Rel))
0.08	0.10	75	(N-Num ⇒ N-Adj) ∧ (Adj-N ⇒ Num-N)
0.07	0.05	66	O-V ⇒ (Adj-N ⇒ Gen-N)
0.07	0.04	74	Adp-NP ⇒ (N-Num ⇒ N-Adj)
0.06	0.07	489	S-O-V ⇒ NP-Adp
0.06	0.05	423	Adp-NP ⇒ (¬ S-O-V ⇒ N-Rel)
0.06	0.06	498	O-V ⇒ $\text{VP}_{\text{subordinate}}$-$\text{V}_{\text{main}}$
0.06	0.08	57	Adj-N ⇒ (Dem-N ∧ Num-N)
0.06	0.05	1334	Adp-NP ⇒ ((Rel-N ⇒ Gen-N) ∧ (Gen-N ⇒ Adj-N) ∧ (Adj-N ⇒ Dem-N))
0.06	0.05	114	N-Dem ⇒ N-Adj
0.05	0.04	60	Adj-Adv ⇒ (N-Adj ∧ V-O)
0.05	0.03	502	V-$\text{O}_{\text{pronominal}}$ ⇒ V-$\text{O}_{\text{nominal}}$
0.05	0.06	88	(N-Dem ⇒ N-Rel) ∧ (Rel-N ⇒ Dem-N)
0.04	0.03	65	S-O-V ⇒ (Adj-N ⇒ Gen-N)
0.04	0.05	126	(N-Num ⇒ N-Rel) ∧ (Rel-N ⇒ Num-N)
0.04	0.04	1	Adp-NP ⇒ N-Gen
0.04	0.04	73	(N-Dem ⇒ N-Adj) ∧ (Adj-N ⇒ Dem-N)
0.03	0.01	174	Adp-NP ⇒ (N-Num ⇒ N-Gen)

e_i			
$\mathcal{V}_F$	$\mathcal{V}_P$	i	Formula
0.03	0.02	72	Adp-NP ⇒ (N-Dem ⇒ N-Adj)
0.03	0.02	173	Adp-NP ⇒ (N-Dem ⇒ N-Gen)
0.03	0.04	176	(N-Gen ⇒ N-Rel) ∧ (Rel-N ⇒ Gen-N)
0.02	0.02	2	NP-Adp ⇒ Gen-N
0.02	0.02	87	Adp-NP ⇒ (N-Dem ⇒ N-Rel)
0.02	0.02	56	V-S-O ⇒ N-Adj
0.02	0.01	175	Adp-NP ⇒ (N-Gen ⇒ N-Rel)
0.02	0.01	69	Adp-NP ⇒ (N-Adj ⇒ N-Gen)
0.02	0.00	121	Adp-NP ⇒ (N-Num ⇒ N-Rel)
0.02	0.01	95	NP-Adp ⇒ ((((Adj-N ∨ Rel-N) ⇒ (Dem-N ∧ Num-N)) ∧ (Dem-N ∨ Num-N)) ⇒ Gen-N)
0.02	0.01	177	NP-Adp ⇒ (Dem-N ⇒ Gen-N)
0.02	0.01	71	NP-Adp ⇒ (Adj-N ⇒ Gen-N)
0.02	0.01	178	NP-Adp ⇒ (Num-N ⇒ Gen-N)
0.02	0.01	70	(Adp-NP ∧ ¬ S-V-O) ⇒ (N-Adj ⇒ N-Gen)
0.01	0.00	91	Adp-NP ⇒ (N-Adj ⇒ N-Rel)
0.01	0.01	1114	V-O ⇒ Comp-Sent
0.01	0.00	68	V-initial ⇒ (N-Adj ⇒ N-Gen)
0.01	0.00	67	V-S-O ⇒ (N-Adj ⇒ N-Gen)
0.01	0.00	5	(S-O-V ∧ N-Gen) ⇒ N-Adj
0.00	0.00	62	Rel-N ⇒ (NP-Adp ∨ Adj-N)
0.00	0.00	55	V-S-O ⇒ Adp-NP

Table 12: Absolute differences between UA score and UD score for individual universals, sorted by the difference when using fuzzy logic; the variable names from the UA are unified (see appendix B.3 for the original formulae and other details).

A closer look is taken at universal 110. There are no known counterexamples listed in the UA, i.e. all languages in $L_{basis}(110) \cap L_{UD}$ should support it and $\hat{s}(\phi_{110}) = 1$. The languages with the greatest differences to 1 are Hungarian, Swedish, Danish, English, Basque, Russian and Polish. Figure 21 shows that the universal has indeed low scores for the language families Uralic, Indo-European and Basque. The universal consists of two implications which are the two directions of an equivalence relation (Rel-N $\Leftrightarrow$ Adj-N). Table 13 breaks down the implications and shows that all of those languages but Basque fail at the second implication since they have relative clauses after nouns and adjectives before nouns. Basque, on the other hand, puts relative clauses before nouns and adjectives after nouns. Universal 110 is also addressed in the linguistic discussion in section 5.3.

ℓ	(Rel-N	$\Rightarrow$	Adj-N)	$\wedge$	(N-Rel	$\Rightarrow$	N-Adj)
Uralic, Ugric, Hungarian	0.00	1.00	1.00	**0.00**	1.00	**0.00**	0.00
IE, Germanic, Swedish	0.00	1.00	0.99	**0.01**	1.00	**0.01**	0.01
IE, Germanic, Danish	0.01	0.99	0.97	**0.03**	0.99	**0.03**	0.03
IE, Germanic, English	0.03	0.98	0.98	**0.03**	0.97	**0.03**	0.02
Basque, Basque	0.99	**0.16**	0.16	**0.16**	0.01	0.99	0.84
IE, Slavic, Russian	0.19	0.98	0.98	**0.19**	0.81	**0.19**	0.02
IE, Slavic, Polish	0.00	1.00	0.66	**0.34**	1.00	**0.34**	0.34

Table 13: Evaluating universal 110 with fuzzy logic. The conjunction in the middle is the minimum of both implications.

- 0.53 All
  - 0.94 Afro-Asiatic
    - 0.94 Egyptian
    - 0.94 Semitic
  - 0.92 Austronesian
    - 0.92 Malayo-Sumbawan
  - **0.16 Basque**
  - **0.36 Code switching**
  - **0.00 Creole**
  - 0.65 Dravidian
    - 1.00 South Central
    - **0.30 Southern**
  - **0.42 IE**
    - 0.69 Armenian
    - **0.30 Baltic**
    - 0.90 Celtic
    - **0.09 Germanic**
    - **0.28 Greek**
    - 0.66 Indic
    - 0.95 Iranian
    - **0.47 Latin**
    - 0.69 Romance
    - **0.16 Slavic**
  - 1.00 Japanese
  - 1.00 Korean
  - 0.75 Mande
  - 0.97 Mongolic
  - 0.96 Niger-Congo
    - 0.92 Defoid
    - 1.00 Northern Atlantic
  - **0.17 Sign Language**
  - 0.85 Sino-Tibetan
  - 0.91 Tai-Kadai
  - 0.98 Tupian
  - 1.00 Turkic
    - 1.00 Northwestern
    - 1.00 Southeastern
    - 0.99 Southwestern
  - **0.25 Uralic**
    - **0.15 Finnic**
    - 0.71 Mordvin
    - **0.42 Permic**
    - **0.26 Sami**
    - **0.00 Ugric**

Figure 21: Average truth value for universal 110, using fuzzy logic, for all language families in the UD treebanks. Language families with a score lower than 0.5 are displayed in boldface.

5 Search for New Universals

The framework introduced in the last section can be used to evaluate formulae of potential universals. Searching new universals means enumerating possible and/or promising formulae and checking whether some of them achieve a high score. In this section, two different searches are performed on the property vectors. Known universals are confirmed and new universals are proposed. At the end there is linguistic analysis, comparing two theories that predict word order. Fuzzy logic and phylogenetic weighting are used for all computations in this section.

5.1 Simple formulae

Simple formulae, more precisely implications between two variables, have been investigated in many previous studies. If one considers $|P|$ variables of which each can occur as positive or negative atom, then there are $(2 \cdot |P|)^2$ possible implications. Since $\phi \Rightarrow \phi$ is not interesting and $\phi_1 \Rightarrow \phi_2 \equiv \neg\phi_2 \Rightarrow \neg\phi_1$, the number of implications of interest is $\binom{2 \cdot |P|}{2}$.

Table 14 shows the values for implications of ten variables when evaluating on all UD languages.[30] Note that $\phi \Leftrightarrow \phi$ does not always evaluate to 1; in fuzzy logic, $\mathcal{V}_F(\phi \Leftrightarrow \phi) = \max\{1 - \mathcal{V}_F(\phi), \mathcal{V}_F(\phi)\}$. It becomes apparent that the subject precedes its head in the majority of UD languages, resulting in high scores for implications where "HEAD-nsubj" is the antecedent or "nsubj-HEAD" is the consequent (principle of explosion). Those implications are responsible for about half of the cells in Table 15, which shows the implications with the highest scores when using the same ten variables as in Table 14.[31]

[30]Throughout this section, $\neg\vec{v}_{\text{HEAD-}R}$ is written as $\vec{v}_{R\text{-HEAD}}$.

[31] Standard deviations interpret as follows. $s(\phi)$ is the (weighted) average of $\mathcal{V}(\phi^\ell)$ over all $\ell \in L$. Since $s(\phi) \in [0, 1]$, the corresponding (weighted) standard deviation σ is not larger than 0.5 (Popoviciu's inequality). The standard deviation gives insights about the truth values on the language level: $\sigma = 0$ iff $\mathcal{V}(\phi^\ell) = s(\phi)$ for all languages, and $\sigma = 0.5$ iff $\mathcal{V}(\phi^\ell) = 0$ for half the languages and $\mathcal{V}(\phi^\ell) = 1$ for the other half.

	HEAD-acl	acl-HEAD	HEAD-advmod	advmod-HEAD	HEAD-amod	amod-HEAD	HEAD-aux	aux-HEAD	HEAD-case	case-HEAD	HEAD-det	det-HEAD	HEAD-nmod	nmod-HEAD	HEAD-nsubj	nsubj-HEAD	HEAD-nummod	nummod-HEAD	HEAD-obj	obj-HEAD
HEAD-acl	0.92	0.41	0.58	0.78	0.72	0.68	0.49	0.87	0.60	0.77	0.56	0.85	0.75	0.64	0.46	0.93	0.57	0.85	0.81	0.58
acl-HEAD	0.59	0.92	0.64	0.90	0.68	0.89	0.87	0.70	0.95	0.60	0.65	0.89	0.60	0.97	0.60	0.96	0.65	0.89	0.64	0.92
HEAD-advmod	0.90	0.78	0.77	0.71	0.86	0.81	0.75	0.90	0.81	0.87	0.74	0.83	0.82	0.82	0.74	0.92	0.76	0.85	0.85	0.78
advmod-HEAD	0.64	0.58	0.29	0.77	0.52	0.71	0.55	0.65	0.71	0.55	0.37	0.81	0.51	0.70	0.31	0.92	0.37	0.83	0.58	0.62
HEAD-amod	0.89	0.68	0.71	0.81	0.91	0.57	0.65	0.88	0.74	0.81	0.72	0.78	0.82	0.71	0.58	0.95	0.72	0.79	0.80	0.72
amod-HEAD	0.68	0.72	0.52	0.86	0.43	0.91	0.70	0.71	0.80	0.61	0.46	0.92	0.55	0.86	0.46	0.94	0.45	0.92	0.65	0.73
HEAD-aux	0.70	0.87	0.65	0.90	0.71	0.88	0.89	0.62	0.93	0.67	0.64	0.92	0.67	0.93	0.64	0.95	0.63	0.92	0.66	0.89
aux-HEAD	0.87	0.49	0.55	0.75	0.70	0.65	0.38	0.89	0.60	0.75	0.56	0.79	0.71	0.64	0.41	0.93	0.55	0.80	0.78	0.52
HEAD-case	0.60	0.77	0.55	0.87	0.61	0.81	0.75	0.67	0.95	0.45	0.56	0.89	0.52	0.91	0.48	0.95	0.56	0.88	0.56	0.87
case-HEAD	0.95	0.60	0.71	0.81	0.80	0.74	0.60	0.93	0.55	0.95	0.65	0.85	0.86	0.67	0.60	0.93	0.65	0.87	0.93	0.60
HEAD-det	0.89	0.85	0.81	0.83	0.92	0.78	0.79	0.92	0.85	0.89	0.84	0.76	0.87	0.84	0.77	0.96	0.83	0.83	0.87	0.85
det-HEAD	0.65	0.56	0.37	0.74	0.46	0.72	0.56	0.64	0.65	0.56	0.24	0.84	0.45	0.74	0.30	0.90	0.30	0.84	0.62	0.57
HEAD-nmod	0.97	0.64	0.70	0.82	0.86	0.71	0.64	0.93	0.67	0.91	0.74	0.84	0.92	0.59	0.63	0.92	0.74	0.86	0.88	0.67
nmod-HEAD	0.60	0.75	0.51	0.82	0.55	0.82	0.71	0.67	0.86	0.52	0.45	0.87	0.41	0.92	0.43	0.94	0.45	0.87	0.59	0.77
HEAD-nsubj	0.96	0.93	0.92	0.92	0.94	0.95	0.93	0.95	0.93	0.95	0.90	0.96	0.94	0.92	0.93	0.90	0.92	0.95	0.94	0.90
nsubj-HEAD	0.60	0.46	0.31	0.74	0.46	0.58	0.41	0.64	0.60	0.48	0.30	0.77	0.43	0.63	0.10	0.93	0.26	0.78	0.52	0.52
HEAD-nummod	0.89	0.85	0.83	0.85	0.92	0.79	0.80	0.92	0.87	0.88	0.84	0.83	0.87	0.86	0.78	0.95	0.87	0.77	0.85	0.86
nummod-HEAD	0.65	0.57	0.37	0.76	0.45	0.72	0.55	0.63	0.65	0.56	0.30	0.83	0.45	0.74	0.26	0.92	0.23	0.87	0.61	0.56
HEAD-obj	0.92	0.58	0.62	0.78	0.73	0.72	0.52	0.89	0.60	0.87	0.57	0.85	0.77	0.67	0.52	0.90	0.56	0.86	0.88	0.48
obj-HEAD	0.64	0.81	0.58	0.85	0.65	0.80	0.78	0.66	0.93	0.56	0.62	0.87	0.59	0.88	0.52	0.94	0.61	0.85	0.52	0.88

Table 14: Implication matrix; row implies column.

Among those, there are some completely uninformative implications, e.g.

(#19) $\vec{v}_{\text{HEAD-advmod}} \Rightarrow \vec{v}_{\text{nsubj-HEAD}}$ and

(#20) $\vec{v}_{\text{advmod-HEAD}} \Rightarrow \vec{v}_{\text{nsubj-HEAD}}$

both achieve a score of 0.92. However, there are also some formulae that were already discovered in different works and are also included in the Universals Archive, such as

(#22) a. $\vec{v}_{\text{amod-HEAD}} \Rightarrow \vec{v}_{\text{nummod-HEAD}}$

b. Adj-N ⇒ Num-N

which is the 18th universal of Greenberg, the 57th in the UA and the 7th in Daumé and Campbell (2007).

#	$s(\phi)$	σ	Formula (UD)	Formula (UA-like)	UA	Greenberg	Daumé and Campbell (2007)	Bjerva et al. (2019b)	Lochbihler (2017)
1	0.97	0.07	$\vec{v}_{\text{acl-HEAD}} \Rightarrow \vec{v}_{\text{nmod-HEAD}}$	Rel-N ⇒ Gen-N	176				
2	0.96	0.07	$\vec{v}_{\text{acl-HEAD}} \Rightarrow \vec{v}_{\text{nsubj-HEAD}}$	Rel-N ⇒ S-V					
3	0.96	0.07	$\vec{v}_{\text{HEAD-det}} \Rightarrow \vec{v}_{\text{nsubj-HEAD}}$	N-Art ⇒ S-V					
4	0.95	0.11	$\vec{v}_{\text{acl-HEAD}} \Rightarrow \vec{v}_{\text{HEAD-case}}$	Rel-N ⇒ NP-Adp	62	24	30		
5	0.95	0.10	$\vec{v}_{\text{HEAD-amod}} \Rightarrow \vec{v}_{\text{nsubj-HEAD}}$	N-Adj ⇒ S-V					
6	0.95	0.07	$\vec{v}_{\text{HEAD-aux}} \Rightarrow \vec{v}_{\text{nsubj-HEAD}}$	V-Aux ⇒ S-V					
7	0.95	0.07	$\vec{v}_{\text{HEAD-case}} \Rightarrow \vec{v}_{\text{nsubj-HEAD}}$	NP-Adp ⇒ S-V					
8	0.95	0.08	$\vec{v}_{\text{HEAD-nsubj}} \Rightarrow \vec{v}_{\text{nummod-HEAD}}$	V-S ⇒ Num-N					
9	0.94	0.09	$\vec{v}_{\text{amod-HEAD}} \Rightarrow \vec{v}_{\text{nsubj-HEAD}}$	Adj-N ⇒ S-V					4[1]
10	0.94	0.09	$\vec{v}_{\text{nmod-HEAD}} \Rightarrow \vec{v}_{\text{nsubj-HEAD}}$	Gen-N ⇒ S-V					
11	0.94	0.09	$\vec{v}_{\text{HEAD-nsubj}} \Rightarrow \vec{v}_{\text{HEAD-obj}}$	V-S ⇒ V-O				3	
12	0.93	0.11	$\vec{v}_{\text{HEAD-acl}} \Rightarrow \vec{v}_{\text{nsubj-HEAD}}$	N-Rel ⇒ S-V					
13	0.93	0.12	$\vec{v}_{\text{HEAD-aux}} \Rightarrow \vec{v}_{\text{HEAD-case}}$	V-Aux ⇒ NP-Adp					
14	0.93	0.11	$\vec{v}_{\text{HEAD-aux}} \Rightarrow \vec{v}_{\text{nmod-HEAD}}$	V-Aux ⇒ Gen-N					
15	0.93	0.10	$\vec{v}_{\text{aux-HEAD}} \Rightarrow \vec{v}_{\text{nsubj-HEAD}}$	Aux-V ⇒ S-V					
16	0.93	0.18	$\vec{v}_{\text{case-HEAD}} \Rightarrow \vec{v}_{\text{nsubj-HEAD}}$	Adp-NP ⇒ S-V					
17	0.93	0.12	$\vec{v}_{\text{case-HEAD}} \Rightarrow \vec{v}_{\text{HEAD-obj}}$	Adp-NP ⇒ V-O			12	2	
18	0.92	0.17	$\vec{v}_{\text{acl-HEAD}} \Rightarrow \vec{v}_{\text{obj-HEAD}}$	Rel-N ⇒ O-V			9		
19	0.92	0.15	$\vec{v}_{\text{HEAD-advmod}} \Rightarrow \vec{v}_{\text{nsubj-HEAD}}$	V-Adv ⇒ S-V					
20	0.92	0.10	$\vec{v}_{\text{advmod-HEAD}} \Rightarrow \vec{v}_{\text{nsubj-HEAD}}$	Adv-V ⇒ S-V					5[1]
21	0.92	0.12	$\vec{v}_{\text{amod-HEAD}} \Rightarrow \vec{v}_{\text{det-HEAD}}$	Adj-N ⇒ Art-N[2]	57	18	13	7	
22	0.92	0.09	$\vec{v}_{\text{amod-HEAD}} \Rightarrow \vec{v}_{\text{nummod-HEAD}}$	Adj-N ⇒ Num-N	57	18	7		
23	0.92	0.17	$\vec{v}_{\text{HEAD-aux}} \Rightarrow \vec{v}_{\text{det-HEAD}}$	V-Aux ⇒ Art-N					
24	0.92	0.12	$\vec{v}_{\text{HEAD-aux}} \Rightarrow \vec{v}_{\text{nummod-HEAD}}$	V-Aux ⇒ Num-N					
25	0.92	0.14	$\vec{v}_{\text{HEAD-nmod}} \Rightarrow \vec{v}_{\text{nsubj-HEAD}}$	N-Gen ⇒ S-V					
26	0.92	0.11	$\vec{v}_{\text{HEAD-nsubj}} \Rightarrow \vec{v}_{\text{HEAD-nummod}}$	V-S ⇒ N-Num					
27	0.91	0.23	$\vec{v}_{\text{HEAD-case}} \Rightarrow \vec{v}_{\text{nmod-HEAD}}$	NP-Adp ⇒ Gen-N	1	2	1	1	
28	0.90	0.19	$\vec{v}_{\text{acl-HEAD}} \Rightarrow \vec{v}_{\text{advmod-HEAD}}$	Rel-N ⇒ Adv-V					
29	0.90	0.12	$\vec{v}_{\text{HEAD-advmod}} \Rightarrow \vec{v}_{\text{aux-HEAD}}$	V-Adv ⇒ Aux-V					
30	0.90	0.19	$\vec{v}_{\text{det-HEAD}} \Rightarrow \vec{v}_{\text{nsubj-HEAD}}$	Art-N ⇒ S-V					

[1] Table 3.5 on page 26 in Lochbihler (2017).

[2] Everywhere else "Dem-N" instead of "Art-N"; the UD relation "det" covers all kinds of determiners, including demonstratives and articles.

Table 15: Top 30 implications of Table 14 with indices in previous works.

The alternative notation, e.g. in (#22) b. and (#1) b., is oriented towards the notation in the UA.[32]

The implication with the highest score,

(#1) a. $\vec{v}_{\text{acl-HEAD}} \Rightarrow \vec{v}_{\text{nmod-HEAD}}$

b. Rel-N ⇒ Gen-N

c. "IF the relative clause precedes the noun, THEN the genitive precedes the noun." (Universals Archive, #176),

concerning the order of adjectival clauses (such as relative clauses) and noun modifiers (such as genitives) and first described by Hawkins (1983), is listed in the UA but was not detected by the computer-assisted models of Daumé and Campbell (2007) and Bjerva et al. (2019b). This shows that corpus statistics can gain advantage over knowledge bases in some cases.

[32] The mapping from dependency relations to sentence elements relies on the most frequent usage of the relation. For example, "case-HEAD" is mapped to "Adp-NP" because the usual head of "case" is an adposition and the usual dependent a noun phrase:

case

(1) a.

DET	NOUN	ADP	DET	NOUN
the	keys	on	the	table

PP

Adp NP

b. the keys on the table

The reader might argue that these mappings are too imprecise since the dependency relation alone does not say anything about the syntactic categories of its head and dependent. In general, this is true. However, the single-link properties without POS are very similar to those with the most frequent POS. For example, $MAE(\vec{v}_{\text{case-HEAD}}, \vec{v}_{\text{case:ADP-HEAD:NOUN}}) = 0.03$ ($\sigma = 0.10$) and $MAE_{overall} = 0.05$ ($\sigma = 0.12$); appendix B.4 lists the MAEs for all relations. For the sake of simplicity, properties without POS information instead of POS-specific properties are used throughout this section.

Last but not least, completely new universals are suggested, e.g. the following four and their composition in (7).

(#13) V-Aux $\Rightarrow$ NP-Adp

(#14) V-Aux $\Rightarrow$ Gen-N

(#23) V-Aux $\Rightarrow$ Art-N

(#24) V-Aux $\Rightarrow$ Num-N

(7) a. V-Aux $\Rightarrow$ (NP-Adp $\wedge$ Gen-N $\wedge$ Art-N $\wedge$ Num-N)

b. IF the auxiliary follows the verb, THEN the adposition follows the noun phrase AND genitive, article and numeral all precede the noun.

5.2 Complex formulae

While given a finite set of variables the number of simple formulae is also finite, the actual number of formulae is infinite. This makes it impossible to enumerate all possibilities and even if only formulae up to a maximum length are considered, the search space becomes too large to test all of them.

5.2.1 Search space

The goal is to find formulae with a high truth value. For that, it is possible to define a search problem where

- formulae Φ are states,
- atomic formulae $\Phi_1 := \{\vec{v}_p : p \in P\} \cup \{\neg\vec{v}_p : p \in P\}$ are start states,
- very true formulae $\Phi_{fin} := \{\phi \in \Phi : s(\phi) \overset{!}{\approx} 1\}$ are final states and
- the successors of $\phi_1 \in \Phi$ are $\{\phi_1 \wedge \phi_2 : \phi_2 \in \Phi\} \cup \{\phi_1 \vee \phi_2 : \phi_2 \in \Phi\}$.

Since $\neg$, $\wedge$ and $\vee$ are functionally complete, it is not necessary to include $\Rightarrow$ or $\Leftrightarrow$ in the successor function to cover all possible truth values. However, the search space still contains many formulae which are syntactically different but semantically equivalent, e.g. $\phi_1 \wedge \phi_2 \equiv \phi_2 \wedge \phi_1$ or $(\phi_1 \wedge \phi_2) \wedge \phi_3 \equiv (\phi_1 \vee \phi_3) \wedge (\phi_2 \vee \phi_3)$ for all $\phi_1, \phi_2, \phi_3 \in \Phi$. Therefore, some limitations of the search space are discussed in the following.

5.2.2 Disjunctive normal forms

First, the search space is restricted to formulae in disjunctive normal form (DNF), i.e. a disjunction of conjunctions of atomic formulae. This does not limit the number of expressable truth values since every formula can be converted to DNF. DNFs have two main advantages over other normal forms:

1) Disjuncts with low truth values can be discarded before enumerating all DNFs (which immensely reduces the search space) since they would have little to no impact on the truth value of a disjunction.
2) DNFs can easily be read as implications since $\phi_1 \vee \phi_2 \equiv \neg\phi_1 \Rightarrow \phi_2 \equiv \neg\phi_2 \Rightarrow \phi_1$ for all $\phi_1, \phi_2 \in \Phi$.

5.2.3 Search complexity and k-best search

The algorithm on page 64 shows the construction of the DNFs. Lines with an asterisk (*) are added in the extended version described in section 5.2.4. Given the maximum formula length (i.e. the maximum number of atoms in a formula) $m \geq 1$, the number of syntactically different DNFs is the number of atoms plus the number of different disjuncts plus the number of different disjunctions not exceeding the length m:

$$\begin{aligned} |\Phi| &= 2 \cdot |P| + \sum_{n=2}^{m} (2 \cdot |P|)^n + \sum_{n=2}^{m} 2^{n-1} \cdot (2 \cdot |P|)^n \\ &\in O\left(2^m \cdot (2 \cdot |P|)^m\right) \end{aligned} \tag{62}$$

```
procedure FIND_BEST_FORMULAE(P, m, k)
    Φ_fin ← {}
    Φ_1 ← {}
    for p ∈ P do
        Φ_1 ← Φ_1 ∪ {v⃗_p}
        Φ_1 ← Φ_1 ∪ {¬v⃗_p}
    end for
  * Φ_1 ← {φ : φ ∈ Φ_1 and ADD(φ)}
  * Φ_fin, Φ_1 ← SORT_OUT_ONES(Φ_1)
    for 2 ≤ n ≤ m do
        Φ_n ← {}
        for φ_1 ∈ Φ_{n−1} do
            for φ_2 ∈ Φ_1 do
              * if ADD(φ_1, φ_2, ∧) then
                    Φ_n ← Φ_n ∪ {φ_1 ∧ φ_2}
              * end if
            end for
        end for
    end for
    Φ ← Φ_1 ∪ . . . ∪ Φ_m
    for 2 ≤ n ≤ m do
      * Φ ← BEST(Φ, k)
        Φ_n ← {}
        for φ_1 ∈ Φ do
            for φ_2 ∈ Φ do
                if |atoms(φ_1 ∨ φ_2)| ≤ m  * and ADD(φ_1, φ_2, ∨) then
                    Φ_n ← Φ_n ∪ {φ_1 ∨ φ_2}
                end if
            end for
        end for
      * Φ_f, Φ_n ← SORT_OUT_ONES(Φ_n)
      * Φ_fin ← Φ_fin ∪ Φ_f
        Φ ← Φ ∪ Φ_n
    end for
    Φ_fin ← Φ_fin ∪ Φ
  * Φ_fin ← BEST(Φ_fin, k)
    return Φ_fin
end procedure
```

Construct atoms

Construct disjuncts (conjunctions)

Construct DNFs (disjunctions)

Hereby, 2^{n-1} is the number of partitions of n without ordering. Since conjunction and disjunction are commutative operations, formulas which solely differ in the ordering of their parts are considered the same state. The number of different DNFs decreases to

$$\begin{aligned} |\Phi| &\approx 2 \cdot |P| + \sum_{n=2}^{m} \binom{2 \cdot |P|}{n} + \sum_{n=2}^{m} \left[\frac{\exp\left(\pi \cdot \sqrt{\frac{2 \cdot n}{3}}\right)}{4 \cdot n \cdot \sqrt{3}} \right] \cdot \binom{2 \cdot |P|}{n} \\ &\in O\left(\frac{2^m}{m} \cdot \frac{(2 \cdot |P|)^m}{m!} \right) \end{aligned} \tag{63}$$

where $\frac{\exp\left(\pi \cdot \sqrt{\frac{2 \cdot n}{3}}\right)}{4 \cdot n \cdot \sqrt{3}}$ is the approximate number of partitions of n with ordering (Hardy and Ramanujan, 1918, p. 79).

The search space is additionally reduced by a k-best search. In each iteration of the DNF construction, Φ is sorted by $\lambda\phi : s(\phi)$ and reduced to the k elements with the highest scores:

$$\begin{aligned} |\Phi| &\leq 2 \cdot |P| + \sum_{n=2}^{m} \binom{2 \cdot |P|}{n} + \sum_{n=2}^{m} k^2 \\ &\in O\left(\frac{(2 \cdot |P|)^m}{m!} \right) \end{aligned} \tag{64}$$

5.2.4 Restrictions

A final state is a formula that is almost true, i.e. the difference between 1 and the average score $s(\phi)$ is less than a certain threshold e_{max}. The function SORT_OUT_ONES separates final states from non-final states:

procedure SORT_OUT_ONES(Φ)
    $\Phi_{fin} \leftarrow \{\phi \in \Phi : 1 - s(\phi) < e_{max}\}$
    return $\Phi_{fin}, \Phi \setminus \Phi_{fin}$
end procedure

The final states are then removed from the search space, i.e. they have no successors, because they do not produce new states of interest by conjunction or disjunction: $\phi \wedge \phi_f \approxeq \phi$ and $\phi \vee \phi_f \approxeq \phi_f$ for all $\phi \in \Phi, \phi_f \in \Phi_{fin}$.

The function ADD (on page 67) checks whether a state should be added to the search space. These restrictions hinder those formulae from being added that are either unnecessarily long, or have a low truth value or a lot of missing values when being evaluated.

5.2.5 Parameter settings

The search parameters are set as follows:

Parameter name	Symbol	Value
languages	L	all 90 UD languages
properties	P	all 37 single-link properties
selected DNFs per iteration	k	500
maximum number of atoms	m	5
maximum error for final states	e_{max}	0.05
maximum percentage of missing values	p_{NULL}	0.25
minimum score for disjuncts	t	0.5
minimum improvement per disjunct	Δ_t	0.1

Table 16: Search parameters

POS information is not used since the 37 base relations are optimally distinct, whereas POS properties would lead to a much larger search space with very similar variables.

5.2.6 Final formatting, selection and ranking

Consider example (8). Final DNFs are transformed into implications if they consist of at least two disjuncts (a. to b.). If $s(\phi)$ is at least 0.95, ϕ is considered a potential universal. However, the truest formulae are not necessarily the most interesting universals. For this reason, the formulae are ranked by their corresponding equivalences' (c.) value. (This makes formulae with both-way implications appear at the top of the table.)

(8) a. $\vec{v}_{\text{HEAD-ccomp}} \vee (\vec{v}_{\text{nmod-HEAD}} \wedge \vec{v}_{\text{nsubj-HEAD}})$

b. $\vec{v}_{\text{ccomp-HEAD}} \Rightarrow (\vec{v}_{\text{nmod-HEAD}} \wedge \vec{v}_{\text{nsubj-HEAD}})$

c. $\vec{v}_{\text{ccomp-HEAD}} \Leftrightarrow (\vec{v}_{\text{nmod-HEAD}} \wedge \vec{v}_{\text{nsubj-HEAD}})$

procedure ADD($\phi_1, \phi_2 = \text{NULL}, \circ = \text{NULL}$)[33]

  if $\circ = \text{NULL}$ **then**

    return $|\{\ell \in L : \mathcal{V}(\phi_1^\ell) = \text{NULL}\}| \,/\, |L| \leq p_{\text{NULL}}$ **and**

    $s(\phi) \geq t$

  else if $\circ = \wedge$ **then**

    return $\phi_2 \notin \text{atoms}(\phi_1)$ **and**

    $\neg\phi_2 \notin \text{atoms}(\phi_1)$ **and**

    $|\{\ell \in L : \mathcal{V}(\phi_1^\ell \wedge \phi_2^\ell) = \text{NULL}\}| \,/\, |L| \leq p_{\text{NULL}}$ **and**

    $s(\phi_1 \wedge \phi_2) \geq t$

  else if $\circ = \vee$ **then**

    return $\text{disjuncts}(\phi_1) \cap \text{disjuncts}(\phi_2) = \varnothing$ **and**

    $\nexists\, \phi_a : \forall \phi \in \text{disjuncts}(\phi_1) \cup \text{disjuncts}(\phi_2) : \phi_a \in \text{atoms}(\phi)$ **and**

    $\nexists\, \phi_a, \phi_b \in \text{disjuncts}(\phi_1) \cup \text{disjuncts}(\phi_2) : \exists \phi :$

      $(\text{atoms}(\phi_a) \smallsetminus \text{atoms}(\phi_b)) \cup (\text{atoms}(\phi_b) \smallsetminus \text{atoms}(\phi_a)) = \{\phi, \neg\phi\}$ **and**

    $|\{\ell \in L : \mathcal{V}(\phi_1^\ell \vee \phi_2^\ell) = \text{NULL}\}| \,/\, |L| \leq p_{\text{NULL}}$ **and**

    $s(\phi_1 \vee \phi_1) \geq \max\{s(\phi_1), s(\phi_2)\} +$

      $(|\text{disjuncts}(\phi_1) \cup \text{disjuncts}(\phi_2)| - 1) \cdot \Delta_s$

  end if

end procedure

---

[33]
- Atoms must
  1) have a percentage of missing values not greater than p_{NULL}.
  2) have an average truth value not lower than t.
- Disjuncts (conjunctions) must
  1) not contain the same atom twice.
  2) not contain the same variable as both positive and negative atom.
  3) have a percentage of missing values not greater than p_{NULL}.
  4) have an average truth value not lower than t.
- DNFs (disjunctions) must
  1) not contain a disjunct more than once.
  2) not contain an atom which is common to all disjuncts.
  3) not contain two disjuncts which are identical except for one variable which is contained as positive atom in one disjunct and as negative atom in the other disjunct.
  4) have a percentage of missing values not greater than p_{NULL}.
  5) improve the average truth value by at least $(n-1)\cdot\Delta_t$ with the n-th disjunct.

5.2.7 Results

23 formulae returned by the k-best search score 0.95 or higher. Table 17 shows them ranked by equivalence score.

#	$\Rightarrow$	σ	$\Leftrightarrow$	σ	Formula (UD)	Formula (UA-like)
1	1.00	0.00	1.00	0.00	$\vec{v}_{\text{HEAD-conj}}$	$\text{Conjunct}_1\text{-Conjunct}_2$
2	1.00	0.00	1.00	0.00	$\vec{v}_{\text{HEAD-fixed}}$	–
3	1.00	0.00	1.00	0.00	$\vec{v}_{\text{HEAD-flat}}$	–
4	1.00	0.00	1.00	0.00	$\vec{v}_{\text{HEAD-root}}$	ROOT-Sent
5	0.97	0.12	0.97	0.12	$\vec{v}_{\text{HEAD-appos}}$	N-N
6	0.95	0.11	0.83	0.26	$\vec{v}_{\text{case-HEAD}} \Rightarrow \vec{v}_{\text{HEAD-acl}}$	Adp-NP $\Rightarrow$ N-Rel
7	0.96	0.08	0.78	0.25	$\vec{v}_{\text{acl-HEAD}} \Rightarrow (\vec{v}_{\text{nmod-HEAD}} \wedge \vec{v}_{\text{nsubj-HEAD}})$	Rel-N $\Rightarrow$ (Gen-N $\wedge$ S-V)
8	0.97	0.07	0.77	0.26	$\vec{v}_{\text{acl-HEAD}} \Rightarrow \vec{v}_{\text{nmod-HEAD}}$	Rel-N $\Rightarrow$ Gen-N
9	0.95	0.11	0.75	0.27	$\vec{v}_{\text{HEAD-nmod}} \Rightarrow (\vec{v}_{\text{HEAD-acl}} \wedge \vec{v}_{\text{HEAD-xcomp}})$	N-Gen $\Rightarrow$ (N-Rel $\wedge$ $\text{V}_{\text{main}}\text{-VP}_{\text{subordinate}}$)
10	0.95	0.09	0.75	0.26	$\vec{v}_{\text{HEAD-nmod}} \Rightarrow (\vec{v}_{\text{HEAD-acl}} \wedge \vec{v}_{\text{HEAD-ccomp}})$	N-Gen $\Rightarrow$ (N-Rel $\wedge$ V-Sent')
11	0.95	0.08	0.75	0.32	$\vec{v}_{\text{ccomp-HEAD}} \Rightarrow (\vec{v}_{\text{nsubj-HEAD}} \wedge \vec{v}_{\text{HEAD-case}})$	Sent'-V $\Rightarrow$ (S-V $\wedge$ NP-Adp)
12	0.96	0.09	0.75	0.26	$\vec{v}_{\text{HEAD-nmod}} \Rightarrow (\vec{v}_{\text{cc-HEAD}} \wedge \vec{v}_{\text{HEAD-acl}})$	N-Gen $\Rightarrow$ (Conj-Conjunct_2 $\wedge$ N-Rel)
13	0.96	0.10	0.74	0.32	$\vec{v}_{\text{case-HEAD}} \Rightarrow \vec{v}_{\text{HEAD-xcomp}}$	Adp-NP $\Rightarrow$ $\text{V}_{\text{main}}\text{-VP}_{\text{subordinate}}$
14	0.97	0.07	0.74	0.34	$\vec{v}_{\text{case-HEAD}} \Rightarrow \vec{v}_{\text{HEAD-ccomp}}$	Adp-NP $\Rightarrow$ V-Sent'
15	0.96	0.13	0.70	0.27	$\vec{v}_{\text{case-HEAD}} \Rightarrow \vec{v}_{\text{HEAD-iobj}}$	Adp-NP $\Rightarrow$ V-iO
16	0.96	0.10	0.69	0.32	$\vec{v}_{\text{HEAD-nmod}} \Rightarrow \vec{v}_{\text{HEAD-xcomp}}$	N-Gen $\Rightarrow$ $\text{V}_{\text{main}}\text{-VP}_{\text{subordinate}}$
17	0.95	0.09	0.68	0.33	$\vec{v}_{\text{ccomp-HEAD}} \Rightarrow (\vec{v}_{\text{nmod-HEAD}} \wedge \vec{v}_{\text{nsubj-HEAD}})$	Sent'-V $\Rightarrow$ (Gen-N $\wedge$ S-V)
18	0.96	0.08	0.66	0.36	$\vec{v}_{\text{HEAD-nmod}} \Rightarrow \vec{v}_{\text{HEAD-ccomp}}$	N-Gen $\Rightarrow$ V-Sent'
19	0.95	0.07	0.64	0.26	$\vec{v}_{\text{HEAD-obl}} \Rightarrow \vec{v}_{\text{HEAD-ccomp}}$	V-PP $\Rightarrow$ V-Sent'
20	0.96	0.06	0.31	0.25	$\vec{v}_{\text{HEAD-nummod}} \Rightarrow \vec{v}_{\text{HEAD-parataxis}}$	N-Num $\Rightarrow$ $\text{Sent}_1\text{-Sent}_2$
21	0.96	0.08	0.25	0.25	$\vec{v}_{\text{HEAD-det}} \Rightarrow \vec{v}_{\text{HEAD-parataxis}}$	N-Art $\Rightarrow$ $\text{Sent}_1\text{-Sent}_2$

#	⇒	σ	⇔	σ	Formula (UD)	Formula (UA-like)
22	0.97	0.06	0.25	0.26	$\vec{v}_{\text{HEAD-nummod}} \Rightarrow \vec{v}_{\text{CC-HEAD}}$	N-Num ⇒ Conj-Conjunct$_2$
23	0.97	0.08	0.19	0.27	$\vec{v}_{\text{HEAD-det}} \Rightarrow \vec{v}_{\text{CC-HEAD}}$	N-Art ⇒ Conj-Conjunct$_2$

Table 17: Formulae returned by the search program with a score of 0.95 or higher, ranked by equivalence score.

The first five formulae are (almost) absolute universals and more a result of the UD annotation guidelines than an interesting linguistic universal. For example, conjuncts (#1) are always connected from left to right. #6–#8 also appear at the top of Table 15 (namely as #1, #2 and #4). Right after those follow some groups of formulae that can be combined:

(9) a. #9, #10, #12, #18:
N-Gen ⇒ (N-Rel ∧ V$_{\text{main}}$-VP$_{\text{subordinate}}$ ∧ V-Sent' ∧ Conj-Conjunct$_2$)

b. #11, #14, #17–#19:
Sent'-V ⇒ (S-V ∧ NP-Adp ∧ Gen-N ∧ PP-V)

c. #13–#15:
Adp-NP ⇒ (V$_{\text{main}}$-VP$_{\text{subordinate}}$ ∧ V-Sent' ∧ V-iO)

d. #13, #16:
VP$_{\text{subordinate}}$-V$_{\text{main}}$ ⇒ (NP-Adp ∧ Gen-N)

The equivalence scores for #9–#19 range between 0.75 and 0.64 and some of them shall receive a closer consideration in the linguistic analysis of section 5.3. The last four formulae of Table 17 are strong one-way implications with equivalence scores between 0.31 and 0.19. The consequent of #20 and #21 concerns the ordering of sentences in a parataxis which is generally annotated from left to right in the UD treebanks. In those few languages where parataxes are mainly annotated from right to left, numerals and determiners happen to precede their governor (except for Amharic, see Table 18), which supports the universal without any linguistic explanation. A similar phenomenon occurs at #22 and #23.

ℓ	HEAD-parataxis	HEAD-nummod	HEAD-det
Turkic, Southeastern, Uyghur	0.02	0.00	0.00
IE, Slavic, Bulgarian	0.03	0.08	0.42
Afro-Asiatic, Semitic, Amharic	0.17	0.03	0.85
IE, Romance, Old French	0.25	0.04	0.00
Turkic, Southwestern, Turkish	0.26	0.01	0.00
IE, Slavic, Polish	0.29	0.01	0.06
IE, Baltic, Lithuanian	0.39	0.11	0.02
Turkic, Northwestern, Kazakh	0.47	0.00	0.00

Table 18: Excerpt from the language-property matrix for languages with $v^{\ell}_{\text{HEAD-parataxis}} < 0.5$, sorted by $v^{\ell}_{\text{HEAD-parataxis}}$.

5.3 Linguistic analysis

Sections 4.6.2, 5.1 and 5.2 provide formulae for old and new universals of varying support in the UD treebanks. The following part introduces two linguistic theories which both make predictions about word order universals and examines whether the predictions can be confirmed in the UD treebanks.

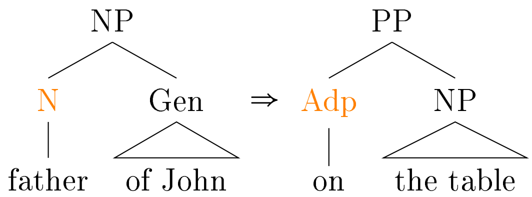
5.3.1 Head-dependent theory

Almost all high-scored formulae from Table 15 and Table 17 exhibit the form head$_1$-dependent$_1$ $\Rightarrow$ head$_2$-dependent$_2$ (or dependent$_2$-head$_1$ $\Rightarrow$ dependent$_2$-head$_2$), where "head" and "dependent" are used in terms of phrase-structure grammar and not in terms of dependency relations. Lochbihler (2017), who also evaluated Greenberg's universals on the UD treebanks and searched for new ones, made the same observation and argued that formulae with high scores support the head-dependent theory (HDT) whereas formulae with low scores don't satisfy it. The HDT suggests that each language consistently orders heads and dependents. For example, both phrases concerned by (the fist part of) Greenberg's universal 2 are left-headed:

(10) a. "IF adpositions follow their NPs (i.e. they are postpositions), THEN head nouns almost always follow their attributive nouns (genitives)." (Universals Archive, #2)

b. NP-Adp ⇒ Gen-N or, equivalently, N-Gen ⇒ Adp-NP

c.



Indeed, Table 17's #9 and #10 contain further consequents of N-Gen order and the involved phrases are left-headed in all of them:

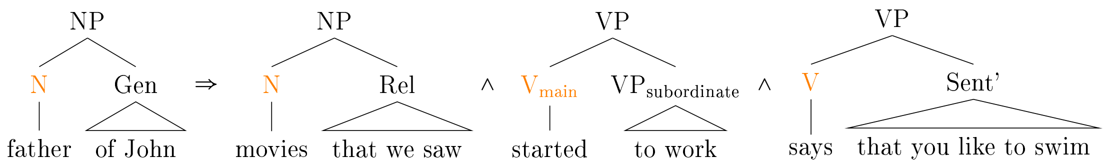


Figure 22: Consequents of N-Gen order.

Lochbihler (2017) distinguished between two types of dependents: complements and modifiers. According to her observations, she postulated that those formulae score low which either involve phrases with different headedness or phrases with different types of dependents. For example, Figure 23 (i) shows an equivalence which involves both a right head (O-V) and a left head (N-Gen) and is not supported by the UD data. Figure 23 (ii) on the other hand shows a merely weakly supported implication between a phrase with a complement (O) and a phrase with a modifier (Adv).

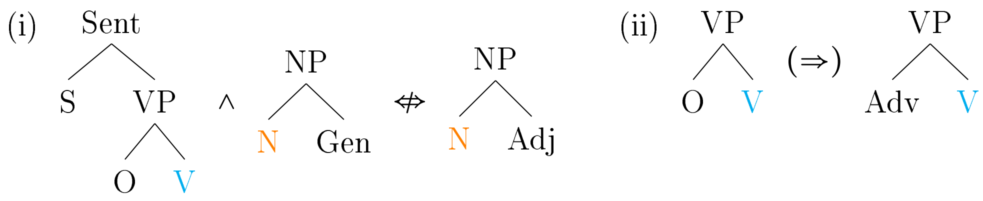


Figure 23: Unsupported universals (Lochbihler, 2017, p. 32)

This theory, however, cannot explain why adjectival clauses such as relative clauses, which are modifiers, occur frequently in the formulae of the search output, e.g. in Figure 22. Or why they behave differently than nonclausal adjectives, as the evaluation of universal 110 at the end of section 4.6.2 showed:

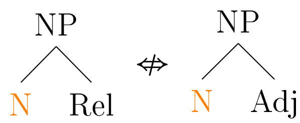


Figure 24: Clausal and single-word modifiers of the noun.

5.3.2 Branching-direction theory

Dryer (1992; 2009a) offers an alternative account, the branching-direction theory (BDT), according to which languages have the tendency to order nonphrasal and phrasal (sibling) constituents—rather than heads and complements—consistently. His investigations usually compare the order of verb and object with another pair of constituents and if there is a significant correlation he calls the latter "verb patterner" and "object patterner". According to the BDT, verb patterners are nonphrasal and object patterners are phrasal. The first part of Table 19 contains all V-O patterners from Dryer (2009a) which have UD equivalents; plus V-Sent' and V_{main}-$VP_{subordinate}$ which I added to the list myself. The equivalence scores measure the correlation between V-O and another constituent pair. The scores range from 0.52 to 0.90, where 0.9 is the strongest correlation (equivalence between V-O and V-O), 0.5 means no correlation and 0.1 would be the strongest negative correlation (equivalence between V-O and O-V). All V-O patterners but the last two, V-Adv and Art-N', show at least a moderate correlation.

Constituents		Example	UD equivalent (p)	s	σ
		nonphrasal + phrasal			
V	O	*ate + the sandwich*	HEAD:VERB-obj:NOUN	0.90	0.15
Cop	Pred	*is + a teacher*	cop:AUX-HEAD:NOUN	0.88	0.17
Adp	NP	*on + the table*	case:ADP-HEAD:NOUN	0.86	0.20
V	Sent'	*says + that you like to swim*	HEAD:VERB-ccomp:VERB	0.78	0.28
N	Rel	*movies + that we saw*	HEAD:NOUN-acl:VERB	0.76	0.31
Aux	VP	*has + eaten dinner*	aux:AUX-HEAD:VERB	0.75	0.32
V_{main}	$VP_{subordinate}$	*started + to work*	HEAD:VERB-xcomp:VERB	0.75	0.30
N	Gen	*father + of John*	HEAD:NOUN-nmod:NOUN	0.75	0.32
V	PP	*slept + on the floor*	HEAD:VERB-obl:NOUN	0.73	0.26
Comp	Sent	*that + John is sick*	mark:SCONJ-HEAD:VERB	0.70	0.35
V	Adv	*ran + slowly*	HEAD:VERB-advmod:ADV	0.60	0.30
Art	N'	*the + tall man*	det:DET-HEAD:NOUN	0.52	0.36
		nonphrasal + nonphrasal			
N	Adj		HEAD:NOUN-amod:ADJ	0.62	0.38
N	Dem		HEAD:NOUN-det:PRON	0.53	0.38
N	Num		HEAD:NOUN-nummod:NUM	0.49	0.36
		phrasal + phrasal			
VP	S		HEAD:VERB-csubj:VERB	0.55	0.40
VP	S		HEAD:VERB-nsubj:NOUN	0.54	0.40

Table 19: Word order correlations, sorted by $s(\vec{v}_{\text{HEAD:VERB-obj:NOUN}} \Leftrightarrow \vec{v}_p)$.

The tendency to order nonphrasal and phrasal constituents consistently may conflict with other cross-linguistic tendencies. For example, the BDT would predict the correlation in Figure 25. In the case of an OV language, however, the consistent branching conflicts with the general tendency that clauses follow their nonphrasal siblings (Dryer, 2003). While there are no known VO languages with Sent'-V order, Sent'-V order and V-Sent' order are about equally common among OV languages, i.e. V-O $\Rightarrow$ V-Sent' but O-V $\not\Rightarrow$ Sent'-V. The V-O $\Rightarrow$ V-Sent' inequivalence does also appear in Table 17: Adp-NP, N-Gen, V-PP all imply V-Sent' (#14, #18, #19) but the equivalence scores are significantly lower. The position of complementisers is affected by this and other principles as well (Dryer, 2009a; 2012).

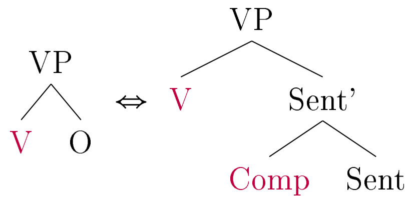


Figure 25: Complement clause (Sent') and complementiser (Comp) as verb patterners.

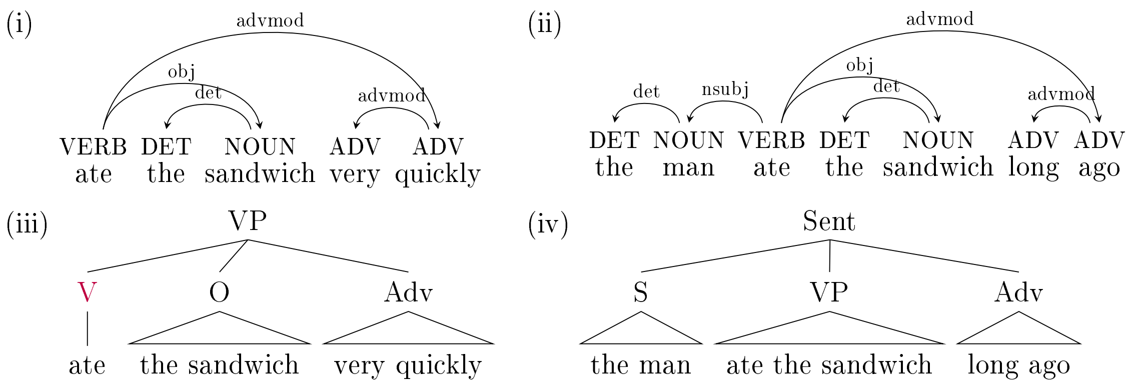


Figure 26: Adverbial modifiers: manner adverb in (i) dependency structure and (iii) phrase structure, and sentence adverb in (ii) dependency structure and (iv) phrase structure.

The case of adverbs is a bit unclear. First of all, an adverb is commonly a single word and thus not phrasal but Dryer (1992) found a high correlation between V-O and V-Adv and thus calls it an object patterner. Figure 26 illustrates different adverbial structures. That the correlation between V-O and V-Adv is not reflected in the UD data might be due to the circumstance that the UD relations make no distinction between adverbs that modify verbs (manner adverbs) as in (i) and adverbs that modify sentences (sentence adverbs) as in (ii), and Dryer (1992) only considers manner adverbs since sentence adverbs have a more flexible word order. If we assume that Adv is phrasal, then it should behave like the object within the VP, shown in (iii), but it has only phrasal siblings on the sentence level and thus no preference for its position, shown in (iv).

As only combinations of a nonphrasal and a phrasal sibling have been discussed so far, this is equal evidence for the BDT as well as the HDT since in the latter the nonphrasal sibling corresponds to the head and the phrasal sibling corresponds to the dependent and both theories make equal predictions about word order. Therefore, let us look at a region where the predictions differ: the complex noun phrase. First, have a look at the structures in Figure 27. All of them produce the same POS sequence but they have

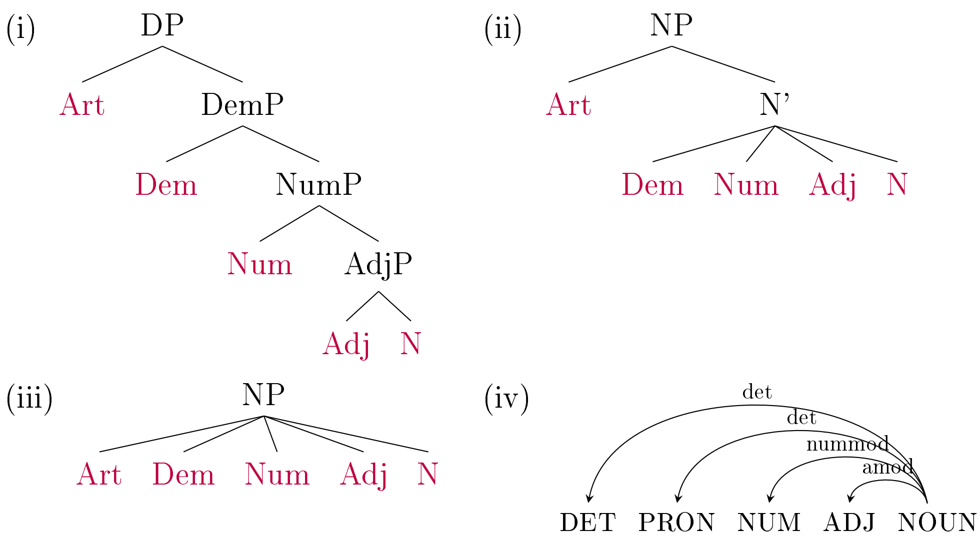


Figure 27: Noun projections in different grammars: (i) generative structure, (ii) flatter structure, (iii) flat structure, (iv) dependency structure.

different levels of complexity. (i) shows a binary structure which is commonly assumed in generative grammar. Although HDT and BDT are not bound to syntactic theories, the HDT is formulated in the terms of the "principles and parameters" framework—with head-directionality being a language-specific parameter—and has been used in X-bar theory and minimalism (Chomsky and Lasnik, 1995). Dryer (2009a), on the other hand, assumes a flatter structure, as in (ii) or (iii), following Culicover and Jackendoff (2005; 2006) who propose a simpler syntax than that in (i). Given the different grammars, HDT and BDT predict the following word orders for a VO language (and the reversed order for an OV language):

	HDT	BDT
(i)	Art-Dem-Num-Adj-N	Art-Dem-Num-{Adj,N}
(ii)	Art-N-{Adj,Dem,Num}	Art-{Adj,Dem,N,Num}
(iii)	N-{Adj,Art,Dem,Num}	{Adj,Art,Dem,N,Num}

Table 20: Predictions of HDT and BDT for a VO language, assuming the different grammars from Figure 27. Braces contain constituents with no predictable ordering.

(i) contains four heads and four dependents which, in a VO language, should be consistently right-branching according to the HDT. Similarly, the phrasal constituents should follow their nonphrasal siblings according to the BDT. This prediction does not hold up under a typological analysis. Leaving out the article, a word-order correlation for N and its dependents, Adj, Dem and Num, is already refuted by Dryer (1992; 2005; 2009a). While Cinque (2005) derives other word orders of Dem, Num, Adj and N from the underlying syntax in (i) by transformation, Dryer (2009a) moves to a same-level ordering of these four elements and attributes different orders to semantically-driven principles (see Dryer, 2018). In his typological investigation in Dryer (1992), he observes Art and N to be patterners of V and O, respectively, and suggests the structure in (ii). In 2009a, however, he moves to the structure in (iii) and explains the NP-peripheral position of the article by principles of constituent recognition and NP-internal semantic hierarchy. This leads us to what can be extracted from the UD treebanks: There is no correlation between V-O order and any of N-Art, N-Dem or N-Num order and at most a moderate correlation between V-O order and N-Adj order. In other words, the relative position of the noun to its dependents cannot be predicted from the order of verb and object. Even a correlation between verb/object and article/noun could not be reproduced, which does only support the structure in (iii) if a simple, non-transformational syntax as in the BDT is assumed.

The BDT can further explain why bare adjectives and adjectival/relative clauses (“amod” and “acl” in the UD treebanks) behave differently since the BDT does not predict a preferred ordering for two nonphrasal constituents, such as N and Adj:

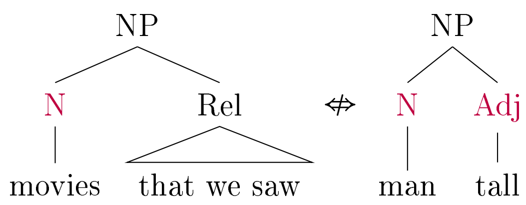


Figure 28: Phrasal and nonphrasal siblings of the noun.

Subjects are neither complements nor modifiers, but specifiers (if at all), and are not considered in the HDT. Since S and VP are both phrasal children of Sent, the BDT would not predict a correlation with the order of V and O. Indeed, the UD data shows no equivalence between these orderings (the last two rows of Table 19). To conclude the linguistic analysis, the UD data supports consistent ordering of words and phrases, i.e. the BDT, rather than consistent ordering of heads and dependents, i.e. the HDT.

6 Conclusion

This thesis pursued three goals:

1) Construct a matrix of continuous typological language vectors. The task specifically focused on continuous vectors with values in the range $[0, 1]$ since the underlying theory builds on a real-valued logic. Therefore, a new set of language vectors was extracted from the Universal Dependencies treebanks (Nivre et al., 2016). The representativity of the extracted vectors was measured using the normalised tree distance. The results showed that the language vectors represent the languages in sufficient detail to build approximate phylogenetic trees, although this property of the vectors is not required for the subsequent tasks. Additional analyses moved from language vectors to treebank vectors and searched for differences between treebanks for the same language. Therefore, the thesis also provides a list of verified annotation inconsistencies in the Universal Dependencies treebanks.

2) Develop a framework to process (logically complex) typological formulae. Transposing the matrix of language vectors yields a matrix of property vectors, e.g. vectors for S-O-V order or NP-Adp order (in which each dimension corresponds to a language). Logical connectives, such as negation, conjunction and implication were implemented in a framework to evaluate arbitrarily complex formulae. The valuation of the connectives is taken from both product logic and fuzzy logic with very similar results. Universal quantifications over languages are found to be best calculated as average with "phylogenetic weighting" which balances the influence of languages from different language families. The framework is useful for exploring universal statements about languages, in which researchers can design and evaluate their own property vectors and formulae. The accuracy of the framework was estimated by applying it to universals from the Universals Archive (Plank and Filimonova, 2000) which are attested to be true/false for many languages in typological studies and the results indicate that the framework works as expected.

3) Use the framework to discover new typological universals. Another interesting task beside the evaluation of existing universals is the search for new universals. Therefore, logical formulae were mapped into a search space and a search was performed to find promising candidates. Although some potential universals were detected with the treebanks, the search algorithm is just a very basic construction and leaves room for improvement; for example, the size of the search space still grows polynomially in the number of typological variables and the restriction parameters are not yet optimised. A linguistic analysis of the supported and unsupported universals completes this work and demonstrates that the treebank statistics support Dryer's (1992) branching-direction theory, which predicts the consistent order of nonphrasal and phrasal sibling constituents.

To summarise, this work presented language vectors and a complete framework for the evaluation of typological formulae of arbitrary complexity. The investigated language vectors are extracted from annotated corpora and can represent the syntactic nature of languages in more detail than typological databases. The framework was validated against traditional and computer-assisted typological studies and its different applications are demonstrated by using it for revisiting proposed universals, searching potential candidates, and finding new evidence in the discussion of word-order theories.

Bibliography

Bergmann, M. (2008). *An introduction to many-valued and fuzzy logic: Semantics, algebras, and derivation systems*. Cambridge University Press.

Bickel, B. (2007). Typology in the 21st century: Major current developments. *Linguistic Typology*, 11(1):239–251.

Bickel, B. (2011). Absolute and statistical universals. In Hogan, P. C., editor, *The Cambridge Encyclopedia of the Language Sciences*, pages 77–79. Cambridge University Press.

Bjerva, J. and Augenstein, I. (2018). From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916, New Orleans, Louisiana. Association for Computational Linguistics.

Bjerva, J., Kementchedjhieva, Y., Cotterell, R., and Augenstein, I. (2019a). A probabilistic generative model of linguistic typology. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1529–1540, Minneapolis, Minnesota. Association for Computational Linguistics.

Bjerva, J., Kementchedjhieva, Y., Cotterell, R., and Augenstein, I. (2019b). Uncovering probabilistic implications in typological knowledge bases. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3924–3930, Florence, Italy. Association for Computational Linguistics.

Bjerva, J., Östling, R., Veiga, M. H., Tiedemann, J., and Augenstein, I. (2019c). What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Bogdanowicz, D., Giaro, K., and Wróbel, B. (2012). TreeCmp: Comparison of trees in polynomial time. *Evolutionary Bioinformatics*, 8:475–487.

Bomhard, A. (2019). The origins of Proto-Indo-European: The Caucasian substrate hypothesis. *Journal of Indo-European Studies*, 47(1&2):9–124.

Chen, X. and Gerdes, K. (2017). Classifying languages by dependency structure. Typologies of delexicalized universal dependency treebanks. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 54–63, Pisa, Italy. Linköping University Electronic Press.

Chen, X. and Gerdes, K. (2019). The relation between dependency distance and frequency. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 75–82, Paris, France. Association for Computational Linguistics.

Chomsky, N. (1993). *Lectures on government and binding: The Pisa lectures.* Number 9 in Studies in Generative Grammar. Walter de Gruyter.

Chomsky, N. and Lasnik, H. (1995). The theory of principles and parameters. In *The minimalist program*, chapter 1, pages 13–127. MIT Press, Cambridge MA. First printed in: J. Jacobs A. von Stechow, W. Sternefeld, and T. Vennemann (eds.). *Syntax: an international handbook of contemporary research*, Walter de Gruyter, Berlin.

Cinque, G. (2005). Deriving Greenberg's universal 20 and its exceptions. *Linguistic inquiry*, 36(3):315–332.

Comrie, B. (1981). *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Culicover, P. W. and Jackendoff, R. (2006). The simpler syntax hypothesis. *Trends in cognitive sciences*, 10(9):413–418.

Culicover, P. W. and Jackendoff, R. S. (2005). *Simpler syntax*. Oxford University Press.

Daumé, III, H. and Campbell, L. (2007). A Bayesian model for discovering typological implications. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 65–72, Prague, Czech Republic. Association for Computational Linguistics.

Dryer, M. S. (1986). Word order consistency and English. In DeLancey, S. and Tomlin, R., editors, *Proceedings of the Second Annual Pacific Linguistics Conference*, pages 97–106, Eugene, Oregon. University of Oregon.

Dryer, M. S. (1988). Object-verb order and adjective-noun order: Dispelling a myth. *Lingua*, 74:185–217.

Dryer, M. S. (1989). Article-noun order. In *Papers from the 25th Regional Meeting of the Chicago Linguistic Society*, pages 83–97.

Dryer, M. S. (1992). The Greenbergian word order correlations. *Language*, 68:81–138.

Dryer, M. S. (2003). Adverbial subordinators and word order asymmetries. In Hawkins, J. A. and Siewierska, A., editors, *Performance principles of word order*, EUROTYP Working Papers, pages 50–67. European Science Foundation.

Dryer, M. S. (2005). Relationship between the order of object and verb and the order of adjective and noun. In Haspelmath, M., Dryer, M. S., Gil, D., and Comrie, B., editors, *World atlas of language structures*, pages 394–397. Oxford University Press.

Dryer, M. S. (2009a). The branching direction theory of word order correlations revisited. In *Universals of language today*, pages 185–207. Springer.

Dryer, M. S. (2009b). Problems testing typological correlations with the online WALS. *Linguistic Typology*, 13(1):121–135.

Dryer, M. S. (2012). On the position of interrogative phrases and the order of complementizer and clause. *Theories of Everything. In Honor of Ed Keenan. UCLA Working Papers*, 17:72–79.

Dryer, M. S. (2013). Order of subject, object and verb. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dryer, M. S. (2018). On the order of demonstrative, numeral, adjective, and noun. *Language*, 94(4):798–833.

Dryer, M. S. and Gensler, O. D. (2013). Order of object, oblique, and verb. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dryer, M. S. and Haspelmath, M., editors (2013). *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Futrell, R. (2019). Information-theoretic locality properties of natural language. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 2–15, Paris, France. Association for Computational Linguistics.

Gerdes, K., Kahane, S., and Chen, X. (2019). Rediscovering Greenberg's word order universals in UD. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 124–131, Paris, France. Association for Computational Linguistics.

Gray, R. D. and Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2:73–113.

Gulordava, K. and Merlo, P. (2015). Structural and lexical factors in adjective placement in complex noun phrases across Romance languages. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 247–257, Beijing, China. Association for Computational Linguistics.

Hardy, G. H. and Ramanujan, S. (1918). Asymptotic formulæ in combinatory analysis. *Proceedings of the London Mathematical Society*, 2(1):75–115.

Haspelmath, M., Dryer, M. S., Gil, D., and Comrie, B., editors (2005). *The World Atlas of Language Structures*. Oxford University Press.

Hawkins, J. A. (1983). *Word Order Universals*. Academic Press, New York.

Hawkins, J. A. (1990). A parsing theory of word order universals. *Linguistic Inquiry*, 21:223–261.

Hempel, C. G. (1945a). Studies in the logic of confirmation (I.). *Mind*, 54(213):1–26.

Hempel, C. G. (1945b). Studies in the logic of confirmation (II.). *Mind*, 54(214):97–121.

Huehnergard, J. and Pat-El, N. (2019). Introduction to the Semitic languages and their history. In *The Semitic languages*, chapter 1. Routledge, 2nd edition.

Itoh, Y. and Ueda, S. (2004). The Ising model for changes in word ordering rules in natural languages. *Physica D: Nonlinear Phenomena*, 198(3-4):333–339.

Jelínek, T. (2017). FicTree: A manually annotated treebank of Czech fiction. In *ITAT 2017 Proceedings*, pages 181–185.

Jøsang, A. and McAnally, D. (2005). Multiplication and comultiplication of beliefs. *International Journal of Approximate Reasoning*, 38(1):19–51.

Kahane, S. and Yan, C. (2019). Advantages of the flux-based interpretation of dependency length minimization. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 98–109, Paris, France. Association for Computational Linguistics.

Keenan, E. L. (1978). The syntax of subject-final languages. In Lehmann, W. P., editor, *Syntactic Typology: Studies in the Phenomenology of Language*, pages 267–328.

Kitchen, A., Ehret, C., Assefa, S., and Mulligan, C. J. (2009). Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proceedings of the Royal Society B: Biological Sciences*, 276(1668):2703–2710.

Kozinsky, I. (1981). *Nekotorye grammaticeskie universalii v podsistemax vyrazenija subjektno-objektnyx otnnosenij. [Some grammatical universals in subsystems of expression of subject-object relations.]*. PhD thesis, Moskovskij gosudarstvennyj universitet.

Kríž, V., Hladká, B., and Urešová, Z. (2016). Czech legal text treebank 1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2387–2392, Portorož, Slovenia. European Language Resources Association (ELRA).

Kroeger, P. (1993). *Phrase structure and grammatical relations in Tagalog.* Center for the Study of Language and Information (CSLI), Stanford, California.

Lehmann, W. P. (1973). A structural principle of languages and its implications. *Language*, 49:47–66.

Li, C. N. and Thompson, S. A. (1981). *Mandarin Chinese: A functional reference grammar.* Berkeley: University of California Press.

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., and Levin, L. (2017). URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Liu, H. (2010). Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578.

Lochbihler, B. C. (2017). Typology across treebanks. Master's thesis, Universität Stuttgart.

Łukasiewicz, J. (1930). Philosophische Bemerkungen zu mehrwertigen Systemen des Aussagenkalküls. *Comptes rendus des séances de la Société des Sciences et des Lettres de Varsovie*, 23, cl. iii:51–77. English translation by H. Weber, "Philosophical Remarks on Many-Valued Systems of Propositional Logic." In Storrs McCall (ed.), Polish Logic: 1920–1939, New York: Oxford University Press, 1967, pages 40–65.

Łukasiewicz, J. and Tarski, A. (1930). Untersuchungen über den Aussagenkalkül. *Comptes rendus des séances de la Société des Sciences et des Lettres de Varsovie*, 23, cl. iii:39–50. English translation by J. H. Woodger, "Investigations into the Sentential Calculus." In: Alfred Tarski, *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*, 2nd ed., Indianapolis: Hackett Publishing Co., 1983, pages 38–59.

Malaviya, C., Neubig, G., and Littell, P. (2017). Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press, online edition.

Matthews, S. and Yip, V. (1994). *Cantonese: A comprehensive grammar*. Routledge, New York.

Mir, A. and Rosselló, F. (2010). The mean value of the squared path-difference distance for rooted phylogenetic trees. *Journal of Mathematical Analysis and Applications*, 371(1):168–176.

Mizraji, E. (1992). Vector logics: The matrix-vector representation of logical calculus. *Fuzzy Sets and Systems*, 50(2):179–185.

Moravcsik, E. A. (1969). Determination. *Working Papers on Language Universals*, 1:63–98.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Noh, Y., Han, J., Oh, T. H., and Kim, H. (2018). Enhancing universal dependencies for Korean. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 108–116, Brussels, Belgium. Association for Computational Linguistics.

Oncevay, A., Haddow, B., and Birch, A. (2019). Towards a multi-view language representation: A shared space of discrete and continuous language features. In *Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP*.

Östling, R. and Tiedemann, J. (2017). Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.

Payne, D. L. (1990). *The Pragmatics of Word Order: Typological Dimensions of Verb Initial Languages*. Mouton de Gruyter, Berlin.

Plank, F. and Filimonova, E. (2000). The Universals Archive: A brief introduction for prospective users. *STUF-Language Typology and Universals*, 53(1):109–123.

Plank, F. and Mayer, T. (2006). The Universals Archive goes do-it-yourself. Walter de Gruyter.

Pullum, G. K. (1977). Word order universals and grammatical relations. In Cole, P. and Sadock, J., editors, *Grammatical Relations*, volume 8 of *Syntax and Semantics*, pages 249–277. Academic Press, New York.

Rabinovich, E., Ordan, N., and Wintner, S. (2017). Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540, Vancouver, Canada. Association for Computational Linguistics.

Serva, M. and Petroni, F. (2008). Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)*, 81(6):68005.

Sharma, K., Swami, K., Shete, A., and Husain, S. (2019). Can Greenbergian universals be induced from language networks? In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 25–37, Paris, France. Association for Computational Linguistics.

Smith, N. J. J. (2012). Many-valued logics. In Fara, D. G. and Russell, G., editors, *The Routledge Companion to Philosophy of Language*, pages 636–651. Routledge.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.

Ultan, R. (1978). Toward a typology of substantival possession. In Greenberg, J. H., editor, *Syntax*, volume 4 of *Universals of human language*, pages 11–50.

Ward, Jr., J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Yu, X., Falenska, A., Haid, M., Vu, N. T., and Kuhn, J. (2019a). IMSurReal: IMS at the surface realization shared task 2019. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*, pages 50–58, Hong Kong, China. Association for Computational Linguistics.

Yu, X., Falenska, A., and Kuhn, J. (2019b). Dependency length minimization vs. word order constraints: An empirical study on 55 treebanks. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 89–97, Paris, France. Association for Computational Linguistics.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, 8(3):338–353.

Zeman, D. (2015). Slavic languages in universal dependencies. *Natural Language Processing, Corpus Linguistics, E-learning (proceedings of SLOVKO 2015)*, pages 151–163.

Zeman, D., Nivre, J., Abrams, M., Aepli, N., Agić, Ž., Ahrenberg, L., Aleksandravičiūtė, G., Antonsen, L., Aplonova, K., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Basmov, V., Batchelor, C., Bauer, J., Bellato, S., Bengoetxea, K., Berzak, Y., Bhat, I. A., Bhat, R. A., Biagetti, E., Bick, E., Bielinskienė, A., Blokland, R., Bobicev, V., Boizou, L., Borges Völker, E., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Boyd, A., Brokaitė, K., Burchardt, A., Candito, M., Caron, B., Caron, G., Cavalcanti, T., Cebiroğlu Eryiğit, G., Cecchini, F. M., Celano, G. G. A., Čéplö, S., Cetin, S., Chalub, F., Choi, J., Cho, Y., Chun, J., Cignarella, A. T., Cinková, S., Collomb, A., Çöltekin, Ç., Connor, M., Courtin, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., de Souza, E., Diaz de Ilarraza, A., Dickerson, C., Dione, B., Dirix, P., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eckhoff, H., Eli, M., Elkahky, A., Ephrem, B., Erina, O., Erjavec, T., Etienne, A., Evelyn, W., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Fujita, K., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Garza, S., Gerdes, K., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., González Saavedra, B., Griciūtė, B., Grioni, M., Grūzītis, N., Guillaume, B., Guillot-Barbance, C., Habash, N., Hajič, J., Hajič jr., J., Hämäläinen, M., Hà Mỹ, L., Han, N.-R., Harris, K., Haug, D., Heinecke, J., Hennig, F., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Hwang, J., Ikeda, T., Ion, R., Irimia, E., Ishola, Ọ., Jelínek, T., Johannsen, A., Jørgensen, F., Juutinen, M., Kaşıkara, H., Kaasen, A., Kabaeva, N., Kahane, S., Kanayama, H., Kanerva, J., Katz, B., Kayadelen, T., Kenney, J., Kettnerová, V., Kirchner, J., Klementieva, E., Köhn, A., Kopacewicz, K., Kotsyba, N., Kovalevskaitė, J., Krek, S., Kwak, S., Laippala, V., Lambertino, L., Lam, L., Lando, T., Larasati, S. D., Lavrentiev, A., Lee, J., Lê H`ông, P., Lenci, A., Lertpradit, S., Leung, H., Li, C. Y., Li, J., Li, K., Lim, K., Liovina, M., Li, Y., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., McGuinness, S., Mendonça, G., Miekka, N., Misirpashayeva, M., Missilä, A., Mititelu, C., Mitrofan, M., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, K. S., Morioka, T., Mori, S., Moro, S., Mortensen, B., Moskalevskyi, B., Muischnek, K., Munro, R., Murawaki, Y., Müürisep, K., Nainwani, P., Navarro Horñiacek, J. I., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nguy˜ên Thị, L., Nguy˜ên Thị Minh, H., Nikaido, Y., Nikolaev, V., Nitisaroj, R., Nurmi, H., Ojala, S., Ojha, A. K., Olúòkun, A., Omura, M., Osenova, P., Östling, R., Øvrelid, L., Partanen, N., Pascual, E., Passarotti, M., Patejuk, A., Paulino-Passos, G., Peljak-Łapińska, A., Peng, S., Perez, C.-A., Perrier, G., Petrova, D., Petrov, S., Phelan, J., Piitulainen, J., Pirinen, T. A., Pitler, E., Plank, B., Poibeau, T., Ponomareva, L., Popel, M., Pretkalniņa, L., Prévost, S., Prokopidis, P., Przepiórkowski, A., Puolakainen, T., Pyysalo, S., Qi, P., Rääbis, A., Rademaker, A., Ramasamy, L., Rama, T., Ramisch, C., Ravishankar, V., Real, L., Reddy, S., Rehm, G., Riabov, I., Rießler, M., Rimkutė, E., Rinaldi, L., Rituma, L., Rocha, L., Romanenko, M., Rosa, R., Rovati, D., Roșca, V., Rudina, O., Rueter, J., Sadde, S., Sagot, B., Saleh, S., Salomoni, A., Samardžić, T., Samson, S., Sanguinetti, M., Särg, D., Saulīte, B., Sawanakunanon, Y., Schneider, N., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Shirasu, H., Shohibussirri, M., Sichinava, D., Silveira, A., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Soares-Bastos, I., Spadine, C., Stella, A., Straka, M., Strnadová, J., Suhr, A., Sulubacak, U., Suzuki, S., Szántó, Z., Taji, D., Takahashi, Y., Tamburini, F., Tanaka, T., Tellier, I., Thomas, G., Torga, L., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., Utka, A., Vajjala, S., van Niekerk, D., van Noord, G., Varga, V., Villemonte de la Clergerie, E., Vincze, V., Wallin, L., Walsh, A., Wang, J. X., Washington, J. N., Wendt, M., Williams, S., Wirén, M., Wittern, C., Woldemariam, T., Wong, T.-s., Wróblewska, A., Yako, M., Yamazaki, N., Yan, C (2019). Universal dependencies 2.5. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, Ç., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., and Li, J. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Appendix A: Language Vectors

A.1 Universal Dependencies treebanks

Language	Treebanks	Tokens
Afro-Asiatic, Egyptian, Coptic	1	40K
Afro-Asiatic, Semitic, Akkadian	1	1K
Afro-Asiatic, Semitic, Amharic	1	10K
Afro-Asiatic, Semitic, Arabic	3	1,042K
Afro-Asiatic, Semitic, Assyrian	1	<1K
Afro-Asiatic, Semitic, Hebrew	1	161K
Afro-Asiatic, Semitic, Maltese	1	44K
Austro-Asiatic, Viet-Muong, Vietnamese	1	43K
Austronesian, Central Philippine, Tagalog	1	<1K
Austronesian, Malayo-Sumbawan, Indonesian	2	141K
Basque, Basque	1	121K
Code switching, Hindi English	1	26K
Creole, Naija	1	12K
Dravidian, South Central, Telugu	1	6K
Dravidian, Southern, Tamil	1	9K
IE, Armenian, Armenian	1	52K
IE, Baltic, Latvian	1	220K
IE, Baltic, Lithuanian	2	75K
IE, Celtic, Breton	1	10K
IE, Celtic, Irish	1	40K
IE, Celtic, Scottish Gaelic	1	42K
IE, Celtic, Welsh	1	16K
IE, Germanic, Afrikaans	1	49K
IE, Germanic, Danish	2	100K
IE, Germanic, Dutch	2	306K
IE, Germanic, English	7	620K

Language	Treebanks	Tokens
IE, Germanic, Faroese	1	10K
IE, Germanic, German	4	3,753K
IE, Germanic, Gothic	1	55K
IE, Germanic, Norwegian	3	666K
IE, Germanic, Swedish	3	206K
IE, Germanic, Swiss German	1	1K
IE, Greek, Ancient Greek	2	416K
IE, Greek, Greek	1	63K
IE, Indic, Bhojpuri	1	4K
IE, Indic, Hindi	2	375K
IE, Indic, Marathi	1	3K
IE, Indic, Sanskrit	1	1K
IE, Indic, Urdu	1	138K
IE, Iranian, Kurmanji	1	10K
IE, Iranian, Persian	1	152K
IE, Latin, Latin	3	582K
IE, Romance, Catalan	1	531K
IE, Romance, French	8	1,157K
IE, Romance, Galician	2	164K
IE, Romance, Italian	6	811K
IE, Romance, Old French	1	170K
IE, Romance, Portuguese	3	570K
IE, Romance, Romanian	3	551K
IE, Romance, Spanish	3	1,004K
IE, Slavic, Belarusian	1	13K
IE, Slavic, Bulgarian	1	156K
IE, Slavic, Croatian	1	199K
IE, Slavic, Czech	5	2,222K
IE, Slavic, Old Church Slavonic	1	57K
IE, Slavic, Old Russian	2	168K

Language	Treebanks	Tokens
IE, Slavic, Polish	3	499K
IE, Slavic, Russian	4	1,263K
IE, Slavic, Serbian	1	97K
IE, Slavic, Slovak	1	106K
IE, Slavic, Slovenian	2	170K
IE, Slavic, Ukrainian	1	122K
IE, Slavic, Upper Sorbian	1	11K
Japanese, Japanese	5	1,688K
Korean, Korean	5	446K
Mande, Bambara	1	13K
Mongolic, Buryat	1	10K
Niger-Congo, Defoid, Yoruba	1	2K
Niger-Congo, Northern Atlantic, Wolof	1	44K
Pama-Nyungan, Warlpiri	1	<1K
Sign Language, Swedish Sign Language	1	1K
Sino-Tibetan, Cantonese	1	13K
Sino-Tibetan, Chinese	5	285K
Sino-Tibetan, Classical Chinese	1	74K
Tai-Kadai, Thai	1	22K
Tupian, Mbya Guarani	2	13K
Turkic, Northwestern, Kazakh	1	10K
Turkic, Southeastern, Uyghur	1	40K
Turkic, Southwestern, Turkish	4	91K
Uralic, Finnic, Estonian	2	465K
Uralic, Finnic, Finnish	3	377K
Uralic, Finnic, Karelian	1	3K
Uralic, Finnic, Livvi	1	1K
Uralic, Mordvin, Erzya	1	15K
Uralic, Mordvin, Moksha	1	<1K
Uralic, Permic, Komi Permyak	1	<1K

Language	Treebanks	Tokens
Uralic, Permic, Komi Zyrian	2	3K
Uralic, Sami, North Sami	1	26K
Uralic, Sami, Skolt Sami	1	<1K
Uralic, Ugric, Hungarian	1	42K

Table 21: Languages in the Universal Dependencies treebanks version 2.5, sorted by phylogenetic classification, with corresponding number of treebanks and number of tokens.

A.2 Single-link statistics

	HEAD-punct	HEAD-nsubj	HEAD-case	HEAD-nmod	HEAD-obl	HEAD-obj	HEAD-advmod	HEAD-amod	HEAD-det	HEAD-cc
Afro-Asiatic, Egyptian, Coptic	0.69	0.10	0.00	1.00	0.95	1.00	0.63	0.90	0.09	0.00
Afro-Asiatic, Semitic, Akkadian	–	0.18	0.01	0.98	0.30	0.47	0.14	0.89	0.95	0.00
Afro-Asiatic, Semitic, Amharic	1.00	0.44	0.53	0.72	0.03	0.12	0.05	0.06	0.85	0.85
Afro-Asiatic, Semitic, Arabic	0.78	0.67	0.00	0.97	0.93	0.98	0.64	1.00	0.08	0.05
Afro-Asiatic, Semitic, Assyrian	0.89	0.03	0.00	0.95	1.00	0.84	0.14	1.00	0.00	0.32
Afro-Asiatic, Semitic, Hebrew	0.76	0.29	0.00	0.96	0.82	0.97	0.41	1.00	0.03	0.00
Afro-Asiatic, Semitic, Maltese	0.88	0.25	0.00	0.99	0.72	0.95	0.22	0.78	0.02	0.00
Austro-Asiatic, Viet-Muong, Vietnamese	0.72	0.02	0.02	0.94	0.62	0.96	0.25	0.88	0.48	0.24
Austronesian, Central Philippine, Tagalog	0.96	0.98	0.08	0.67	–	1.00	0.78	–	0.02	–
Austronesian, Malayo-Sumbawan, Indonesian	0.73	0.03	0.03	0.96	0.81	0.95	0.12	0.88	0.56	0.05
Basque, Basque	0.65	0.19	1.00	0.10	0.26	0.15	0.19	0.72	0.62	0.07
Code switching, Hindi English	0.76	0.04	0.71	0.21	0.37	0.42	0.39	0.03	0.00	0.01
Creole, Naija	0.53	0.01	0.01	0.29	0.90	0.99	0.44	0.02	0.09	0.00
Dravidian, South Central, Telugu	0.92	0.01	1.00	0.04	0.01	0.01	0.04	0.01	0.04	0.43
Dravidian, Southern, Tamil	0.64	0.02	1.00	0.01	0.00	0.00	0.36	0.00	0.00	0.56
IE, Armenian, Armenian	0.37	0.15	0.84	0.03	0.34	0.41	0.14	0.01	0.00	0.01
IE, Baltic, Latvian	0.51	0.24	0.02	0.16	0.45	0.68	0.13	0.00	0.01	0.00
IE, Baltic, Lithuanian	0.51	0.28	0.01	0.17	0.60	0.65	0.11	0.03	0.02	0.02
IE, Celtic, Breton	0.81	0.51	0.00	0.99	0.83	0.90	0.60	0.89	0.00	0.03
IE, Celtic, Irish	0.47	0.88	0.00	0.88	0.88	0.61	0.71	0.98	0.19	0.00
IE, Celtic, Scottish Gaelic	0.62	0.98	0.00	0.80	0.94	0.62	0.70	0.85	0.07	0.04
IE, Celtic, Welsh	0.64	0.70	0.00	0.86	0.91	0.78	0.39	0.81	0.04	0.00
IE, Germanic, Afrikaans	0.93	0.02	0.03	0.91	0.53	0.23	0.34	0.29	0.00	0.99
IE, Germanic, Danish	0.79	0.16	0.03	0.77	0.82	0.95	0.36	0.13	0.00	0.07

	HEAD-punct	HEAD-nsubj	HEAD-case	HEAD-nmod	HEAD-obl	HEAD-obj	HEAD-advmod	HEAD-amod	HEAD-det	HEAD-cc
IE, Germanic, Dutch	0.62	0.21	0.02	0.81	0.46	0.41	0.24	0.04	0.00	0.01
IE, Germanic, English	0.71	0.04	0.04	0.63	0.85	0.96	0.24	0.03	0.00	0.00
IE, Germanic, Faroese	0.88	0.13	0.01	0.92	0.80	0.97	0.36	0.01	0.00	0.00
IE, Germanic, German	0.60	0.22	0.01	0.92	0.31	0.38	0.21	0.00	0.00	0.00
IE, Germanic, Gothic	–	0.33	0.00	0.91	0.73	0.62	0.18	0.53	0.21	0.61
IE, Germanic, Norwegian	0.66	0.17	0.01	0.72	0.80	0.93	0.32	0.05	0.02	0.01
IE, Germanic, Swedish	0.78	0.17	0.01	0.56	0.82	0.94	0.34	0.03	0.00	0.02
IE, Germanic, Swiss German	0.87	0.06	0.01	0.77	0.19	0.15	0.16	0.04	0.00	0.00
IE, Greek, Ancient Greek	0.84	0.32	0.03	0.55	0.57	0.47	0.22	0.54	0.12	0.77
IE, Greek, Greek	0.59	0.27	0.00	0.95	0.78	0.84	0.35	0.03	0.02	0.01
IE, Indic, Bhojpuri	0.73	0.03	1.00	0.02	0.04	0.11	0.00	0.04	0.02	0.53
IE, Indic, Hindi	0.80	0.01	0.98	0.02	0.00	0.23	0.12	0.03	0.00	0.00
IE, Indic, Marathi	0.84	0.03	1.00	0.01	0.03	0.03	0.08	0.00	0.00	0.02
IE, Indic, Sanskrit	0.97	0.25	0.57	0.11	0.14	0.10	0.15	0.11	0.05	0.53
IE, Indic, Urdu	0.87	0.01	0.97	0.06	0.01	0.25	0.10	0.01	0.00	0.00
IE, Iranian, Kurmanji	0.71	0.00	0.30	0.55	0.00	0.05	0.02	0.96	0.00	0.11
IE, Iranian, Persian	0.65	0.01	0.13	0.88	0.02	0.02	0.04	0.93	0.00	0.01
IE, Latin, Latin	0.61	0.22	0.01	0.75	0.48	0.44	0.14	0.48	0.21	0.36
IE, Romance, Catalan	0.59	0.17	0.02	0.93	0.71	0.71	0.33	0.78	0.02	0.00
IE, Romance, French	0.67	0.06	0.00	0.98	0.83	0.85	0.34	0.70	0.00	0.00
IE, Romance, Galician	0.68	0.17	0.06	0.99	0.77	0.81	0.23	0.80	0.01	0.80
IE, Romance, Italian	0.77	0.21	0.00	0.97	0.76	0.87	0.23	0.68	0.00	0.01
IE, Romance, Old French	–	0.19	0.00	0.92	–	0.40	0.22	0.31	0.00	0.00
IE, Romance, Portuguese	0.62	0.10	0.01	0.93	0.79	0.91	0.24	0.69	0.01	0.01
IE, Romance, Romanian	0.54	0.27	0.01	0.93	0.76	0.66	0.20	0.88	0.19	0.01
IE, Romance, Spanish	0.59	0.17	0.01	0.97	0.77	0.79	0.28	0.73	0.01	0.01
IE, Slavic, Belarusian	0.45	0.23	0.00	0.99	0.65	0.88	0.12	0.01	0.01	0.00
IE, Slavic, Bulgarian	0.62	0.24	0.00	0.98	0.66	0.81	0.19	0.03	0.42	0.00
IE, Slavic, Croatian	0.57	0.23	0.00	0.97	0.65	0.75	0.10	0.02	0.01	0.01
IE, Slavic, Czech	0.56	0.33	0.00	0.96	0.58	0.69	0.18	0.08	0.02	0.03
IE, Slavic, Old Church Slavonic	–	0.36	0.02	0.93	0.74	0.68	0.20	0.79	0.59	0.67
IE, Slavic, Old Russian	0.39	0.43	0.02	0.75	0.70	0.64	0.20	0.39	0.39	0.43
IE, Slavic, Polish	0.60	0.27	0.00	0.93	0.73	0.82	0.19	0.35	0.06	0.01
IE, Slavic, Russian	0.53	0.25	0.00	0.97	0.60	0.80	0.16	0.03	0.05	0.00
IE, Slavic, Serbian	0.58	0.22	0.00	0.99	0.76	0.86	0.08	0.00	0.00	0.01
IE, Slavic, Slovak	0.69	0.32	0.00	0.85	0.61	0.67	0.17	0.04	0.02	0.01
IE, Slavic, Slovenian	0.55	0.27	0.00	0.95	0.55	0.59	0.23	0.03	0.00	0.00
IE, Slavic, Ukrainian	0.48	0.22	0.00	0.99	0.66	0.74	0.14	0.05	0.03	0.00
IE, Slavic, Upper Sorbian	0.61	0.21	0.00	0.81	0.43	0.62	0.12	0.01	0.01	0.01
Japanese, Japanese	0.96	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.73
Korean, Korean	0.92	0.00	0.99	0.00	0.00	0.01	0.01	0.00	0.01	0.01
Mande, Bambara	0.99	0.00	0.93	0.00	0.94	0.00	0.74	1.00	0.84	0.02

	HEAD-punct	HEAD-nsubj	HEAD-case	HEAD-nmod	HEAD-obl	HEAD-obj	HEAD-advmod	HEAD-amod	HEAD-det	HEAD-cc
Mongolic, Buryat	0.85	0.00	0.96	0.02	–	0.01	0.06	0.02	0.06	0.80
Niger-Congo, Defoid, Yoruba	0.45	0.02	0.22	0.96	0.89	0.91	0.31	0.73	0.32	0.00
Niger-Congo, Northern Atlantic, Wolof	0.46	0.00	0.00	0.99	0.77	0.70	0.51	1.00	0.58	0.00
Pama-Nyungan, Warlpiri	1.00	0.09	1.00	1.00	0.33	0.18	0.00	0.40	–	–
Sign Language, Swedish Sign Language	–	0.19	0.10	0.37	0.77	0.79	0.39	0.12	0.38	0.00
Sino-Tibetan, Cantonese	0.85	0.02	0.65	0.00	0.16	0.97	0.08	0.00	0.00	0.04
Sino-Tibetan, Chinese	0.56	0.00	–	0.01	0.10	1.00	0.03	0.00	0.00	0.01
Sino-Tibetan, Classical Chinese	–	0.00	0.58	0.00	0.48	0.96	0.00	0.01	0.00	0.00
Tai-Kadai, Thai	0.47	0.00	0.01	0.98	0.84	0.92	0.56	0.91	0.59	0.01
Tupian, Mbya Guarani	0.76	0.13	1.00	0.06	0.26	0.47	0.40	1.00	0.00	0.08
Turkic, Northwestern, Kazakh	0.71	0.01	0.98	0.00	0.00	0.00	0.23	0.01	0.00	0.04
Turkic, Southeastern, Uyghur	0.93	0.04	0.96	0.00	0.01	0.00	0.02	0.00	0.00	0.01
Turkic, Southwestern, Turkish	0.75	0.05	0.98	0.02	0.04	0.03	0.30	0.02	0.00	0.06
Uralic, Finnic, Estonian	0.56	0.33	0.79	0.15	0.48	0.60	0.24	0.01	0.02	0.00
Uralic, Finnic, Finnish	0.75	0.16	0.87	0.37	0.60	0.70	0.34	0.01	0.01	0.00
Uralic, Finnic, Karelian	0.59	0.15	0.95	0.03	0.57	0.77	0.40	0.01	0.11	0.00
Uralic, Finnic, Livvi	0.48	0.23	1.00	0.04	0.57	0.59	0.30	0.02	0.00	0.00
Uralic, Mordvin, Erzya	0.55	0.36	0.99	0.07	0.54	0.59	0.16	0.01	0.03	0.01
Uralic, Mordvin, Moksha	0.58	0.05	1.00	0.00	0.53	0.79	0.13	0.00	0.00	0.00
Uralic, Permic, Komi Permyak	0.68	0.14	1.00	0.00	0.40	0.58	0.26	0.00	0.00	0.00
Uralic, Permic, Komi Zyrian	0.57	0.20	0.85	0.18	0.58	0.55	0.24	0.02	0.00	0.02
Uralic, Sami, North Sami	0.92	0.20	0.91	0.14	0.69	0.81	0.41	0.01	0.07	0.88
Uralic, Sami, Skolt Sami	0.64	0.17	0.00	0.00	0.60	0.75	0.23	0.00	0.00	0.00
Uralic, Ugric, Hungarian	0.68	0.24	1.00	0.16	0.16	0.43	0.10	0.01	0.00	0.21

Table 22: Relative frequencies for the 10 most frequent single-link properties, excluding the root relation, for all languages in the Universal Dependencies treebanks.

A.3 Dendrograms

A.3.1 17 languages

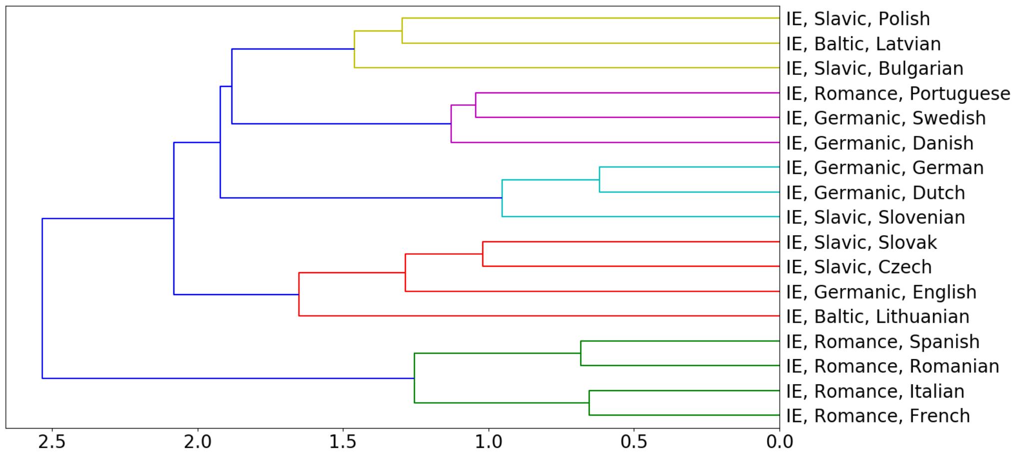


Figure 29: Constructed tree for 17 languages using single links.

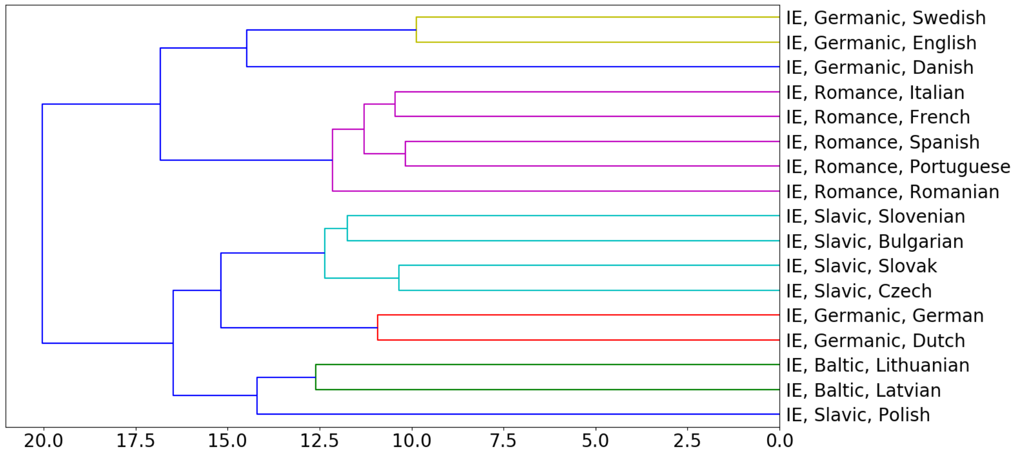


Figure 30: Constructed tree for 17 languages using double links.

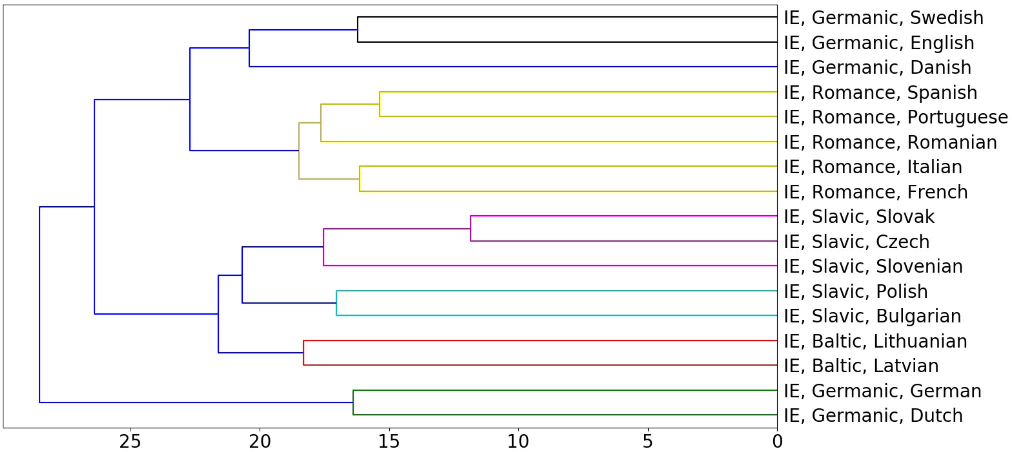


Figure 31: Constructed tree for 17 languages using chain links.

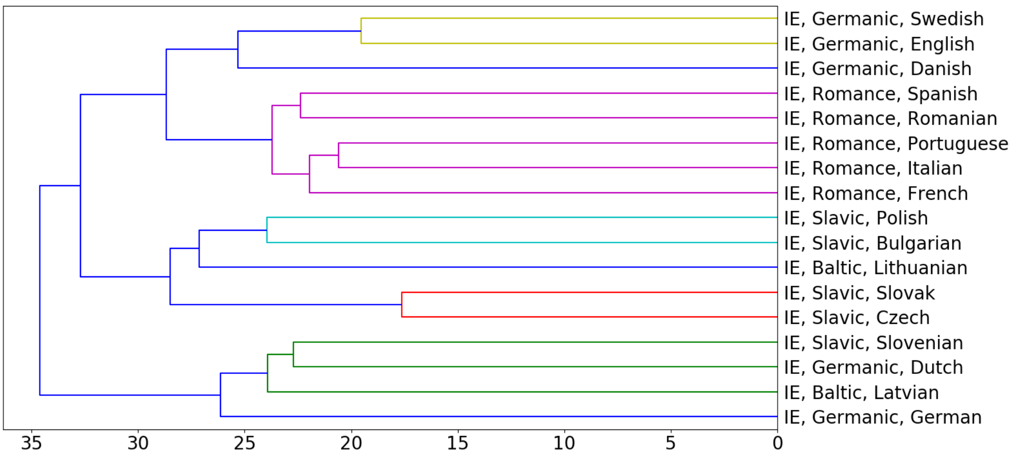


Figure 32: Constructed tree for 17 languages using single links + POS.

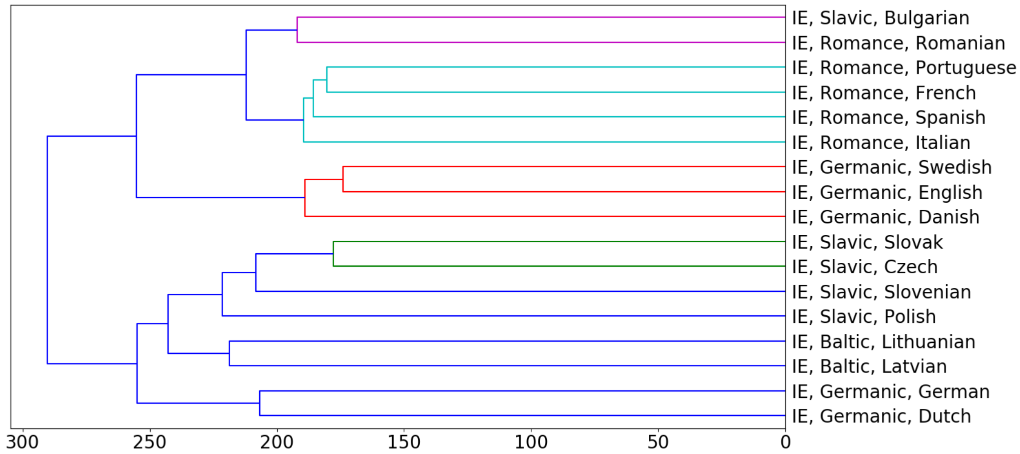


Figure 33: Constructed tree for 17 languages using double links + POS.

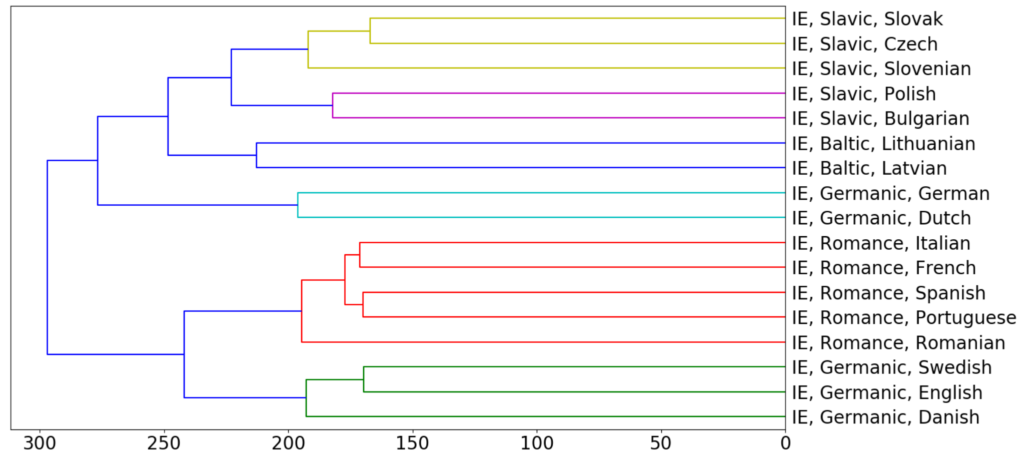


Figure 34: Constructed tree for 17 languages using chain links + POS.

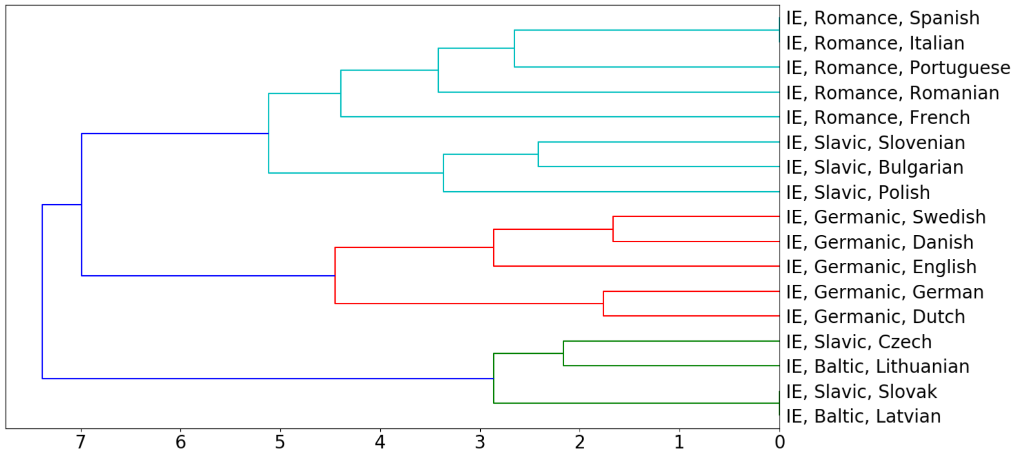


Figure 35: Constructed tree for 17 languages using WALS features.

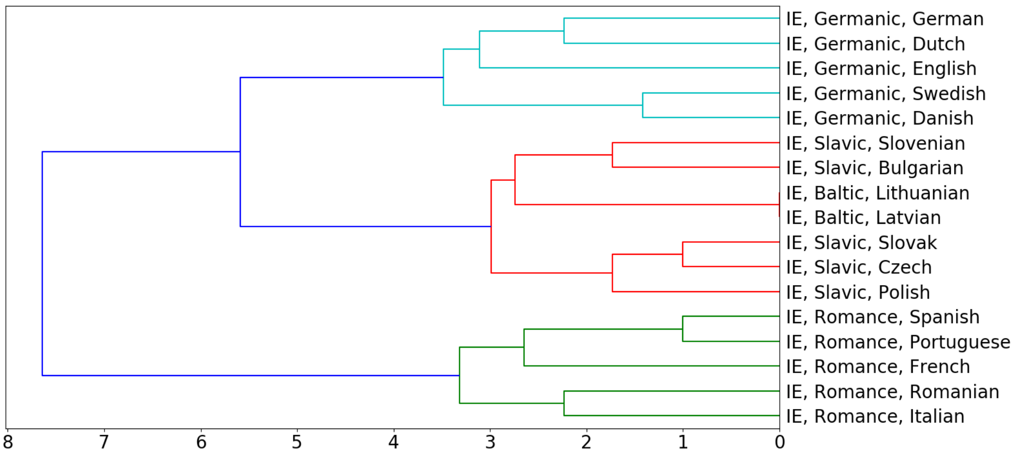


Figure 36: Constructed tree for 17 languages using phylogenetic features.

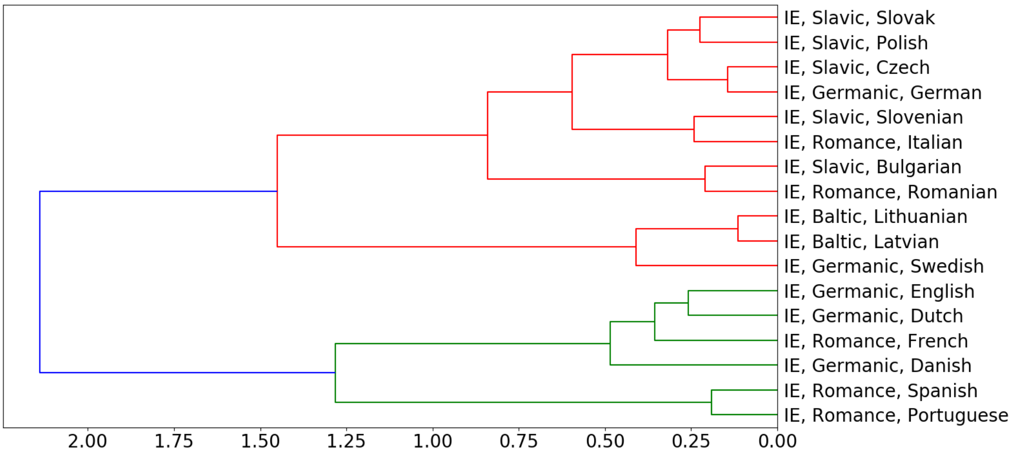


Figure 37: Constructed tree for 17 languages using geographical features.

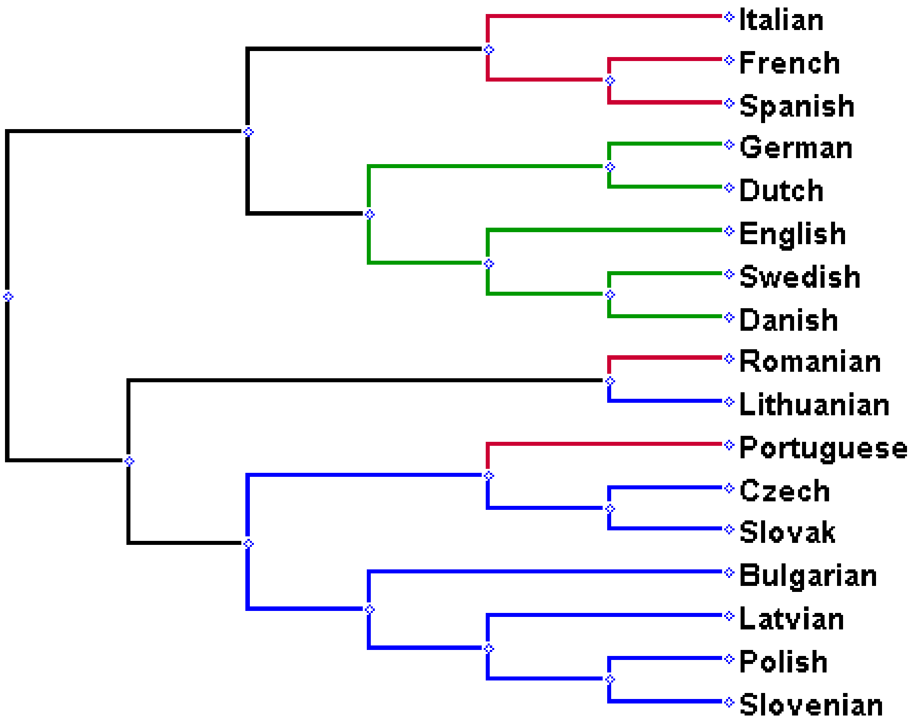


Figure 38: Constructed tree for 17 languages using POS (Rabinovich et al., 2017).

A.3.2 66 languages

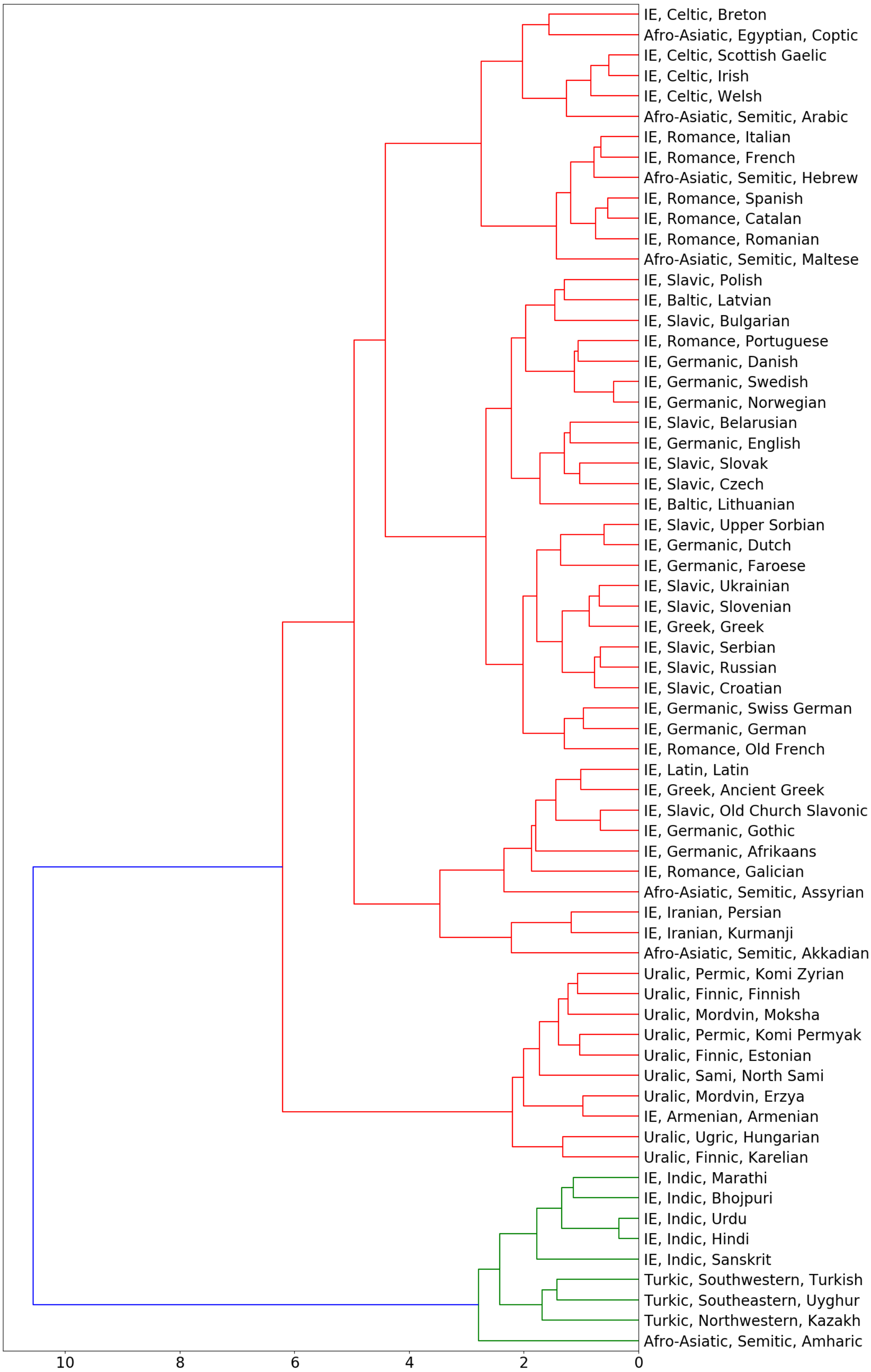


Figure 39: Constructed tree for 66 languages using single links.

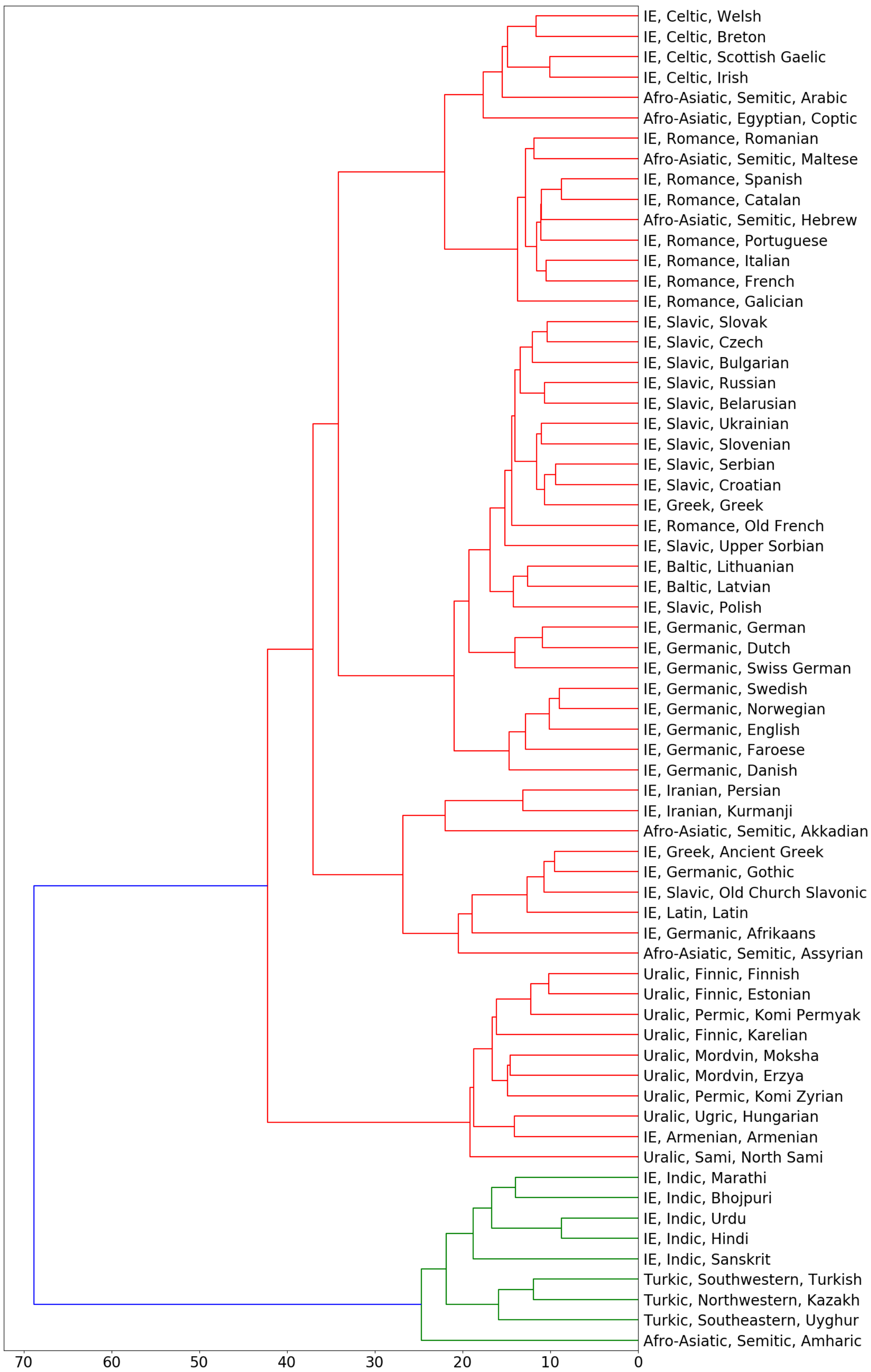


Figure 40: Constructed tree for 66 languages using double links.

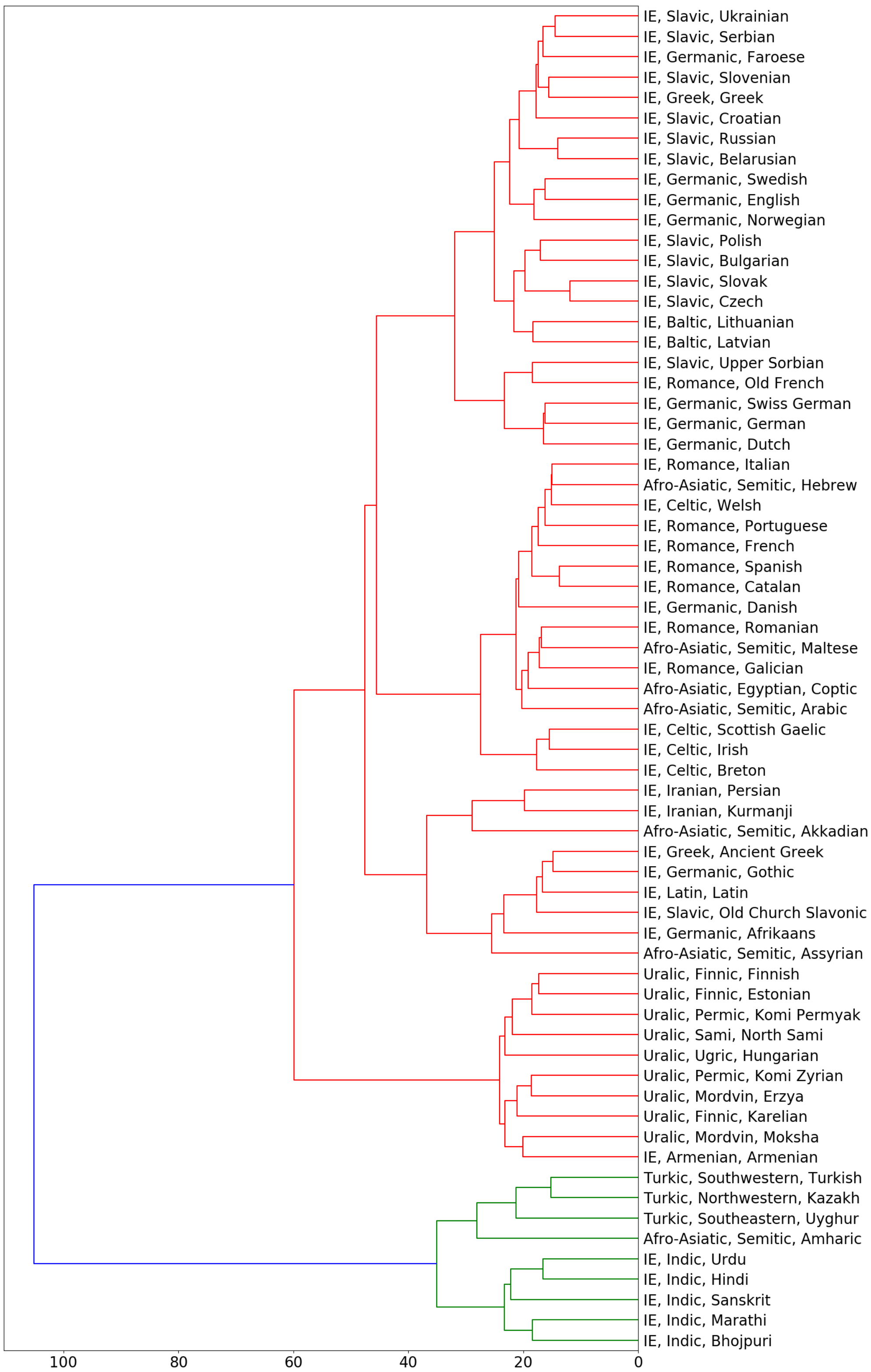


Figure 41: Constructed tree for 66 languages using chain links.

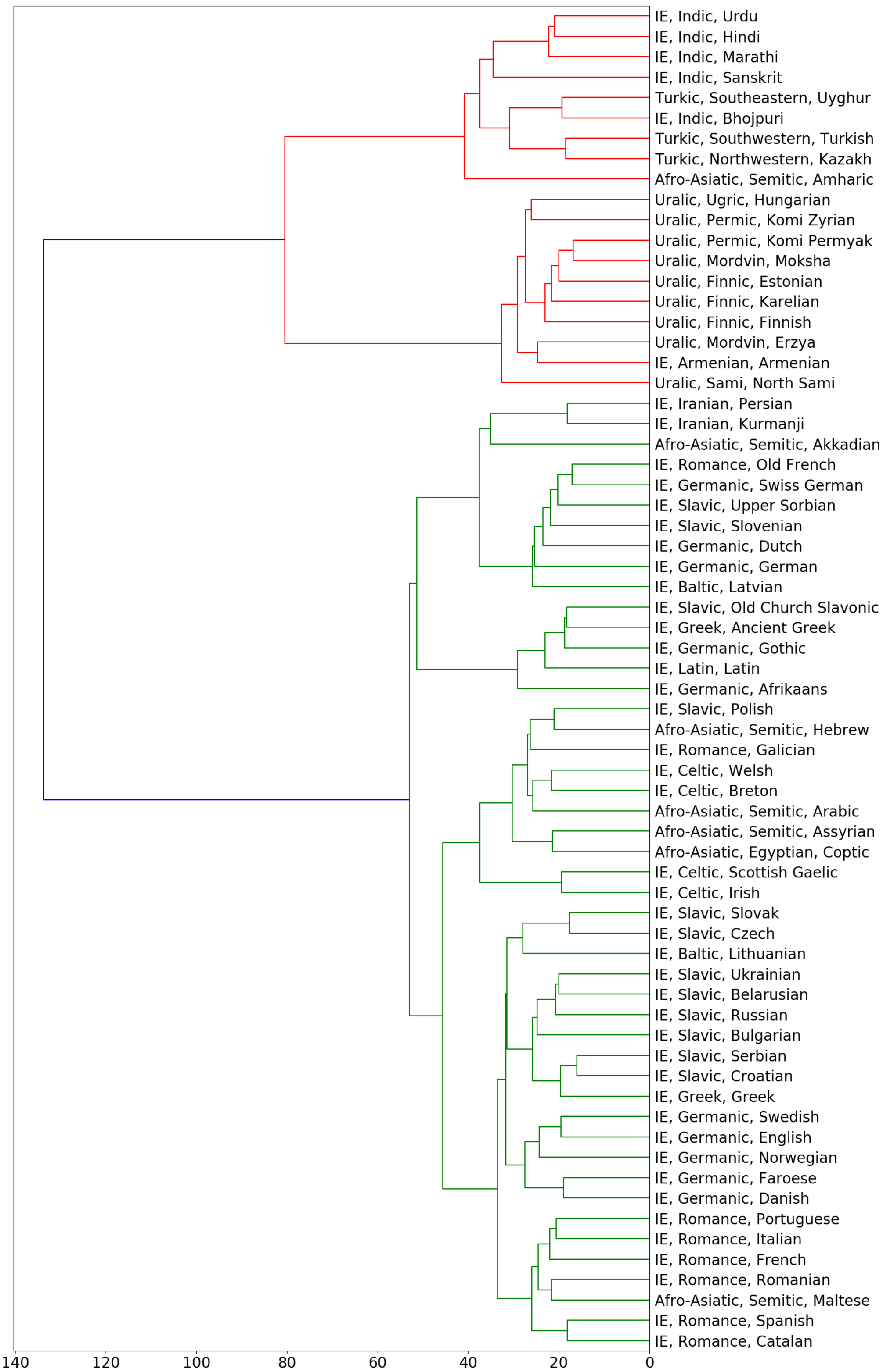


Figure 42: Constructed tree for 66 languages using single links + POS.

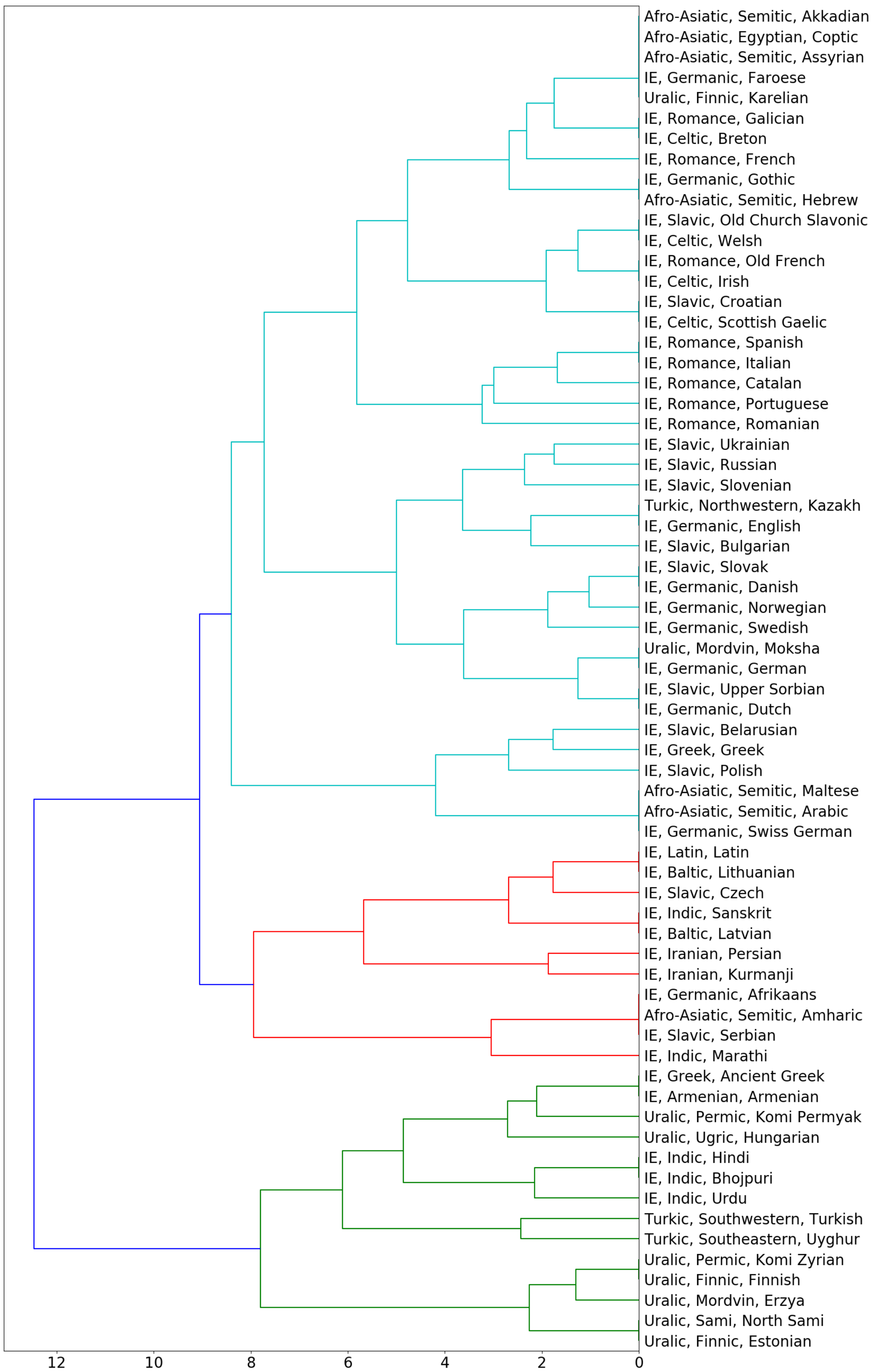


Figure 43: Constructed tree for 66 languages using WALS features.

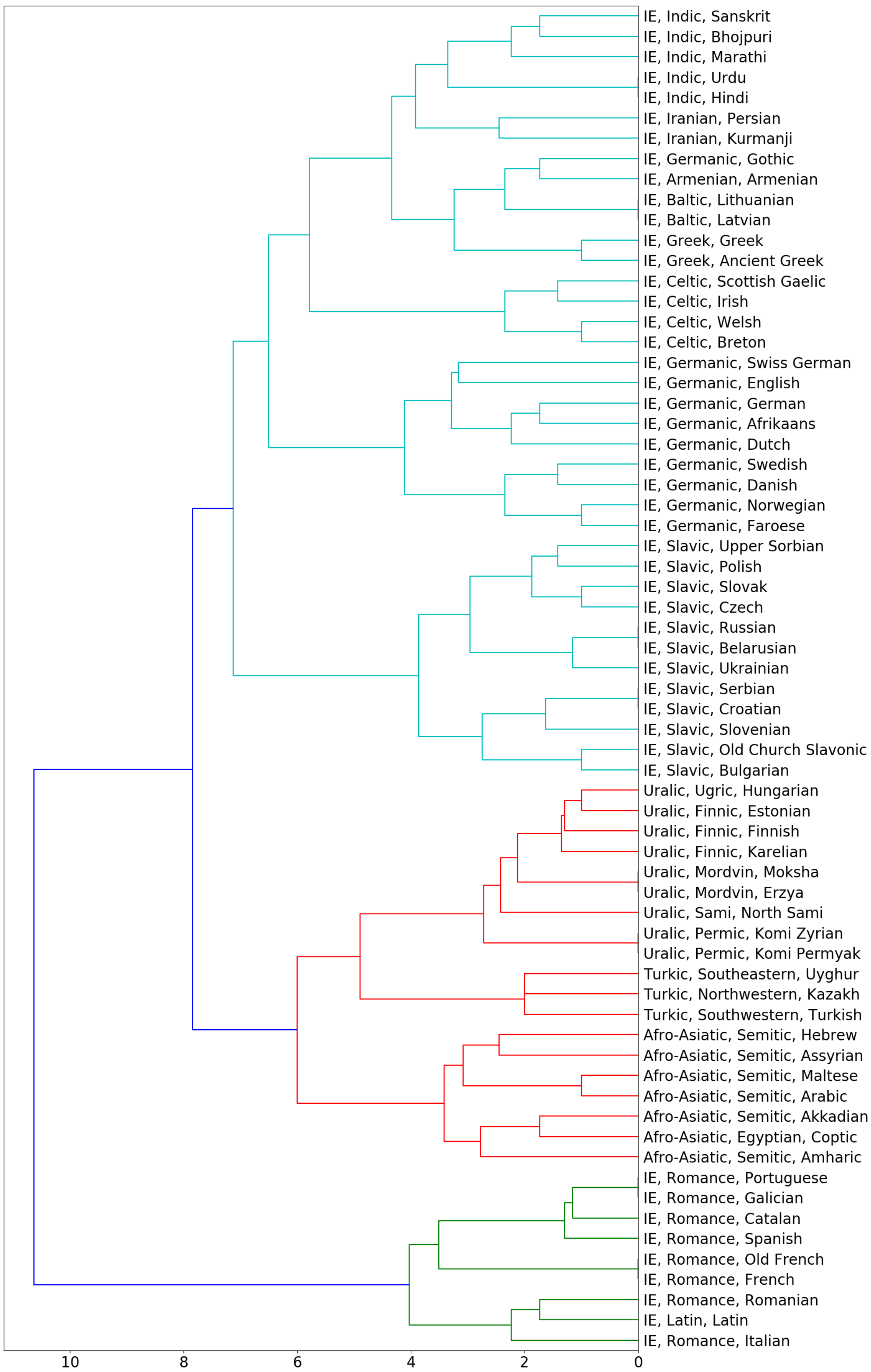


Figure 44: Constructed tree for 66 languages using phylogenetic features.

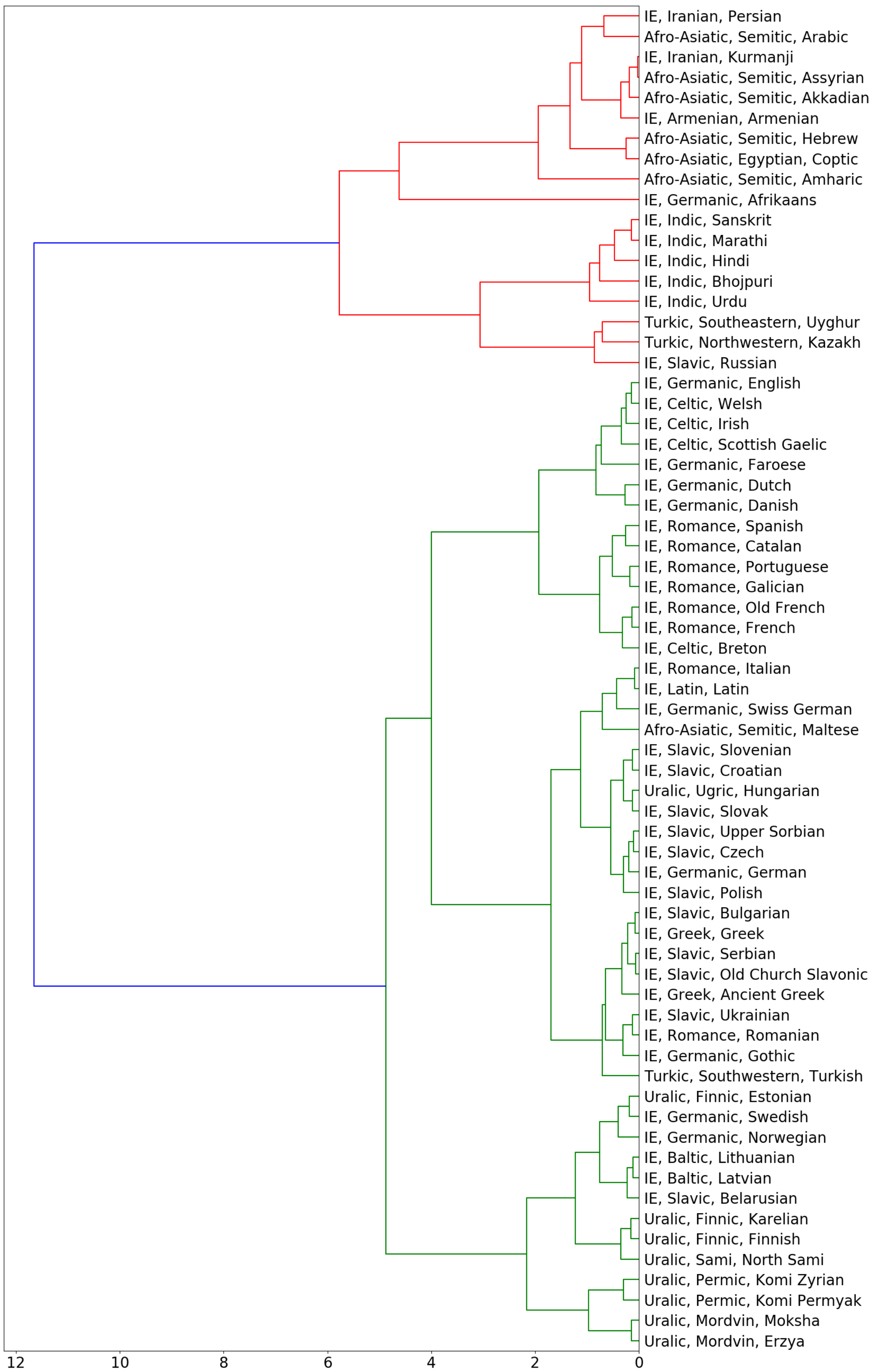


Figure 45: Constructed tree for 66 languages using geographical features.

Appendix B: Property Vectors

B.1 UD vs. WALS

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.41	0.23	$\vec{v}_{\text{OBLIQUE\_AFTER\_OBJECT}}$	$\vec{v}_{\textsc{head}:\text{VERB-obj:NOUN-obl:NOUN}}$ $+\vec{v}_{\text{obj:NOUN-}\textsc{head}\text{:VERB-obl:NOUN}}$ $+\vec{v}_{\text{obj:NOUN-obl:NOUN-}\textsc{head}\text{:VERB}}$

$e^{\ell}_{u,p}$	v^{ℓ}_u	v^{ℓ}_p	ℓ
0.99	1.00	0.01	Sino-Tibetan, Chinese
0.90	1.00	0.10	Sino-Tibetan, Cantonese
0.70	1.00	0.30	Afro-Asiatic, Semitic, Amharic
0.57	1.00	0.43	Uralic, Finnic, Estonian
0.52	1.00	0.48	Austro-Asiatic, Viet-Muong, Vietnamese
0.51	1.00	0.49	Uralic, Permic, Komi Zyrian

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.37	0.21	$\vec{v}_{\text{OBLIQUE\_BEFORE\_OBJECT}}$	$\vec{v}_{\textsc{head}:\text{VERB-obl:NOUN-obj:NOUN}}$ $+\vec{v}_{\text{obl:NOUN-}\textsc{head}\text{:VERB-obj:NOUN}}$ $+\vec{v}_{\text{obl:NOUN-obj:NOUN-}\textsc{head}\text{:VERB}}$

$e^{\ell}_{u,p}$	v^{ℓ}_u	v^{ℓ}_p	ℓ
0.90	0.00	0.90	Sino-Tibetan, Cantonese
0.70	0.00	0.70	Afro-Asiatic, Semitic, Amharic
0.57	0.00	0.57	Uralic, Finnic, Estonian
0.52	0.00	0.52	Austro-Asiatic, Viet-Muong, Vietnamese
0.51	0.00	0.51	Uralic, Permic, Komi Zyrian

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.33	0.22	$\vec{v}_{\text{VOX}}$	$\vec{v}_{\textsc{head}:\text{VERB-obj:NOUN-obl:NOUN}}$

$e^{\ell}_{u,p}$	v^{ℓ}_u	v^{ℓ}_p	ℓ
0.77	1.00	0.23	Uralic, Finnic, Estonian
0.77	1.00	0.23	Uralic, Permic, Komi Zyrian
0.66	1.00	0.34	Uralic, Finnic, Finnish
0.56	1.00	0.44	Austro-Asiatic, Viet-Muong, Vietnamese
0.53	1.00	0.47	IE, Slavic, Bulgarian

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.24	0.19	$\vec{v}_{\text{XVO}}$	$\vec{v}_{\text{obl:NOUN-}\textsc{head}\text{:VERB-obj:NOUN}}$

$e^\ell_{u,p}$	v^ℓ_u	v^ℓ_p	ℓ
0.88	0.00	0.88	Sino-Tibetan, Cantonese

0.23	0.25	$\vec{v}_{\text{SUBJECT\_AFTER\_VERB}}$	$\vec{v}_{\textsc{head}\text{:VERB-nsubj:NOUN}}$

$e^\ell_{u,p}$	v^ℓ_u	v^ℓ_p	ℓ
0.93	1.00	0.07	Pama-Nyungan, Warlpiri
0.76	1.00	0.24	IE, Romance, Spanish
0.74	1.00	0.26	IE, Romance, Catalan
0.72	1.00	0.28	IE, Romance, Italian
0.71	1.00	0.29	IE, Slavic, Belarusian
0.69	1.00	0.31	IE, Slavic, Bulgarian
0.69	1.00	0.31	IE, Slavic, Polish
0.64	1.00	0.36	IE, Greek, Greek
0.61	1.00	0.39	IE, Baltic, Latvian

0.19	0.11	$\vec{v}_{\text{OBLIQUE\_AFTER\_VERB}}$	$\vec{v}_{\textsc{head}\text{:VERB-obl:NOUN}}$
0.19	0.11	$\vec{v}_{\text{OBLIQUE\_BEFORE\_VERB}}$	$\neg\vec{v}_{\textsc{head}\text{:VERB-obl:NOUN}}$
0.17	0.16	$\vec{v}_{\text{SUBJECT\_BEFORE\_VERB}}$	$\neg\vec{v}_{\textsc{head}\text{:VERB-nsubj:NOUN}}$

$e^\ell_{u,p}$	v^ℓ_u	v^ℓ_p	ℓ
0.56	1.00	0.44	IE, Celtic, Breton

0.14	0.17	$\vec{v}_{\text{SVO}}$	$\vec{v}_{\text{nsubj:NOUN-}\textsc{head}\text{:VERB-obj:NOUN}}$

$e^\ell_{u,p}$	v^ℓ_u	v^ℓ_p	ℓ
0.69	1.00	0.31	IE, Germanic, German
0.55	1.00	0.45	IE, Germanic, Dutch
0.55	1.00	0.45	Uralic, Ugric, Hungarian

0.13	0.31	$\vec{v}_{\text{OXV}}$	$\vec{v}_{\text{obj:NOUN-obl:NOUN-}\textsc{head}\text{:VERB}}$

MAE	σ	$\vec{v}_u$	$\vec{v}_p$

$e^\ell_{u,p}$	v^ℓ_u	v^ℓ_p	ℓ
1.00	1.00	0.00	Sino-Tibetan, Chinese
0.96	1.00	0.04	Sino-Tibetan, Cantonese
0.73	1.00	0.27	Afro-Asiatic, Semitic, Amharic

0.12	0.19	$\vec{v}_{\text{OBJECT\_BEFORE\_VERB}}$	$\neg\vec{v}_{\text{HEAD:VERB-obj:NOUN}}$

$e^\ell_{u,p}$	v^ℓ_u	v^ℓ_p	ℓ
0.96	1.00	0.04	IE, Slavic, Belarusian
0.71	1.00	0.29	Uralic, Permic, Komi Permyak
0.51	1.00	0.49	Uralic, Ugric, Hungarian

0.12	0.15	$\vec{v}_{\text{NUMERAL\_BEFORE\_NOUN}}$	$\neg\vec{v}_{\text{HEAD:NOUN-nummod:NUM}}$

$e^\ell_{u,p}$	v^ℓ_u	v^ℓ_p	ℓ
0.62	0.00	0.62	Tai-Kadai, Thai

0.12	0.15	$\vec{v}_{\text{NUMERAL\_AFTER\_NOUN}}$	$\vec{v}_{\text{HEAD:NOUN-nummod:NUM}}$

$e^\ell_{u,p}$	v^ℓ_u	v^ℓ_p	ℓ
0.62	1.00	0.38	Tai-Kadai, Thai

0.11	0.18	$\vec{v}_{\text{OBJECT\_AFTER\_VERB}}$	$\vec{v}_{\text{HEAD:VERB-obj:NOUN}}$

$e^\ell_{u,p}$	v^ℓ_u	v^ℓ_p	ℓ
0.84	1.00	0.16	Pama-Nyungan, Warlpiri
0.62	1.00	0.38	IE, Germanic, German
0.57	1.00	0.43	IE, Armenian, Armenian
0.56	1.00	0.44	IE, Germanic, Dutch

0.09	0.21	$\vec{v}_{\text{RELATIVE\_BEFORE\_NOUN}}$	$\neg\vec{v}_{\text{HEAD:NOUN-acl:VERB}}$

MAE	σ	$\vec{v}_u$	$\vec{v}_p$

$e^{\ell}_{u,p}$	v^{ℓ}_u	v^{ℓ}_p	ℓ
1.00	1.00	0.00	Uralic, Ugric, Hungarian
0.70	1.00	0.30	Dravidian, Southern, Tamil
0.60	0.00	0.60	Uralic, Finnic, Finnish
0.60	0.00	0.60	IE, Baltic, Lithuanian

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.09	0.14	$\vec{v}_{\text{SUBJECT\_AFTER\_OBJECT}}$	$\vec{v}_{\text{HEAD:VERB-obj:NOUN-nsubj:NOUN}}$ $+\vec{v}_{\text{obj:NOUN-HEAD:VERB-nsubj:NOUN}}$ $+\vec{v}_{\text{obj:NOUN-nsubj:NOUN-HEAD:VERB}}$

$e^{\ell}_{u,p}$	v^{ℓ}_u	v^{ℓ}_p	ℓ
0.94	0.00	0.94	Austronesian, Central Philippine, Tagalog

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.09	0.14	$\vec{v}_{\text{SUBJECT\_BEFORE\_OBJECT}}$	$\vec{v}_{\text{HEAD:VERB-nsubj:NOUN-obj:NOUN}}$ $+\vec{v}_{\text{nsubj:NOUN-HEAD:VERB-obj:NOUN}}$ $+\vec{v}_{\text{nsubj:NOUN-obj:NOUN-HEAD:VERB}}$

$e^{\ell}_{u,p}$	v^{ℓ}_u	v^{ℓ}_p	ℓ
0.94	1.00	0.06	Austronesian, Central Philippine, Tagalog

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.09	0.17	$\vec{v}_{\text{SOV}}$	$\vec{v}_{\text{nsubj:NOUN-obj:NOUN-HEAD:VERB}}$

$e^{\ell}_{u,p}$	v^{ℓ}_u	v^{ℓ}_p	ℓ
0.73	1.00	0.27	Uralic, Ugric, Hungarian
0.62	1.00	0.38	IE, Germanic, Dutch
0.61	1.00	0.39	IE, Armenian, Armenian
0.55	1.00	0.45	IE, Germanic, German

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.08	0.18	$\vec{v}_{\text{RELATIVE\_AFTER\_NOUN}}$	$\vec{v}_{\text{HEAD:NOUN-acl:VERB}}$

$e^{\ell}_{u,p}$	v^{ℓ}_u	v^{ℓ}_p	ℓ
0.70	0.00	0.70	Dravidian, Southern, Tamil
0.69	1.00	0.31	IE, Armenian, Armenian
0.60	1.00	0.40	Uralic, Finnic, Finnish
0.60	1.00	0.40	IE, Baltic, Lithuanian

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.07	0.19	$\vec{v}_{\text{VSO}}$	$\vec{v}_{\textsc{head}\text{:VERB-nsubj:NOUN-obj:NOUN}}$

$e^{\ell}_{u,p}$	v^{ℓ}_{u}	v^{ℓ}_{p}	ℓ
1.00	1.00	0.00	Austronesian, Central Philippine, Tagalog
0.97	1.00	0.03	IE, Greek, Greek

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.07	0.20	$\vec{v}_{\text{ADPOSITION\_AFTER\_NOUN}}$	$\vec{v}_{\textsc{head}\text{:NOUN-case:ADP}}$

$e^{\ell}_{u,p}$	v^{ℓ}_{u}	v^{ℓ}_{p}	ℓ
0.93	1.00	0.07	Sino-Tibetan, Chinese
0.87	1.00	0.13	Sino-Tibetan, Cantonese
0.83	1.00	0.17	Afro-Asiatic, Semitic, Amharic

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.07	0.11	$\vec{v}_{\text{ADJECTIVE\_BEFORE\_NOUN}}$	$\neg\vec{v}_{\textsc{head}\text{:NOUN-amod:ADJ}}$

$e^{\ell}_{u,p}$	v^{ℓ}_{u}	v^{ℓ}_{p}	ℓ
0.60	0.00	0.60	Pama-Nyungan, Warlpiri

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.07	0.11	$\vec{v}_{\text{ADJECTIVE\_AFTER\_NOUN}}$	$\vec{v}_{\textsc{head}\text{:NOUN-amod:ADJ}}$

$e^{\ell}_{u,p}$	v^{ℓ}_{u}	v^{ℓ}_{p}	ℓ
0.60	1.00	0.40	Pama-Nyungan, Warlpiri

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.05	0.19	$\vec{v}_{\text{SUBORDINATOR\_WORD\_BEFORE\_CLAUSE}}$	$\neg\vec{v}_{\textsc{head}\text{:VERB-mark:SCONJ}}$

$e^{\ell}_{u,p}$	v^{ℓ}_{u}	v^{ℓ}_{p}	ℓ
0.94	0.00	0.94	Afro-Asiatic, Semitic, Amharic
0.81	0.00	0.81	IE, Indic, Marathi

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.05	0.15	$\vec{v}_{\text{XOV}}$	$\vec{v}_{\text{obl:NOUN-obj:NOUN-}\textsc{head}\text{:VERB}}$

$e^{\ell}_{u,p}$	v^{ℓ}_{u}	v^{ℓ}_{p}	ℓ
0.70	0.00	0.70	Afro-Asiatic, Semitic, Amharic

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.04	0.05	$\vec{v}_{\text{OVS}}$	$\vec{v}_{\text{obj:NOUN-HEAD:VERB-nsubj:NOUN}}$
0.04	0.06	$\vec{v}_{\text{OVX}}$	$\vec{v}_{\text{obj:NOUN-HEAD:VERB-obl:NOUN}}$
0.03	0.15	$\vec{v}_{\text{SUBORDINATOR\_WORD\_AFTER\_CLAUSE}}$	$\vec{v}_{\text{HEAD:VERB-mark:SCONJ}}$

$e^\ell_{u,p}$	v^ℓ_u	v^ℓ_p	ℓ
0.94	1.00	0.06	Afro-Asiatic, Semitic, Amharic

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.03	0.06	$\vec{v}_{\text{ADPOSITION\_BEFORE\_NOUN}}$	$\neg\vec{v}_{\text{HEAD:NOUN-case:ADP}}$
0.03	0.12	$\vec{v}_{\text{VOS}}$	$\vec{v}_{\text{HEAD:VERB-obj:NOUN-nsubj:NOUN}}$

$e^\ell_{u,p}$	v^ℓ_u	v^ℓ_p	ℓ
0.94	0.00	0.94	Austronesian, Central Philippine, Tagalog

MAE	σ	$\vec{v}_u$	$\vec{v}_p$
0.02	0.04	$\vec{v}_{\text{OSV}}$	$\vec{v}_{\text{obj:NOUN-nsubj:NOUN-HEAD:VERB}}$

Table 23: Mean absolute error (MAE) between WALS and UD properties and standard deviation, sorted by MAE; languages with an absolute error greater than or equal to 0.5 are shown within minitables directly beneath the corresponding main row.

B.2 UD vs. UA

e_i $\mathcal{V}_F$	$\mathcal{V}_P$	i	Formula
0.39	0.55	1346	(PP-V $\Leftrightarrow$ Adv-V) $\wedge$ (V-PP $\Leftrightarrow$ V-Adv)

UA	$\mathcal{V}_F(\phi^\ell_i)$	$\mathcal{V}_P(\phi^\ell_i)$	ℓ
1.00	0.19	0.06	Austronesian, Malayo-Sumbawan, Indonesian
1.00	0.25	0.11	IE, Slavic, Serbian
1.00	0.25	0.15	IE, Romance, Romanian
1.00	0.28	0.08	Uralic, Permic, Komi Permyak
1.00	0.30	0.20	IE, Slavic, Polish

e_i			
$\mathcal{V}_F$	$\mathcal{V}_P$	i	Formula

1.00	0.31	0.19	IE, Slavic, Ukrainian
1.00	0.32	0.15	IE, Slavic, Croatian
1.00	0.35	0.18	IE, Baltic, Lithuanian
1.00	0.35	0.19	IE, Romance, Spanish
1.00	0.35	0.20	IE, Slavic, Russian
1.00	0.36	0.24	IE, Slavic, Bulgarian
1.00	0.38	0.25	IE, Slavic, Czech
1.00	0.38	0.19	IE, Germanic, English
1.00	0.40	0.18	Niger-Congo, Defoid, Yoruba
1.00	0.43	0.22	IE, Romance, French
1.00	0.43	0.24	IE, Germanic, Danish
1.00	0.45	0.24	IE, Latin, Latin

0.34 0.34 110 (Rel-N $\Rightarrow$ Adj-N) $\wedge$ (N-Rel $\Rightarrow$ N-Adj)

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.00	0.00	Uralic, Ugric, Hungarian
1.00	0.01	0.01	IE, Germanic, Swedish
1.00	0.03	0.04	IE, Germanic, Danish
1.00	0.03	0.05	IE, Germanic, English
1.00	0.16	0.17	Basque, Basque
1.00	0.19	0.20	IE, Slavic, Russian
1.00	0.34	0.34	IE, Slavic, Polish

0.34 0.34 964 (V-O $\Rightarrow$ Art-N) $\wedge$ (O-V $\Rightarrow$ N-Art)

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.00	0.00	Dravidian, Southern, Tamil
1.00	0.00	0.00	Japanese, Japanese
1.00	0.00	0.00	Korean, Korean
1.00	0.01	0.01	IE, Indic, Marathi
1.00	0.01	0.01	IE, Iranian, Persian
1.00	0.03	0.03	Turkic, Southwestern, Turkish
1.00	0.03	0.03	IE, Indic, Urdu
1.00	0.04	0.04	Dravidian, South Central, Telugu
1.00	0.04	0.04	IE, Indic, Hindi
1.00	0.05	0.06	Mongolic, Buryat

e_i			
$\mathcal{V}_F$	$\mathcal{V}_P$	i	Formula

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
0.00	0.77	0.73	IE, Romance, Romanian
1.00	0.38	0.38	IE, Germanic, German
1.00	0.43	0.44	Tai-Kadai, Thai
1.00	0.43	0.43	IE, Armenian, Armenian
1.00	0.44	0.44	IE, Germanic, Dutch
1.00	0.49	0.49	Austronesian, Malayo-Sumbawan, Indonesian

0.33 0.33 10 Common-Proper $\Rightarrow$ Gen-N

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.00	0.00	IE, Celtic, Welsh
1.00	0.02	0.02	IE, Romance, Italian
1.00	0.03	0.03	Tai-Kadai, Thai
1.00	0.04	0.04	IE, Greek, Greek
1.00	0.04	0.04	Niger-Congo, Defoid, Yoruba
1.00	0.04	0.04	Austronesian, Malayo-Sumbawan, Indonesian
1.00	0.05	0.05	Afro-Asiatic, Semitic, Hebrew
1.00	0.17	0.17	IE, Germanic, Norwegian

0.27 0.28 107 (V-O $\Rightarrow$ (N-Rel $\wedge$ N-Adj $\wedge$ N-Gen)) $\wedge$ (O-V $\Rightarrow$ (Rel-N $\wedge$ Adj-N $\wedge$ Gen-N))

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
0.00	0.97	0.95	Turkic, Southwestern, Turkish
1.00	0.03	0.05	IE, Greek, Greek
1.00	0.04	0.05	IE, Germanic, Norwegian
1.00	0.05	0.05	IE, Slavic, Serbian
1.00	0.16	0.24	Basque, Basque
1.00	0.25	0.21	Uralic, Finnic, Finnish
1.00	0.30	0.29	Dravidian, Southern, Tamil

0.25 0.26 1343 (O-V $\Rightarrow$ PP-V) $\wedge$ (V-O $\Rightarrow$ V-PP)

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.05	0.05	Mande, Bambara
1.00	0.08	0.09	Sino-Tibetan, Chinese
1.00	0.16	0.16	Pama-Nyungan, Warlpiri
1.00	0.20	0.23	Sino-Tibetan, Cantonese

e_i			
$\mathcal{V}_F$	$\mathcal{V}_P$	i	Formula

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.49	0.54	Uralic, Ugric, Hungarian

0.25 0.25 9 (Proper-Common $\Rightarrow$ N-Gen) $\wedge$ (Common-Proper $\Rightarrow$ Gen-N)

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.00	0.00	IE, Celtic, Welsh
1.00	0.02	0.02	IE, Romance, Italian
1.00	0.03	0.03	Tai-Kadai, Thai
1.00	0.04	0.04	IE, Greek, Greek
1.00	0.04	0.04	Austronesian, Malayo-Sumbawan, Indonesian
1.00	0.17	0.17	IE, Germanic, Norwegian

0.19 0.16 112 ((V-O $\wedge$ Adv-V) $\Rightarrow$ Adj-N) $\wedge$ ((O-V $\wedge$ V-Adv) $\Rightarrow$ N-Adj)

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.08	0.14	Austronesian, Malayo-Sumbawan, Indonesian
1.00	0.39	0.40	Afro-Asiatic, Semitic, Hebrew
1.00	0.40	0.48	Niger-Congo, Defoid, Yoruba

0.19 0.15 1153 Rel-N $\Rightarrow$ N-Rel

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.01	0.01	Basque, Basque
1.00	0.02	0.04	Sino-Tibetan, Chinese
1.00	0.04	0.09	Afro-Asiatic, Semitic, Amharic
1.00	0.31	0.52	IE, Armenian, Armenian

0.14 0.12 111 V-Adv $\Rightarrow$ N-Adj

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.47	0.47	Uralic, Finnic, Finnish
1.00	0.48	0.48	IE, Germanic, Swedish

0.14 0.15 1419 (O-V $\Rightarrow$ Pred-Cop) $\wedge$ (V-O $\Rightarrow$ Cop-Pred)

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.38	0.52	IE, Germanic, German
1.00	0.44	0.51	IE, Germanic, Dutch
1.00	0.49	0.54	Uralic, Ugric, Hungarian

$\mathcal{V}_F$ (e_i)	$\mathcal{V}_P$ (e_i)	i	Formula
0.14	0.12	501	(V-S-O $\Rightarrow$ Aux-V) $\wedge$ (S-O-V $\Rightarrow$ V-Aux)

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.01	0.01	Mande, Bambara
1.00	0.05	0.05	IE, Iranian, Persian
1.00	0.07	0.07	IE, Celtic, Welsh
1.00	0.21	0.26	Pama-Nyungan, Warlpiri

0.10	0.08	92	Adp-NP $\Rightarrow$ (((N-Dem $\vee$ N-Num) $\Rightarrow$ N-Adj) $\wedge$ (N-Adj $\Rightarrow$ N-Gen) $\wedge$ (N-Gen $\Rightarrow$ N-Rel))

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.17	0.32	Afro-Asiatic, Semitic, Amharic

0.08	0.10	75	(N-Num $\Rightarrow$ N-Adj) $\wedge$ (Adj-N $\Rightarrow$ Num-N)
0.07	0.05	66	O-V $\Rightarrow$ (Adj-N $\Rightarrow$ Gen-N)

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.38	0.39	IE, Germanic, German
1.00	0.44	0.46	IE, Germanic, Dutch

0.07	0.04	74	Adp-NP $\Rightarrow$ (N-Num $\Rightarrow$ N-Adj)
0.06	0.07	489	S-O-V $\Rightarrow$ NP-Adp

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
0.00	1.00	1.00	Dravidian, South Central, Telugu

0.06	0.05	423	Adp-NP $\Rightarrow$ ($\neg$ S-O-V $\Rightarrow$ N-Rel)

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.07	0.09	Sino-Tibetan, Chinese
1.00	0.40	0.45	IE, Baltic, Lithuanian

0.06	0.06	498	O-V $\Rightarrow$ $\mathrm{VP}_{\mathrm{subordinate}}$-$\mathrm{V}_{\mathrm{main}}$
0.06	0.08	57	Adj-N $\Rightarrow$ (Dem-N $\wedge$ Num-N)

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
0.00	1.00	1.00	Niger-Congo, Northern Atlantic, Wolof
0.00	0.92	0.92	Niger-Congo, Defoid, Yoruba

e_i			
$\mathcal{V}_F$	$\mathcal{V}_P$	i	Formula
0.06	0.05	1334	Adp-NP $\Rightarrow$ ((Rel-N $\Rightarrow$ Gen-N) $\wedge$ (Gen-N $\Rightarrow$ Adj-N) $\wedge$ (Adj-N $\Rightarrow$ Dem-N))

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.17	0.32	Afro-Asiatic, Semitic, Amharic

0.06	0.05	114	N-Dem $\Rightarrow$ N-Adj

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.14	0.21	Afro-Asiatic, Semitic, Amharic

0.05	0.04	60	Adj-Adv $\Rightarrow$ (N-Adj $\wedge$ V-O)
0.05	0.03	502	V-O$_{\text{pronominal}}$ $\Rightarrow$ V-O$_{\text{nominal}}$
0.05	0.06	88	(N-Dem $\Rightarrow$ N-Rel) $\wedge$ (Rel-N $\Rightarrow$ Dem-N)

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.14	0.03	Afro-Asiatic, Semitic, Amharic
1.00	0.21	0.04	Basque, Basque

0.04	0.03	65	S-O-V $\Rightarrow$ (Adj-N $\Rightarrow$ Gen-N)
0.04	0.05	126	(N-Num $\Rightarrow$ N-Rel) $\wedge$ (Rel-N $\Rightarrow$ Num-N)
0.04	0.04	1	Adp-NP $\Rightarrow$ N-Gen
0.04	0.04	73	(N-Dem $\Rightarrow$ N-Adj) $\wedge$ (Adj-N $\Rightarrow$ Dem-N)

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.14	0.04	Afro-Asiatic, Semitic, Amharic

0.03	0.01	174	Adp-NP $\Rightarrow$ (N-Num $\Rightarrow$ N-Gen)
0.03	0.02	72	Adp-NP $\Rightarrow$ (N-Dem $\Rightarrow$ N-Adj)

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.17	0.34	Afro-Asiatic, Semitic, Amharic

0.03	0.02	173	Adp-NP $\Rightarrow$ (N-Dem $\Rightarrow$ N-Gen)

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.17	0.32	Afro-Asiatic, Semitic, Amharic

0.03	0.04	176	(N-Gen $\Rightarrow$ N-Rel) $\wedge$ (Rel-N $\Rightarrow$ Gen-N)
0.02	0.02	2	NP-Adp $\Rightarrow$ Gen-N

e_i			
$\mathcal{V}_F$	$\mathcal{V}_P$	i	Formula
0.02	0.02	87	Adp-NP ⇒ (N-Dem ⇒ N-Rel)

UA	$\mathcal{V}_F(\phi_i^\ell)$	$\mathcal{V}_P(\phi_i^\ell)$	ℓ
1.00	0.17	0.32	Afro-Asiatic, Semitic, Amharic

$\mathcal{V}_F$	$\mathcal{V}_P$	i	Formula
0.02	0.02	56	V-S-O ⇒ N-Adj
0.02	0.01	175	Adp-NP ⇒ (N-Gen ⇒ N-Rel)
0.02	0.01	69	Adp-NP ⇒ (N-Adj ⇒ N-Gen)
0.02	0.00	121	Adp-NP ⇒ (N-Num ⇒ N-Rel)
0.02	0.01	95	NP-Adp ⇒ ((((Adj-N ∨ Rel-N) ⇒ (Dem-N ∧ Num-N)) ∧ (Dem-N ∨ Num-N)) ⇒ Gen-N)
0.02	0.01	177	NP-Adp ⇒ (Dem-N ⇒ Gen-N)
0.02	0.01	71	NP-Adp ⇒ (Adj-N ⇒ Gen-N)
0.02	0.01	178	NP-Adp ⇒ (Num-N ⇒ Gen-N)
0.02	0.01	70	(Adp-NP ∧ ¬ S-V-O) ⇒ (N-Adj ⇒ N-Gen)
0.01	0.00	91	Adp-NP ⇒ (N-Adj ⇒ N-Rel)
0.01	0.01	1114	V-O ⇒ Comp-Sent
0.01	0.00	68	V-initial ⇒ (N-Adj ⇒ N-Gen)
0.01	0.00	67	V-S-O ⇒ (N-Adj ⇒ N-Gen)
0.01	0.00	5	(S-O-V ∧ N-Gen) ⇒ N-Adj
0.00	0.00	62	Rel-N ⇒ (NP-Adp ∨ Adj-N)
0.00	0.00	55	V-S-O ⇒ Adp-NP

Table 24: Absolute differences between UA score and UD score for individual universals, sorted by the difference when using fuzzy logic; languages with an absolute difference between the expected score (according to the UA) and the calculated score (fuzzy) greater than or equal to 0.5 are shown within minitables directly beneath the corresponding main row.

B.3 Crawled universals

No.	Formula	Pro	Con	Score	Reference
1	Adp NP ⇒ N G	30	1	0.97	Greenberg
	$\neg v^{\ell}_{\text{HEAD:NOUN-case:ADP}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}}$	14	0	1.00	(1963) #2
	Basque, Finnish, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Norwegian, Serbian, Thai, Turkish, Welsh, Yoruba; –				
2	Postposition ⇒ G N	30	0	1.00	Greenberg
	$v^{\ell}_{\text{HEAD:NOUN-case:ADP}} \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}}$	14	0	1.00	(1963) #2
	Basque, Finnish, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Norwegian, Serbian, Thai, Turkish, Welsh, Yoruba; –				
5	SOV & N Gen ⇒ N Adj	30	1	0.97	Greenberg
	$(v^{\ell}_{\text{nsubj:NOUN-obj:NOUN-HEAD:VERB}} \wedge v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}}) \Rightarrow v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}}$	14	0	1.00	(1963) #5
	Basque, Finnish, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Norwegian, Serbian, Thai, Turkish, Welsh, Yoruba; –				
9	Proper Common ⇒ N G, Common Proper ⇒ G N	16	14	0.53	Greenberg
	$(\neg v^{\ell}_{\text{HEAD:NOUN-flat:PROPN}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}}) \wedge (v^{\ell}_{\text{HEAD:NOUN-flat:PROPN}} \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}})$	10	13	0.43	(1963) #23
	Basque, Finnish, Greek, Indonesian, Italian, Japanese, Norwegian, Thai, Turkish, Welsh; Belarusian, Bulgarian, Croatian, Czech, Old Church Slavonic, Old Russian, Polish, Russian, Serbian, Slovak, Slovenian, Ukrainian, Upper Sorbian				
10	Common Proper ⇒ G N	29	13	0.69	Greenberg
	$v^{\ell}_{\text{HEAD:NOUN-flat:PROPN}} \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}}$	13	13	0.50	(1963) #23
	Basque, Finnish, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Norwegian, Thai, Turkish, Welsh, Yoruba; Belarusian, Bulgarian, Croatian, Czech, Old Church Slavonic, Old Russian, Polish, Russian, Serbian, Slovak, Slovenian, Ukrainian, Upper Sorbian				
55	VSO ⇒ prepositions	623	10	0.98	Greenberg
	$v^{\ell}_{\text{HEAD:VERB-nsubj:NOUN-obj:NOUN}} \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-case:ADP}}$	51	0	1.00	(1963) #3,
	Amharic, Arabic, Armenian, Bambara, Basque, Breton, Bulgarian, Buryat, Cantonese, Chinese, Croatian, Czech, Danish, Dutch, English, Finnish, French,				Dryer (1992)

	German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Komi Permyak, Korean, Latin, Lithuanian, Marathi, Norwegian, Persian, Polish, Romanian, Russian, Scottish Gaelic, Serbian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Vietnamese, Warlpiri, Welsh, Yoruba; –				
56	VSO ⇒ N Adj	30	20	0.60	Greenberg
	$v^{\ell}_{\text{HEAD:VERB-nsubj:NOUN-obj:NOUN}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}}$	14	0	1.00	(1963) #17
	Basque, Finnish, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Norwegian, Serbian, Thai, Turkish, Welsh, Yoruba; –				
57	Adj N ⇒ Dem N & Num N	29	39	0.43	Greenberg
	$\neg v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}} \Rightarrow (\neg v^{\ell}_{\text{HEAD:NOUN-det:DET}} \wedge \neg v^{\ell}_{\text{HEAD:NOUN-nummod:NUM}})$	13	2	0.87	(1963) #18
	Basque, Finnish, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Norwegian, Serbian, Thai, Turkish, Welsh; Wolof, Yoruba				
60	Adj Adv ⇒ N Adj & VO	30	53	0.36	Greenberg
	$v^{\ell}_{\text{HEAD:ADJ-advmod:ADV}} \Rightarrow (v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}} \wedge v^{\ell}_{\text{HEAD:VERB-obj:NOUN}})$	14	0	1.00	(1963) #21
	Basque, Finnish, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Norwegian, Serbian, Thai, Turkish, Welsh, Yoruba; –				
62	Rel N ⇒ Postposition &/v Adj N	30	0	1.00	Greenberg
	$\neg v^{\ell}_{\text{HEAD:NOUN-acl:VERB}} \Rightarrow (v^{\ell}_{\text{HEAD:NOUN-case:ADP}} \vee \neg v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}})$	14	0	1.00	(1963) #24
	Basque, Finnish, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Norwegian, Serbian, Thai, Turkish, Welsh, Yoruba; –				
65	SOV ⇒ (A N ⇒ G N)	328	1	1.00	Hawkins
	$v^{\ell}_{\text{nsubj:NOUN-obj:NOUN-HEAD:VERB}} \Rightarrow (\neg v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}} \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}})$	47	0	1.00	(1983)
	Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –				
66	OV ⇒ (A N ⇒ GN)	328	1	1.00	Hawkins (1983)

	$\neg v^{\ell}_{\text{HEAD:VERB-obj:NOUN}} \Rightarrow (\neg v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}} \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}})$ Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –	47	0	1.00	
67	VSO $\Rightarrow$ (N A $\Rightarrow$ N G) $v^{\ell}_{\text{HEAD:VERB-nsubj:NOUN-obj:NOUN}} \Rightarrow (v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}})$ Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –	328 47	2 0	0.99 1.00	Hawkins (1983)
68	#V $\Rightarrow$ (N A $\Rightarrow$ N G) $(v^{\ell}_{\text{HEAD:VERB-nsubj:NOUN-obj:NOUN}} + v^{\ell}_{\text{HEAD:VERB-obj:NOUN-nsubj:NOUN}}) \Rightarrow (v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}})$ Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –	327 47	5 0	0.98 1.00	Hawkins (1983)
69	Prep $\Rightarrow$ (N A $\Rightarrow$ N G) $\neg v^{\ell}_{\text{HEAD:NOUN-case:ADP}} \Rightarrow (v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}})$ Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew,	325 47	23 0	0.93 1.00	Hawkins (1983)

	Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –				
70	Prep & ¬SVO ⇒ (N A ⇒ N G) $(\neg v^{\ell}_{\text{HEAD:NOUN-case:ADP}} \wedge \neg v^{\ell}_{\text{nsubj:NOUN-HEAD:VERB-obj:NOUN}}) \Rightarrow (v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}})$ Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –	328 47	3 0	0.99 1.00	Hawkins (1983)
71	Postp ⇒ (A N ⇒ G N) $v^{\ell}_{\text{HEAD:NOUN-case:ADP}} \Rightarrow (\neg v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}} \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}})$ Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –	328 47	1 0	1.00 1.00	Hawkins (1983)
72	Prep ⇒ (N Dem ⇒ N A) $\neg v^{\ell}_{\text{HEAD:NOUN-case:ADP}} \Rightarrow (v^{\ell}_{\text{HEAD:NOUN-det:DET}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}})$ Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –	326 47	13 0	0.96 1.00	Hawkins (1983)
73	N Dem ⇒ N A; equivalently: A N ⇒ Dem N	326	20	0.94	Hawkins (1983)

	$(v^{\ell}_{\text{HEAD:NOUN-det:DET}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}}) \wedge (\neg v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}} \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-det:DET}})$	47	0	1.00	
	Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –				
74	Prep $\Rightarrow$ (N Num $\Rightarrow$ N A)	326	9	0.97	Hawkins
	$\neg v^{\ell}_{\text{HEAD:NOUN-case:ADP}} \Rightarrow (v^{\ell}_{\text{HEAD:NOUN-nummod:NUM}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}})$	47	0	1.00	(1983)
	Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –				
75	N Num $\Rightarrow$ N A; equivalently: A N $\Rightarrow$ Num N	328	0	1.00	Hawkins
	$(v^{\ell}_{\text{HEAD:NOUN-nummod:NUM}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}}) \wedge (\neg v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}} \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-nummod:NUM}})$	47	0	1.00	(1983)
	Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –				
87	Prep $\Rightarrow$ (N Dem $\Rightarrow$ N Rel)	328	0	1.00	Hawkins
	$\neg v^{\ell}_{\text{HEAD:NOUN-case:ADP}} \Rightarrow (v^{\ell}_{\text{HEAD:NOUN-det:DET}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-acl:VERB}})$	47	0	1.00	(1983)
	Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian,				

	Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –				
88	N Dem ⇒ N Rel; equivalently: Rel N ⇒ Dem N $(v^{\ell}_{\text{HEAD:NOUN-det:DET}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-acl:VERB}}) \wedge (\neg v^{\ell}_{\text{HEAD:NOUN-acl:VERB}} \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-det:DET}})$ Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –	328 47	14 0	0.96 1.00	Hawkins (1983)
91	Prep ⇒ (N A ⇒ N Rel) $\neg v^{\ell}_{\text{HEAD:NOUN-case:ADP}} \Rightarrow (v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-acl:VERB}})$ Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –	328 47	0 0	1.00 1.00	Hawkins (1983)
92	Prep ⇒ ((N Dem v N Num ⇒ N A) & (N A ⇒ N G) & (N G ⇒ N Rel)) $\neg v^{\ell}_{\text{HEAD:NOUN-case:ADP}} \Rightarrow (((v^{\ell}_{\text{HEAD:NOUN-det:DET}} \vee v^{\ell}_{\text{HEAD:NOUN-nummod:NUM}}) \Rightarrow v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}}) \wedge (v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}}) \wedge (v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-acl:VERB}}))$ Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –	328 47	0 0	1.00 1.00	Hawkins (1983)

95	Postp ⇒ ((A N v Rel N ⇒ Dem N & Num N) & (Dem N v Num N) ⇒ G N))	328	0	1.00	Hawkins (1983)
	$v^{\ell}_{\text{HEAD:NOUN-case:ADP}} \Rightarrow ((((\neg v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}} \vee \neg v^{\ell}_{\text{HEAD:NOUN-acl:VERB}}) \Rightarrow (\neg v^{\ell}_{\text{HEAD:NOUN-det:DET}} \wedge \neg v^{\ell}_{\text{HEAD:NOUN-nummod:NUM}})) \wedge (\neg v^{\ell}_{\text{HEAD:NOUN-det:DET}} \vee \neg v^{\ell}_{\text{HEAD:NOUN-nummod:NUM}})) \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}})$	47	0	1.00	
	Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –				
107	VO ⇒ N Rel & N Adj & N Gen; OV ⇒ Rel N & Adj N & Gen N	32	15	0.68	Greenberg (1963),
	$(v^{\ell}_{\text{HEAD:VERB-obj:NOUN}} \Rightarrow (v^{\ell}_{\text{HEAD:NOUN-acl:VERB}} \wedge v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}} \wedge v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}})) \wedge (\neg v^{\ell}_{\text{HEAD:VERB-obj:NOUN}} \Rightarrow (\neg v^{\ell}_{\text{HEAD:NOUN-acl:VERB}} \wedge \neg v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}} \wedge \neg v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}}))$	15	4	0.79	Lehmann (1973)
	Basque, Finnish, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Norwegian, Portuguese, Serbian, Tamil, Thai, Welsh, Yoruba; Chinese, Persian, Tagalog, Turkish				
110	Rel N ⇒ Adj N; N Rel ⇒ N Adj	282	0	1.00	Dryer
	$(\neg v^{\ell}_{\text{HEAD:NOUN-acl:VERB}} \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}}) \wedge (v^{\ell}_{\text{HEAD:NOUN-acl:VERB}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}})$	26	0	1.00	(1988)
	Amharic, Arabic, Bambara, Basque, Buryat, Cantonese, Chinese, Danish, English, Finnish, Hebrew, Hindi, Hungarian, Indonesian, Korean, Lithuanian, Persian, Polish, Russian, Swedish, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –				
111	V Adv ⇒ N Adj	282	0	1.00	Dryer
	$v^{\ell}_{\text{HEAD:VERB-advmod:ADV}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}}$	26	0	1.00	(1988)
	Amharic, Arabic, Bambara, Basque, Buryat, Cantonese, Chinese, Danish, English, Finnish, Hebrew, Hindi, Hungarian, Indonesian, Korean, Lithuanian, Persian, Polish, Russian, Swedish, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –				
112	(V O & Adv V) ⇒ Adj N; (O V & V Adv) ⇒ N Adj	282	1	1.00	Dryer (1988)

	$((v^{\ell}_{\text{HEAD:VERB-obj:NOUN}} \wedge \neg v^{\ell}_{\text{HEAD:VERB-advmod:ADV}}) \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}}) \wedge ((\neg v^{\ell}_{\text{HEAD:VERB-obj:NOUN}} \wedge v^{\ell}_{\text{HEAD:VERB-advmod:ADV}}) \Rightarrow v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}})$ Amharic, Arabic, Bambara, Basque, Buryat, Cantonese, Chinese, Danish, English, Finnish, Hebrew, Hindi, Hungarian, Indonesian, Korean, Lithuanian, Persian, Polish, Russian, Swedish, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –	26	0	1.00	
114	N Dem ⇒ N Adj $v^{\ell}_{\text{HEAD:NOUN-det:DET}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}}$ Amharic, Arabic, Bambara, Basque, Buryat, Cantonese, Chinese, Danish, English, Finnish, Hebrew, Hindi, Hungarian, Indonesian, Korean, Lithuanian, Persian, Polish, Russian, Swedish, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –	274 26	21 0	0.93 1.00	Dryer (1988)
121	Prep ⇒ (N Num ⇒ N Rel) $\neg v^{\ell}_{\text{HEAD:NOUN-case:ADP}} \Rightarrow (v^{\ell}_{\text{HEAD:NOUN-nummod:NUM}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-acl:VERB}})$ Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –	328 47	0 0	1.00 1.00	Hawkins (1983)
126	N Num ⇒ N Rel; equivalently: Rel N ⇒ Num N $(v^{\ell}_{\text{HEAD:NOUN-nummod:NUM}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-acl:VERB}}) \wedge (\neg v^{\ell}_{\text{HEAD:NOUN-acl:VERB}} \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-nummod:NUM}})$ Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –	326 47	2 0	0.99 1.00	Hawkins (1983)
173	Prep ⇒ (N Dem ⇒ N G) $\neg v^{\ell}_{\text{HEAD:NOUN-case:ADP}} \Rightarrow (v^{\ell}_{\text{HEAD:NOUN-det:DET}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}})$	327 47	2 0	0.99 1.00	Hawkins (1983)

	Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –				
174	Prep ⇒ (N Num ⇒ N G) $\neg v^{\ell}_{\text{HEAD:NOUN-case:ADP}} \Rightarrow (v^{\ell}_{\text{HEAD:NOUN-nummod:NUM}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}})$	327 47	2 0	0.99 1.00	Hawkins (1983)
	Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –				
175	Prep ⇒ (N G ⇒ N Rel) $\neg v^{\ell}_{\text{HEAD:NOUN-case:ADP}} \Rightarrow (v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-acl:VERB}})$	328 47	0 0	1.00 1.00	Hawkins (1983)
	Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –				
176	N G ⇒ N Rel; equivalently: Rel N ⇒ G N $(v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-acl:VERB}}) \wedge (\neg v^{\ell}_{\text{HEAD:NOUN-acl:VERB}} \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}})$	328 47	0 0	1.00 1.00	Hawkins (1983)
	Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –				

177	Postp ⇒ (Dem N ⇒ G N) $v^{\ell}_{\textsc{head}:\text{NOUN-case:ADP}} \Rightarrow (\neg v^{\ell}_{\textsc{head}:\text{NOUN-det:DET}} \Rightarrow \neg v^{\ell}_{\textsc{head}:\text{NOUN-nmod:NOUN}})$ Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –	328 47	0 0	1.00 1.00	Hawkins (1983)
178	Postp ⇒ (Num N ⇒ G N) $v^{\ell}_{\textsc{head}:\text{NOUN-case:ADP}} \Rightarrow (\neg v^{\ell}_{\textsc{head}:\text{NOUN-nummod:NUM}} \Rightarrow \neg v^{\ell}_{\textsc{head}:\text{NOUN-nmod:NOUN}})$ Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –	328 47	0 0	1.00 1.00	Hawkins (1983)
345	Nominal_predicate Subject ⇒ Verbal_predicate Subject $v^{\ell}_{\textsc{head}:\text{NOUN-nsubj:NOUN}} \Rightarrow v^{\ell}_{\textsc{head}:\text{VERB-nsubj:NOUN}}$	– –	– –	– –	*Kozinsky (1981)*
423	Prep ⇒ (¬ SOV ⇒ N Rel) $\neg v^{\ell}_{\textsc{head}:\text{NOUN-case:ADP}} \Rightarrow (\neg v^{\ell}_{\text{nsubj:NOUN-obj:NOUN-}\textsc{head}:\text{VERB}} \Rightarrow v^{\ell}_{\textsc{head}:\text{NOUN-acl:VERB}})$ Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –	328 47	0 0	1.00 1.00	Hawkins (1983)
449	Gen N (nominally possessed construction) ⇒ Gen Pro (pronominal construction) $\neg v^{\ell}_{\textsc{head}:\text{NOUN-nmod:NOUN}} \Rightarrow \neg v^{\ell}_{\textsc{head}:\text{PRON-nmod:NOUN}}$	– –	– –	– –	*Ultan (1978)*

489	SOV ⇒ postposition	30	2	0.94	Greenberg
	$v^{\ell}_{\text{nsubj:NOUN-obj:NOUN-HEAD:VERB}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-case:ADP}}$	14	1	0.93	(1963) #4
	Basque, Finnish, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Norwegian, Serbian, Thai, Turkish, Welsh, Yoruba; Telugu				
498	OV ⇒ Subordinate-Main verb	30	0	1.00	Greenberg
	$\neg v^{\ell}_{\text{HEAD:VERB-obj:NOUN}} \Rightarrow \neg v^{\ell}_{\text{HEAD:VERB-xcomp:VERB}}$	14	0	1.00	(1963) #13
	Basque, Finnish, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Norwegian, Serbian, Thai, Turkish, Welsh, Yoruba; –				
501	VSO ⇒ Aux V; SOV ⇒ V Aux	628	0	1.00	Greenberg
	$(v^{\ell}_{\text{HEAD:VERB-nsubj:NOUN-obj:NOUN}} \Rightarrow \neg v^{\ell}_{\text{HEAD:VERB-aux:AUX}}) \wedge$	51	0	1.00	(1963) #16,
	$(v^{\ell}_{\text{nsubj:NOUN-obj:NOUN-HEAD:VERB}} \Rightarrow v^{\ell}_{\text{HEAD:VERB-aux:AUX}})$				Dryer
	Amharic, Arabic, Armenian, Bambara, Basque, Breton, Bulgarian, Buryat, Cantonese, Chinese, Croatian, Czech, Danish, Dutch, English, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Komi Permyak, Korean, Latin, Lithuanian, Marathi, Norwegian, Persian, Polish, Romanian, Russian, Scottish Gaelic, Serbian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Vietnamese, Warlpiri, Welsh, Yoruba; –				(1992)
502	pronominal object follows the verb ⇒ nominal object follows the verb	30	0	1.00	Greenberg (1963) #25
	$v^{\ell}_{\text{HEAD:VERB-obj:PRON}} \Rightarrow v^{\ell}_{\text{HEAD:VERB-obj:NOUN}}$	14	0	1.00	
	Basque, Finnish, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Norwegian, Serbian, Thai, Turkish, Welsh, Yoruba; –				
964	VO ⇒ Art N; OV ⇒ N Art	613	6	0.99	*Dryer*
	$(v^{\ell}_{\text{HEAD:VERB-obj:NOUN}} \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-det:DET}}) \wedge$	47	1	0.98	*(1989)*,
	$(\neg v^{\ell}_{\text{HEAD:VERB-obj:NOUN}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-det:DET}})$				Dryer
	Amharic, Arabic, Armenian, Bambara, Basque, Breton, Bulgarian, Buryat, Cantonese, Chinese, Croatian, Czech, Danish, Dutch, English, Finnish, French, German, Greek, Hindi, Hungarian, Indonesian, Irish, Japanese, Komi Permyak, Korean, Latin, Lithuanian, Marathi, Persian, Polish, Russian, Scottish Gaelic, Serbian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Vietnamese, Warlpiri, Welsh, Yoruba; Romanian				(1992)
1017	verb-initial ⇒ N Gen	–	–	–	Dryer (1986)

	$(v^{\ell}_{\text{HEAD:VERB-nsubj:NOUN-obj:NOUN}} + v^{\ell}_{\text{HEAD:VERB-obj:NOUN-nsubj:NOUN}}) \Rightarrow v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}}$	–	–	–	
1114	VO $\Rightarrow$ CompS; $v^{\ell}_{\text{HEAD:VERB-obj:NOUN}} \Rightarrow \neg v^{\ell}_{\text{HEAD:VERB-mark:SCONJ}}$ Bambara, Basque, Chinese, English, German, Japanese, Korean, Persian, Tamil, Telugu, Thai, Turkish; –	23 12	0 0	1.00 1.00	Hawkins (1990)
1153	Rel N $\Rightarrow$ N Rel $\neg v^{\ell}_{\text{HEAD:NOUN-acl:VERB}} \Rightarrow v^{\ell}_{\text{HEAD:NOUN-acl:VERB}}$ Amharic, Ancient Greek, Armenian, Bambara, Basque, Bulgarian, Chinese, Coptic, Danish, English, French, German, Hebrew, Hungarian, Indonesian, Irish, Italian, Portuguese, Romanian, Russian, Spanish, Tagalog, Welsh, Wolof; –	107 24	0 0	1.00 1.00	Moravcsik (1969)
1183	OS $\Rightarrow$ #V $(v^{\ell}_{\text{HEAD:VERB-obj:NOUN-nsubj:NOUN}} + v^{\ell}_{\text{obj:NOUN-HEAD:VERB-nsubj:NOUN}} + v^{\ell}_{\text{obj:NOUN-nsubj:NOUN-HEAD:VERB}}) \Rightarrow (v^{\ell}_{\text{HEAD:VERB-nsubj:NOUN-obj:NOUN}} + v^{\ell}_{\text{HEAD:VERB-obj:NOUN-nsubj:NOUN}})$	– –	– –	– –	*Pullum (1977)*
1334	Prep $\Rightarrow$ ((RelN $\Rightarrow$ GenN) & (GenN $\Rightarrow$ AdjN) & (AdjN $\Rightarrow$ DemN)) $\neg v^{\ell}_{\text{HEAD:NOUN-case:ADP}} \Rightarrow ((\neg v^{\ell}_{\text{HEAD:NOUN-acl:VERB}} \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}}) \wedge (\neg v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}} \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}}) \wedge (\neg v^{\ell}_{\text{HEAD:NOUN-amod:ADJ}} \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-det:DET}}))$ Akkadian, Amharic, Arabic, Armenian, Bambara, Basque, Buryat, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Lithuanian, Norwegian, Persian, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Slovenian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Vietnamese, Welsh, Yoruba; –	328 47	0 0	1.00 1.00	Hawkins (1983)
1343	O V $\Rightarrow$ PP V; V O $\Rightarrow$ V PP	616	0	1.00	Dryer (1992)

	$(\neg v^{\ell}_{\textsc{head}\text{:VERB-obj:NOUN}} \Rightarrow (v^{\ell}_{\text{case:ADP-obl:}\textsc{head}_2\text{:NOUN-}\textsc{head}_1\text{:VERB}} + v^{\ell}_{\text{obl:}\textsc{head}_2\text{:NOUN-case:ADP-}\textsc{head}_1\text{:VERB}})) \wedge (v^{\ell}_{\textsc{head}\text{:VERB-obj:NOUN}} \Rightarrow (v^{\ell}_{\textsc{head}_1\text{:VERB-case:ADP-obl:}\textsc{head}_2\text{:NOUN}} + v^{\ell}_{\textsc{head}_1\text{:VERB-obl:}\textsc{head}_2\text{:NOUN-case:ADP}}))$ Amharic, Arabic, Armenian, Bambara, Basque, Breton, Bulgarian, Buryat, Cantonese, Chinese, Croatian, Czech, Danish, Dutch, English, Finnish, French, German, Greek, Hindi, Hungarian, Indonesian, Irish, Japanese, Komi Permyak, Korean, Latin, Lithuanian, Marathi, Persian, Polish, Romanian, Russian, Scottish Gaelic, Serbian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Vietnamese, Warlpiri, Welsh, Yoruba; –	48	0	1.00	
1346	PP-V ⇔ Adv V; V-PP ⇔ V Adv $((v^{\ell}_{\text{case:ADP-obl:}\textsc{head}_2\text{:NOUN-}\textsc{head}_1\text{:VERB}} + v^{\ell}_{\text{obl:}\textsc{head}_2\text{:NOUN-case:ADP-}\textsc{head}_1\text{:VERB}}) \Leftrightarrow \neg v^{\ell}_{\textsc{head}\text{:VERB-advmod:ADV}}) \wedge ((v^{\ell}_{\textsc{head}_1\text{:VERB-case:ADP-obl:}\textsc{head}_2\text{:NOUN}} + v^{\ell}_{\textsc{head}_1\text{:VERB-obl:}\textsc{head}_2\text{:NOUN-case:ADP}}) \Leftrightarrow v^{\ell}_{\textsc{head}\text{:VERB-advmod:ADV}})$ Amharic, Arabic, Armenian, Bambara, Basque, Breton, Bulgarian, Buryat, Cantonese, Chinese, Croatian, Czech, Danish, Dutch, English, Finnish, French, German, Greek, Hindi, Hungarian, Indonesian, Irish, Japanese, Komi Permyak, Korean, Latin, Lithuanian, Marathi, Persian, Polish, Romanian, Russian, Scottish Gaelic, Serbian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Vietnamese, Warlpiri, Welsh, Yoruba; –	616 48	0 0	1.00 1.00	Dryer (1992)
1419	OV ⇒ Predicate Copula; VO ⇒ Copula Predicate $(\neg v^{\ell}_{\textsc{head}\text{:VERB-obj:NOUN}} \Rightarrow v^{\ell}_{\textsc{head}\text{:NOUN-cop:AUX}}) \wedge (v^{\ell}_{\textsc{head}\text{:VERB-obj:NOUN}} \Rightarrow \neg v^{\ell}_{\textsc{head}\text{:NOUN-cop:AUX}})$ Amharic, Arabic, Armenian, Bambara, Basque, Breton, Bulgarian, Buryat, Cantonese, Chinese, Croatian, Czech, Danish, Dutch, English, Finnish, French, German, Greek, Hindi, Hungarian, Indonesian, Irish, Japanese, Komi Permyak, Korean, Latin, Lithuanian, Marathi, Persian, Polish, Romanian, Russian, Scottish Gaelic, Serbian,	616 48	0 0	1.00 1.00	Dryer (1992)

	Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Vietnamese, Warlpiri, Welsh, Yoruba; –				
1512	OS ⇒ #V $(v^{\ell}_{\textsc{head}\text{:VERB-obj:NOUN-nsubj:NOUN}} + v^{\ell}_{\text{obj:NOUN-}\textsc{head}\text{:VERB-nsubj:NOUN}} + v^{\ell}_{\text{obj:NOUN-nsubj:NOUN-}\textsc{head}\text{:VERB}}) \Rightarrow (v^{\ell}_{\textsc{head}\text{:VERB-nsubj:NOUN-obj:NOUN}} + v^{\ell}_{\textsc{head}\text{:VERB-obj:NOUN-nsubj:NOUN}})$	11 0	0 0	1.00 –	Keenan (1978)
1517	OS ⇒ prepositions $(v^{\ell}_{\textsc{head}\text{:VERB-obj:NOUN-nsubj:NOUN}} + v^{\ell}_{\text{obj:NOUN-}\textsc{head}\text{:VERB-nsubj:NOUN}} + v^{\ell}_{\text{obj:NOUN-nsubj:NOUN-}\textsc{head}\text{:VERB}}) \Rightarrow \neg v^{\ell}_{\textsc{head}\text{:NOUN-case:ADP}}$	11 0	0 0	1.00 –	Keenan (1978)
1520	OS ⇒ N Gen $(v^{\ell}_{\textsc{head}\text{:VERB-obj:NOUN-nsubj:NOUN}} + v^{\ell}_{\text{obj:NOUN-}\textsc{head}\text{:VERB-nsubj:NOUN}} + v^{\ell}_{\text{obj:NOUN-nsubj:NOUN-}\textsc{head}\text{:VERB}}) \Rightarrow v^{\ell}_{\textsc{head}\text{:NOUN-nmod:NOUN}}$	11 0	0 0	1.00 –	Keenan (1978)
1521	OS ⇒ N Rel $(v^{\ell}_{\textsc{head}\text{:VERB-obj:NOUN-nsubj:NOUN}} + v^{\ell}_{\text{obj:NOUN-}\textsc{head}\text{:VERB-nsubj:NOUN}} + v^{\ell}_{\text{obj:NOUN-nsubj:NOUN-}\textsc{head}\text{:VERB}}) \Rightarrow v^{\ell}_{\textsc{head}\text{:NOUN-acl:VERB}}$	10 0	1 0	0.91 –	Keenan (1978)
1524	OS ⇒ Num N $(v^{\ell}_{\textsc{head}\text{:VERB-obj:NOUN-nsubj:NOUN}} + v^{\ell}_{\text{obj:NOUN-}\textsc{head}\text{:VERB-nsubj:NOUN}} + v^{\ell}_{\text{obj:NOUN-nsubj:NOUN-}\textsc{head}\text{:VERB}}) \Rightarrow \neg v^{\ell}_{\textsc{head}\text{:NOUN-nummod:NUM}}$	11 0	0 0	1.00 –	Keenan (1978)
1525	OS ⇒ Art N $(v^{\ell}_{\textsc{head}\text{:VERB-obj:NOUN-nsubj:NOUN}} + v^{\ell}_{\text{obj:NOUN-}\textsc{head}\text{:VERB-nsubj:NOUN}} + v^{\ell}_{\text{obj:NOUN-nsubj:NOUN-}\textsc{head}\text{:VERB}}) \Rightarrow \neg v^{\ell}_{\textsc{head}\text{:NOUN-det:DET}}$	11 0	0 0	1.00 –	Keenan (1978)
1539	verb-initial ⇒ [S and S]; verb-initial ⇒ ¬ [S, S and]	–	–	–	*Payne (1990)*

	$((v^{\ell}_{\text{HEAD:VERB-nsubj:NOUN-obj:NOUN}} + v^{\ell}_{\text{HEAD:VERB-obj:NOUN-nsubj:NOUN}}) \Rightarrow v^{\ell}_{\text{HEAD}_1\text{:VERB-cc:CCONJ-conj:HEAD}_2\text{:VERB}}) \wedge ((v^{\ell}_{\text{HEAD:VERB-nsubj:NOUN-obj:NOUN}} + v^{\ell}_{\text{HEAD:VERB-obj:NOUN-nsubj:NOUN}}) \Rightarrow \neg v^{\ell}_{\text{HEAD}_1\text{:VERB-conj:HEAD}_2\text{:VERB-cc:CCONJ}})$	–	–	–	
1543	verb-initial ⇒ preposition	–	–	–	*Payne*
	$(v^{\ell}_{\text{HEAD:VERB-nsubj:NOUN-obj:NOUN}} + v^{\ell}_{\text{HEAD:VERB-obj:NOUN-nsubj:NOUN}}) \Rightarrow \neg v^{\ell}_{\text{HEAD:NOUN-case:ADP}}$	–	–	–	*(1990)*
1549	verb-initial ⇒ N Gen	–	–	–	*Payne*
	$(v^{\ell}_{\text{HEAD:VERB-nsubj:NOUN-obj:NOUN}} + v^{\ell}_{\text{HEAD:VERB-obj:NOUN-nsubj:NOUN}}) \Rightarrow v^{\ell}_{\text{HEAD:NOUN-nmod:NOUN}}$	–	–	–	*(1990)*
1550	verb-initial ⇒ N Rel	–	–	–	*Payne*
	$(v^{\ell}_{\text{HEAD:VERB-nsubj:NOUN-obj:NOUN}} + v^{\ell}_{\text{HEAD:VERB-obj:NOUN-nsubj:NOUN}}) \Rightarrow v^{\ell}_{\text{HEAD:NOUN-acl:VERB}}$	–	–	–	*(1990)*
1557	verb-initial ⇒ Main verb-Subordinate	–	–	–	*Payne*
	$(v^{\ell}_{\text{HEAD:VERB-nsubj:NOUN-obj:NOUN}} + v^{\ell}_{\text{HEAD:VERB-obj:NOUN-nsubj:NOUN}}) \Rightarrow v^{\ell}_{\text{HEAD:VERB-xcomp:VERB}}$	–	–	–	*(1990)*

Table 25: Extracted universals from the Universals Archive: number in the Universals Archive; formula, number of supportive languages, number of counterexamples, percentage of supportive languages (above: original paper(s); below: UD treebanks); original paper(s) (italics indicate that a list of languages surveyed in the paper or the paper itself was not available; if the original paper is Greenberg (1963), the corresponding number is also given); list of supportive languages and counterexamples in the UD treebanks (if any).

B.4 POS influence

MAE	σ	$\vec{v}_1$	$\vec{v}_2$
0.00	0.00	$\vec{v}_{\text{HEAD-conj}}$	$\vec{v}_{\text{HEAD:NOUN-conj:NOUN}}$
0.00	0.00	$\vec{v}_{\text{HEAD-fixed}}$	$\vec{v}_{\text{HEAD:ADP-fixed:NOUN}}$
0.00	0.00	$\vec{v}_{\text{HEAD-flat}}$	$\vec{v}_{\text{HEAD:PROPN-flat:PROPN}}$
0.00	0.00	$\vec{v}_{\text{HEAD-root}}$	$\vec{v}_{\text{HEAD:ROOT-root:VERB}}$
0.01	0.02	$\vec{v}_{\text{HEAD-appos}}$	$\vec{v}_{\text{HEAD:NOUN-appos:NOUN}}$
0.01	0.02	$\vec{v}_{\text{HEAD-clf}}$	$\vec{v}_{\text{HEAD:NOUN-clf:NOUN}}$
0.02	0.04	$\vec{v}_{\text{HEAD-ccomp}}$	$\vec{v}_{\text{HEAD:VERB-ccomp:VERB}}$
0.02	0.04	$\vec{v}_{\text{HEAD-amod}}$	$\vec{v}_{\text{HEAD:NOUN-amod:ADJ}}$
0.02	0.05	$\vec{v}_{\text{HEAD-expl}}$	$\vec{v}_{\text{HEAD:VERB-expl:PRON}}$
0.03	0.06	$\vec{v}_{\text{HEAD-list}}$	$\vec{v}_{\text{HEAD:PROPN-list:PROPN}}$
0.03	0.09	$\vec{v}_{\text{HEAD-cc}}$	$\vec{v}_{\text{HEAD:VERB-cc:CCONJ}}$
0.03	0.10	$\vec{v}_{\text{HEAD-case}}$	$\vec{v}_{\text{HEAD:NOUN-case:ADP}}$
0.03	0.06	$\vec{v}_{\text{HEAD-nummod}}$	$\vec{v}_{\text{HEAD:NOUN-nummod:NUM}}$
0.04	0.12	$\vec{v}_{\text{HEAD-det}}$	$\vec{v}_{\text{HEAD:NOUN-det:DET}}$
0.04	0.06	$\vec{v}_{\text{HEAD-advcl}}$	$\vec{v}_{\text{HEAD:VERB-advcl:VERB}}$
0.04	0.06	$\vec{v}_{\text{HEAD-obl}}$	$\vec{v}_{\text{HEAD:VERB-obl:NOUN}}$
0.04	0.05	$\vec{v}_{\text{HEAD-cop}}$	$\vec{v}_{\text{HEAD:NOUN-cop:AUX}}$
0.04	0.10	$\vec{v}_{\text{HEAD-acl}}$	$\vec{v}_{\text{HEAD:NOUN-acl:VERB}}$
0.05	0.06	$\vec{v}_{\text{HEAD-parataxis}}$	$\vec{v}_{\text{HEAD:VERB-parataxis:VERB}}$
0.05	0.18	$\vec{v}_{\text{HEAD-aux}}$	$\vec{v}_{\text{HEAD:VERB-aux:AUX}}$
0.06	0.10	$\vec{v}_{\text{HEAD-nmod}}$	$\vec{v}_{\text{HEAD:NOUN-nmod:NOUN}}$
0.06	0.13	$\vec{v}_{\text{HEAD-xcomp}}$	$\vec{v}_{\text{HEAD:VERB-xcomp:VERB}}$
0.06	0.07	$\vec{v}_{\text{HEAD-nsubj}}$	$\vec{v}_{\text{HEAD:VERB-nsubj:NOUN}}$
0.06	0.14	$\vec{v}_{\text{HEAD-mark}}$	$\vec{v}_{\text{HEAD:VERB-mark:SCONJ}}$
0.06	0.05	$\vec{v}_{\text{HEAD-punct}}$	$\vec{v}_{\text{HEAD:VERB-punct:PUNCT}}$
0.06	0.09	$\vec{v}_{\text{HEAD-csubj}}$	$\vec{v}_{\text{HEAD:VERB-csubj:VERB}}$
0.06	0.18	$\vec{v}_{\text{HEAD-goeswith}}$	$\vec{v}_{\text{HEAD:NUM-goeswith:NUM}}$
0.07	0.08	$\vec{v}_{\text{HEAD-vocative}}$	$\vec{v}_{\text{HEAD:VERB-vocative:NOUN}}$

MAE	σ	$\vec{v}_1$	$\vec{v}_2$
0.08	0.06	$\vec{v}_{\text{HEAD-advmod}}$	$\vec{v}_{\text{HEAD:VERB-advmod:ADV}}$
0.08	0.08	$\vec{v}_{\text{HEAD-obj}}$	$\vec{v}_{\text{HEAD:VERB-obj:NOUN}}$
0.11	0.19	$\vec{v}_{\text{HEAD-reparandum}}$	$\vec{v}_{\text{HEAD:VERB-reparandum:VERB}}$
0.11	0.15	$\vec{v}_{\text{HEAD-orphan}}$	$\vec{v}_{\text{HEAD:NOUN-orphan:NOUN}}$
0.12	0.19	$\vec{v}_{\text{HEAD-iobj}}$	$\vec{v}_{\text{HEAD:VERB-iobj:NOUN}}$
0.16	0.24	$\vec{v}_{\text{HEAD-dislocated}}$	$\vec{v}_{\text{HEAD:VERB-dislocated:NOUN}}$
0.19	0.24	$\vec{v}_{\text{HEAD-discourse}}$	$\vec{v}_{\text{HEAD:VERB-discourse:ADV}}$
0.20	0.25	$\vec{v}_{\text{HEAD-compound}}$	$\vec{v}_{\text{HEAD:NOUN-compound:NOUN}}$
0.24	0.23	$\vec{v}_{\text{HEAD-dep}}$	$\vec{v}_{\text{HEAD:NOUN-dep:SYM}}$
0.05	0.12	overall	

Table 26: Mean absolute errors between property vectors without and with POS information.