Natalia Levshina*

Token-based typology and word order entropy: A study based on Universal Dependencies

https://doi.org/10.1515/lingty-2019-0025
Received December 27, 2017; revised March 15, 2019

Abstract: The present paper discusses the benefits and challenges of token-based typology, which takes into account the frequencies of words and constructions in language use. This approach makes it possible to introduce new criteria for language classification, which would be difficult or impossible to achieve with the traditional, type-based approach. This point is illustrated by several quantitative studies of word order variation, which can be measured as entropy at different levels of granularity. I argue that this variation can be explained by general functional mechanisms and pressures, which manifest themselves in language use, such as optimization of processing (including avoidance of ambiguity) and grammaticalization of predictable units occurring in chunks. The case studies are based on multilingual corpora, which have been parsed using the Universal Dependencies annotation scheme.

Keywords: corpora, word order, frequency, Universal Dependencies, entropy, token-based typology, language classification

1 Token-based typology and word order variation

This paper discusses the role of usage frequencies with regard to the fundamental goals of typology: language classification, and identification and explanation of cross-linguistic generalizations (cf. Croft 2003: 1–2). In most typological research, languages have been treated as single data points with a categorical value (e.g. OV or VO, prepositional or postpositional). The overwhelming majority of typological universals (e.g. the ones found in The Universals Archive at the University of Konstanz)[1] are of this kind. Similarly, the influential *World Atlas of Language Structures* (WALS) (Dryer & Haspelmath 2013) contains only

1 https://typo.uni-konstanz.de/archive/intro/index.php, Accessed on 2019–03–13.

***Corresponding author: Natalia Levshina [nɐˈtalʲjə lʲefʃɪnə],** Institute of British Studies, Leipzig University, (IPF 141199), Nikolaistraße 8-10, Leipzig 04109, Germany, E-mail: natalevs@gmail.com

categorical or ordinal variables, which characterize language types. I will refer to this approach as **type-based**.

In contrast, **token-based typology** makes generalizations and classifies languages using the tokens of specific linguistic units or structures observed in language use, as approximated by corpora. It can also use aggregate variables derived from the distributions of usage tokens, such as entropy, complexity, average dependency length, etc. Unlike in the type-based approach, the variables are continuous and reflect language-internal variation. This is a growing area of research, which has been boosted by the increasing number of multilingual corpora becoming available nowadays. These corpora range from translations of popular children's books like *Harry Potter*, to the Universal Declaration of Human Rights, and from Wikipedia articles to collections of spoken traditional narratives (see examples in the Supplementary Materials: *Multilingual Corpora*). Corpora have been used for diverse goals in recent typological and comparative research, such as the following:

- testing of general functional laws and principles, e.g. the correspondence between frequency and amount of coding (Haspelmath et al. 2014), Zipf's law of abbreviation (Zipf 1935; Bentz & Ferrer-i-Cancho 2015) and effects of average surprisal on word length (Piantadosi et al. 2011);
- explanation of common cross-linguistic patterns, e.g. discussion of ergativity in terms of preferred argument structure (e.g. Du Bois et al. 2003; Haig & Schnell 2016b);
- induction of cross-linguistically salient dimensions of conceptual variation with the help of probabilistic semantic maps (e.g. Wälchli & Cysouw 2012; Levshina 2015);
- language classification based on quantitative indices derived from corpora, e.g. different measures of linguistic complexity on the basis of Kolmogorov complexity and entropy (e.g. Juola 1998; Koplenig et al. 2017).

By using continuous variables instead of categorical ones, it is possible to capture intra-linguistic variation, which is ubiquitous in language, at the same time avoiding the existing bias towards a restricted set of linguistic patterns, which display low language-internal variability and cross-linguistic bimodal distributions (Wälchli 2009; see also Section 3.3). As put by Diessel, "language consists of fluid structures and probabilistic constraints that are shaped by communication, memory, and processing" (2017: 2). In this paper I want to demonstrate the typological relevance of intra-linguistic variation, which emerges as the result of such probabilistic constraints and different performance pressures.

For illustration, I will focus on word order. This domain has been thoroughly investigated on the basis of categorical typological data, especially the correlations between different word order patterns (e.g. Greenberg 1963; Vennemann 1974; Lehmann 1978; Dryer 1992; Dunn et al. 2011, to name just a few). There has been corpus-based work on word order, as well. In particular, Wälchli (2009) studied the verb–locative order in New Testament translations, focusing on specific contexts, such as the imperative domain. Liu (2010) used treebanks to investigate the continuum between head-initial and head-final languages. In Östling (2015), the outcomes of token-based corpus analyses were compared to the type-based classifications from the WALS (Dryer & Haspelmath 2013). Guzmán Naranjo and Becker (2018) focused on correlations between verb-headed and noun-headed dependencies in the Universal Dependencies corpora. All these studies are examples of the token-based approach.

In the present paper, I will focus on one aspect of word order typology that is impossible without the token-based approach, namely, word order variability. This aspect has received less attention in general linguistics and typology than word order correlations, although it was quite prominent in formal linguistics in the 1980s. It was argued, in particular, that languages belong to two main categories: configurational and non-configurational. Unlike configurational languages (the standard example being English, as usual), non-configurational languages exhibit relatively free word order along with other characteristics, such as extensive use of zero anaphoras (or pro-drop), the use of discontinuous expressions and rich case systems (Hale 1982, Hale 1983). Examples are Warlpiri (Hale 1983), Japanese (Chomsky 1981) and Hungarian (Kiss 1987). At the same time, individual languages tend to exhibit these properties to different degrees, which makes one doubt that configurationality as a single typological parameter exists (Rögnvaldsson 1995). Moreover, as will be shown in this paper, there is no such thing as absolutely rigid or absolutely free word order, some patterns in one language being more fixed, others more flexible. In addition, even configurational languages like English allow for some variation, e.g. the subject – verb inversion in *In the middle of the room stood an antique mahogany table*. At the same time, word order in non-configurational languages is not completely free, either. For example, as will be shown in Section 3.2, the subject – object order in written Japanese, identified as a non-configurational language, does not exhibit substantially more variation than the same pattern in English, and has in fact more rigid order in several other patterns. From this it follows that the notion of configurationality is not particularly useful for predicting actual language use, which is the primary concern of the present paper.

On the functional side, there has also been some work on word order variability. One of the central questions has been to what extent pragmatic factors influence the order in different languages. In some languages (e.g. English) the role of syntactic functions in determining the word order is greater than in others (e.g. Czech, Warlpiri and Ojibwa). In languages of the latter type, the role of pragmatic functions is more prominent, such as topic and comment, definiteness and indefiniteness of the referents, given and new information, newsworthiness, etc. (Givón 1984: Ch. 6; Payne 1992). Among such languages, it is also possible to make more subtle distinctions. For instance, Czech speakers are highly sensitive to what constitutes 'natural' word order, while there is no such order in languages like Cayuga (Iroquoian, spoken in the USA, ISO 'cay') and Ngandi (Gunwinyguan, spoken in Australia, ISO 'nid') (Mithun 1992). Instead of using some marked word order for the purposes of focus, as one would do in Czech, the speakers will use additional morphological marking.

These studies usually investigate a limited number of languages and only a selection of word order patterns, due to the high costs of manual analysis of such data. They are mostly qualitative and corpus-illustrated, rather than quantitative and corpus-driven. With the growing number of available corpora, however, one can work out objective criteria and procedures for quantification of word order variation. It is also possible to compute and compare variability measures on different levels of abstraction for a large number of languages and word order patterns. Moreover, corpora give us an opportunity to understand how this variability is related to other linguistic parameters and universal functional constraints, and test the hypotheses statistically. These are the goals of the present paper.

The data for the case studies are taken from different corpora, which are annotated lexically, morphologically and syntactically according to the Universal Dependencies annotation scheme (Nivre et al. 2017). The datasets are provided in the Supplementary Materials: *Datasets*. All statistical analyses presented below were performed with the help of R, free statistical software (R Core Team 2018).

It is necessary to mention that word order freedom has already been discussed from a corpus-based perspective by Futrell et al. (2015). They used several aggregate measures based on conditional entropy of whether a head is to the right or left of a dependent. The entropy measures were conditioned on such features as relation type, part of speech of head and dependent, and syntactic subtree. Each language was represented by only one aggregate score of each type. In contrast to this 'black box' approach, the present paper discusses the metrics that are computed for each syntactic dependency individually, in the spirit of the late data aggregation approach (Zakharko et al. 2017).

This allows us to examine the intra- and cross-linguistic variation in greater detail, as well as to distinguish between different factors that determine the amount of variability.

The remaining part of the paper is structured as follows. Section 2 provides information about the data from the Universal Dependencies corpora and the main quantitative concepts. In Section 3, I present the results of exploratory analyses of word order entropy both for individual languages, and for individual word order patterns, and provide explanations of this variation. Section 4 compares the word order variation at different levels of granularity, showing that some (but not all) variation at the level of abstract syntactic dependencies can be explained by the variation at the level of wordforms. In Section 5, I test several usage-based functional explanations that play a role in determining the variation of word order. In Section 6, I provide a summary of the findings and an outlook.

2 Word order patterns based on the Universal Dependencies

2.1 Dependencies and co-dependencies

The case studies presented in this paper are based on the frequencies of different word order patterns in corpora annotated using the Universal Dependencies (UD) approach. The cross-linguistic syntactic categories and parts of speech used in the UD framework are the result of a long evolution. One of the main goals and challenges of the UD project is to keep the annotation design satisfactory both for language-specific analyses and for typological comparison.[2]

I use the frequencies of so-called heads and dependent elements, as well as some co-dependencies (e.g. subject and object of the same verbal predicate). The full list is given in Table 1. This selection includes the main dependencies in nominal phrases, verbal phrases, simple clauses and complex sentences. Nominal and pronominal subjects, objects, obliques and modifiers of nouns were treated separately, since the order of nominal and pronominal constituents is often different. Subordinators in adverbial and complement clauses were extracted separately, as well, and so were adverbial modifiers of adjectives and verbs. Note that in the case of interclausal dependencies, the head is the

2 See more information at http://universaldependencies.org/introduction.html.

Table 1: Word order patterns (syntactic dependencies and co-dependencies) examined in the present paper.

Type	Label in this study	Label in UD	Dependent	Head	Example
Nominals and their heads	nsubjNoun_Pred, nsubjPron_Pred	nsubj	Subject (noun or pronoun)	Predicate of the main clause (root)	$Jane_{NSUBJ}$ $reads_{HEAD}$ many books.
	objNoun_Pred, objPron_Pred	obj	Direct object (noun or pronoun)	Predicate of the main clause (root)	Jane $reads_{HEAD}$ many $books_{OBJ}$.
	oblNoun_Pred, oblPron_Pred	obl	Oblique phrase, i.e. NP encoded differently from the core arguments (noun or pronoun).	Predicate of the main clause (root)	Jane $looks_{HEAD}$ at the $stars_{OBL}$.
	nmodNoun_Noun, nmodPron_Noun	nmod	Nominal dependent (noun or pronoun)	Noun	Jane is reading a $book_{HEAD}$ on quantum $mechanics_{NMOD}$.
Co-dependent nominals	nsubj_obj	nsubj, obj	Nominal subject and nominal object	–	$Jane_{NSUBJ}$ is reading a $book_{OBJ}$.
	obj_obl	obj, obl	Nominal object and nominal oblique phrase	–	Jane is reading a $book_{OBJ}$ in the $library_{OBL}$.
Modifiers and their heads	nummod_Noun	nummod	Numeric modifier	Noun	Jane published ten_{NUMMOD} $books_{HEAD}$.
	amod_Noun	amod	Adjectival modifier	Noun	Jane is a $famous_{AMOD}$ $scholar_{HEAD}$.
	advmod_Verb, advmod_Adj	advmod	Adverbial modifier	Verb or adjective	Jane $writes_{HEAD}$ $clearly_{ADVMOD}$. $genetically_{ADVMOD}$ $modified_{HEAD}$ food

Function words and their heads	det_Noun	det	Determiner (article, possessive pronoun, demonstrative pronouns, etc.)	Noun	Jane is reading a_{DET} $book_{HEAD}$.
	adp_Noun	case	Adposition or clitic case marker	Noun	$Jane_{HEAD}$'s_{CASE} recent book; a book on_{CASE} $syntax_{HEAD}$
	aux_Verb	aux	Auxiliary (tense, mood, aspect, voice or evidentiality)	Verb	Jane is_{AUX} $reading_{HEAD}$.
	cop_Pred	cop	Copula	Any nominal	Jane is_{COP} $honest_{HEAD}$.
	mark_ccomp, mark_advcl	mark	Subordinators (complementizers or subordinating conjunctions)	Predicate of complement clause (mark_ccomp) or adverbial clause (mark_advcl)	Jane hopes $that_{MARK}$ her paper $appears_{HEAD}$ soon.
Clauses and their heads	csubj_Main	csubj	Clausal subject	Predicate of the main clause	What she $said_{CSUBJ}$ $makes_{HEAD}$ sense.
	ccomp_Main	ccomp	Clausal complement	Predicate of the main clause	Jane $thinks_{HEAD}$ Peter $cooks_{CCOMP}$ very well.
	acl_Noun	acl	Finite and non-finite clausal modifier (adjectival clause)	Noun	a $book_{HEAD}$ that Jane $wrote_{ACL}$
	advcl_Main	advcl	Adverbial clause modifier	Predicate of the main clause	Jane was sad_{HEAD} when I $talked_{ADVCL}$ to her.

predicate of the main clause or the head noun (for relative clauses), and the dependent element is the predicate of the subordinate clause. The dependencies related to punctuation, coordination, various discourse markers, vocatives and multiword expressions were not taken into account. I also excluded the so-called root (the predicate of the main clause), which does not depend on anything, and dependencies which are infrequent in the UD sample of languages (e.g. classifiers). Some contextual restrictions were made, as well. In particular, questions and exclamatory sentences were disregarded. Subjects, objects and obliques were only counted when they occurred in the main clause.

Note that the understanding of heads and dependents in the UD corpora is different from some syntactic theories in that functional elements (adpositions, auxiliaries, subordinators) are coded as dependents, not as heads.[3] One may agree or disagree with that approach, but it plays no role in the analyses that follow. The reason is that the central measure in this study, namely, entropy of the order of two elements X and Y, remains the same regardless of whether X is the head and Y is the dependent, or the other way round (see Section 2.2 for more details).

2.2 Shannon entropy

The main measure used in this study is Shannon entropy (Shannon 1948). It represents variation of word order in the twenty-four dependencies and co-dependencies described in the previous section. For each of the word order patterns in a corpus, I computed the entropy using the formula in (1):

$$H(X) = -\sum_{i=1}^{2} P(x_i)\log_2 P(x_i) \quad (1)$$

where X is a binary variable representing two possible word orders, e.g. Determiner + Noun and Noun + Determiner. $P(x_i)$ is the probability of one of the orders, which equals its relative frequency (proportion) in a given corpus. If the proportion of one word order (e.g. Determiner + Noun) is 1, and the proportion of the reverse order (e.g. Noun + Determiner) is 0, or the other way round, the entropy H is equal to zero. There is no variation. If the proportion of each of the possible word orders is 0.5, the entropy takes the maximum value of 1. If both orders are attested, and one of them is more frequent than the other, then the entropy lies

3 See more at http://universaldependencies.org/u/overview/syntax.html.

between 0 and 1. For example, if the proportion of Determiner + Noun in a specific corpus is 0.9 (or 90%), and the proportion of Noun + Determiner is 0.1 (or 10%), then the entropy of this dependency is $H = -(0.9 \times log_2 0.9 + 0.1 \times log_2 0.1) = 0.47$. These numbers suggest that the relationship between probability and entropy is non-linear. As shown in Figure 1, this is indeed the case. Already a small amount of variation is sufficient to obtain relatively high values of entropy.

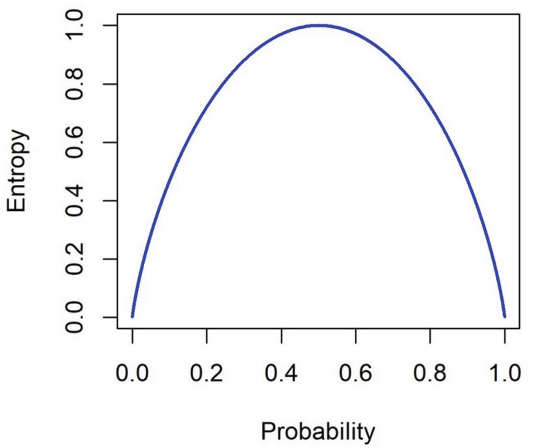

Figure 1: Non-linear relationship between probability of a word order pattern A followed by B (as opposed to B followed by A) and entropy.

Entropy can be computed at different levels of granularity. For example, one can use the level of abstract dependencies, as shown above, or compute entropy for individual wordforms that represent dependencies, e.g. the entropy of *actively_ADV* + *VERB* and *VERB* + *actively_ADV*, where *actively* is an adverbial modifier of a verb. This approach will be discussed in Section 4.

Word order entropy, as it is defined here, is a purely descriptive and data-driven measure based on actual word order patterns. Its most important advantage is that it is objectively defined, unlike such notions as word order "freedom" or "flexibility", which usually reflect intuitions about the acceptability of different word order variants in a pre-defined set of contexts. Although such intuitions may be of substantial interest, this subjective kind of variability is much more difficult to define, measure and compare cross-linguistically than corpus-based entropy.

It is also necessary to mention a methodological caveat. One of the main challenges of working with cross-linguistic corpora is their representativeness

and comparability of the text types (cf. Croft 2003: 112). There are no multilingual corpora at the moment that could be compared to the carefully sampled British National Corpus or similar standards. The existing multilingual corpora can represent one text (e.g. *Le Petit Prince* used by Stolz et al. 2017) or numerous texts of a few rather specific genres (e.g. OPUS by Tiedemann 2012 with film subtitles, European Parliament transcripts and technical documentation). Comparable spoken corpora are also available, but for a limited number of languages (e.g. Haig & Schnell 2016a; Dingemanse et al. 2013).

To what extent text types are important for cross-linguistic investigations of word order entropy is an empirical question. In order to check if entropy scores are strongly influenced by text types, I compared different text types in eight languages from the Leipzig Corpora Collection (Goldhahn et al. 2012),[4] which consists of comparable corpora in diverse languages. The procedure and the results are described in the Appendix. The comparison suggests that the differences between different text types in one language are small.

3 Case Studies 1a and 1b: Word order entropy at the level of (co-)dependencies

3.1 Data for Case Studies 1a and 1b

This section discusses word order entropy in the Universal Dependencies corpora 2.1 (Nivre et al. 2017). This is a collection of 102 treebanks (i.e. syntactically parsed corpora) representing 60 languages from 24 genera and twelve different families, including one sign language and several isolates (Basque, Japanese and Korean). More details can be found in the Supplementary Materials: *UDCorpora*. First, the frequencies of every (co-)dependency in each possible order (e.g. Determiner + Noun and Noun + Determiner) in every corpus were extracted with the help of a Python script. If a language was represented by several corpora, all counts were summarized. Dependencies with frequency less than 20 (usually due to the small size of some corpora) were excluded from the subsequent analyses. Next, the proportions of word order pairs (e.g. Determiner + Noun and Noun + Determiner) were computed for each language and (co-)dependency. Finally, the entropy scores were computed as described in Section 2.2.

4 http://wortschatz.uni-leipzig.de/de/download (Accessed on 2018–11–11).

3.2 Case Study 1a. A probabilistic typology of languages based on word order entropy

This subsection presents a language classification based on word order entropy. Several languages with small frequencies were excluded (Buryat, Cantonese, Korean, Kurmanji, Sanskrit and Swedish Sign Language). The entropy values of the 24 (co-)dependencies were averaged in every language. Figure 2 presents the mean entropy scores for the languages in the UD corpora. The horizontal axis represents the average entropy of the head-dependent patterns. The vertical axis shows the average entropy in the order of co-dependents (SO/OS and OX/XO, where X stands for oblique). The higher the score, the more variable the word order is on average in a given language. Trying out combinations of other parameters (e.g. entropy of function words or clauses) did not reveal additional interesting patterns or clusters. The reason is that most of the entropy values are positively correlated.[5]

Figure 2 shows that quite a few languages have high entropy on both dimensions (the top right corner). These are morphologically rich European languages (Basque, Finnic and Slavic), including several ancient Indo-European languages (Ancient Greek, Gothic, Latin and Old Church Slavonic). They are followed by the more analytical Romance, Germanic and Semitic languages, which have moderate scores. Notably, if languages have very low entropy, they have either low variation of the head-dependent patterns, or low variation of the co-dependent patterns, but not both. In particular, the Altaic and Dravidian languages and Japanese have very low entropy of head-dependent patterns, but the order of co-dependents is to some extent variable (cf. Jamashita & Chang 2000). In contrast, Irish, Coptic and especially Mandarin have the most rigid orders of co-dependents, but moderately rigid orders of heads and dependents, in comparison with the other languages. One may wonder if there exists a language in the world that allows no variation in word order at all.

5 See the Supplementary Materials: *Mean Entropy Per Genus Correlations*. An exception is the position of subordinators, which is highly variable in low-entropy Mandarin Chinese, and to some extent in Vietnamese, Indic and Dravidian languages. In contrast, subordinators tend to be rigid in the other languages, most of which have more variable word orders in general. As a consequence, there are negative correlations between the entropy values of subordinators and most of the other dependencies. According to Diessel (2001), languages with flexible position of adverbial clauses usually have adverbial subordinators in the beginning of the subordinate clause. This observation is fully supported by the UD data. Variable and final adverbial subordinators are only observed in the languages where the adverbial clauses always precede the main clause. For complement clauses and complementizers, the picture is not so clear, however.

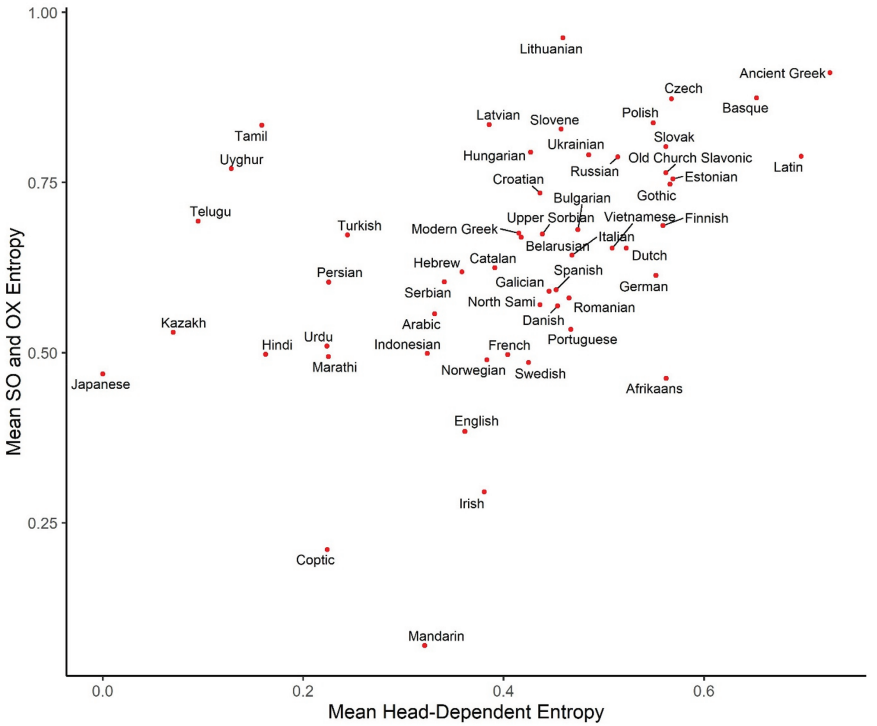

Figure 2: Mean head-dependent order entropy and co-dependent entropy in the UD corpora.

How can we interpret this variation? Let us begin with the horizontal dimension, which shows the head-dependent entropy. Many morphologically rich, highly synthetic languages are located on the right. The analytic languages tend to be located close to the middle. The left-hand part of the figure with very low entropy includes verb-final languages (Turkish, Japanese, etc.), which are synthetic, as well (e.g. Hawkins 2014: 139–146). This supports the results presented by Futrell et al. (2015), who found that languages which have highly predictable head direction (i.e. low head – dependent entropy) tend to be mostly head-final. The low variation of the head-dependent orders in verb-final languages may be a manifestation of their 'tight fit' with regard to argument and predicate frame differentiation (Hawkins 2014: Section 7.2). OV languages usually have a morphological case system, narrow semantic range of basic grammatical relationships, no raising or wh-extractions. All these restrictions, together with rigid word order, help the hearer to identify the argument structure early, so that there is no need to make corrections after the verb comes.

The vertical dimension reflects mostly variation in the order of subject and object (SO entropy) because the entropy of objects and obliques (OX) is higher than 0.6 in most languages (with the exception of Mandarin, Coptic and Irish). The languages closer to the top have rich nominal morphology. For example, Lithuanian, an extremely conservative Indo-European language as far as the nominal declension system is concerned (Erhart 1987: 129), has the highest score on the vertical dimension. This suggests that high SO entropy is associated with the presence of case marking.

Indeed, there is a common view that rigid word order can 'step in' when there are no morphological cues that help to disambiguate between the main arguments, and that the availability of case morphology is associated with more flexible word order. This relationship is well-known in linguistics (e.g. Sapir 1921: 66; Jakobson 1971[1936]: 28). For example, Blake (2001: 15) considers word order 'an alternative to case marking' in languages like English, Indonesian, Thai and Vietnamese. Those languages in which the main arguments (A and O) are not distinguishable by means of case flagging or indexing will tend to have on average more rigid word order of the arguments than the languages in which A and O are distinguishable. Typological data in support of this idea have been provided by Siewierska (1998) and Sinnemäki (2008). Corpus evidence can be found in Futrell et al. (2015), who show that languages with flexible word order tend to have case marking. One can also find experimental evidence in Fedzechkina et al. (2016).

The corpus data allow us to formulate and test a more nuanced hypothesis than the previous accounts. The hypothesis can be formulated as follows: the amount of entropy in the order of S and O negatively correlates with the amount of formal overlap between S and O wordforms. In other words, the more confusable the forms on average, the more rigid the word order. It is more specific than the previous ideas because existence of case marking in a language does not guarantee that the forms are always distinctive. Many languages exhibit case syncretism in specific nominal forms or differential marking of subject and object, which depends on the semantic and pragmatic properties of the arguments (see the quantitative data in Levshina 2018: Ch. 6). Therefore, the categorical variable [+case] or [-case] may not be exact enough to represent the actual need for disambiguation.

To test the hypothesis, I computed for each language the number of individual nouns that were identical in the subject and object functions. I also compared the nouns in the function of an object and an oblique nominal phrase. Case marking with adpositions was taken into account on a par with case inflections, so that the Spanish form *a Juan* was just as different from *Juan* as the Russian accusative form *Ivan-a* is different from the nominative

form *Ivan*. Only non-plural forms were taken into account, i.e. those that were not marked as 'Plural' in the morphological properties slot. Two confusability indices were computed. The first one was the proportion of lemmata that have identical forms as subjects and objects in the total number of lemmata found both in the subject and object positions in a given language. The second one was the proportion of lemmata with the same forms in the object and oblique functions.

The analyses revealed very little overlap between object forms and oblique forms. The languages differ mostly in the degree to which they distinguish formally the subject from the object. This variation can be seen in Figure 3, where the proportion of formally identical subjects and objects (per lemma) is plotted against the entropy of subject and object in a given language.[6] Overall,

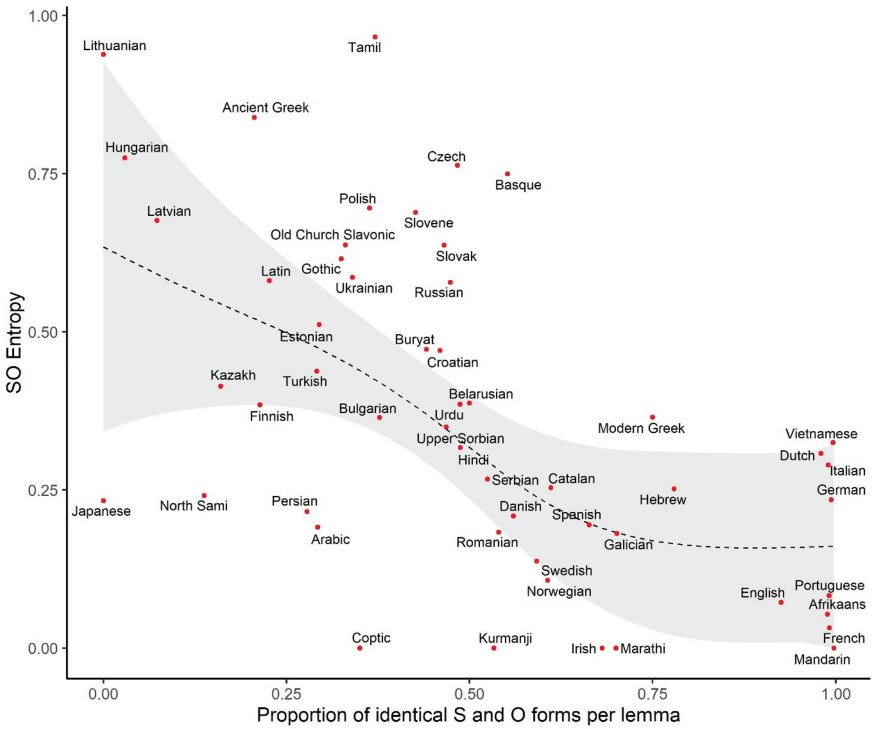

Figure 3: Negative correlation between SO entropy and the proportion of identical S and O forms.

6 Some comments are due. German has very many confusable forms because the determiners (articles, etc.) were not taken into account. English does not have a perfect confusability score

we see a negative correlation between the average confusability of the subject and object and the entropy of their order. A Generalized Additive Model with the genera as random intercepts reveals a non-linear negative effect displayed as the curved line in Figure 3 (p = 0.0003 for 2.8 estimated degrees of freedom).[7] The shaded area is the 95% confidence band. The model explains the data well (83.1% of deviance, i.e. variation, is explained; the adjusted R^2, which penalizes a model for having too many parameters, is 0.74). The model output can be accessed in the Supplementary Materials: *Regression Output*. Importantly, the correlation is between continuous variables, rather than binary features, such as [± case] or [± fixed word order]. This kind of relationship can only be tested on the basis of corpora.

Interestingly, the top right corner in Figure 3 contains no languages. This is the area with high confusability of Subject and Object, and with high variability of their order.[8] Compare that with the bottom left corner. There are no languages with zero confusability and zero entropy, but a few languages have low values on both dimensions (i.e. Japanese, North Sami, Persian, Arabic and Coptic). This suggests that the relationship is not only a correlation, but also to some extent a one-way implication: high confusability implies low entropy, but low confusability does not necessarily lead to high entropy. Similar ideas have been expressed on the level of languages as types (e.g. Kiparsky 1997; McFadden 2003).[9] This finding can be explained by diachronic processes: loss of case marking leads to word order fixation due to avoidance of ambiguity, as happened in Middle English (e.g. McFadden 2003), but development of case marking does not (necessarily) lead to more flexible word order. As Kiparsky put it (1997: 490),

> A language may lose its inflections, but it cannot 'lose its word order' in the same sense: it must go on putting one word after another, even when it does not grammatically exploit or constrain word order.

---

because it allows the use of some oblique case forms as core argument, e.g. *While Mary's dream was to become a scientist, Jane's was to marry a rich man*. Also, the names of companies, restaurants, basilicas, etc. (Papa John's, St Peter's) can contain the genitive. Judging from the corpus data, written Japanese has almost no confusable forms, although a substantial share of case markers can be omitted in casual conversations (e.g. Kurumada & Jaeger 2015).

7 A likelihood ratio test suggested that the families were not worth including as random effects.

8 I thank an anonymous reviewer for pointing this out.

9 On the level of types, Kiparsky formulated an exceptionless one-way implication: "lack of inflectional morphology implies fixed order of direct nominal arguments" (1997: 461). According to McFadden (2003: 301), the implication is not exceptionless, but represents a strong tendency. This fact and other evidence, such as intra-linguistic variation in the use of markers, lead him to conclude that the relationship should be explained by a performance-based account, i.e. the principles of language use, acquisition and change.

Obviously, more research is needed in order to gain a full picture, which should also include the role of agreement markers, intonation and other resources for disambiguation.

3.3 Case Study 1b. Entropy of individual word order patterns

This section discusses variation of the word order patterns within and across languages. Two types of average measures were computed. The average **intra-linguistic variability** of an individual (co-)dependency is represented by its mean entropy. In order to take into account the hierarchical structure of the UD corpora, where some genera are represented by many languages and some by only one, I averaged the entropy first within each genus, and then computed mean entropy scores across the genera.

The **cross-linguistic variability** of a specific word order pattern is represented by the standard deviations of proportions of the word orders across the UD genera. They are based on the average proportions of the dependent element being before the head (e.g. the average proportion of auxiliaries before the main verb in a genus), and of the SO and OX order. These proportions were first computed for each language, and then averaged for each genus. Finally, the standard deviations were computed that reflect the cross-linguistic variability of the proportions across the genera.

These two types of variability, intra-linguistic and cross-linguistic, are shown in Figure 4. The horizontal axis displays intra-linguistic variation of word order in the (co)-dependencies, measured as mean entropy. The patterns on the left-hand side have low entropy. This means that the order is fixed in most languages. For example, a language has either prepositions or postpositions. The right-hand side of the plot contains the word orders with greater intra-linguistic variability. In an individual language, the position of pronominal objects and obliques, adverbial clauses and adverbial modifiers of verbs, obliques with regard to objects and predicates tends to be the most flexible, whereas the position of complementizers, adverbial subordinators and adpositions is rigid in most languages. Various modifiers of nouns (adjectival and nominal modifiers, determiners, and attributive clauses) usually have limited variability, as well as auxiliaries, copulas, pronominal subjects, complement and subject clauses. Nominal subjects and numeral modifiers are more variable.

The vertical axis represents the cross-linguistic variability of the word order patterns specified by the labels. The numbers show the standard deviations of the proportions of the corresponding word orders. Adpositions, adjectival modifiers, subordinators, auxiliaries, objects, as well as attributive, complement and

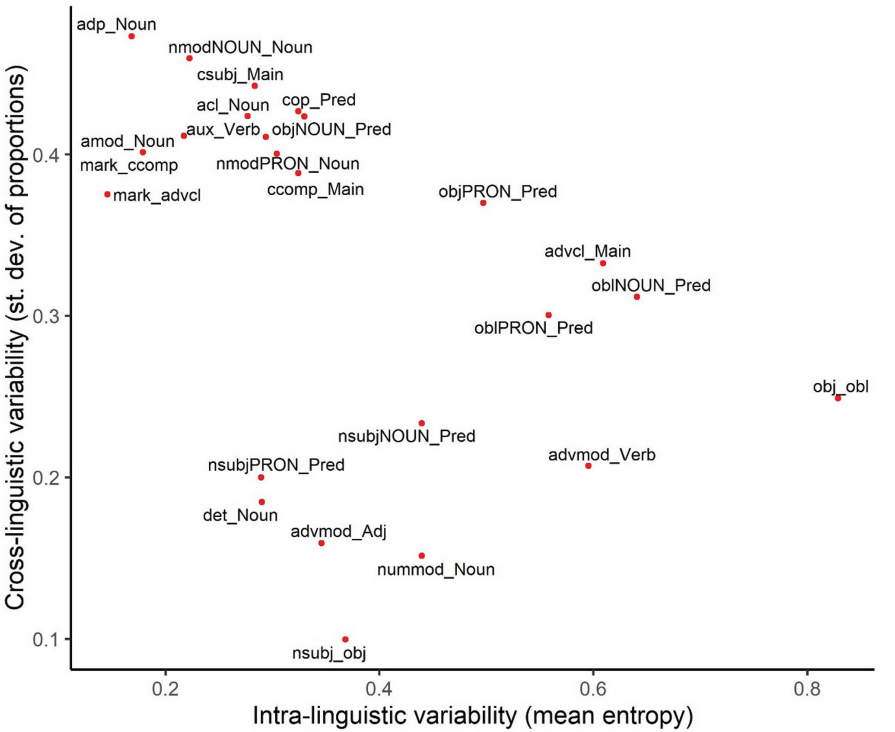

Figure 4: Intra-linguistic variability (horizontal axis) and cross-linguistic variability (vertical axis) of selected syntactic dependencies from the UD corpora.

subject clauses, display substantial cross-linguistic variation, as one can judge from their position at the top of the plot. In contrast, the languages agree with regard to the position of subjects, determiners, numerals and adverbial modifiers, which are located below.

This information has important implications for word order typology as a domain of inquiry. The patterns in the top left corner of the plot include the ones that have received the greatest attention in the literature: adpositions, nominal objects, adjectives and nominal modifiers (predominantly genitives). This figure explains why: these patterns exhibit high cross-linguistic variation, but low intra-linguistic variation, which makes them perfect candidates for typological investigations based on reference grammars. Although this bias is understandable from a practical point of view, it also means a considerable data reduction when one focuses only on those patterns (cf. Wälchli 2009).

The patterns in the bottom, on the left, may be less interesting for language classification because they exhibit less variation, but they are still important for

typological purposes because they allow us to formulate cross-linguistic generalizations. Most languages tend to put subjects before objects. This is not new (e.g. Hawkins 2014: 5). What is more interesting, perhaps, is the fact that the variation in the position of pronominal subjects with regard to verbs within a language is lower than the variation in the position of nominal subjects. This may have to do with the fact that nominal subjects can be used to introduce new information, as in the example *In the middle of the room stood a table*. In contrast, pronominal subjects are topical and given, and are used before the verb in topic-first languages (cf. Givón 1984: Section 6.5; see also Section 5.1). Also, the order of numerals, determiners and adverbial modifiers of adjectives is to some extent similar cross-linguistically. They tend to occur predominantly before their heads. However, a closer look reveals that this is an artefact of the UD language sample, as far as the order of numerals and determiners is concerned. According to Dryer (2013a, 2013b), numerals and demonstratives (a subclass of determiners) exhibit strong areal patterns. They are pre-nominal in Eurasia (where most of the corpora come from), but can be post-nominal in some other areas, such as Africa and Southeast Asia. The low intra-linguistic entropy of determiners, however, is quite informative. It can be explained by their high level of grammaticalization and loss of syntagmatic variability (cf. Lehmann 2015: Section 4.3.3; see also Section 5.1).

Finally, the obliques and adverbials are located on the right, which means that they exhibit high entropy in individual languages. This variability can be explained by their different functions in discourse. For example, adverbial clauses of condition tend to precede the main clause, while adverbial clauses of result and purpose usually follow the main clause (Diessel 2001). This is explained by universal cognitive and discourse-pragmatic factors, in particular, by iconicity of temporal order. Similar explanations can also be offered for obliques and adverbial modifiers of verbs. For instance, a closer look at English, Finnish and Russian obliques and adverbials reveals striking similarities between the languages. The adjuncts on the left from the predicate are often those which introduce causal links with the previous discourse (e.g. *therefore, thus*), mark the relative position of the statement in the rhetorical structure (e.g. *also*, *moreover, finally*), express the epistemic stance towards the entire preposition (e.g. *possibly*, *reportedly, certainly*) or the speaker's emotional attitude (e.g. *hopefully*). The adjuncts on the right often have the directional meaning (*go* + *abroad, to the market*), expressing the potential outcome of the action. These word order patterns have received less attention in typology than the others, but they are also important because they allow us to formulate probabilistic universals based on the general discourse-pragmatic principles.

To summarize, although word order typology has mainly focused on the patterns with low intra-linguistic and high cross-linguistic variability, we also need to explore the other patterns. It is not unreasonable to expect that the high-entropy patterns can be particularly valuable for providing a window into the universal cognitive and communicative biases.

4 Case Study 2. Word order entropy at the lexically specific level

4.1 Aims of this section

In the previous section it was mentioned that high entropy of particular word order patterns may be explained by their diverse functions in discourse. This functional diversity can be approximated by taking into account the individual wordforms. For example, *on Monday* is typically a temporal adjunct, whereas *therefore* is a causal adverbial modifier. To what extent will the variability of the word order patterns change if we compute the entropy on the lexical level? Will we see a completely different distribution of languages and individual dependencies? The goal of this section is to answer these questions.

4.2 Data

I took large news corpora of eleven languages (100,000 sentences in each) from the Leipzig Corpora Collection (Goldhahn et al. 2012) and parsed them with the UDPipe software (Straka & Straková 2017), implemented in the R package *udpipe* (Wijffels 2018), which provides tokenization, lemmatization, part-of-speech annotation and syntactic parsing. The languages were Arabic, Basque, English, Finnish, Hindi, Indonesian, Irish, Mandarin, Russian, Tamil and Turkish. All of them belong to different genera. The reason for using the additional larger corpora is that one needs high frequencies in order to obtain estimates of entropy for individual lexemes.

From each of the corpora, I extracted the head-dependent patterns, using the same approach as outlined in Section 2, plus the information about the wordform of the dependent element, e.g. *big* + Noun, where *big* is an adjectival modifier *amod*, which modifies a noun. These types of patterns will be called here lexically specific dependencies. Only the dependencies in the simple clause were considered. The dependencies related to complex clauses were excluded because it is

difficult to imagine that the lexical predicate of the subordinate clause can be highly relevant for the semantic and pragmatic function of the clause.

The wordforms of the dependent elements also included adpositions with the function 'case' because these functional units play an important role in determining the function of the lexeme in a sentence. For example, *John*, *of John*, *John's*, *to John*, etc. were all considered different wordforms and were treated separately, similar to Russian *Ivan*, *Ivan-a* 'I.-ACC/GEN', *Ivan-u* 'Ivan-DAT', *k Ivan-u* 'to Ivan-DAT', etc. For each lexically specific dependency I computed the frequencies of occurrence of this wordform before and after the head (e.g. the form *table* as a nominal subject before and after the predicate). Only the wordforms with a frequency above 20 were considered. On the basis of those frequencies, I computed the measure of entropy for each individual lexically specific dependency using the method described in Section 2.2. In addition, the entropy scores for each syntactic dependency without the lexical information were computed, as was done in the case studies discussed in Section 3.

4.3 Results

Figure 5 displays the mean entropy measures for the eleven languages, based on the sixteen dependencies in simple clauses. The horizontal axis shows the mean entropies at the level of abstract dependencies without lexical information. The vertical axis displays the entropies based on the lexically specific dependencies. The entropy scores were first computed for each wordform, then averaged for each dependency (e.g. *det_Noun*, *amod_Noun*, etc.) and finally averaged across the languages.

Importantly, the entropies based on the abstract dependencies are higher than the ones based on the lexically specific ones, as the reader can infer from the values on the axes. This means that some part of the variation can be explained by the variation between individual wordforms. At the same time, there is a strong correlation between the two sets of scores: $r = 0.978133$, $p < 0.0001$. This means that the entropy-based classification of the languages remains essentially the same, regardless of the level of granularity.

The languages in which the averaged dependencies show the largest differences are Finnish (the difference between the mean abstract dependency entropy and mean lexically informed entropy is 0.24) and Russian (0.23). The smallest differences are observed, as one would expect, in the low-entropy languages Hindi and Tamil (0.04 both).

As for the individual dependencies, consider Figure 6. Again, the horizontal axis shows the mean entropies at the level of abstract dependencies, whereas

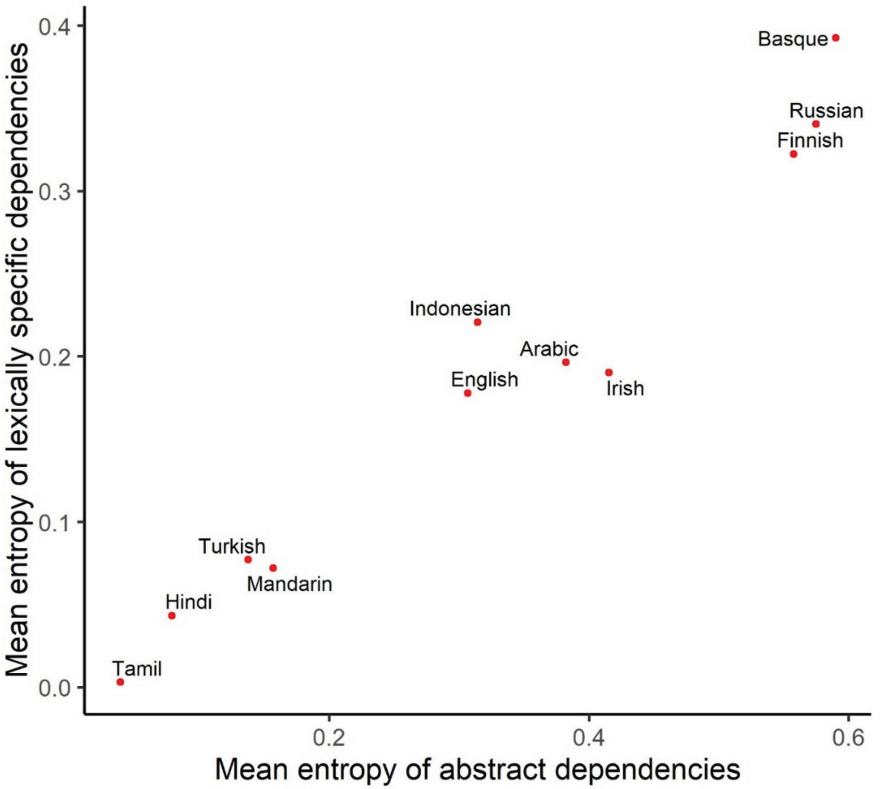

Figure 5: Mean head-dependent entropy in a language. Horizontal axis: mean head-dependent entropy without lexical information; vertical axis: mean head-dependent entropy with lexical information.

the vertical axis displays the mean entropies at the level of wordforms, averaged across the languages.

Again, we find that the entropies are strongly correlated: $r = 0.91$, $p < 0.0001$. Adpositions, auxiliaries, adjectives and nominal modifiers of other nouns have low entropies regardless of the method. The highest entropy scores on both axes are observed for obliques and for adverbial modifiers of verbs, similar to the results reported in Section 3.3. However, the correspondence is not perfect. On average, the dependencies that show the greatest differences between the entropies with and without lexical information are *advmod_Verb* (the difference is 0.24), *det_Noun* (0.23) and *nummod_Noun* (0.21). These are the dependencies where the greatest functional specialization is expected. This is not surprising. Some examples of adverbial modifiers of verbs were given in the previous section.

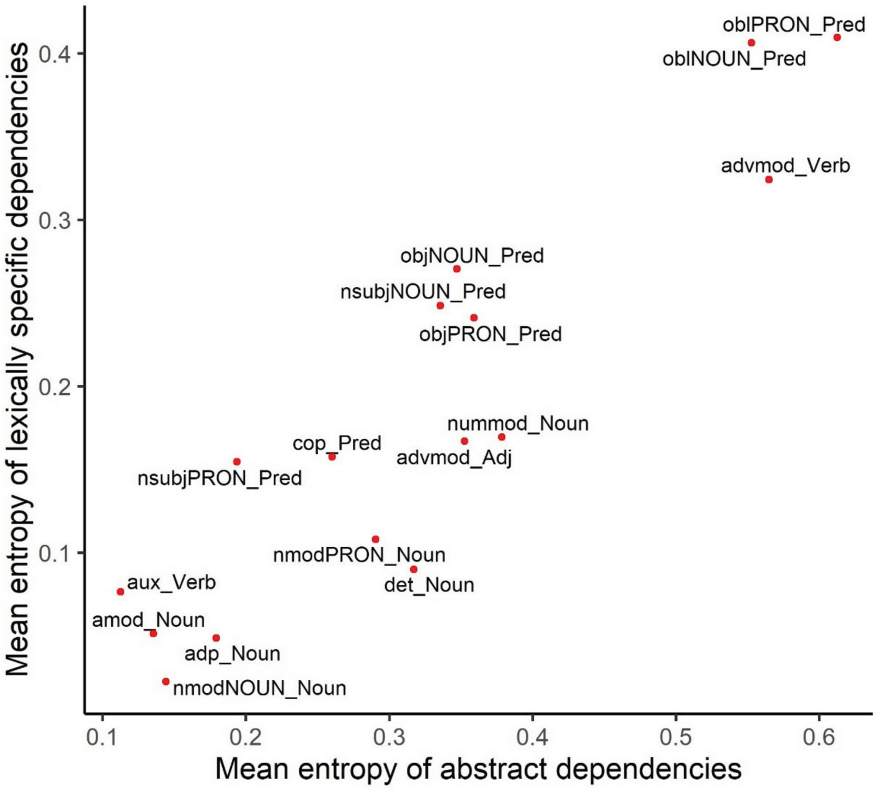

Figure 6: Mean entropies of dependencies averaged across eleven languages. Horizontal axis: mean head-dependent entropy without lexical information; vertical axis: mean head-dependent entropy with lexical information.

Their position often depends on their function. As for determiners, they constitute a very heterogeneous category: demonstrative, possessive, negative and indefinite pronouns, articles, etc. Their positions with regard to the head noun can vary. For example, in Irish, the definite article and possessive pronouns precede the noun, e.g. *an* 'the' + *fear* 'man' and *do* 'your' + *chara* 'friend', whereas demonstrative pronouns follow it, e.g. *an* 'the' + *bhean* 'woman' + *seo* 'this'. As for numeral modifiers, they can also perform diverse functions. For example, they can specify the quantity, e.g. 5 books, but appear in dates (e.g. the year 2019), telephone numbers (e.g. the emergency number 112) or addresses (e.g. 10 Downing Street). The annotation of these instances in a corpus is not trivial. The dependencies with the smallest differences are *aux_Verb* and *nsubjPRON_Pred* (both 0.04). They also display low entropy on average.

If we look at specific languages and individual dependencies, the differences between the approaches can be quite striking, however. For instance, English nominal modifiers are often genitive constructions, which occur both before and after the head noun, e.g. *the emperor's new dress* and *the new dress of the emperor*. Their position is strictly determined by the type of the genitive construction: the Saxon genitive always precedes the noun, and the Norman genitive always follows it. This is why English nominal modifiers have much lower entropy when the wordforms are taken into account than when the wordforms are ignored: 0.024 vs. 0.524.

The highest discrepancy between abstract and lexically-specific entropy is observed for Mandarin adpositions (0.99 vs. 0.01). They can be postpositions, prepositions, or elements of circumpositions, but their place with regard to the head noun depends very strictly on the lexeme. For example, the locative marker *zài* 'at, in' and the directional marker *dào* 'to' are always used before the noun, whereas the possessive markers *de* and *zhī* are used after the noun. Next come numeral modifiers in Basque (0.93 vs. 0.19), where the numeral *bat* 'one' in different forms, unlike all other cardinal numerals, takes the postnominal position (Hualde & de Urbina 2003: 118). This is a highly frequent numeral, which is also the indefinite article with the meaning 'a certain'. It is difficult to distinguish between the functions in the corpus data. Similarly, Indonesian contains different types of determiners (0.87 vs. 0.29). Quantifiers like *semua* 'all' and *beberapa* 'some' are used before the head noun, while possessive and demonstrative pronouns follow the head noun. The smallest differences are observed in those cases where the entropy based on abstract dependencies is close to 0.

There are also some cases where the lexically specific entropy is in fact higher than the entropy based on the abstract dependencies, although the difference is usually very small. For example, English pronominal subjects have on average slightly higher lexically specific entropy due to the pronouns *something* and *nothing*, which are often used in the construction with the presentative adverb *there*, e.g. *there is something in the air tonight*. However, this variation nearly disappears when the entropy is computed at the level of the abstract dependency because of the high-frequency personal pronouns, which exhibit almost no variability. Such discrepancies are also observed in the other pronominal arguments (subjects and modifiers of nouns) and in adpositions in some languages. Thus, the presence of high-frequency exemplars with low variability tends to decrease the entropy on the level of abstract dependencies, whereas the presence of functional subcategories is responsible for its increase.

From all this follows that the lexically 'naïve' estimates of entropy give a reliable idea of the general magnitude of word order variation in a language,

although they tend to overestimate the variation. This works only when many dependencies are taken, however. If one performs a classification that involves a small selection of word order patterns, he or she would be well advised to check the subcategories of dependencies and individual lexemes.

5 Case Study 3. Functional explanations of word order entropy

5.1 Causal factors that influence word order variation

If we want to explain the variation of word order patterns, several potential factors come to mind. One of them has already been discussed. Low word order entropy helps to disambiguate potentially confusable arguments. One should also mention the typical information-theoretic role of constituents. If a constituent frequently performs a particular role, which is associated with a specific position in discourse, then the position of the constituent will also be fixed. For example, constituents expressing topical, non-surprising and given information tend to have a fixed position in discourse. They occur either before the verb, close to the left, as in Mandarin or Russian, or after the verb, more to the right, as in Ute or Early Biblical Hebrew (Givón 1984: Section 6.5). Since subjects are often associated with givenness and topicality, especially transitive ones (Du Bois et al. 2003; Lambrecht 1994), they are expected to have a fixed position. We have seen already that pronominal subjects exhibit little cross-linguistic and intra-linguistic variation in their position with regard to the predicate (see Section 3.3). This can be explained by their topical role.

Another important factor seems to be grammaticalization, which limits the syntagmatic variability of linguistic units, i.e. the ease with which a word can be shifted around in context (Lehmann 2015: 167). For example, full verbs are more flexible than auxiliaries with regard to the verbal phrase with which they combine. Compare the verb 'have' in Classical Latin and Italian. In Classical Latin, the parts of the construction *epistulam scriptam habeo* 'I have a letter written' could appear in any order. Compare this with Italian *ho scritto una lettera* 'I have written a letter', where the auxiliary always precedes the main verb (Lehmann 2015: 168). From this follows that more grammaticalized functional elements should have more rigid positions with regard to their heads (i.e. nouns for adpositions, main verbs for auxiliaries) than less grammaticalized ones (e.g. nominal arguments or adjuncts).

However, grammaticalization itself is only an umbrella term for several different processes; it is not a functional pressure on its own. So, why are more grammaticalized units more fixed positionally? This can be explained by the fact that more predictable units have higher chances of grammaticalization to begin with. Units are more predictable when they often co-occur together with other units and become elements of highly entrenched chunks (cf. Bybee 2010). If these lexically specific chunks overlap with other chunks (e.g. *my car, my book, her book, her phone, this phone,* etc.), the pattern will be schematized as an abstract construction (e.g. [Determiner + Noun]), which can serve as the basis for emergence of syntactic constituents (Bybee 2002). Therefore, in order to obtain an abstract syntactic dependency with low entropy, one needs high frequency of two types: high token frequency of lexically specific chunks, and high type frequency of those chunks. On the other hand, grammaticalization (due to such processes as obligatorification and semantic bleaching) also boosts the frequency of the corresponding chunks and more abstract schemata, which leads to further increase in the relative frequency of the dominant word order and entropy reduction.

Next, one should mention processing constraints. In particular, the length of constituents has been shown to play an important role in word order. The preference to put long phrases after short ones was observed already in the classical rhetorical tradition (cf. Behaghel 1909/1910: 137–138). Behaghel (1909/1910) provided a first systematic account based on texts in Greek, Latin and German and called this preference *Das Gesetz der wachsenden Glieder* (the law of the growing elements). According to Hawkins (1994, 2014), human parsers prefer short constituent recognition domains. In a nutshell, these domains are minimized when shorter constituents, which are faster to process, are placed closer to the head than longer ones. This principle is known as "Minimize Domains", or Early Immediate Constituents. Another principle is "Maximize Online Processing", which helps to avoid garden path effects and delays in online property assignment. All this can be important for word order variability. For instance, VO languages tend to minimize domains and maximize online processing by putting complement clauses after the predicate. At the same time, the position of shorter dependent elements (e.g. one-word adjectives) is less crucial for domain minimization (Hawkins 2014: 101). This may affect the potential for variability of the corresponding word order patterns.

These factors and principles can be in conflict. For instance, the position of short elements is less important for the principle "Minimize Domains". This will lead to high variability of their position before or after the head in comparison with long constituents. At the same time, short elements can also be highly frequent and grammaticalized, which means that their position will be more

fixed. As was shown in Section 3.3, short functional elements (e,g. adpositions, auxiliaries and subordinators) usually have a rigid position in a language, as well as long clauses (with the exception of adverbial clauses). The position of mid-length constituents is usually more flexible. In the remaining part of this section, I will present a case study where these opposing factors are tested on corpus data.

5.2 Data and variables

In order to test the ideas discussed in the previous section, I used the data with the head-dependent patterns in the UD corpora (see Section 3.1 for more details). The entropy of a specific word order in an individual UD corpus was treated as the response variable. There were three categorical predictors, which are described below.

The first predictor described the functions of the dependent elements, which were classified into four categories:

- function elements (adpositions, subordinators, auxiliaries, copulas, determiners);
- core arguments and functionally similar clauses (subjects, objects, complement and subject clauses);
- obliques and adverbials (oblique nominal and pronominal phrases, adverbial modifiers, adverbial clauses);[10]
- modifiers of nouns and adjectives in a nominal phrase, with the exception of determiners (adjectival modifiers, attributive clauses, numerals and nominal modifiers of nouns).

Following the considerations discussed in the previous section, I expected function words to exhibit the lowest entropy due to their high level of entrenchment in the combination with their heads. I also expected adverbials and non-argument obliques to have the highest entropy, due to their multifunctionality.

Heaviness was operationalized in a binary way, whether the dependent element was a clause or not. This has to do with the fact that the lengths in orthographic words are difficult to compare cross-linguistically because the word is a problematic comparative concept (see the discussion in Section 6). At the same time, it is uncontroversial that clausal constituents represent the heaviest elements cross-linguistically. Their position is determined by the general processing principles (Hawkins 1994, Hawkins 2014), which were mentioned above.

10 In the UD corpora, obliques include both adjuncts and non-core arguments.

Importantly, the previous studies show that the effect of length depends on whether the language is predominantly OV or VO. In order to take that into account, I created three categories, based on the proportion of OV in the language. The distribution is strongly bimodal, so it was useful to represent it as a categorical variable. The first category was "VO", with the proportions of nominal objects followed by predicates ranging from 0 to 20%. The second category was "Flexible", with the proportions from 20 to 80%. Note that this intermediate category was biased towards the languages with a mild preference towards VO. The third category was "OV", with the proportions from 80 to 100%.

5.3 Mixed-effects regression analysis

This section reports the main results of fitting the linear mixed model with the variables discussed in the previous section.[11] I tested the pairwise interactions between the function and the OV/VO order, and between heaviness and the OV/VO order, in order to check whether the effects of the function and length vary in OV and VO languages.[12] These interactions are highly significant. The language, genus and dependency type were tested as random intercepts. Only the language and dependency type proved to be useful, according to the likelihood ratio tests. The table of coefficients and other information are provided in the Supplementary Materials: *Regression Output*.

Figure 7 shows the effect of the functional type of dependencies on the entropy for OV, VO and flexible word order. Not surprisingly, the flexible languages have the highest entropy values across core arguments, adverbials/obliques and nominal phrase elements, whereas the OV languages have the lowest values. Adverbials/obliques have the highest entropy, whereas function words have the lowest entropy, also in the OV languages. In the flexible and VO languages, we observe the following order: adverbials/obliques with the highest

11 The linear model ignores the fact that entropy of a binary variable ranges from 0 to 1. Unfortunately, there is no straightforward way of modeling such data. A beta regression does not allow values 0 and 1. In order to fit a beta model, I modified the data by replacing 0 and 1 with 0.001 and 0.999, respectively, and fitted a beta model with the R package glmmTMB. The results are very similar to the ones reported below. The correlation between the fitted and actual values of entropy is 0.52, which is identical to a linear model fitted on the same dataset.

12 In addition, an interaction between the function and heaviness was tested on the data without function elements (since they cannot be clauses). The interaction was not significant ($p = 0.49$).

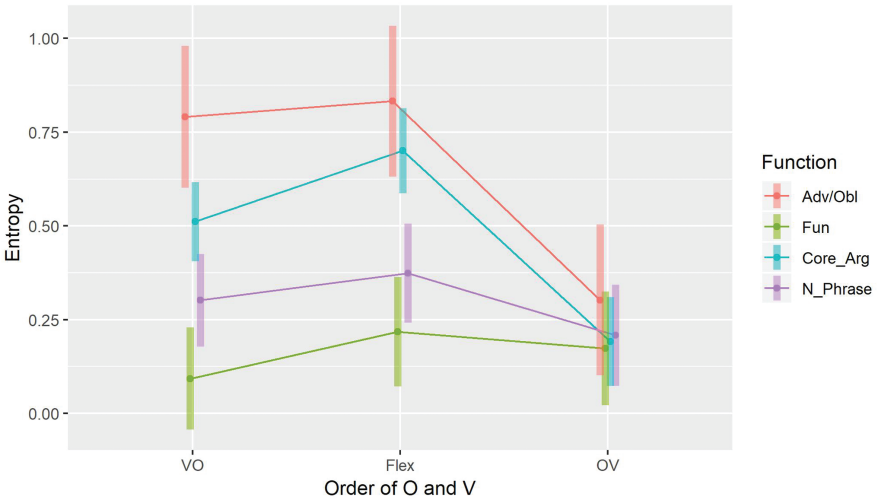

Figure 7: Interaction between the function type of dependencies and the order of O and V. The bars represent 95% confidence intervals.

entropy are followed by core arguments, then by nominal phrase elements and finally by function words.

A post-hoc analysis of all pairs of estimates (see the Supplementary Materials: *Regression Output*) shows that there are no statistically significant differences between the functions in the OV languages. In the VO and flexible languages, function words are significantly different from adverbials/obliques and core arguments, but not from NP elements. Core arguments, adverbials/obliques and NP elements are significantly more variable in the VO and flexible languages than in the OV languages. There are no differences between function words in the three language types.

Next, let us interpret the interaction between the heaviness (clause vs. non-clause) and the order of object and predicate, which is displayed in Figure 8. The plot suggests that the heaviness effect is observed only in the VO and flexible languages. The clauses are indeed less variable than non-clauses, although they still exhibit substantial variation. In the OV languages, we observe a reversed order: the clauses are slightly more variable than non-clauses, but the difference is very small. The post-hoc tests indicate that the differences between clauses and non-clauses are not statistically significant. At the same time, the interaction in general is statistically significant and should therefore be kept in the model. This means that the effect of heaviness indeed varies across the languages with different OV/VO orders.

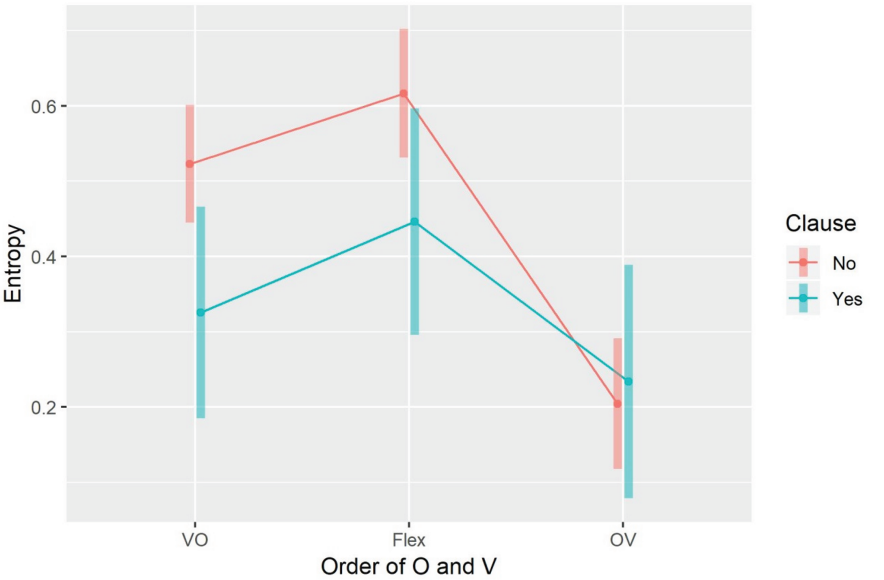

Figure 8: Interaction between heaviness and the order of O and V. The bars represent 95% confidence intervals.

5.4 Interpretation of the results

We can draw the following conclusions. First, not surprisingly, the languages without a clear preference for OV or VO also exhibit the highest entropy in all conditions. The predominantly OV languages display the lowest entropy.

Second, function words have the lowest entropy in all languages, and adverbials/obliques have the highest, although the differences are statistically significant only in the VO and flexible languages. Therefore, entrenchment due to high frequency can make a difference, provided there is enough variability in a language. The adverbials/obliques have the highest entropy, most probably, due to their multifunctionality.

Third, there is a significant difference in the heaviness effect on entropy between the OV languages and the VO and flexible languages. There is a tendency in the latter to have lower entropy of clauses in comparison with non-clauses, whereas the OV languages display hardly any difference. This supports the processing-related explanation, which was mentioned in Section 5.1. According to Hawkins (1994, 2014), object clauses with the subordinator in the initial position, as is the case in VO languages, tend to follow the main clause. This

order makes the processing optimal based on both principles, "Minimize Domains", and "Maximize Online Processing". As for OV languages, these principles are in conflict. Therefore, there is no perfect solution, although typological data suggest that the postverbal position is the more popular one (Schmidtke-Bode & Diessel 2017). The same logic applies to relative clauses (Hawkins 2014: Ch. 7), which are usually postnominal in VO languages, but can be either postnominal or prenominal in OV languages.

Importantly, Hawkins' principles were formulated on the level of languages and their preferred strategies. The regression model shows that the differences in the effect of heaviness between the VO and VO languages are discernible also at the level of usage tokens and probabilistic tendencies.[13]

6 Conclusions and perspectives

The aim of the present paper was to show the importance of token frequencies for the main tasks of typological research: language classification, as well as identification and explanation of cross-linguistic generalizations. Case Study 1a reported a classification of languages based on their average entropy scores. The data show that morphologically rich European VO and OV languages (e.g. Basque, Estonian and Slovak) tend to have high entropy of head-dependent orders. Notably, the ancient Indo-European languages (Ancient Greek, Gothic, Latin, and Old Church Slavonic) have high word order variation. Their more analytical descendants and relatives (with the exception of some Slavic languages) exhibit lower entropy. Syntheticity, which is usually accompanied by abundance of grammatical markers, seems to be a necessary, but not sufficient condition for high entropy. This is so because morphologically rich Asian OV languages have a very fixed head–dependent order. It seems that there are no languages with very rigid head–dependent orders and very rigid orders of co-dependent subjects, objects and obliques with regard to one another. This may be due to the necessity to have some means for the management of information flow in discourse.

The entropy of the order of Subject and Object correlates significantly with the proportion of distinct forms of nominal subjects and objects in a language.

13 This does not mean, of course, that there are no heaviness effects in usage in OV languages, but these effects manifest themselves in the order of co-dependent elements, e.g. objects and obliques (Hawkins 2014: 96–98), rather than in the order of heads and dependents, which is what is analyzed here.

This supports the idea that low word order entropy has a disambiguating function. This idea has been around since time immemorial, but the regression model presented in Section 3.2 adds a new flavor to this old idea. Namely, it shows that the correlation is observed when the formal distinctness and word order variability are treated as probabilistic, gradable parameters. This result demonstrates how amazingly subtle and fine-grained grammar adjustments can be for the purposes of efficient communication.

As for the order of heads and dependents, we observe a correlation between syntheticity and entropy only for VO and flexible languages. As already mentioned, the predominantly OV languages, which are far from being isolating, exhibit very low entropy of heads and dependents. This can be explained by the 'tight fit' of arguments for the purposes of early and correct recognition of the constituents, which is required when the verb appears only in the end of a clause. However, this explanation is not unproblematic, since we do find some variability in the order of core arguments and obliques in the verb-final languages, which would contradict this explanation. More research is needed to solve this puzzle.

Case Study 1b focused on the intra-linguistic and cross-linguistic variability of individual dependencies. It made very clear that the traditional word order typology is based on those dependencies that exhibit low intra-linguistic variability and high cross-linguistic variation, e.g. the order of adpositions and nouns, nominal objects and verbs, adjectives and nouns. This does not mean that the other types are irrelevant, however. The other dependencies represent valuable material for universals related to processing and organization of information flow in discourse, e.g. the pronominal subject in most languages tends to be in front of the predicate.

The next question was whether the semantic and pragmatic functions of different word classes within the dependencies may explain word order variability. This was tested in Case Study 2 on a sample of eleven languages represented by large corpora of online news. To some extent, this is indeed so. The average entropy is lower when the individual wordforms are taken into account, especially for some dependencies, such as determiners, numeral modifiers and adverbial modifiers. The reason is the functional diversity of those categories. However, there still remains a large amount of variation. Overall, there is a strong positive correlation between the lexically specific and non-specific measures of entropy for the languages and the individual dependencies.

The paper has also discussed the universal factors that influence the variability of word order. This was the focus of Case Study 3. In contrast to the previous explanations, which employed single parameters, such as configurationality, branching direction or head-dependent order – which may be

useful as descriptive measures – I believe that word order directionality and variability can be explained directly by general cognitive and communicative principles, which determine language use. No additional theoretical layers are required. We can name several usage-based factors. First, language users tend to optimize processing by minimizing domains and avoiding ambiguity. Here we should also mention the fundamental principles of organizing the flow of information in discourse, in particular, how old information is linked with new. Another important factor is entrenchment of word combinations due to high frequency, which leads to grammaticalization of the units involved.

These factors were investigated in a regression model based on the UD corpora, which modeled the effects of length, function and predominant order of verb and object on word order entropy of head-dependent patterns. The model revealed the following cline:

(2)
Function elements < Modifiers of a noun < Core Arguments < Adverbials/Obliques

where the function elements have the lowest entropy, and oblique phrases and adverbials have the highest entropy in the languages where there is sufficient word order variation. Function words are the least flexible because grammaticalization both implies and triggers high co-occurrence frequency of the head and the dependent in a specific order. Adjuncts are the most variable because of their multifunctionality, which can be explained by their retrospective or prospective orientation in discourse. In addition, there is an effect of heaviness, which is related to optimization of processing. Clauses, which are the heaviest constituents, tend to have less freedom than lighter elements and usually occur after the head (the verb or noun in the main clause) in the VO and flexible languages. There is no strong effect in the OV languages, however.

Thus, the token-based approach provides us with new criteria for typological classification, ideas for cross-linguistic generalizations, an opportunity for analyses at different levels of lexical granularity, and a testing ground for universal functional constraints. We can also reformulate some of the existing type-based generalizations at the level of tokens. This helps us to understand better the interaction between language use and language structure, which can manifest itself either as 'soft' or 'hard' constraints in individual languages (Bresnan et al. 2001).

In addition to these advantages, there are also other benefits. In particular, a linguist can use mixed types of languages and linguistic categories, without forcing them into a Procrustean bed. The researcher is no longer required to group languages into types, using some arbitrary cut-off points. Moreover, the

token-based approach will force typologists to formulate their generalizations and comparative concepts more precisely, in a way that enables these to be tested on corpus data. The task for future research is to include various semantic, pragmatic and structural variables in order to induce universal correlations and implicational relationships directly from usage events, in line with Multivariate and Distributional Typology (Bickel 2010, Bickel 2015).

There are a few challenges that should be mentioned, as well. First of all, we need corpora representing a large number of diverse languages with rich and maximally uniform linguistic annotation, including semantic and pragmatic variables. At present, the largest collection of multilingual texts is the parallel corpus of Bible translations (Mayer & Cysouw 2014), which currently contains more than 1850 translations that represent more than 1400 languages (i.e. unique ISO-codes). However, they do not have any linguistic annotation, with the exception of information about verse alignment (see the Supplementary Materials: *Multilingual Corpora*). Similarly, the Leipzig Corpora Collection only contains sentences. In contrast, the UD corpora have a bias towards the Indo-European and Eurasian languages, but contain a lot of useful linguistic information, such as syntactic functions and morphological categories. Multi-CAST (Haig & Schnell 2016a) has semantic annotation of NPs and their syntactic roles and even information about zero elements in discourse, but covers only a few languages at the moment.

The examination of the lexically specific dependencies has also revealed that some word order variability is due to different subclasses of words, which have different positional preferences. For example, different types of determiners and adverbial modifiers exhibit different preferences in some languages in the sample. This makes one wonder, what level of granularity is optimal for cross-linguistic comparisons. A possible solution could be as follows: the researcher should choose the level of abstraction at which the differences between this level and a more specific one are minimized across the languages in the sample.

There are some conceptual issues, as well. Traditionally, the fundamental unit of analysis in corpus linguistics is the word. One speaks of key words, key key words, word frequency lists, etc. The main criterion for delimiting words is orthography. However, since this criterion is problematic (Haspelmath 2011), analyses based on orthographic words can be misleading. In fact, there have been some practical attempts to take this into account. For example, some UD corpora provide information about multiword expressions (MWE), such as *New York* and *in spite of*, both analyzed as a whole and at the level of individual components. Some corpora offer a two-tier analysis of clitics, as one word and as two units, e.g. Italian *inquinandolo* = *inquinando* “polluting” + *lo* “it”, leaving the decision to the linguist which solution to use (or both).

Still, in spite of these practical and theoretical challenges, we can be optimistic about the future of the token-based approach. The number of diverse corpora and corpus tools is increasing rapidly. Hopefully, cooperation between typologists and corpus linguists will result in new insights and creative solutions of conceptual problems.

Acknowledgements: This study is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 670985). The author is very grateful to the editor and anonymous reviewers for their insightful suggestions, constructive criticisms, and patience.

Appendix. Comparison of different text types

From the Leipzig Corpora Collection, I took the corpora that represent eight languages: Arabic, English, Finnish, Hindi, Indonesian, Russian, Turkish and Vietnamese. Three types of texts were selected: Wikipedia articles, online news and miscellaneous web content. Samples of 10,000 sentences of each text type in every language were annotated using the UDPipe software (see Section 4.2). Next, the word order frequencies were extracted using the general procedure described above, and the proportions of different word orders were computed, e.g. the proportion of adjectival modifiers followed by nouns in the total occurrences of adjectival modifiers before or after the head noun. (Co-)Dependencies with frequency less than 20 (usually due to the small size of some corpora) were excluded from the subsequent analyses.

The correlations between the word order proportions in the three text types within each language were very high, ranging from 0.85 to 0.995 (Pearson's product-moment correlation). I also computed the average entropy measures across the dependencies and co-dependencies for each subcorpus. The dot chart in Figure 9 displays the mean entropy scores. The results show that there is little difference. The greatest discrepancy is in Russian, where Wikipedia shows lower word order entropies on average than the two other text types, but the difference is still relatively modest.

These results support the conclusion made by Liu (2010), who argues on the basis of empirical data that genre differences are not strong enough to influence the conclusions about dominant word order.

Some languages in the UD corpora are also represented by texts from different historical periods. For example, Latin combines numerous sources,

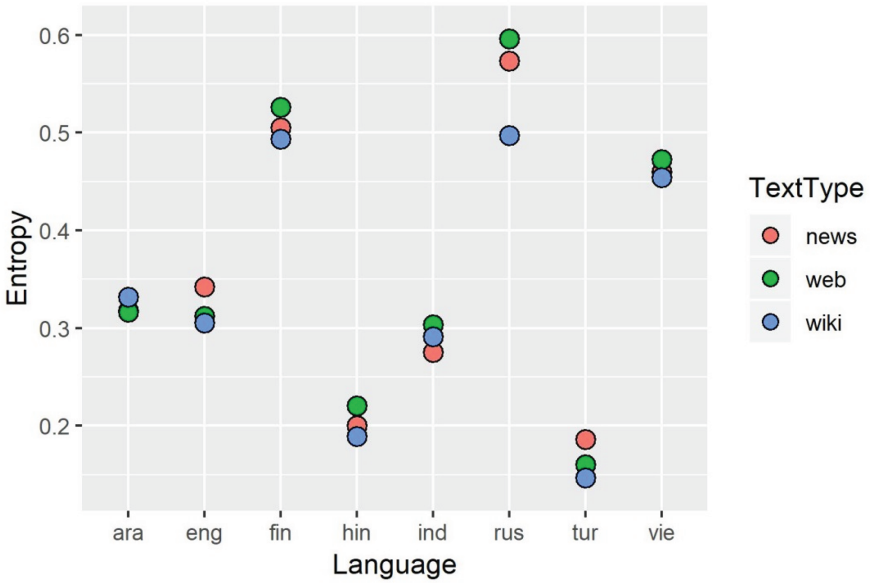

Figure 9: Mean entropy scores for different text types in eight languages (data from the Leipzig Corpora Collection).

which include the Vulgate New Testament translations, Caesar, Cicero, Thomas Aquinas, etc. High word order entropy in Latin (see Section 3.2) may be due to the fact that different authors who lived in different periods might have had different word order preferences. In order to check the individual text variation in Latin, I took several Classical and Medieval Latin texts from the online publications on Wikisource.[14] The texts and their mean entropies are shown in Table 2. The numbers are high, which suggests that the high entropy in Latin is not an artefact of combining different texts from different time periods. In particular, all texts exhibited variation in the position of copulas, adjectival modifiers, nominal modifiers and nominal objects with regard to their heads. Especially variable was the relative position of the co-dependents (SO/OS and OX/XO, where X stands for the oblique nominal phrase), as one can see from the entropies that are very close to 1.

14 https://la.wikisource.org/wiki/Pagina_prima, Accessed on 2018–11–12.

Table 2: Entropy in Classical and Medieval Latin texts. The number in brackets specifies the entropy without those dependencies that exhibited low frequencies in at least one of the texts and were discarded.

Author	Text	Period	Mean Entropy Head-Dep	Mean Entropy SO, OX
Caesar	Commentarii de Bello Gallico	58–49 BC	0.61 (0.66)	0.97
Cicero	De Fato	44 BC	0.77 (0.77)	0.95
Cicero	Cato Maior de Senectute	44 BC	0.76 (0.78)	0.97
Tacitus	De origine et situ Germanorum (Germania)	appr. 98 AD	0.73 (0.75)	0.95
Tacitus	Dialogus de oratoribus	appr. 102 AD	0.75 (0.77)	0.99
Seneca	De Clementia	55–56 AD	0.78 (0.80)	0.96
St Augustine	Confessiones	397–400 AD	0.76 (0.86)	0.99
Venerable Bede	Historia ecclesiastica gentis Anglorum (a fragment)	appr. 731 AD	0.70 (0.77)	0.98
Thomas Aquinas	Summa Theologiae (a fragment)	1265–1274	0.71 (0.81)	0.90
Dante Alighieri	De Monarchia	1312–1313	0.73 (0.81)	0.94

References

Behaghel, Otto. 1909/1910. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25. 110–142.

Bentz, Christian & Ramon Ferrer-i-Cancho. 2015. Zipf's law of abbreviation as a language universal. Capturing Phylogenetic Algorithms for Linguistics. Lorentz Center Workshop, Leiden, October 2015.

Bickel, Balthasar. 2010. Capturing particulars and universals in clause linkage: A multivariate analysis. In Isabelle Bril (ed.), *Clause-hierarchy and clause-linking: The syntax and pragmatics interface*, 51–101. Amsterdam: Benjamins.

Bickel, Balthasar. 2015. Distributional typology: Statistical inquiries into the dynamics of linguistic diversity. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*, 2nd edn., 901–923. Oxford: Oxford University Press.

Blake, Barry J. 2001. *Case*. Cambridge: Cambridge University Press.

Bresnan, Joan, Shipra Dingare & Christopher D. Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG01 Conference*, 13–32. Stanford: CSLI publications.

Bybee, Joan. 2002. Sequentiality as the basis of constituent structure. In Talmy Givón & Bertram F. Malle (eds.), *The evolution of language out of Pre-Language*, 109–132. Amsterdam/Philadelphia: John Benjamins.

Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris.

Croft, William. 2003. *Typology and universals*, 2nd edn. Cambridge: Cambridge University Press.

Diessel, Holger. 2001. The ordering distribution of main and subordinate clauses: A typological study. *Language* 77(3). 433–455.

Diessel, Holger. 2017. Usage-based linguistics. In Mark Aronoff (ed.), *Oxford research encyclopedia of linguistics*. New York: Oxford University Press. http://linguistics.oxfordre.com/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e363?rskey=ivWwgv&result=2.

Dingemanse, Mark, Francisco Torreira & N. J. Enfield. 2013. Is "Huh?" a Universal Word? Conversational infrastructure and the convergent evolution of linguistic items. *PLoS ONE* 8(11). e78273. doi:10.1371/journal.pone.0078273.

Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language* 68. 81–138.

Dryer, Matthew, S. 2013a. Order of demonstrative and noun. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/88 (accessed 04 January 2019).

Dryer, Matthew, S. 2013b. Order of numeral and noun. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/89, (accessed 04 January 2019).

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info (accessed 05 July 2018).

Du Bois, W. John, Lorraine E. Kumpf & William J. Ashby (eds.). 2003. *Preferred argument structure: Grammar as architecture for function*. (Studies in discourse and grammar, 14). Amsterdam: John Benjamins.

Dunn, Michael, Simon Greenhill, Stephen C. Levinson & Russell Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473. 79–82.

Erhart, Adolf. 1987. Die Nichtunterscheidung von Numerus in der 3. Person des Baltischen Verbs und ihre indoeuropäischen Grundlagen. *Baltistica* XXIII(2). 126–130.

Fedzechkina, Maryia, Elissa L. Newport & T. Florian Jaeger. 2016. Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. *Cognitive Science* 41(2). 416–446.

Futrell, Richard, Kyle Mahowald & Edward Gibson. 2015. Quantifying word order freedom in dependency corpora. *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, 91–100, Uppsala, Sweden, 24–26 August 2015.

Givón, Talmy. 1984. *Syntax: A functional-typological introduction*, vol. I. Amsterdam/Philadelphia: John Benjamins.

Goldhahn, Dirk, Thomas Eckart & Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. *Proceedings of the 8th International Language Resources and Evaluation (LREC'12), 2012*.

Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg (ed.), *Universals of human language*, 73–113. Cambridge, Mass: MIT Press.

Guzmán Naranjo, Matías & Laura Becker. 2018. Quantitative word order typology with UD. *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, Issue 155, 91–104. Oslo University, Norway, 13–14 December 2018.

Haig, Geoffrey & Stefan Schnell (eds.). 2016a. Multi-CAST (Multilingual Corpus of Annotated Spoken Texts). https://lac.uni-koeln.de/multicast/ (accessed 25 October 2017).
Haig, Geoffrey & Stefan Schnell. 2016b. The discourse basis of ergativity revisited. *Language* 92(3). 591–618.
Hale, Ken. 1982. Preliminary remarks on configurationality. *Proceedings of the North Eastern Linguistic Society* 12. 86–96.
Hale, Ken. 1983. Warlpiri and the grammar of non-configurational languages. *Natural Language and Linguistic Theory* 1. 5–47.
Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1). 31–80.
Haspelmath, Martin, Andreea Calude, Michael Spagnol, Heiko Narrog & Elif Bamyacı. 2014. Coding causal-noncausal verb alternations: A form-frequency correspondence explanation. *Journal of Linguistics* 50(3). 587–625.
Hawkins, John A. 1994. *A performance theory of order and constituency*. (Cambridge Studies in Linguistics, 73.) Cambridge: Cambridge University Press.
Hawkins, John A. 2014. *Cross-linguistic variation and efficiency*. Oxford: Oxford University Press.
Hualde, José Ignacio & Jon Ortiz de Urbina (eds.). 2003. *A grammar of Basque*. Berlin: De Gruyter Mouton.
Jakobson, Roman. 1971[1936]. Beitrag zur allgemeinen Kasuslehre. In Roman Jakobson (ed.), *Selected writings. Vol. II. Word and language*, 23–71. The Hague/Paris: Mouton.
Jamashita, Hiroko & Franklin Chang. 2000. "Long before short" preference in the production of a head-final language. *Cognition* 81. B45–B55.
Juola, Patrick. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics* 5(3). 206–213.
Kiparsky, Paul. 1997. The rise of positional licensing. In Ans van Kemenade & Nigel Vincent (eds.), *Parameters of morphosyntactic change*, 460–494. Cambridge: Cambridge University Press.
Kiss, Katalin É. 1987. *Configurationality in Hungarian*. Dordrecht: Reidel.
Koplenig, Alexander, P., Meyer, S. Wolfer & C. Müller-Spitzer. 2017. The statistical trade-off between word order and word structure – Large-scale evidence for the principle of least effort. *Plos One* 2017(12). e0173614. https://doi.org/10.1371/journal.pone.0173614.
Kurumada, Chigusa & Florian Jaeger. 2015. Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language* 83. 152–178.
Lambrecht, Knud. 1994. *Information structure and sentence form: Topic, focus, and the mental representation of discourse referents*. Cambridge: Cambridge University Press.
Lehmann, Christian. 2015. *Thoughts on grammaticalization*, 3rd edn. Berlin: Language Science Press.
Lehmann, Winfred P. 1978. The great underlying ground-plans. In Winfred P Lehmann (ed.), *Syntactic typology: Studies in the phenomenology of language*. Austin: University of Texas Press.
Levshina, Natalia. 2015. European analytic causatives as a comparative concept. Evidence from a parallel corpus of film subtitles. *Folia Linguistica* 49(2). 487–520.
Levshina, Natalia. 2018. *Towards a theory of communicative efficiency in human languages*. Leipzig University Habilitation thesis. http://doi.org/10.5281/zenodo.1542857.
Liu, Haitao. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua* 120(6). 1567–1578.

Mayer, Thomas & Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 3158–3163, Reykjavik.

McFadden, Thomas. 2003. On morphological case and word-order freedom. *Proceedings of the Twenty-Ninth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Phonetic Sources of Phonological Patterns: Synchronic and Diachronic Explanations*, 295–306.

Mithun, Marianne. 1992. Is basic word order universal? In Doris L. Payne (ed.), *Pragmatics of word order flexibility*, 15–61. Amsterdam: John Benjamins.

Nivre, Joakim, Željko Agić, Lars Ahrenberg et al. 2017. Universal dependencies 2.0 – CoNLL 2017 shared task development and test data. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University. http://hdl.handle.net/11234/1-2184. See also http://universaldependencies.org/ (accessed 14 December 2017).

Östling, Robert. 2015. Word order typology through multilingual word alignment. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, 205–211. Beijing, China, 26–31 July 2015.

Payne, Doris L. (ed.). 1992. *Pragmatics of word order flexibility*. Amsterdam: John Benjamins.

Piantadosi, Steven, Harry Tily & Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108(9). 3526.

R Core Team. 2018. R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. Version 3.5.2. https://www.R-project.org/.

Rögnvaldsson, Eiríkur. 1995. Old Icelandic: A non-configurational language? *NOWELE* 26. 3–29.

Sapir, Edward. 1921. *Language, an introduction to the study of speech*. New York: Harcourt, Brace and Co.

Schmidtke-Bode, Karsten & Holger Diessel. 2017. Cross-linguistic patterns in the form, function and position of (object) complement clauses. *Linguistics* 55(1). 1–38.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(3). 379–423.

Siewierska, Anna. 1998. Variation in major constituent order: A global and a European perspective. In Anna Siewierska (ed.), *Constituent order in the languages of Europe*, 475–552. Berlin: De Gruyter Mouton.

Sinnemäki, Kaius. 2008. Complexity trade-offs in core argument marking. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity: Typology, contact, change*, 67–88. Amsterdam: John Benjamins.

Stolz, Thomas, Nataliya Levkovych, Aina Urdze, Julia Nintemann & Maja Robbers. 2017. *Spatial interrogatives in Europe and beyond: Where, whither, whence*. Berlin: De Gruyter Mouton.

Straka, Milan & Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, August 2017. Vancouver, Canada.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, 2214–2218.

Vennemann, Theo. 1974. Theoretical word order studies: Results and problems. *Papiere zur Linguistik* 7. 5–25.

Wälchli, Bernhard. 2009. Data reduction typology and the bimodal distribution bias. *Linguistic Typology* 13. 77–94.

Wälchli, Bernhard & Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50(3). 671–710.
Wijffels, Jan. 2018. udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the UDPipe NLP Toolkit. R package version 0.7. https://CRAN.R-project.org/package=udpipe.
Zakharko, Taras, Alena Witzlack-Makerevich, Johanna Nichols & Balthasar Bickel. 2017. Late aggregation as a design principle for typological databases. ALT Workshop on Design Principles of Typological Databases, 15 December 2017.
Zipf, George. 1935. *The psychobiology of language: An introduction to dynamic philology*. Cambridge, MA: MIT Press.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/lingty-2019-0025).