Quantitative word order typology with UD

Matías Guzmán Naranjo[1], *Laura Becker*[2]

(1) Heinrich Heine Universität Düsseldorf
(2) University of Erlangen-Nürnberg, Leipzig University

guzmanna@hhu.de, laura.becker@uni-leipzig.de

Abstract

Cross-linguistic universals of word order correlations often based on the distinction basic VO and OV orders have received a lot of attention since the seminal work by Greenberg (1963), followed by e.g. Dryer (1991, 1989); Hawkins (1979, 1980, 1983); Lehmann (1973, 1974); Vennemann (1974, 1975). However, there have been quantitative studies (e.g. Chen and Gerdes, 2017; Dunn et al., 2011; Liu, 2010) focusing on a small number of languages (Celano, 2014), or insisting on canonical word order for every language. The aim usually is to find crosslinguistic word order correlations on the basis of this canonical order. How to determine the latter for any language is, however, highly problematic and potentially misleading for a number of languages, as was already argued convincingly in Mithun (1992): it means that stricter OV order languages such as Japanese are treated like flexible OV order languages such as German. Despite some strong crosslinguistic correlations based on canonical word order that could be confirmed in independent samples, it is still not clear whether these effects can reliably be modelled as categorical or whether we should rather treat them as gradient. This is what we propose in the present study: We explore the question of whether word order tendencies between the verb and its arguments may have some influence on the orders between nouns and their dependents, and whether these tendencies are cross-linguistic or language specific.

KEYWORDS: word order universals, word order correlations, gradient tendencies.

1 Background and motivation

Since Greenberg (1963), crosslinguistic word order correlation and related questions have received a lot of attention in language typology (Cristofaro, 2018; Dryer, 1992, 2009, 2019; Hawkins, 1994, 2014; Payne, 1992; Siewierska, 1988; Song, 2009), to name just a few. Examples of robust crosslinguistic generalizations include the following correlations between the verb-object order and the order of other elements in the clause (taken from Dryer (1991, 1992, 2009):

VO	OV
prepositions	postpositions
postnominal relative clause	prenominal genitive
prenominal article	postnominal article
verb - adverb	adverb - verb
clause-initial complementizer	clause-final complementizer

Table 1: Examples of crosslinguistic word order correlations.

Some of these correlations were argued to be based on a general preference for a head-dependent order within a given language. For instance, Hawkins (1983) argued for a so-called "cross-category harmony", meaning that we often find verb-initial languages with mostly all of the dependents following their heads, while verb-final languages should mostly have all dependents preceding their heads. SVO languages were situated in the middle and expected to feature some dependents before and some following their heads. Dryer (1992, 2009), on the other hand, shows that the explanations in terms of head-dependent orders cannot capture all the crosslinguistic patterns found. Therefore, he argues for a "branching directory theory", according to which word order correlations reflect a tendency for languages to be consistently left-branching or right-branching. Two other main types of explanations for the correlation patterns found have to be mentioned: Hawkins (1994, 2014) offers parsing-based explanations in terms of structures with shortest constituent recognition domains. Bybee (1988); Aristar (1991); Cristofaro (2018), on the other hand, show that what may be taken as a correlation between two structures on the synchronic level, are rather diachronically related structures which is why we cannot speak of correlations at all in some cases (e.g. the orders between the noun and genitives as well as with relative clauses). Nevertheless, the main idea of a basic order of different elements for the sake of consistency within a given language is taken up in Dryer (2019), who proposes 5 surface principals that account for the crosslinguistic trends, one of which being being "intracategorial harmony", i.e. the general preference of different types of nominal elements occurring on one side of the noun within a language.

What this brief overview shows is that even though the concept of basic word orders has been discussed and criticised (Mithun, 1992; Payne, 1992; Siewierska, 1988), more recent studies that seek functional explanations for word order correlations still take it for granted that we can determine the basic word order of any given language (e.g. Dunn et al., 2011; Dryer, 2019; Hawkins, 2014). To give an example, languages like German or Spanish are then assigned SOV or SVO orders (respectively), even though both languages allow for most possible orderings. This leads to a situation where a language like Japanese, which is consistently a OV language ends up in the same group as German, which despite being an OV language, allows for all possible orderings (with respect to S, V and O). Similarly, it is often the case that languages have both pre- and postpositions, but this is often not taken into account. In other words, languages are classified categorically with respect to certain word order properties, even though we almost always find gradient variation for these features within single languages.

The aim of this study is to present a new method which could help remedy this shortcoming. The main point is that we do not need to make such categorical choices; rather, we should look at the proportions with which languages use VO vs. OV orders, and compare those proportions with the proportions of other word orders (e.g. pre- vs. postpositions). Thus, instead of saying that a language is OV, we can say that it uses OV structures in 90% of cases, while 10% of cases have a VO structure.

Even though there are a number of quantitative studies on word order variation within and across languages using dependency treebanks, most of these have slightly different objectives, e.g. examining word order freedom and zooming in on its correlation with the availability of case marking across languages (Futrell et al., 2015), or investigating the evidence for dependency length minimization for different types of head-dependent relations (Gulordava, 2018). Another study that uses dependency treebanks for crosslinguistic comparison is Liu (2010), which investigates 20 languages for the prevalence of head initial vs head final in them. We argue that this methodology can be further expanded to include more fine-grained comparisons, and establish correlations across orderings of different dependents, returning to the starting point of word order typology. Somewhat related is the study by Chen and Gerdes (2017), who classify languages according to dependency structure also using the Universal Dependencies Treebank for 20 languages. This study, however, does not explore word order universals, focusing on the distance between languages.

2 Datasets

To explore gradient word order correlations we use the Universal Dependencies Treebank (Nivre et al., 2016) version 2.2, which as of July 1st, 2018 contains 122 treebanks for 71 languages.

We are aware of the fact that this dataset has several shortcomings when used for language typology. First, there is relatively little subfamily variation. Most languages in the sample belong to an Indo-European subfamilies, and the corpora for Non-Indo-European languages are smaller (e.g. Bambara with 13K tokens) than the datasets for languages like English (586K tokens) or Russian (1247K tokens). Typological studies usually take a lot more care in selecting a balanced sample of languages (Bickel, 2008; Dryer, 1989, 2019). However, despite this clear issue, the results we obtain from looking at the Universal Dependency dataset serve as a good starting point for future work on quantitative word order correlations. As treebanks for more languages from other families and geographical locations become more readily available, one can easily expand on this study, and see whether results confirm or disprove these initial findings. The aim of this study is to present what we believe is an innovative technique, even if our results hold for the language sample of the UD treebanks and cannot be generalized as true *universal* tendencies. The second potential objection is that we entirely depend on the annotation schemes used by the creators of the UD treebanks, which is not yet perfectly consistent. However, as the UD project aims at having an annotation scheme which is applicable to different languages and comparable across them, the UD treebanks certainly offer a robust crosslinguistically comparable annotation.

3 Methodology

The UD project offers treebanks for 70 languages of 20 sub families, 8 of which are Indo European. We first combined the available treebanks for all languages. The families and the number of languages in each subfamily were the following: Afro-Asiatic (4), Altaic (6), Armenian

(1), Austronesian (2), Baltic (2), Basque (1), Celtic (2), Creole (1), Defoid (1), Dravidian (2), Germanic (9), Greek (2), Indo-Iranian (6), Pama-Nyungan (1), Romance (9), Swedish Sign Language (1), Sinitic (2), Slavic (12), Uralic (5), Viet-Muong (1).

We extracted the dependents from the treebanks for each noun, for each verb, and whether these dependents preceded or followed their heads. We only considered verb dependents with one of the following part-of-speech tags: NOUN, VERB, PROPN (proper noun), PRON (pronoun) and AUX (auxiliary). We considered all noun dependents. We made this decision in order to restrict correlations to content words, which seemed more likely to occur crosslinguistically. This gives us a count for each language: for instance, of how many times the determiner follows and precedes a noun, or of how often objects follow or precede the verb, etc. From these absolute occurrences of different types of head-preceding and head-following dependents, we calculated the proportion of a given dependent following its head (noun or verb).

We took into account the following types of verb dependents:

- *advcl*: adverbial clause modifiers
- *advmod*: adverbial modifiers (non clausal)
- *nsubj*: nominal subject (noun phrase which acts as subject of the verb), first core argument of the clause
- *obj*: (direct) object of a verb, second core argument of the clause
- *obl*: oblique, or non-core argument of the verb

The noun dependents we considered are the following:

- *advcl*[1]: adverbial clause modifiers
- *acl*: clausal modifiers of nouns
- *amod*: adjectival modifiers
- *case*: used for any case-marking element which is treated as a separate syntactic word (mostly prepositions, but also postpositions, and clitic case markers)
- *compound*: relation used to mark noun compounding
- *det*: nominal determiners
- *nmod*: nominal modifiers of other nouns (not appositional)
- *nummod*: numeral modifiers of nouns

Clearly, the use of these dependency relations has some benefits as well as potential issues. An advantage is that the UD treebanks inherently aim at defining these relations in such a way that they are crosslinguistically applicable. However, it is not the case that, at this point in the development of the UD treebanks, all treebanks use these relations consistently, and since checking this is prohibitively complicated and time consuming, we have to assume that the relations used and the annotation schemes are comparable to some extent.

[1] We sometimes mark this as *n_advcl* to distinguish it from the adverbial clause modifiers for verbs.

4 Results

We explore several questions in this section. First, we examine the distribution of proportions for each of the dependents that are included in this study. The distribution of proportions is a first sanity check, to make sure our data aligns with what we know about these categories from previous studies. Second, we examine the intracategorial harmony (Dryer, 2009, 2019) of orderings of the noun dependents and verb dependents. The final question is whether noun dependent ordering proportions are predictable from verb dependent ordering proportions, and vice versa. Although related to the crosscategorial harmony proposed in Hawkins (1983), it is a novel question which directly extends the classical implicational universals that have been established in word order typology. Since we examine proportions of head-dependent orders, single languages correspond to single data points. Therefore, we address the languages collectively and not separately in this section in order to assess crosslinguistic tendencies of word order correlations between different types of heads and dependents. For brevity, we sometimes refer to *dependent ordering proportions* simply as *noun dependent* or *verb dependent*.

4.1 Distributions

We first explore the distribution of all dependents and their position with respect to their heads. Figures 1 and 2 present the distribution of verb and noun dependents for all languages, with the highest proportions of preceding dependents on the left and the highest proportions of following dependents on the right. We can observe some clear global trends. First, for verb dependents, there is a pronounced preference for subjects to be preverbal. This is likely due to the fact that subjects are often topics and thus given information, which has been shown to generally precede new information in the sentence (Gundel, 1988; Lambrecht, 1994; Arnold et al., 2000; Taboada and Wiesemann, 2010; Junge et al., 2015). For both direct and oblique objects, on the other hand, we see a bimodal distribution with a preference for both categories following the verb, and obliques being somewhat more flexible. Adverbial clauses show a similar preference for postverbal position, but less pronounced than objects and obliques. Finally, adverbial modifiers are predominantly preverbal.

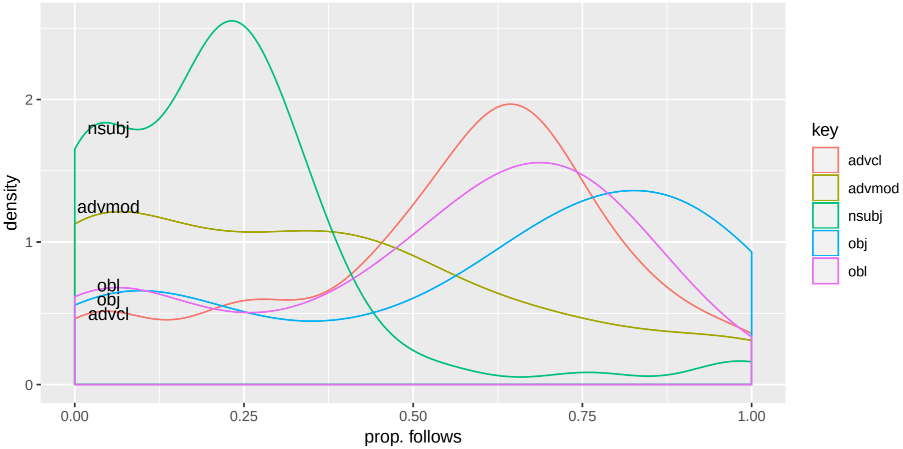


Figure 1: Density distribution for verb dependents.

For noun dependents, Figure 2 shows that the situation is somewhat different. The top plot illustrates the proportions of prenominal (left) and postnominal (right) clausal, adjectival modifiers, nominal, numeral, and adverbial clause modifiers of nouns. All these dependents have a preference for being either post- or preverbal, but they also appear in the other position, respectively. In the bottom plot, on the other hand, we see two other types of distributions: case marking words (i.e. mainly adpositions) and compounds have no clear preference for either position; while determiners show a very strong preference within single languages as well as across languages to precede the head with only few exceptions. Thus, there seems to be a clear difference between these two distribution types of dependents: those with and those without strong preferences for prenominal or postnominal occurrence.

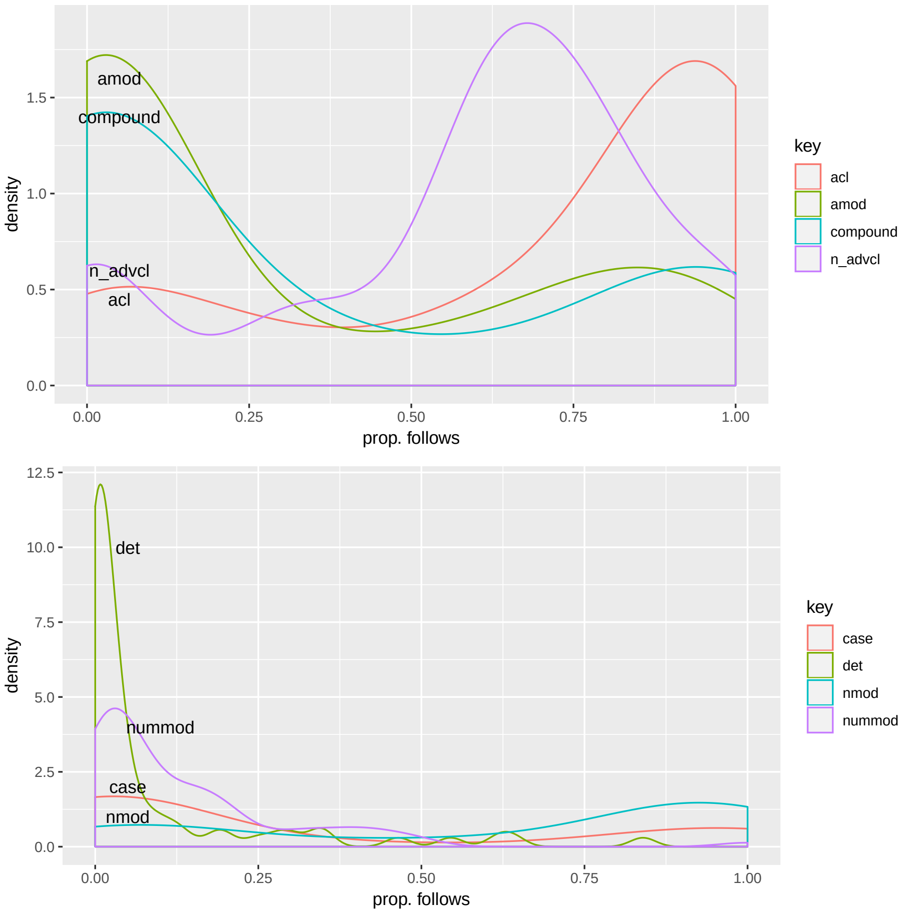
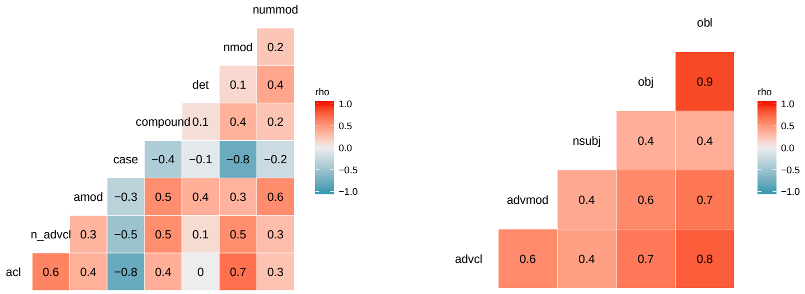


Figure 2: Density distribution for noun dependents.

4.2 Basic correlations

First, we examine the intracategorial correlations of head-dependent orders. Figure 3 presents the correlations between dependents of the verb as well as of the noun for all languages in our dataset. For nouns (a), we observe a strong positive correlation (red) between the position of the nominal modifier (*nmod*) and clausal modifiers (*acl*). This means that both indeed tend



(a) Correlations between noun dependents.

(b) Correlations between verb dependents.

Figure 3: Correlations between dependents.

to occur on the same side of the noun. We also see negative correlations (blue) between case marking elements (*case*) and both clausal and nominal modifiers, meaning these tend to occur at the opposite side of the noun. What this means is that we often find structures in which the *case* element connects the head noun and the dependent noun (*nmod*, as in “the house of the major”, in which the case element (“of”) precedes its head (“major”), which is at the same time the nominal modifier of the head noun “house” and follows the latter. On the other hand, structures with e.g. a following *nmod* and *case* element (like “the house the major of”) are infrequent in our dataset. This observation relates to Himmelmann (1997, 159-188), who discusses a class of nominal linking elements (called “linking articles”) that are used in a number of languages to indicate nominal modification by other nouns, adjectives, and clauses. Interestingly, this linking element always occurs between the two elements, and not at the edge of the noun phrase.

For verb dependents, we see that there are no negative correlations, but at least two strong positive correlations, namely between direct objects (*obj*) and oblique objects (*obl*), as well as obliques and adverbial clause modifiers. That the different types of objects and clausal modifiers tend to occur on the same side of the noun supports that there is a general intracategorial harmony. On the other hand, the order of subjects being less strongly correlated to other verb dependents again points towards a stronger information structural effect motivating the position of subjects.

The next step is to consider intercategorial correlations between noun and verb dependents, as is shown in Figure 4. We find a strong correlation (red) between the position of oblique objects (*obl*) with respect to the verb, and the direction of adverbial clauses modifying a noun (*n_advcl*), clausal modifiers of the noun (*acl*), and nominal modifiers (*nmod*). Similarly, there is a strong negative correlation (white) between obliques and the position of case markers. The position of the object with respect to the verb (*obj*) also correlates with the position of clausal modifiers and nominal modifiers.

Perhaps somewhat interesting is that we do not see any strong correlations between the position of the subject with respect to the verb and other head-dependent features. This is not completely surprising in the light of Vennemann (1974, 1975); Lehmann (1973, 1974); Dryer (1991), but it is a nice corroboration that subject position has more to do with information structure, than

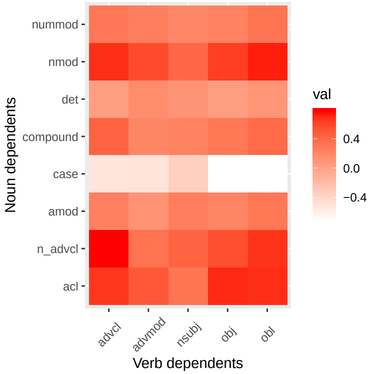


Figure 4: Correlations between noun and verb dependents.

other things. We confirm this observation in the models in the next section.

4.3 Models

In order to test for deeper correlations between these variables, we fitted beta regression models to the data, using the subfamily of the languages as a random effect. This should, at least to a certain extent, control for subfamily biases. Since in some cases the proportions of follows or precedes were equal to 1 or 0, and because beta regression does not allow for a dependent variable containing values of 1 or 0, but only values between 1 and 0, we transformed the dependent variable for every model using the technique described in (Smithson and Verkuilen, 2006). For each model, we calculated the marginal and conditional R2 values following the method developed by (Nakagawa and Schielzeth, 2013; Nakagawa et al., 2017). The marginal R2 value is the proportion of the variance explained the fixed effects alone, and the conditional R2 value is the proportion of the data explained by both the random and fixed effects. Using these two metrics, we can examine how much the fixed effects correlate with the dependent variable, and how much the variable is explained by subfamily bias.

Table 2 contains the models for noun dependents. Each row represents one model with the dependent variable in the leftmost column, then the intercept, and then the coefficients of the significant predictors. The cells marked in gray correspond to the predictors which did not reach statistical significance in the models. Strikingly, *obl* is the most frequent significant predictor, appearing in 7 out of the 10 models, even above *obj*, which proved to be significant as a predictor in only 4 of the models. This is a somewhat unexpected result, since most previous work on word order typology focused on the position of the direct object with respect to the verb, rather than the position of oblique objects with respect to the verb. One possible explanation for this fact is that in 41 out of the 70 languages, obliques are more frequent than direct objects. Another potential explanation for a difference between direct and oblique objects in predictive power could be a difference in realization as lexical noun or proform. This remains to be tested, however.

Another important result from the models in Table 2 is that there is a large difference in the predictability of the proportions of the noun dependents. The models for clausal noun modifiers (*acl*) and noun adverbial clause modifiers (*advcl*) explain a large amount of variance just with the fixed effects. These are factors which correlate with other variables in the corpus, but for

predicted	intercept	advcl	nsubj	nsbuj:obj	obj	obj^2	obj:obl	obl	obl^2	R2_m	R2_c
acl	0.02	2.02	-1.43		6.39	-3.81				0.462	0.462
n_advcl	-1.29				0.94	-5.45		3.25		0.428	0.555
amod	-1.59							1.56		0.076	0.362
case	0.5							-2.48		0.099	0.67
compound	-1.63	1.99								0.111	0.285
det	-2.88		0.74	-9.36	-0.11			2.10	3.26	0.170	0.170
nmod	-0.95	3.71			-5.31		7.20	-1.36		0.246	0.720
nummod	-2.66							1.64		0.079	0.409

Table 2: Coefficients and R2 values for models predicting noun dependents.

which there are no strong subfamily biases. In the models for adjectival modifiers (*amod*), nominal modifiers (*nmod*) and numeric modifiers (*nummod*), the fixed effects explain a small portion of the variance, while the random effects (subfamily) explains a relatively large amount of variance. These are cases where the subfamily is the main explanatory factor. Finally, in the models for case marking elements (*case*) and determiners (*det*), neither the fixed nor the random effects explain much of the variance.

Table 3 shows the models predicting verb dependents from noun dependents. Here, we see that the noun adverbial clause modifier (*n_advcl*) is the most common significant predictor (we excluded it as a predictor when *advcl* was also the dependent variable). The other important observation is that the models for both direct objects (*obj*) and oblique objects (*obl*) are very similar. They have similar coefficients, and the same two significant predictors (*advcl* and *case*), and their R2 values are also close to each other. The main difference is that for *obj*, the effect of the language subfamily seems to be larger. Subjects are the least predictable verb dependents given other verb dependents as predictors. This is most likely due to the fact that the subject position is heavily influenced by the information structure of the sentence.

predicted	intercept	acl	n_advcl	case	compound	nmod	R2_m	R2_c
advcl	-0.76				0.72	1.57	0.15	0.528
advmod	-2.07		1.65			0.97	0.240	0.240
nsubj	-1.17	-1.54	2.27	-1.26			0.161	0.320
obj	-0.30		2.86	-2.15			0.433	0.634
obl	-1.05		2.92	-1.64			0.445	0.513

Table 3: Coefficients and R2 values for models predicting verb dependents.

Finally, Figure 5 presents the three best models for verb (left column) and noun (right column) dependents: predicting *acl*, *nmod*, *advcl* (in both), *obj* and *obl*. We see the observed vs fitted values for all six models. First, the models for verb dependents are a better fit to the data. For noun dependents, we see that there are several outliers for *advcl* and *acl* from mostly Non-European languages, the overall fit still being relatively good.

5 Conclusion and Outlook

This study had two main objectives. First, and most importantly, we present a new method for investigating word order universals by making use of treebanks. Even though our results may still be somewhat influenced by the bias in the UD Treebank towards Indo-European languages, a similar approach with a more balanced, and larger corpus can provide more accurate and differentiated results than previous studies based on categorical word order distinctions. This objection not withstanding, we want to emphasize that our results generally agree with previous observations in the literature. This fact leads us to be relatively confident that our results should

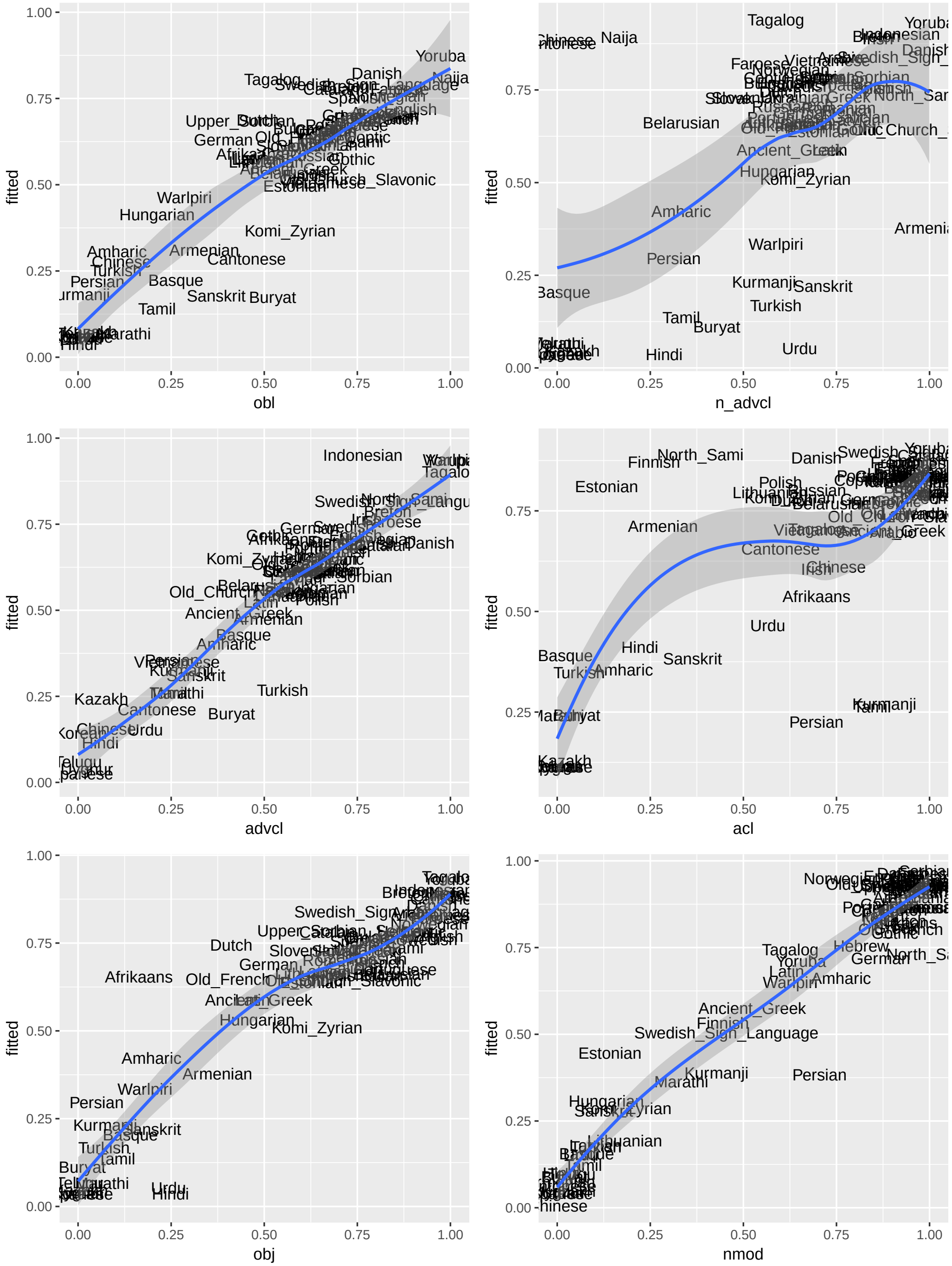


Figure 5: Observed vs fitted predictions for six models. The left column shows verb dependent models for *obl*, *advcl* and *obj*, and the right column shows noun dependent models for (n_)*advcl*, *acl* and *nmod*.

be generalizable to a larger sample.

Secondly, we show that at least some word order universals are not categorical, but in fact gradient. For instance, it is not that OV languages favour postpositions, it is that the proportion of OV vs VO structures in the language correlates with the proportion of post and prepositions. This is a more nuanced claim. As far as we are aware, this is a new observation, and it may help us to rethink the explanations for these phenomena.

A possibility for future work is to distinguish between main and subordinate clauses. We know that word order can vary between main and subordinate sentences, with the later often having a stricter word order like in German or French (Bybee, 2002). Similarly, we could try to distinguish between different types of noun phrases (nominal vs. pronominal), noun phrases of different lengths, and different elements within noun phrases. Also, even though we saw robust crosslinguistic trends such as the relative independence of the position of subjects, some languages of our sample are usually considered to be VSO languages. A more detailed look at specific languages or subfamilies for certain head-dependent orders could show to what extend this corpus is in line with previous language-specific word order observations. To avoid the bias towards Indo-European languages, it might also be helpful to exclude these from the sample and see if the results still hold for the Non-Indo-European subset of the UD treebanks.

Another possible path to take in future work is to try and convert the UD treebanks into different annotation schemes. There is work on converting dependency treebanks into LFG representations (Haug, 2012), for example. If one could convert the UD dependencies into some other theory, this might provide us with structures that make it possible to explore other relations crosslinguistically.

References

Aristar, A. R. (1991). On Diachronic Sources and Synchronic Pattern: An Investigation into the Origin of Linguistic Universals. *Language*, 67(1):1–33.

Arnold, J. E., Losongco, A., Wasow, T., and Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1):28–55.

Bickel, B. (2008). A refined sampling procedure for genealogical control. *Sprachtypologie und Universalienforschung*, 61:221–233.

Bybee, J. (1988). The diachronic dimension in explanation. In *Explaining Language Universals*, pages 350–379. Blackwell, Oxford.

Bybee, J. L. (2002). Main clauses are innovative, subordinate clauses are conservative: Consequences for the nature of constructions. In Bybee, J. L. and Noonan, M., editors, *Complex Sentences in Grammar and Discourse Essays in Honor of Sandra A. Thompson*, pages 1–18. Benjamins, Amsterdam.

Celano, G. G. A. (2014). A computational study on preverbal and postverbal accusative object nouns and pronouns in Ancient Greek. *The Prague Bulletin of Mathematical Linguistics*, 101(1):97–110.

Chen, X. and Gerdes, K. (2017). Classifying languages by dependency structure. Typologies of delexicalized universal dependency treebanks. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), September 18-20, 2017, Università Di Pisa, Italy*, pages 54–63.

Cristofaro, S. (2018). Processing explanations of word order universals and diachrony: Relative clause order and possessor order.

Dryer, M. S. (1989). Article-noun order. *Chicago Linguistic Society*, 25:83–97.

Dryer, M. S. (1991). SVO languages and the OV : VO typology. *Journal of Linguistics*, 27(2):443–482.

Dryer, M. S. (1992). The Greenbergian word order correlations. *Language*, 68(1):81–138.

Dryer, M. S. (2009). The branching direction theory of word order correlations revisited. In Scalise, S., Magni, E., and Bisetto, A., editors, *Universals of Language Today*, Studies in Natural Language and Linguistic Theory, pages 185–207. Springer, Dordrecht.

Dryer, M. S. (2019). On the order of demonstrative, numeral, adjective and noun. *Language*.

Dunn, M., Greenhill, S. J., Levinson, S. C., and Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82.

Futrell, R., Mahowald, K., and Gibson, E. (2015). Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100, Uppsala.

Greenberg, J. H., editor (1963). *Universals of Language*. MIT Press, Cambridge, MA.

Gulordava, K. (2018). *Word Order Variation and Dependency Length Minimisation: A Cross-Linguistic Computational Approach*. PhD thesis, University of Geneva.

Gundel, J. K. (1988). Universals of topic-comment structure. In Hammond, M., Moravcsik, E. A., and Wirth, J., editors, *Studies in Syntactic Typology*, pages 209–239. Benjamins, Amsterdam.

Haug, D. T. T. (2012). From dependency structures to LFG representations. In *Proceedings of the LFG12 Conference*, pages 271–291.

Hawkins, J. A. (1979). Implicational universals as predictors of word order change. *Language*, 55(3):618–648.

Hawkins, J. A. (1980). On implicational and distributional universals of word order. *Journal of Linguistics*, 16(2):193–235.

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

Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge University Press, Cambridge.

Hawkins, J. A. (2014). *Cross-Linguistic Variation and Efficiency*.

Himmelmann, N. P. (1997). *Deiktikon, Artikel, Nominalphrase: Zur Emergenz Syntaktischer Struktur*. Niemeyer, Tübingen.

Junge, B., Theakston, A. L., and Lieven, E. (2015). Given–new/new–given? Children's sensitivity to the ordering of information in complex sentences. *Applied Psycholinguistics*, 36(3):589–612.

Lambrecht, K. (1994). *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge University Press, Cambridge.

Lehmann, W. P. (1973). A structural principle of language and its implications. *Language*, 49(1):47–66.

Lehmann, W. P. (1974). *Proto-Indo-European Syntax*. University of Texas Press, Austin.

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

Mithun, M. (1992). Is basic word order universal? In Payne, D. L., editor, *Pragmatics of Word Order Flexibility*, pages 15–61. Benjamins, Amsterdam.

Nakagawa, S., Johnson, P. C., and Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, 14(134).

Nakagawa, S. and Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R. T., Petrov, S., Pyysalo, S., Silveira, N., and others (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *LREC*.

Payne, D. L., editor (1992). *Pragmatics of Word Order Flexibility*. Benjamins, Amsterdam.

Siewierska, A. (1988). *Word Order Rules*. Croom Helm, London.

Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, 11(1):54.

Song, J. J. (2009). Word order patterns and principles: An overview. *Language and Linguistics Compass*, 3(5):1328–1341.

Taboada, M. and Wiesemann, L. (2010). Subjects and topics in conversation. *Journal of Pragmatics*, 42(7):1816–1828.

Vennemann, T. (1974). Topics, subjects and word order: From SXV to SVX via TVX. In Anderson, J. and Jones, C., editors, *Proceedings of the First International Congress of Historical Linguistics, Edinburgh, September 1973*, pages 339–376. North-Holland, Amsterdam.

Vennemann, T. (1975). An explanation of drift. In Li, C. N., editor, *Word Order and Word Order Change*, pages 269–305. University of Texas Press, Austin, TX.