Word order correlations from a quantitative perspective

Matías Guzmán Naranjo and Laura Becker

15.11.2018, Paris

Word order typology

Since Greenberg (1963), crosslinguistic word order correlation and related questions have received a lot of attention in language typology (e.g. Cristofaro, 2018; Dryer, 1992, 2009, 2019; Hawkins, 1994, 2014; Payne, 1992; Siewierska, 1988; Song, 2009).

Some examples of robust crosslinguistic generalizations concerning the verb-object order and the order of other elements in the clause (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

Word order typology: types of explanations

"Cross-category harmony" (Hawkins, 1983)

a general preference for a head-dependent order within a given language

→ find verb-initial languages with mostly all of the dependents following their heads

→ verb-final languages should mostly have all dependents preceding their heads

"Branching directory theory" (Dryer, 1992, 2009)

Word order correlations reflect a tendency for languages to be consistently left-branching or right-branching.

Issues

This traditional approach, based on categorical decisions concerning word order is problematic:

- It is difficult to determine the main word order of a language.
- While some languages show rigid word order, others are much more flexible.

→ This approach treats these two types of languages equally.

→ For languages with flexible word order, other minor patterns are disregarded.

We can overcome this problem if we take a corpus based approach instead, and model word order tendencies as gradient.

Dataset

Universal Dependencies Treebank 2.2 (Nivre et al., 2016)

- we removed those treebanks without complete annotations

→ treebanks for 70 languages of 20 subfamilies (8 are Indo-European)

- Afro-Asiatic, (4)
- Altaic (6),
- Austronesian (2),
- Basque (1),
- Defoid (1),
- Dravidian (2),
- Indo-European (Armenian (1), Baltic (2), Celtic (2), Germanic (9), Greek (2), Romance (9), Slavic (12))
- Indo-Iranian (6),
- Pama-Nyungan (1),
- Sinitic (2),
- Uralic (5),
- Viet-Muong (1)
- Creole (1), Swedish Sign Language (1)

Dataset: disclaimer

We are aware some shortcomings of this dataset:

- There is relatively little family variation.
- The corpora for non Indo-European languages are smaller than the datasets for languages like Czech or Russian.
- We entirely depend on the annotation schemes used by the treebank creators.

Typological studies usually take a lot more care in selecting a balanced sample of languages (Bickel, 2008; Dryer, 1989, 2019).

Despite this clear issue, the results we obtain from looking at the Universal Dependency dataset serve as a robust starting point for future work on quantitative word order correlations.

Extracted dependencies

We extracted the dependents from the treebanks for each noun and each verb, and distinguish between their relative order with the head

- head – dependent (following)
- dependent – head (preceding)

We then calculated the proportion of a given dependent **following** its head (noun or verb).

Extracted dependencies

For verb dependents the following part-of-speech tags were considered:

- NOUN
- VERB
- PROPN (proper noun)
- PRON (pronoun)
- AUX (auxiliary)

For noun dependents we considered all part-of-speech tags.

Verb dependents

We took into account the following types of verb dependents:

advcl	adverbial clause modifiers *He talked to him in order to secure the account.*
advmod	adverbial modifiers (non clausal) *genetically modified food*
nsubj	nominal subject (noun phrase which acts as subject of the verb), first core argument of the clause *There is a ghost in the room.*
obj	(direct) object of a verb, second core argument of the clause *She gave me a raise.*
obl	oblique, or non-core argument of the verb *Last night , I swam in the pool.* *give the toys to the children*

Noun dependents

advcl	adverbial clause modifiers *He was the one present when it happened.*
acl	clausal modifiers of nouns *There are many online sites offering booking facilities.* *the issues as he sees them*
amod	adjectival modifiers *Sam eats red meat*
case	used for any case-marking element which is treated as a separate syntactic word (mostly prepositions, but also postpositions, and clitic case markers) *the office of the Chair*
compound	relation used to mark noun compounding *phone book*
det	nominal determiners *which book, the woman*
nmod	nominal modifiers of other nouns (not appositional) *the dog's bone*
nummod	numeral modifiers of nouns *Sam ate 3 potatoes*

Results

We explore three questions in this section, exploring the proportions of head-following dependents:

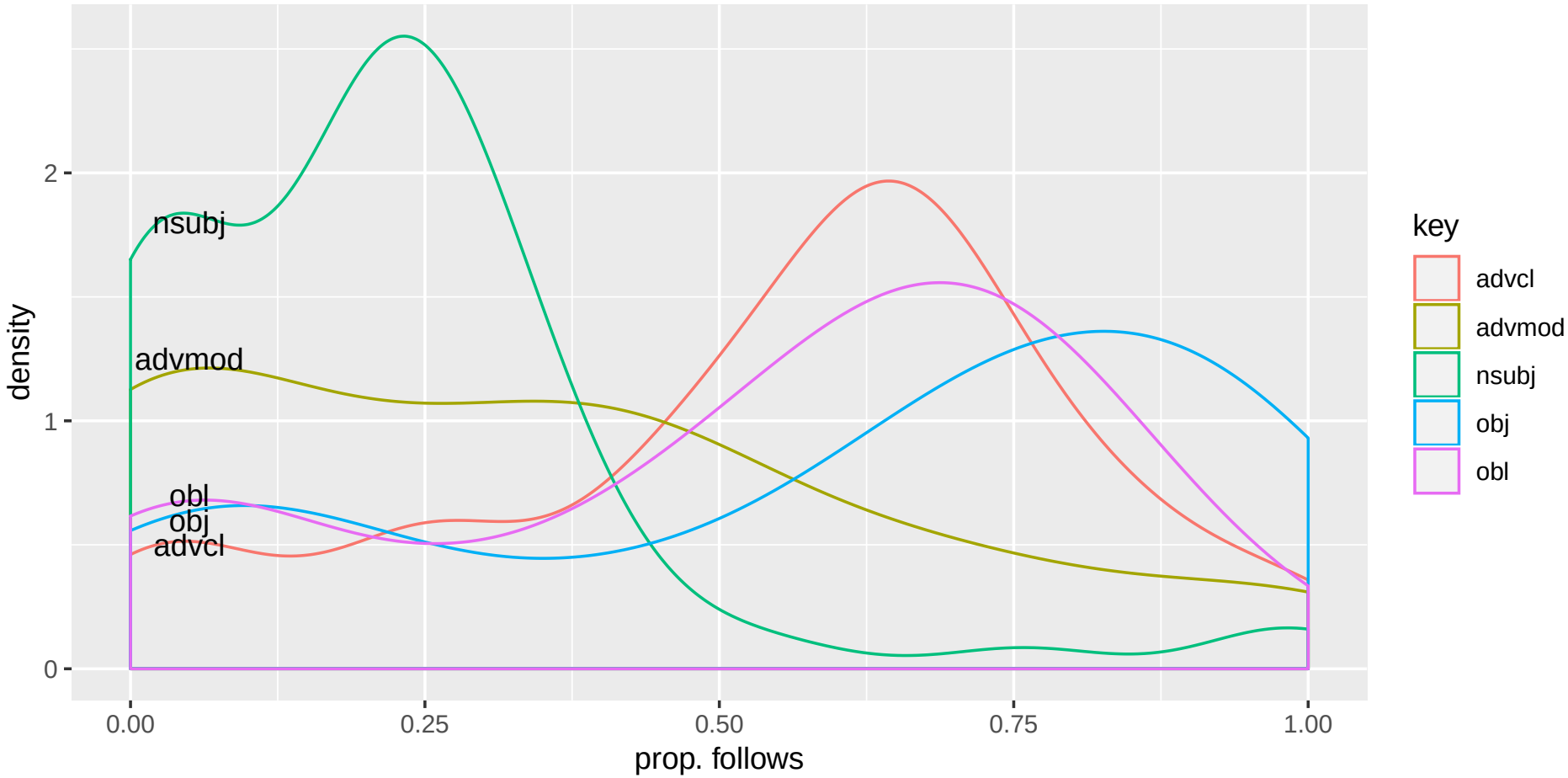
1. the density distributions of head-following dependents
2. the order correlations among noun dependents as well as among verb dependents
   (intra-categorial correlations)
3. predictability of noun dependent orders from verb dependent orders and vice versa
   (cross-categorial correlations)

Distributions

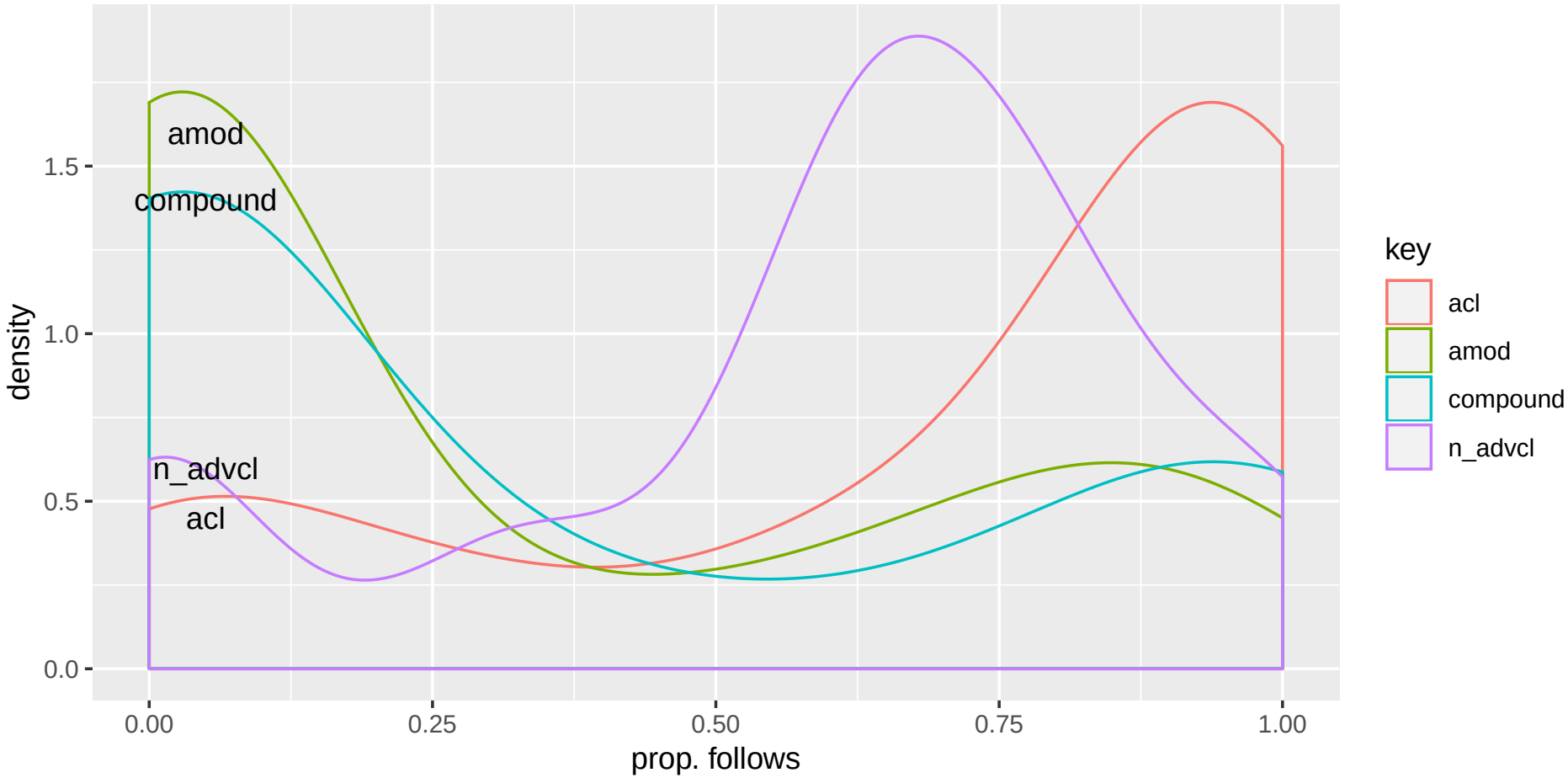
We first explore the distribution of all dependents and their position with respect to their heads.

We look at the density of the proportion of `follows` for each dependent.

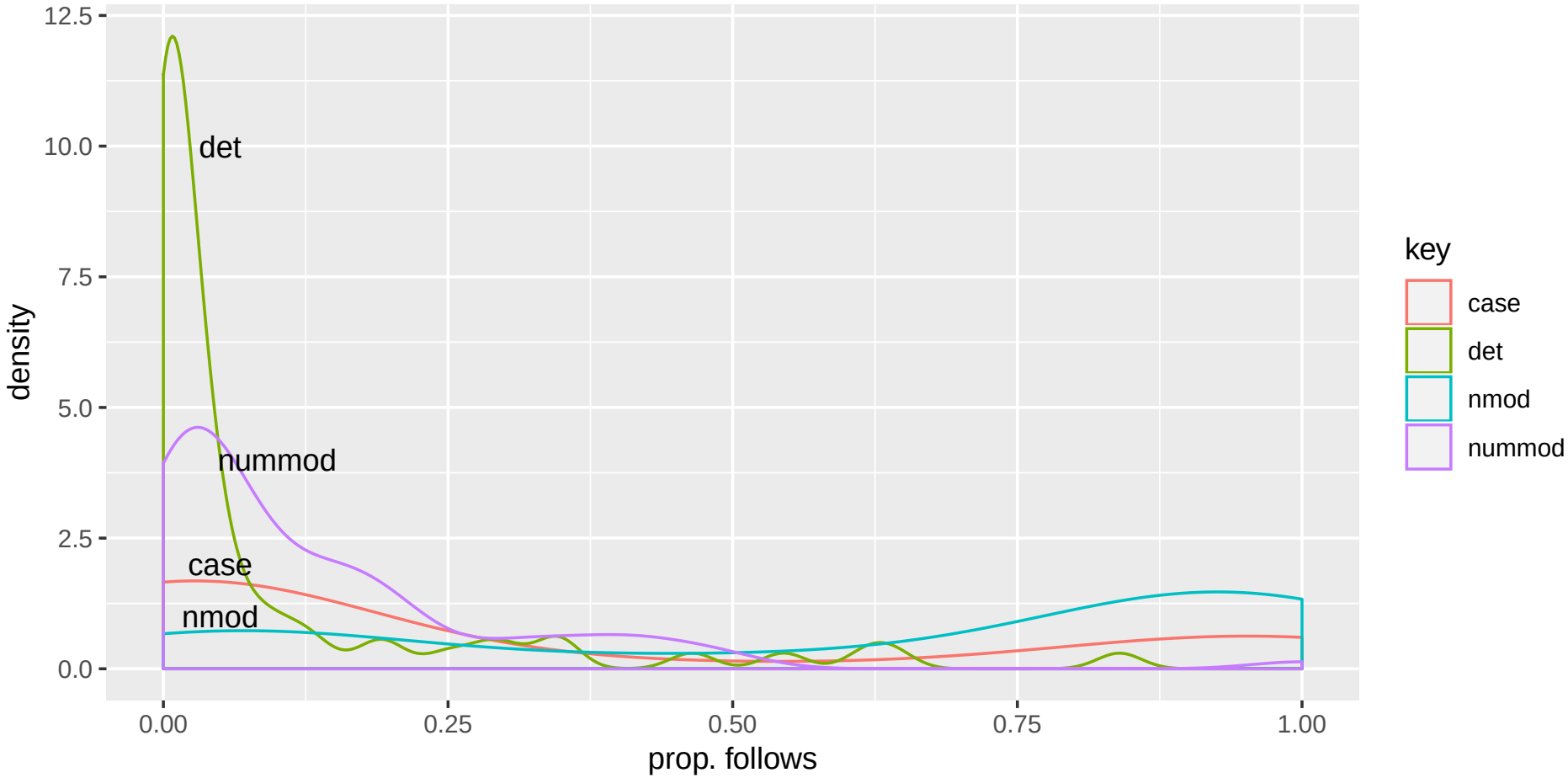
Density distribution: verb dependents



Density distribution: noun dependents



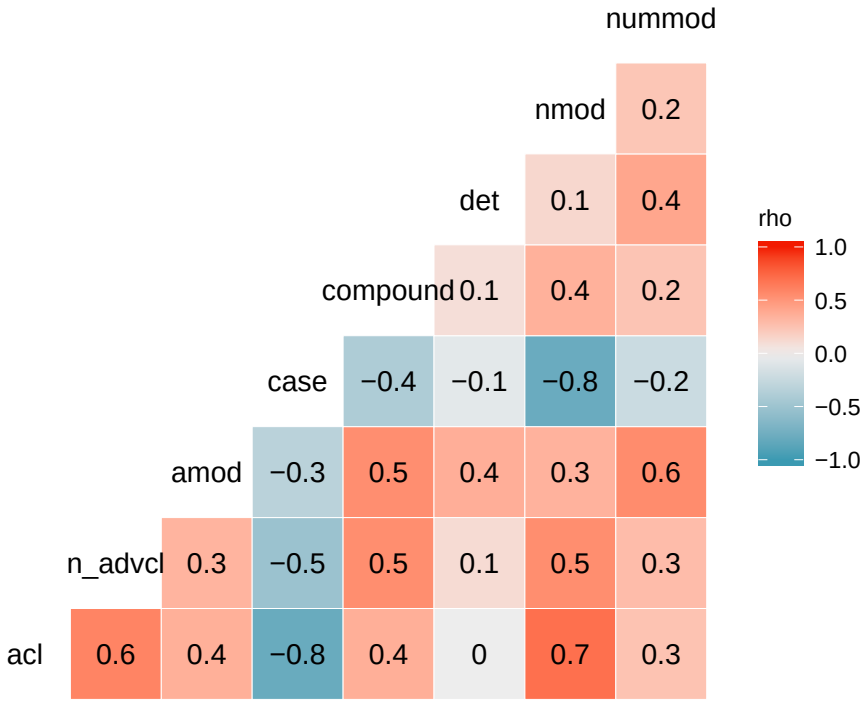
Density distribution: noun dependents



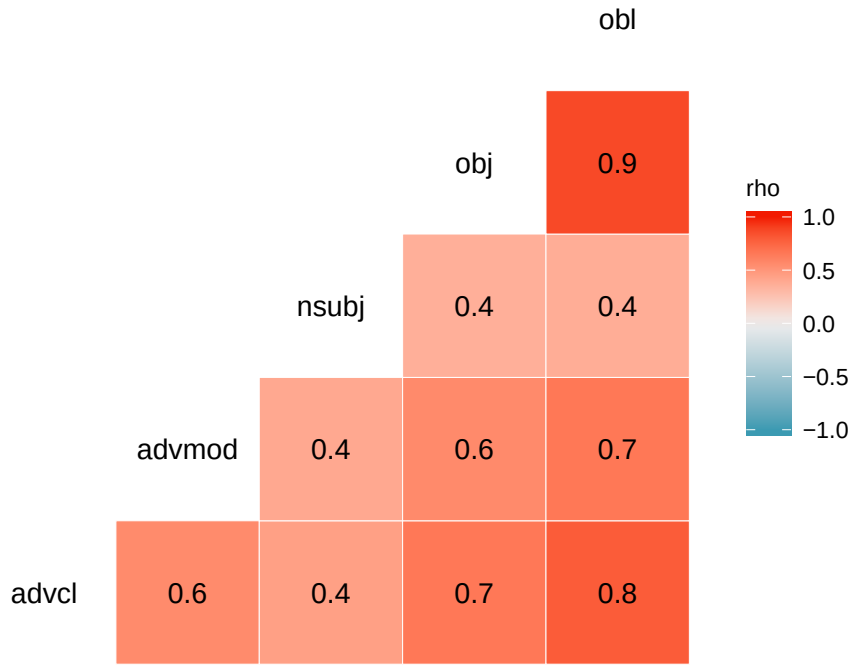
Intra-categorial correlations

We see how verb dependents and noun dependents are correlated among them.

noun dependents

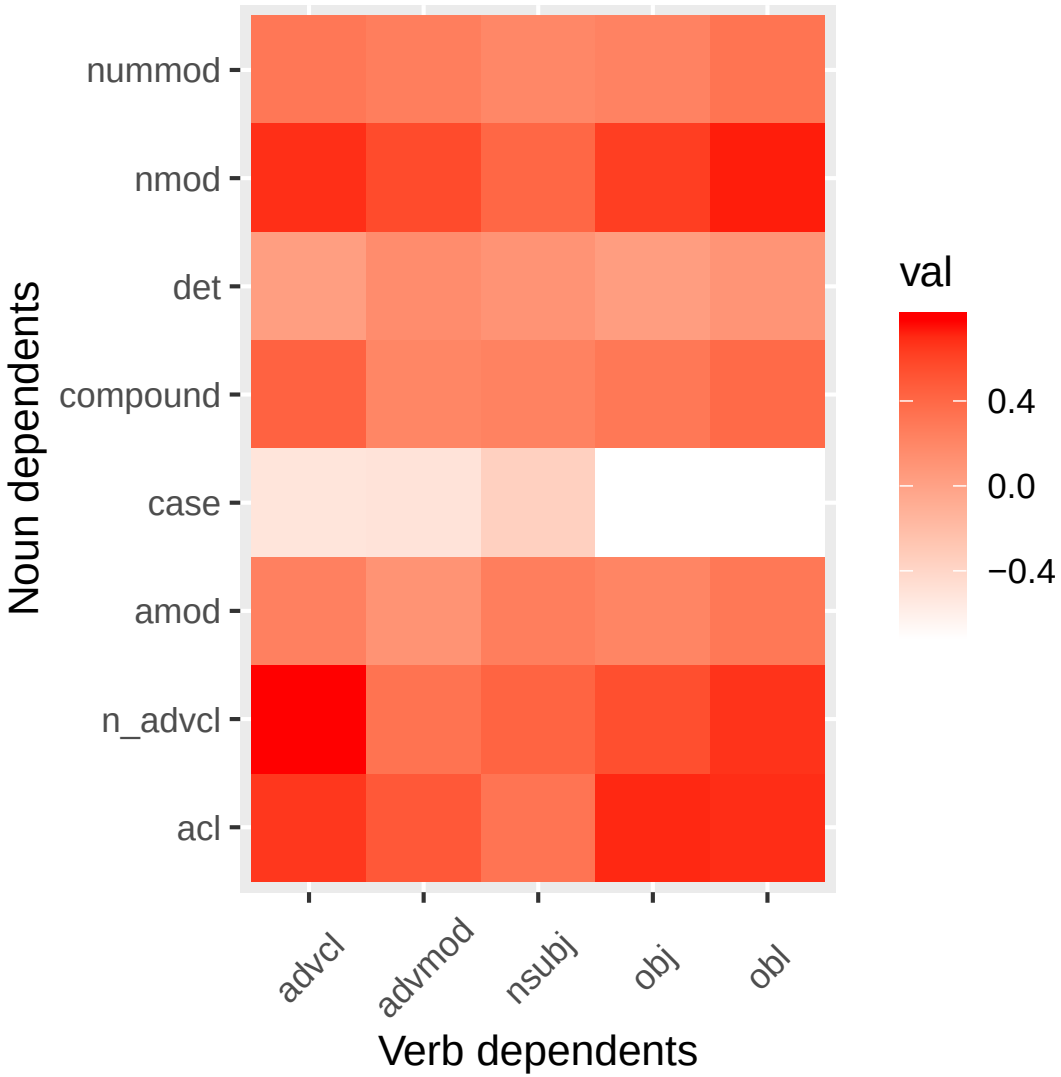


verb dependents



Cross-categorial correlations

We see how verb noun dependents are correlated with verb dependents.



Models

We fitted beta regression models for each factor (verb or noun dependent) as a dependent variable, and using `family` as a random effect.

To prevent overfitting we carried out stepwise factor elimination.

For each model, we calculated the marginal and conditional R^2 values following the method developed by (Nakagawa, Johnson, and Schielzeth, 2017; Nakagawa and Schielzeth, 2013).

R^2

We used:

- Marginal R^2: Portion of the data explained by the fixed effects (dependents).
- Conditional R^2: Portion of the data explained by the fixed (dependents) and random (families) effects.

This is a reasonable way to evaluate model performance, as well as to know how much of the variation is due to factor correlations, and how much to family biases.

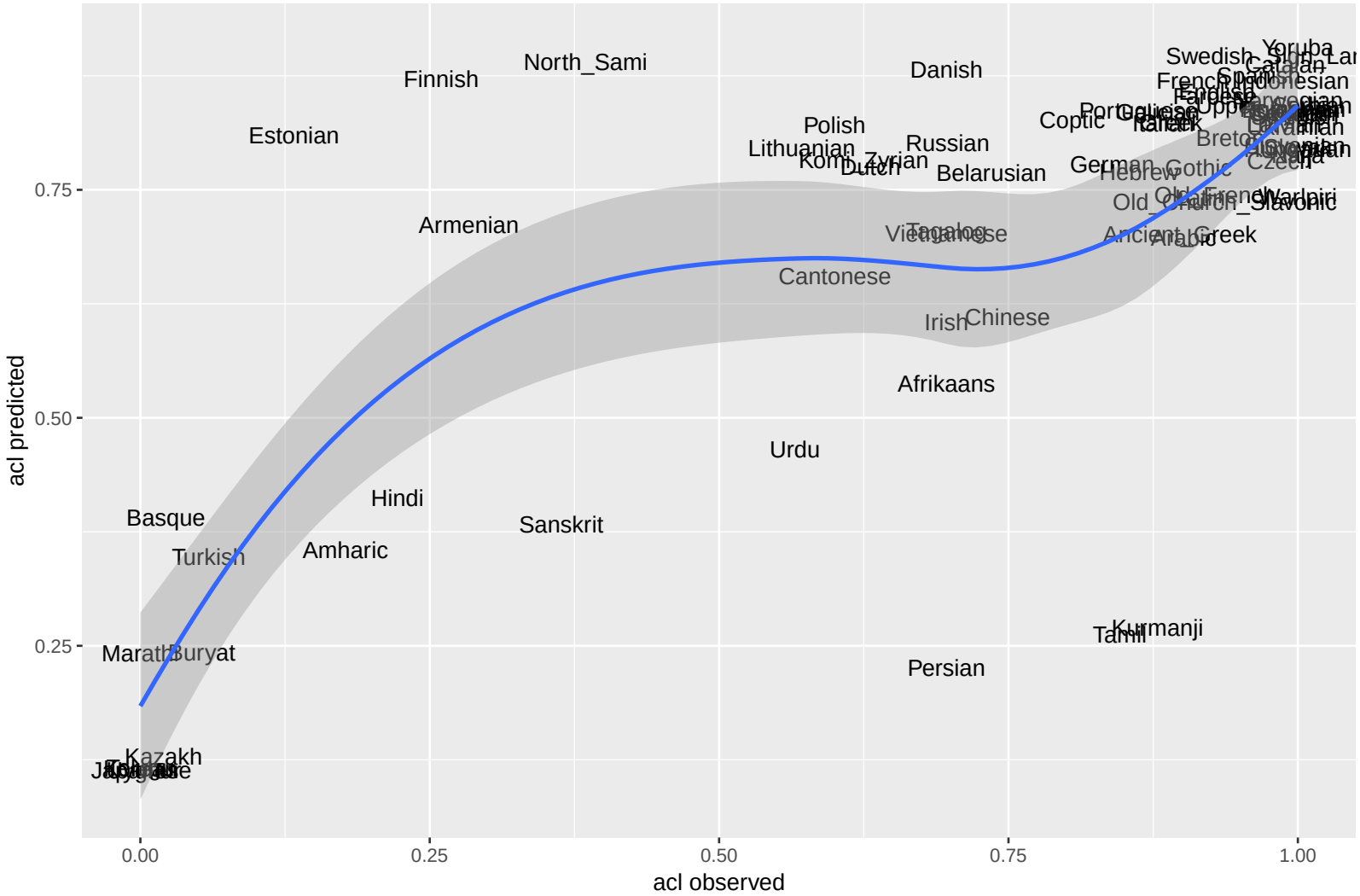
Models predicting noun dependents

predicted	intercept	advcl	nsubj	nsubj:obj	obj	obj:obl	obl	R2_m	R2_c
acl	0.02	2.02	-1.43		6.39			0.462	0.462
advcl	-1.29				0.94		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	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

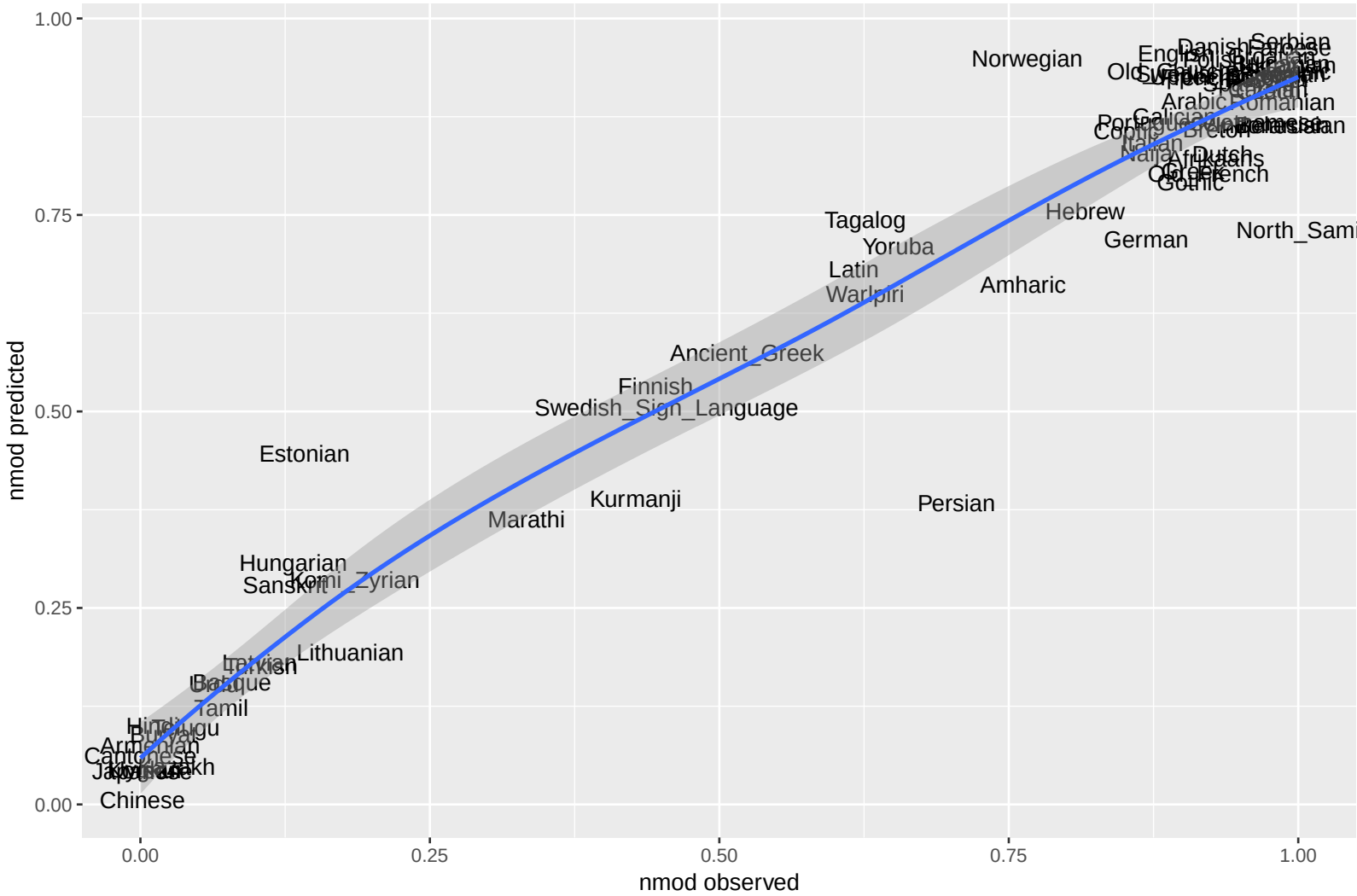
Models predicting noun dependents

predicted	intercept	acl	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

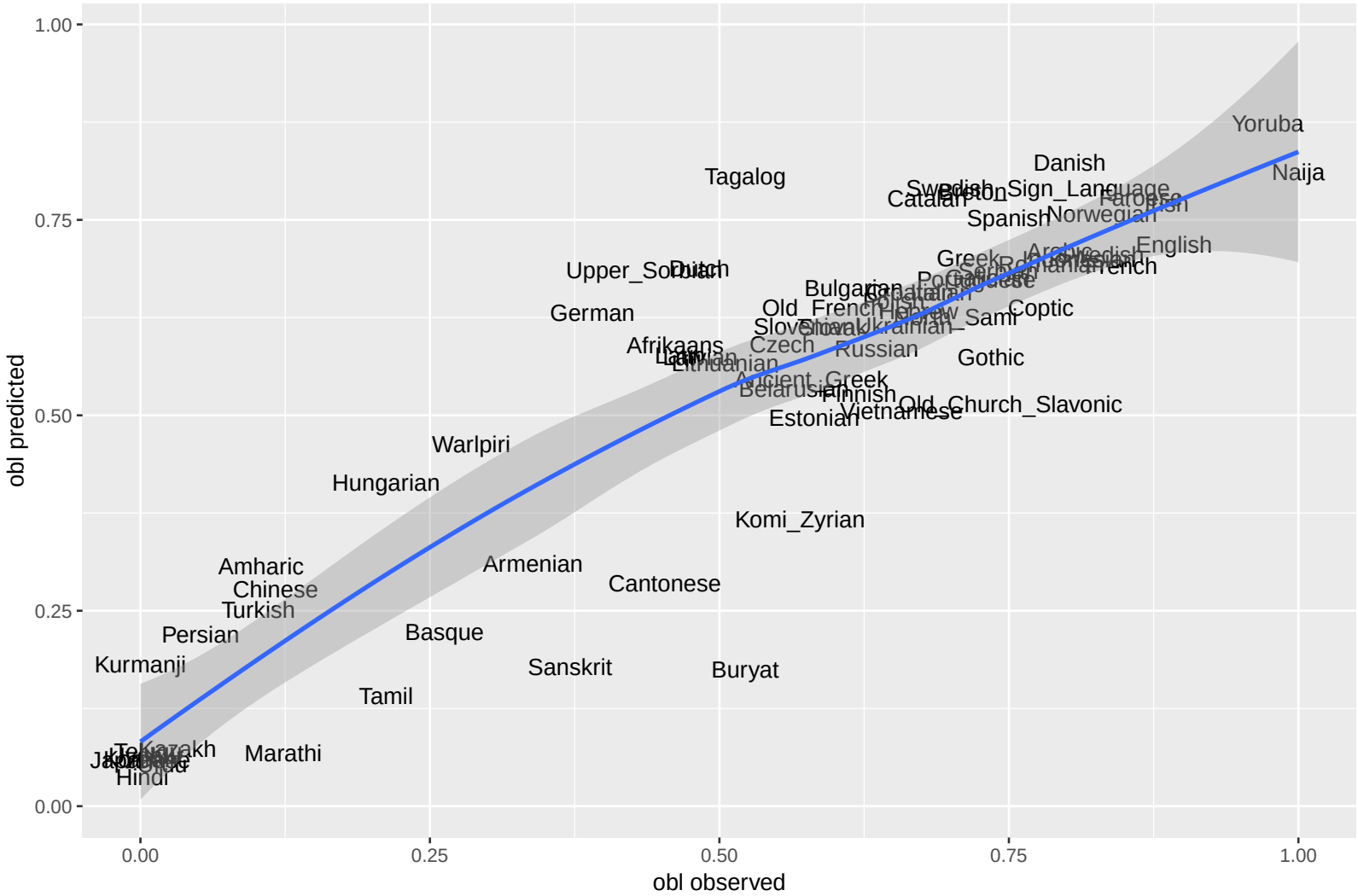
Fitted models



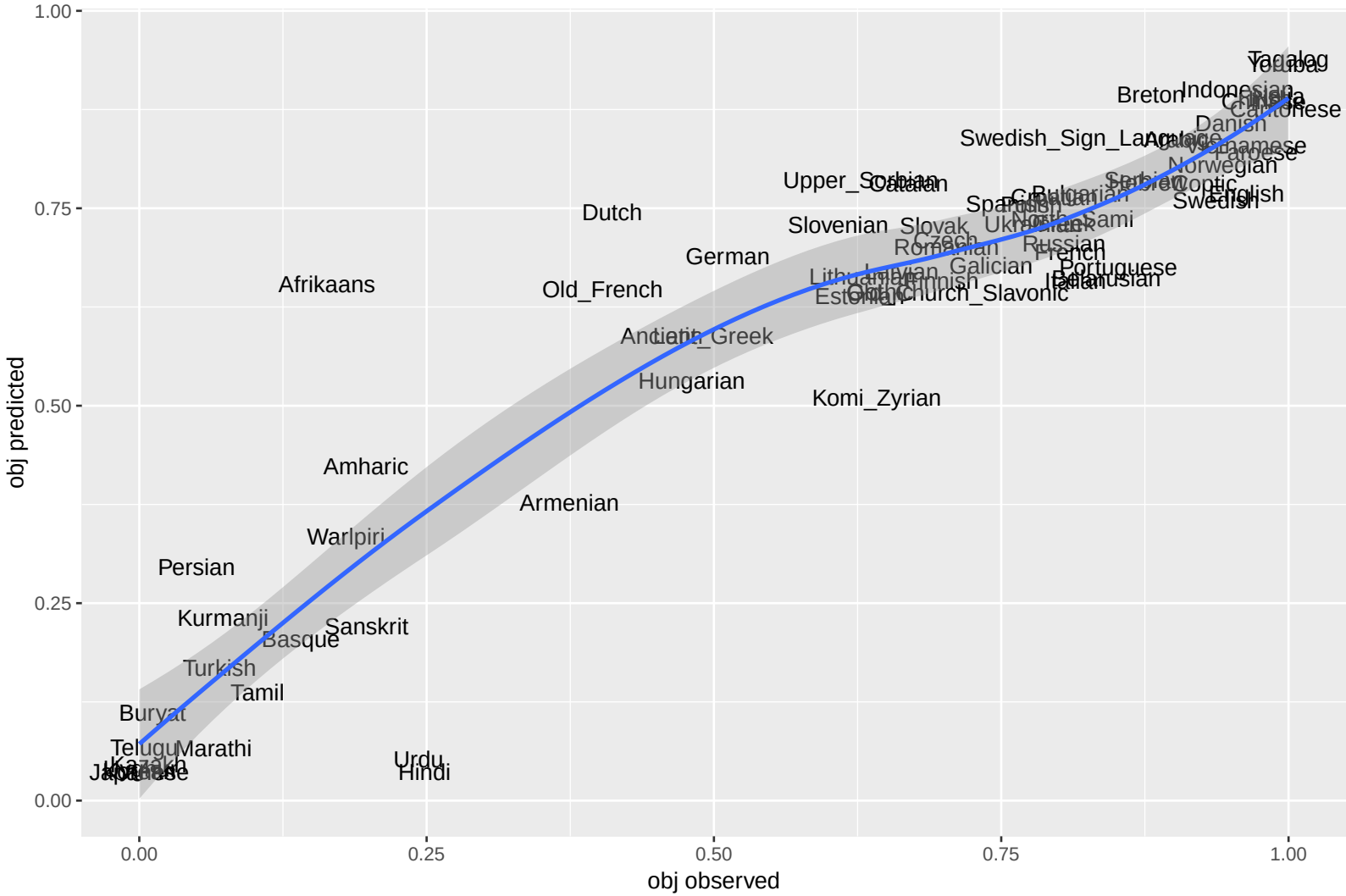
Fitted models



Fitted models



Fitted models



Concluding remarks

Using treebanks helps to gain new insights on word order typology.

- **Gradience** We should rethink the classic word order correlations as being gradient instead of categorical.
- **Order consistency** The consistency of dependent orders vary across different types of dependents across languages: some dependents (e.g. `det`, `acl`) show a clear tendency towards preceding or following the head, while others (e.g. `nmod`, `case` do not.
- **Intra-categorical correlations** For both verb and noun dependents, we find some strong intra-categorical order correlations; as well as negative correlations between `case` and other nominal dependents.
- **Cross-categorical correlations** Dependents that are good predictors are not necesseraly well predicted themselvs (`obl` is a better predictor than `obj` for nominal dependent orders, but both can be predicted equally well).

  Different types of dependent orders are more (e.g. `case`) or less (e.g. `det`) sensitive to family biases.

Future work

We see two potential paths for future work:

- distinguish between different main and subordinate clauses, since subordinate clauses have been shown to be more conservative syntactically (e.g. Bybee, 2002)
- convert the UD format to some other linguistic annotation (HPSG, LFG, TAG, etc.) and see whether the theoretical elements of these theories improve the cross-linguistic patterns

Thank you!

Bibliography I

Bickel, Balthasar (2008). "A Refined Sampling Procedure for Genealogical Control". In: *Sprachtypologie und Universalienforschung* 61, pp. 221–233.

Bybee, Joan L. (2002). "Main Clauses Are Innovative, Subordinate Clauses Are Conservative: Consequences for the Nature of Constructions". In: *Complex Sentences in Grammar and Discourse Essays in Honor of Sandra A. Thompson*. Ed. by Joan L. Bybee and Michael Noonan. Amsterdam: Benjamins, pp. 1–18.

Cristofaro, Sonia (2018). *Processing Explanations of Word Order Universals and Diachrony: Relative Clause Order and Possessor Order*. Paris, INALCO.

Dryer, Matthew S. (1989). "Article-Noun Order". In: *Chicago Linguistic Society* 25, pp. 83–97.

– (1991). "SVO Languages and the OV : VO Typology". In: *Journal of Linguistics* 27.2, pp. 443–482.

– (1992). "The Greenbergian Word Order Correlations". In: *Language* 68.1, pp. 81–138.

Bibliography II

Dryer, Matthew S. (2009). "The Branching Direction Theory of Word Order Correlations Revisited". In: *Universals of Language Today*. Ed. by Sergio Scalise, Elisabetta Magni, and Antonietta Bisetto. Studies in Natural Language and Linguistic Theory. Dordrecht: Springer, pp. 185–207.

– (2019). "On the Order of Demonstrative, Numeral, Adjective and Noun". In: *Language*.

Greenberg, Joseph Harold, ed. (1963). *Universals of Language*. Cambridge, MA: MIT Press.

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

– (1994). *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

– (2014). *Cross-Linguistic Variation and Efficiency*.

Nakagawa, Shinichi, Paul CD Johnson, and Holger Schielzeth (2017). "The Coefficient of Determination R2 and Intra-Class Correlation Coefficient from Generalized Linear Mixed-Effects Models Revisited and Expanded". In: *Journal of the Royal Society Interface* 14.134.

Bibliography III

Nakagawa, Shinichi and Holger Schielzeth (2013). "A General and Simple Method for Obtaining R2 from Generalized Linear Mixed-Effects Models". In: *Methods in Ecology and Evolution* 4.2, pp. 133–142.

Nivre, Joakim et al. (2016). "Universal Dependencies v1: A Multilingual Treebank Collection.". In: *LREC*.

Payne, Doris L., ed. (1992). *Pragmatics of Word Order Flexibility*. Amsterdam: Benjamins.

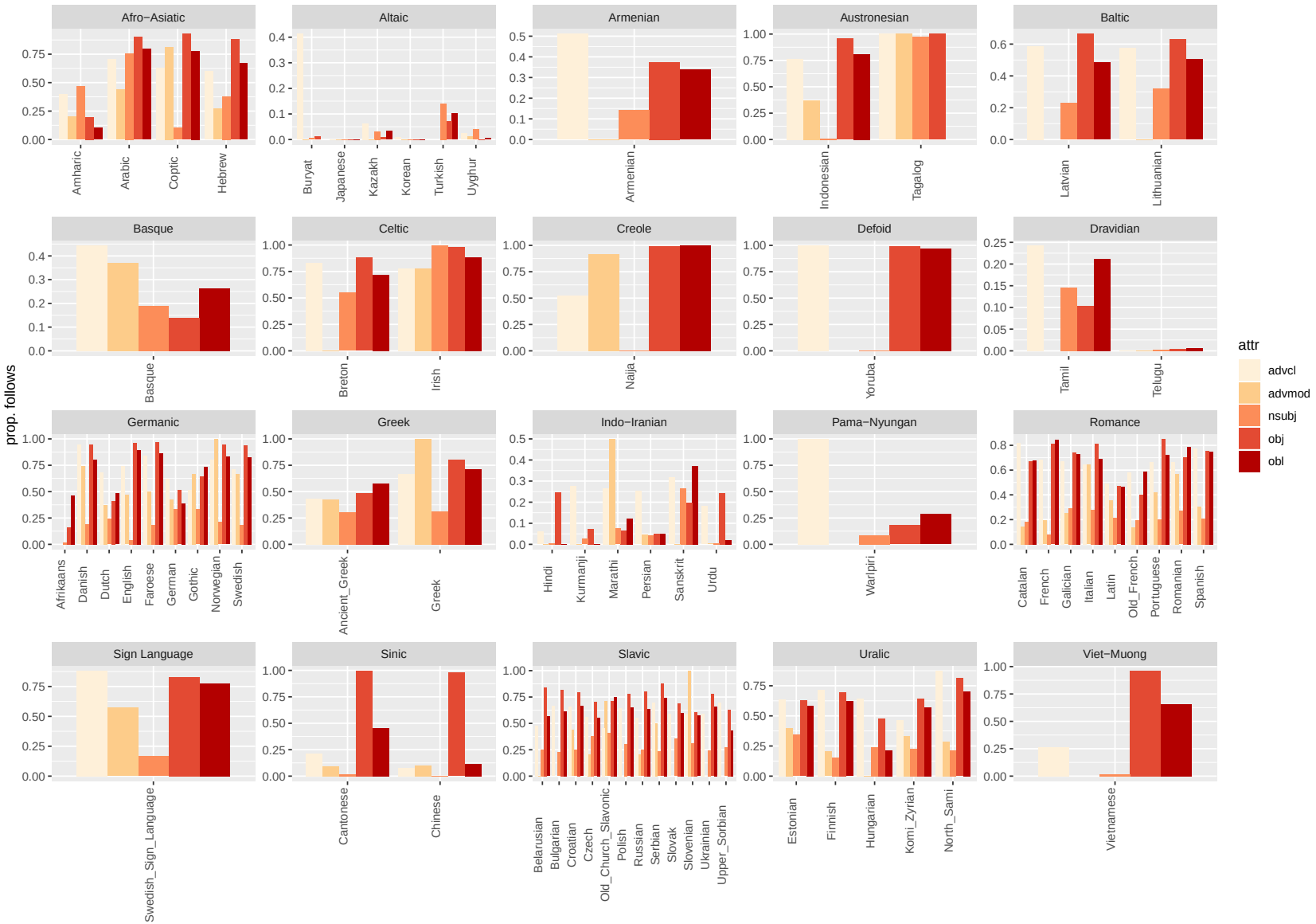
Siewierska, Anna (1988). *Word Order Rules*. London: Croom Helm.

Song, Jae Jung (2009). "Word Order Patterns and Principles: An Overview". In: *Language and Linguistics Compass* 3.5, pp. 1328–1341.

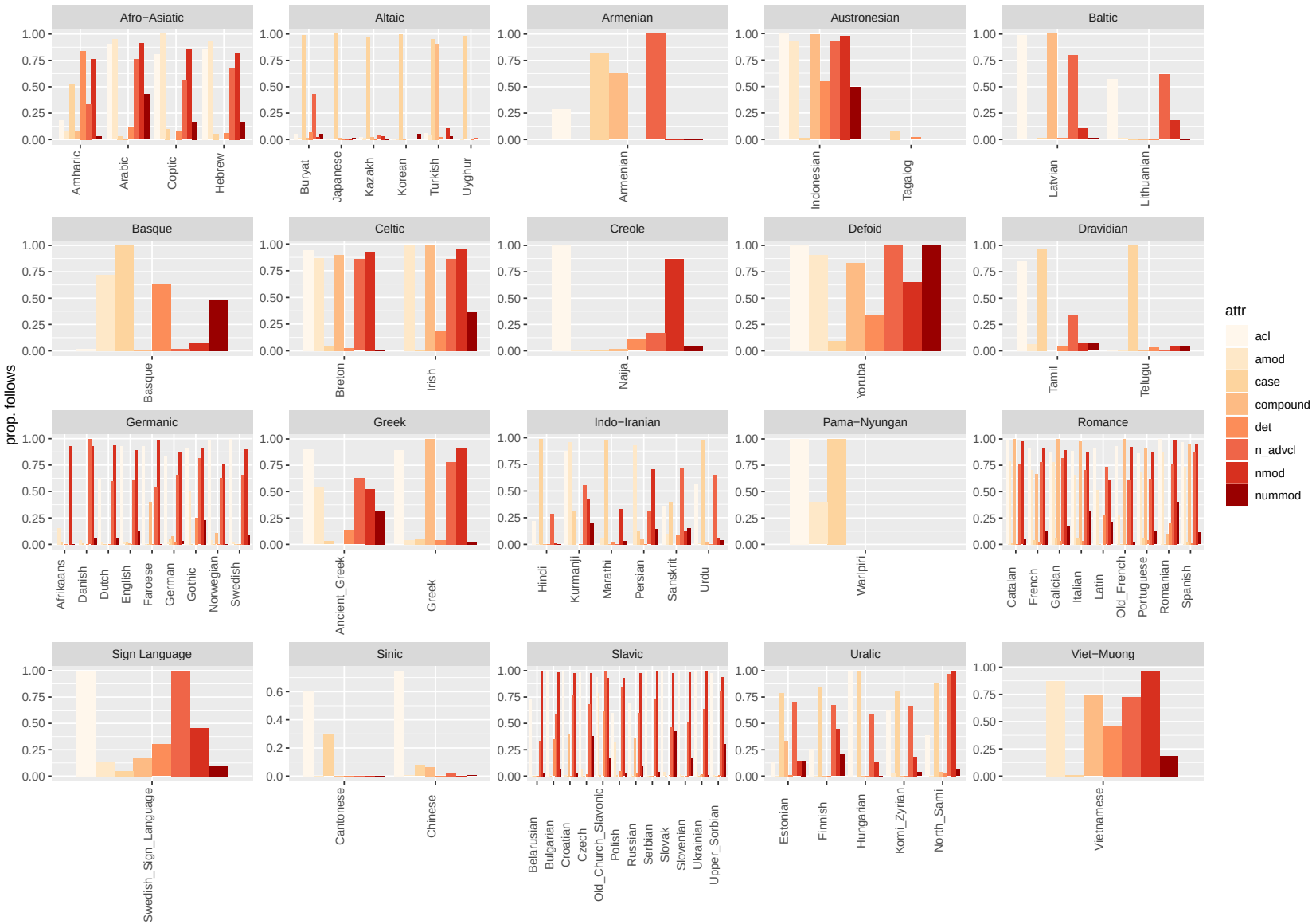
Languages

Afrikaans, Amharic, Ancient Greek, Arabic, Armenian, Bambara, Basque, Belarusian, Breton, Bulgarian, Buryat, Cantonese, Catalan, Chinese, Coptic, Croatian, Czech, Danish, Dutch, English, Erzya, Estonian, Faroese, Finnish, French, Galician, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Komi Zyrian, Korean, Kurmanji, Latin, Latvian, Lithuanian, Maltese, Marathi, Naija, North Sami, Norwegian, Old Church Slavonic, Old French, Persian, Polish, Portuguese, Romanian, Russian, Sanskrit, Serbian, Slovak, Slovenian, Spanish, Swedish, Swedish Sign Language, Tagalog, Tamil, Telugu, Thai, Turkish, Ukrainian, Upper Sorbian, Urdu, Uyghur, Vietnamese, Warlpiri, Yoruba

Verb dependents



Noun dependents



Models predicting noun dependents

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
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