Title

Verb-argument lability and its correlations with other typological parameters: a quantitative corpus-based study

Authors

Hawkins, John

Levshina, Natalia

Publication Date

2022

Peer reviewed

Linguistic Typology at the Crossroads 1-1 (2021): XXX ISSNXXXX-XXXX

Verb-argument lability and its correlations with other typological parameters: a quantitative corpus-based study

ONCE THE PAPER HAS BEEN ACCEPTED, INSERT YOUR NAME IN CHARIS SIL POINT 12 SINGLE SPACING ALL CAPS CENTRED[1]

[1]ONCE THE PAPER HAS BEEN ACCEPTED, INSERT YOUR AFFILIATION HERE IN CHARIS SIL POINT 11 SINGLE SPACING SMALL CAPS CENTRED

Submitted: we will insert dates Revised version: we will insert dates
Accepted: we will insert dates Published: we will insert dates

Abstract

We investigate the correlations between A- and P-lability for verbal arguments with other typological parameters using large, syntactically annotated corpora of online news in 28 languages. To estimate how much lability is observed in a language, we measure associations between Verbs or Verb + Noun combinations and the alternating constructions in which they occur. Our correlational analyses show that high P-lability scores correlate strongly with the following parameters: little or no case marking; weaker associations between lexemes and the grammatical roles A and P; rigid order of Subject and Object; and a high proportion of verb-medial clauses (SVO). Low P-lability correlates with the presence of case marking, stronger associations between nouns and grammatical roles, relatively flexible ordering of Subject and Object, and verb-final order. As for A-lability, it is not correlated with any other parameters. A possible reason is that A-lability is a result of more universal discourse processes, such as deprofiling of the object, and also exhibits numerous lexical and semantic idiosyncrasies. The fact that P-lability is strongly correlated with other parameters can be interpreted as evidence for a more general typology of languages, in which some tend to have highly informative morphosyntactic and lexical cues, whereas others rely predominantly on contextual environment, which is possibly due to fixed word order. We also find that P-lability is more strongly correlated with the other parameters than any of these parameters are with each other, which means that it can be a very useful typological variable.

Keywords: verb-argument lability; corpora; Universal Dependencies; word order; case marking; tight-fit and loose-fit languages.

1. Theoretical background

The usefulness of a typological parameter depends on how many other parameters it helps us to predict. Greenberg's (1963) word order correlations have been such a major achievement in linguistics because they connected many diverse and seemingly unrelated word order patterns. In this paper we demonstrate that the strength of attraction of verbs (as well as their arguments) to specific subcategorization frames, which can be defined in terms of verb-argument lability, can be a useful parameter, because it is strongly correlated with many others.

https://doi.org/....

The attraction of verbs to specific subcategorization frames has been argued to be a part of the typology of tight-fit versus loose-fit languages. The terms were coined by Hawkins (1986: 121– 127, 1995; see also Müller-Gotama 1994). Generally speaking, tight-fit languages have unique surface forms that map onto more constrained meanings, whereas loose-fit languages have vaguer forms with less constrained meanings. More specifically, grammatical roles in tight-fit languages have a narrower semantic range than grammatical roles in loose-fit languages. For example, the languages Jakaltek and Halkomelem strictly exclude inanimate subjects in transitive clauses (Aissen 2003), while English and Swedish merely strongly disprefer them (Dahl 2000). There are also more gradient distinctions. For example, while English and German allow for different kinds of subjects, English is still considered looser than German, and also than Russian and Korean.

The strength of the associations with grammatical roles is correlated with other linguistic parameters, including more explicit grammatical coding (e.g., formal case marking and use of complementizers and relativizers), avoidance of raisings and long distance WH-movements. Tight-fit languages have fewer cases of category ambiguity. For example, the English word *book* can be both a noun and a verb, while in German the corresponding noun and verb have different forms, *Buch* – *buchen*. Moreover, verb-final languages are often semantically tight.

If these parameters change, they often change together. English is a well-known case (Hawkins 1986). The loss of morphology correlated, in particular, with the emergence of SVO order, long distance movement and raising, greater category ambiguity and other features, including fewer restrictions on the semantics of syntactic arguments. In contrast, German is more conservative. It preserves case marking, verb-final order (for all verbs in subordinate clauses and for non-finite verbs in main clauses) and it still has some variability in the order of Subject and Object. In addition, German has fewer instances of category ambiguity, tighter associations between semantics and roles, and very limited examples of raising. Generally speaking, English is more structurally ambiguous than German (Hawkins 2019). For example, raising and control constructions are not distinguished formally in surface structure. Compare *Sue happened to win the lottery* (raising) and *Sue hoped to win the lottery* (control). In German, these are distinguished by formally different constructions. In Hawkins' terminology (2019), English relies more on word-external properties to derive meanings from ambiguous or vague surface forms, whereas German relies more on distinct grammatical and lexical patterns and on word-internal properties.

Taking the perspective of processing typology (Hawkins 1994, 2004), we can explain the correlations discussed above on the basis of different strategies for optimizing language processing during communication. Languages that have the verb at the end, need to rely on semantic and formal cues for the assignment of thematic roles early in the sentence. Otherwise, the hearer will need to perform a costly reanalysis. Relatively flexible order of Subject and Object also necessitates the use of case marking and semantic restrictions for different roles. In contrast, rigid

https://doi.org/....

word order with the verb in the middle helps language users to distinguish between Subject and Object better in a noisy channel (Gibson et al. 2013), making case markers dispensible, and it also helps to resolve other ambiguities (Levshina 2021).

Importantly for our study, verbs in loose-fit languages have a broader set of subcategorization frames than in tight-fit languages. For example, the English verb *open* can be both transitive (e.g, *I open the door*) and intransitive (*The door opened*), while German distinguishes formally between the transitive *öffnen* 'open (tr.)' and the reflexive verb *sich öffnen* 'open (intr.)'.

In this paper we will consider two types of verb-argument lability, which are known as A-lability and P-lability. We speak of A-lability when the A argument of a transitive clause remains the same, but the P argument can be removed. In other words, with the same verb, the A-argument can turn into an S-argument: A=S (Dixon 1994). Examples are the unspecified object alternation (1a), the understood body-part alternation (1b) and the characteristic property alternation (1c) (Levin 1993).

(1) a. Unspecified object alternation
Jack ate the cake. - Jack ate.
b. Understood body-part alternation
The Queen waved her hand at the crowd. - The Queen waved at the crowed.
c. Characteristic property alternation
The dog bites strangers. - The dog bites.

P-lability is observed when the same argument can be used as intransitive subject (S) and as direct object (P) with the same verb, or S=P (Dixon 1994). Examples are the causative-inchoative alternation (2a), the middle alternation (2b) and the induced action alternation (2c) (Levin 1993).

(2) a. Causative-inchoative alternation
The boy broke the vase. - The vase broke.
b. Middle alternation
The publisher sells the book. - The book sells well.
c. Induced action alternation
She jumped the horse over the fence. - The horse jumped over the fence.

The above-mentioned contrast between English *open* (transitive, intransitive) and German *öffnen* 'open (tr.)' and intransitive reflexive verb *sich öffnen* 'open (intr.)' suggests that English has more P-lability (causative-inchoative alternations, in particular) than German. However, this has not yet been examined in corpora and using quantitative measures.

In this paper we fill this gap, measuring A- and P-lability in languages with the help of large corpora, which are described in Section 2. We compute the Mutual Information between verbs, or combinations of verbs and nouns, and the alternating constructions in which they occur. The

https://doi.org/....

procedure and the scores are presented in Section 3. Then, we test the correlations between different measures of A- and P-lability and four other variables which have been used in the literature on tight-fit and loose-fit languages and more generally: word order rigidity; the position of the verb in the sentence; case marking; and the strength of associations between nouns and the grammatical roles of Subject and Object (Section 4). Finally, in Section 5 we discuss our findings and conclusions.

2. Data and method

We used the Leipzig Corpus Collection (Goldhahn et al. 2012)[1]. We first selected 30 online news corpora with 1M sentences in each of: Arabic, Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek (modern), Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Latvian, Lithuanian, Persian, Portuguese, Romanian, Russian, Slovenian, Spanish, Swedish, Tamil, Turkish and Vietnamese. The corpora were annotated with the Universal Dependencies pipeline *udpipe* (Wijffels, Straka & Straková 2018), which allowed us to extract the subject, the direct object, the predicate, as well as their lemmas, part of speech and morphological features, as well as other useful information. Due to our doubts about the quality and consistency of verb lemmatization in the data from Tamil and Turkish, these languages were later excluded. This is why we had 28 languages in the final sample.

In order to find patterns of A-lability, we extracted the frequencies of all verb lemmas with the same noun in subject position (represented by the Universal Dependency 'nsubj') with and without any kind of nominal or pronominal direct object (the Universal Dependency 'obj'). Consider the examples in Table 1.

A-lability Frequencies			
Verb	**Subject**	**Transitive**	**Intransitive**
be	idea	0	140
learn	student	21	35
play	team	55	47

Table 1: Examples of frequencies relevant for A-lability.

The table shows that the verb *be* with the noun *idea* as subject occurs 140 times (e.g., the idea was...), only in intransitive clauses. This is not surprising. The combination *student + learn* occurs 21 times with a direct object (e.g., *the students learn languages*) and 35 times without (e.g., *the students learn*). This is an example of A-lability.

In order to identify examples of P-lability, we extracted the frequencies of all verb lemmas (only predicates of main clauses) with the same noun occurring as direct object and as intransitive subject. Consider the examples in Table 2.

P-lability Frequencies

[1] http://wortschatz.uni leipzig.de/en/download/

https://doi.org/....

Verb	Noun	Intr. subject + Verb	Verb + Object
die	people	64	0
open	door	36	149
begin	work	35	33

Table 2: Examples of frequencies relevant for P-lability.

The numbers should be read as follows. The verb *die* occurs with the noun *people* only as an intransitive subject (64 times), and never as an object. The verb *open* with the noun *door* as intransitive subject (*The door opened*) occurs 36 times, and as a direct object (*I opened the door*) 149 times. This is an example of P-lability.

If we simply counted intransitive and transitive uses of verbs, it would be impossible to distinguish A-lability from P-lability. As will be shown below, making this distinction is crucial, and it is why it was necessary to control for the nouns as A, P or S.

Note that we only selected the verbs that served as predicates of main clauses. Particle verbs and verbs with separable prefixes were treated as one lemma (e.g., *break+out*, *um+leiten*). We also excluded verbs with reflexive, passive, antipassive, middle morphology or auxiliaries because of the substantial cross-linguistic differences in their semantics, formal properties and annotation. One consequence of this decision is that we are primarily measuring looseness vs. non-looseness (the formal marking of which can be quite variable across languages). We also excluded ditransitive clauses. The measures of lability presented below are based only on combinations of verbs and nouns that occur ten times or more in a corpus, in order to make sure that the zero occurrences of nouns in certain alternations were not due to data sparseness.

Lability measures were computed using two methods. According to the first, we controlled for both the verb and the noun, which means that our measures took into account not only the flexibility of the verb with regard to the alternation variants, but also the flexibility of the noun with regard to the roles of A or S (in cases of A-lability) and S or P (in cases of P-lability). In the second method, we took into account the verbs only, adding up the frequencies of all nouns occurring as A and S, or as S and O with a given verb.

3. Measures of lability

3.1. Mutual Information related to A-lability

Using the kinds of frequencies shown in Table 1, we computed Mutual Information (MI) related to A-lability for twenty-eight languages. For Verb + Noun combinations, the formula was as follows:

https://doi.org/....

(3) $$I(V \wedge N; Cx) = \sum_{i,j} p(V \wedge N_i, Cx_j) \log \frac{p(V \wedge N_i, Cx_j)}{p¿¿}$$

where *V&N* represents Verb + Noun combinations, and *Cx* stands for the constructional alternation, which includes the transitive construction ('nsubj' + Verb + some object) and the intransitive construction ('nsubj' + Verb). The higher MI, the stronger the association between the Verb + Noun combination and a particular construction. Therefore, high MI scores suggest weak lability, characteristic of a tight-fit language, and low MI scores correspond to strong lability, characteristic of a loose-fit language.

For verbs only, the formula was as follows, where V stands for a verb:

(4) $$I(V; Cx) = \sum_{i,j} p(V_i, Cx_j) \log \frac{p(V_i, Cx_j)}{p¿¿}$$

Both types of scores are shown in Figure 1. The languages are ordered by their MI scores based on Verb + Noun combinations, but the two types of scores are strongly correlated: Spearman's rank-based correlation coefficient is 0.97, and the *p*-value < 0.0001 (but see a more precise measure with genetic dependencies taken into account in Section 4). This means that the measures represent very similar information. The scores based on verbs only are lower in all languages, but the ordering is more or less the same, as the high correlation coefficient suggests. The highest scores are found in Portuguese, followed by Italian, Hindi, English and Slovene. The lowest score belongs to Lithuanian, followed by Vietnamese, Korean, Arabic and Persian. This ranking is not predictable from any typological, genealogical or areal properties of the languages.

https://doi.org/....

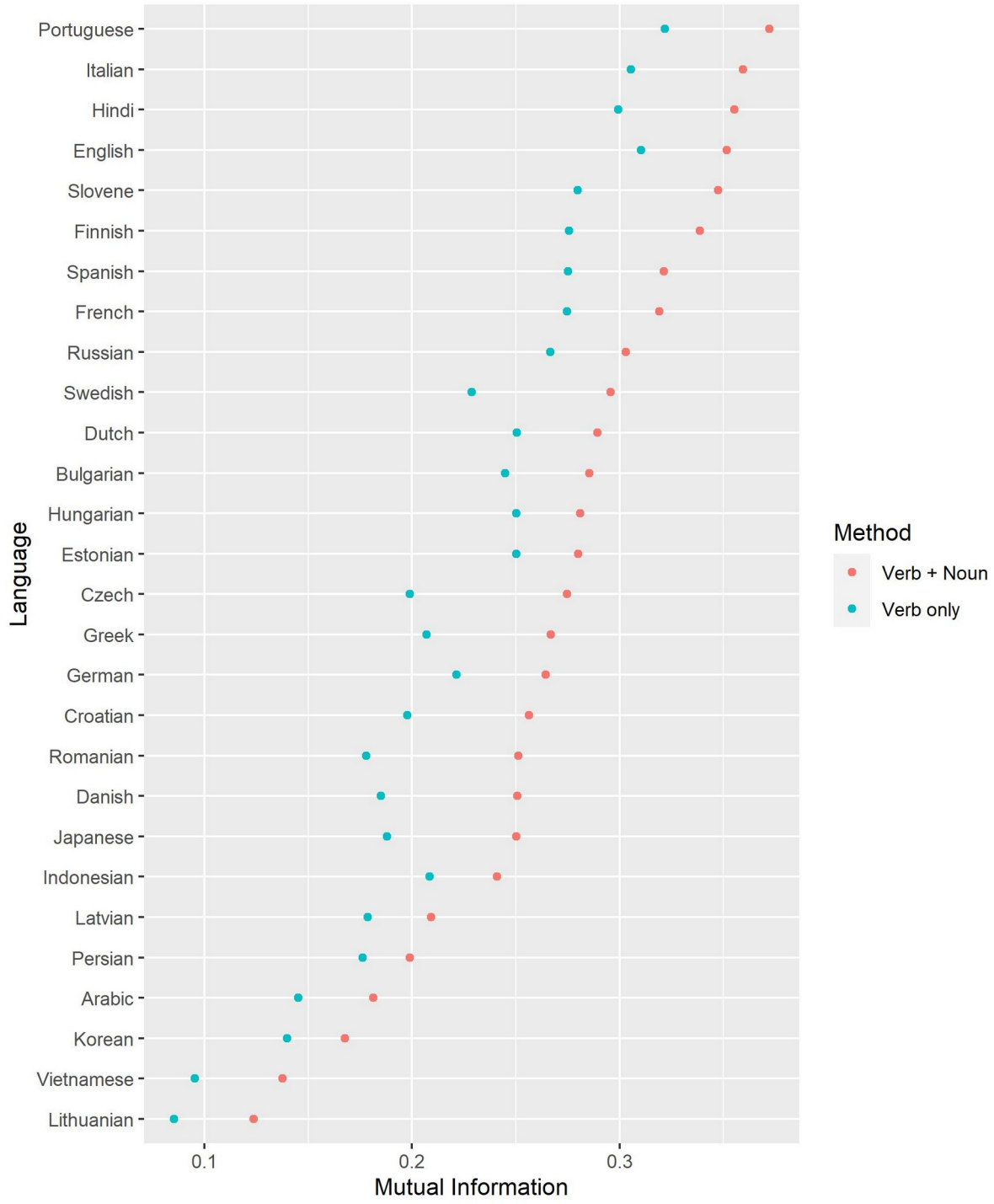

Figure 1: Distribution of MI scores representing A-lability. The greater the score, the weaker this type of lability in a language.

3.2. Mutual Information related to P-lability

To compute MI related to P-lability, we used the same approach as for A-lability, but took the frequencies of verbs and nouns in the construction 'nsubj' + Verb without object and the construction Verb + 'obj' (regardless of the presence or absence of any subject). The two methods, Verb +

https://doi.org/....

Noun (as 'nsubj' or 'obj') and Verb only, yield scores that are highly correlated: Spearman's correlation coefficient rho is 0.96, with the *p*-value < 0.001.

Figure 2 displays both types of MI scores. The top scores belong to Hungarian, Russian, Estonian, Latvian, Korean and Finnish. The high scores mean that the languages have strong associations between the Verb + Noun combinations and the constructions in which they appear as 'nsubj' or 'obj' respectively, characteristic of tight-fit languages. These languages also have formal case marking and relatively free word order of the core arguments. Many of the languages at the top are verb-final, or at least allow for the V-final order. The two languages at the bottom are Indonesian and Vietnamese, followed by English, French and Romanian. These have weaker associations between the Verb + Noun combinations and the constructions in which they appear as 'nsubj' or 'obj'. So they display stronger P-lability characteristic of loose-fit languages. They also have fairly rigid SVO order and no case morphology.

https://doi.org/....

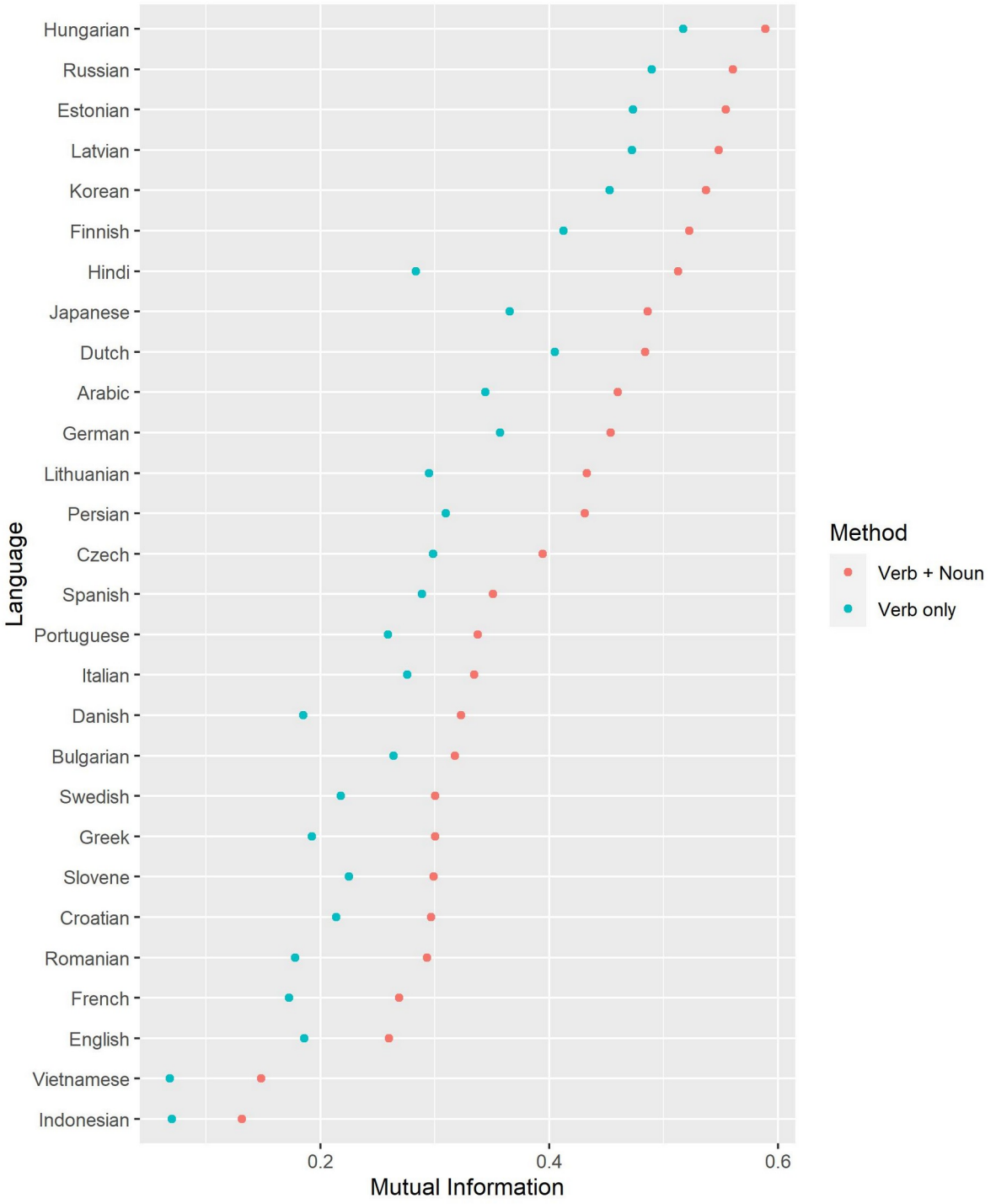

Figure 2: Distribution of MI scores representing P-lability. The greater the score, the weaker this type of lability in a language.

If we compare the range of values in Figure 1 and Figure 2, we see that the MI scores related to A-lability are on average lower than the MI scores related to P-lability. This impression is supported by paired Wilcoxon tests. The difference between the A- and P-lability scores is significant for both

https://doi.org/....

methods (p = 0.028 for verbs only, and p < 0.001 for Verb + Noun combinations). This means that languages are more tolerant with regard to A-lability in general. In addition, the spread of the P-lability scores is greater, which suggests more substantial cross-linguistic differences.

4. Correlations with other typological parameters

In this section we test for correlations between these A-lability and P-lability scores and the following parameters: rigidity of Subject and Object order; position of the lexical verb in the clause; case marking; and associations between lexemes and grammatical roles, which serves as a proxy for semantic tightness. We recycle the data from Levshina (2021), where the parameters were estimated by using the same online news corpora. More specifically, rigidity of Subject and Object order was computed as 1 minus entropy of SO and OS orders, following Shannon (1948), as shown below:

(5) $$H(X) = -\left(P(SO)\log_2 P(SO) + P(OS)\log_2 P(OS)\right)$$

The proportions of SO and OS orders in transitive clauses were computed first based on the corpora, and then these entropy scores were computed (see Levshina 2019). If the proportions of SO and OS orders are equal (0.5), this leads to entropy of 1. If only one of the orders is used (either SO or OS), this leads to zero entropy. Since entropy represents word order variability, we subtracted the entropy scores from one in order to obtain measures of word order rigidity. Lithuanian, Hungarian, Latvian, Czech and Estonian had the lowest scores and therefore the most variable orders, and Indonesian, French, English, Danish and Swedish had the highest scores and thus the most rigid orders. Note that in all languages, the SO order was the more frequent one. So we can speak about the rigidity of SO order. This variable was called "Rigid Order (SO)".

Another measure was the proportion of main clauses with a lexical verb between the Subject and Object. As expected, it was near-zero in verb-final languages, such as Japanese, Korean, Persian and Hindi, and close to one in Indonesian, English, French, Vietnamese and Portuguese. This variable was labelled "Verb between Subj and Obj".

We also took into account how much case marking was present to help in identifying the Subject and the Object. In Levshina (2021), the scores represented Mutual Information between case and the corresponding grammatical roles. For languages with adpositional case marking, the data were extracted automatically. As an illustration, consider the frequencies for Spanish in Table 3.

Case Marking and Grammatical roles in Spanish

https://doi.org/....

Case	Transitive Subject	Direct Object
Zero marking	126,736	569,252
Preposition *a*	0	55,442

Table 3: Frequencies of zero case marking and the direct object marker *a* for Subject and Object in Spanish.

For languages with case morphology, random samples were drawn and analyzed manually. Next, the counts were extrapolated to all occurrences of Subjects and Objects in transitive clauses in a corpus. Consider an illustration in Table 4, which contains frequencies for Russian. In languages with distinct forms for Subject and Object and also forms with case syncretism, as in Russian, these three situations were represented by separate rows.

Case Marking and Grammatical roles in Finnish		
Case	**Transitive Subject**	**Direct Object**
Nominative	47,521	0
Accusative	0	93,520
Nominative/ Accusative (case syncretism)	42,884	246,361

Table 4: Frequencies of Nominative, Accusative and case syncretism forms in Russian

German was a special case, where all feminine, neuter and plural forms were treated as ambiguous, since their Nominative and Accusative forms are formally indistinguishable, whereas masculine nouns were analyzed as Nominative or Accusative only in the presence of determiners or adjectives, which normally carry the distinct marking in combination with the noun. See more details about the procedure in Levshina (2021).

Based on numbers like the ones displayed in Tables 3 and 4, we computed the Mutual Information between cases and grammatical roles for each language. The higher the Mutual Information, the more strongly the case forms are associated with the grammatical roles in question. Languages with zero scores had no case marking on Subject and Object (Danish, Dutch, English, Indonesian, Swedish and Vietnamese). Languages with the highest scores were those with rich morphological case marking (Lithuanian, Hungarian, Latvian, Estonian and Japanese). Languages with some type of differential, lexically restricted or optional marking were in-between (the Slavic languages, Hindi, Korean, German, Persian and Turkish). The variable with these scores was called "Case Marking".

Finally, we took the Mutual Information between nouns and the grammatical role of Subject and Object as a proxy for semantic tightness. If the proportions with which a noun is found as a transitive Subject and Object are similar to the baseline proportions of Subject and Object, this

https://doi.org/....

contributes to the semantic looseness of a language. If a noun is strongly biased towards one of these roles, this increases its semantic tightness (see Levshina 2021 for more details). The higher the Mutual Information, the tighter the language. The languages with the highest scores were Hindi, Korean, Russian, Hungarian and Japanese. They are known as tight-fit languages in the literature with regard to the relationships between arguments and their semantics. Indonesian had the lowest score, followed by English and Spanish. These were the loosest languages in our sample. This tightness measure was labelled as "MI Nouns".

The correlation analyses were based on Spearman's rank-based correlations. In order to control for the genealogical dependencies in our data (i.e., the fact that many languages come from one and the same genus), we used a sampling procedure, where we created 1,000 samples. For every sample, we drew randomly only one language per genus and computed the correlation coefficient (ρ) and the p-value based on 100 permutations. After we had these data for all samples, we averaged the coefficients and the *p*-values.

Figure 3 represents the correlation coefficients between the parameters. It shows that both types of P-lability scores (Verbs and Verb + Noun) are correlated with the other typological parameters. The correlation between both types of MI scores related to P-lability and case marking is strong and positive. This means that languages with systematic case marking have high MI scores and therefore low P-lability. There is also a positive correlation between MI related to P-lability and MI based on nouns only. At the same time, P-lability scores are negatively correlated with verb-medialness and word order rigidity (low MI between the verb and the grammatical role(s) of Its Noun argument(s) correlates with SVO and rigid SO). This means, in turn, that languages with SVO and rigid SO order have more P-lability. All these scores are statistically significant at the conventional level ($p < 0.05$). Judging from the magnitude of the coefficients, we can also see that the P-lability scores based on verbs only are overall less strongly correlated with the other typological parameters than the P-lability scores based on Verb + Noun combinations.

We also observe significant negative correlations between rigid SO order and case marking, and between verb-medialness and case marking. In addition, there is a strong and significant negative correlation between Mutual Information based on associations between nouns and grammatical roles, and verb-medialness.

All other correlations are not statistically significant. This means that we do not find evidence that A-lability is correlated with any of these typological parameters.

https://doi.org/....

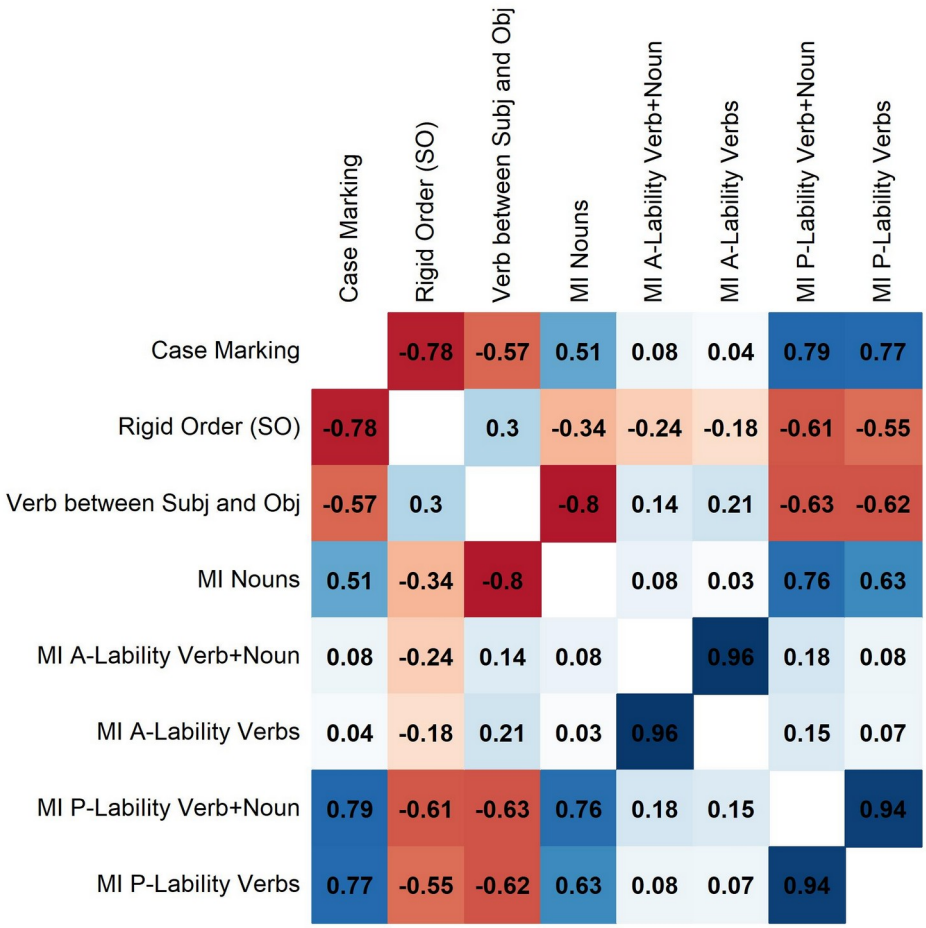

Figure 3: Correlations between the typological parameters and lability scores. The colour intensity represents the strength of the correlation. Blue cells stand for positive correlations. Red cells display negative correlations.

5. Conclusions

Our quantitative analyses reveal that P-lability scores are systematically correlated with the other parameters related to tight and loose fit. Languages with low P-lability (and high MI scores) tend to have case marking, stronger associations between nouns and grammatical roles, relatively flexible order of Subject and Object, and verb-final order. These features are associated with tight-fit languages. In contrast, languages with high P-lability (and low MI scores) tend to have little or no case marking, quite rigid SVO order, and weaker associations between nouns and grammatical roles. These features are associated with loose-fit languages. Therefore, our data support Hawkins' (1986, 1995) prediction that verbs in loose-fit languages are used in more diverse subcategorization frames.

More specifically, the P-lability scores based on verbs only are overall less strongly correlated with the other typological parameters than the P-lability scores based on Verb + Noun combinations. This is not surprising, because these latter scores also include the attraction of nouns to different grammatical roles. At the same time, both of these scores are

https://doi.org/....

more strongly correlated with word order rigidity and case marking than semantic tightness scores based on nouns only, but the latter has a stronger correlation with verb-final order (since the languages in the sample, except for Arabic, are either verb-medial, or verb-final). This may have to do with the fact that the attraction of nouns to one or the other role helps to avoid costly reanalysis when the verb comes last. Whether or not the verb has special marking depending on the roles of its arguments is less important for that purpose.

It is remarkable that our P-lability scores are more strongly correlated with the other typological parameters than the latter are among themselves. This is an unexpected finding, but it can be explained by the fact that P-lability scores convey information not only about the verbs, but also (explicitly or implicitly) about the nouns in different roles. These scores can thus be a useful indicator of the word-external or word-internal orientation of the language in question (Hawkins 2019).

In contrast, the A-lability scores are not correlated with any of those properties. A-lability is also found more frequently in our corpora than P-lability, as we see from the lower MI values in the former. A possible explanation for this is that A-lability is often driven by general pragmatic factors. For example, the object can be omitted due to its high accessibility, e.g., *And Italy wins [the final]!* Many objects are omitted due to specific conventionalized inferences, *e.g. He drinks again [liquor].* Object omission is also possible if the focus is on the action, while the object has low discourse prominence, e.g. *She chopped and chopped [e.g., meat]* (Goldberg 2005). Other reasons are cultural. For example, the object can be omitted when it is taboo, e.g., *Pat sneezed [mucus] onto the computer screen*, or for feelings of tact, *I contributed [$1,000] to UNICEF* (Goldberg 2005). In addition, many rules allowing for object omission are also lexically and semantically specific (Fillmore 1986). All these pragmatic factors and lexical idiosyncrasies explain the lack of systematic correlations between A-lability and the other typological properties of the languages in question.

The findings of this study should be tested on a larger and more diverse sample of languages and genres. A further question is whether there are causal relationships between these parameters, and what they look like. A causal analysis in Levshina (2021) showed that case marking is more likely to be affected by other typological parameters (word order and associations between lexemes and syntactic roles) than the other way round. We need a larger sample of languages in order to answer this question and test all possible causal links. It would also be interesting to add other parameters, such as the frequency of long-distance syntactic dependencies or categorial ambiguity, and test their relationships with the ones examined here.

Acknowledgements
To be added.

References

Aissen, Judith. 2003. Differential object marking: Iconicity vs. economy. *Natural Language and Linguistic Theory* 21. 435-483.

Dahl, Östen. 2000. Egophoricity in discourse and syntax. *Functions of Language* 7(1). 37-77.

Dixon, R.M.W. 1994. *Ergativity*. Cambridge: Cambridge University Press.

Fillmore, Charles J. 1986. Pragmatically Controlled Zero Anaphora. *Proceedings of the Berkeley Linguistics Society* 12. 95-107.

Gibson, Edward, Steven T. Piantadosi, Kimberly Brink, Leon Bergen, Eunice Lim and Rebecca Saxe. 2013. A Noisy-Channel Account of Crosslinguistic Word-Order Variation. *Psychological Science* 24(7). 1079–1088. DOI https://doi.org/10.1177/0956797612463705

Goldberg, Adele E. 2005. *Argument realization: the role of constructions, lexical semantics and discourse factors*. In Jan-Ola Östman & Miriam Fried (eds.), *Construction Grammars: Cognitive grounding and theoretical extensions*, 17 44. Amsterdam: John Benjamins.

Goldhahn, Dirk, Thomas Eckart & Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck et al. (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 759-765. Istanbul: ELRA. URL http://www.lrec conf.org/proceedings/lrec2012/pdf/327_Paper.pdf.

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.

Hawkins, John A. 1986. *A Comparative Typology of English and German: Unifying the Contrasts*. London: Croom Helm.

Hawkins, John A. 1995. Argument-predicate structure in grammar and performance: A comparison of English and German. In Irmengard Rauch & Gerald F. Carr (eds.), *Insights in Germanic Linguistics, Vol. 1 Methodology in Transition*, 127-44. Berlin: Mouton de Gruyter.

Hawkins, John A. 2019. Word external properties in a typology of Modern English: A comparison with German. *English Language and Linguistics* 23(3). 701-723.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

Levshina, Natalia. Token-based typology and word order entropy. *Linguistic Typology* 23(3). 533-572. DOI https://doi.org/10.1515/lingty-2019-0025

Levshina, Natalia. 2020. How tight is your language? A semantic typology based on Mutual Information. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, 70-78. Düsseldorf: ACL. URL https://www.aclweb.org/anthology/2020.tlt 1.7.pdf

Levshina, Natalia. 2021. Cross linguistic trade-offs and causal relationships between cues to grammatical subject and object, and the

problem of efficiency-related explanations. *Frontiers in Psychology* 12. 648200. DOI 10.3389/fpsyg.2021.648200.
Müller-Gotama, Franz. 1994. *Grammatical Relations: A Cross-Linguistic Perspective on Their Syntax and Semantics*. Berlin: Mouton de Gruyter.
Plank, Frans. 1984. Verbs and objects in semantic agreement: Minor differences between English and German might that might suggest a major one. *Journal of Semantics* 3(4). 305-360.
Shannon, Claude E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27. 379-423, 623-656.
Wijffels, Jan, Milon Straka & Jana Straková. 2018. udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the UDPipe NLP Toolkit. R package version 0.7. URL https://CRAN.R project.org/package=udpipe.

Corpora

Leipzig Corpora Collection
https://wortschatz.uni-leipzig.de/en/download

CONTACT

Email address of corresponding author
(Email address of a co-author)