How tight is your language?

A semantic typology based on Mutual Information

Natalia Levshina
MPI for Psycholinguistics, Nijmegen
Neurobiology of Language Department
natalia.levshina@mpi.nl

Abstract

Languages differ in the degree of semantic flexibility of their syntactic roles. For example, English and Indonesian are considered more flexible with regard to the semantics of subjects, whereas German and Japanese are less flexible. In Hawkins' classification, more flexible languages are said to have a loose fit, and less flexible ones are those that have a tight fit. This classification has been based on manual inspection of example sentences. The present paper proposes a new, quantitative approach to deriving the measures of looseness and tightness from corpora. We use corpora of online news from the Leipzig Corpora Collection in thirty typologically and genealogically diverse languages and parse them syntactically with the help of the Universal Dependencies annotation software. Next, we compute Mutual Information scores for each language using the matrices of lexical lemmas and four syntactic dependencies (intransitive subjects, transitive subject, objects and obliques). The new approach allows us not only to reproduce the results of previous investigations, but also to extend the typology to new languages. We also demonstrate that verb-final languages tend to have a tighter relationship between lexemes and syntactic roles, which helps language users to recognize thematic roles early during comprehension.

1 Theoretical background and aims of the paper

This paper proposes a quantitative bottom-up corpus-based approach to cross-linguistic comparison, determining how tightly or loosely different lexemes can be mapped on basic syntactic roles. The idea goes back to Hawkins (1986: 121– 127, 1995; see also Müller-Gotama 1994), who coined the terms 'tight-fit' and 'loose-fit' languages. The former have unique surface forms that map onto more constrained meanings, whereas the latter have more vague forms with less constrained meanings. For instance, Present-Day English has fewer semantic restrictions on the subject and object than Old English, German or Russian. Consider several examples below.

(1) a. Locative: *This tent sleeps four.*
b. Temporal: *2020 witnessed a spread of the highly infectious coronavirus disease.*
c. Instrument: *10 Euros will buy you a meal.*
d. Source: *The roof leaks water.*

While these sentences are perfectly acceptable in English, their German or Russian equivalents would be unacceptable or strange. This means that subjects in English are less semantically restricted than subjects in German and Russian (see also Plank 1984).

Tightness and looseness have several components. Semantic flexibility of arguments is only one of them. Other features of tight languages include formal case marking, avoidance of raisings and long WH-movements and lower reliance on context in interpretation.

Languages can change their degree of tightness. English is a well-known example of shifting from tight to loose (Hawkins 1986). As the case was lost, the zero-marked NPs in Middle English became more dependent on the verb for theta-role assignment. This is why the rigid SVO order emerged, which helps language users to understand correctly who did what to whom. Also, new instrumental and locative subjects as in (1) became possible, which used to be the case only in prepositional phrases.

In the previous work, the judgements about tightness and looseness were made introspectively and qualitatively. This paper presents a method that allows one to quantify these differences objectively with the help of corpus data. We only focus on the fit between syntactic roles and semantics of lexemes in this study. As a proxy for semantics, we extract frequencies of lexemes in different syntactic roles from syntactically parsed corpora (see Section 2). Next, we compute how much these frequencies diverge from the total frequencies of the roles with the help of the Mutual Information metric. The higher a score, the tighter the language (see Section 3). The scores are then compared with the existing classification of languages. We find a close correspondence between the scores (see Section 4). The scores are computed for lemmas alone and for lemmas plus multiword units. We also investigate the correlation between tightness and the proportion of verb-final frames in a corpus (Section 5). Section 6 provides the conclusions and an outlook.

2 Data

In order to extract the distributional information, one needs large corpora. Available cross-linguistic syntactically annotated collections, such as the Universal Dependencies corpora (Zeman et al. 2020), are too small for the purposes of the present study. The solution was to use the Leipzig Corpora Collection (Goldhahn et al. 2012), which contains freely downloadable web-based corpora of reasonable size in more than 200 languages. The language sample used for the present study includes thirty languages, which are listed in Table 1. Each language is represented by one million sentences from online news (categories ‘news’ and ‘newscrawl’). The corpora contain sentences in random order. The choice of languages was determined by the availability of sufficient data and a reasonably good language model in the UDPipe annotation tools.

The sentences were tokenized, lemmatized and morphologically and syntactically annotated with the help of the UD corpus tools (Straka & Straková 2017) in the R package *udpipe* (Wijffels et al. 2019). The language models, which were trained on the UD corpora (Zeman et al. 2020), provide, among other things, universal parts-of-speech tags and dependency relations, which can be compared across different languages. This is crucial for the purposes of the present study.

One should be aware of risks involved in using automatic parsers for cross-linguistic data analysis. Manual evaluation of the annotation was impossible, given the size and diversity of the data. However, ongoing research (Levshina, Submitted) indicates very strong correlations between diverse morphological and word-order parameters based on the same annotated corpora and on the training corpora from the Universal Dependencies collection, as far as the core arguments are concerned. This gives us some confidence in the results.

The following universal dependencies, which represent syntactic arguments, were extracted from the annotated corpora:

- *nsubj* (lexical, or non-clausal subject), e.g. *<u>The student</u> is reading*. Subjects in transitive and intransitive clauses were treated separately. A head verb was considered transitive if it had an overt object.

- *obj* (object), e.g. *I see <u>the student</u>.*

- *obl* (oblique, i.e. any non-core nominal argument or adjunct), e.g. *I’m talking <u>with a student;</u> She’s reading <u>in the library</u>.*

The UD approach does not distinguish between oblique arguments and adjuncts. In addition, many languages do not have indirect object (*iobj*) as a separate dependency. This is why indirect objects, which were not very numerous, were counted as a joined category of indirect objects + obliques for the sake of cross-linguistic comparability. The more detailed tags in the dependencies, such as *nsubj:pass* (subject of a passive clause) were treated as simply *nsubj*, *obj* or *obl*. The reason is that such extended tags are language-specific and not used in a unified way across the languages.

Language	Genus	Family	UD model	Lemmas
Arabic	Semitic	Afro-Asiatic	arabic-padt-ud-2.4	16,799
Bulgarian	Slavic	Indo-European	bulgarian-btb-ud-2.4	11,924
Croatian	Slavic	Indo-European	croatian-set-ud-2.4	13,791
Czech	Slavic	Indo-European	czech-pdt-ud-2.4	11,783
Danish	Germanic	Indo-European	danish-ddt-ud-2.4	16,340
Dutch	Germanic	Indo-European	dutch-alpino-ud-2.4	13,334
English	Germanic	Indo-European	english-ewt-ud-2.4	10,480
Estonian	Finnic	Uralic	estonian-edt-ud-2.4	20,231
Finnish	Finnic	Uralic	finnish-tdt-ud-2.4	20,822
French	Romance	Indo-European	french-gsd-ud-2.4	9,386
German	Germanic	Indo-European	german-gsd-ud-2.4	16,729
Greek (modern)	Greek	Indo-European	greek-gdt-ud-2.4	13,789
Hindi	Indic	Indo-European	hindi-hdtb-ud-2.4	10,546
Hungarian	Ugric	Uralic	hungarian-szeged-ud-2.4	13,931
Indonesian	Malayo-Sumbawan	Austronesian	indonesian-gsd-ud-2.4	9,820
Italian	Romance	Indo-European	italian-isdt-ud-2.4	10,643
Japanese	Japanese	Japanese	japanese-gsd-ud-2.4	19,198
Korean	Korean	Korean	korean-gsd-ud-2.4	29,017
Latvian	Baltic	Indo-European	latvian-lvtb-ud-2.4	12,062
Lithuanian	Baltic	Indo-European	lithuanian-hse-ud-2.4	17,652
Persian	Iranian	Indo-European	persian-seraji-ud-2.4	11,440
Portuguese	Romance	Indo-European	portuguese-bosque-ud-2.4	9.663
Romanian	Romance	Indo-European	romanian-rrt-ud-2.4	12,962
Russian	Slavic	Indo-European	russian-syntagrus-ud-2.4	10,092
Slovenian	Slavic	Indo-European	slovenian-ssj-ud-2.4	13,094
Spanish	Romance	Indo-European	spanish-gsd-ud-2.4	10,317
Swedish	Germanic	Indo-European	swedish-talbanken-ud-2.4	16,096
Tamil	Southern Dravidian	Dravidian	tamil-ttb-ud-2.4	14,737
Turkish	Turkic	Altaic	turkish-imst-ud-2.4	12,554
Vietnamese	Viet-Muong	Austro-Asiatic	vietnamese-vtb-ud-2.4	16,552

Table 1: Languages and UD language models used in the present study.

Next, the lexemes (lemmas) performing these syntactic roles were extracted. The analyses presented below are based only on common nouns, following the tradition of word order research in typology, but the scores for a wider range of lexemes were computed, as well, including proper nouns, verbs, adjectives, symbols and numerals. Pronouns were excluded because of the lack of anaphora resolution in the corpora and the fact that the languages have vastly different pronominal systems with different pro-drop rates. The correlations between the Mutual Information scores based on these lexemes and the ones based on common nouns only are very strong and positive: $r = 0.914$, $p < 0.0001$ for lemmas only and r = 0.944, p < 0.0001 for lemmas and MWE.

If there was coordination (e.g. *Students and teachers came to the party*), the subsequent coordinated elements marked with the dependency 'conj' (i.e. *teachers* in the example) were treated as having the same dependency as the first coordinate member (i.e. *students*). The cleaning procedure involved removing punctuation marks in the beginning and at the end of the strings and normalizing the case. The lemmas with the frequency of 10 and less were left out because they were often analyzed erroneously.

An important issue in language comparison is what to count as a word (Haspelmath 2011). For example, in English, the phrase *art history* consists of two words, but its German equivalent *Kunstgeschichte* is only one word. In order to counterbalance the influence of orthographic conventions, we also computed the scores treating multiword units like *art history* as one lexeme. In order to identify multiword expressions (MWE), we used the following dependencies in the UD annotation: *compound*, *fixed* and *flat*. The dependency *compound* is used to identify parts of compounds, e.g. *art history* or *frying* *pan*. The dependency *fixed* helps to identify grammaticalized MWE, e.g. *in spite of*. Finally, the UD annotation has the dependency *flat*, which helps to identify complex proper names, such as *Angela Merkel*.[1]

3 Information-theoretic measures of semantic fit

For every lexeme, its actual and relative frequencies were computed in each of the four main syntactic roles: subject of an intransitive clause, subject of a transitive clause, object and oblique. Some examples are displayed in Table 2.

Lexeme	**Intransitive subject**	**Transitive subject**	**Object**	**Oblique**
hunter/NOUN	64	40	22	30
evening/NOUN	100	38	150	1145
street/NOUN	155	34	466	1331
t-shirt/NOUN	7	3	118	36

Table 2: A fragment of the lexeme – dependency matrix for English.

On the basis of these matrices, the Mutual Information (MI) scores were computed for each language. This metric represents the degree by which the relative frequencies of the syntactic roles performed by individual lexemes differ from the relative frequencies of these roles in the corpus. The formula for computing the measures based on a matrix of probabilities is given below.

$$I\,(Lex;\ Dep) = \sum\nolimits_{i,j} p\left(lex_i, dep_j\right) log\ \frac{p\left(lex_i, dep_j\right)}{p\left(lex_i\right) p\ \left(dep_j\right)}$$

where *Lex* stands for lexemes (lemmas) and *Dep* represents the four selected syntactic dependencies.

The greater this divergence, the more biased the lexemes on average towards a particular role, and therefore the tighter the fit between the lexemes and the syntactic dependencies. For instance, human nouns tend to be biased towards the role of intransitive and transitive subjects (e.g. *hunter*), inanimate objects frequently occur in the object role (e.g. *t-shirt*), whereas temporal and locative nouns (e.g. *evening, street*) are frequent in the oblique role.

4 Estimation of tight and loose fit

Figure 1 displays the MI scores in the thirty languages, based on lemmas only and on lemmas plus MWE. The correlation between these scores is high: $r = 0.929$ ($p < 0.001$). For English, Hindi, Indonesian, Japanese, Korean and Vietnamese, the scores based on lemmas plus MWE are higher than the scores based on lemmas only.

The English corpus has the lowest divergence. This means that on average the lexemes in that corpus are 'promiscuous' with regard to the roles. The other Germanic languages, from Swedish and German to Dutch and Danish have higher scores. The Romance languages are loose; they have relatively low scores, with Spanish being the loosest and Portuguese the tightest. Modern Greek and Bulgarian (the most analytic Slavic language) are loose, as well. The other Slavic languages have moderate scores, with Slovene being the tightest. The two Baltic languages (Latvian and Lithuanian) are on the loose-to-

[1] More information on multiword expressions in the UD can be found here: https://universaldependencies.org/u/overview/specific-syntax.html#multiword-expressions.

moderate side of the distribution. The three Uralic languages (Finnish, Estonian and Hungarian) have high scores, especially Finnish, which is among the tightest languages, together with Hindi and Korean. Hungarian is the loosest language of the Uralic languages.

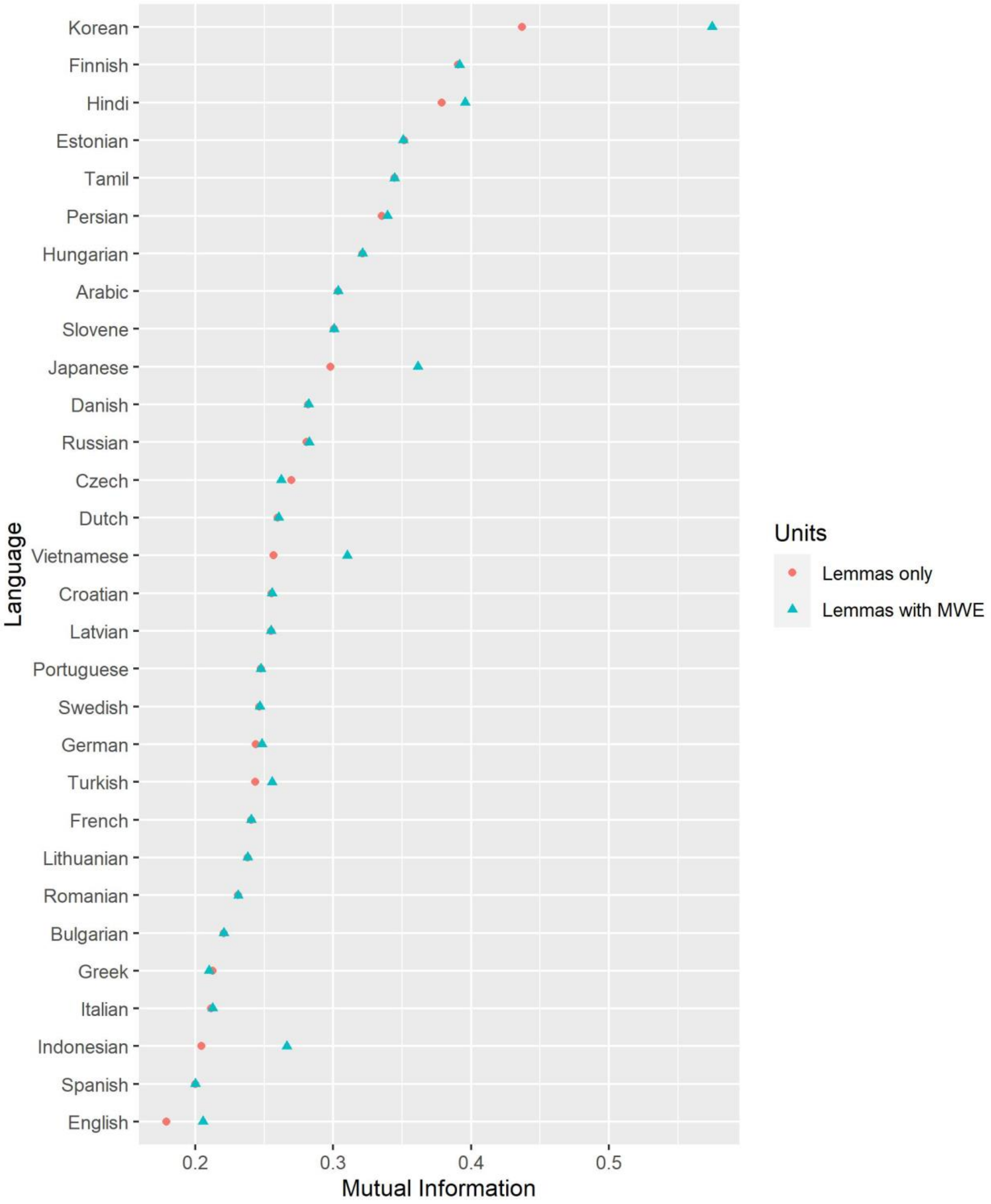

Figure 1: Mutual Information of lexemes and syntactic dependencies in 30 languages.

Overall, the previous observations about loose and tight languages are met. Among the languages represented in our sample, English has been evaluated in the literature as the most flexible, followed by Indonesian, and further by German, Japanese, Korean, Russian and Turkish (Hawkins (1986: 121– 127, 1995) and Müller-Gotama (1994). However, Indonesian is slightly tighter than German or Turkish, contrary to the previous reports (see Section 4), if we take into account MWE. This is also what we see in the data at the levels of lemmas. We also see that there is large variability within the languages that were considered tight, with German and Turkish having moderate scores and Korean having a very large score.

5 Correlation between tightness scores and word order

An important question is, how can we explain the cross-linguistic differences in tightness and looseness? There are substantial differences even among genetically related languages, so this factor does not seem to play an important role. A possible explanation may be related to processing constraints. If a language has the SVO order, the verb is accessed early. As a result, the addressee can use the semantic information in the verb to identify the roles of the other constituents in the clause (in particular, the thematic roles, such as Agent, Patient or Instrument). There is some experimental support of this claim. In particular, when asked to describe events in pantomime, people tend to avoid SOV in favour of SVO if the transitive event is reversible, that is, if each participant can be subject or object, e.g. “The mother hugs the boy” or “The boy hugs the mother” (Hall et al. 2014).

If a verb occurs in the end of the sentence, as in SOV languages, the thematic roles of nouns are more difficult to assign early. In order to mitigate the risk of incorrect interpretation of the frame and to avoid the costs of reanalysis, verb-final languages rely on semantic tightness of the arguments, as well as on case marking and other features of tight-fit languages (see Section 1). This is why, according to Hawkins (1995), the languages with verb-final structures (e.g. Japanese or German) exhibit greater predicate frame differentiation than languages like English.

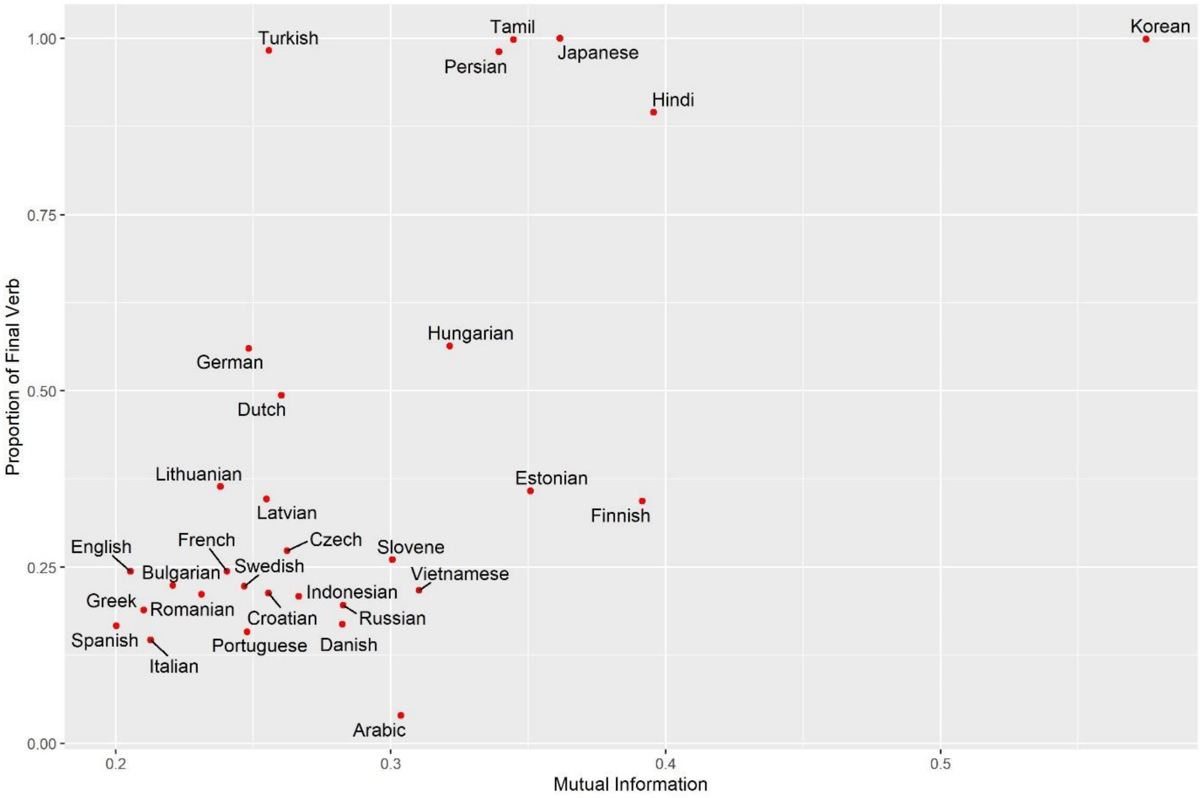

Figure 2: Mutual Information and proportion of verb-final sentences

This explanation, however, has not been systematically tested. In order to fill in this gap, we computed what we call a 'verb-finalness' score for each language. The procedure was as follows. We looked for all verbs with following dependencies: subject, object, oblique (with the exception of adverbs) and/or indirect object, where available. Each verb with at least one dependency from the list was counted as one frame. If a verb was used after all these dependent elements, then the frame was considered verb-final. The verb-finalness score was computed for each language by dividing the number of verb-final frames by the total number of frames. Arguments of nominal predicates were not taken into account.

Figure 2 displays the MI scores based on words and multiword expressions against the verb-finalness scores. The plot suggests that the correlation is positive. That is, the tighter a language, the more frequently the verb is final and therefore the more difficult it is to infer thematic roles from the start.

A Bayesian mixed-effect model with genera (see Table 1) as random intercepts, verb-finalness as the response variable and MI as the fixed effect shows that the effect of verb-finalness is positive, with the estimate *b* = 1.63 and the 95% credible interval between 0.06 and 3.19. This confirms our expectations. The Bayesian R^2 is 0.85, with the 95% credible interval between 0.66 and 0.93, which suggests a strong relationship between semantic tightness and verb-finalness.

If we take the divergence scores based on lemmas only, the effect of verb-finalness is slightly weaker (the estimate *b* = 1.58, with the 95% credible interval between -0.10 and 3.57). The credible interval is this time wider and includes zero, so we can be less confident in this result. The Bayesian R^2 is 0.85 again, with the 95% credible interval between 0.63 and 0.93.

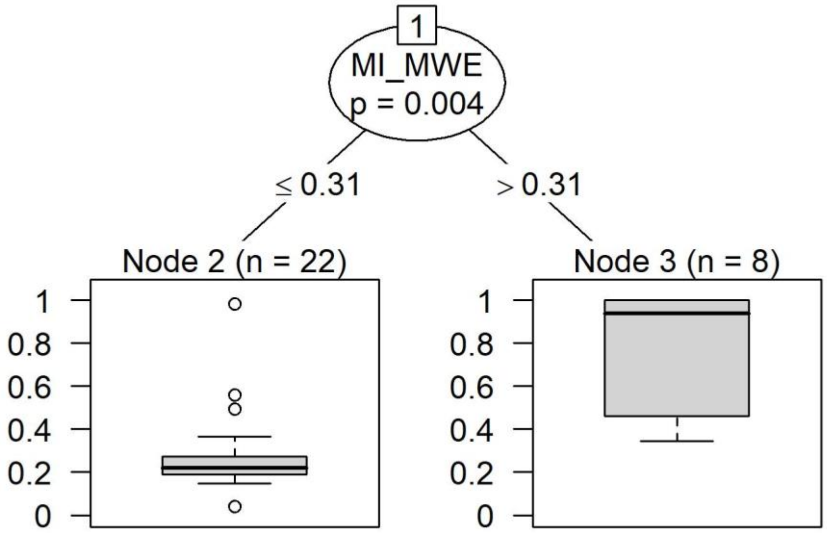

Figure 3: A conditional inference tree predicting verb-finalness

Since the data indicate a heteroscedastic relationship, with more variation in the MI scores as the verb-finalness scores increase, we also used a non-parametric method of conditional inference trees in order to make sure that our conclusions are valid. The method tests the null hypothesis of conditional independence of the response variable given a predictor. Conditional inference trees involve recursive binary partitioning of the data (Hothorn et al. 2006). The algorithm tries to identify the predictor that has the strongest association with the response variable and makes a binary split in that variable. After that, the procedure is repeated for each subset of the data until no further split can be made. In order to make a split, a set of criteria should be met, such as the level of significance at 0.05. Using this method, we can predict the verb-finalness scores (the response variable) from the two types of MI scores – based on lemmas and lemmas plus MWE. The genus was also tested.

Figure 3 displays the conditional inference tree model predicting verb-finalness. It shows that the MWE-based MI scores allow us to predict verb-finalness, and the other variables are not important. If MI is less than or equal to 0.31, then the word order is not likely to be verb-final, as shown by the box plot in Node 2. If MI is higher than 0.31, then we are likely to have a verb-final language. The genealogical factors do not play a significant role because they do not participate in any splits. Adding the family as a predictor does not change the results, either.

Therefore, there is a strong association between verb-finalness and MI. Also, taking into account composite nouns and other MWE leads to a stronger association between word order and semantic tightness.

6 Conclusions

In this paper we have demonstrated how one can use information about attraction between lexemes and syntactic roles (dependencies) measured with the help of Mutual Information for the purposes of language comparison. We have reproduced most of previous observations about languages with tight and loose fit between lexemes and arguments, and computed scores for many new languages. One should also be aware that the ranking changes somewhat depending on whether one takes single lemmas or also takes into account multiword expressions, which usually make the MI scores higher, and the language tighter. This is not surprising because composite nouns can be more semantically specific (e.g. *computer mouse* vs. *field mouse*) than simple lemmas (e.g. *mouse*) and therefore their syntactic behaviour can be more restricted.

The regression analysis also indicates that semantic tightness is associated with the final position of the verb. This relationship is more credible if the divergence scores take into account multiword expressions.

In the future, the results of this study should be tested on new data representing other registers and text types. One can expect substantial intra-linguistic variation. In addition, it would be interesting to investigate correlational and causal relations between tightness and other cues for understanding who did what to whom. One of the most important cues is case marking. As we can see in Figure 1, languages with low tightness scores tend to have fewer nominal cases than languages with high scores, although there are a few exceptions, such as Lithuanian, which has rich case morphology but loose fit between lexemes and dependencies. One should also consider verb agreement, which can help in identification of roles (cf. De Vogelaer 2007). Finally, it would be interesting to test word order entropy (Futrell et al. 2015; Levshina 2019), since rigid word order with low entropy can also be used as a cue for mapping the roles and participants. Other potential factors of interest are extralinguistic. For example, one can imagine tighter semantic relationships in languages with few speakers and closely knit communities, where the semantic restrictions can be easier to maintain and transfer, similar to other high-complexy features, and with few L2 speakers, who can have difficulties acquiring the semantic restrictions.

Acknowledgements

The research in this paper was funded by the Netherlands Organisation for Scientific Research (NWO) under Gravitation grant *Language in Interaction*, grant number 024.001.006. The author is grateful for insightful comments made by anonymous reviewers. All remaining mistakes are her own.

References

Gunther De Vogelaer. 2007. Extending Hawkins' comparative typology: Case, word order, and verb agreement in the Germanic languages. *Nordlyd,* 34 (special issue on Scandinavian Dialect Syntax), 167-182.

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

Dirk Goldhahn, Thomas Eckart, and 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.

Matthew L. Hall, Rachel Mayberry, and Victor S. Ferreira. 2013. Cognitive constraints on constituent order: evidence from elicited pantomime. *Cognition*, 129(1), 1-17. DOI https://doi.org/10.1016/j.cognition.2013.05.004

Martin Haspelmath. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 45(1): 31–80. DOI https://doi.org/10.1515/flin.2011.002.

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

John A. Hawkins. 1995. Argument-predicate structure in grammar and performance: A comparison of English and German. In Irmengard Rauch, and Gerald F. Carr (eds.), *Insights in Germanic Linguistics.* Vol. 1: Methodology in Transition, 127–144. Mouton de Gruyter, Berlin.

Torsten Hothorn, Kurt Hornik and Achim Zeileis. 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3): 651--674.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology,* 23(3): 533–572.

Natalia Levshina. Submitted. Efficient trade-offs as explanations in functional linguistics: some problems and an alternative proposal.

Franz Müller-Gotama. 1994. *Grammatical Relations: A Cross-Linguistic Perspective on Their Syntax and Semantics*. Mouton de Gruyter, Berlin.

Frans Plank. 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.

Milan Straka, and 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*, Vancouver, Canada, August 2017.

Jan Wijffels, Milan Straka, and Jana Straková. 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.

Daniel Zeman, Joakim Nivre, Mitchell Abrams et al. 2020. Universal Dependencies 2.6, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3226. See also http://universaldependencies.org