Explaining universal semantic properties in the adjectival/adverbial domain

Irina Nemtcova
Center for Mind/Brain Sciences,
University of Trento
nemtzova.irina@gmail.com

Jakub Szymanik
Center for Mind/Brain Sciences,
Dep. of Information Engineering and Computer Science
University of Trento
jakub.szymanik@gmail.com

Abstract

Native speakers would generally agree that certain patterns in the language are preferable to others. The order of several adjectives modifying the same noun is the best example: various potential sequences of adjectives are all equivalent in meaning, yet speakers often prefer specific orders. Recently, with the availability of large quantities of data and significant development of processing tools and statistical methods, researchers have started to study word order preferences using quantitative approaches based on "efficiency-based" communication. In our work, we test three of these theories: pointwise mutual information (PMI) (Futrell, Qian, et al. 2019), integration complexity (IC) (Dyer 2017), and information gain (IG) (Futrell, Dyer, and Scontras 2020). We start by replicating the results obtained by (Futrell, Dyer, and Scontras 2020) on adjective order preferences in English. To verify the cross-linguistic consistency of the results, we analyze adjective order tendencies in Russian. Since the theories in the study provide insights into base word order, not limited to adjectives only, we continue by investigating in detail if the predictions made by these theories can generalize and be applied to explain the order of adverbs.

1 Introduction

Native speakers of a language often show a preference for the position of words in specific constructions. The order of several adjectives modifying the same noun is the best example: speakers consistently show preferences for a particular order of adjectives, and interestingly, these preferences are observed cross-linguistically. For instance, in a situation where English speakers would have to describe a chair that is both *small* and *red*, they would show a preference for *a small red chair* compared to *a red small chair* (Scontras, Degen, and Goodman 2017). The very same preferences are found not only in English but also in other languages where adjectives occupy the prenominal position, such as Greek, Polish, or Russian. The tendency for two or more consecutive adjectives to follow some order is also observed in postnominal languages, where adjectives come after the noun they modify. Various potential sequences of adjectives are all equivalent in meaning, yet specific orders are preferred to others.

Cross-linguistic similarities have also been found in the order of adverbs (Cinque 1999). Their distribution is challenging to study due to the high variation in adverbs' positioning. As long as there are no syntactic structure constraints, adverbs can be placed in various positions in the sentence, often without a change in meaning.

Many theories, primarily relying on descriptive approaches, have been proposed in an attempt to find a measure that governs the word order tendencies in the adjectival and adverbial domains (Danks and Glucksberg 1971; Scontras, Degen, and Goodman 2017; Cinque 1999.) However, these approaches often lack generalization and frequently describe correlation rather than actual cause.

Recent approaches that try to make more generic predictions that would hold cross-linguistically use the argument of "functionality." In other words, the communication process

is efficient-based and constrained by the brain: it tries to maximize information transfer while minimizing cognitive cost (Hawkins 2004). Efficiency can be achieved with simplicity by minimizing cognitive effort for both sides involved in communication: the sender and the receiver. It can, for instance, be quantified using the length of the transmitted message (Gibson et al. 2019). Thus, the reason we find patterns that are repeated across languages can be due to these constraints of cognition on communication.

Efficiency-based approaches have already been studied separately before on cross-linguistic data. However, except for a few studies, the results of their predictions have been analyzed in isolation and have been missing systematic comparisons. Following Futrell, Dyer, and Scontras 2020, our research aims to address this question and test the predictions from four efficiency-based theories: integration complexity (Dyer 2017), information gain, pointwise mutual information (Futrell, Dyer, and Scontras 2020), and subjectivity (Scontras, Degen, and Goodman 2017).

1.1 Predictors

Each theory in our analysis measures cognitive effort by examining a sentence's relationship between a headword and its dependents.

Integration complexity measurement is based on the co-occurrence of the dependent word and the syntactic category of its head. The theory predicts that a dependent with a narrow, predictable range of possible heads should be placed closest to the head since we are more likely to predict its head's syntactic category. On the other hand, a dependent with no clear tendency regarding the syntactic category of the head will be placed further. The range of possible heads of the depended is measured with Shannon's entropy using the probabilities of each possible outcome (x_i), as shown in Equation 1:

$$H(X) = -\sum_{i=1}^{n} P(x_i) * log_2 P(x_i). \tag{1}$$

PMI states that an efficient language will minimize the linear distance between elements with high pointwise mutual information (PMI) - an information-theoretic measure of how strongly two words predict each other. Thus, the dependent (x) that has high PMI with its head (y) will tend to be closer to the head:

$$PMI(x, y) = log \frac{p(x, y)}{p(x)p(y)}. \tag{2}$$

Information gain looks at a phrase as a decision tree for identifying a referent where each partitioning of a phrase is associated with some information gain. The proposal is that a word with smaller information gain will be placed earlier so that each word gradually narrows the set of referents. The information gain of a word is measured as a difference between the starting entropy of a headword $H[N]$ and the sum of the entropy of each partition of a dependent N_x (a subset of heads that take a given word as a dependent) and a negative evidence N_x^c (which in turn, do not have a given word as a dependent), as shown in the Equation 3:

$$IG(x) = H[N] - \left[\frac{N_x}{N}\ H[N_x] + \frac{N_x^c}{N}\ H[N_x^c]\right] \tag{3}$$

And lastly, **subjectivity** predicts that less subjective adjectives occur linearly closer to the nouns they modify (Scontras, Degen, and Goodman 2017). An empirical study of the English language supported the claim. Participants were shown a series of adjectives and had to indicate whether an adjective was objective or subjective using a sliding scale. After comparing subjectivity ratings with the sequences of several consecutive adjectives, it turned out that less subjective adjectives are placed closer to the noun they modify. The reason might be that the placement of adjectives from most to least subjective facilitates successful reference resolution.

2 Proposal

We propose to compare the accuracies of these theories and to find the predictor that would be the best in identifying the word order in phrases with two stacked adjectives/adverbs in pre-head position in typologically diverse languages: English and Russian.

First, we will calculate predictors based on the scoring function S(H, D) applied to a dependent (D) and a head (H). As we saw earlier, the estimated predictors come with theories describing their effect on word order. Thus, the difference in scoring functions between $S(D_1, H)$ and $S(D_2, H)$ can be enough to predict the order: dependents that have low PMI or low IG are placed father from the head; whereas dependents with low IC or subjectivity are placed closer to the head.

However, we can go further and fit a logistic regression classifier to predict a word order in a sequence D_1D_2H, given the difference between $\mathrm{S}(D_1, H)$ and $\mathrm{S}(D_2, H)$. The goal of the classifier is to predict the order from the unordered set of adjectives (A_1, A_2) and a noun; and adverbs (Adv_1, Adv_2) and a verb. Based on empirical results from previous studies, we expect that the fitted values will be negative for PMI and IG, and positive for IC and subjectivity. We can compare predictors in terms of accuracy by analyzing the classifier's results.

3 Data

We estimated predictors using an automatically annotated subsection of the Wikipedia corpus. First, the corpus was cleaned, and only sentences with alphanumeric characters and punctuation, with no less than 5 and no more than 35 items per sentence, were left. The limit on the sentence length was needed because of the particularities of the Wikipedia corpus to filter sentences that contained, for example, long enumeration lists, which did not have any meaningful information useful for the study. After cleaning the data, we got 22,534,266 tokens in English and 11,429,524 in Russian. The extracted text was then parsed with dependency annotations using Stanza parser (Qi et al. 2020).

From the corpus, we extracted:

– Russian, adjectives: For the Russian language, we worked with lemmas. First, we extracted adjective-noun (AN) pairs, a set of pairs $< A, N >$ where A is an adjective dependent on the noun N with dependency type *amod*. We extracted 895,485 AN pairs, out of which 345,675 unique pairs. AN pairs were used to calculate the values of corpus-based predictors and for clustering.

  Second, we extracted AAN triples, a set of triples $< A_1, A_2, N >$ where A_1 and A_2 are adjectives with relation type *amod*, dependent on a single noun head *N*. Additionally, A_1 and A_2 have no other dependents except *N*, must appear in order A_1A_2N with no intervening words in between. We extracted 56,801 unique triples, which we used to fit logistic regression from predictors to word order and evaluate the model.

– English, adjectives: We extracted 990,453 AN pairs in total, 430,698 unique ones, and 41,617 unique AAN triples following the above procedure. The analysis of the English language was done on word forms.

– A similar procedure was used for extracting adverb-verb pairs and adverb-adverb-verb triples, where the adverb is dependent on the verb and has a dependency type *advmod*. In English, we got 93,479 unique Adv-Verb pairs (279,509 total) and 3,320 Adv-Adv-Verb triples. In Russian, we got 71,073 unique Adv-Verb pairs (159,596 in total) and 1,172 Adv-Adv-Verb triples.

We estimated predictors on wordforms (for English) and lemmas (for Russian) as well as over clustering of words in embedding space.

To work with clusters, first, we substitute words with indexes of their respective clusters and then calculate the co-occurrences of these cluster indices (Figure 1). We use cluster analysis to deal with data sparsity problems (which might be present when working with the distribution of word forms). Additionally, this analysis can show us what information the predictors are sensitive to. For instance, clusters contain semantic information about the words that form the cluster.

$$\text{huge, big, massive} \longrightarrow Adj_{35} \ ; \ \text{house, building} \longrightarrow N_{21} \ ; \quad Adj_{35}N_{21}$$

Figure 1: Cluster analysis example: words are replaced with the respective number of their cluster, and then co-occurrences of these indices are calculated in pairs and triples.

We performed clustering with sklearn.cluster.KMeans (Pedregosa et al. 2011). The number of clusters was determined by choosing the largest k, which did not result in singleton clusters. For English, we used KMeans applied to a pre-trained set of 400k 300-dimension GloVe vectors generated from Wikipedia (Pennington, Socher, and Manning 2014). For Russian, we used Navec embeddings[1]. Adjectives were clustered into $k_A = 300$ and nouns $k_N = 1000$ clusters, both in English and Russian. Adverbs were clustered into $k_{Ad} = 50$ and verbs into $k_V = 150$ clusters in English. In Russian, we had $k_{Ad} = 50$ adverb and $k_V = 300$ verb clusters. The variation in the number of clusters for adverbs and verbs might reflect the inherent semantic differences between the two languages.

We calculated corpus-based predictors (PMI, IG, IC) using pairs of Adj-Noun and Adv-Verb. For English adjectives, we also include the subjectivity measure. The scores for 398 adjectives were provided by Futrell, Dyer, and Scontras 2020 following Scontras, Degen, and Goodman 2017 methodology: where 264 English-speaking participants indicated with the slider scale the subjectivity, by adjusting a slider (in a range from “completely objectivity" to “completely subjective"). Each out of 398 adjectives received an average of 20 ratings.

For comparative reasons, we keep triples where D_1 and D_2 have scores for all predictors; at the end, for testing a classifier, we have 8,752 AAN and 2,857 Adv-Adv-V triples in English, and 61,337 AAN and 1,085 Adv-Adv-V in Russian, which we use for fitting the model and evaluation.

4 Results

In both languages, the results differ based on whether the predictor is estimated on a wordform or a cluster, visualized in Figures 2 and 3. Still, we can point out that a specific pattern is present in both languages. When applied to adjective order, PMI is the best predictor when tested on word forms. However, when tested on clusters, its accuracy drops. In this case, integration complexity and information gain show better results. Interestingly, this pattern is present with adverbs too. This might be seen as a confirmation that PMI is good at tracing meaning closeness but may be less effective in capturing the general properties of the cluster to which a word belongs. Additionally, we can see that integration complexity shows better results when tested on clusters. The reason might be that IC is good at capturing general semantic properties of the cluster since it focuses not only on the distribution of words but also on the syntactic category of the heads. The same might be valid for information gain. This pattern aligns with the observations made by Futrell, Dyer, and Scontras 2020.

[1] `https://github.com/natasha/navec`

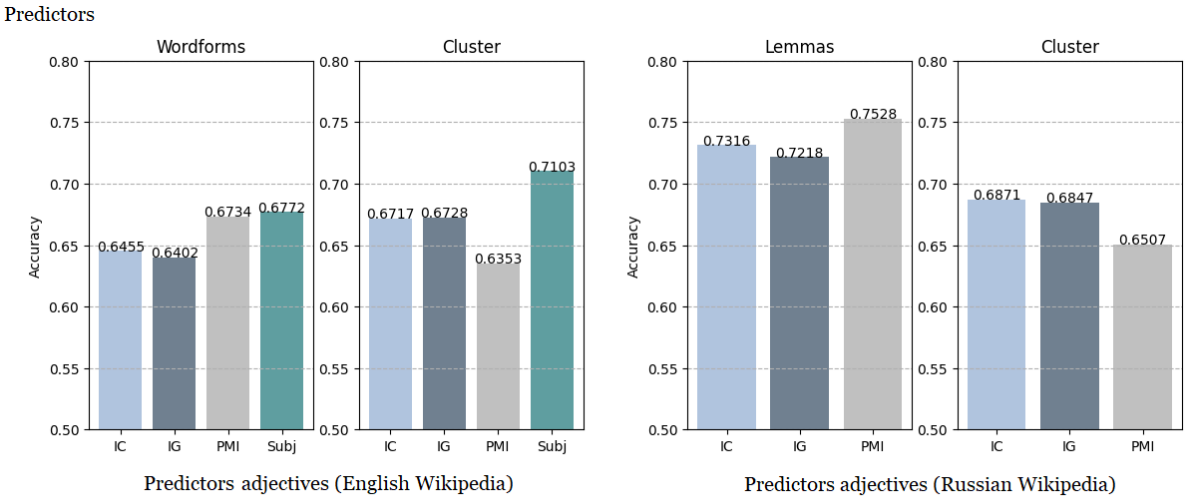

Figure 2: Test accuracy of logistic regression derived from 80:20 train/test split of 8,752 AAN triples extracted from a subpart of English Wikipedia and 61,337 AAN triples from subpart of Russian Wikipedia.

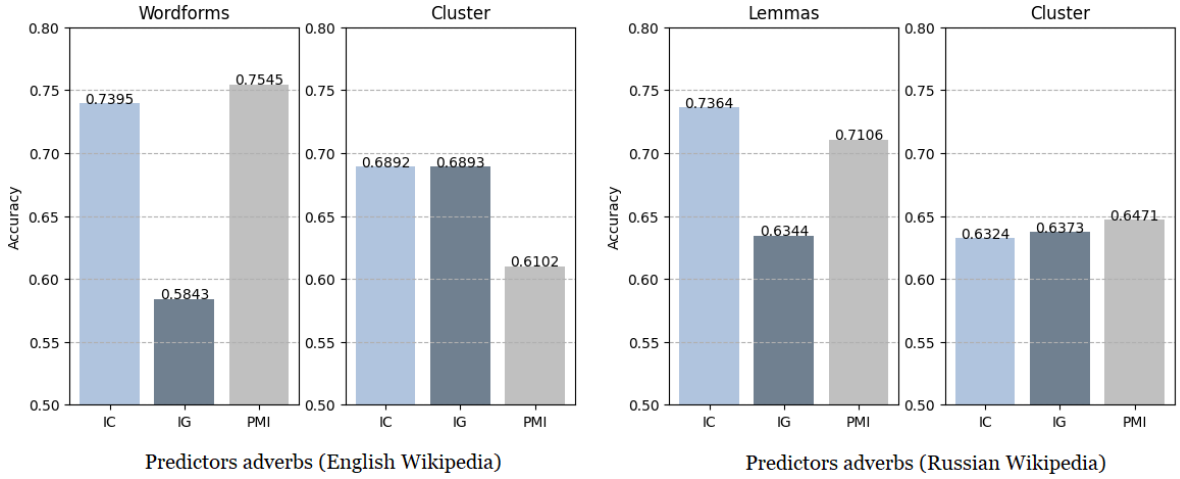

Figure 3: Test accuracy of logistic regression derived from 80:20 train/test split of 2,857 Adv-Adv-V triples extracted from a subpart of English and 1,085 Adv-Adv-V from subpart of Russian Wikipedia.

5 Discussion

The theories integrating syntax and processing efficiency can be tested to explain a broader linguistic phenomena. In our study, we started by replicating the results of Futrell, Dyer, and Scontras 2020 cross-linguistically (on two morphologically different languages) and across two different domains (adjectives and adverbs order). However, certain aspects should be taken into consideration when analyzing the results. First, the study's predictors are text-based, rely purely on syntactic annotations, and thus cannot differentiate between meanings. For future investigation one may want to focus more on the sematic aspect, for instance, by including more specific phrase annotation. Second, it was shown that subjectivity acts as the best predictor of adjective order in the prenominal position. However, collecting significant subjectivity ratings can not always be straightforward, as in the case of the English language. Future work for collecting subjectivity ratings is needed: either manually or including LLMs to get reliable scores

(Jumelet et al. 2024). Lastly, our study concentrates on the base order and shows tendencies rather than universals. We can see these tendencies as "soft biases" towards a base order that can be violated depending on the context and the meaning the speaker wants to convey.

6 Conclusion

In our work, we investigated whether processing efficiency theories can explain the tendencies in the attested order of adjectives and adverbs. Overall, the method of using the dependency structure of a sentence to study word order can be tested on other languages, provided they have a big enough corpus from which to extract pairs of heads and dependents, which are needed to calculate predictors. This research has many future directions: including more languages or examining the explanatory power coming from a combination of several predictors. Future work will be important for thoroughly understanding how information theoretic approaches explain word order.

References

Cinque, G. (1999). *Adverbs and Functional Heads: A Cross-Linguistic perspective.* Oxford University Press.

Danks, J. H. and S. Glucksberg (1971). "Psychological scaling of adjective orders." In: *Journal of Verbal Learning and Verbal Behavior, 10(1)*, pp. 63–67.

Dyer, W. (2017). "Minimizing Integration Cost: A General Theory of Constituent Order". PhD thesis. University of California, Davis.

Futrell, R., W. Dyer, and G. Scontras (2020). "What determines the order of adjectives in English? Comparing efficiency-based theories using dependency treebanks." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2003–2012.

Futrell, R., P. Qian, et al. (2019). "Syntactic dependencies correspond to word pairs with high mutual information". In: *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019).*

Gibson, E. et al. (2019). "How Efficiency Shapes Human Language". In: *Trends in Cognitive Sciences, 23(5)*, pp. 389–407. DOI: https://doi.org/10.1016/j.tics.2019.02.003.

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars.* Oxford, UK: Oxford University Press.

Jumelet, Jaap et al. (2024). *Black Big Boxes: Do Language Models Hide a Theory of Adjective Order?* arXiv: 2407.02136 [cs.CL]. URL: https://arxiv.org/abs/2407.02136.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research 12*, pp. 2825–2830.

Pennington, J., R. Socher, and C. D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Qi, P. et al. (2020). "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.*

Scontras, G., J. Degen, and N. D. Goodman (2017). "Subjectivity predicts adjective ordering preferences." In: *Open Mind: Discoveries in Cognitive Science, 1(1)*, pp. 53–65.