Word order affects the frequency of adjective use across languages

Zeinab Kachakeche, Richard Futrell, and Gregory Scontras
{zkachake, rfutrell, g.scontras} @uci.edu
Department of Language Science
3151 Social Science Plaza A
University of California-Irvine
Irvine, CA 92697-5100

Abstract

Recent research has proposed that adjective form (i.e., whether adjectives typically occur before or after the nouns they modify) interacts with considerations from efficient communication to determine the rate at which we use adjectives to resolve reference to objects. According to this efficiency hypothesis, languages with pre-nominal adjectives use modifying adjectives at a higher rate in an effort to aid incremental reference resolution. We test this broad typological prediction in a large-scale corpus analysis of 74 languages, finding that languages that favor pre-nominal adjectives indeed do exhibit higher rates of adjectival modification than languages that favor post-nominal adjectives.

Keywords: efficient communication; adjective use; reference resolution; incremental processing; cross-linguistic corpus analysis

Introduction

Linguists have long studied the interplay between **form** (properties of linguistic structure) and **function** (the purpose for using individual forms), entertaining the hypothesis that humans use language structures that communicate their intended meaning in a way that maximizes **efficiency**. The current paper investigates this broad hypothesis by taking a magnifying glass to the use of adjectives. In terms of function, adjectives have at least two universal roles: (i) aiding in the resolution of reference by providing additional information about the intended referent and (ii) revealing aspects of the speaker's subjective evaluation. For example, in *the big wooden box*, the adjectives *big* and *wooden* tell us properties of the *box* being referenced. While there is little reason to suspect that adjective function differs across languages, adjective form does: in pre-nominal languages like English, adjectives come before the modified noun, and in post-nominal languages like Arabic, adjectives follow the noun.

Recent research has proposed that adjective form (i.e., whether adjectives typically occur before or after the nouns they modify) interacts with considerations from efficient communication to determine the rate at which we use adjectives to resolve reference to objects. Rubio-Fernández (2016) argues that, when adjectives precede nouns, they serve an additional function of *incrementally* aiding in a visual search for referents, with evidence from visual-world-paradigm experiments. This account claims to explain why speakers often use adjectives over-informatively, for example describing a box as a *brown box* even when there is only one box in a visual scene. In other words, overmodification arises when using adjectives in contexts where they do not provide extra referential information for the nouns they modify. Critically, adjectives would *not* serve this function of incrementally aiding visual search when they occur after nouns, since the noun already establishes reference to the unique box. Accordingly, Rubio-Fernandez et al. (2020) and Wu & Gibson (2020) find that speakers of Spanish (a post-nominal language) use overinformative adjectives at a reduced rate than English speakers in controlled behavioral experiments. Waldon & Degen (2021) also modeled this phenomenon within the RSA framework and predicted that Spanish speakers should use redundant color adjectives less frequently than English speakers, supporting Rubio-Fernández's behavioral results.

The functional consideration that adjectives might be more useful pre-nominally than post-nominally leads us to a typological prediction about language form: adjectives should be used more frequently in pre-nominal languages, where they are more useful. The goal of this work is to test this prediction in massively multilingual corpus analyses of adjective use in 74 languages.

The paper is structured as follows. First, we introduce and motivate our efficiency hypothesis, which predicts that adjectives should be used more frequently in pre-nominal languages. Next, we perform an analysis of adjective use in Wikipedia corpora from three languages: English, Spanish, and Arabic. Finding inconclusive evidence for or against the efficiency hypothesis in our Wikipedia analyses, we then expand our sights to 74 languages in the Universal Dependencies corpora. There, we find robust evidence supporting the efficiency hypothesis: languages with pre-nominal adjectives feature more adjectival modification. Importantly, our cross-linguistic analysis also includes Spanish, English, and Arabic, and we replicate the finding from our Wikipedia analysis that these three languages feature similar rates of adjectival modification. We conclude with a discussion of the efficiency hypothesis in light of our findings, as well as discussion of the merits of large-scale cross-linguistic investigations such as ours.

Background

Humans interpret incoming language in a linear and incremental way (e.g., Tanenhaus et al., 1995), processing words as they are encountered in linear time. For example, in the English phrase *the big wooden box*, listeners hear *big* before *wooden*, and also before *box*. Eye-tracking data show that the adjectives allow a listener in a referential context to incrementally restrict the space of possible referents, first based on their sizes, excluding the not-big items, and then on the basis of their material, excluding the not-wooden items, before arriving at the correct object of reference, namely the box (e.g., Sedivy et al., 1999; although see Qing et al., 2018, for a more nuanced interpretation of eye-tracking data of this sort). In Arabic, a language with post-nominal adjectives, the equivalent phrase starts with the noun 'box' and then the adjectives follow: *al-sundūq al-xašabi al-kabīr* 'the-box the-wooden the-big'; given the incremental nature of language processing, information about the intended referent may be communicated already by the noun, rendering some of the information contributed by the post-nominal adjectives superfluous to reference resolution.

While establishing reference is one of the most common uses of language, studies have shown that adjectives used in referential settings are commonly overinformative with respect to reference resolution (Rubio-Fernández, 2016; Wu & Gibson, 2020; for a recent overview, see Degen et al., 2020). For example, a speaker might choose to say *the brown box* in a case where all of the relevant boxes are brown, and thus the use of the color adjective is redundant—and hence overinformative. According to Grice (1975), a speaker must be maximally informative, without being overly informative. Thus, the use of overinformative adjectives violates the maxim of quantity. One might wonder why speakers consistently violate this maxim, and what communicative goals are acheived by using overinformative adjectives.

One resolution to this apparent paradox of overinformativity relies crucially on the incremental nature of language processing. When overinformative adjectives are used pre-nominally, they enable the listener to perform an efficient, incremental visual search for the intended referent (Rubio-Fernández, 2016; see also Degen et al., 2020, for another analysis of the rationale behind overinformative uses). However, when adjectives occur post-nominally, then they cannot serve this function—reference has already been fully established incrementally by the noun. In this connection, Rubio-Fernández (2016) suggests that Spanish speakers (Spanish is a language with post-nominal adjectives) use overinformative color adjectives less frequently than English speakers; she proposes that these results should extend to all languages with post-nominal adjectives.

Rubio-Fernández's (2016) eye-tracking experiments involved visual displays of two types: either monochrome (where all the objects in the display are of the same color; Figure 1 *left*) or polychrome (where different objects in the display have different colors; Figure 1 *right*). In both cases, using the color adjectives in nominal descriptions is redundant, but English's pre-nominal adjectives aid in the incremental visual search, especially in the polychrome condition. In both monochrome and polychrome contexts, English speakers produced redundant color adjectives more often than Spanish speakers (*monochrome*: 37% for English, 5% for Spanish; *polychrome*: 95% for English, 59% for Spanish). In a related study, Wu & Gibson (2020) compared the use of color adjectives and number words in English and Spanish, taking into account that number words occur pre-nominally in Spanish (like in English), and thus the presence of number words should be as useful in an incremental search as they are in English. Wu & Gibson replicate the basic result from Rubio-Fernández: English speakers use overinformative color adjectives more than Spanish speakers. Importantly, the study also showed no difference in the rate of number word usage between English and Spanish speakers—as one would expect given the efficiency hypothesis, since number words are pre-nominal in both languages.

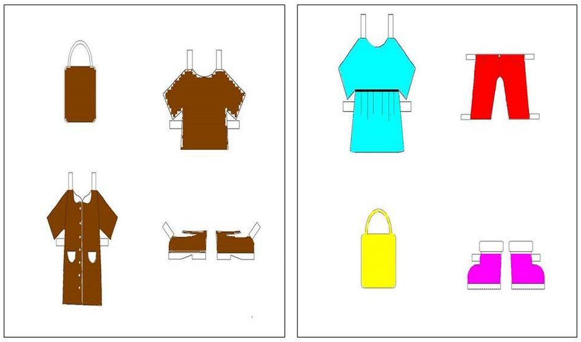

Figure 1: An example of the visual reference paradigm used by Rubio-Fernández (2016). Within the monochrome scene on the left, referring to the shoes as *the brown shoes* would be an overinformative use of the adjective *brown*.

These previous studies look at a very specific case of carefully-controlled referential contexts, namely object requests; the experimental results support the hypothesis that adjectives aid in referential resolution when they occur before the noun, and thus the use of adjectives is less efficient after the noun. We might wonder how broadly the pressure toward efficiency applies in a language, such that adjective modification rates may be lower overall in post- vs. pre-nominal languages.

These ideas about adjective usage are related to more general ideas about the role of efficiency in shaping the form of human language. Gibson et al. (2019) define efficiency from an information-theoretic perspective, meaning that languages allow humans to communicate successfully with minimal effort, subject to cognitive and environmental constraints. If a speaker's intended message is equal to the message received by the listener, then the communication was successful; if this success was achieved with minimal effort, then it was efficient. Rubio-Fernández's hypothesis about adjective use posits that speakers do not use overinformative adjectives post-nominally because such a usage would represent effort without any payoff. If true, the hypothesis would establish a connection between language function and language form that is asymmetrical between post-nominal and pre-nominal positions.

However, the evidence for Rubio-Fernández's hypothesis currently comes only from controlled visual-world experiments, raising the possibility that these patterns of usage might not generalize to other more naturalistic communicative settings, where the communicative environment and thus the functions of language are different. The current paper examines at what rates adjectives are used in both pre-nominal and post-nominal structures in written and spoken corpora, together with the factors that play a role in determining these rates.

Hypothesis

Our hypothesis is that adjectives are used at a higher rate in pre-nominal languages compared to post-nominal languages. As described above, this hypothesis is based on the idea that adjective usage is driven by efficiency considerations. We explore the extent to which this efficiency-based pressure toward more adjectives pre-nominally extends beyond controlled cases of overinformative reference resolution in experiments to the contexts reflected in corpora of naturally-occurring language. Assuming that incremental reference resolution is at least sometimes an objective in written text, then we expect efficiency-based pressures to apply also in these corpora, such that languages with pre-nominal adjectives use adjectival modification more frequently than languages with post-nominal adjectives.

Our analysis consists of two parts: an in-depth analysis of three languages, and a broad analysis of 74 languages. Our first analysis uses Wikipedia corpora from Spanish and English, two languages involved that have been studied with respect to the efficiency hypothesis, as well as Arabic, a language that is more clearly post-nominal (Spanish allows some pre-nominal adjectives). We choose Wikipedia because it instantiates a similar genre across the three languages, and even covers similar topics (i.e., Wikipedia entries). By controlling for genre and topic, we hope to get a fair comparison of modification rates. The Wikipedia corpus also provides us with a written context that is different from the referential experimental contexts utilized in the literature in the original motivation for the efficiency hypothesis.

The broad corpus analysis is conducted on the Universal Dependencies (UD) Treebanks, which include more than 90 languages. These languages are manually tagged for Part-of-Speech and syntactic dependencies. The text in the UD corpus comes from several different genres, such as spoken, weblogs, newspapers, and emails. Comparing the UD results with the results of our more targeted Wikipedia analysis allows for a more robust test of the efficiency hypothesis: do these trends extend beyond the handful of languages typically investigated in psycholinguistic experiments?

Wikipedia analysis

We begin with our analysis of the Wikipedia corpora from Spanish, English, and Arabic.

Tagging the Wikipedia data

We downloaded the Wikipedia corpus from the Wikimedia Dumps,[1] a large data set of encyclopedia articles, open-sourced by Wikipedia. Our work takes advantage of several tools for data collection and analysis. The WikiExtractor (Attardi, 2012) was used to extract plain text from the xml-formatted Wikipedia data. The English corpus consisted of 12 GB of data, while the Arabic and Spanish corpora were 1.2 and 3.7 GB, respectively. To identify nouns and the adjectives that modify them, we first need part-of-speech information. We used the StanfordCoreNLP Part-Of-Speech tagger (Manning et al., 2014) to tag the Spanish and English corpora. For Arabic, we used the Farasa Part-Of-Speech tagger, which is made specifically for Arabic, and achieves state-of-the -art performance on POS benchmarks while being faster (Abdelali et al., 2016). An initial attempt using the Stanford POS tagger took much longer and led to lower accuracy. With the Farasa tagger, we noticed performance on a par with the Stanford tagger's performance on Spanish and English.

Estimating the rate of adjective modification

Once we had part-of-speech tags for each language, we identified all of the nouns (tagged as NOUN, NN, or NNS). We then collected all the adjectives (tagged as ADJ, JJ, or JJS) that are adjacent to the noun. The logic is as follows: we sequentially iterate over all the words in our corpus, and if we hit a noun, we check the two adjacent words (to the right and to the left). The noun will then fall into one of four buckets: (1) not modified, where there are no adjacent adjectives, such as the three-word segment *the box was*, (2) pre-nominally modified (AN), such as the three-word segment *blue box on*, (3) post-nominally modified (NA), such as *ra'aytu al-sundūq al-xašabi* (the equivalent of saw-I the-box the-brown) in Arabic, and (4) mixed modified (ANA), where the noun is modified both pre- and post-nominally, such as *gran caja marrón* (big box brown) in Spanish.

This method of identifying adjectives modifying nouns is only approximate, as (1) we miss any adjectives not directly adjacent to nouns, and (2) we might unintentionally include counts of non-attributive adjectives that happen to be adjacent to a noun by chance. However, this method has the advantage that it can be scaled up to large amounts of text and does not require parsing of structure.

Once we had the nouns and adjectives counts, we then calculated the rate at which adjectives are used to modify nouns in each of these languages. We calculated the rate by dividing the number of nouns with any adjectival modification (one of the three types mentioned above) by the total number of nouns.

Results

The results of our Wikipedia corpus analysis appear in Figure 2 and are summarized in Table 1. The estimated modification rate is nearly identical across the three languages.

[1] https://dumps.wikimedia.org/

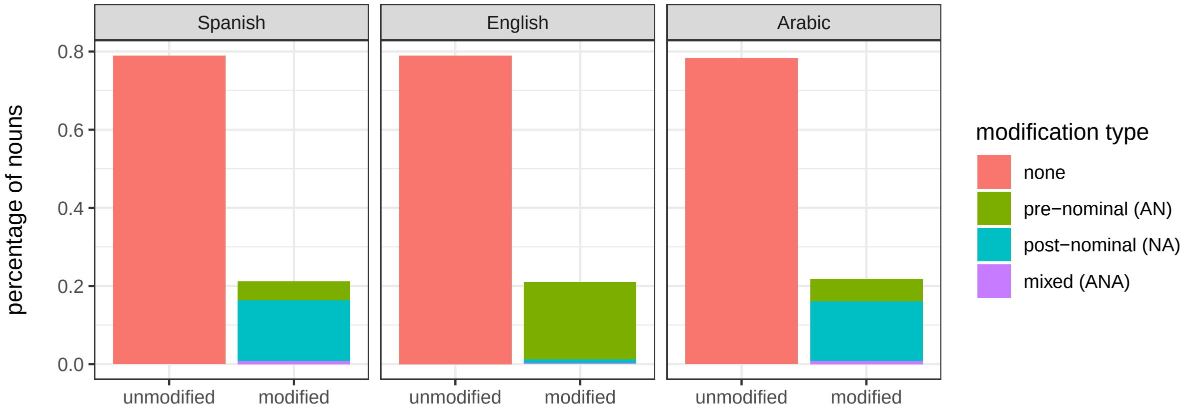

Figure 2: Wikipedia analysis results: percentage of noun modification is similar across the three languages.

Failing to find a reliable difference in modification rates across the three languages in our analysis, we therefore fail to find evidence in support of the efficiency hypothesis. However, there are a couple of caveats. First, the Wikipedia text we investigate may commonly feature translations from one language's articles (often English) into another language. These translations could contribute to similarities across languages that would not arise in the absence of translated content, thereby artificially leveling differences in modification rates. Second, by only investigating three languages, the analysis is severely limited and may therefore miss important generalizations that only become apparent in the consideration of a larger set of languages. In the next section, we attempt to address both of these concerns through our analysis of UD corpora.

Universal Dependencies analysis

The UD corpora come manually tagged with part-of-speech information. Using the same logic as our Wikipedia query, we collected all the nouns and adjacent adjectives. Nouns were again classified into one of four categories: not-modified, pre-nominally, post-nominally, and mixed modified. We excluded languages with zero adjectival modification (Tagalog and Classical Chinese), as well as languages with fewer than 1000 nouns in our search (14 languages total); these criteria yielded data from 74 languages.[2] Using these counts, we then calculated the rate at which adjectives are used to modify nouns in each of the 74 languages by dividing the number of nouns with any adjectival modification by the total number of nouns.

To test the efficiency hypothesis, we need a way of characterizing a language as either pre- vs. post-nominal in its use of attributive adjectives. However, a quick look at the data revealed that languages rarely fit neatly into either category: although a language may favor pre-nominal modification, it is common for there to also be post-nominal uses, and vice-versa.

Rather than relying on a potentially-arbitrary binary classification of our languages as pre- vs. post-nominal, we decided to characterize languages in terms of their propensity toward post-nominal adjectival modification: among the nouns with adjectival modification, what proportion of those nouns feature post-nominal adjectives? We calculate a language's post-nominal propensity by dividing the number of nouns with post-nominal adjectives (note that this number will include both nouns with strictly post-nominal adjectives, and mixed pre- and post-nominal adjectives, as in an ANA template common to languages like Spanish) by the number of nouns with any adjectives at all.[3]

Following the hypothesis of Rubio-Fernández (2016), and if the function of adjectives in writing is sufficiently similar to their function in visual-world-paradigm experiments, we expect languages with a higher post-nominal propensity to feature fewer instances of adjectival modification. In Figure 3, we plot the results of our analysis, where we see clearly that this prediction holds: as a language features more post-nominal modification, the proportion of modified nouns decreases; post-nominal propensity accounts for 18% of the variance in the proportion of noun modification ($r^2 = 0.18$, 95% CI [0.054, 0.325], Spearman's $\rho = -0.41$). Although the effect may be small, it is reliable. We fit a mixed-effects linear regression predicting proportion of nouns with modification by proportion of nouns with post-nominal modification (i.e., post-nominal propensity), with random intercepts by language family. We find a significant effect of post-nominal propensity ($\beta = -0.12$, $t = -3.30$, $p < 0.01$).

One might worry that this trend is driven by the ten languages in the upper-left quadrant of Figure 3 with low post-nominal propensity and high modification rates; these languages—Hungarian, Czech, Upper Sorbian, Belarusian, Croatian, Serbian, Slovenian, Slovak, Ukrainian, and Russian—all have nominal modification rates above 0.33. Importantly, if even if we exclude these ten potential outliers, the pattern

[2] We suspect that the reason no adjectival modification appears in our data for Tagalog and Classical Chinese is that these languages commonly feature a linking particle between adjectives and the nouns they modify (Scontras & Nicolae, 2014).

[3] We used an analogous 'pre-nominal propensity' measure and got nearly-identical quantitative results (i.e., a robust statistical correlation between pre-nominal propensity and modification rate).

Language	# Nouns	%AN	%NA	%ANA	Total % Mod.
English	435,089,686	19.8	1.0	0.2	21.1
Spanish	128,946,337	4.9	15.4	0.8	21.1
Arabic	92,218,932	5.7	15.2	0.8	21.8

Table 1: Results of Wikipedia analyses for English, Spanish, and Arabic. # Nouns gives the total number of nouns extracted. Columns AN, NA, and ANA give the percentage of nouns with adjacent adjectives of the specified type. Total mod. gives the percentage of modified nouns which are adjacent to any adjective.

persists, such that post-nominal propensity predicts nominal modification rates for the 64 languages that remain ($r^2 = 0.12$, 95% CI [0.008, 0.299], Spearman's $\rho = -0.40$). We therefore find strong support for the efficiency hypothesis.

Discussion

We performed an in-depth analysis on Spanish, English, and Arabic Wikipedia text covering more data in these three languages, and a large-scale corpus analysis on Universal Dependencies treebanks, covering a wide range of languages. In the Wikipedia analysis, we failed to find a difference between the rates of adjective modification between the targeted three languages, and thus no evidence supporting the incremental efficiency hypothesis. When looking at the broader analysis of the UD treebanks, however, we see a robust trend consistent with the hypothesis: the rates of post-nominal adjective modification predict the rates of noun modification overall.

One might worry that the results of our two analyses are in conflict: no clear trend in the Wikipedia analysis, but a clear trend in UD. Yet upon closer examination we see that Spanish, English, and Arabic have similar modification rates also in our UD results. In other words, if we had only looked at these three languages also in our UD analysis, we might have (mistakenly) concluded that there is no reliable cross-linguistic relationship between the pre- vs. post-nominal form a language takes and the rate at which it uses adjectives. By conducting a large-scale corpus analysis of dozens of languages, most of them under-studied in psycholinguistic research, we find that this relationship does in fact exist.

Another concern involving our results is that our counts of adjectives in different nominal orders might be skewed. This skewing might be due to errors arising from identifying adjectives that are adjacent to nouns as modifying those nouns. Skewing of this sort could potentially explain why we get such high rates of mixed ANA orders in Arabic, a strongly head-initial language, and a medium post-nominal propensity in languages such as Korean that are strictly head-final. However, it is unclear how the errors introduced by this technique would bias the data towards or against our hypothesis. We chose to use the POS-tag-based approach because it would allow us to generalize between the Wikipedia and UD data, and enable extension to further large datasets that might easily admit accurate POS tagging but not easy or accurate parsing.

Despite these considerations, we appear to have robust support for the efficiency hypothesis: pre-nominal languages are more likely to overmodify nouns given the contribution pre-nominal adjectives make to incremental reference resolution; as a result of this overmodifcation, pre-nominal languages have higher overall rates of adjectival modification. This finding suggests that overmodification happens rather frequently in pre-nominal languages; assuming that pre- and post-nominal languages have an equal rate of adjectival modification in cases where adjectives are necessary for the successful communication of a message, then the increases we observe in pre-nominal languages arise due to overmodification. Given that overmodification arises in cases where adjectives are used in the determination of reference, our results further suggest that the adjectival function of reference resolution commonly arises in both controlled experiments and in corpora of spoken and written language.

One might worry that instances of referential language use—and overmodification specifically—are relatively rare in the corpora we analyze. It would seem unlikely, then, that the effects we observe (i.e., greater adjective use in pre-nominal languages) are driven entirely by cases of overmodification. However, it is possible that the more frequent use of pre-nominal adjectives in genuine contexts of overmodification bleeds over to the use of adjectives overall, such that pre-nominal adjectives are used across the board at a higher rate. Initially small biases of this kind can strengthen over time into robust categorical trends (Kirby, 2017).

Moving forward, we are currently looking into extending our analyses using dependency parses and better-typologically-controlled datasets. It may not be the case that the result we find here survives in a dataset of strictly typologically-controlled languages. And by taking into account the structural information from dependency parses, we can refine our queries to more accurately identify attributive adjectives and the nouns they modify.

References

Abdelali, A., Darwish, K., Durrani, N., & Mubarak, H. (2016). Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations* (pp. 11–16).

Attardi, G. (2012). *Wikiextractor.* `https://https://github.com/attardi/wikiextractor`. GitHub.

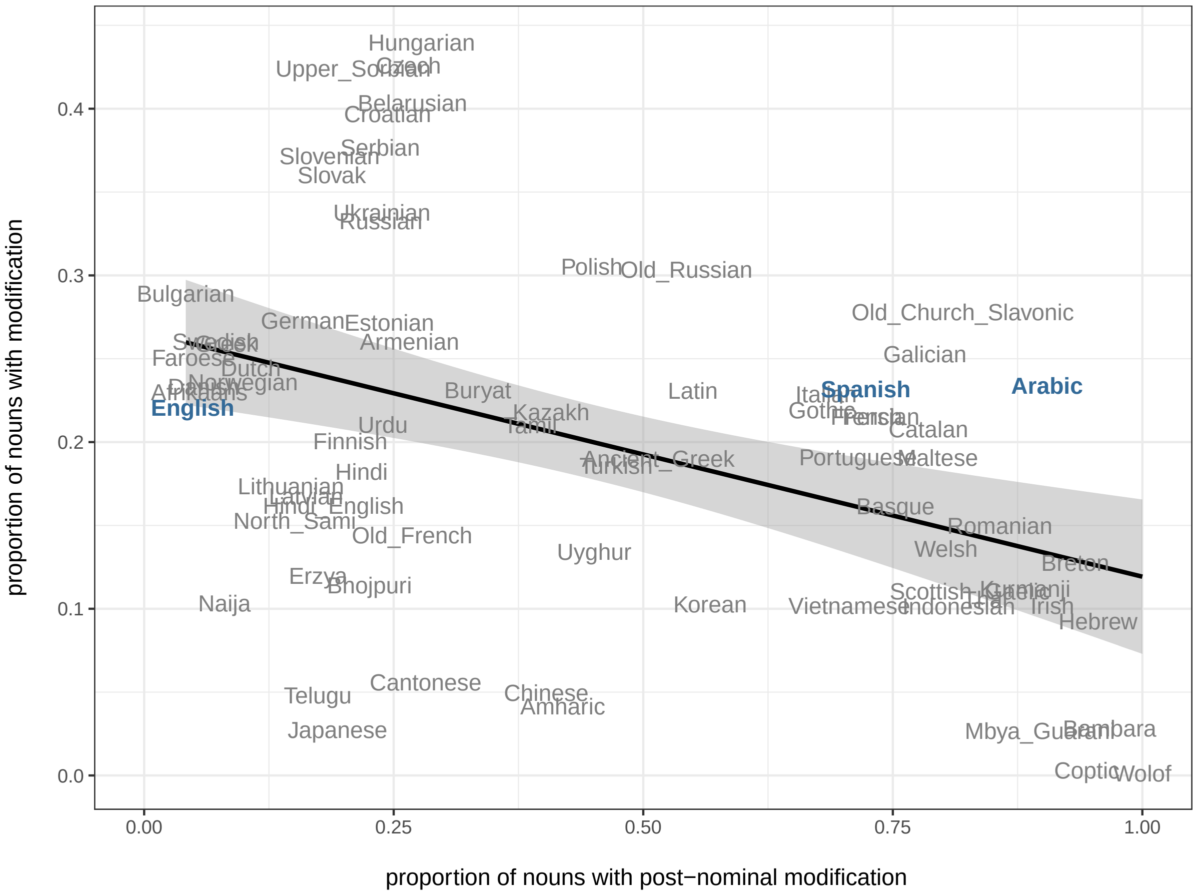

Figure 3: UD analysis results: proportion of nouns with any adjectival modification compared with the proportion of modified nouns with post-nominal adjectives ($r^2 = 0.18$; 95% CI [0.05, 0.32]).

Degen, J., Hawkins, R., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is rational: A bayesian approach to 'overinformative' referring expressions. *Psychological Review*. doi: 10.1037/rev0000186

Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*.

Grice, H. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics, Vol. 3, Speech Acts* (pp. 41–58). New York: Academic Press.

Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychonomic Bulletin & Review*, *24*(1), 118–137.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60).

Qing, C., Lassiter, D., & Degen, J. (2018). What do eye movements in the visual world reflect? a case study from adjectives. In *Proceedings of 40th annual conference of the Cognitive Science Society* (pp. 2297–2302). London: Cognitive Science Society.

Rubio-Fernandez, P., Mollica, F., & Jara-Ettinger, J. (2020). Speakers and listeners exploit word order for communicative efficiency: A cross-linguistic investigation. *Journal of Experimental Psychology: General*.

Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology*, *7*, 153.

Scontras, G., & Nicolae, A. C. (2014). Saturating syntax: Linkers and modification in Tagalog. *Lingua*, *149*, 17-33. doi: 10.1016/j.lingua.2014.05.005

Sedivy, J. C., Tanenhaus, M., Chambers, C., & Carlson, G. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Waldon, B., & Degen, J. (2021). Modeling cross-linguistic production of referring expressions. *Proceedings of the Society for Computation in Linguistics*, *4*(1), 206–215.

Wu, S. A., & Gibson, E. (2020). Word order predicts cross-linguistic differences in the production of redundant color and number modifiers.