Communicative efficiency and syntactic predictability: A cross-linguistic study based on the Universal Dependencies corpora

Natalia Levshina
Leipzig University
Nikolaistraße 6–10
04109 Leipzig
natalia.levshina@uni-leipzig.de

Abstract

There is ample evidence that human communication is organized efficiently: more predictable information is usually encoded by shorter linguistic forms and less predictable information is represented by longer forms. The present study, which is based on the Universal Dependencies corpora, investigates if the length of words can be predicted from the average syntactic information content, which is defined as the average information content of a word given its counterpart in a dyadic syntactic relationship. The effect of this variable is tested on the data from nine typologically diverse languages while controlling for a number of other well-known parameters: word frequency and average word predictability based on the preceding and following words. Poisson generalized linear models and conditional random forests show that the words with higher average syntactic informativity are usually longer in most languages, although this effect is often found in interactions with average information content based on the neighbouring words. The results of this study demonstrate that syntactic predictability should be considered as a separate factor in future work on communicative efficiency.

1 Research hypothesis

It is well known that more predictable information tends to be presented by shorter forms and less coding material, whereas less predictable information is expressed by longer forms and more coding material. This form-function mapping allows for efficient communication. A famous example is the inverse correlation between the frequency of a linguistic unit and its length discovered by Zipf (1935[1968]). The main cause is an underlying law of economy, saving time and effort (*Ibid*: 38).

In the domain of grammar, Greenberg (1966) provided substantial cross-linguistic evidence that relative frequencies of unmarked members of grammatical categories (e.g. singular number or present tense) are more frequent than their marked counterparts (e.g. dual/plural or future/past, respectively). This idea has been developed further by Haspelmath (2008), who provides numerous examples of coding asymmetries in which the more frequent morphosyntactic forms are shorter than the functionally comparable less frequent ones. These asymmetries can be explained by the tendency of language users to make communication efficient: “The overall number of formal units that speakers need to produce in communication is reduced when the more frequent and expected property values are assigned zero” (Hawkins, 2014: 16).

While the accounts mentioned above are based on context-free probability of linguistic units, some other approaches, which go back to Shannon’s (1948) information theory, take into consideration the conditional probability of a unit given its context. The measures computed from these conditional probabilities are often called information content, surprisal, or informativity. There is ample evidence of ‘online’ word reduction in speech production based on contextual predictability (e.g. Aylett and Turk, 2004; Bell et al., 2009). In addition, one has found ‘offline’ effects of average informativity on formal length in written corpora: the more predictable a word is on average, the shorter it is (Piatandosi et al., 2011). One of the explanations of such correlations is known as the hypothesis of Uniform Information Density (Levy and Jaeger, 2007), which says that information tends to be distributed uniformly across the speech

Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), pages 72–78,
Gothenburg, Sweden, 22 May 2017.

signal, so that less predictable elements, which carry more information, get more formal coding, and more predictable elements, which carry less information, get less coding.

In information-theoretic studies, informativity is usually computed from the co-occurrence frequency of a word with the immediately preceding or following word(s) and the frequency of the neighbouring word(s). Corpora of *n*-grams, such as the Google Books Ngrams, are often used for this purpose. The present study goes beyond the *n*-gram approach and investigates if formal length can be predicted from the syntactic dependencies between words, regardless of the order in which the latter occur. A more specific hypothesis, which is tested in the present paper, is that the length of a word can be predicted from its average syntactic informativity. This hypothesis is based on the following intuition. Consider, for example, the English article *the*, which is shorter than most nouns it accompanies. At the same time, it is also more predictable from the specific nouns than vice versa. If one hears the noun *table*, there is a relatively high probability that it will be used with the definite article. In contrast, when one hears the article *the*, one is less likely to expect that it defines the specific noun *table*, simply because there are very many other nouns that can be used with the article. This asymmetry is shown in (1):

(1) $P(the|table) > P(table|the)$

From these contextual predictabilities one can compute syntactic informativity scores. Syntactic informativity is defined here as the negative log-transformed conditional probability of x given y:

(2) $I = -\log_2 P(x|y)$

In our example, the noun *table* is less predictable and therefore more informative than the article *the*. One can expect the words that are more syntactically informative in general to be longer than the less surprising ones. For the purposes of our study, one can define average syntactic informativity as shown in (3). This measure is computed as the sum of all syntactic information content scores of the word in a corpus divided by the number of syntactic dependencies n where it occurs:

(3) $\bar{I} = -1/n \sum \log_2 P(x|y)$

In the present study, average syntactic information content, which is referred to as ASIC in the remaining part of the paper, is computed on the basis of the data from nine typologically diverse languages represented by the Universal Dependencies corpora. The UD corpora provide an advantage of having the same or highly similar syntactic annotation in different languages, which makes the statistical models directly comparable.

2 Data

For this case study, I selected nine languages, which are represented by relatively large Universal Dependencies 2.0 corpora (Nivre et al., 2017): Arabic, Chinese, English, Finnish, German, Hindi, Persian, Russian and Spanish. These languages correspond to different points on the synthetic–analytic continuum and have different writing systems.

The procedure of data extraction was as follows. As a first step, I used a Python script to extract all triples that included a dependent, its head and the syntactic dependency that connects them, such as NSUBJ or AUX. The heads and dependents were represented as wordforms associated with a certain part of speech, e.g. *the*/DET, *table*/NOUN or *goes*/VERB. For the sake of simplicity, these expressions will be referred to as ‘words’ in the remaining part of the paper, with exception of ‘word length’, when the tags are disregarded. Punctuation marks and special symbols were not taken into account. The frequencies of all unique triples were summarized. I also extracted the frequency of every word in a corpus. From these measures, I computed the ASIC score of every word using the formula in (3).

Word length (without the POS tags) was measured in characters, as the UTF-8 string length. Needless to say, this is only a rough approximation of the effort required by speakers to produce the words.

Finally, the data were cleaned up: the words with token frequency less than five were removed, in order to mitigate problems with data sparseness, which arise in small-size corpora. I also removed numeric expressions because they represent a separate semiotic system. Table 1 displays the number of unique words in each corpus after this procedure, as well the source UD subcorpora. One can see that the Chinese data set is the smallest and the Russian one is the largest.

Language	**UD Corpora**	**Number of unique wordforms**
Arabic	UD_Arabic	4,708
Chinese	UD_Chinese	2,246
English	UD_English	3,851
Finnish	UD_Finnish, UD_Finnish-FTB	6,511
German	UD_German	4,269
Hindi	UD_Hindi	4,760
Persian	UD_Persian	3,075
Russian	UD_Russian-SynTagRus	17,811
Spanish	UD_Spanish, UD_Spanish-AnCora	12,832

Table 1. UD corpora and number of unique lemmas for each language.

In addition, I computed the average information content of every word in the data based on its predictability from the word on the left and the word on the right, using the standard procedure described in the literature (e.g. Piatandosi et al., 2011). More exactly, I used the frequencies of the bigrams, which constituted a) the word on the left from the target word and the target word itself, and b) the target word followed by the next word. To compute the average information content, the frequencies of these bigrams were divided by the frequency of the neighbouring word on the left/right, and these proportions were averaged across all occurrences of the target word in the given corpus. Due to the small size of the corpora, it did not make sense to compute the probabilities based on longer *n*-grams.

The next section presents the results of statistical analyses, which were carried out with the help of R (R Core Team, 2016), including add-on packages *car* (Fox & Weisberg, 2011), *party* (Strobl et al., 2007) and *visreg* (Breheny & Burchett, 2016).

3 Statistical analyses

3.1 ASIC across parts of speech and bivariate correlations

As a first step, I performed a descriptive analysis and compared the informativity of different parts of speech in order to test the original intuition. As an illustration, box-and-whisker plots for the English data are displayed in Figure 1. The plot shows that the English determiners are on average less informative than the English nouns, in accordance with the expectations. The analyses reveal that content words tend to be more informative than functional ones across the languages. In particular, adpositions are less informative than nouns, and auxiliaries are less informative than verbs. These observations support the original intuition behind the present study.

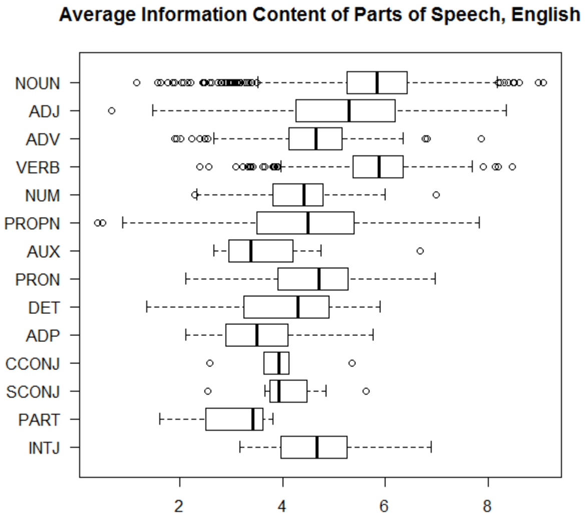

Figure 1. Distribution of average syntactic informativity scores across parts of speech in English.

In addition, I computed Spearman correlation coefficients between ASIC and word length. There are positive highly significant correlations in seven languages: Chinese (Spearman coefficient rho $\rho = 0.29$, $p < 0.0001$), English ($\rho = 0.106$, $p < 0.0001$), German ($\rho = 0.179$, $p < 0.0001$), Hindi ($\rho = 0.24$, $p < 0.0001$), Persian ($\rho = 0.114$, $p < 0.0001$), Russian ($\rho = 0.105$, $p < 0.0001$) and Spanish ($\rho = 0.144$, $p < 0.0001$). In two languages, one finds significant negative correlations: Arabic ($\rho = -0.073$, $p < 0.0001$) and Finnish ($\rho = -0.045$, $p = 0.0002$).

3.2 Poisson generalized linear regression models

The next step was to investigate the relationship between ASIC and word length when taking into account the other frequency-related measures.

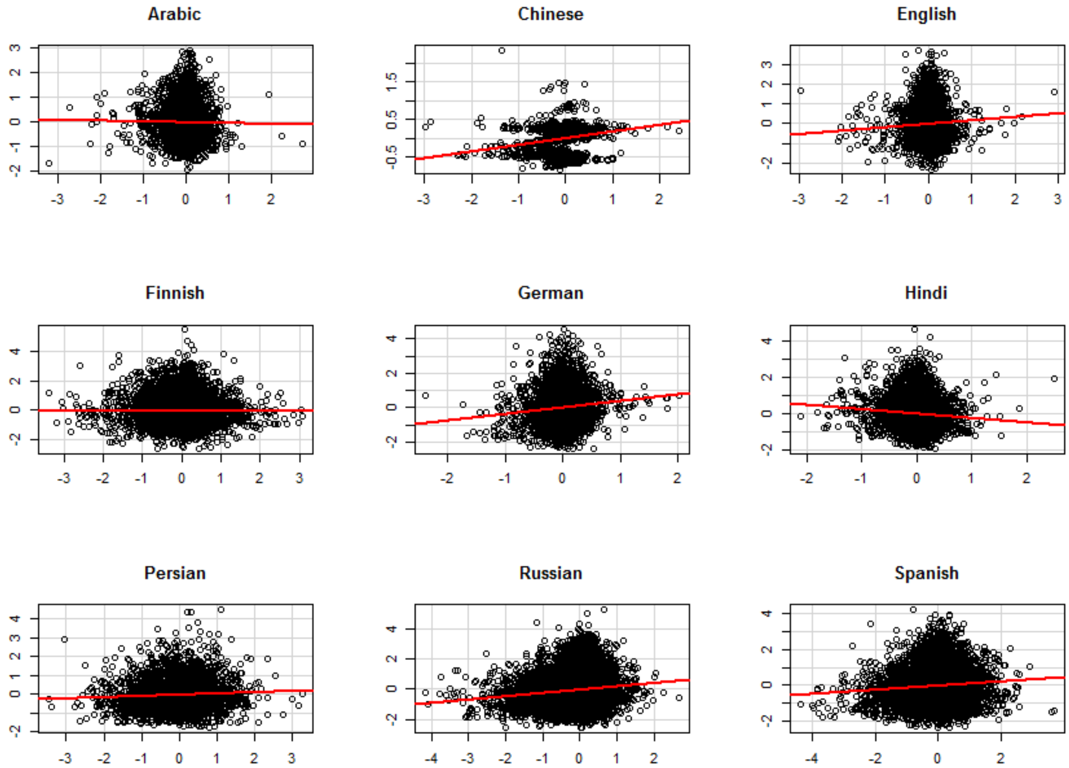

Figure 2. Partial regression plots, which show the direction of the effect of ASIC in nine languages, other variables being controlled for. The axes represent the partial residuals.

For this purpose, I fitted generalized linear regression models for each of the languages with word length (POS tags disregards) as the response and ASIC, as well as average information content based on the preceding word, average information content based on the following word and log-transformed frequency of the target word as predictors. Previous studies have shown that these scores have a substantial effect on word length cross-linguistically (Zipf, 1935[1968]; Piatandosi et al., 2011; Bentz and Ferrer-i-Cancho, 2016). Multiple regression is used here in order to control for the effect of these variables. If ASIC has a separate effect on word length, its estimate will be statistically significant in the presence of all other variables. Poisson models were fitted because the response variable (word length) is always positive, and has a rather skewed distribution.

Figure 2 demonstrates the general effects of ASIC on word length when the effect of other variables is taken into account (so-called added variable, or partial regression plots). The effect is mildly positive in most languages. Exceptions are Arabic and Hindi, where the effect is in fact negative, and there is virtually no effect in Finnish.

However, these results should be taken with caution because of a large number of significant interactions between the variables. To identify interactions, I fitted models with all possible pairwise interactions between the predictors, and then removed those with the *p*-values above the conventional significance level ($\alpha = 0.05$). The remaining interactions were interpreted visually with the help of interaction plots (e.g. see Figure 3). The conclusions based on these plots are presented below.

In Chinese, English, German and Russian, there was a significant interaction between ASIC and average information content based on the preceding word. The interaction is shown in Figure 2. ASIC correlates positively with word length when the interacting variable has smaller values (panels on the left and in the centre), and negatively when it has higher values (see the panel on the right).

In Arabic, German and Spanish, similar interactions are also found with information content based on the following words. There is a significant interaction in German with log-transformed frequency, as well, which follows the same pattern. In Arabic, however, one finds a negative effect of ASIC on the length for most

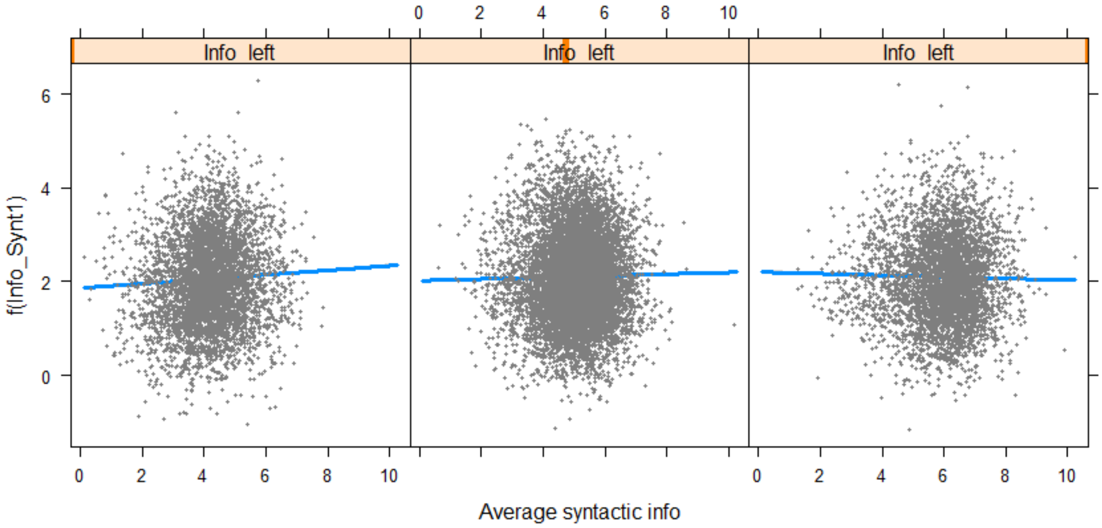

Figure 3. Interaction between ASIC and informative content based on the preceding word in Russian.

values of informative content based on the previous word.

In Hindi, one can find a reverse pattern, when the effect of ASIC becomes stronger and positive as both *n*-gram measures increase. Again, the effect of ASIC is positive in almost all situations. This result is at odds with the negative effect shown in the partial regression plot in Figure 2.

In Finnish, there is a very mild interaction between ASIC and information content based on the following word. The negative effect is observed only for higher values of the interacting variable.

In English, the positive effect of ASIC is only observed for the words which are highly predictable from the left and right context (i.e. have low informativity based on *n*-grams. The positive effect of ASIC is also stronger in highly frequent words.

3.3. Random forests and conditional variable importance

This subsection investigates whether the effect of syntactic informativity is greater or smaller than that of the well-known variables. It is difficult to estimate the importance in the presence of numerous interactions. This is why I used conditional random forests to compute the variable importance scores for each of the variables. This method for regression and classification based on binary recursive partitioning of data allows one to compare strongly intercorrelated and interacting variables (Strobl et al., 2007). Table 3 displays the results. Note that the scores are not comparable between the languages, and the sign does not mean a negative direction of the correlation. These results are based on random samples of 500 words without replacement drawn for every language because the calculation of conditional variable importance is computationally intensive. The random forests were grown from 1,000 trees. Several samples with different random number seeds were tried in order to make sure that the results are stable.

Lang.	**log freq.**	**Synt. info**	**Info given previous**	**Info given next**
Arabic	0.147	0.004	0.022	0.019
Chinese	0.019	0.006	0.016	0.003
English	0.399	0.066	0.032	0.165
Finnish	0.612	-0.007	0.051	0.03
German	0.86	0.192	0.09	0.04
Hindi	0.062	0.206	-0.003	0.114
Persian	0.35	0.026	-0.001	0.007
Russian	0.565	0.028	0.06	0.29
Spanish	0.268	0.021	-0.002	0.308

Table 2. Conditional variable importance scores based on random forests.

The results indicate that average syntactic information is more important than one of the information measures based on the preceding (English, German, Hindi, Persian, Spanish) and/or following words (Chinese, German, Hindi, Persian). There is no effect of ASIC in Finnish, as was already shown in Section 3.2, and it is very close to zero in Arabic, but this variable shows up as the most important one in Hindi.

4 Conclusions

The present study has investigated whether the average predictability of a word given the syntactic dependencies where it occurs can be useful for predicting word length. The analyses of data from nine typologically diverse languages based on bivariate correlations and multivariate Poisson regression reveal that words with higher syntactic informativity (or lower syntactic predictability) tend to be longer in most languages, in accordance with the theoretical predictions. These conclusions also mostly hold when the traditional frequency-based measures, which do not take into account the syntactic information, are controlled for. However, we observed negative or absent correlations in Finnish and Arabic. For Hindi, the evidence based on different methods is somewhat discordant, which requires further investigation.

With the exception of Persian, where syntactic predictability serves as an independent factor, the relationships between the frequency-based measures are usually quite complex. In particular, ASIC often plays a role in the contexts where the other information-theoretical measures have low values. The results of conditional random forest modelling reveal that syntactic predictability often outperforms other frequency-based measures in determining word length.

Although the exact nature of the relationships between different types of lexical and syntactic, context-independent and context-dependent information needs to be further investigated, the results of the present study demonstrate that dependency-based syntactic predictability should be taken into account in future investigations of ‘offline’ communicative efficiency in different languages. Whether it helps to explain formal reduction in ‘online’ language production is a question for future research. Another question the impact of corpus size. One may wonder whether the effects will become stronger and less dependent on the other variables if the measures are computed from the data with higher and therefore more reliable co-occurrence frequencies. Hopefully, the future growth and development of the Universal Dependencies corpora will provide researchers with new opportunities of measuring the effects of syntactic informativity with the help of increasingly large and diverse linguistic data.

Acknowledgments

The author is grateful to the anonymous reviewers, whose constructive comments have helped to improve the paper significantly. The support of the European Research Council (ERC Advanced Grant 670985, Grammatical Universals) is gratefully acknowledged.

References

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31-56.

Alan Bell, Jason Brenier, Michelle Gregory, Cynthia Girand and Dan Jurafsky. 2009. Predictability Effects on Durations of Content and Function Words in Conversational English. *Journal of Memory and Language*, 60(1): 92-111.

Christian Bentz and Ramon Ferrer-i-Cancho. 2016. Zipf’s law of abbreviation as a language universal. In Bentz, Christian, Gerhard Jäger and Igor Yanovich (eds.), *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. University of Tubingen, online publication system: https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68558.

Patrick Breheny and Woodrow Burchett. 2016. visreg: Visualization of Regression Models. R package version 2.3-0. https://CRAN.R-project.org/package=visreg.

John Fox and Sanford Weisberg. 2011. *An R Companion to Applied Regression*. 2nd ed. Thousand Oaks, CA: Sage, http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

Joseph Greenberg. 1966. *Language universals, with special reference to feature hierarchies*. The Hague: Mouton.

Martin Haspelmath. 2008. Frequencies vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics*, 19(1): 1–33.

John A. Hawkins. 2014. *Cross-linguistic Variation and Efficiency*. Oxford: OUP.

Roger Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In Bernhard Schlökopf, John Platt & Thomas Hoffman (eds.), *Advances in neural information processing systems (NIPS)* Vol. 19, 849–856. Cambridge, MA: MIT Press.

Joakim Nivre, Željko Agić, Lars Ahrenberg et al. 2017. Universal Dependencies 2.0, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles

University in Prague, http://hdl.handle.net/11234/1-1983.

Steven T. Piantadosi, Harry Tily and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *PNAS,* 108(9). www.pnas.org/cgi/doi/10.1073/pnas.1012551108

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Claude E. Shannon. 1948. A Mathematical Theory of Communication, *Bell System Technical Journal*, 27: 379–423 & 623–656.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis & Torsten Hothorn. 2007. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*, 8, 25, http://www.biomedcentral.com/1471-2105/8/25.

George K. Zipf. 1935 [1968]. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Cambridge, MA: MIT Press.