

# Uniform Density in Linguistic Information Derived from Dependency Structures

Michael Richter<sup>1</sup>, Maria Bardají I Farré<sup>2</sup>, Max Kölbl<sup>1</sup>, Yuki Kyogoku<sup>1</sup>,  
J. Nathanael Philipp<sup>1</sup>, Tariq Yousef<sup>1</sup>, Gerhard Heyer<sup>1</sup> and Nikolaus P. Himmelmann<sup>2</sup>

<sup>1</sup>*Institute of Computer Science, Natural Language Processing Group, Leipzig University, Germany*

<sup>2</sup>*Institute of Linguistics, University of Cologne, Germany*

**Keywords:** Dependency Structures, Uniform Information Density, Universal Dependencies.

**Abstract:** This pilot study addresses the question of whether the Uniform Information Density principle (UID) can be proved for eight typologically diverse languages. The lexical information of words is derived from dependency structures both in sentences preceding the sentences and within the sentence in which the target word occurs. Dependency structures are a realisation of extra-sentential contexts for deriving information as formulated in the *surprisal model*. Only subject, object and oblique, i.e., the level directly below the verbal root node, were considered. UID says that in natural language, the variance of information and information jumps from word to word should be small so as not to make the processing of a linguistic message an insurmountable hurdle. We observed cross-linguistically different information distributions but an almost identical UID, which provides evidence for the UID hypothesis and assumes that dependency structures can function as proxies for extra-sentential contexts. However, for the dependency structures chosen as contexts, the information distributions in some languages were not statistically significantly different from distributions from a random corpus. This might be an effect of too low complexity of our model's dependency structures, so lower hierarchical levels (e.g. phrases) should be considered.

## 1 INTRODUCTION


The current work is a pilot study based on the hypothesis that dependency structures may serve as proxies for the contextual factors relevant in computing Uniform Information Density (UID). The primary research question is: can the UID-principle in linguistic utterances be proved by information derived from dependency structures in extra-sentential contexts of a target word?


The UID principle claims of a universal condition of successful linguistic communication. It says that the flow of information in any natural language should be uniform, that is to say, without extreme peaks and troughs of information, in order not to overload the


communication channel's capacity. In case of an information overload, the processing of the message - and thus successful communication - is threatened, for instance, the comprehension of a linguistic message. (Jaeger, 2010, 25) formulates the following cognitive principle:


“Within the bounds defined by grammar, speakers prefer utterances that distribute information uniformly across the signal (information density). Where speakers have a choice between several variants to encode their message, they prefer the variant with more uniform information density (*ceteris paribus*).”


If the density of information in a linguistic utterance is “dangerously high” (Levy and Jaeger, 2007), that is, if the information structure within that utterance exhibits extreme peaks, this might cause massive problems in comprehension. Consider, for example, the garden path sentence with dropped relative pronoun *the horse raced past the barn fell*, which, as (Crocker and Demberg, 2015) demonstrate, is ex-


<sup>a</sup> <https://orcid.org/0000-0001-7460-4139>

<sup>b</sup> <https://orcid.org/0000-0003-3512-0435>

<sup>c</sup> <https://orcid.org/0000-0002-5715-4508>

<sup>d</sup> <https://orcid.org/0000-0003-0577-7831>

<sup>e</sup> <https://orcid.org/0000-0001-6136-3970>

<sup>f</sup> <https://orcid.org/0000-0002-4385-8395>

tremely hard to process since the sentence-final fell is highly surprising and has thus a high information value, thereby forming an information peak. UID, therefore, seems to be an essential principle of language processing. This study aims to check whether the UID principle holds when extra-sentential contexts for calculating the information content of words are considered. Formal definitions of UID (given in 6 and 7 below) consider both the variance of information in messages, for example, in sentences, and the change of information from sign to sign in messages, for instance, from word to word in sentences (Collins, 2014; Jain et al., 2018). Our prediction for the eight languages in focus is that the variance and information change per word will be small on average and per sentence.

In contrast to the assumption of the cross-linguistically valid UID principle, we assume that information derived from dependency structures is language-dependent. Therefore we exploit corpora from typologically different languages: the empirical testing ground for this study is a convenience sample of the non-European languages Indonesian and Arabic and some European languages from different language subfamilies, i.e. Russian (Slavic), Spanish, French (Romance), Swedish and German (Germanic). Including more than one language from the same (sub)family allows us to see whether the Romanic or the Germanic languages, respectively, behave similarly.

Shannon defines information as the likelihood of a sign  $s$  (Shannon, 1948). Shannon Information (SI), in bits, is the log-transformation of the sign's probability whereby  $s$  represents a sign, given equation 1:

$$SI(s) = -\log_2(P(s)) \quad (1)$$

Intuitively, the number of bits corresponds to the number of 'yes/no'-questions to determine a possible state in a probability space, and it is important to clarify that information in Shannon's theory is different from the concept of 'information' in linguistics and also in everyday language use. Initially, the meaning of messages was not of any interest for Shannon, since "[...] semantic aspects of communication are irrelevant to the engineering problem[...]" (Shannon and Weaver, 1949), i.e. for the optimal coding and transmission of messages. In particular since the seminal work of (Dretske, 1981), the relationship between Shannon information and natural language understanding came into focus (see for instance (Resnik, 1995; Melamed, 1997; Bennett and Goodman, 2018)). This is also important for our study, since the UID deals with principles of language understanding. In surprisal theory (Hale, 2001), infor-

mation is derived from conditional probabilities, i.e., given a context. (Levy, 2008) equals information content of a sign with its surprisal, which, in turn, is proportional to the processing effort it causes (Hale, 2001; Levy, 2008): the more surprising a sign  $s$  is, that is to say, the smaller its probability is in a context. The more informative  $s$  is, the higher the effort is to process it. This relationship is given in 2:

$$difficulty \propto surprisal \quad (2)$$

This corresponds to Zipf's law (Zipf, 2013), which describes the negative correlation between frequency and length of linguistic signs and, in addition, the *principle of least effort* (Zipf, 1949): frequently occurring signs tend to be short, rarely occurring ones tend to be longer and tend to have higher information content. (Levy, 2008) points out that in the *Surprisal Model* model of language comprehension that employs information theory, large, extra-sentential contexts need to be considered to estimate a word's information content (IC). This is represented in 3 by the variable *CONTEXT*:

$$SI(w_i) = -\log_2 P(w_i|w_1...w_{i-1}, \text{CONTEXT}) \quad (3)$$

However, (Levy, 2008) gives no clear definition of what a context is. This makes the notion of extra-sentential context somewhat challenging to grasp and might explain that, to our best knowledge, there are no studies on the calculation of information utilising large contexts. In this paper, we will explicitly take up the idea of extra-sentential contexts by using dependency structures on the highest hierarchy-level (directly below the verbal root-node) in sentences that precede the target word and in the actual sentence a target word occurs. Thereby we take up an idea from (Levshina, 2017), who estimated lexical information from dependency structures. However, in contrast to Levshina, we consider complete syntactic dependency patterns in sentences. We assume that the languages in focus differ in terms of their dependency structures: differences in the position of subjects and objects are to be expected because our set of languages contains both strongly inflected and weakly inflected languages, with the former tending to have greater freedom in word position in the sentence. Consequently, when deriving lexical information from language-specific dependency structures, there should be differences in the distribution of information between the languages.

The present study uses the existing – partly quite small - corpora in the Universal Dependency Treebanks<sup>1</sup>; in this respect, our study is based on convenience sampling. For some languages like Indonesian, larger Dependency treebanks as models need to

<sup>1</sup><https://universaldependencies.org>

be annotated to prepare testing corpora of sufficient sizes (a task we are currently engaged in). Our intention is to subject the usability of a syntactic-semantic context type for deriving the information content of words to a first empirical test based on these data.

## 2 RELATED WORK

(Levy, 2008; Levy, 2018) found a positive correlation between surprisal and processing effort of signs, which underpins the relevance of UID. Processing effort was operationalized by measuring reading times: surprising words in sentences need more time to be read. In their study on the omission of the relative pronoun in relative clauses (RC) in English, (Levy and Jaeger, 2007) showed that *that* as a relative pronoun is omitted if RC is less informative. However, in unexpected and (too) high informative RC, that is not omitted: The relative pronoun signals to the human processor that a relative sentence follows, thus reducing the amount of surprisal and information. The study of (Horch and Reich, 2016) provided evidence that for article omission in German, UID holds on the level of non-terminal POS-tags and that POS-tags provide even a better basis for explaining article-omission than terminal symbols. Exploring Universal Dependency corpora from 30 languages, (Richter et al., 2019) observed UID within two types of lexical information of single words, i.e., lexical information from pure unigram frequencies and lexical information from conditional probabilities (n-grams): the two lexical information values correlate negatively, but the variance of information is not high, that is to say, the information density is uniform (for the definition of UID see equation 6 and equation 7) (Richter and Yousef, 2019) came to the same result in their study on verbal information content in six Slavic languages, which is illustrated in figure 1 for Polish, Slovenian and Latvian (The a-axis gives the UID-values, the y-axis depicts the raw frequencies of the values).

## 3 CORPORA AND METHOD

As data resource, we utilised eight UD treebanks (number of sentences in brackets) (version 2.8), i.e., ar\_padt-ud-train.conllu.txt (7664), en\_ewt-ud-train.conllu (16621), fr\_gsd-ud-train.conllu (16341), de\_gsd-ud-train.conllu (15590), id\_gsd-ud-train.conllu (5593), ru\_gsd-ud-train.conllu (5030), es\_gsd-ud-train.conllu (16013), sv\_pud-ud-test.conllu (1000). We consider subject, direct object, indirect object and oblique as syntactic complements (in the

case of several oblique elements, we took only the first ) in sentence frames on the top hierarchical level below the verb root node. We calculated first the individual information values of the word forms, given in equation 4:

$$I(w) = -\frac{1}{N} \sum_{i=1}^N \log_2 (P(w) * P(w|c_i)) \\ = -\log_2 P(w) - \frac{1}{N} \sum_{i=1}^N \log_2 P(w|c_i) \quad (4)$$

The information content of a word  $w$  is the average of its information  $I$ , which is made up of Shannon Information and Surprisal of  $w$ . Equation 4 expresses the concatenation of two types of information: information of a given word in relation to all alternative words and information in its contexts  $c_1 \dots c_n$ , i.e., the dependency structures in the environment of the target word. A short fictitious example, starting from figure 2, may clarify the idea of deriving information from a dependency frame as context. Note that the example just covers  $P(w|c_i)$  in equation 4 above:

Figure 2 depicts a sort of dependency structure with which the verb *esse*<sub>[1st person singular]</sub> ‘eat’ occurs. The labels on the top level, directly below the node TOP\_ROOT / S are SUBJ (=subject), OBJA (=direct object) and ROOT (which indicates punctuation). Without ROOT, the dependency structure is SUBJ-OBJA. Let us say, that is dependency structure #1. Let us further assume that in the entire corpus *esse* occurs in 100 sentences, 50 times with dependency structure #1, 30 times with dependency structure #2, which is, say, SUBJ-OBJA-OBJD(=indirect object) and 20 times with dependency structure #3, say, SUBJ-OBJA-OBL(=oblique). The dependency structures #1, #2, and #3 occur each 1000 times in the corpus. The average information of *esse* derived from these dependency contexts is given in equation 5:

$$\bar{I}(\text{esse}) = -\frac{1}{3} (\log_2 0.05 + \log_2 0.03 + \log_2 0.02) \\ = 5.01 \text{ bits} \quad (5)$$

From the information values, we calculated the Uniform Information Density (Collins, 2014; Jain et al., 2018), that is in particular (i) the Global Uniform Information Density ( $UID_{global}$ ), i.e., the average variation of information within the sentences of a language in a corpus (calculated with 6) and finally (ii) the Local Uniform Information Density ( $UID_{local}$ ), i.e., the average degree of information changes from word to word within the sentences (equation 7). Where  $IC$  is the information content of a word form,  $N$  is the number of sentences in the corpus,  $M$  is the number of tokens in a sentence, and is  $\mu$

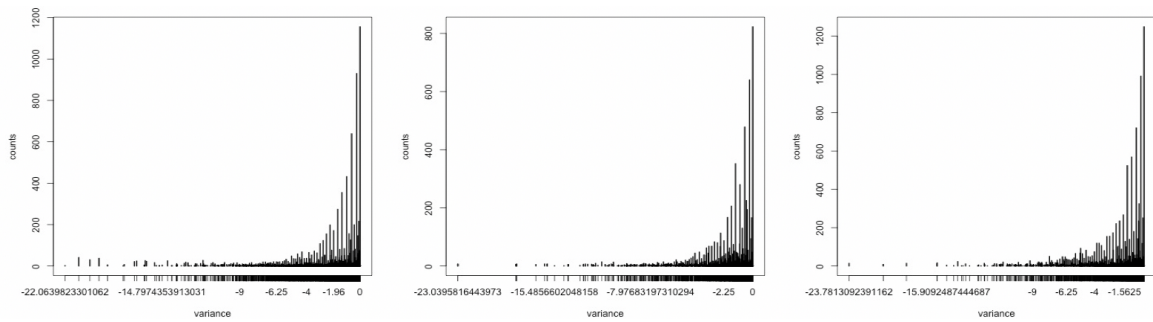


Figure 1: P Uniform Information Density in Polish, Slovenian and Latvian.

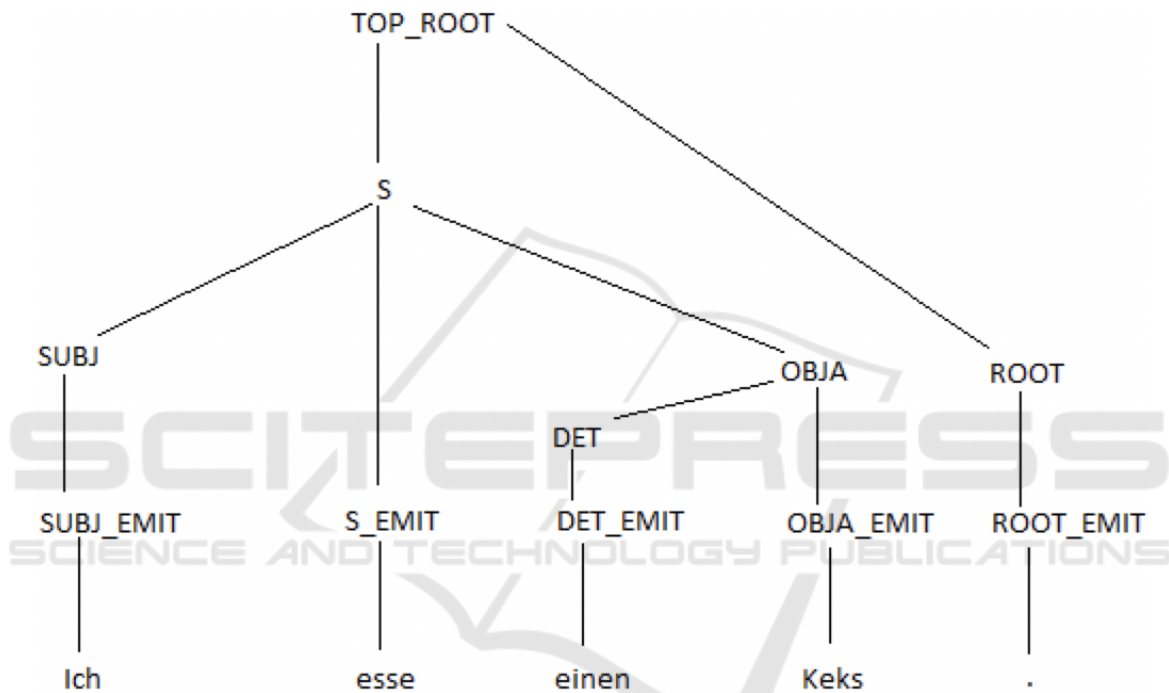


Figure 2: A type of dependency structure of the sentence *ich esse einen Keks* 'I eat a cookie'.

the average information content of a sentence.

$$UID_{GLOBAL} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M (IC_{ij} - \mu)^2 \quad (6)$$

$$UID_{LOCAL} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M (IC_{ij} - IC_{ij-1})^2 \quad (7)$$

## 4 RESULTS

Figure 3 gives six figures for each language: from left to right, these are (1) the density of the distribution (the area under the curve is 1) of the information values derived from the dependency structures of the previous sentence, (2) the density of the information distribution from the dependency structures

of the current sentence the target word occurs in, (3) the density distribution of  $UID_{global}$  of the previous sentence, (4)  $UID_{global}$  of the current sentence, (5)  $UID_{local}$  of the previous sentence, (6)  $UID_{local}$  of the current sentence. The x-axis in the plots of the first two columns in figure 2 depicts the information values. The y-axis depicts the distribution of information values. In columns three and four, the x-axis depicts the variance of information in sentences, i.e.,  $UID_{global}$ . Columns five and six depict the difference in information from word to word in sentences. The y-axis depicts the density in all plots, and the area under the curve should be 1.

Figure 3 shows differences and similarities in the distribution of information between languages (plots in columns 1 and 2 from left). The genetically areally related languages Germanic languages German and

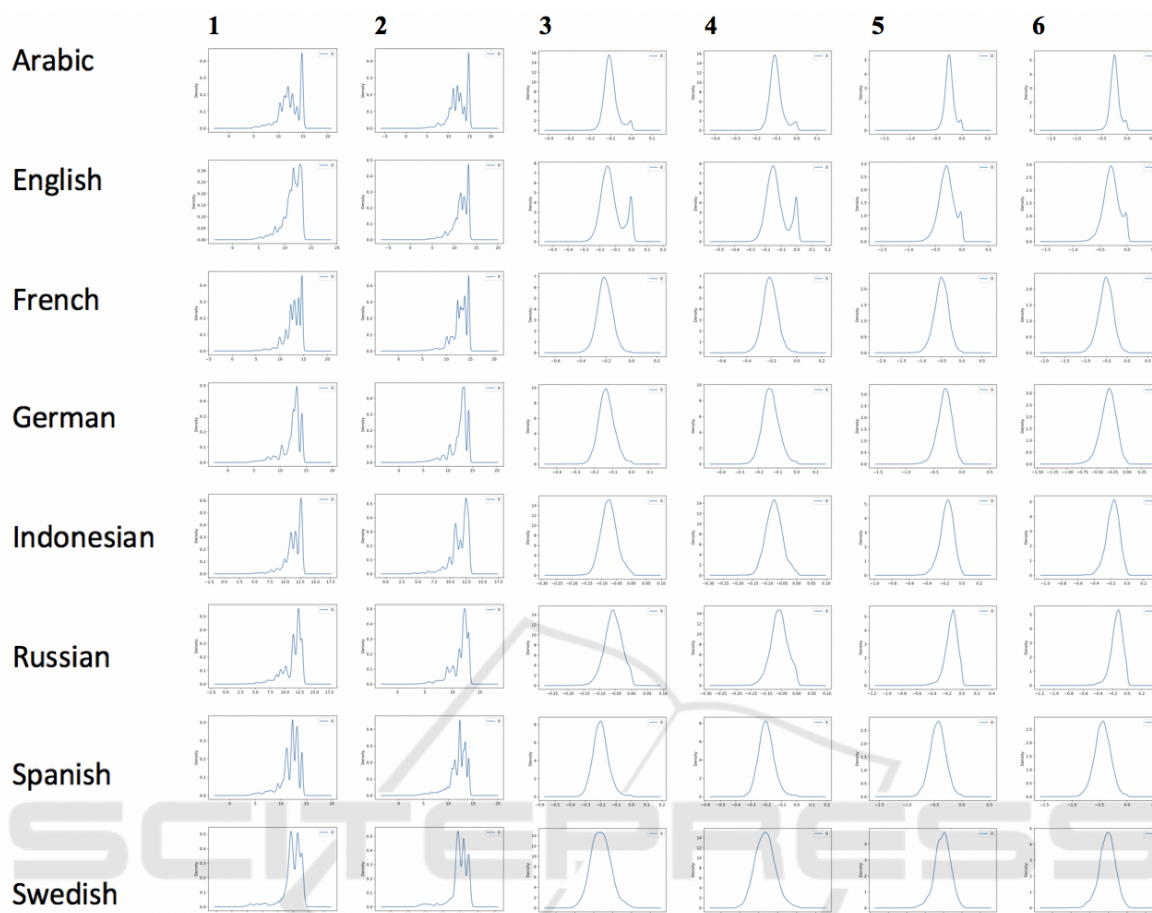


Figure 3: Distribution of Information and  $UID_{global}$  and  $UID_{local}$  in eight languages. Columns 3 and 4 contain the plots of  $UID_{global}$ , and columns 5 and 6 contain the plots of  $UID_{local}$ , for both UIDs within previous and current sentences, respectively. In columns 1 and 2, the x-axis gives the information values, in columns 3 - 5, the x-axis depicts the UID-values. In all plots, the y-axis depicts the distribution of the respective values whereby the area under the curve is 1.

Swedish cluster together and show similar curves. In contrast, Russian stands somewhat out from the other languages. However, this does not appear to be necessarily so since we observed similarities between Arabic, English, Indonesian and the two Romance languages, French and Spanish. The UID curves (plots 3 - 6 in figure 3 above) show strong similarities between the languages. With  $UID_{global}$ , English stands out a little, but with  $UID_{local}$ , almost identical density curves are shown for all languages. We also note that the values for both  $UID_{global}$  and  $UID_{local}$  are close to 0. That is to say, the sentence-wise variance of information in the eight languages and the average change of information from word to word is small across the languages. In order to assess the relevance of dependency structures at the top hierarchical level as contexts for the information value calculation, we compared the single language information values in pairs with values from a random corpus using T-Tests

and Mann-Whitney-U-Tests.

We created the random corpus by extracting each sentence frame and the set of words that belong to the sentence frame. They are stored respectively in a set of sentence frames and in a set of word sets, in which the word order is maintained as in the sentence. Then the set of sentence frames and the set of word sets are randomly combined to change the original arrangements between sentence frames and word sets. The elements of a sentence frame must be fewer than those of a word set. If a sentence frame is combined with a word set whose elements are fewer than those of the sentence frame, they are rearranged. For example, the sentence frame "[nsubj, obj, obl]" cannot be combined with a word set "[I, sleep]", because it is impossible for the word set with two elements to possess three elements as a sentence frame.

The second and third columns of table 1 show the p-values of the tests. A " $\sqrt$ " means the assumption of

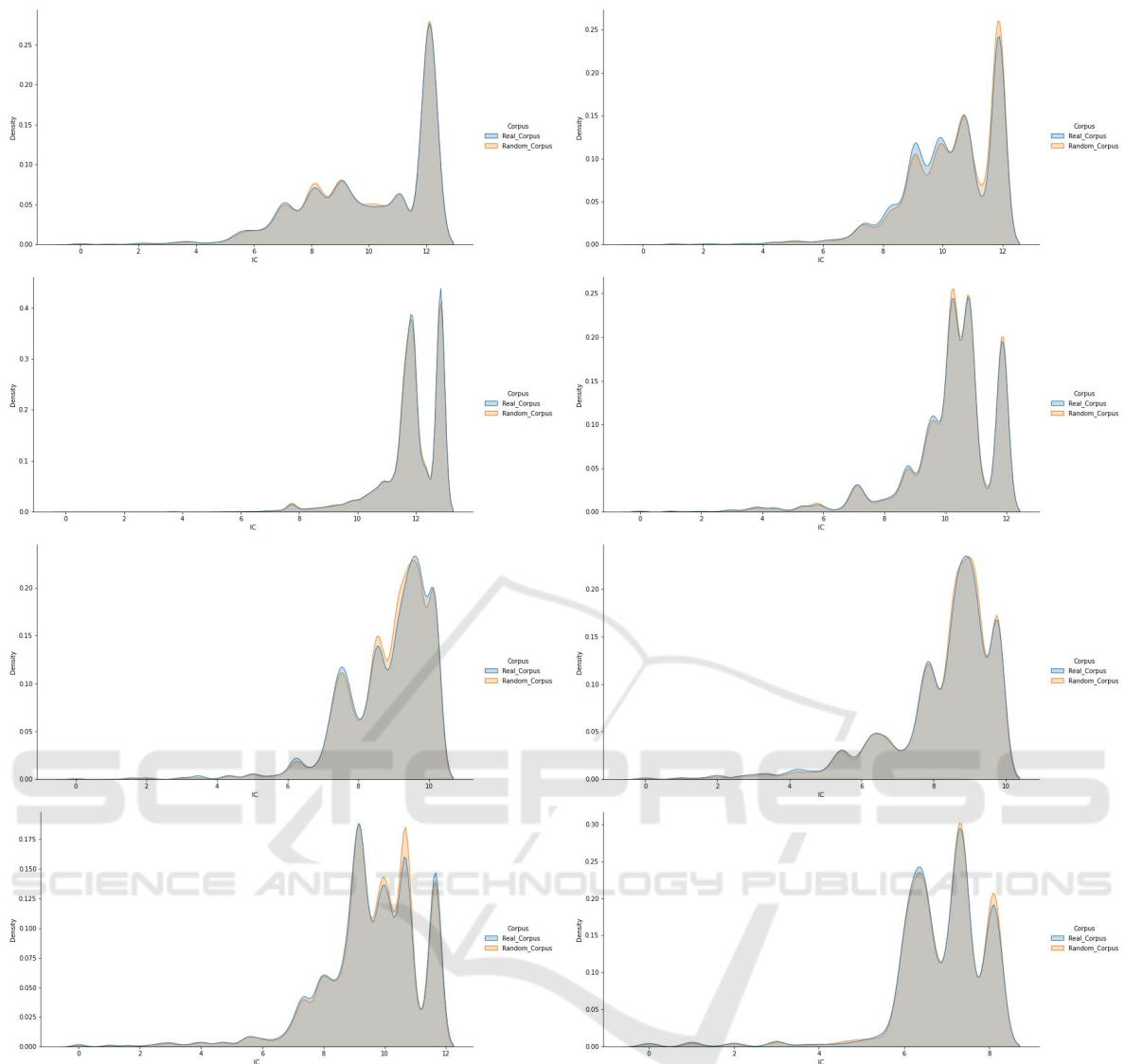


Figure 4: Density plots of the distribution of information derived from dependency structures in UD-corpora and in random corpora in the eight languages. (Top row from left to right: Arabic, English; second row: French, German; third row Indonesian, Russian; bottom row Spanish and Swedish). The x-axis gives the information values, the y-axis depicts their distribution whereby the area under the curve is 1.

the null hypothesis, i.e., that both sets of information values (language corpus-random corpus) are equally distributed.

In English, French, Russian and Spanish, both the null hypothesis of equality of the distribution function and the equality of the means are rejected. In Arabic and German, both the null hypothesis of equality of the distribution function and the equality of the means are confirmed. In Indonesian and Swedish, the equality of the distribution function is confirmed only by the Mann-Whitney U-Test. These results suggest that dependency structures at the highest hierarchical level cannot derive different information courses in all

languages. This makes it necessary to include dependency structures at lower hierarchical levels to derive information values (see section 5). Figure 4 depicts density plots of the distributions of information values derived from dependency structures in (top row from left to right) Arabic, English, French and (bottom row from left to right) Indonesian, Russian, Spanish and Swedish. The information values from the real corpus are highlighted in blue, and the random corpus values in orange. The grey colour indicates their overlapping range. Even in languages where the H0 hypothesis was rejected, such as English, the curves are very similar. This observation again underlines the

Table 1: Test for equality of the distribution functions of the information values and equality of the means of natural language corpora random corpus by the Mann-Whitney-U-Test and the T-Test.

| Language   | Mann-Whitney-U-test<br>p-values | T-test |
|------------|---------------------------------|--------|
| Arabic     | 0.85 ✓                          | 0.37 ✓ |
| English    | ~ 0                             | ~ 0    |
| French     | ~ 0                             | ~ 0    |
| German     | 0.34 ✓                          | 0.09 ✓ |
| Indonesian | 0.51 ✓                          | 0.03   |
| Russian    | ~ 0                             | ~ 0    |
| Spanish    | ~ 0                             | ~ 0    |
| Swedish    | 0.06 ✓                          | 0.02   |

necessity of including dependency structures of lower hierarchies to calculate the information content.

## 5 CONCLUSION AND DISCUSSION

Due to the sparseness of the data, the results of our study are to be interpreted with all due caution. We got evidence for the UID principle in linguistic utterances with information of words that is derived from dependency structures. We observed across the eight languages in focus almost identical distributions of the density values, which strengthen the hypothesis of UID as a universal linguistic principle to enable and facilitate language processing. However, the question of whether dependency structures are suitable as contexts could not yet be answered conclusively: There are differences in lexical information between the languages, but for some languages, we found the HO-hypothesis to be valid, that is, there was no statistical difference between information from the actual corpus and a random corpus. We draw the preliminary conclusion that deriving lexical information content from complex sentence dependency structures seems to be a promising approach - after all, the UID principle could be substantiated. Nonetheless, future research will have to address the inclusion of more complex dependency structures from lower levels in the sentence, such as phrase structures, to derive words' information, and we will also expand the database. Additionally, for an operationalising of the complete surprisal-model, additional extra-sentential contexts will have to be tested, for example, semantic contexts in the discourse of a target word (Kölbl. et al., 2020; Kölbl et al., 2021; Richter and Yousef, 2020), and so future research will address the concatenation of the different types of extra-sentential context types in the surprisal model.

## ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number: 442315837.

## REFERENCES

- Bennett, E. D. and Goodman, N. D. (2018). Extremely costly intensifiers are stronger than quite costly ones. *Cognition*, 178:147–161.
- Collins, M. X. (2014). Information density and dependency length as complementary cognitive models. *Journal of psycholinguistic research*, 43(5):651–681.
- Crocker, M. and Demberg, V. (2015). Surprisal theory and empirical evidence. <https://www.coli.unisaarland.de/~vera/InfoTheoryLecture3.pdf>.
- Dretske, F. (1981). *Knowledge and the Flow of Information*. MIT Press.
- Hale, J. (2001). A probabilistic early parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.
- Horch, E. and Reich, I. (2016). On “article omission” in german and the “uniform information density hypothesis”. *Bochumer Linguistische Arbeitsberichte*, page 125.
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62.
- Jain, A., Singh, V., Ranjan, S., Rajkumar, R., and Agarwal, S. (2018). Uniform information density effects on syntactic choice in hindi. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 38–48.
- Kölbl, M., Kyogoku, Y., Philipp, J. N., Richter, M., Rietdorf, C., and Yousef, T. (2020). Keyword extraction in german: Information-theory vs. deep learning. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1: NLPinAI*, pages 459–464. INSTICC, SciTePress.
- Kölbl, M., Kyogoku, Y., Philipp, J. N., Richter, M., Rietdorf, C., and Yousef, T. (2021). *The Semantic Level of Shannon Information: Are Highly Informative Words Good Keywords? A Study on German*, volume 939 of *Studies in Computational Intelligence (SCI)*, pages 139–161. Springer International Publishing.
- Levshina, N. (2017). Communicative efficiency and syntactic predictability: A cross-linguistic study based on the universal dependencies corpora. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies, 22 May, Gothenburg Sweden*, number 135, pages 72–78. Linköping University Electronic Press.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

- Levy, R. and Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19:849.
- Levy, R. P. (2018). Communicative efficiency, uniform information density, and the rational speech act theory. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, page 684–689.
- Melamed, I. D. (1997). Measuring semantic entropy. In *Tagging Text with Lexical Semantics: Why, What, and How?*
- Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.
- Richter, M., Kyogoku, Y., and Kölbl, M. (2019). Interaction of information content and frequency as predictors of verbs' lengths. In *International Conference on Business Information Systems*, pages 271–282. Springer.
- Richter, M. and Yousef, T. (2019). Predicting default and non-default aspectual coding: Impact and density of information feature. In *Proceedings of the 3rd Workshop on Natural Language for Artificial Intelligence co-located with the 18th International Conference of the Italian Association for Artificial Intelligence*.
- Richter, M. and Yousef, T. (2020). Information from topic contexts: the prediction of aspectual coding of verbs in russian. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.
- Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press.
- Zipf, G. K. (1949). Human behavior and the principle of least effort: an introd. to human ecology.
- Zipf, G. K. (2013). *The psycho-biology of language: An introduction to dynamic philology*. Routledge.