

Information theory unravels the subtext in Chekhov

J. Nathanael Philipp <nathanael_at_philipp_dot_land>, Sächsische Akademie der Wissenschaften zu Leipzig  <https://orcid.org/0000-0003-0577-7831>

Michael Richter <mprichter_at_gmail_dot_com>, Universität Leipzig  <https://orcid.org/0000-0001-7460-4139>

Olav Mueller-Reichau <reichau_at_uni-leipzig_dot_de>, Universität Leipzig

Matthias Irmer <irmer_at_conceptmining_dot_de>, Digital Science  <https://orcid.org/0000-0002-7719-9594>

Abstract

The present paper reports a pilot study to approach the subtext in Chekhov's short-story "Ward No. 6" by means of information theory. The original text is enriched by glosses by which we intend to make explicit the implicit knowledge conveyed by the original text, i.e., the subtext. We generated several text variants with meaningful enrichments and one fake variant that served as a baseline. We could not observe that semantic surprisal as a feature of words and uniform information density have a subtext effect throughout all text variants. However, it turned out that kurtosis and skewness are suitable classification criteria to distinguish meaningful enrichments from fake enrichments.

1. Introduction

This paper is a first attempt to approach the subtext by means of information theory [Shannon 1948]. The object of study is a text by the Russian writer Anton P. Chekhov, namely the short-story "Ward No. 6" ("Palata No. 6" in the original). 1

The message conveyed by any text is composed of two kinds of meaning. The first is the explicit meaning, i.e. the meaning that is explicitly coded by the linguistic material appearing black on white. The second is the meaning that a cooperative interpreter will infer from the explicit text following Gricean or Neo-Gricean reasoning [Grice 1989], making use of both linguistic knowledge and non-linguistic resources (world knowledge, conversational context). Even the most explicit text leaves open numerous informational gaps, which must be filled by the interpreter. These "bridging inferences" [Irmer 2011] constitute the implicit meaning of the text, for which we utilise the term subtext here. Another metaphor which is sometimes employed is the visible text as opposed to *dark text* (e.g. [Hinrichs 2014]). 2

The discourse about the subtext is rich (see [Lelis 2011], [Lelis 2013] for a survey), and the notion is controversial. According to the Oxford Dictionary of literary terms [Baldick 2015], the subtext is "any meaning or set of meanings which is implied rather than explicitly stated in a literary work, especially in a play." For the Russian Literary encyclopedia of terms and concepts [Nikoljukin 2003], the subtext ("podtekst" in Russian) represents the "hidden sense of an utterance, stemming from the interaction of the literary meanings, the context, and the speech situation." As can be seen, there are basically two ways of understanding what subtext is ([Ermatova 2010], p. 72.) Under the first reading, the one described above, advocated by [Baldick 2015], the term is basically synonymous to "pragmatic inferences." Besides that, the term subtext is used in the literature also to refer to the ultimate sense of a literary text, i.e. to the interpretation that is intended by the author and that the reader has to decipher (e.g. [Nikoljukin 2003], [Myrkin 1976]). In that second respect, the meaning of subtext comes close to the "moral of the story." In what follows, we use subtext only in the first of these two readings. 3

To approach the implicit meanings of a text, we will investigate the Russian short-story "Palata No. 6", written by Anton P. Chekhov, published in 1892. Chekhov is considered to have originated the role of the subtext, in his plays but also in his prose (e.g. [Gatrall 2003], [McSweeney 2004]). Using a narrative text for our purposes has the advantage that the informational impact of the conversational context is reduced because the addressee (reader) does not have to calculate her own position as well as the position of the speaker (author). 4

Related to the previous point, Chekhov is known for having developed a specific literary technique of "concentration and shortness" ([Kluge 1995], p. 48-49), manifesting itself in the language of the narrator, which usually shows simple syntax, parataxis, short sentences. This and the sparse use of adjectives, comparisons and metaphors is functional, for Chekhov's intention is to challenge the reader intellectually and to encourage critical reading ([Kluge 1995], p. 67) [Whyman 2010]. The reader is asked to actively interpret by filling the informational gaps of the texts. Therefore, we may expect a Chekhov story to entail a drastic discrepancy between what is said explicitly and what is meant by the text. If the subtext is "a level of speech between the lines" [Freise 1997], we may thus safely expect a lot of speech between the lines in the work of Chekhov. 5

The general language of the Chekhovian short-story is the Russian literary language ([Kluge 1995], p. 49), which is important for our methodology to be introduced below. The particular text "Ward No. 6" has been chosen, moreover, because of its considerable length compared to other short-stories of the author. For our methodology to work, we need a text of a certain cardinality of words.

2. Outline

In this paper, we presuppose a probabilistic notion of linguistic meaning. The basic assumption is that the meaning of a word influences the probability with which other words accompany that word in a text. Given a sufficiently large number of text passages in which a word appears, it is in principle possible to obtain a co-occurrence profile that is characteristic of the word. We consider this profile, the *co-text* of the word, to be a function of its meaning. ^[1]

Our point of departure is that information and the flow of information (FoI) are characterising features of texts. By *FoI* we mean the distribution of information per word over time in sentences. Our approach is inspired by [Dretske (1981)] who, in a Shannon information theoretic framework, was concerned with the relationship between knowledge and information in linguistic messages [Dretske (1981)] states that Shannon information (SI) forms a framework for conveying meaning and its evolvment: in his view, information is a key component of knowledge, and knowledge acquisition is a result of the way information is processed. In the present paper, we exploit this idea, likewise assuming that the flow of information is a classifying semantic feature of texts.

We use contextualised information, i.e., *surprisal* [Tribus 1961], [Hale 2001], [Levy 2008] as a lexical feature of a word *w*.

In order to approach the actual meaning of a word, in other words, we measure the effect that the meaning of the word has on its co-text.

This idea is at the heart of the Topic Context Model (TCM), see Section 3.2. TCM calculates the amount of *semantic surprisal* of a word, i.e., the amount of SI, respectively, that can be derived from a semantic context.

In the due course, we use the terms *information* and *information value* as equivalents of *surprisal* and *surprisal value*, respectively, and we speak of the flow of surprisal (FoS) and the flow of information (FoI).

Since the surprisal of a word *w* depends on the text in which it appears, it should make a difference whether the information carried by *w* is calculated only on the basis of those words that literally constitute the novel (bare narration), or whether it is calculated on the basis of the narration enriched by additional word material. We take methodological advantage of this fact in the following way.

In addition to the bare narration, we set up a contextualised narration, which consists of the original text words plus additional texts that contextualise the original text words by adding encyclopedic knowledge about them. These additional texts, which we also refer to as glosses, serve as a (very simple) model of the subtext of the novel. Details on how the text is enriched by additional knowledge are given in Section 4. In order to check whether the expected change in lexical information through contextualisation, i.e. through the enrichment of the text through meaningful glossing, is a random effect, the original text was also enriched with text fragments that have no relation to the Chekhovian story in terms of content, that is, a *fake glossing*. After that, in a first work step, we carry out two measurements: (i) we determine the surprisal values of the words in the bare narration, (ii) we measure them again, but this time both in the meaningful and fake glossed narration.

From FoS, we determine the information density in the text's paragraphs. Information density is another notion that we need to introduce to explain our methodology. The *Uniform Information Density* (UID) principle [Fenk and Fenk 1980], [Jaeger and Levy 2006] says that the sender of a message prefers to distribute surprisal evenly and smoothly across a sequence of linguistic units in a sentence, utterance or in a text. FoS should not be slowed down or even brought to a standstill by large jumps in information. This implies that the more the information in sentences or statements swings up or down, the higher the peaks and troughs of information, the more difficult it is for the recipient to process the information. Excessive information fluctuations can even prevent processing altogether.

Information density characterises linguistic communication. Therefore, if we assume that the subtext plays an essential role in Chekhov's prose (which we do, see above for motivation), we can expect that factoring in the subtext will lead to a more balanced FoI. In other words, we interpret density as the decisive discriminating feature in order to distinguish the narration enriched both by meaningful and fake subtexts from the bare narration.

The measurement of meaningfully enriched text should show less peaks and troughs than the one of the bare narration. From that our first prediction follows: in meaningfully enriched texts, the density will be optimised compared to the original text. In contrast, with fake enrichments, the density is predicted to deteriorate drastically.

H1 Adding meaningful glosses to the original text leads to a more balanced FoS.

H2 Adding fake glosses to the original text leads to a less balanced FoS.

What does that mean in terms of the principle of UID? In Section 3.3, we give the definition of UID that we follow in this paper: information

density gets more *balanced*, that is, better, when its value shifts towards zero. This represents on average small surprisal jumps from word to word and it represents uniformity. In contrast, when a UID value moves away from zero, it is a deterioration since this means that there are larger surprisal gaps between words. The dispersion of UID values should not be large. If this were the case, it would indicate arbitrariness in the UID distribution. We formulate two hypotheses: a meaningful subtext will reduce the scatter of UID values compared to the original text. Consequently, a fake subtext will lead to stronger scattering.

Models that employ information theory are rarely applied in digital humanities, and to the best of our knowledge, there are no studies that aim to disclose the subtext in the framework of information theory.

In the next sections, we introduce some fundamentals of information and surprisal theory, and we present the Topic Context Model. Subsequently, we describe our methodology and the different enrichment strategies (glossings) that we employed. Finally, we report the experimental results and discuss the outcomes of our study.

3. Information theoretic background

3.1 Information theory and information as a lexical feature

Shannon's information theory [Shannon 1948], [Shannon and Weaver 1949] models the transmission of information from a sender to a receiver. The information theoretic model aims to output a code in the form of consecutive 0 and 1 that compresses a message as much as possible without loss of information. Shannon developed his theory in an engineering context, and from the 1950s onwards the theory was applied to natural language.

Shannon Information (SI) is measured in bits. 1 bit is the information of two equally probable possibilities. For example, a toss with a fair coin carries 1 bit of information because there are two possible outcomes of a toss, i.e., "front of the coin," for instance represented by '1', or "back of the coin," represented by '0'. The amount of bits equals the number of yes/no decisions needed to determine results of processes such as tossing a coin or determining a word in a chain, given a good algorithm.

For the sake of illustration, imagine a toy language L that comprises solely eight words, that is, $L = \{Anton, Chekhov, was, a, very, talented, Russian, writer\}$. Let us agree that in L any sentence must consist of eight words, and there are no rules of linear precedence in a sentence, so that for instance both *Anton a was very Chekhov talented writer Russian* or *a was Chekhov Russian Anton talented writer very* are possible sentences with the same probability. For a sentence in the toy language it needs three yes/no questions to determine a word w uniquely as shown in Figure 1.

As can easily be seen, the word Chekhov as terminal element in Figure 1 carries three bits of information. We get the same result when Formula 1 is applied. This expression is the standard definition of SI, i.e., the negative log-likelihood of a linguistic unit w :

$$SI(w) = -\log_2 P(w)$$

(1)

Since the probability of a word w in L is $P(w) = \frac{1}{8}$, we get for instance for the word Anton in Figure 1 below:

$$SI(\text{Anton}) = -\log_2 P\left(\frac{1}{8}\right) = 3 \text{ bits}$$

(2)

The information type Surprisal [Hale 2001], [Levy 2008] is basically a psycholinguistic concept ^[2]:

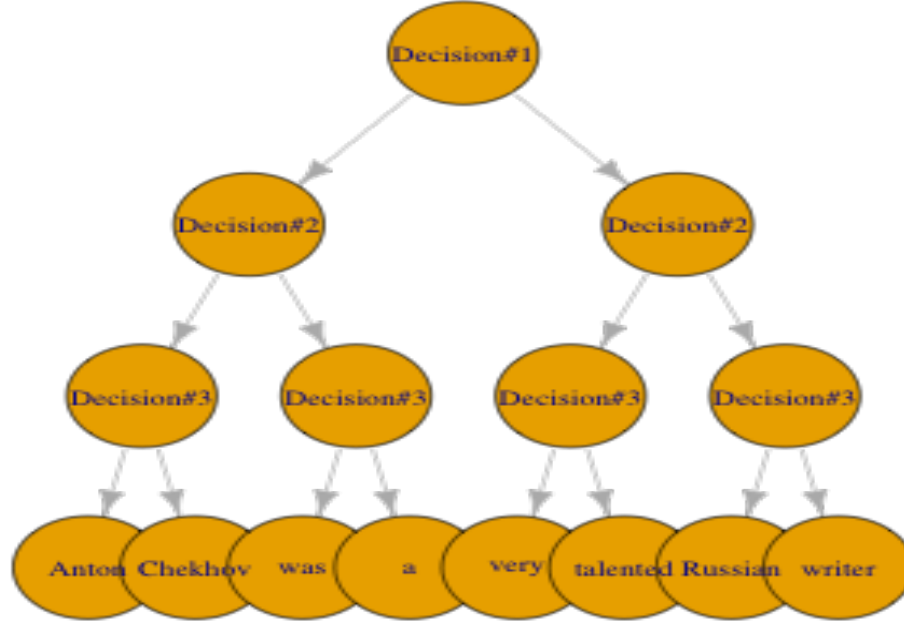


Figure 1. Information amount in bits of words in a sentence represented in a decision tree.

surprisal of a linguistic unit is proportional to the mental processing effort that it causes [Hale 2001]. If the probability of a word is high in a given context, its surprisal and the effort for the processing of that word are low. For instance, given the word *Barack* in a text, the probability is presumably quite high that Obama follows. In that case, Obama would carry low surprisal. An example of high surprisal because of the semantic unexpectedness of words is Chomsky's famous sentence *colorless green ideas sleep furiously*, with which he wanted to make clear that semantically uninterpretable sentences can nevertheless be syntactically well-formed [Chomsky 1957]. To give another example: the sentence final *verb* fell in the famous Garden Path-sentence *the horse raced past the barn fell* is extremely unexpected and carries high surprisal [Hale 2001]. Consequently, this requires a considerable processing. A language processor would probably assume that *barn* is the sentence final word given the preceding context. The expression in Formula 3 defines surprisal, as a special case of SI as in Formula 1, as negative log-likelihood of a linguistic unit. The basis is its conditional probability which is in Formula 3 represented by the vertical bar.

$$surprisal(w_i) = -\log_2 P(w_i | w_1, \dots, w_{i-1}, CONTEXT)$$

(3)

The variables $w_1, \dots, w_{i-1}$ represent co-occurrences of any kind of the target word w_i within a sentence. The variable *CONTEXT* represents extra-sentential contexts [Levy 2008].

Technically, contexts for the calculation of surprisal can be co-occurrences of target linguistic units, such as n-grams of any type (e.g. terminal symbols and of part-of-speech tags) [Horch and Reich 2016], syntactic structures [Celano et al. 2018], [Rubino et al. 2016], [Levshina 2017], [Richter et al. 2022], [Richter et al. 2019] and also semantic contexts [Kölbl et al. 2020], [Kölbl et al. 2021], [Philipp et al. 2022], [Philipp et al. 2023], [Venhuizen et al. 2019].

3.2 Topic Context Model

The Topic Context Model (TCM) [Kölbl et al. 2020], [Kölbl et al. 2021], [Philipp et al. 2022], [Philipp et al. 2023], [Philipp and Kölbl et al. 2023] [3]. calculates the semantic surprisal of words relative to an extra-sentential context in a given corpus and predicts that a word carries high surprisal when its mean probability, that is, the expected value, in the topics of a document is low. The prerequisite for calculating the information content of a word in a text is the knowledge of the topics that the text contains. In general terms, a *topic* is what

a text document is about. Depending on how coarse or fine-grained we look at a text, the text can deal with more or less topics. In this study we calculate the surprisal on a paragraph basis, meaning that we use the different paragraphs as the basis for calculating topics.

The topics that a paragraph covers are established by the process of topic detection. In this study, TCM relies on the topic detection model *Latent Dirichlet Allocation* (LDA) [Blei et al. 2003] which is a widely used, successful model, but other models can also be used in TCM such as *Latent Semantic Analysis*.

The assumption behind LDA is that (i) each text document contains a statistical mixture of topics, that (ii) similar topics correlate with similar probability distributions of words, and that consequently (iii) each topic is characterised by a specific distribution of the words. How many topics LDA should detect can be chosen before calculation by means of the topic detection starts.

Assume that LDA has disclosed two topics in a document, as illustrated in Figure 2:

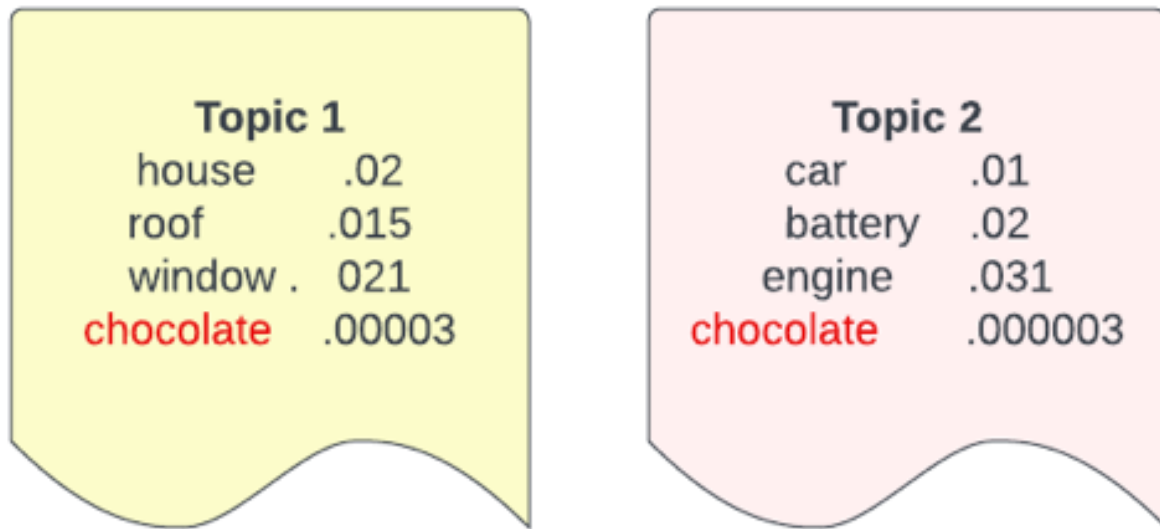


Figure 2. Topics as probability distributions of words.

In the yellow-coloured topic, high probabilities are assigned to words in the semantic field of, say, human created buildings. The pink-coloured topic contains highly probable words from the automobile sector. It is hardly surprising that a word from the nutrition such as chocolate has a low probability in both topics. TCM calculates the surprisal of chocolate per topic according to the Formulae 1 and 3 and then takes the mean value. In our example, in $P_{t1}(\text{chocolate}) = 0.00003$ and $\text{surprisal}(\text{chocolate}) = -\log_2 0.00003 = 15.02$ bits and $P_{t2}(\text{chocolate}) = 0.000003$ and $\text{surprisal}(\text{chocolate}) = -\log_2 0.000003 = 18.35$ bits, and the mean is 16.69, see Formula 4.

$$\text{surprisal}(\text{chocolate}) = \frac{-\log_2(0.00003) + (-\log_2(0.000003))}{2} = \frac{15.02 + 18.35}{2} = 16.69 \text{ bits}$$

(4)

The surprisal for a language processor who encounters the word *chocolate* in the context of the two topics *buildings* and *automobiles* is therefore quite high.

Figure 3 illustrates the working of the TCM.

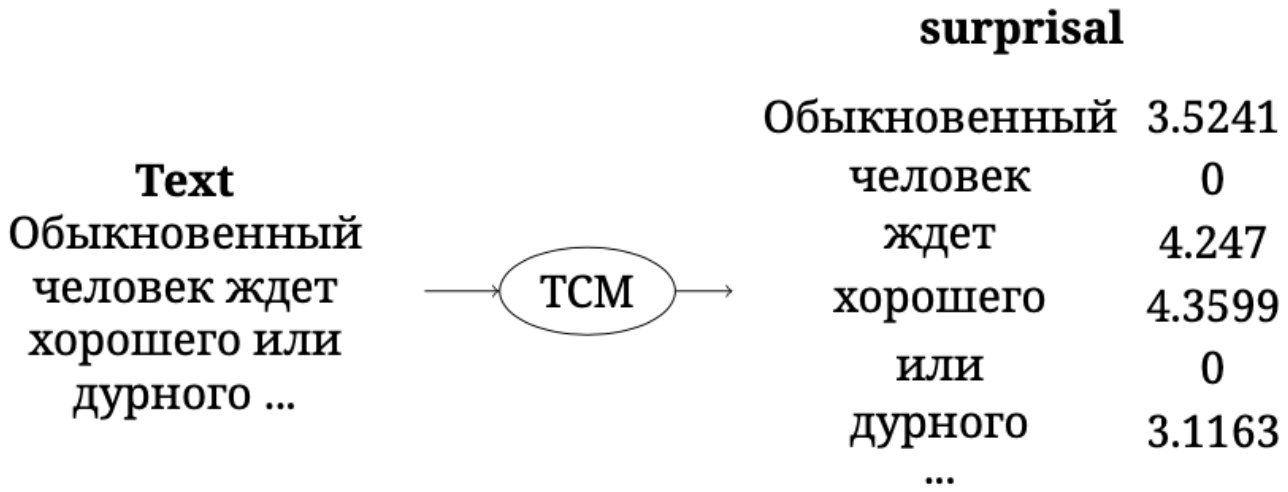


Figure 3. Illustration of the Topic Context Model.

3.3 Uniform information density

Uniform Information Density (UID) is an important principle in linguistic communication of humans [Levy and Jaeger 2007], [Jaeger 2010] but it is also important in artificial generation of language: in any linguistic product, information peaks and troughs must not be too extreme, so as not to make it too difficult for the recipient of a message to process it. In this study, we utilise the operationalisation of UID in [Collins 2014].^[5]

UID_{LOCAL} is the measure of the average (squared) information change from word to word in a sentence. Formula 5 gives its definition: id is the information / surprisal of a word, n represents the total number of words in a sentence. Instead of UID_{LOCAL} we use the term $UID_{wordwise}$ [Scheffler et al. 2023], [Philipp and Richter et al. 2024].

$$UID_{wordwise} = -\frac{1}{n-1} \sum_{i=2}^n (id_i - id_{i-1})^2$$

(5)

$UID_{wordwise}$ is negative by definition, and therefore a $UID_{wordwise}$ value close to zero indicates a high uniformity of the information density distribution, that is, on average smaller information jumps from word to word.

In the Garden Path-example from above information does not flow very smoothly as becomes clear in Figure 4 from [Hale 2001]. The information is distributed fairly evenly over the first five words. However, the jump in information in the sentence-final word *fell* is enormous:

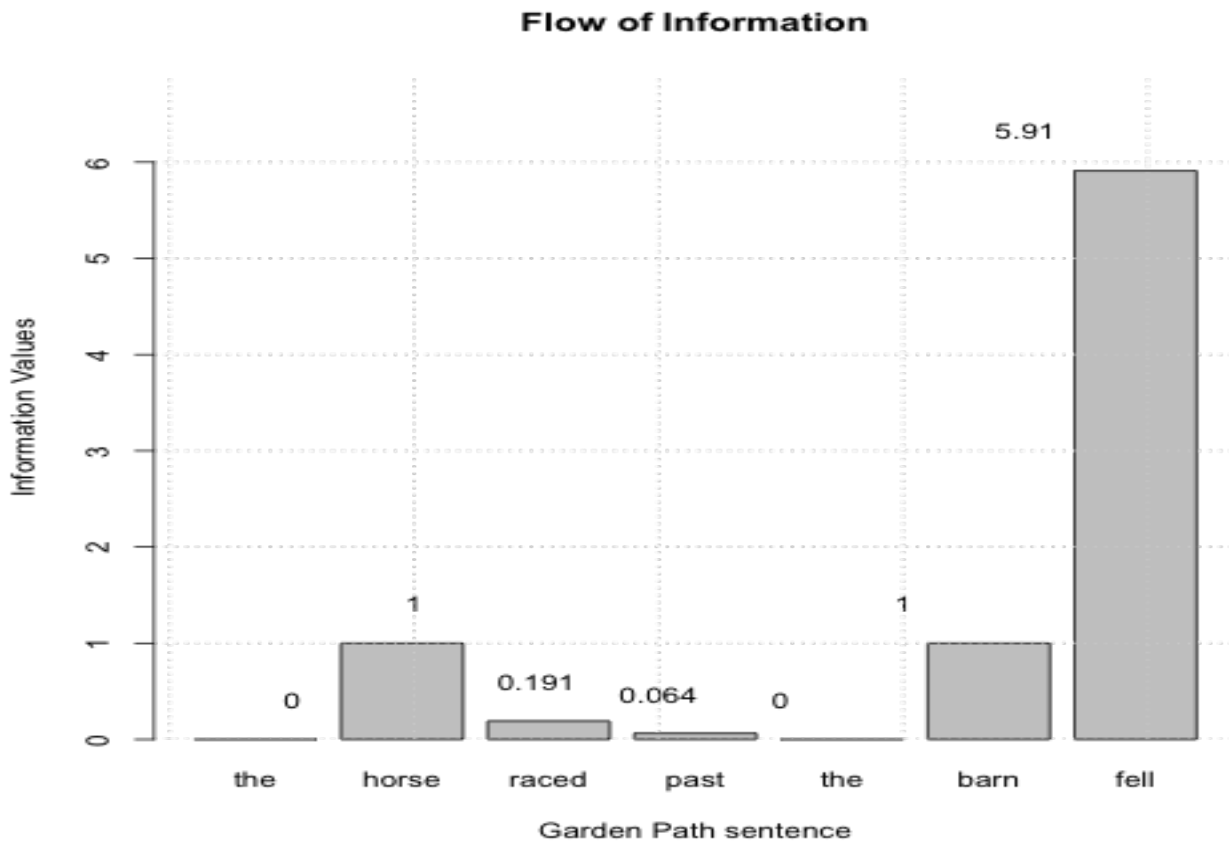


Figure 4. Flow of information in a Garden Path-sentence.

The steep sentence final information peak threatens the successful processing of the sentence.

The comparison of $UID_{wordwise}$ in the "expectable" sequence and in the garden path clearly shows that in the latter, the value of $UID_{wordwise}$ is further away from 0 (see (b) below) than in the expected sequence (see (a) below). This represents the lower uniformity of the information density in the garden path sentence. We use the information values from Figure 4 as input of expression 5:

(a)

$$UID(\text{the horse raced past the barn}) = \frac{-(0 - 1)^2 + (1 - .191)^2 + (.191 - .064)^2 + (.064 - 0)^2 + (0 - 1)^2}{5} = -0.53$$

(b)

$$UID(\text{the horse raced past the barn fell}) = \frac{-(0 - 1)^2 + (1 - .191)^2 + (.191 - .064)^2 + (.064 - 0)^2 + (0 - 1)^2 + (1 - 5.91)^2}{6} = -4.46$$

In the present study, we use UID as a measure of the "communicative naturalness" of the Chekhovian original text and its manipulated versions. As should have become clear, we hypothesise that the original text alone, i.e. without its subtext, performs worse with respect to this measure than the original text enriched by its subtext. In the next section we explain how we set up the enrichments.

4. Enrichment

As noted above, the general methodology is to perform three different measurements of surprisal of the text words in Ward No. 6. First, we measure the surprisal of the words based on the word material of the narration alone. In a second measurement (or set of measurements, see below), we enrich the narration by a text addendum which is semantically related to the original text words. Finally, we use an additional text which is not content-related to the original text in any obvious way. In the present section, we describe how we set up the meaningful text addendum, i.e. the meaningful glosses.

Recall that we conceive of the meaningful glosses as a toy model of the subtext of the narration. Accordingly, the enrichment shall make explicit the implicit knowledge conveyed by the text. To determine appropriate subtext explications, we make use of the idea brought

forward by *Frame Semantics* [Fillmore 1976] that lexical material in the form of content words evokes, when used in a text, conceptual frames that represent generalised conceptual knowledge associated with the respective content word [Busse 2012].

Specifically, we use *BabelNet* [Ferrucci et al. 2009], a multilingual semantic network compiled from a variety of sources such as *WordNet* and *Wikipedia*, as an encyclopedic resource providing the conceptual knowledge associated with lexical material. *BabelNet* provides a network of conceptual nodes (“senses”) interconnected by semantic relations of multiple types, e.g. hypernyms, hyperonyms, meronyms etc. The nodes contain translations to multiple languages and also glosses providing definitions or descriptions of the concepts.

We set up a pipeline that automatically enriches a given text by adding information associated with content words available from *BabelNet*. The pipeline is based on Apache UIMA [Ferrucci et al. 2009], uses open source software components and comprises the following steps:

1. **Part-of-speech (POS) Tagging:** In this step, POS-tags are assigned to all word forms using *DKPro HunPosTagger* [Halácsy et al. 2007].
2. **Lemmatisation:** using *DK Pro Language Tool Lemmatizer* [DKPro Core Component Reference], we lemmatised the word forms. Further, stopwords, punctuation and duplicate words within paragraphs are removed.
3. **Word Sense Disambiguation (WSD):** we employed *DKPro WSD* [DKPro Core Component Reference] in combination with an extension that enables the use of *BabelNet* [Codina 2018] to resolve ambiguous lemmas to their most probable sense. A number of disambiguation methods have been tried, with varying results.
 - **Lesk algorithm:** this method looks up *BabelNet* glosses and decides upon textual overlap of those in order to disambiguate.
 - **Graph connectivity:** this method creates a graph from *BabelNet* neighbours; achieves poor results with all *BabelNet* sense types, and better results with only hyper/hypo/mero/holonyms as neighbours and only *WordNet* senses.
4. **Enrichment Generation:** we used the following method to enrich the text:
 - **Glosses:** by adding *BabelNet* glosses (i.e. descriptions of concepts) to the original text. Addition of glosses is done in parentheses right after the content words triggering them.

Overall, we created a number of enrichment variations resulting from varying the following parameters:

- BabelNet lexicon
 - ALL: all available lexicons
 - WNTR: only WordNet translations
- WSD algorithm
 - LESK: overlap of glosses
 - GRAPH: graph connectivity
- Enrichment type
 - GLOSSES: BabelNet glosses in parentheses after triggering lemma

The data resource for the “fake” glosses was the Wikipedia based “rus_news_2020_1M” corpus (1M sentences) from the “Wortschatz Leipzig” corpora collection. ^[6]

The core concept of our methodology is to observe the difference in the surprisal for each text word before and after meaningful or fake enrichments. Therefore, we employed TCM to calculate the surprisal for every token in the original text and for the four enrichments of each type. Table 1 gives an overview of the used enrichments and the size of the corpora.

When running LDA, we use the lemmatised tokens and ignored punctuations, stopwords and tokens shorter than three letters. Defining the number of topics is essential for LDA and has been determined experimentally for TCM calculation so far. We therefore calculated the surprisal values three times using three different numbers of topics, see Section 3.2 for the calculations. We then averaged the three resulting surprisal values for the final surprisal value. Further, we normalised the surprisal values to have a consistent range of values among different experiments. Then, we calculated the information density distribution for every enrichment and the original text to observe how enrichments change the information density.

5. Results

Firstly, we tested whether the different text variants differ measurably in their information density.

	#Paragraphs	#Tokens	#Unique lemmas
Original Text	186	8398	3336
ALL LESK	186	41914	7417
WNTR LESK	187	31094	5874
ALL GRAPH	186	25853	5569
WINTR GRAPH	187	25531	5272
ALL LESK FAKE	187	41601	7594
ALL GRAPH FAKE	187	40868	7491
WNTR LESK FAKE	188	36146	7506
WNTR GRAPH FAKE	188	39253	7528

Table 1. Overview of the different enrichments used in the experiment.

To do this, we compared the meaningful values in the sets and used the non-parametric *Mann-Whitney*-test.^[7] The results of the *Mann-Whitney*-tests are given in Table 2. A significance level of $p \leq 0.05$ means that there are strong, non-random differences between the sets.

All mean differences of information density between the original text and the texts enriched with fake glossings are highly significant. In each case $p \approx 0$, the hypothesis of equality of means has thus to be rejected. That is to say, the information density in the original texts differs strongly from the density in fake glossed texts. Statistically speaking, original texts and fake glossed texts are therefore different sets.

Weaker pronounced is the difference between the information density of the original text and the meaningfully glossed text. Approximately half of the comparisons lack significance, and the comparison of the original text with meaningful glossing sourced from all lexicons and graph connectivity is highly non-significant $p = 0.30$.

Date	Mann-Whitney U	p-value
Meaningful glosses		
original - ALL GRAPH	16229.0	0.3028
original - ALL LESK	27039.0	≈ 0
original - WNTR GRAPH	19302.0	0.0665
original - WNTR LESK	21425.0	0.0001
Fake glosses		
original - WNTR LESK	34560.0	≈ 0
original - WNTR GRAPH	34545.0	≈ 0
original - ALL GRAPH	34395.0	≈ 0
original - ALL LESK	34284.0	≈ 0

Table 2. Mann-Whitney U test for mean differences in UIDs and p-values.

The following figures compare the mean values of the $UID_{wordwise}$ -values and their dispersion in original, meaningful and fake glossed texts. A mean tends towards 0 means that many of the sentences in a data set show little wordwise surprisal-changes which means that the information density is tendentially uniform. Recall that $UID_{wordwise}$ is negative by definition.

Figure 5 displays that for the All-GRAPH-method employing word sense disambiguation (WSD) via graph connectivity, meaningful glossing make the density of surprisal more uniform compared to the original text, albeit only a little bit (5b). Conversely, in fake glosses (5b) information density in sentences is much less uniform. Meaningful glossing reduces the dispersion of UID-values compared to the original text (5a), whereas fake glossing results in a substantially stronger dispersion of the UID-values (5b).

Considering only LESK disambiguation, the scenario alters clearly: meaningful glosses fail to make the information density more uniform, even worsening it while, however, decreasing the dispersion (5c). Fake glosses make information density less uniform, accompanied by

heightened dispersion (5d).

For the WNTR-glossed method with GRAPH connectivity, Figure 6 shows as well that meaningful glossing does not manage to make information distribution more uniform, rather the contrary is the case. However the dispersion is reduced (6a) which is an improvement. The same is the case with LESK disambiguation (6c). With the fake glosses, the information density gets less uniform while, in contrast to meaningful glossing, the dispersion of the density values increases (6b/6d).

64

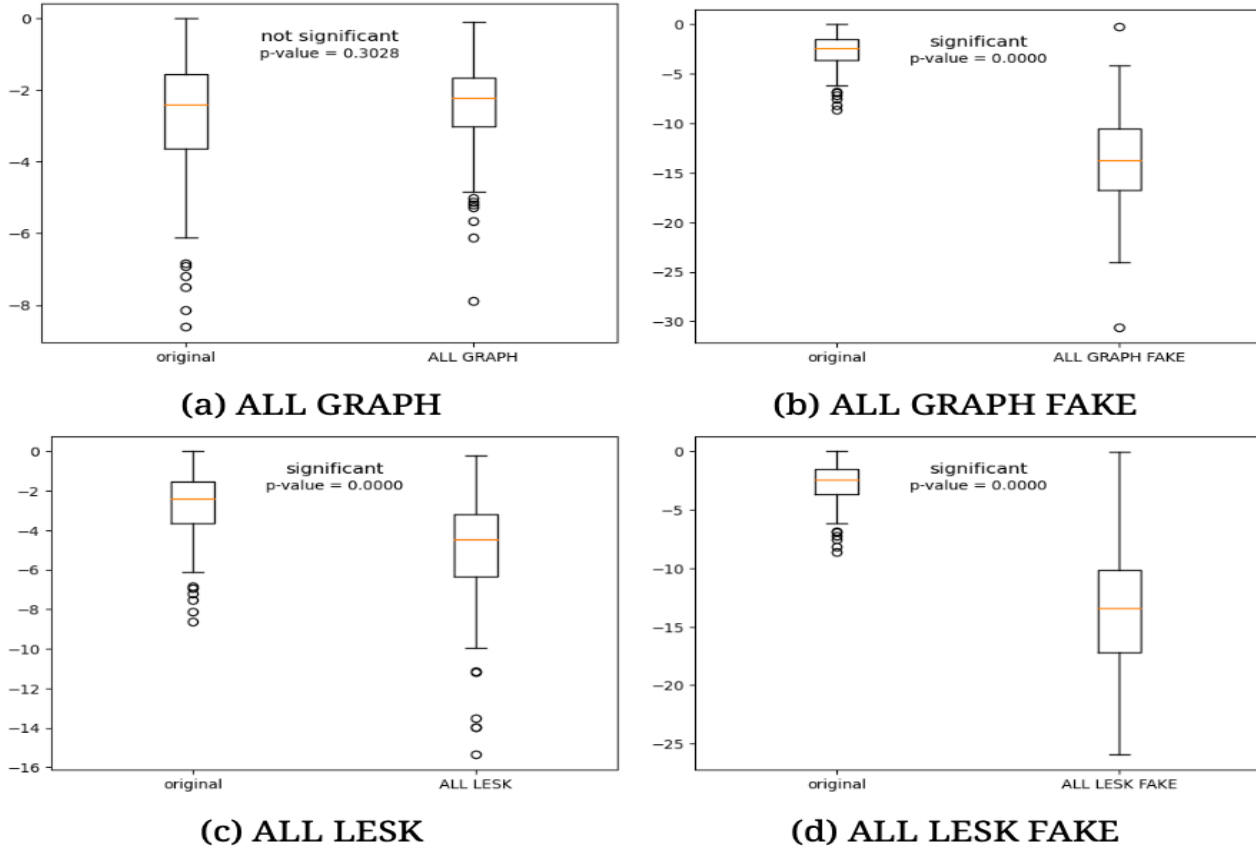


Figure 5. Comparison of means of original and meaningful / fake glossed texts based on UIDs for ALL-glossed method.

The plots in the Figures 7 and 8 depict probability distributions of UID values for the glossing types ALL GRAPH (ALL GRAPH method with WSD via graph connectivity), ALL LESK, WNTR GRAPH and WNTR LESK data. ^[8] The x-axis gives the $UID_{wordwise}$ values, the y-axis gives the probabilities.

65

Firstly, all plots show very clearly that fake glossed texts have distributions of UID-values reminiscent of a flattened normal distribution halfway, so to speak, to a uniform distribution: we see flattened peaks, and the edges of the distributions run wide to the left and right meaning the probability of extreme values is high. The distribution of the fake glossed text exhibits low kurtosis, i.e., peakiness.

66

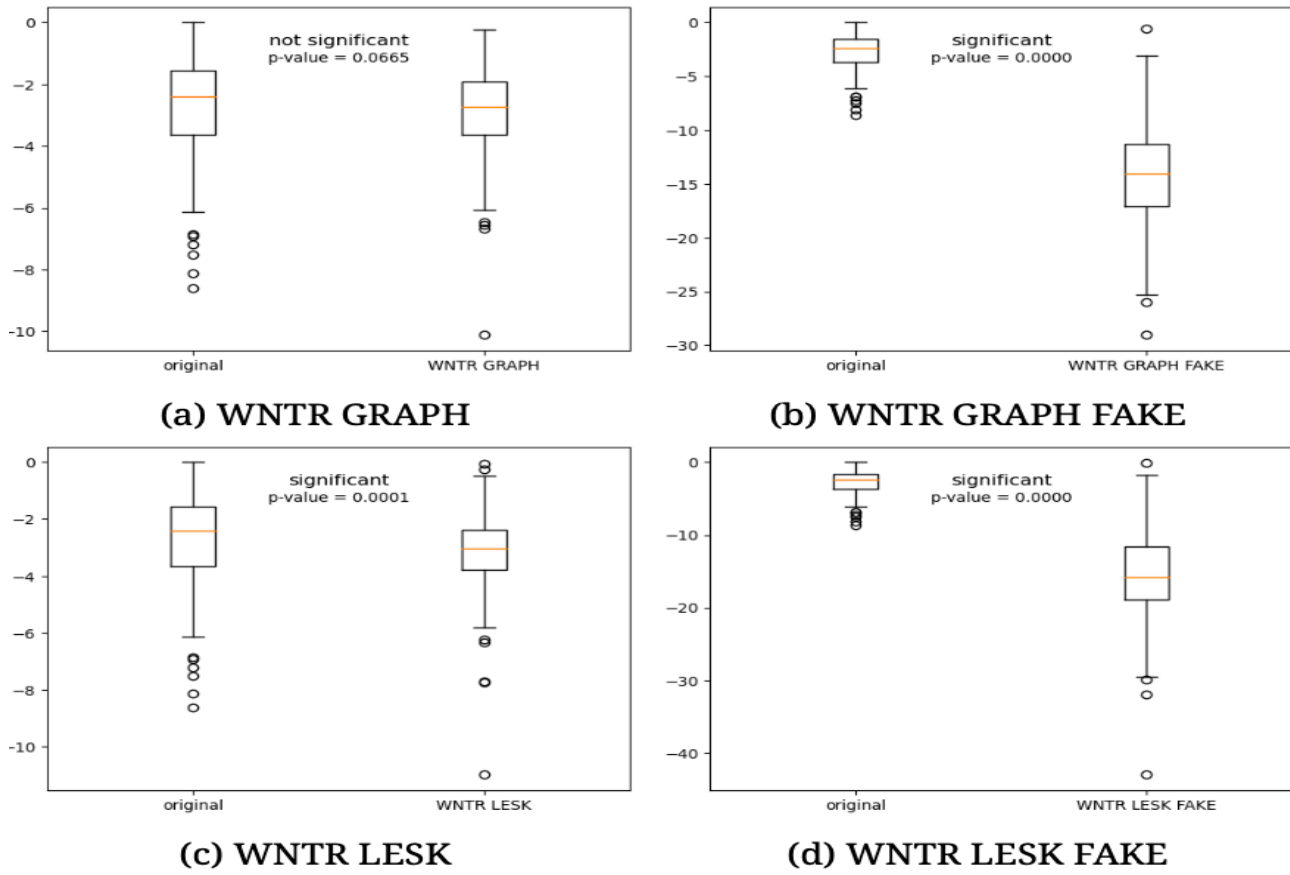
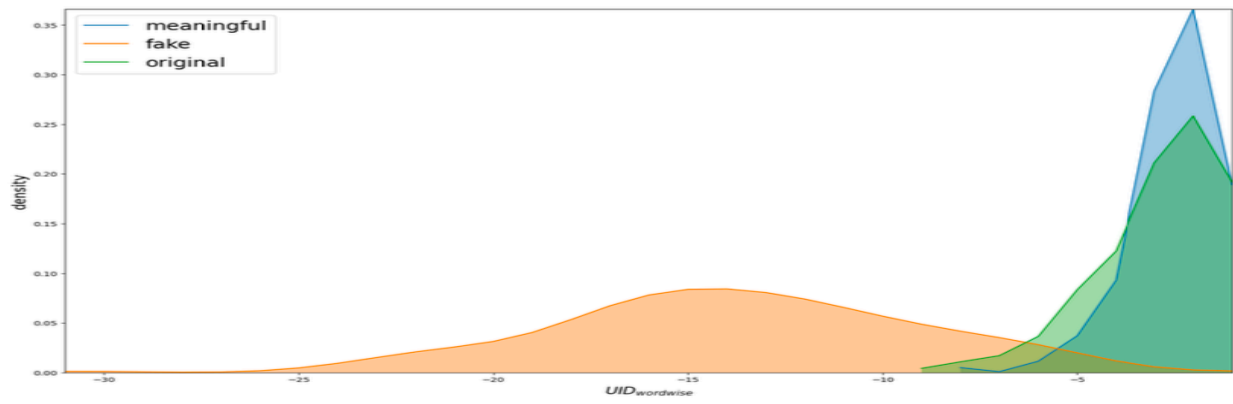


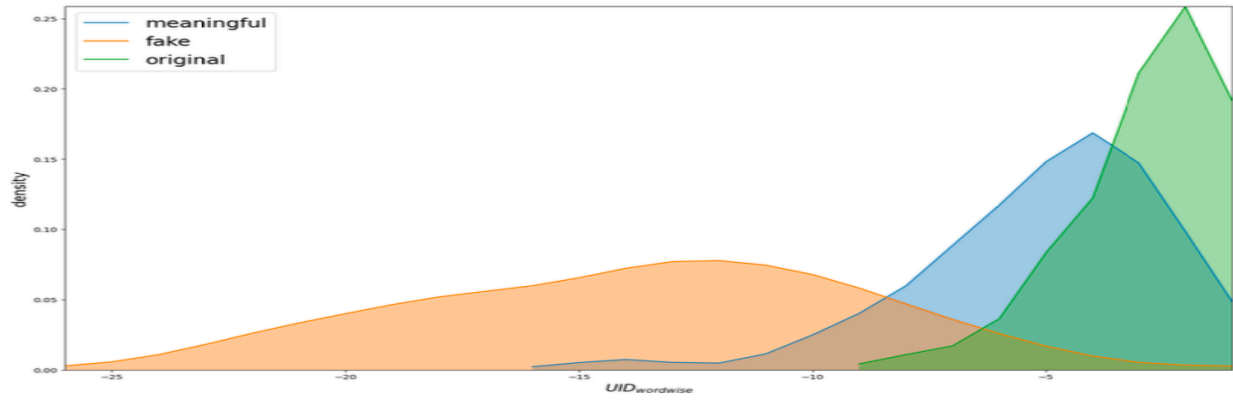
Figure 6. Comparison of means of original and meaningful / fake glossed texts based on UIDs for WNTR-glossed method.

[9]

Secondly, the plots in the Figures 7a and 8 visualise that meaningfully enriched texts exhibit higher *kurtosis* compared to the original text relative to fake enriched texts, that is to say, they have a higher, steeper peak. However, this does not seem to hold when LESK algorithm is employed as becomes visible in Figure 7b: here, the density distribution in the original text seems to be more peaky than in the meaningful enriched text, and the distribution in the latter runs out far to the left and right. Figure 7b seems to show that the generalisation about kurtosis, i.e., that a meaningful enriched texts has a more peaky density distribution than the original text, does not hold. We aimed to confirm this suspicion and chose an alternative graphical representation of the density distribution in the LESK text with a *Cullen and Frey-Graph* in Figure 9.



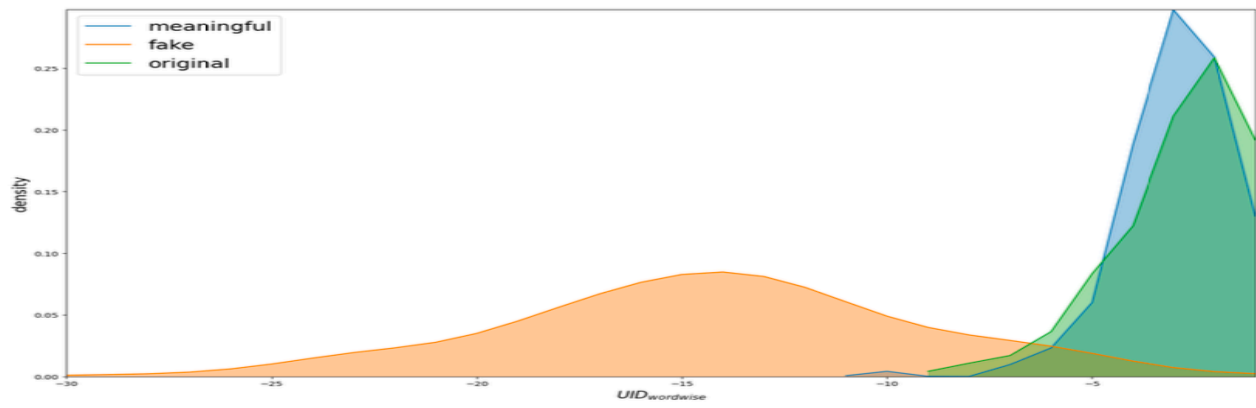
(a) ALL GRAPH



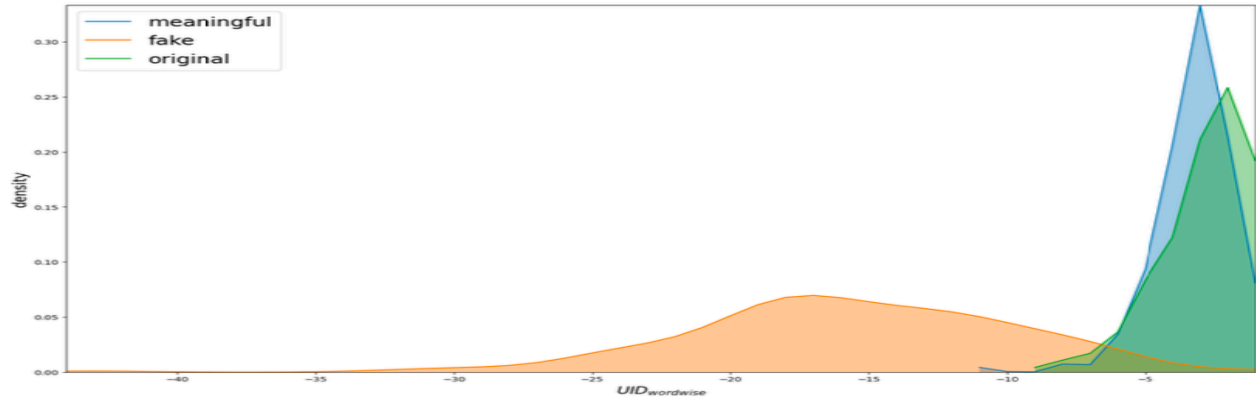
(b) ALL LESK

Figure 7. Information density of meaningfully/fake enriched texts and original texts.

It visualises the proximity and distance of LESK distributions with a set of theoretical distributions. The x-axis gives the degree of skewness of a distribution, and the y-axis gives the *kurtosis*.^[10] Figure 9 discloses that the ALL LESK density-distribution



(a) WNTR GRAPH



(b) WNTR LESK

Figure 8. Information density of meaningfully/fake enriched texts and original texts.

has higher kurtosis than the distribution in the original text, and the plot thus confirms the generalisation from above that the density distribution in meaningfully enriched texts is more peaky than the distribution in the original text. In addition, Figure 7b shows very clearly the proximity of the fake distribution to the uniform distribution that we observed already in the Figures 7a and 8. So, can we infer a subtext effect from these results? We will return to this research question posed above in the Section 6.

Cullen and Frey graph

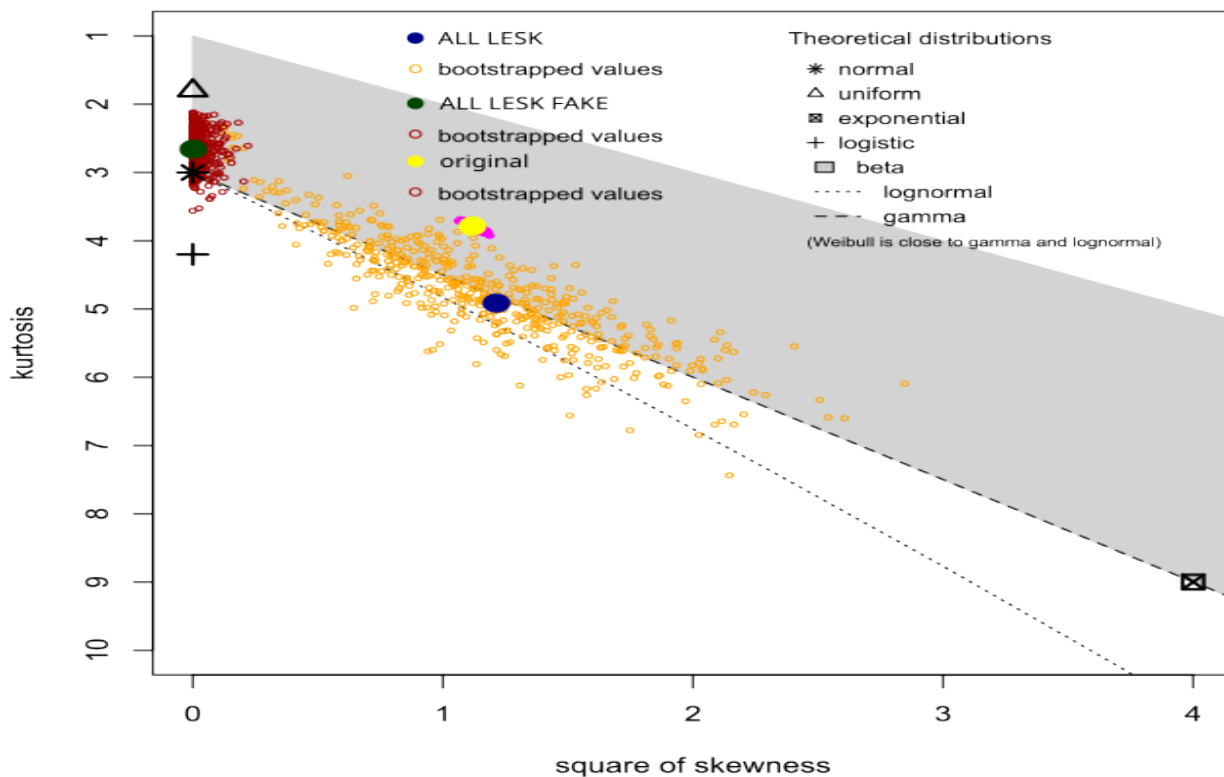


Figure 9. Cullen and Frey graph for ALL LESK GLOSSES with fake and meaningful enrichment's and original text.

6. Discussion and conclusion

The study confirms H2, and it confirms H1 solely for the enrichment combination ALL GRAPH (i.e. all available lexicons and the graphtheoretical disambiguation model), albeit the improvement is only marginal. Just for this case we observe that meaningful glossing improves the FoI, and that, in addition, the dispersion of density values decreases. Thus, with ALL GRAPH enrichments, the expected effect, albeit weak, occurs. In contrast, no such effect occurs with the rest of the meaningful enrichments: although the dispersion generally decreases (except for the LESK algorithm on all lexicons), the information density generally gets worse. So, we have to draw an overall negative conclusion: taking into calculation the otherwise invisible subtext, which we implemented by different versions of glosses semantically related to the original text, does not lead to a more uniform information density.

Our study also produced a very positive result, however, because we managed to determine a parameter that does correspond positively to the addition of meaningful glosses and negatively to the addition of fake glosses: peakiness/kurtosis. When fake enrichments are added to the original text, the distribution of UID-values flattens, which means that it approaches the contour of the uniform distribution (the uniform distribution results from a maximally fake text in which every word is chosen arbitrarily, see Figure 10). When meaningfully related glosses of the same cardinality are added to the original text, by contrast, the resulting contour shows a higher peak compared to the one of the original text.

The described finding comes out clearly in Figures 7a, 8a, and 8b. It is only with respect to enrichment type ALL LESK that the difference cannot be read from the respective plot, see 7b. However, the alternative visual representation in Figure 9 shows that also in this case meaningful enrichments lead to a higher kurtosis value whereas fake enrichments lead to a lower kurtosis value compared to the kurtosis value of the original text. This can be taken as a confirmation that the text features *surprisal* and *information density*, so to speak, measure semantic properties. We may confidently conclude that we attested a subtext effect.

Although we have shown that information theory may unravel the subtext in Chekhov, the observed subtext effects are subtle. This may be due to that the glosses that we used in the documented experiment are too primitive to reproduce any more clear subtext effect. Recall the basic idea underlying this study: if the subtext is an inalienable part of the semantics of a meaningful text, measuring the surprisal values of original texts alone will produce values that reflect the actual content of the text only incompletely. To measure the true surprisal of a text, one would have to enrich the original text by the perfect explication of the subtext.

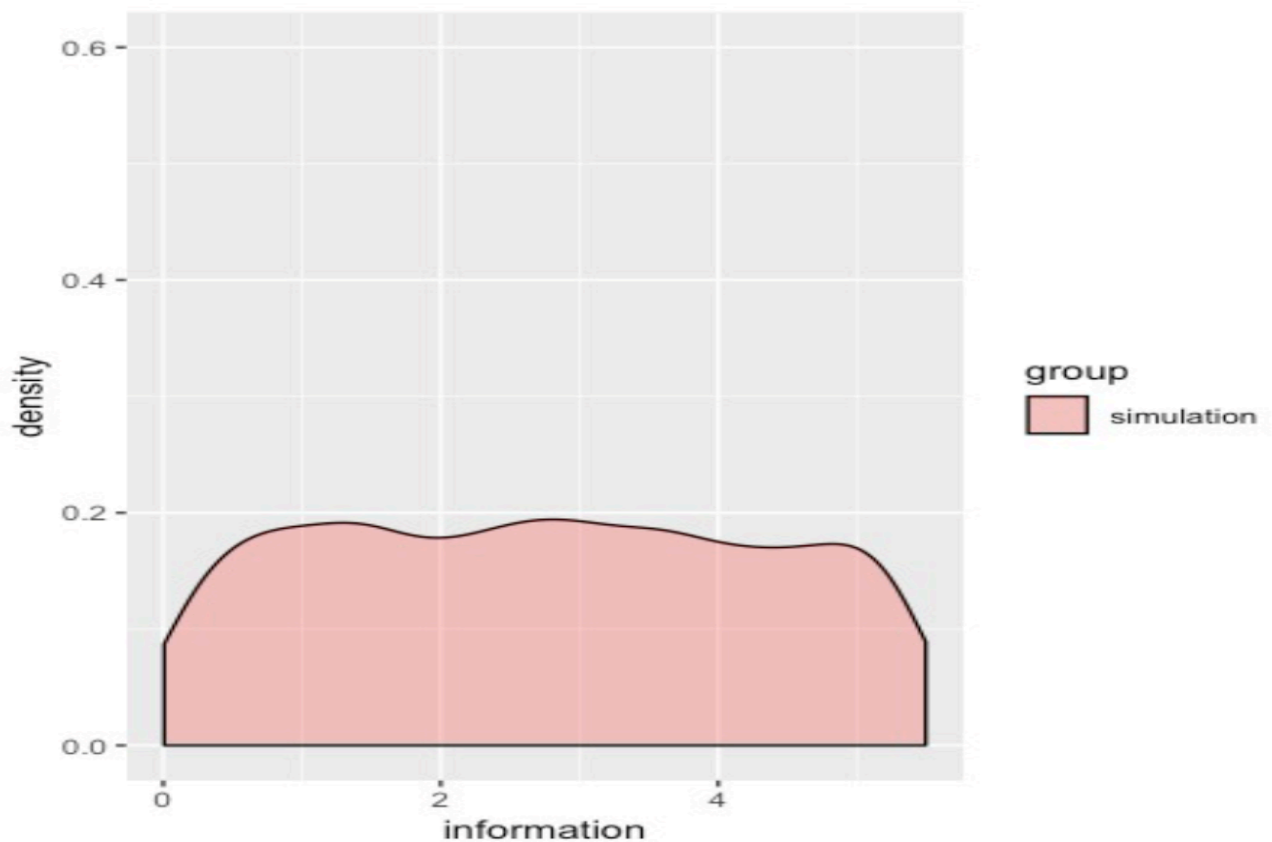


Figure 10. Uniform Distribution generated from 10,000 random numbers.

The (sub)textual prostheses that we actually used in our experiments are obviously very primitive enrichments, far from being perfect. Future research will therefore have to focus on the technique of enrichment.

While in this study, the generation of the subtexts was based in principle on the Fillmore model, future experiments will be carried out statistically based by employing neural networks. Doing so while keeping the features *information density* and *kurtosis/skewness*, we hope to achieve and observe a more significant subtext effect than in the current study.

72

Notes

[1] This idea is taken from the Distributional Semantics. Its hypothesis is that words with similar meanings tend to appear in similar contexts which enables to measure their semantic relationships, see for instance [Firth 1957], [Harris 1954], [Rubenstein and Goodenough 1965], [Landauer and Dumais 1997], [Turney and Pantel 2010], [Mikolov et al. 2013].

[2] Empirical evidence comes amongst others from [DeLong et al. 2005], [Bentum 2021].

[3] Available at: <https://github.com/jnphilipp/tcm>

[4] The above explanations and the calculation 4 are, of course, a simplified description TCM. For more details, see [Kölbl et al. 2020], [Kölbl et al. 2021], [Philipp et al. 2022], [Philipp et al. 2023], [Philipp and Kölbl et al. 2023].

[5] Available at: <https://github.com/jnphilipp/uid>

[6] Available at: <https://wortschatz.uni-leipzig.de/de>

[7] *Non-parametric* means that the test makes no or just minimal assumptions about the distribution within the data. An example may illustrate the question posed by statistical tests of mean-differences: the Dutch and the Danes are among the tallest populations in the world. On average, there is hardly any difference between them. In terms of height, the Dutch and Danes belong to the same group, that is, the group of the tallest people. In Malta, for example, the average height of the population is much lower. The difference in average height between the Dutch and the Maltese is likely to be highly significant. This means that the Dutch and the Maltese do not belong to the same height group.

[8] The area under the curves must be approximately 1 in each case.

[9] Available at: <https://en.wikipedia.org/wiki/Kurtosis>

[10] Available at: <https://en.wikipedia.org/wiki/Skewness>

Works Cited

Baldick 2015 Baldick, C. (2015) *The Oxford dictionary of literary terms*. Oxford: Oxford University Press.

- Baroni et al. 2014** Baroni, M., Dinu, G., and Kruszewski, G. (2014) "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1, pp. 238-247.
- Bentum 2021** Bentum, M. (2021) "Listening with great expectations: A study of predictive natural speech processing. Ph.D. thesis", *Radbound Repository*.
- Blei et al. 2003** Bleu, D.M., Ng, A.Y., and Jordan, M.I. (2003) "Latent dirichlet allocation", *Journal of machine Learning research*, 3(Jan), pp. 993-1022.
- Busse 2012** Busse, D. (2012) *Frame-Semantik: Ein Kompendium*. Berlin: deGruyter.
- Celano et al. 2018** Celano, G.G., Richter, M., Voll, R., and Heyer, G. (2018) "Aspect coding asymmetries of verbs: The case of Russian", in *PROCEEDINGS of the 14th Conference on Natural Language Processing*, pp. 34-39.
- Chomsky 1957** Chomsky, N. (1957) "Syntactic structures". The Hague: Mouton and Co.
- Codina 2018** Codina, J. (2018) OpenMinted BableNet (2018) Available at: https://github.com/TalnUPF/OpenMinted_BabelNet. (Accessed 01 September 2022).
- Collins 2014** Collins, M.X. (2014) "Information density and dependency length as complementary cognitive models", *Journal of psycholinguistic research*, 43(5), pp. 651-681.
- DKPro Core Component Reference** DKPro Core Component Reference (2017) Available at: <https://dkpro.github.io/>. (Accessed 01 September 2022).
- Delong et al. 2005** Delong, K.A., Urbach, T.P., and Kutas, M. (2005) "Probabilistic word preactivation during language comprehension inferred from electrical brain activity", *Nature neuroscience*, 8(8), pp. 1117-1121.
- Dretske (1981)** Dretske, F. (1981) *Knowledge and the Flow of Information*. Cambridge: MIT Press.
- Ermatova 2010** Ermatova, E.V. (2010) "Implicitnost' v chudožestvennom tekste", *Izdatel'stvo Saratovskogo universiteta*.
- Fenk and Fenk 1980** Fenk, A. and Fenk, G. (1980) "Konstanz im Kurzzeitgedächtnis-Konstanz im sprachlichen Informationsfluß", *Zeitschrift für experimentelle und angewandte Psychologie*, 27(3), pp. 400-414.
- Ferrucci et al. 2009** Ferrucci, D., Lally, A. Verspoor, K., and Nyberg, E. (2009) "Unstructured information management architecture (UIMA) version 1.0", *OASIS Standard*, March 2009. Available at: <https://docs.oasis-open.org/uima/v1.0/uima-v1.0.html>.
- Fillmore 1976** Fillmore, C. (1976) "Frame semantics and the nature of language", in *Annals of the NY Academy of Sciences: Conf. on the Origin and Development of Language and Speech*, 280, pp. 20-32.
- Firth 1957** Firth, R. (1957) *Two papers on linguistic theory: Studies in linguistic analysis (Special volume of the philological society)*. Oxford: Blackwell.
- Freise 1997** Freise, M. (1997) *Die Prosa Anton Čechovs: eine Untersuchung im Ausgang von Einzelanalysen*. Amsterdam: Rodopi.
- Gatrall 2003** Gatrall, J.J. (2003) "The paradox of melancholy insight: Reading the medical subtext in Chekhov's "a boring story"", *Slavic Review*, 62(2), pp. 258-277.
- Grice 1989** Grice, P. (1989) *Studies in the way of words*. Cambridge: Harvard University Press.
- Hale 2001** Hale, J. (2001) "A probabilistic Earley 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, Association for Computational Linguistics*, pp. 1-8.
- Halácsy et al. 2007** Halácsy, P., Kornai, A., and Oravecz, C. (2007) "HunPos- an open source trigram tagger", in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague: Czech Republic, Association for Computational Linguistics, pp. 209-217. Available at: <https://aclanthology.org/P07-2053>.
- Harris 1954** Harris, Z.S. (1954) "Distributional structure", *Word*, 10(2-3), pp. 146-162.
- Hinrichs 2014** Hinrichs, U. (2014) *Die dunkle Materie des Wissens: Über Leerstellen wissenschaftlicher Erkenntnis*. Giessen, Germany: Psychosozial-Verlag.
- Horch and Reich 2016** Horch, E. and Reich, I. (2016) "On "article omission" in German and the "uniform information density hypothesis"", in *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016): Bochumer Linguistische Arbeitsberichte*, 16, pp. 125-127.
- Irmer 2011** Irmer, M. (2011) *Bridging inferences*. Boston: de Gruyter.
- Jaeger 2010** Jaeger, T.F. (2010) "Redundancy and reduction: Speakers manage syntactic information density", *Cognitive psychology*, 61(1), pp. 23-62. Available at: <https://doi.org/10.1016/j.cogpsych.2010.02.002>
- Jaeger and Levy 2006** Jaeger, T. and Levy, R. (2006) "Speakers optimize information density through syntactic reduction", *Advances in neural information processing systems*, 19.
- Kluge 1995** Kluge, R.D. (1995) *Anton P. Čechov: eine Einführung in Leben und Werk*. Darmstadt: Wissenschaftliche Buchgesellschaft.

- Kölbl et al. 2020** Kölbl, M., Kyogoku, Y., Philipp, J., 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 INSTICC*. Setúbal, Portugal: SciTePress, pp. 459-464. Available at: <https://doi.org/10.5220/0009374704590464>.
- Kölbl et al. 2021** 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", *Vol. 939 of Studies in Computational Intelligence (SCI)*. London: Springer International Publishing, pp. 139-161. Available at: https://doi.org/10.1007/978-3-030-63787-3_5.
- Landauer and Dumais 1997** Landauer, T.K. and Dumais, S.T. (1997) "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge", *Psychological Review*, 104(2), pp. 211-240.
- Lelis 2011** Lelis, E.I. (2011) "Podtekst i smežnye javljenja", *Istorija i filologija*, 4, pp. 143-151.
- Lelis 2013** Lelis, E.I. (2013) *Podtekst kak lingvopoétičes kajakategorija v proze*. Iževsk, Russia: A.P. Čechova.
- Levshina 2017** 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* Gothenburg, Sweden: Linköping University Electronic Press, pp. 72-78.
- Levy 2008** Levy, R. (2008) "Expectation-based syntactic comprehension", *Cognition*, 106(3), pp. 1126-1177. <https://doi.org/10.1016/j.cognition.2007.05.06>.
- Levy and Jaeger 2007** Levy, R. and Jaeger, T. (2007) "Speakers optimize information density through syntactic reduction", *Advances in neural information processing systems*, 19, pp. 849-856.
- McSweeney 2004** McSweeney, K. (2004) "Effects or subtexts?", *Modern Language Studies*, 34(1-2), pp. 42-51.
- Mikolov et al. 2013** Mikolov, T., Chen, K., Corrado, J., and Dean, J. (2013) "Efficient estimation of word representations in vector space", *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Myrkin 1976** Myrkin, V.J. (1976) "Tekst, podtekst i kontekst", *Voprosy jazykoznanija*, 2.
- Navigli and Ponzetto 2012** Navigli, R. and Ponzetto, S.P. (2012) "Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network", *Artificial Intelligence*, 193(0), pp. 217-250. Available at: <https://www.sciencedirect.com/science/article/pii/S0004370212000793>.
- Nikoljukin 2003** Nikoljukin, A.N. (2003) *Literaturnaja énciklopedija terminov iponjatij*. Moskva, Russia: NPK "Intelvak".
- Philipp and Kölbl et al. 2023** Philipp, J.N., Kölbl, M., Daas, E., Kyogoku, Y., and Richter, M. (2023) "Perplexed by idioms?", in *Knowledge Graphs: Semantics, Machine Learning, and Languages*. Amsterdam: IOS Press, pp. 70-76.
- Philipp and Richter et al. 2024** Philipp, J.N.P., Richter, M., Scheffler, T., and van Hout, R. (2024) "The role of information in modeling German intensifiers", in Robin Lemke, Lisa Schäfer, Ingo Reich (eds.), *Information structure and information theory*, Berlin: Language Science Press, pp. 117-145. Available at: <https://zenodo.org/records/12784266>.
- Philipp et al. 2022** Philipp, J.N., Kölbl, M., Kyogoku, Y., Yousef, T., and Richter, M. (2022) "One step beyond: Keyword extraction in German utilising surprisal from topic contexts", in K. Arai (ed.) *Intelligent Computing*. London: Springer International Publishing, pp. 774-786. Available at: https://doi.org/10.1007/978-3-031-10464-0_53.
- Philipp et al. 2023** Philipp, J.N., Richter, M., Daas, E., and Kölbl, M. (2023) "Are idioms surprising?", in M. Georges, A. Herygers, A. Friedrich, B. Roth (eds.) *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, Association for Computational Linguistics, pp. 149-154. Available at: <https://aclanthology.org/2023.konvens-main.15>.
- Richter et al. 2019** Richter, M., Kyogoku, Y., and Kölbl, M. (2019) "Estimation of average information content: Comparison of impact of contexts", in *Proceedings of SAI Intelligent Systems Conference*. London: Springer, pp. 1251-1257. Available at: https://doi.org/10.1007/978-3-030-29513-4_91.
- Richter et al. 2022** Richter, M., Farré, M.B., Kölbl, M., Kyogoku, Y., Philipp, J.N., Yousef, T., Heyer, G., and Himmelmann, N.P. (2022) "Uniform density in linguistic information derived from dependency structures", in *Proceedings of the 14th International Conference on Agents and Artificial Intelligence - Volume 1: NLPinAI INSTICC*. Setúbal, Portugal: SciTePress, pp. 496-503. Available at: <https://doi.org/10.5220/0000155600003116>.
- Rubenstein and Goodenough 1965** Rubenstein, H. and Goodenough, J.B. (1965) "Contextual correlates of synonymy", *Communications of the ACM*, 8(10), pp. 627-633.
- Rubino et al. 2016** Rubino, R., Lapshinova-Koltunski, E., and Van Genabith, J. (2016) "Information density and quality estimation features as translationese in dicators for human translation classification", in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pp. 960-970.
- Scheffler et al. 2023** Scheffler, T., Richter, M., and van Hout, R. (2023) "Tracing and classifying German intensifiers via information theory", *Language Sciences*, 96, 101535.
- Shannon 1948** Shannon, C.E. (1948) "A mathematical theory of communication", *The Bell system technical journal* 27(3), pp. 379-423.
- Shannon and Weaver 1949** Shannon, C.E. and Weaver, W. (1949) *The mathematical theory of communication*. Champaign: University of Illinois Press.
- Tribus 1961** Tribus, M. (1961) "Information theory as the basis for thermostatics and thermodynamics", *Journal of Applied Mechanics*, 28(1),

pp. 1-8.

Turney and Pantel 2010 Turney, P.D. and Pantel, P. (2010) "From frequency to meaning: Vector space models of semantics", *Journal of Artificial Intelligence Research*, 37, pp. 141-188.

Venhuizen et al. 2019 Venhuizen, N.J., Crocker, M.W., and Brouwer, H. (2019) "Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience", *Discourse Processes*, 56(3), pp. 229-255.

Whyman 2010 R. Whyman, R. (2010) *Anton Chekhov*. London: Routledge.



This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.