Can information theory unravel the subtext in a Chekhovian short story?

J. Nathanael Philipp
Sächsische Akademie der Wissenschaften zu Leipzig
Karl-Tauchnitz-Str. 1
04107 Leipzig, Germany
nathanael@philipp.land

Olav Mueller-Reichau
Leipzig University
Beethovenstraße 15
04107 Leipzig, Germany
reichau@uni-leipzig.de

Matthias Irmer
Digital Science & Research Solutions
6 Briset Street, Farringdon
London, EC1M 5NR, UK
irmer@conceptmining.de

Michael Richter
Leipzig University
Augustusplatz 10
04109 Leipzig, Germany
mprrichter@gmail.com

Max Kölbl
Osaka University
1-5 Yamadaoka, Suita
565-0871 Osaka, Japan
max.w.koelbl@gmail.com

Abstract

In this study, we investigate whether information-theoretic measures such as surprisal can quantify the elusive notion of subtext in a Chekhovian short story. Specifically, we conduct a series of experiments for which we enrich the original text once with (different types of) meaningful glosses and once with fake glosses. For the different texts thus created, we calculate the surprisal values using two methods: using either a bag-of-words model or a large language model. We observe enrichment effects depending on the method, but no interpretable subtext effect.

1 Introduction

The meaning conveyed by any text has two layers: (i) explicit content encoded in linguistic form and (ii) an implicit layer inferred through Gricean reasoning (Grice, 1989), world knowledge and context (Irmer, 2011). To the best of our knowledge there is no method to measure the meaning of the implicit layer. In this study, we present such a method by trying to measure the effect of the implicit text (subtext) in Anton Chekhov's story *Ward No. 6* (*Palata No. 6* in the original) using information-theoretic measures (Shannon, 1948).[1]

To identify the subtextual structure, we enrich the Russian original with *glosses*, thus modelling implicit content explicitly. Some of these glosses are meaningful and explicate background knowledge; others are content-unrelated "fake glosses".

Our analysis relies on two information theoretic concepts: (i) contextualised information of words, that is, *surprisal* (Hale, 2001) and (ii) *Uniform Information Density (UID)* to capture differences in the *flow of information (FoI)*. FoI is made measurable by surprisal and UID: smooth information flow supports easier processing, while abrupt fluctuations hinder it (Fenk and Fenk, 1980; Jaeger, 2010). The UID principle posits that speakers tend to balance information distribution in messages to optimise comprehension.

For the calculation of wordwise surprisal in sentences, we use two models: the Large Language Model (LLM) **Llama 3.2-3B** (= M1) from Meta-Platforms as a computational engine for word-by-word text generation that calculates surprisal from an incrementally growing context, and the Topic Context Model (TCM) (= M2) (Kölbl et al., 2021; Philipp et al., 2022, 2023), which is an extended bag of word-topic model that calculates surprisal from words-topic probability distributions. We will test the hypothesis **H1** and its alternative hypothesis **H2**:

H1: Adding meaningful glosses reduces surprisal and leads to a well-balanced flow of information (UID values close to 0).

H2: Adding fake glosses leads to greater surprisal fluctuations (UID values diverge from 0).

To test this, we compute surprisal values for three text versions: the original, one with

[1] Subtext has been defined as both pragmatic inference (Baldick, 2015) and as a deeper authorial meaning (Nikoljukin, 2003; Myrkin, 1976). Chekhov's style, marked by brevity and a minimalist use of figurative language (Whyman, 2010; Kluge, 1995), invites an interpretive effort, making his prose ideal for subtext analysis (Lelis, 2016).

Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), pages 84–90

meaningful glosses, and one with fake glosses. UID serves as a diagnostic metric to determine whether glossing brings the text's Flow of Information closer to or further from optimal processing conditions.[2]

2 Information

2.1 Information indices

Shannon's information theory (Shannon, 1948; Shannon and Weaver, 1949) models the transmission of information from a sender to a receiver. Surprisal (Hale, 2001; Levy, 2008) builds on Shannon's theory and is contextualised Shannon information, linking information to cognitive processing effort. Surprisal s of a word w depends on its conditional probability in a given context:

$$\mathrm{s}(w_i) = -\log_2 P(w_i \mid w_1, \ldots, w_{i-1}, \mathrm{CONTEXT}) \quad (1)$$

In Equation 1, $w_{<i}$ represents co-occurrences, and $CONTEXT$ extra-sentential context that, in this study, is defined as semantic topics, from which semantic surprisal is derived.

For **M1**, we employ **Llama 3.2-3B**. Text is first segmented into *AI-tokens* i.e. character sequences whose length ranges from single characters to entire words, but never extend across word boundaries. Then, Llama tries to predict each token with respect to the previous text. In this way, we get probability values for every token, which we extend to probabilities of entire words by multiplying them.

As **M2**, the Topic Context Model (TCM) (Kölbl. et al., 2020; Kölbl et al., 2021; Philipp et al., 2022, 2023) is used:[3] TCM estimates the surprisal of a word from its probabilities in topics in a document, a paragraph, or even a single sentence. In order to detect topics in a text, TCM needs a topic detection model. We use Latent Dirichlet Allocation (LDA) (Blei et al., 2003). This generative model assigns probability distributions to topics in a document and to words within topics.

2.2 Uniform information density

The principle of *Uniform Information Density* is initially put by Fenk and Fenk (1980, p. 402): *In an effective and economical communication system, the information transmitted should be distributed as uniformly as possible across small time spans, and the average level of information transmitted per time should not exceed capacity limits.*

The UID principle describes a smoothing mechanism in linguistic messages that serves to reduce processing effort while enhancing communicative efficiency (Levy and Jaeger, 2007; Jaeger, 2010). Models of UID disclose (Meister et al., 2021) (i) a superlinear relationship between surprisal and processing effort since processing effort does not increase linearly with surprisal, sharp peaks in information load become disproportionately costly, and a more uniform distribution softens this effect; (ii) a tendency toward regression to the mean in information flow, implying that UID promotes convergence toward an average surprisal value (for instance on sentence-, text- or corpus-level); and (iii) the local smoothing of the Flow of Information in sentences. In this study, we use the operationalisation of UID in Collins (2014) and Meister et al. (2021):[4].

UID is the measure of the average (squared) information change from word to word in a sentence. In Formula 2, $I(w_i)$ is the information / surprisal of a word, n is the number of words in a sentence.

$$UID = -\frac{1}{n-1}\sum_{i=2}^{n}(I(w_i) - I(w_{i-1}))^2 \quad (2)$$

In order to make the determination of UID a maximisation problem, (Jain et al., 2018) define UID as negative. Therefore, a UID value close to zero indicates a 'good' information density distribution, that is, on average a smooth Flow of Information in sentences.

[2]We are not aware of studies on subtext in an information-theoretic framework. However, there are studies on subtext that deal with information, although not quantifiable. Taking Sims and Bamman (2020) as an example who are concerned with the propagation of information in literary texts. But this is about propositional knowledge, not probabilistically modelled, that is to say, information theoretic measures are not employed.

[3]For a Python implementation see https://github.com/jnphilipp/tcm.

[4]Source code available at https://github.com/jnphilipp/uid.

3 The study

3.1 Models and techniques

In general, we employ six enrichment techniques and two information models M1 and M2 yielding eight experimental conditions. The bag-of-words-model TCM does not consider word order. In contrast, Llama 3.2-3B is an incremental model that recalculates the probabilities of words with each new context word that is added.

Table 1 illustrates the eight conditions. The output of each condition is surprisal values of words (**OT**: original text; **LLM**: large language model; **MG**: meaningful glosses):

enrichment	**information models**	
	M1:Llama	M2: TCM
OT	words surprisal	
OT + MG: NLP (Irmer et al., to appear)	words surprisal	
OT + MG: LLM	words surprisal	
OT + fake glosses	words surprisal	

Table 1: Enrichment types and information models.

3.2 Techniques

3.2.1 Enrichments

Our methodology serves to observe fundamental differences in the surprisal for each text word before and after meaningful or fake enrichments. For glossing, we used (i) a traditional NLP technique described in Irmer et al. (to appear), (ii) an enrichment based on a large language model (LLM) and (iii) a fake enrichment.

(i) **NLP glosses**: The original text was enriched by inserting BabelNet-based glosses for content words.[5] The following processing steps are involved: first POS tagging, lemmatisation and filtering take place. Subsequently, **Word Sense Disambiguation** (WSD) identifies the most probable sense for each lemma, which is then looked up in BabelNet (Navigli and Ponzetto, 2012). Two disambiguation strategies are applied: **Lesk algorithm** is based on textual overlap of BabelNet glosses, while **Graph connectivity** builds a BabelNet neighbour graph.

Four enrichment variants result from varying the lexicon used for look-up (ALL BabelNet lexicons vs. WNTR, WordNet-translations only)[6] and varying the WSD algorithm: LESK (overlap of glosses) vs. GRAPH (graph connectivity).

(ii) **LLM-based enrichment**: We used different LLMs provided by Google Vertex AI: initially, for glosses after paragraph text-bison-@001 was used (in the following referred to as **Bison**), and for inline glosses gemini-2.5-flash-preview-04-17 (in the following **Gemini**). For the latter, we applied the following system prompt: *Find content words (nouns, adjectives, adverbs) in the text given by the user prompt and provide a Russian gloss explaining them. The gloss should be a description or explanation in about 10 words in Russian language. Replicate the original text exactly (including exact preservation of line breaks and empty lines), only adding the glosses in parentheses after the corresponding word.* The original Chekhov text was then given as a user prompt.

(iii) **Fake glosses**: For comparison, we produced "fake" glosses by adding a pseudo-enrichment consisting of random sentences from the *rus_news_2020_1M* corpus (1M sentences) from the *Wortschatz Leipzig* corpora collection[7].

All texts used in the experiments consist of 186 paragraphs. The original text consists of 8398 tokens corresponding to 3336 unique lemmas. The texts enriched by the BabelNet pipeline consist on average of 31098 tokens, 6033 lemmas, the fake text of 39467 tokens, 7530 lemmas. The fake news text has 31648 tokens and 9032 lemmas, the Bison-generated text has 18057 tokens and 4539 lemmas, and the Gemini-generated text 20242 tokens and 7365 lemmas.

As an example, we give the first sentence of the first paragraph of the original text together

[5] Implemented using Apache UIMA (Ferrucci et al., 2009) and open-source DKPro components (dkp, 2017), including DKPro HunPosTagger (Halácsy et al., 2007) and DKPro LanguageToolLemmatizer,

[6] Regarding *WordNet*-ontology, see Miller (1994)

[7] https://wortschatz.uni-leipzig.de/de

with an English translation:[8] В больничном дворе стоит небольшой флигель, окруженный целым лесом репейника, крапивы и дикой конопли. 'In the hospital yard stands a small wing surrounded by a whole forest of burr, nettle and wild hemp.'
Then lemmatized with the glosses from ALL GRAPH where the glosses are in brackets:
больничный двор стоить небольшой флигель (пристройка) окруженный целый лес репейник кра пива дикий конопля (марихуана). 'medical yard stand small wing (annex) surrounded whole forest burr nettle wild hemp (marijuana).'

4 Results

The plots in Figure 1 compare UID-distributions across all conditions (see Table 1 above). At first glance it turns out that the Llama-based UID-distributions (M1) differ fundamentally from the TCM-based ones (M2). This is probably due to the low probabilities and thus high information values of the Llama model, which operates in a much larger probability space, i.e. the entire vocabulary of the training texts, than TCM, which is limited to a single text of even only a paragraph.

With **M1**, we observe in the conditions **OT**, **OT + MG: NLP**, **OT + fake glosses** near-normal distributions (see Figures 1a, 1c,1e,1g). Both MG and fake glosses yield better UID-distributions (=closer to zero) than OT. In case of MG, H1 is confirmed, but surprisingly H1 also seems to hold for fake glosses.

With **M2** and the conditions **OT**, **OT + MG: NLP**, **OT + fake glosses** (see Figures 1b, 1d, 1f, 1h) all distributions exhibit much higher *kurtosis* (peakiness) and *skewness* that is to say, the UID values are concentrated within small intervals, and the distributions are asymmetric. The plots do not provide evidence for H1. Rather, H2 is confirmed, as fake glosses have peaks slightly more distant from zero than the original text. However, surprisingly, this holds also for MG. This is the reverse scenario of the experimental conditions with M1 above. With M2, all text manipulations, be it with MG, be it with fake glosses, lead to (slightly) less uniform distributions.

For the condition **OT + MG: LLM** using **Gemini**-enrichment, and employing **M1**, near-normal distributions as in the conditions above come to light (see Figure 1i): the Gemini-enriched text has a slightly higher, better, density of UID then the original text, however the former is located between the OT-distribution and fake-distributions which corresponds to the observations with M1 above.

For the condition **OT + MG: LLM** using **Bison**-enrichment, and employing **M2**, the picture changes (see Figure 1j): H2 is confirmed since fake glossing has a less favourable distribution of UID than OT but here, as above, the meaningful glossing is positioned between OT and fake glossing which contradicts H2. With regard to the confirmation of H1 and H2, the glossing technique, i.e. TCM vs. LLM, is not relevant.

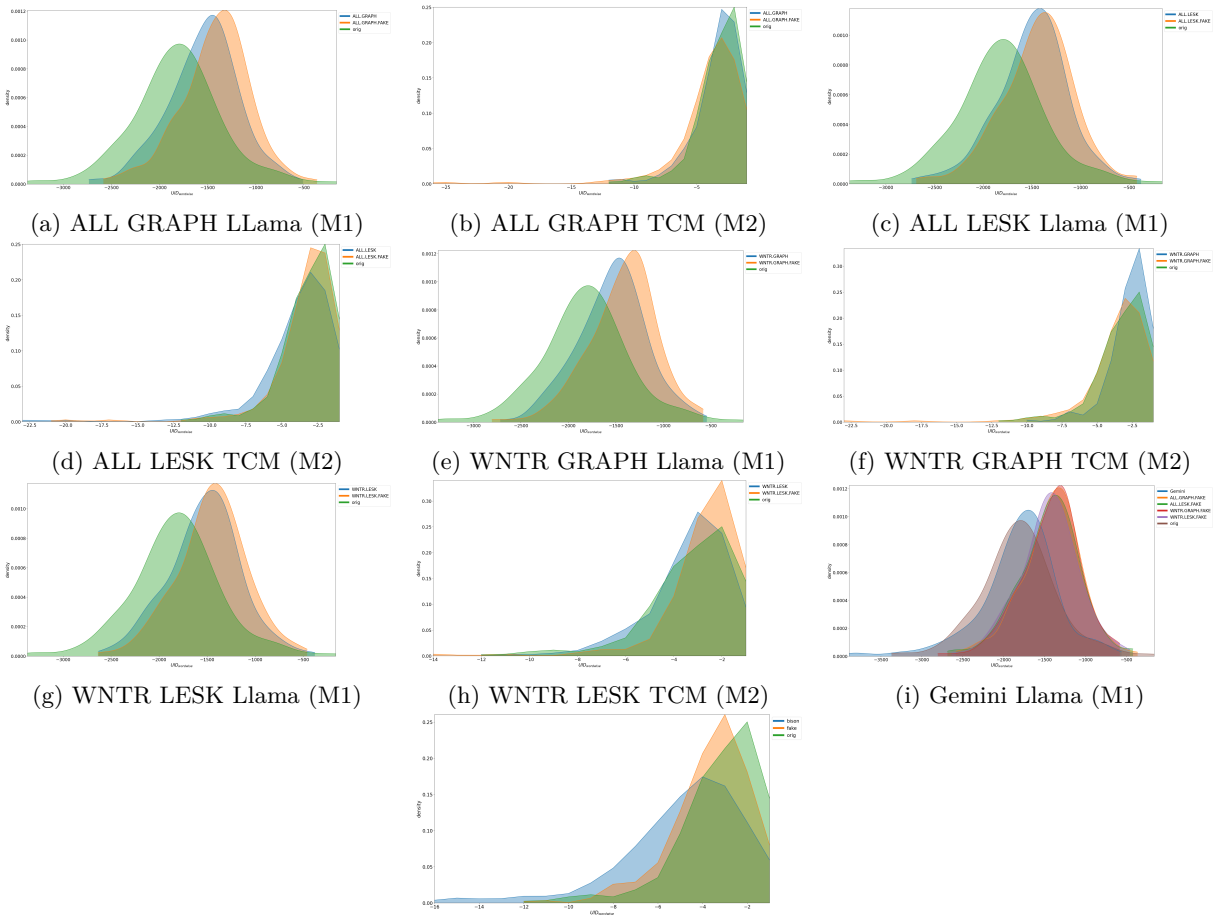
5 Discussion

Under the experimental conditions both with NLP and Bison-glossing and employing M2, our hypotheses could not be confirmed at the same time: H2 turned out to be true, while H1 did not. Meaningful and fake enrichments could be distinguished from OT through UID-distributions, however, both fake-glossing and MG had a lower (=worse) density of UID than OT.

We observed the reverse situation in the test series based on M1: here H1 came out to be true, while H2 did not. Again, the results were different from what we had expected since the effect that we had hoped for from MG (and which would have justified viewing them as models of the subtext), namely an approximation of the UID values to zero (cf. H1), occurred most strongly with the fake-glossed text. Assuming that the Gemini-based enrichment represents a good or maybe even a human-like of the subtext, the results show that, in semantic respect, the more remote the enrichment is from the original text, the better the UID density becomes.

In general, we observe that enrichments of any type lead to UID distributions that differ form OT's UID-distribution, but without differentiating between MG and fake glossing. Hence, we have an enrichment effect but not a subtext effect.

[8] The full data can be found under https://github.com/jnphilipp/chekhov-data.



(a) ALL GRAPH LLama (M1) (b) ALL GRAPH TCM (M2) (c) ALL LESK Llama (M1)

(d) ALL LESK TCM (M2) (e) WNTR GRAPH Llama (M1) (f) WNTR GRAPH TCM (M2)

(g) WNTR LESK Llama (M1) (h) WNTR LESK TCM (M2) (i) Gemini Llama (M1)

(j) Bison TCM (M2)

Figure 1: Density plots from UID-distributions.

There are two main ways to interpret these findings: either our "meta-hypothesis" is incorrect, i.e., UID is *not* an inadequate measure for quantifying text clarity, or our attempt at adding explicitised subtext does not achieve its intended goal. However, it is also possible that both are true or that the answer lies somewhere in between. It is undeniable that the distributions of the meaningful enrichments are different from those of both the unenriched and the fake-enriched texts. Hence, it is thinkable that a subtext effect exists, but it was overshadowed by the enrichment effect due to an inadequate experimental setup.

In any case, future research is needed to explain and interpret these effects in relation to the subtext.

Limitations

- The enrichments are machine generated texts and cannot be considered a genuine subtext in literary or communicative sense,
- our pilot study is based on a single story, which may constrain the generalisability of our observations,
- no human raters were involved in the evaluation of the glosses,
- the glosses vary in length, particularly in terms of the number of words.

References

2017. DKPro Core Component Reference. https://dkpro.github.io/dkpro-core/releases/2.2.0/docs/component-reference.html. Accessed: 2022-09-01.

Chris Baldick. 2015. *The Oxford dictionary of literary terms*. Oxford University Press.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Michael Xavier Collins. 2014. Information density and dependency length as complementary cognitive models. *Journal of psycholinguistic research*, 43(5):651–681.

August Fenk and Gertraud Fenk. 1980. Konstanz im kurzzeitgedächtnis-konstanz im sprachlichen informationsfluß. *Zeitschrift für experimentelle und angewandte Psychologie*, 27(3):400–414.

David Ferrucci, Adam Lally, Karin Verspoor, and Eric Nyberg. 2009. Unstructured information management architecture (UIMA) version 1.0. OASIS Standard.

Paul Grice. 1989. *Studies in the way of words*. Harvard University Press.

Péter Halácsy, András Kornai, and Csaba Oravecz. 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*, pages 209–212, Prague, Czech Republic. Association for Computational Linguistics.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the ACL on Language technologies*, pages 1–8.

Matthias Irmer. 2011. *Bridging inferences*. de Gruyter.

Matthias Irmer, Olav Mueller-Reichau, J. Nathanael Philipp, and Michael Richter. to appear. In quest of the subtext: Information theory measures the implicit in chekhov. *Digital Humanities Quarterly*. Forthcoming.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62.

Ayush Jain, Vishal Singh, Sidharth Ranjan, Rajakrishnan Rajkumar, and Sumeet Agarwal. 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, Santa Fe, New-Mexico. Association for Computational Linguistics.

Rolf-Dieter Kluge. 1995. *Anton P. Čechov: eine Einführung in Leben und Werk*. Wissenschaftliche Buchgesellschaft.

Max Kölbl., Yuki Kyogoku., J. Philipp., Michael Richter., Clemens Rietdorf., and Tariq Yousef. 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.

Max Kölbl, Yuki Kyogoku, J. Nathanael Philipp, Michael Richter, Clemens Rietdorf, and Tariq Yousef. 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.

Elena I. Lelis. 2016. Leksičeskie sredstva formirovanija podteksta v proze A.P. Čechova. *Slavjanskie čtenija*, 8(14):120–133.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Roger Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19:849–856.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. *arXiv preprint arXiv:2109.11635*.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Viktor Ja. Myrkin. 1976. Tekst, podtekst i kontekst. *Voprosy jazykoznanija*, 2:86–93.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193(0):217 – 250.

Aleksandr Nikolaevič Nikoljukin. 2003. *Literaturnaja ėnciklopedija terminov i ponjatij*. NPK "Intelvak".

J. Nathanael Philipp, Max Kölbl, Yuki Kyogoku, Tariq Yousef, and Michael Richter. 2022. One step beyond: Keyword extraction in german utilising surprisal from topic contexts. In *Intelligent Computing*, pages 774–786, Cham. Springer International Publishing.

J. Nathanael Philipp, Michael Richter, Erik Daas, and Max Kölbl. 2023. Are idioms surprising? In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 149–154, Ingolstadt, Germany. Association for Computational Linguistics.

Claude E. Shannon and Warren Weaver. 1949. *The Mathematical Theory of Communication*. University of Illinois Press.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Matthew Sims and David Bamman. 2020. Measuring information propagation in literary social networks. *arXiv preprint arXiv:2004.13980*.

Rose Whyman. 2010. *Anton Chekhov*. Routledge.