Reinterpreting the surprisal calculation of the Topic Context Model

J. Nathanael Philipp
Saxon Academy of Sciences and Humanities in Leipzig
Karl-Tauchnitz-Str. 1
04107 Leipzig
jnathanael@philipp.land

Abstract

In this paper, we propose and demonstrate an novel computation of wordwise surprisal within the information-theoretic Topic Context Model (TCM). Rather than calculating the average surprisal as in previous studies, the TCM is redefined to output a simple, non-averaged surprisal value. This adjustment facilitates a more straightforward comparison with other common surprisal models and measures, such as n-gram surprisal.

1 Introduction

In previous studies, the Topic Context Model (TCM) has been successfully applied to a wide range of tasks, including keyword extraction (Kölbl et al., 2020, 2021; Philipp et al., 2022), identification of idioms (Philipp et al., 2023), the analysis of expressivity of intensifiers (Philipp et al., 2024), subtext detection (Philipp et al., 2025b,c), and calculation of semantic information (Philipp et al., 2026). More recently, Philipp (in press) conducted a detailed examination of the TCM's surprisal formulas, uncovering several noteworthy patterns and interrelations in the behaviour of the surprisal values. However, all prior approaches share a common characteristic: the TCM computes an average surprisal, whereby a separate surprisal score is calculated for each topic and subsequently averaged. In this paper, we propose a reinterpretation of the TCM in which the calculation of average surprisal is replaced by a single, unified surprisal value. As demonstrated in Philipp (in press), the mean level of the average surprisal is systematically elevated, an effect that arises solely from the number of topics in the model. Furthermore, Philipp et al. (2024) demonstrated a substantial discrepancy in surprisal values between the TCM and unigram models, which has hitherto complicated direct comparisons between them. The new formulation is evaluated through a comparative analysis analogous to that presented in Philipp (in press). This analysis contrasts the behaviour of the revised surprisal measure with that of the original approach.

2 Topic Context Model

The Topic Context Model (TCM) (Kölbl et al., 2020, 2021; Philipp et al., 2022, 2023, 2024, 2025a,b,c, 2026; Philipp, in press)[1] is based on the Surprisal Theory of Hale (2001) and Levy (2008). It builds on Shannon's information theory (Shannon, 1948, 1951) and on Dretske (1981), who distinguished between the quantitative measure of information in Shannon's sense and semantic information, which transfers knowledge. Both types of information have an impact on brain activity. This is the starting point for the Surprisal Theory, which links information to the cognitive processing effort required (see, the convincing studies of Bentum et al. (2019), Wilcox et al. (2023)). TCM measures surprisal, a quantitative measure which, in addition, due to its context, also incorporates the linguistic property of meaning.

The surprisal of a word w depends on its conditional probability in a given context, see Formula 1. Currently the TCM calculates the $\overline{\text{surprisal}}(w_d)$ for a word w in a document d, as the average over the topics of the underlying topic model, see Formula 2 (see Philipp (in press) for the evolution of the TCM formulas).

Here, $c_d(w)$ is the frequency of a word w given a document d, $|d|$ is the total number of words in the document d, so $\frac{c_d(w_d)}{|d|}$ denotes the relative word frequency in a document. Further $\phi \in \{0, 1\}$ defines whether to use the relative word frequency or exclude it from the calculations, T_d is the topic distribution T of the document d and let it be indexed such that $P(t_i|d) \geq P(t_j|d)$ if and only if $i \leq j$. Then pick a $\mu \in (0, 1]$, also called

[1] https://github.com/jnphilipp/tcm

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

$$\overline{\text{surprisal}}(w_d) = -\phi \log_2 \frac{c_d(w_d)}{|d|} - \frac{1}{|T_d|} \sum_{i \in T_d} \log_2 \left(P(t_i|d) f(w_d, t_i)\right) \tag{2}$$

threshold, and let k_μ be the largest integer such that $\sum_{i=1}^{k_\mu} P(t_i|d) \geq \mu$. For $k_\mu = 0$ we still require $T_d^\mu = \{1\}$ in order to keep the sets non-empty. See Formula 3 for the definition of $f(w_d, t_i)$, which gives the probability $P(w|t_i)$.

Here WT is the normalised word topic distribution $WT \in \mathbb{R}^{|T| \times N}$ of the Latent Dirichlet Allocation (LDA, the underlying topic model, Blei et al. (2003)[2]) and N is the number of words. Further, we denote the entries of WT by WT_{w_d,t_i} and the columns in $\mathbb{R}^N$, which correspond to the topics, by WT_{t_i}. The notation $w \in WT$ means that there exists a row of WT corresponding to the word w. The case that $w \notin WT$ only occurs when using a TCM that was trained on different words than using to calculate surprisal for, i.e., a trained TCM is used on texts that were not part of the training corpus with new words.

So currently the underlying probability $P(w_d|t_i)$ is calculated as in Formula 4.

$$P(w_d|t_i) = P(t_i|d) f(w_d, t_i) \tag{4}$$

The first change proposed is to reformulate the probability $P(w)$ and calculate it according to Formula 5.

$$P(w_d) = \sum_{i \in T_d} P(t_i|d) f(w_d, t_i) \tag{5}$$

Estimating that $P(w_d)$ has the following bounds: $\min_{t_i} P(w|t_i) \leq P(w) \leq \max_{t_i} P(w|t_i)$. The relative word frequency $\frac{c_d(w_d)}{|d|}$, will no longer be part of the surprisal calculation, as it behaved as a summand (see Formula 2). Thus the surprisal for the TCM can now be calculated according to Formula 6.

$$\text{surprisal}(w_d) = -\log_2 \sum_{i \in T_d} P(t_i|d) f(w_d, t_i) \tag{6}$$

[2] `model.components_ / model.components_.sum(axis=1)[:, np.newaxis]` as suggested by https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

Note that using the TCM with just one topic is roughly equivalent to an unigram surprisal model.

3 Text Corpora

In total, five corpora are used in this analysis. Four corpora from the Leipzig Corpora Collection (Goldhahn et al., 2012)[3], the German news corpora from the years 2020 to 2023, each with 1 million sentences. The other corpus is a collection of Tweets and Youtube comments, which were grouped together into longer text.

4 Results

For each of the five corpora, four TCMs were trained, with 10, 20, 50 and 100 topics respectively, with the threshold $\mu = 1$. Figure 1 shows the average surprisal distribution with $\mu = 1$ for all topic-corpus-distributions and Figure 2 shows the same distributions for suprisal based on Formula 6.

Figure 1 shows that for the average surprisal, the increase of the number of topics the overall level of the average surprisal values also increases, while the spread of the average surprisal values stays the same. In contrast Figure 2 shows that for the new surprisal formulas the overall level of the surprisal values is significantly lower than for the average surprisal. It also shows that increasing the number of topics the overall level of the surprisal values decreases, but far less than it increases for the average surprisal. The spread of the surprisal values stays also the same.

Figure 3 and Figure 4 show the spread of the surprisal values for each word in the corpora. For the Wortschatz news corpora, Figure 3 shows, that the spread per word of the average surprisal values is quite low, for the Twitter/YouTube corpus it is about double that. But overall the spread per word stays consistent for the different number of topics.

Conversely, Figure 4 discloses a considerably broader distribution of surprisal values for 10 subjects, marked by a substantial number of elevated outliers. This pronounced variability is observed almost exclusively in the Wortschatz corpora. For

[3] https://corpora.uni-leipzig.de

$$P(w|t_i) = f(w_d, t_i) = \begin{cases} WT_{w_d,t_i} & \text{for } w \in WT \\ \frac{1}{N} & \text{for } w \notin WT \end{cases} \tag{3}$$

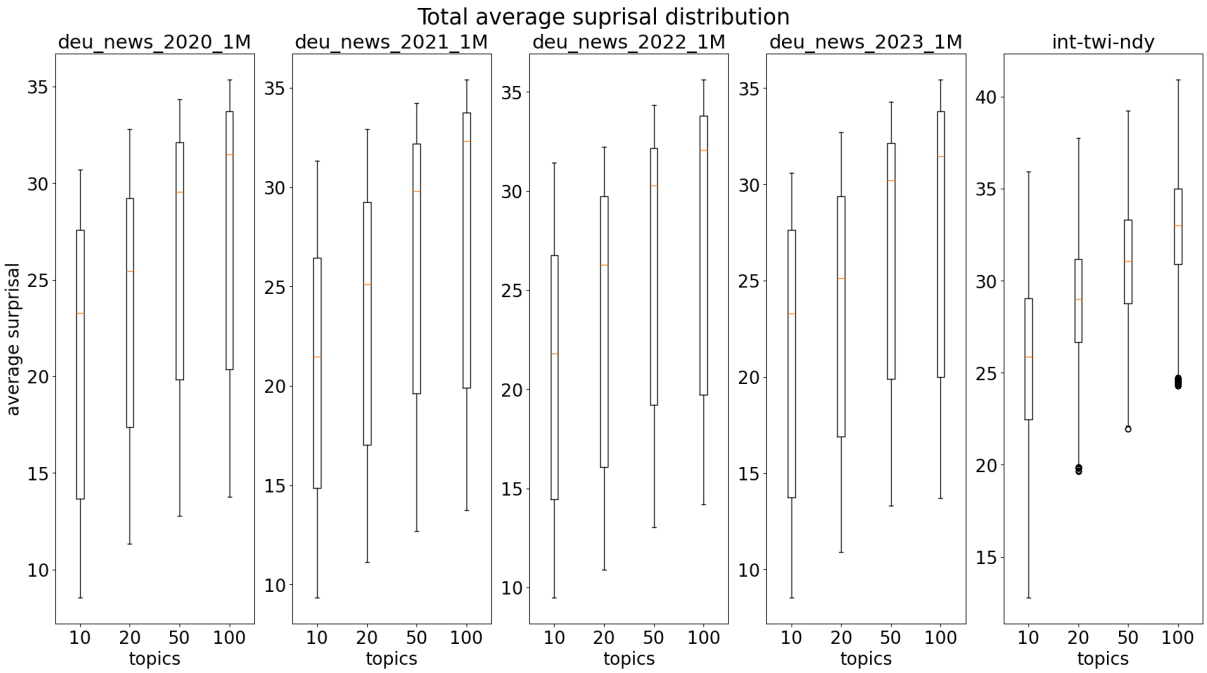

Figure 1: Average surprisal distribution based on Formula 2, from Philipp (in press).

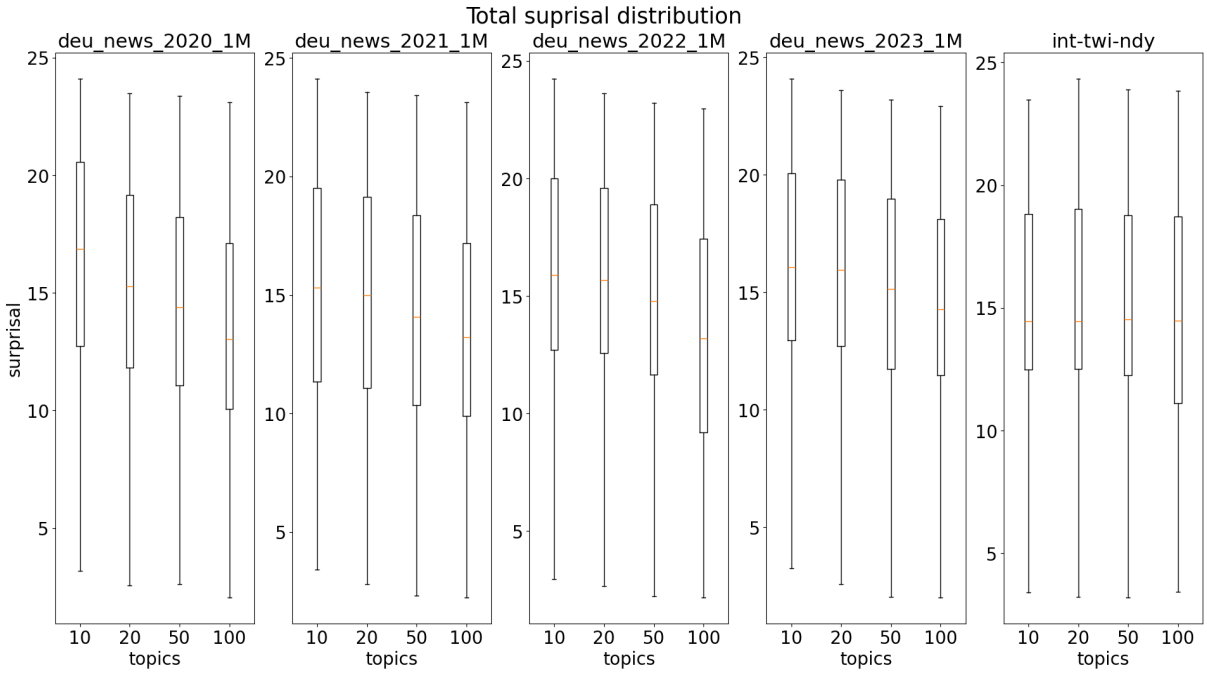

Figure 2: Surprisal distribution based on Formula 6.

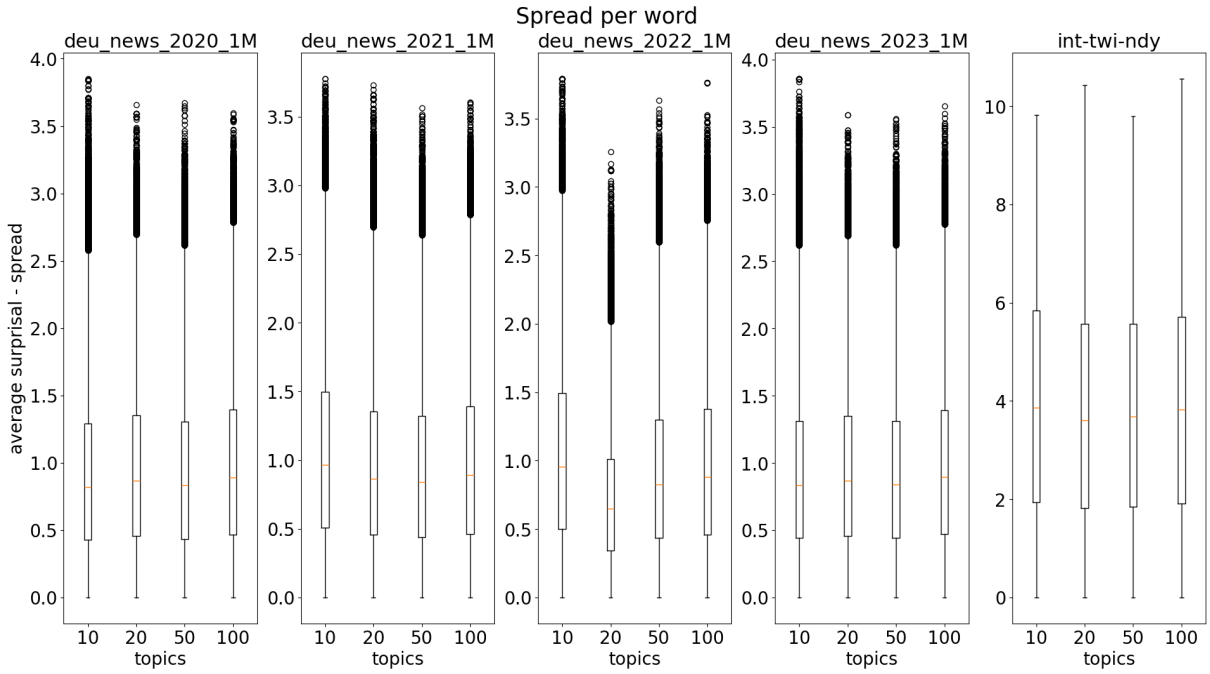

Figure 3: Average surprisal spread distribution

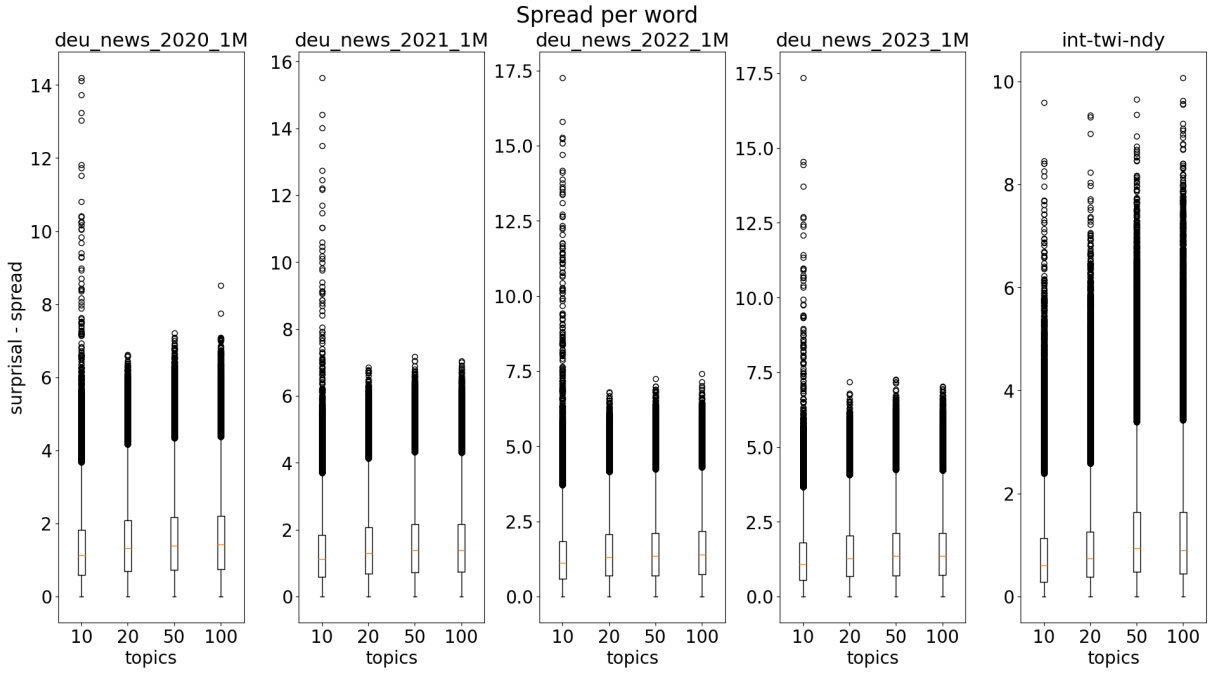

Figure 4: Surprisal spread distribution

the remaining corpora, the spread of surprisal values is considerably smaller. There are also hardly any differences in the distribution of the surprisal values across the other topics. As demonstrated in Figure 4, the surprisal values exhibit a wider distribution than in Figure 3. However, for the Twitter/YouTube corpus, the spread remains comparable. Notably, Figure 4 highlights a higher prevalence of outliers.

5 Conclusion

In this paper, we have demonstrated that replacing the average surprisal with a unified, non-averaged surprisal measure in the Topic Context Model (TCM) leads to substantially lower surprisal values while simultaneously increasing the spread of surprisal per word. One should note that increasing the number of topics has no significant effect on the overall distribution and spread of the surprisal values across the corpus. The number of topics still has negligible systematic influence on the resulting surprisal scores, just like it does with the average surprisal. It is important to note, however, that the present study focuses exclusively on a direct comparison between the original and the revised TCM surprisal formulations. Consequently, future research should extend beyond this methodological comparison and evaluate the practical implications and performance of the new surprisal measure in downstream NLP tasks and real-world applications.

Limitations

This paper contains only a short overall analysis of the new TCM formula, with only the threshold $\mu = 1$, see Philipp (in press) for the influence of the different thresholds on the original formula.

Secondly, the new unified surprisal calculation is only been compared analytically in this paper and not with downstream tasks as in previous papers.

References

Martijn Bentum, Louis Ten Bosch, A. Van den Bosch, and Mirjam Ernestus. 2019. Listening with great expectations: An investigation of word form anticipations in naturalistic speech. In *Proceedings of Interspeech 2019*, pages 2265–2269, Graz, Austria. Interspeech 2019 : 20th Annual Conference of the International Speech Communication Association.

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

Fred Dretske. 1981. *Knowledge and the Flow of Information*. MIT Press.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Max Kölbl, Yuki Kyogoku, J. Nathanael 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. In Roussanka Loukanova, editor, *Natural Language Processing in Artificial Intelligence - NLPinAI 2020*, volume 939 of *Studies in Computational Intelligence (SCI)*, pages 139–161. Springer International Publishing.

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

J. Nathanael Philipp. in press. Surprising results in the surprisal of the Topic Context Model. Walter de Gruyter.

J. Nathanael Philipp, Max Kölbl, and Michael Richter. 2025a. Surprisal in action: A comparative study of LDA and LSA for keyword extraction. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, pages 1–11, Hannover, Germany. HsH Applied Academics.

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, Max Kölbl, and Michael Richter. 2026. Semantic Information: A Difference That Makes a Difference. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 11300–11308, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

J. Nathanael Philipp, Olav Mueller-Reichau, Matthias Irmer, Michael Richter, and Max Kölbl. 2025b. Can information theory unravel the subtext in a chekhovian short story? In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 84–90, Vienna, Austria. Association for Computational Linguistics.

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.

J. Nathanael Philipp, Michael Richter, Olav Mueller-Reichau, and Matthias Irmer. 2025c. Information theory unravels the subtext in Chekhov. *Digital Humanities Quarterly*, 19(2).

J. Nathanael Philipp, Michael Richter, Tatjana Scheffler, and Roeland van Hout. 2024. The role of information in modeling German intensifiers. In *Information structure and information theory*, pages 117–145. Language Science Press.

Claude E Shannon. 1951. Prediction and entropy of printed English. *Bell system technical journal*, 30(1):50–64.

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

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.