What does surprisal have to do with information status?

Andrew Dyer
Language Science and Technology
Saarland University, Germany
andrew.dyer@uni-saarland.de

Abstract

It is common in cognitive computational linguistics to use language model surprisal as a measure of the information content of units in language production. From here, it is tempting to then apply this to information structure and status, considering surprising mentions to be *new* and unsurprising ones to be given, providing us with a ready-made continuous metric of information givenness/newness. To see if this conflation is appropriate, we perform regression experiments to see if surprisal is actually well predicted by information status as manually annotated, and if so, if this effect is separable from more trivial linguistic information such as parts of speech and word frequency. We find that information status alone is at best a very weak predictor of surprisal, and that surprisal can be much better predicted by the effect of parts of speech, which are highly correlated with both information status and surprisal; and word frequency. We conclude that surprisal should not be used as a continuous representation of information status by itself.

1 Introduction

Language model surprisal – a measure of the unexpectedness of a following word in a sequence – is commonly used as a measure of the difficulty of processing language, and is surprisingly adaptable to a wide range of tasks (Goldstein et al., 2022), most notably reading times (de Varda and Marelli, 2022; Wilcox et al., 2023), among others. While the concept of surprisal is agnostic to the architecture of the language model used, increasingly neural, and particularly transformer-based language models are used.

Transformer-based language models are able to pay attention to previous context when making their next-word predictions (Vaswani et al., 2017). For this reason, they are ubiquitous in coreference and anaphora resolution, a task which requires connecting an often ambiguous, closed-class mention of a referent to its previously mentioned antecedent (Ogrodniczuk et al., 2025). This makes them promising in classification of information status – the givenness or newness of mentions and the referents they refer to (Chafe, 1976). Probes on transformer language models have also shown that their hidden representations can be used to predict information status, without finetuning for the task (Loáiciga et al., 2022).

With this in mind, it is tempting to then consider surprisal as a measure of that which which is "novel and unexpected" (Xu and Futrell, 2024), and this would fit with the conception of given information as that which is predictable or topical in context (Givón, 1983). It would certainly be intuitive that, if a mention refers to an entity that has been encountered in discourse previously, it should be less surprising for a language model that can encode this context.

In this study, we aim to test whether this intuition holds by measuring the extent to which surprisal is predicted by information status of mentions in a multilingual corpus. In doing so, we also examine whether information status itself is predicted by two morphosynactic cues: parts of speech and dependency relations, so that we can see whether the effect of information status rises to the surface in surprisal beyond the context-free information provided by these cues. We report results for nine languages: Chinese (Mandarin), English, German, Greek, Hindi, Indonesian, Portuguese, Turkish and Ukrainian.

2 Related work

Loáiciga et al. (2022) conducted a probing experiment on two English-language transformer-based language models (Transformer-XL and GPT2) to determine whether their parameters can be used to predict given- or newness of mentions. They found that hidden representations of tokens in contextual language models were indeed sufficient to clas-

Proceedings of the 8th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, pages 26–31
March 29, 2026 ©2026 Association for Computational Linguistics

sify a mention (or mention head) as given or new, though they were less good at extracting mentions from text. They used a part-of-speech baseline in their study, whereby pronouns and definite noun phrases were considered given, and found that the systems regularly beat this baseline.

3 Experimental setup

3.1 Design

Our experiments are simple regression experiments to investigate the following:

1. First, to see whether information status itself is predicted by two syntactic features: universal parts of speech (UPOS) and dependency relations (deprel). We do this using logistic regression, with information status as the response variable and UPOS and deprel as explanatory variables. F1 score is reported in this experiment.
2. Second, to measure the extent to which information status predicts surprisal, independently of the two syntactic predictors (UPOS and deprel) and word frequency. R2 score is reported in this experiment.

In both experiments, we run analyses with different input features, including combinations of features, to see which contribute to the regression model and which can be removed without degrading the fit of the model.

3.2 Data

For our experiments, we use CiepInf (Dyer et al., 2024), a parallel corpus of modern prose built on top of the mini-CIEP+ corpus (Verkerk and Talamo, 2024), annotated into Universal Dependencies (Nivre et al., 2020) using UDPipe v.2 (Straka, 2018), and annotated for information status. The corpus is available in conllu format, and mention annotation follows the CorefUD format (Nedoluzhko et al., 2022). The corpus currently has annotated data in nine languages: English, Chinese (Mandarin), German, Greek, Hindi, Indonesian, Portuguese, Turkish, and Ukrainian. We show the data sizes in Table 1.

We extract mentions from CiepInf using Udapi (Popel et al., 2017). As mentions can have varying lengths, we extract surprisal of the syntactic head of the mention, rather than full spans.

	Sentences (approx)	Mentions	
		New	Old
Chinese	3500	8294	7940
English	6380	18998	14560
German	560	810	1161
Greek	900	1361	2086
Hindi	3100	6963	3857
Indonesian	1880	3459	3176
Portuguese	2830	4252	4121
Turkish	580	714	813
Ukrainian	900	1310	1042

Table 1: Table of approximate number of sentences and mentions in CiepInf.

3.3 Resources

For all regressions, we use the sk-learn package in Python (Buitinck et al., 2013). For the first experiment (predicting information status from parts of speech and dependency relations) we use logistic regression. For the second experiment (predicting surprisal from given predictors) we use poisson regression, as surprisal is non-negative and has a long tailed distribution. In all cases, we apply the default L2 regularisation to constrain weights and reduce overfitting. We also use eight-fold cross-validation, and report averaged results across runs.

For surprisal, we use the Huggingface mGPT (Shliazhko et al., 2024) model[1], a multilingual model which covers all the languages that are of interest to us, and is autoregressive (i.e. unidirectional), which is suitable for measuring surprisal as the unexpectedness of following words. We tokenize using the mGPT tokenizer with a context size of 256 tokens, and a stride of 128. Because pretrained mGPT models are trained using byte-pair encoding, which produces token boundaries incongruous with Universal Dependencies token boundaries, we enforce byte-pair splitting only within UD token boundaries, and after inference merge these subtokens into their parent UD token, along with their surprisals, which are summed.

For word frequencies, we use the wordfreq package in Python (Speer, 2022) to get word frequencies, and specifically the *zipf_frequency* function to scale frequencies between languages. Since the Chinese lookup is not functional in this package at time of writing, we exclude Chinese from this part of the analysis.

[1] https://huggingface.co/ai-forever/mGPT

	UPOS	deprel	UPOS + deprel	UPOS ∗ deprel
Chinese	.84	.68	.84	.84
English	.82	.68	.82	.82
German	.98	.8	.98	.98
Greek	.88	.76	.88	.88
Hindi	.79	.43	.8	.8
Indonesian	.76	.71	.77	.77
Portuguese	.75	.69	.76	.76
Turkish	.78	.65	.78	.78
Ukrainian	.83	.65	.84	.84

Table 2: logistic regression scores (F1) of information status by UPOS and dependency relations. + means two features being used in a model independently, while ∗ means the interaction between the two of them.

4 Results

4.1 Experiment 1: Correlates of information status

Table 2 shows the F1 scores of the first experiment. We run regression models with different combinations of features: UPOS alone, deprel alone, UPOS and deprel as independent features, and the interaction between UPOS and deprel.

In all languages, we find that information status is well predicted by parts of speech and dependency relations; particularly the former. Dependency relations are a weaker predictor, and when combined with parts of speech in an interaction, the F1 is no higher, suggesting that this feature adds little information not already captured by parts of speech.

Figure 1 shows the weights (coefficients) of UPOS as predictors of surprisal between languages (from the analysis using only UPOS). Pronouns tend to be given, while nouns tend to be new. Proper nouns have a generally weak effect: they can be given or new. The trend is relatively consistent between our nine languages, though Greek shows an unusually higher tendency of proper nouns to signal new information.

So far, this is in line with the observation that pronouns, as reduced, closed class mentions, are indicative of given referents, and open class, full referring expressions are indicative of new ones (Gundel et al., 1993). It also tells us that when it comes to information status, there is a colinear effect of parts of speech that is hard to separate.

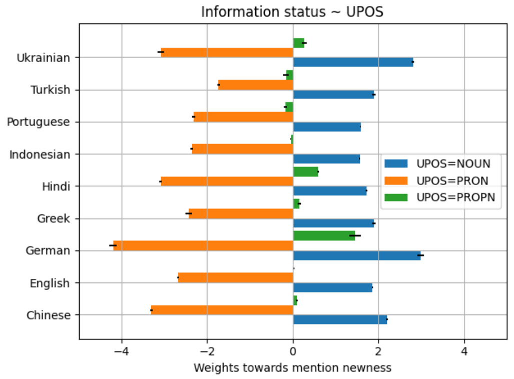

Figure 1: Weights towards information status newness by UPOS per language.

4.2 Experiment 2: Predictors of surprisal

Next, we explore the extent to which surprisal responds to information status as compared to other predictors, including parts of speech. Table 3 shows the results of the poisson regressions. Again, we run analyses with different features. The first two analyses are with UPOS and information status alone, respectively. The third is the interaction between information status and UPOS. The last two use word frequency: in a multiple model with the infstat-UPOS interaction, and alone, respectively.

We see that information status alone seems to have very little predictive power, shown by the low R2 scores. Once again, the effect it has is surpassed by UPOS, and there is little gain from the interaction of the two, telling us that surprisal responds much more to UPOS than information status. The strongest predictor of all appears to be word frequency, and it gains little from the interaction between parts of speech and information status being added to the model.

It is to be expected that a continuous variable (frequency) should provide the best fit to a continuous response variable (surprisal). But it is also remarkable that parts of speech alone are so predictive of surprisal, and considerably more so than information status.

5 Discussion

Our first finding – and one that is fairly intuitive – is that information status is itself well correlated with parts of speech. For example, given referents are very often pronominalised, while fully referring noun phrases are more frequently used for new referents (Gundel et al., 1993). This provides us with an intuitive baseline: if parts of speech are in-

	UPOS	infstat	infstat $*$ UPOS	infstat*UPOS + frequency	frequency
Chinese	.19	.03	.17	–	–
English	.17	.1	.17	.38	.38
German	.14	.16	.16	.36	.36
Greek	.09	.08	.09	.42	.42
Hindi	.21	.14	.21	.42	.42
Indonesian	.12	.09	.13	.22	.22
Portuguese	.02	.02	.03	.23	.23
Turkish	.03	.04	.03	.22	.22
Ukrainian	.17	.12	.19	.4	.4

Table 3: Poisson regression scores (R2) of surprisal by information status, UPOS, and frequency (and interactions). Once again, + means two features being used in a model independently, while $*$ means the interaction between the two of them. Chinese is excluded from all frequency experiments due to limitations in the Python package.

formative of mention givenness/newness, then any proposed or demonstrated sensitivity of surprisal to information status must be shown to be clearly independent from this.

As for surprisal, this also appears to be sensitive to parts of speech, but also, to simple word frequency. Both of these add predictive power to the regression model, but information status itself adds little. The response to parts of speech is intuitive to us as surprisal is, in the end, an output of the probabilities of next-word prediction, and where an open-class part of speech is expected there is a wider range of possible completions. The response to word frequency is also unsurprising from a language-modeling perspective, as language models have previously been shown to increasingly fit to this, to the detriment of fit to measures such as reading time (Oh et al., 2024).

6 Future work

There are baseline systems we could include in our study. For example, this study finds that information status is only weakly predictive of surprisal, despite transformer language models theoretically being able to encode long distance contexts. But do they at least outperform baselines such as LSTM language models, which have some ability to store previous context in a memory representation but no transformer architecture to "look back" at previous context; or a statistical language model such as a Kneser-Ney n-gram language model, which operates within a fixed window size? If the answer to this is *no*, then it appears even more inappropriate to consider surprisal from these language models to be reflective of information status.

We performed our analyses using simple linear regression models in Python, where feature interactions had to be manually programmed. But we see some variables which would be more appropriately modeled as mixed effects (for example, frequency could be grouped by UPOS), and a mixed effects regression may be more appropriate for this use case.

Finally, we may like to extend Loáiciga et al.'s probing experiment to the multilingual case and experiment more to see how well information status is encoded in the embeddings of large language models, separate from simple morphosyntactic cues. Information status may, as they found, be encoded in layers of the model, but not rise to the surface in surprisal, which is a single number derived only from the output layer.

7 Conclusion

Our experiments indicate that language model surprisal, while correlating well with many cognitive linguistic measures, is also well predicted by fairly mundane linguistic information such as parts of speech and word frequency, and only minimally by information status alone, making it problematic to use it as a stand-in for information status, which itself is also easily predicted by the same mundane syntactic information. This does not mean that information status is irrelevant to surprisal, but we do not see the effect shining through when simply looking at the surprisal of mentions. Though this may change with innovations in language models, we maintain that disentangling the effect of information status from that of more trivial and context-free linguistic cues is a must in evaluation.

Limitations

The generalisability of our study is limited by the use of a single language model and hyperparameter set. We originally set out to repeat this experiment over multiple context and stride sizes, but were unable to do this due to time and compute. Further experiments using different hyperparameters and/or models would be beneficial.

The imbalance between language data sizes in CiepInf is a problem, though the consistency of the results suggests that these findings are robust to data size.

Ethics Statement

We are unaware of any concrete harms towards individuals or communities arising as a result of our study.

We use a publicly available large language model, which we ran locally. We do not train or finetune this model. Our use of this model is within its intended use, and there is no possibility of personally identifiable information becoming available from our use, nor of harm to individuals or communities.

The corpus we use is available to share for the purposes of scientific study upon request, though it is not open-source due to copyrighted material contained within it.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

References

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Wallace L. Chafe. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Charles N. Li, editor, *Subject and Topic*, pages 25–55. Academic Press, New York.

Andrea de Varda and Marco Marelli. 2022. The effects of surprisal across languages: Results from native and non-native reading. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 138–144, Online only. Association for Computational Linguistics.

Andrew Dyer, Ruveyda Betul Bahceci, Maryam Rajestari, Andreas Rouvalis, Aarushi Singhal, Syahidah Asma Umniyati Yuliya Stodolinska, and Helena Rodrigues Menezes de Oliveira Vaz. 2024. A multilingual parallel corpus for coreference resolution and information status in the literary domain. In *Proceedings of the 22nd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2024)*, Hamburg, Germany. Association for Computational Linguistics.

Talmy Givón. 1983. Topic continuity in discourse: An introductions. *Topic continuity in discourse: A quantitative cross-language study.*

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380. Publisher: Nature Publishing Group.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive Status and the Form of Referring Expressions in Discourse. *Language*, 69(2):274–307. Publisher: Linguistic Society of America.

Sharid Loáiciga, Anne Beyer, and David Schlangen. 2022. New or old? exploring how pre-trained language models represent discourse entities. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 875–886, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Maciej Ogrodniczuk, Michal Novak, Massimo Poesio, Sameer Pradhan, and Vincent Ng, editors. 2025. *Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference*. Association for Computational Linguistics, Suzhou, China.

Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663, St. Julian's, Malta. Association for Computational Linguistics.

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden. Association for Computational Linguistics.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Robyn Speer. 2022. wordfreq: v3.0.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Annemarie Verkerk and Luigi Talamo. 2024. mini-CIEP+ : A shareable parallel corpus of prose. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*, pages 135–143, Torino, Italia. ELRA and ICCL.

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.

Weijie Xu and Richard Futrell. 2024. Syntactic dependency length shaped by strategic memory allocation. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 1–9, St. Julian's, Malta. Association for Computational Linguistics.

A Appendix

Appendix material may be provided in the camera-ready submission.