What does surprisal have to do with information status?

Andrew Dyer

Language Science and Technology
Universität des Saarlandes
Germany

andrew.dyer@uni-saarland.de

IDeaL
SFB 1102

UNIVERSITÄT
DES
SAARLANDES

TL;DW

- Suprisal from transformer language models is tempting to use as a continuous measure of **information status**. *However*, we find...
- Surprisal is barely at all predicted by information status; more trivial linguistic information such as parts of speech and word frequency are much better predictors.
- Information status itself is well predicted by parts of speech, which provides an intuitive baseline.

Introduction I

- Language model surprisal often used as a measure of difficulty of processing language (Goldstein et al., 2022; Wilcox et al., 2023).
- More surprising tokens held to correspond to more difficult or contentful units of speech.
- Often extended to "novel and unexpected" information (Xu and Futrell, 2024), with implicit linking hypothesis that new information is more surprising.

Introduction II

- Transformer-based language models use specific attention to previous, long-distance context to model next-word probabilities.
- This makes them promising to be able to look back to determine whether a mention is of a referent that is **given** (has been mentioned before) or **new** (is being introduced for the first time).
- Loáiciga et al. (2022) probed two large language models to test whether their hidden layer parameters could be used to predict the information status of mentions in English – and found that they could.

Research questions

1. As a preliminary, is information status itself is predicted by context-free features?
2. Does language model surprisal respond to the information status of mentions, as opposed to context-free features?

Experiment structure

We conduct the following regression experiments:

1. **Logistic regression:** UPOS and deprel as predictors of information status. Reported measure: F1-score.
2. **Poisson regression:** UPOS, deprel, information status, and word frequency as predictors of surprisal. R2 score is reported in this experiment.

Data and resources

Data

We use CiepInf (Dyer et al., 2024) for our experiments.

- Parallel multilingual corpus of coreference resolution and information status.
- CorefUD format, similar annotation scheme to GUM.
- Data currently annotated in Chinese, English, Hindi, Indonesian, Portuguese, Turkish, Ukrainian.

Model

- **mGPT** (Shliazhko et al., 2024) – multilingual autoregressive model
- Context size = 512, Stride = 256

Results: Experiment I

	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: 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.

Results: Experiment I

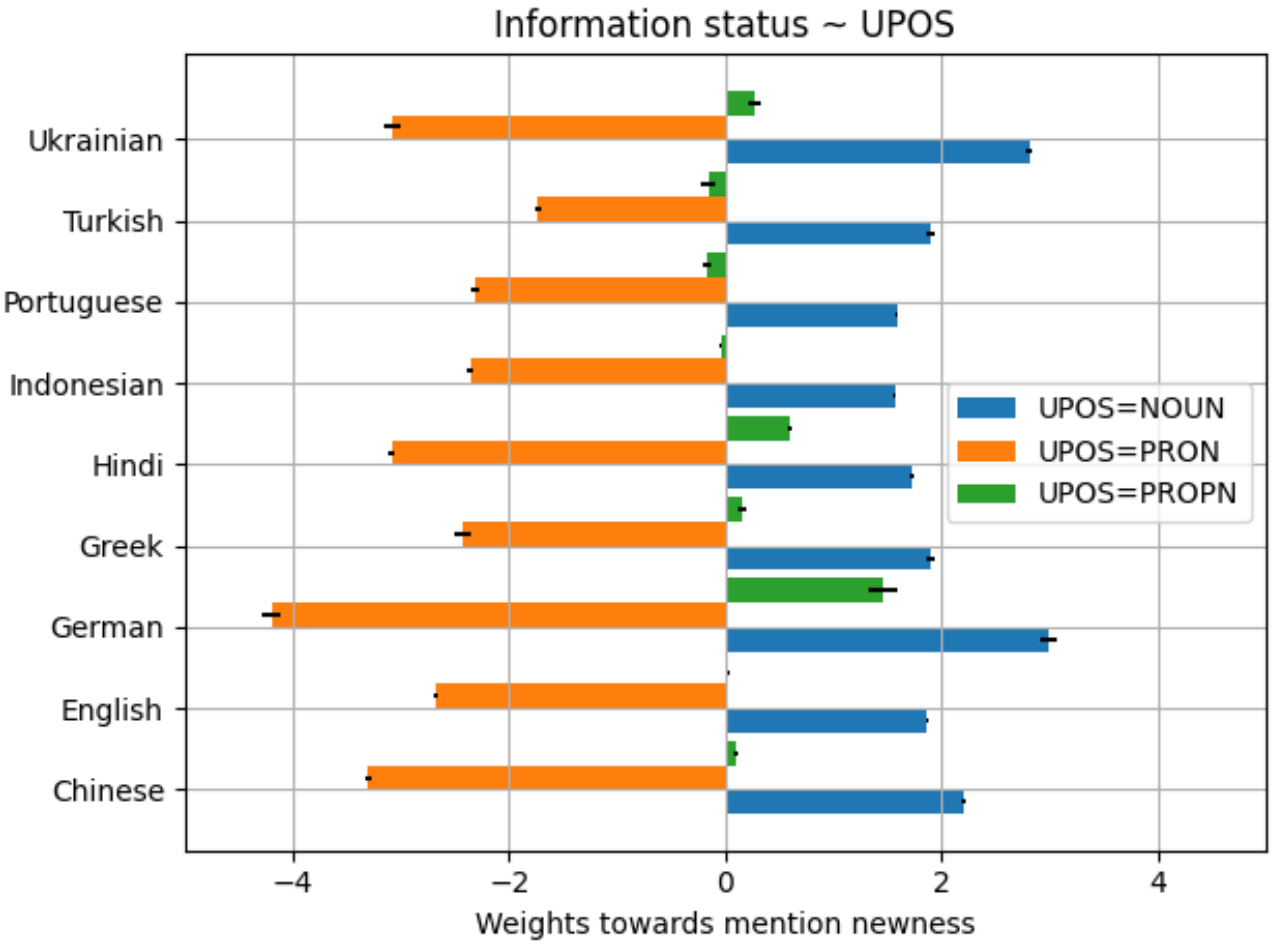

Figure: Weights towards information status newness by UPOS per language.

Results: Experiment II

	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: 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.

Conclusion

- Despite correlating well with other cognitive linguistic measures, language model surprisal is not a reliable continuous measure of information status.
- More context-free information such as parts of speech and word frequency are much better predictors of surprisal, and parts of speech in particular are strong predictors of information status.
- Disentangling information status itself from this context-free information is a must, as with other predictors.

End

Thank you!

Something

InfStat	UPOS		
	PRON	PROPN	NOUN
given	12123	1809	4488
new	369	840	13145

References

Andrew Dyer, Ruveyda Betul Bahceci, Maryam Rajestari, Andreas Rouvalis, Aarushi Singhal, Syahidah Asma Umniyati Yuliya Stodolinska, and Helena Rodrigues Menezes de Oliveira Vaz. 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, December 2024. Association for Computational Linguistics.

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. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3): 369–380, March 2022. ISSN 1546-1726. doi: 10.1038/s41593-022-01026-4. URL https://www.nature.com/articles/s41593-022-01026-4. Publisher: Nature Publishing Group.

Sharid Loáiciga, Anne Beyer, and David Schlangen. New or Old? Exploring How Pre-Trained Language Models Represent Discourse Entities. In Nicoletta