Evaluating the Interplay of Information Status and Information Content in a Multilingual Parallel Corpus

Julius Steuer[2], **Andrew Dyer**[1], **Toshiki Nakai**[1],
Luigi Talamo[1], **Annemarie Verkerk**[1]
[1]Department of Language Science and Technology Saarland University
[2]Heidelberg Institute for Theoretical Studies
Correspondence: julius.steuer@h-its.org

Abstract

The uniform information density (UID) hypothesis postulates that linguistic units are distributed in a text in such a way that the variance around an average information density is minimized. The relationship between information density and information status (IS) is so far underexplored. In this ongoing work, we project IS annotations on the English section of the CIEP+ corpus (Verkerk and Talamo, 2024) to parallel sections in other languages. We then use the projected annotations to evaluate the relationship between IS and information content in a typologically diverse sample of languages. Our preliminary findings indicate that there is an effect of information status on information density, with the directionality of the effect depending on language and part of speech.

1 Introduction

Research on information status (IS) and information content has the same aim: assessing how information is distributed across words in sentences and larger discourse units; sometimes, special attention is paid to word order and communicative efficiency, e.g., Tsipidi et al. (2024). IS refers to whether a listener or speaker should look for the referents of a word or phrase among a set of already mentioned or given entities, or add a new entity to this set (Chafe, 1976), e.g., in centering theory (Gundel, 1997). IS has an effect on the placement of words: new items are preferentially placed earlier in the utterance than given items (Clark and Clark, 1978). As an example, consider the following sequence of sentences:

1.a The Hobbit, or there and back again.

1.b In a hole in the ground there lived a hobbit.

1.c It was a hobbit-hole, and that means comfort.

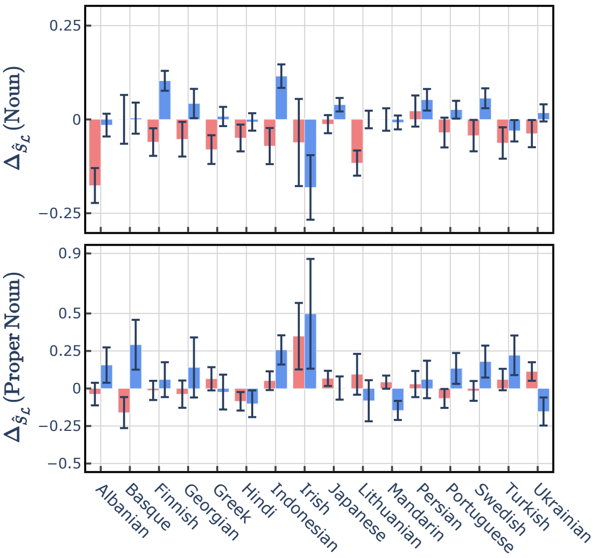

Figure 1: Mean deviation $\Delta_{\hat{S}_\mathcal{L}}$ of surprisal from channel capacity $\hat{S}_\mathcal{L}$ for nouns and proper nouns for new and given items. Error bars show standard error.

Here a new entity *hole* is mentioned at the beginning of **1.b**, and referred to by co-referring pronoun *it* at the beginning of **1.c**. This coreference relationship establishes the entity referred to by *it* as given. *Hobbit* in **1.b** is already given (as it was already mentioned in **1.a**), and thus placed at the end of **1.b**.

While information status refers to the mental state of the speaker or listener *prior* to encountering a new or given word, information content, measured in bits per item (Shannon, 1948), as the basis of surprisal theory (Hale, 2001; Levy, 2008) refers to the amount of information gained by the listener *after* encountering it. Surprisal, like IS, has been shown to impact word and morpheme order (e.g., Hahn et al. 2021 and Cuskley et al. 2021): *Hobbit* may have been placed at the end of **1.b** because (for a first-time reader of *The Hobbit*), this word would have high surprisal.

Thus, surprisal theory and IS use the term "information" in different ways, which becomes clear

Proceedings of the 8th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, pages 18–25
March 29, 2026 ©2026 Association for Computational Linguistics

if we try to reformulate the content of each in the verbiage of the other: Surprisal measures *how* new a word is in context *after* it has been encountered, as the information gained at each word is by definition new information. IS, in contrast, assigns one of two or more labels to a mental list of entities that are part of a discourse, with new entities expected to yield more information (in bits) than given entities. To our knowledge, this conjectured relationship between surprisal and IS has not been evaluated before. In this work, we are taking a first step towards an evaluation of this relationship in a multilingual setting.

Our starting point is the uniform information density hypothesis (UID) (Levy and Jaeger, 2007), which posits that speakers prefer sentences in which the information content of words stays close to their channel capacity (Collins (2014) inter alia), that is, the average information rate at which language transmission occurs. In order to evaluate UID and its interaction with IS in a multilingual setting, we use the English section of miniCIEP+ (Verkerk and Talamo, 2024) annotated for IS by (Dyer et al., 2024) to automatically annotate parallel data in miniCIEP+.

The remainder of the paper is structured as follows: Section 2 introduces a formal definition of the surprisal of new and given items, and briefly reviews existing work on UID and the interaction of information content and word order. Section 3 describes the annotation projection from English to other languages. In Section 4, we evaluate the annotation projection and discuss preliminary results: we find that while surprisal of nouns stays close to channel capacity independent of IS, surprisal of given pronouns consistently falls below the channel capacity. For new proper nouns, we find that in some languages surprisal on the average falls below channel capacity, while in others, channel capacity is usually exceeded.

2 Surprisal, information status and UID

As already laid out in the introduction, from the vantage point of surprisal theory, we expect new and given items to behave in fundamentally different ways: Since new items introduce entities to the discourse, they should be less predictable from context and their surprisal should be higher than that of given items. UID, in contrast, predicts that new and given items are distributed in such a way that their surprisal deviation from channel capacity is minimized. In this section, we will introduce formal definitions of surprisal of new and given items, and a corresponding operationalization of UID.

2.1 Surprisal of new and given entities

Following the nomenclature of Dyer et al. (2024), we express that an entity $\mathbf{e}$ was *mentioned* in the discourse segment or preceding context $\mathbf{c}$ as $\mathcal{M}(\mathbf{c}, \mathbf{e})$, and the opposite case, i.e., if $\mathbf{e}$ was not mentioned in $\mathbf{c}$ and thus has IS new, as $\neg\mathcal{M}(\mathbf{c}, \mathbf{e})$.

We now want to formalize our intuition about the information content of new and given entities. We expect that the information gained by observing a string representation $\mathbf{s}$ of mentions of entity $\mathbf{e}$ given $\mathbf{c}$ will be lower if condition $\mathcal{M}(\mathbf{c}, \mathbf{e})$ holds, i.e., that on average already mentioned entities receive lower surprisal S:

$$S[\mathbf{s}|\mathbf{c}, \mathcal{M}(\mathbf{c}, \mathbf{e})] \leq S[\mathbf{s}|\mathbf{c}, \neg\mathcal{M}(\mathbf{c}, \mathbf{e})] \qquad (1)$$

Our intuition is based on the fact that, by definition, there must be *some* information about (a referent of) $\mathbf{e}$ in $\mathbf{c}$ if $\mathbf{e}$ is given, but not necessarily if $\mathbf{e}$ is new. However, this does not mean that the context $\mathbf{c}$ is entirely devoid of information about entity $\mathbf{e}$ if it is new, for example in bridging entities (*I went to the hospital. The doctor took my blood pressure.*). It is this relationship between an entity's IS and the relevance of its context in predictive processing that we want to unravel:

Research Question

Do given items receive lower surprisal than new items?

2.2 UID and information status

As already mentioned in Section 1, UID makes a very different prediction for the average surprisal of new and given entities: All else equal, a speaker will accommodate words in such a way that the variance of surprisal from the channel capacity[1] is minimized, and IS is only one of the factors at play. Thus, from the vantage point of UID, we would expect a different relationship between new and given entities, i.e., that on average the effects of newness and givenness cancel out:

$$S[\mathbf{s}|\mathbf{c}, \mathcal{M}(\mathbf{c}, \mathbf{e})] \approx S[\mathbf{s}|\mathbf{c}, \neg\mathcal{M}(\mathbf{c}, \mathbf{e})] \qquad (2)$$

UID has been shown to hold for English both from the perspective of dependency length (Collins,

[1] Of the speaker, but also the listener CITE

2014) and word order (Cuskley et al., 2021). However, both methods calculate the deviation from channel capacity as the word-to-word variance of unigram surprisal. In contrast, we estimate surprisal from a causal transformer language model (see Appendix B for details).

A more comprehensive evaluation of different formalizations of UID is offered by Meister et al. (2021), who systematically vary the scope with which the channel capacity is calculated, finding that the language level aligns better with human data. Based on their work, for a language $\mathcal{L}$, we will calculate channel capacity $\hat{S}_{\mathcal{L}}$ on the language level as average surprisal over all N words in that language's section of miniCIEP+. We refer to this formalization of UID as UID$_{\mathcal{L}}$:

$$\mathrm{UID}_{\mathcal{L}} = \frac{1}{N-1}\sum_{i=2}^{N}(S(\mathbf{s}_i|\mathbf{c}) - \hat{S}_{\mathcal{L}})^2 \qquad (3)$$

Here, $\mathbf{c}$ expands into a prefix of words of fixed length T, $\mathbf{c} = \mathbf{s}_{i-T-1}, ..., \mathbf{s}_{i-1}$. We can now rephrase our initial research question it in terms of UID$_{\mathcal{L}}$:

Research Question*

Does UID$_{\mathcal{L}}$ hold for given and new items, i.e., is there a difference in surprisal between new and given items when viewing deviation from channel capacity?

3 Annotation projection on miniCIEP+

3.1 Data

We start out from English CiepInf (Dyer et al., 2024), which comprises a subcorpus of miniCIEP+ (Verkerk and Talamo, 2024). We use the annotations for IS on the English to automatically annotate other languages in miniCIEP+ by projecting English IS labels to word-aligned parallel data. Although there are more fine-grained notions of IS (e.g., Gundel et al. 1993; Markert et al. 2012), we restrict ourselves to a dichotomy between new and given entities for the sake of simplicity. We evaluate the projected annotations by comparing them to hand-aligned gold-standard data from CiepInf.

3.2 Projection

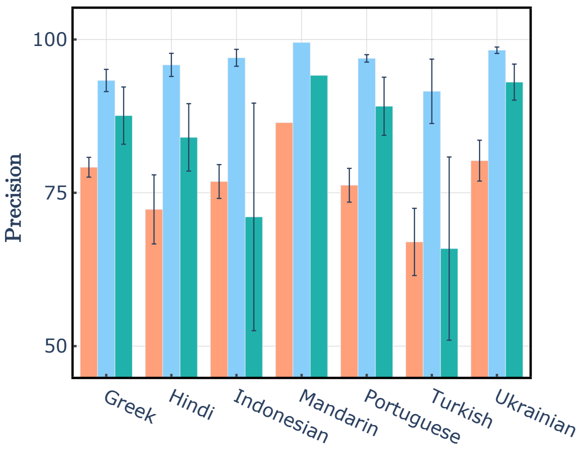

Figure 2: Precision of information status projection on gold data for nouns, pronouns, and proper nouns per language. We calculate precision as accuracy on words that have a label in the gold and projected data. Error bars show standard error over books.

In CiepInf, entity tags are associated with the spans of noun phrases (NPs), and each entity is annotated with an IS label. Because our projection operates at the token level, we extract only the syntactic head token of the span. We then project IS labels from English to each target language in two stages. We first obtain sentence-level alignments with Bertalign (Liu and Zhu, 2023). Then, for every aligned sentence pair, we compute word alignments with awesome-align (Dou and Neubig, 2021). Whenever an aligned English token carries an IS label, we copy that label to its aligned target token(s).

We do not fine-tune either aligner for any language. While Bertalign officially supports 25 languages and awesome-align reports evaluation on five languages, both approaches are based on LaBSE (Feng et al., 2022) and mBERT (Devlin et al., 2019) respectively and therefore remain applicable to a wider set of languages, though quality may degrade for typologically distant and/or lower-resource languages. We therefore evaluate projection quality on 7 languages with gold IS labels taken from CiepInf (Chinese, Greek, Hindi, Indonesian, Portuguese, Turkish, Ukrainian).

Evaluation with gold labels To measure intrinsic projection quality, we evaluate token-level agreement on the subset of tokens that have *both* a gold IS label and a projected IS label:

$$\text{Precision} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\hat{y}_i = y_i]\,,$$

where y_i is the gold IS label, $\hat{y}_i$ is the projected label, and N is the number of evaluated tokens (intersection of gold-labeled and projected-labeled tokens). $\mathbb{I}$ denotes the indicator function that gives

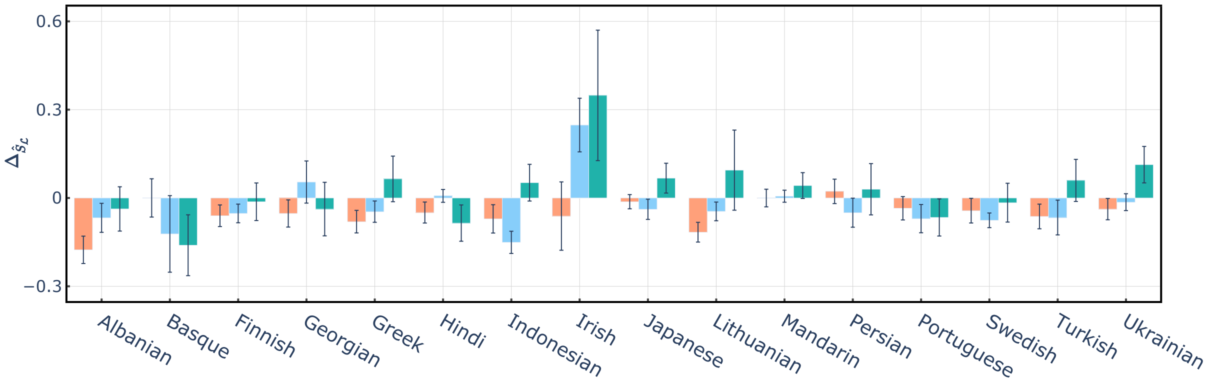

Figure 3: Mean deviation $\Delta_{\hat{S}_{\mathcal{L}}}$ of surprisal from channel capacity $\hat{S}_{\mathcal{L}}$ for given **nouns**, **pronouns**, and **proper nouns** per language. Error bars show standard error. Mandarin lacks error bars because the evaluation is based on a single book.

1 if the condition is satisfied, and 0 if not.

Evaluation without gold labels For unlabeled languages, we quantify how frequently projection assigns IS labels by computing (per POS category defined in miniCIEP+) the proportion of tokens that receive any projected IS label:

$$\mathrm{coverage}(\mathrm{pos}) = \frac{N_{\mathrm{projIS}}(\mathrm{pos})}{N(\mathrm{pos})}.$$

where $N(\mathrm{pos})$ is the number of tokens in POS category pos, and $N_{\mathrm{projIS}}(\mathrm{pos})$ is the number of such tokens that have received an IS annotation tag through projection.

Expected label sparsity Since English is the sole source of supervision and projection is one-way (English → target), projected corpora typically contain fewer IS labels than the English gold annotations. Additional label loss arises when sentence/word alignment fails to link an English labeled token to a target token, compounding effects from linguistic distance and aligner quality.

4 Preliminary results

4.1 Projection quality

Figure 2 reports the precision of projected IS labels against gold annotations. **Pronouns** achieve consistently high precision across all evaluated languages, which is expected because pronouns are strongly biased towards given, and their closed-class nature makes it easier to align them. In contrast, common **nouns** show substantially lower precision in Hindi and Turkish. A plausible explanation is increased syntactic divergence from English (both are predominantly SOV), which can degrade word-alignment quality and increase noise in projection. **Proper nouns** exhibit the largest cross-linguistic variability, potentially reflecting differences in orthographic conventions, named-entity formation, and language-specific usage patterns that affect both their presence in the target text and their potential alignment. We evaluate the quality of the projected annotations to languages for which we did not collect gold data in Appendix A.

4.2 Surprisal and information status

For our preliminary analysis, we report results for the 7 languages for which we collected gold data, and 9 more languages we selected based on their genealogical diversity: Albanian, Basque, Finnish, Georgian, Irish, Japanese, Lithuanian, Persian, and Swedish. We excluded new pronouns from our analysis because true cataphora, i.e., forward-pointing personal pronouns as in

2. As she was searching for her$_{\mathrm{M}_1}$ hat, Mary$_{\mathrm{M}_2}$ came upon a long-lost letter.

are rare; consequently, new pronouns are rare and almost exclusively equivalents of ‘something’, ‘nobody’, and other such indefinite pronouns.

UID$_{\mathcal{L}}$ Comparing new and given words in this set of languages, we find that surprisal of **nouns** usually stays close to channel capacity independently of IS, while there is an apparent effect of IS on the surprisal of **proper nouns**: Figure 1 (on first page) shows that the surprisal of new proper nouns exceeds channel capacity in Basque, Indonesian, Portuguese, Swedish and Turkish, while it falls below channel capacity in Ukrainian. Thus, nouns are more conformant to $\text{UID}_{\mathcal{L}}$ than proper nouns, whose usage may be influenced by other considerations than $\text{UID}_{\mathcal{L}}$.

Givenness Figure 3 shows that for all languages in our sample, surprisal of given nouns and pronouns consistently falls below channel capacity, while that of proper nouns usually exceeds it. Together with our finding for $\text{UID}_{\mathcal{L}}$, this indicates a differing use of proper nouns as per the *mediated* IS in Markert et al. (2012), a distinction which is not reflected in CiepInf.

5 Conclusion

In this work, we laid out our approach to annotate miniCIEP+ for IS by projecting gold annotations from English to parallel data in other languages and presented preliminary results on the interaction of IS and UID. While our results point to an interaction of IS and UID, the quality of the projected annotations is affected by the quality of alignment models, the improvement of which will be a first step to higher-quality projections and further analysis.

Limitations & future work

The work presented in this paper relies heavily on quality of projected annotations, which in turn relies on the quality of BertAlign (Liu and Zhu, 2023) and awesome-align (Dou and Neubig, 2021). We applied these tools to our parallel corpus without any modifications, yielding a lossy projection as evident from Figure 4, e.g., for Arabic, Irish and Latin. We want to address this performance gap by fine-tuning both alignment models.

Secondly, the IS annotations in CiepInf are rather coarse-grained: We only distinguish new and given words, while most IS schemata have intermediate states (e.g., brand-new vs. new and mediated vs. new/given) or distinguish between the hearer/reader and the discourse level (Prince, 1992).

Lastly, the distinction between reader and discourse is only meaningful if the reader has some prior knowledge about the world in which the discourse is situated. While this is arguably the case in a multilingual language model like mGPT (see, e.g., (Li et al., 2021)), it is not possible to know *beforehand*, e.g., if and how the string *Alice* is associated with the main character in *Through the Looking Glass*, i.e., what is part of the discourse the language model is aware of.

References

Wallace L Chafe. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. *Subject and topic*. Publisher: Academic Press.

Eve Vivienne Clark and Herbert H. Clark. 1978. Universals, relativity, and language processing. In Joseph H. Greenberg, editor, *Universals of human language. 1: Method & theory*. Stanford Univ. Press, Stanford, California. Num Pages: 286.

Michael Xavier Collins. 2014. Information Density and Dependency Length as Complementary Cognitive Models. *Journal of Psycholinguistic Research*, 43(5):651–681.

Christine Cuskley, Rachael Bailes, and Joel Wallenberg. 2021. Noise resistance in communication: Quantifying uniformity and optimality. *Cognition*, 214:104754.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Andrew Dyer, Ruveyda Betul Bahceci, Maryam Rajestari, Andreas Rouvalis, Aarushi Singhal, Yuliya Stodolinska, Syahidah Asma Umniyati, 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 Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 55–64, Hamburg,Germany. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Jeanette K. Gundel. 1997. Centering theory and the givenness hierarchy: Towards a synthesis. In *Centering Theory In Discourse*. Oxford University Press.

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.

Michael Hahn, Judith Degen, and Richard Futrell. 2021. Modeling word and morpheme order in natural language as an efficient trade-off of memory and surprisal. *Psychological Review*, 128(4):726–756.

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*.

Carina Kauf and Anna Ivanova. 2023. A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

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. In Bernhard Schölkopf, John Platt, and Thomas Hofmann, editors, *Advances in Neural Information Processing Systems 19*, pages 849–856. The MIT Press.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online. Association for Computational Linguistics.

Lei Liu and Min Zhu. 2023. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 795–804, Jeju Island, Korea. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ellen F. Prince. 1992. The ZPG Letter: Subjects, Definiteness, and Information-status. In William C. Mann and Sandra A. Thompson, editors, *Pragmatics & Beyond New Series*, volume 16, page 295. John Benjamins Publishing Company, Amsterdam.

C. E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.

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.

Eleftheria Tsipidi, Franz Nowak, Ryan Cotterell, Ethan Wilcox, Mario Giulianelli, and Alex Warstadt. 2024. Surprise! Uniform Information Density isn't the whole story: Predicting surprisal contours in long-form discourse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18820–18836, Miami, Florida, USA. Association for Computational Linguistics.

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.

A Projection coverage on unlabeled data

Figure 4 shows POS-wise coverage of projected IS on the unlabeled portion of the corpus. We observe markedly lower coverage for languages such as Arabic, Armenian, Georgian, Irish, Kurmanji, Latin, Urdu, and Welsh, compared to several higher-resource languages. This pattern is consistent with noisier or sparser alignments (e.g., reduced lexical overlap and weaker cross-lingual representations), which results in fewer confidently aligned tokens and therefore fewer projected IS instances.

B Surprisal estimation

We calculated surprisal on all words that received an IS annotation with mGPT (Shliazhko et al., 2024) using the minicons[2] Python library (Kauf and Ivanova, 2023). For all languages we ran mGPT concurrently on 4 H100 GPUs, with a batch size of 128 per language and a fixed context size of $T = 256$. Code and data needed to run the experiments will be made available on GitHub upon publication. For each language, we calculate channel capacity $\hat{S}_{\mathcal{L}}$ as average surprisal over all words in that language's section of the corpus.

C Corpus statistics

Table 1 reports corpus statistics for all languages considered in this study. For Chinese, Greek, Hindi, Indonesian, Portuguese, Turkish, and Ukrainian, surprisal is computed using manually annotated Information Status labels. For the remaining nine languages, we rely on automatically projected annotations in order to broaden the typological and script diversity of the dataset.

[2] https://github.com/kanishkamisra/minicons

Language	**IS (NOUN)**	**IS (PRON)**	**IS (PROPN)**	**#Tokens**	**#Sents**
Human annotation					
Chinese	1,296	484	257	10,286	370
Greek	736	827	59	18,100	892
Hindi	2,558	3,281	558	31,591	1,853
Indonesian	1,331	568	227	8,923	540
Portuguese	4,955	1,647	763	46,607	2,803
Turkish	846	196	160	6,983	499
Ukrainian	1,140	791	211	11,257	838
Projection					
Albanian	11,465	3,449	1,956	133,690	6,401
Arabic	7,911	3,068	4	106,319	7,547
Basque	4,482	310	671	55,120	3,949
Finnish	11,101	5,483	1,755	101,193	6,169
Georgian	4,454	830	548	44,337	2,225
Irish	1,447	849	161	20,535	958
Japanese	14,650	2,758	1,793	199,575	6,637
Lithuanian	11,987	4,604	573	105,226	6,800
Persian	12,509	3,694	1,312	129,444	5,317
Swedish	12,322	10,004	1,975	123,708	5,970

Table 1: Corpus statistics by language. Rows are grouped by annotation type (human annotation vs. automatic projection). IS columns report the number of tokens annotated with Information Status for each POS category (noun, pronoun, proper noun). #Tokens and #Sents denote the total number of tokens and sentences in the corpus, respectively.

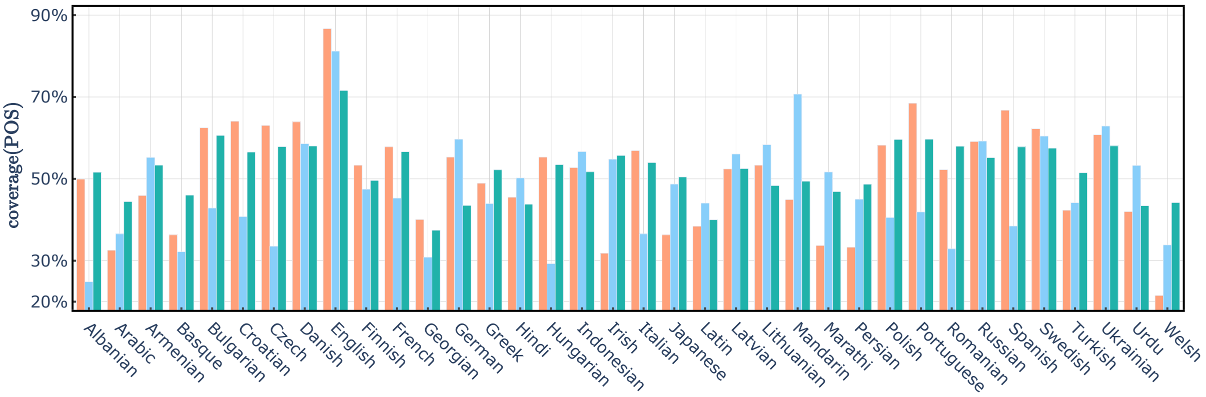

Figure 4: Percentage of **nouns**, **pronouns**, and **proper nouns** that received a POS tag through annotation projection, per language.