Lexical Surprisal from Large Language Models: Cautionary Notes for Legal Discourse

Silvie Cinková, Ivana Kvapilíková, Jan Černý
Faculty of Mathematics and Physics
Charles University, Czech Republic
{cinkova, kvapilikova}@ufal.mff.cuni.cz, cerny@matfyz.cz

Abstract

Lexical surprisal correlates with human reading times and is easy to extract from large language models (LLMs), thus appearing useful for readability assessment. However, recent evidence suggests that LLMs tend to assign very low surprisal to texts that are demonstrably difficult for lay readers. In this paper, we investigate the stability and genre sensitivity of lexical surprisal estimates for Czech across five language models. Using diverse single-genre corpora, we analyze word-level surprisal distributions and inter-model correlations, observing inter-model agreement to be generally low and strongly modulated by genre. Detailed distributional analyses reveal sharp differences between generative and n-gram models, particularly in legal genres, where Transformer-based models assign disproportionately low surprisal to technical nouns. Correlations between lexical surprisal and TF-IDF further suggest that large models are more sensitive to structural patterns than to domain-specific lexical difficulty. These findings caution against using lexical surprisal from general-purpose LLMs for readability assessment in institutional and legal texts, highlighting the need for genre-aware and cognitively grounded evaluation.

Keywords: readability; lexical surprisal; language models

1 Introduction

Lexical surprisal is a linguistic concept adopted from Information Theory (Shannon, 1948), which models the predictability of a text by determining the probability of each word given its preceding context. There is ample empirical evidence of lexical surprisal reflecting the cognitive load on a reader, operationalized as reading time (Levy, 2008; Hale, 2001; Smith and Levy, 2013; Wilcox et al., 2023).

Assuming that there is an optimum pattern of surprisal distribution (Frank and Jaeger, 2008; Clark et al., 2023), lexical surprisal lends itself as an aid to automatic readability assessment.

Novotná et al. (2025); Černý (2025) describe their experiments with lexical surprisal (computed by GPT-2) to assess the readability of Czech legal and administrative texts in an application. Interestingly, their initial observations did not reveal an interpretable pattern. The token-wise values were uniformly very low, suggesting that the texts would be both information-light and easy to read, which could by no means comply with the judgment of a common reader. We speculate that GPT-2 is too well-trained in legal writing to model the mental lexicon of a legal layman. This initial finding has motivated our current research questions:

1. Do language models agree on lexical surprisal values in Czech texts?

2. Is the agreement affected by text genre?

2 Related work

We build on prior work that established the relationship between lexical surprisal and text readability, using probabilistic language models to quantify surprisal (Demberg and Keller, 2008; Pimentel et al., 2023; Sun and Wang, 2025). Whereas these studies confirm the correlation between word surprisals and their reading times in the English language, Wilcox et al. (2023) extend the evidence for surprisal theory to multiple languages.

Fernandez Monsalve et al. (2012); Goodkind and Bicknell (2018) postulate that the predictive power of word surprisal on reading times increases with the decreasing model perplexity. In contrast, Oh and Schuler (2023) demonstrate that for Transformer models, the predictive accuracy declines

when the model was exposed to too much data (over 2 billion tokens) during training.

Regarding surprisal across different text styles, Gehrmann et al. 2019 use word predictability to discern natural texts from automatically generated ones using the fact that the latter are based on probabilistic distribution samples and hence maintain particularly low surprisal levels.

3 Experiment setup

In this work, we focus exclusively on the Czech language and analyze the surprisal distributions across different genres.

3.1 Surprisal Computation

We prepared seven Czech corpora to represent different genres (see 4). We ran each language model from our selection (as listed in 3.2) to compute the lexical surprisal on each text, using an earlier in-house re-implementation of *gltr* (Gehrmann et al., 2019; Černý, 2025). Word surprisal is approximated as the negative log probability assigned to each word by the model given its preceding context. For Transformer models, the context is limited to the previous 512 tokens, whereas n-gram models, by construction, condition only on the four preceding words.

As a next step, we mapped the model-specific subword tokens on word boundaries as defined by a Czech tokenizer (Straka et al., 2016), transforming the lexical surprisal to word-wise sums of their negative log probabilities (Pimentel and Meister, 2024).

3.2 Language models

When selecting language models, we included two Transformer models trained on Czech data and one multilingual model. Suspecting that heavily trained Transformer models may exhibit *super-reader* capabilities, as observed by Oh and Schuler (2023), we also constructed two n-gram models on datasets over which we have greater control. The model overview is displayed in table 1.

Model	Params	Train Tokens	Language
gpt2_small	117M	5B	cs
gpt2_XL	1.5B	11B	cs
qwen	7B	18T	multiling
ngram_web	–	1B	cs
ngram_cnk	–	5B	cs

Table 1: Overview of language models used in the experiments

3.2.1 Qwen2.5

Qwen2.5 (Qwen et al., 2025) represents one of the most advanced multilingual language models. The base models were pretrained on a large-scale, high-quality mixture of web text, books, code across languages and curated synthetic datasets totaling 18 trillions of tokens. No fine-tuning for Czech was performed. We used the variant with 7 billion parameters which already has the ability to generate texts in the Czech language and its size is feasible for running inference on a GPU with 16 GB of GPU VRAM.

3.2.2 Czech-GPT-2-XL-133k

Czech-GPT-2-XL-133k (*gpt2_XL*) was trained on a 11-billion word corpus of Czech obtained by web crawling. Its architecture was adapted from GPT-2 XL by replacing the tokenizer, corresponding embeddings, and copying over 1,000 English representations corresponding to the 1,000 most frequent tokens into new embeddings based on a bilingual dictionary. (BUT-FIT, 2024).

3.2.3 General GPT-2 Czech

The General GPT-2 Czech model (Chaloupský, 2022) (*gpt2_small*) represents a first-generation model fine-tuned to Czech. It is based on the English pre-trained GPT-2 (Radford et al., 2019), fine-tuned with a Czech tokenizer and word embeddings, and trained on the Czech subset of the OSCAR dataset (Ortiz Su'arez et al., 2020). The dataset is derived from the Web Crawl corpus and the authors filtered the output to only include longer web articles (over 1,200 characters).

3.2.4 n-gram models: NewsCrawl and CNC

For this experiment, two n-gram models were trained to provide contrast to the neural models: *ngram_web* and *ngram_cnc*. The NewCrawl[1] corpus consists of texts crawled between the years 2007 and 2019 from online news sites. The Czech National Corpus (CNC) is the public release of the balanced Czech corpus, where the texts are divided into randomly shuffled chunks of approximately 100 words (respecting sentence boundaries), for copyright reasons[2].

[1] https://data.statmt.org/news-crawl/cs/

[2] https://wiki.korpus.cz/doku.php/en:cnk:syn:verze9

4 Data

4.1 Czech Documents Corpus (*newswire*)

The Czech Text Document Corpus v 2.0 (Král and Lenc, 2018) is a collection of newswires and press releases provided by the Czech News Agency, spanning more than sixty labeled topics. For our experiment, we used a random sample of 500 documents, which we dub *newswire*.

4.2 Prague Dependency Treebank Consolidated 2.0 (*press*)

PDTC 2.0 (Hajič et al., 2024) consists mainly of Czech press texts from the 1990s and Czech translations of the Penn Treebank (Santorini, 1990) (Wall Street Journal articles), with a small proportion of narrative speech transcriptions. It represents a somewhat more diverse press corpus than the Czech Documents Corpus.

4.3 FicTree (*fiction*)

FicTree (Jelínek et al., 2017) consists of eight fiction texts (books) from the Czech National Corpus. The texts are shuffled in random chunks of a maximum of 100 words (respecting sentence boundaries). We divided the train section of its Universal Dependencies release (Zeman et al., 2025) into the original shuffled chunks and randomly selected 300 items.

4.4 Czech Court Decisions Corpus (*justice*)

The Czech Court Decisions Corpus (CzCDC) is a corpus of whole decisions from three top-tier Czech courts (Supreme, Supreme Administrative, and Constitutional court) (Novotná and Harašta, 2019). We used a random sample of 500 documents, which we dub *justice*. The typical text would contain an abundance of file reference numbers, addresses, dates, and numbered headings. New lines are used to emphasize the main points in the decision sentence. The register is extremely formal and practically unreadable for a legal layman (cf. Example 1, machine-translated).

(1) By the defendants decision (hereinafter also referred to as the complainant) of 6 May 2008, Ref. No. OAM347/VL07112008, the claimants application for the granting of international protection was rejected as manifestly unfounded pursuant to Section 16(2) of Act No. 325/1999 Coll., on Asylum and on Amendments to Act No. 283/1991 Coll., on the Police of the Czech Republic, as amended (hereinafter the Asylum Act).
The Regional Court decided on the costs of the proceedings by ruling that none of the parties is entitled to reimbursement thereof.

4.5 KUKY 1.0 (*legal_admin*)

KUKY 1.0 (Cinková et al., 2024) is a corpus of 226 diverse legal and administrative texts, mostly findings and communications of the Office of the Czech Public Defender of Rights, generic legal advice from the Frank Bold law firm, and various court and office filings. In the experiment, it is labeled as *legal_admin*. This corpus contrasts several legal writing styles, the distinction being readability (assessed by experts in plain legal language). Unlike court decisions in CzCDC, these documents also capture communication with the general public, often with observable intentions to be comprehensible to the layman recipient. However, they are still official documents written by legal experts to withstand scrutiny from peers (Example 2, machine-translated).

(2) I am responding to your submission dated 12 September 2011, which concerns the procedure followed by the Ministry of Justice in handling your application for compensation for damage pursuant to the Act on Liability for Damage Caused in the Exercise of Public Authority.
The said file was sent to me on 28 November 2011. In reviewing this file, I focused in particular on the question of whether delays occurred in your matter in decisionmaking under Section 419 of Act No. 40/2009 Coll., the Criminal Code, as amended (hereinafter also referred to as the Criminal Code).

4.6 ParCzech 4.0 (*parl_debates*)

The ParlaCzech 4.0 corpus (Kopp, 2024) is a corpus of transcribed debates in the Czech parliament, corresponding to the Czech subset of the ParlaMint corpora (Erjavec et al., 2025). We used a random sample of 500 items, called *parl_debates*.

4.7 CAC (*academic*)

The Czech Academic Corpus 2.0 (CAC) (Hladká et al., 2008) consists of 650,000 words from various newspapers, magazines, and TV broadcast

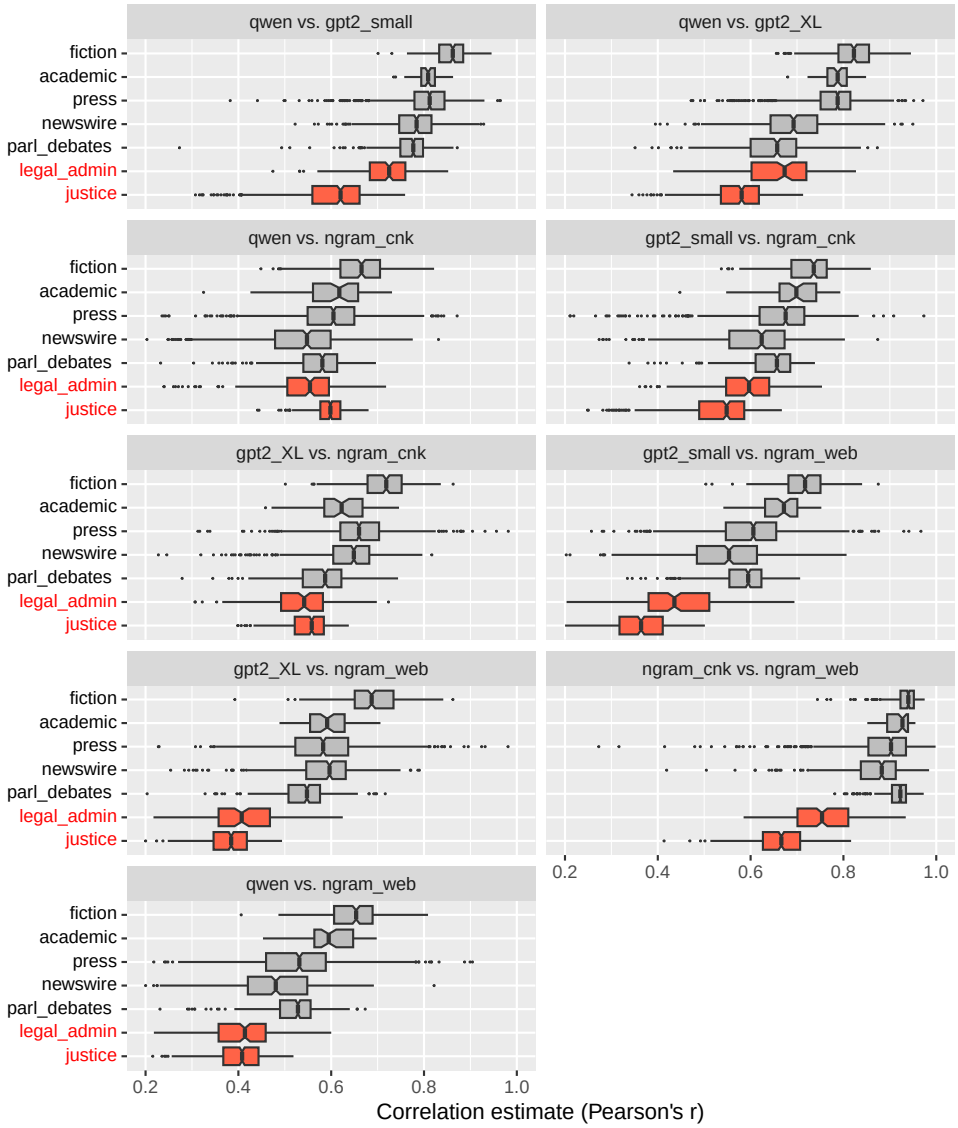

Figure 1: Correlations of lexical surprisal (negative log probability) in model pairs

transcripts gathered by a team of linguists between 1971 and 1985 to reflect the so-called "practical style". The "practical style" ("věcný styl") in Czech stylistics describes the consultative and formal register used in academia and public service journalism. We used all (68) documents labeled as "scientific" by the corpus creators, dubbing this sample *academic*.

5 Results

5.1 Pairwise correlations between models

We first analyzed the consistency of surprisal estimates across different models by examining pairwise correlations between model outputs. For each document within a pair of models, we obtained seven distributions of correlation measurements (Fig. 1). Each subplot represents one pair of models, e.g. *qwen* vs. *gpt2_XL*. Each boxplot represents one genre corpus (labeled on the Y-axis). Each data point underlying a boxplot represents one document in the given genre corpus. Its value is the Pearson correlation estimate of the word-wise aggregated lexical surprisal values by two models.

We observe that the correlations in Fig. 1:

1. are considerably low
2. differ between genres.

The genre-wise correlation distributions are significantly different from each other in each pair of language models (Kruskal-Wallis rank sum test). The post-hoc Dunn's test of pairwise differences

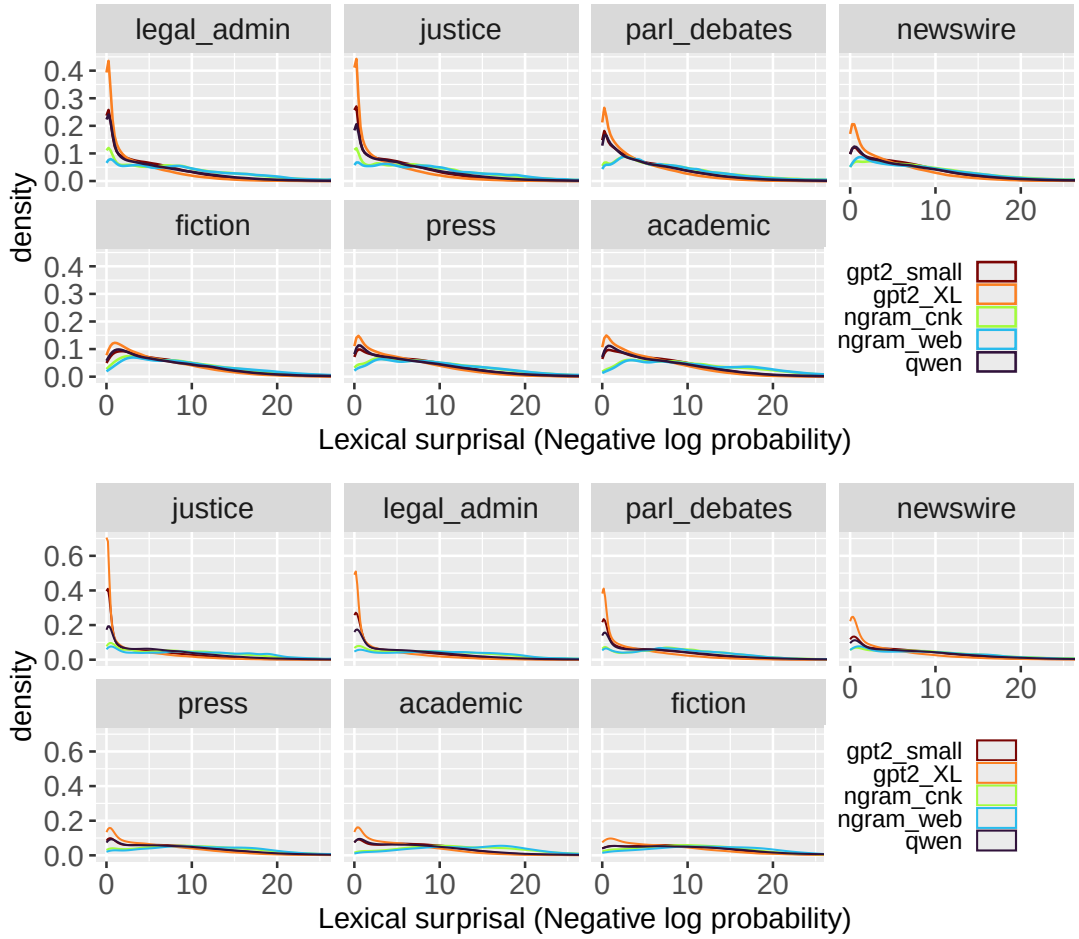

Figure 2: Distributions of negative log probabilities of lexical surprisals of all word types (top) and common and proper nouns (bottom) on individual corpora by different language models

between groups (with Holm's p-value adjustment) reveals statistical significance between all genres except *parl_debates* and *newswire*. This applies to all model pairs. Table 2 shows a prominent effect of genre on the correlation in all model pairs except *qwen* vs. *ngram_cnk*.

The genre ranking was almost consistent across the model pairs: the highest correlation occurred in the *fiction* corpus and the lowest in the corpora *legal_admin* and *justice*. The best correlating pair were the two Czech n-gram models, but even here there occurred a gap between the two legal corpora (max 0.75) and the other genres, which clustered around 0.9.

5.2 Comparison of the individual distributions

Although individual lexical surprisal distributions are all right-skewed across models and genres, a closer look reveals a difference between the large generative models (*gpt2_XL*, *gpt2_small* and

model pair	ε^2	lower ci	upper ci
gpt2_XL/ngram_web	0.560	0.536	0.584
gpt2_small/ngram_web	0.580	0.558	0.602
qwen/ngram_web	0.440	0.414	0.465
ngram_cnk/ngram_web	0.548	0.523	0.571
gpt2_XL/ngram_cnk	0.447	0.420	0.473
gpt2_small/ngram_cnk	0.373	0.345	0.398
qwen/ngram_cnk	0.187	0.159	0.212
qwe/gpt2_XL	0.616	0.592	0.639
qwen/gpt2_small	0.566	0.541	0.589

Table 2: Effect size of corpus (genre) on the correlation between lexical surprisal judgments provided by two models for each document, with a 95% confidence interval

qwen) and the n-gram models. In all corpora, the generative models tend to assess most contexts as highly predictable with long right tails, whereas the n-gram based models exhibit considerably flatter distributions across all genres. The difference is particularly strong in *legal_admin* and *justice*. Among generative models, *gpt2_XL* has the steepest right slope in most genres. The n-gram models have a very high correlation with all models, except *legal_admin* and *justice*.

In Fig. 2, the tails were cut to zoom in on the initial slope. The actual maxima were mostly around 75. The absolute maximum was 122.4 (*gpt2_small* on *justice*). Both the slope and the tail are in this case best reflected by kurtosis (Fig. 3).

In order to better understand the surprisal patterns in text, we examine the distributions by different parts of speech. In particular, we look at how difficult it is for the models to predict nouns and verbs, based on their preceding context. We observe across all models that the distributions of lexical surprisal in nouns largely mirror the complete distributions, with more extreme values in *justice*, *legal_admin*, and *parl_debates* where nouns seem to be over-represented among the most predictable words (see Fig. 2)[3]. This contrasts with expectations regarding readability, as many of these low-surprisal terms consist of technical legal jargon. The effect is most pronounced for the *gpt2_XL*. This result is consistent with Oh and Schuler (2023) who found that the surprisal estimates of nouns and adjectives produced by large Transformer models have a severe tendency to underpredict their reading times.

When analyzing the surprisal of verbs, we observe completely different patterns (see Fig. 4). In particular, the distribution for content verbs exhibits a higher mean and a more symmetric, bell-shaped distribution. In our data, it is, on average, easier for all models to predict nouns and than content verbs. However, this raises the question of whether technical nouns are also easier for humans to process.

The observed high variation in verbs is consistent with the findings of Gennari and Poeppel (2003) on reading times of different semantic classes of verbs in context. Since processing the verb means processing the entire event, verbs naturally carry more information, and the information is even scattered in context, verbs denoting complex events naturally tend to have higher reaction times in self-paced reading. From the perspective of nouns, different verb senses pose different collocability constraints on their noun arguments, which in turn makes the nouns more predictable (e. g., the set of typical objects of *read* is at least conceptually easy to oversee)[4].

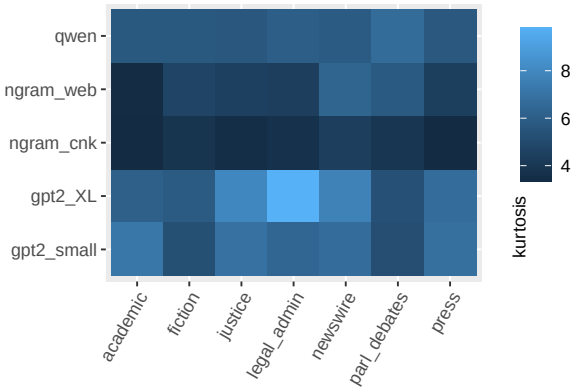

Figure 3: Kurtosis of the distributions of negative log probabilities of lexical surprisals on individual corpora by different language models

The curves of auxiliary (function) words resemble the noun curves, but with substantially lower kurtosis. The difference is particularly prominent in GPT-2-based models on *justice* and *legal_admin*. As auxiliary words typically conform to grammatical structure, their higher predictability is expected. However, the probability density of lexical surprisal for auxiliary words indicates that, in GPT-2-based models, they are still less predictable than nouns, particularly in legal language.

5.3 Correlation between TF-IDF and lexical surprisal

While lexical surprisal approximates the contextual unpredictability of a word given its preceding context, TF-IDF measures its importance within a document relative to a corpus. It is therefore useful to combine the two, as they capture complementary aspects of linguistic information.

We observe weak-to-moderate positive correlation effects of TF-IDF (Silge and Robinson, 2016) and lexical surprisal in n-gram-based models on all corpora. All models exhibit weak-to-moderate positive correlation on *fiction* and, with the exception of gpt2_XL, on *academic* (see Fig. 5). TF-IDF was computed on all word forms converted to

[3]The subplots come in a slightly different order in the two figures, as they are always sorted according to the median of the negative log probability across all models).

[4]see also (Hanks, 2013)

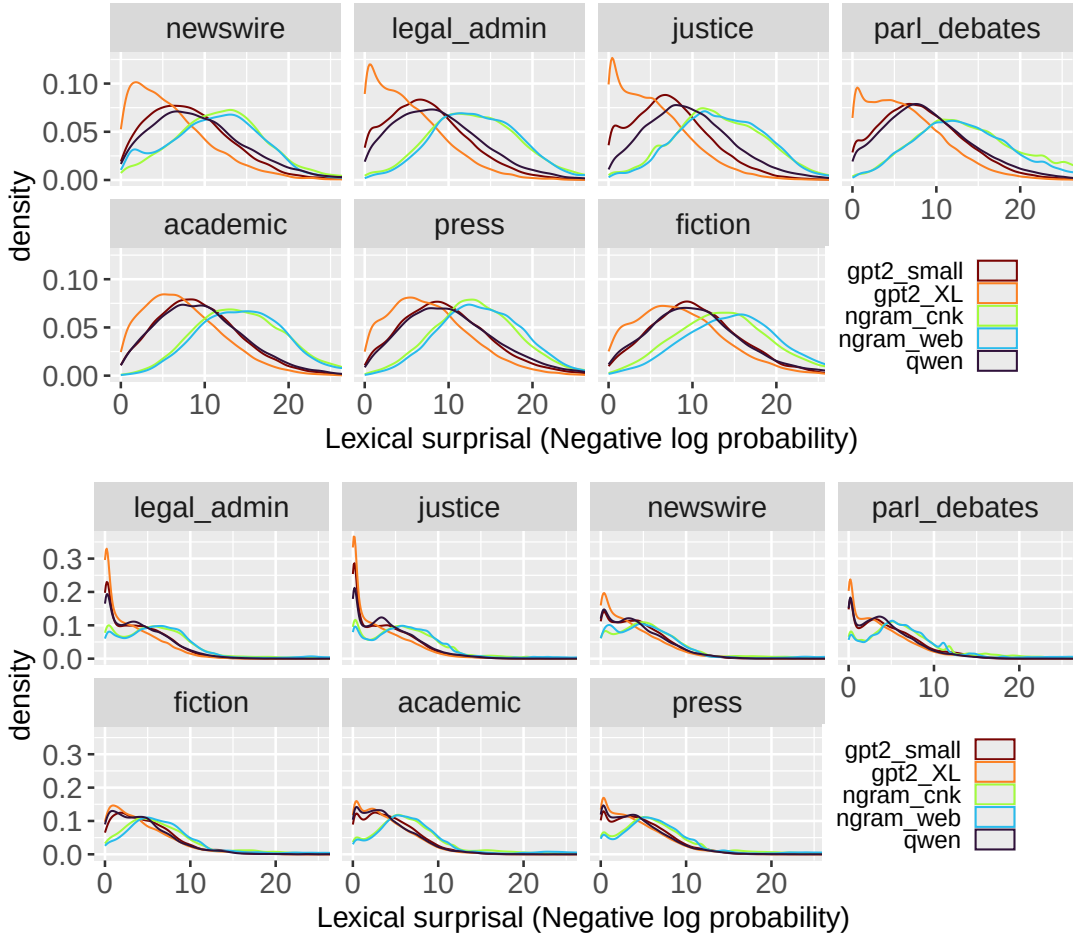

Figure 4: Distributions of negative log probabilities of lexical surprisals of content verbs (top) and lexical surprisals of auxiliary verbs, prepositions, and conjunctions (bottom) on individual corpora by different language models

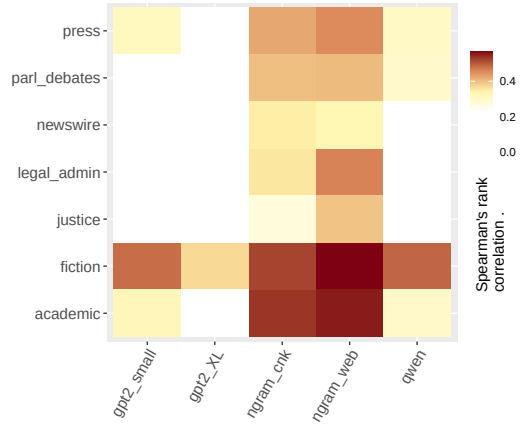

Figure 5: Weak to moderate positive correlation between lexical surprisal (negative log probability) and TF-IDF of all words in combinations of models and corpora

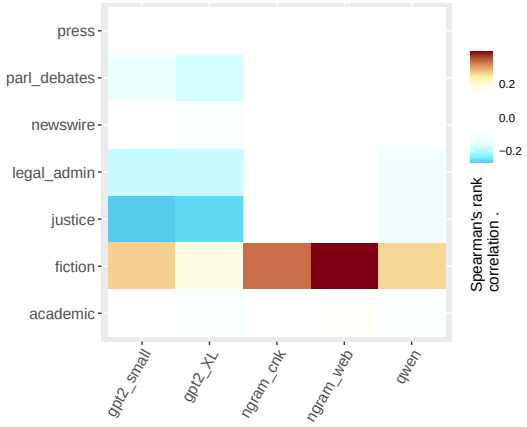

Figure 6: Weak positive and negative correlation effects between lexical surprisal (negative log probability) and TF-IDF of content words in combinations of models and corpora

lower case, including symbols and function words but not punctuation, which had already been filtered away from lexical surprisal values. Each corpus represented one document.

With TF-IDF computed only on content words, excluding stopwords (Cinkova and Eder, 2021), the positive correlation effects in fiction prevailed, whereas a very weak weak negative correlation was observed in judgments by GPT-2-based models in *legal_admin* and *justice* (Fig. 6).

All correlation values (Spearman's rank correlation ρ) in these two experiments were highly statistically significant.

6 Discussion

We observe differing surprisal patterns across genres[5]. Fiction exhibits consistent alignment between surprisal estimates across different models. In contrast, legal and institutional texts show weak correlations between Transformer-based and n-gram models, likely due to the high predictability of formulaic technical language for models extensively exposed to such data during training. Furthermore, technical jargon in legal texts typically consists of content words with high TF-IDF values. The negative correlations shown in Fig. 6 suggest that GPT 2-based models assign particularly low surprisal to these terms.

As each model reflects the characteristics of its training data, the divergence is expected: the n-gram models were trained in controlled settings on a balanced corpus (*ngram_cnk*) and on news-domain text (*ngram_web*) whereas the Transformer models were trained on large-scale data scraped from web. However, in the context of readability estimation, it is crucial to select the model that best approximates human reader behavior. Our findings suggest that model capacity, training data, and domain together play a crucial role in determining the cognitive plausibility of surprisal estimates.

7 Future work

So far, we have explored five models for Czech texts, which represented: large multilingual generative models fine-tuned to Czech, a purely multilingual generative model, and n-gram based models trained on different data sets. If we do not learn about similar work for other languages shortly, we will need to conduct experiments on other languages and comparable datasets. This requires the proper construction of a set of domain-specific corpora for each language, as well as a set of comparable models.

Assuming that we are going to need a human reference, we are preparing an Self-Paced-Reading experiment on samples of the legal and the press corpus. We are going to approximate the lexical surprisal with reading time and hope to find models whose lexical surprisal judgments display a satisfactory correlation with the human reading times.

8 Conclusion

In contrast to previous work, we point out that there are certain kinds of text that are highly predictable for the language models but incomprehensible for human readers, such as *legalese* in Czech. We have gathered evidence that, in Czech texts, there are substantial differences in lexical surprisal judgments by different LLMs and, on top of that, that they depend on genre, with the legal domain being particularly sensitive to model choice. Hence, we assume that the performance of language models is not only language-specific but also specific to a given genre.

9 Acknowledgements

This output was created with the support of the Czech Ministry of Education, Youth and Sports and the Operational Program Johannes Amos Comenius within the project Open Science II (reg. No. CZ.02.01.01/00/24_030/0015041, "HumanAId"). The authors acknowledge the support of the National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO. The work described here has also been using data, tools, and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

BUT-FIT. 2024. Czech-gpt-2-xl-133k. https://huggingface.co/BUT-FIT/Czech-GPT-2-XL-133k. Model card published on HuggingFace.

Jan Černý. 2025. *Measuring Lexical Surprisal in Legal Texts*. Bachelor thesis, Charles University, Faculty

[5]We consider genre in a broad sense, disregarding the interplay of style and content domain for the moment.

of Mathematics and Physics, Prague, Czech Republic.

Lukáš Chaloupský. 2022. Automatic generation of medical reports from chest X-rays in Czech. Master's thesis, Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic.

Silvie Cinkova and Maciej Eder. 2021. *tidystopwords: Customisable Stop-Words in 110 Languages*. R package version 0.9.1.

Silvie Cinková, Michal Kuk, Jana Šamánková, Barbora Kubíková, Přemysl Pospíšil, Jiří Mírovský, Barbora Hladká, and Tereza Novotná. 2024. KUKY1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy. 2023. A cross-linguistic pressure for uniform information density in word order. *Transactions of the Association for Computational Linguistics*, 11:1048–1065. Tex.eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00589/2154495/tacl_a_00589.pdf.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Tomaž Erjavec, Matyáš Kopp, Taja Kuzman Pungeršek, Nikola Ljubešić, Maciej Ogrodniczuk, Petya Osenova, Manex Agirrezabal, Tommaso Agnoloni, José Aires, Monica Albini, Jon Alkorta, Iván Antiba-Cartazo, Ekain Arrieta, Mario Barcala, Daniel Bardanca, Starkaður Barkarson, Roberto Bartolini, Roberto Battistoni, Nuria Bel, Maria del Mar Bonet Ramos, María Calzada Pérez, Aida Cardoso, Çağrı Çöltekin, Matthew Coole, Roberts Daris, Ruben de Libano, Griet Depoorter, Sascha Diwersy, Réka Dodé, Kike Fernandez, Elisa Fernández Rei, Francesca Frontini, Marcos Garcia, Noelia García Díaz, Pedro García Louzao, Maria Gavriilidou, Dimitris Gkoumas, Ilko Grigorov, Vladislava Grigorova, Dorte Haltrup Hansen, Mikel Iruskieta, Johan Jarlbrink, Kinga Jelencsik-Mátyus, Bart Jongejan, Neeme Kahusk, Martin Kirnbauer, Anna Kryvenko, Noémi Ligeti-Nagy, Giancarlo Luxardo, Carmen Magariños, Måns Magnusson, Carlo Marchetti, Maarten Marx, Katja Meden, Amália Mendes, Michal Mochtak, Martin Mölder, Simonetta Montemagni, Costanza Navarretta, Bartłomiej Nitoń, Fredrik Mohammadi Norén, Amanda Nwadukwe, Mihael Ojsteršek, Andrej Pančur, Vassilis Papavassiliou, Rui Pereira, María Pérez Lago, Stelios Piperidis, Hannes Pirker, Marilina Pisani, Henk van der Pol, Prokopis Prokopidis, Valeria Quochi, Paul Rayson, Xosé Luís Regueira, Andriana Rii, Michał Rudolf, Manuela Ruisi, Peter Rupnik, Daniel Schopper, Kiril Simov, Laura Sinikallio, Jure Skubic, Lars Magne Tungland, Jouni Tuominen, Ruben van Heusden, Zsófia Varga, Marta Vázquez Abuín, Giulia Venturi, Adrián Vidal Miguéns, Kadri Vider, Ainhoa Vivel Couso, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, Rodolfo Zevallos, and Darja Fišer. 2025. Multilingual comparable corpora of parliamentary debates ParlaMint 5.0. Slovenian language resource repository CLARIN.SI.

Irene Fernandez Monsalve, Stefan L. Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408, Avignon, France. Association for Computational Linguistics.

A. Frank and T. Florian Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, pages 939–944.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Silvia Gennari and David Poeppel. 2003. Processing correlates of lexical semantic complexity. *Cognition*, 89(1):B27–B41.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Jan Hajič, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Fučíková, Eva Hajičová, Jiří Havelka, Jaroslava Hlaváčová, Petr Homola, Pavel Ircing, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, David Mareček, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Michal Novák, Petr Pajas, Jarmila Panevová, Nino Peterek, Lucie Poláková, Martin Popel, Jan Popelka, Jan Romportl, Magdaléna Rysová, Jiří Semecký, Petr Sgall, Johanka Spoustová, Milan Straka, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jana Šindlerová, Jan Štěpánek, Barbora Štěpánková, Josef Toman, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. 2024. Prague dependency treebank - consolidated 2.0 (PDT-c 2.0). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

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.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. MIT Press.

Barbora Vidová Hladká, Jan Hajič, Jiří Hana, Jaroslava Hlaváčová, Jiří Mírovský, and Jan Raab. 2008. Czech academic corpus 2.0.

Tomáš Jelínek, Milena Hnátková, and Hana Skoumalová. 2017. FicTree 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Matyáš Kopp. 2024. ParCzech 4.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Pavel Král and Ladislav Lenc. 2018. Czech text document corpus v 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

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

Tereza Novotná, Jan Černý, Ivan Kraus, Ivana Kvapilíková, Jiří Mírovský, Arnold Stanovský, and Barbora Hladká. 2025. PONK: Tool for client-oriented legal writing in czech. In *JURIX 2025: The Thirty-eighth Annual Conference*, volume 416 of *Frontiers in Artificial Intelligence and Applications*, pages 330–336, Amsterdam, Netherlands. University of Turin, Italy, IOS Press.

Tereza Novotná and Jakub Harašta. 2019. Czech court decisions corpus (CzCDC 1.0). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Byung-Doh Oh and William Schuler. 2023. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921, Singapore. Association for Computational Linguistics.

Pedro Javier Ortiz Su'arez, Laurent Romary, and Benoit Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Conference on empirical methods in natural language processing*.

Tiago Pimentel, Clara Meister, Ethan G. Wilcox, Roger P. Levy, and Ryan Cotterell. 2023. On the effect of anticipation on reading times. *Transactions of the Association for Computational Linguistics*, 11:1624–1642.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Beatrice Santorini. 1990. Part-of-Speech Tagging Guidelines for the Penn Treebank Project. *University of Pennsylvania 3rd Revision 2nd Printing*, (MS-CIS-90-47):33. 00376 tex.ids= santoriniPartofSpeechTaggingGuidelines1990a.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Julia Silge and David Robinson. 2016. tidytext: Text mining and analysis using tidy data principles in r. *JOSS*, 1(3).

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Milan Straka, Jan Haji, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297. European Language Resources Association (ELRA). Place: Portoro, Slovenia.

K. Sun and R. Wang. 2025. Computational sentence-level metrics of reading speed and its ramifications for sentence comprehension. *Cognitive Science*, 49(7):e70092.

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.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Jephtey Adolphe, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Arofat Akhundjanova, Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Antonios Anastasopoulos, Erik Andersen, Kirk Andrews, Matthew Andrews, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, Hórunn Arnardóttir, Wirote

Aroonmanakun, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Hiwa Asadpour, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Jana Bajorat, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Bryan Khelven da Silva Barbosa, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Nathanaël Beiner, Juan Belieni, Alevtina Bémová, Kepa Bengoetxea, brahim Benli, Yifat Ben Moshe, Marie Benzerrak, Aleksandrs Berdicevskis, Ansu Berg, Gözde Berk, Delphine Bernhard, Astrid Berntsson Ingelstam, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Esma Fatıma Bilgin Taşdemir, Helin Binici, Kristín Bjarnadóttir, Verena Blaschke, Rogier Blokland, Nina Böbel, Victoria Bobicev, Loïc Boizou, Stavros Bompolas, Johnatan Bonilla, Emanuel Borges Völker, Lars Borin, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Myriam Bras, Théo Brillet, Kristina Brokaitė, Lanni Bu, Eva Buráňová, Aljoscha Burchardt, Carmen Cabeza, Natalia Cáceres Arandia, Olesea Caftanatov, Marisa Campos, Marie Candito, Caterina Maria Cappello, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Anila Çepani, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Claudine Chamoreau, Shweta Chauhan, Yifei Chen, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Bermet Chontaeva, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Esther Cocco, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Claudia Corbetta, Daniela Corbetta, Francisco Costa, Marine Courtin, Benoît Crabbé, Mihaela Cristescu, Vladimir Cvetkoski, Netanel Dahan, Ingerid Løyning Dale, Sabrina D'Alì, Philemon Daniel, Khensa Daoudi, Bijayalaxmi Dash, Satya Ranjan Dash, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Ahmet Demir, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Roberto Antonio Díaz Hernández, Carly Dickerson, Ariani Di Felippo, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Hoa Do, Kaja Dobrovoljc, Caroline Döhmer, Adrian Doyle, Timothy Dozat, Kira Droganova, Magali Sanches Duran, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Roald Eiselen, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Louise Esher, Soudabeh Eslami, Farah Essaidi, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Ján Faryad, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Barbara Francioni, Theodorus Fransen, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Edith Galy, Federica Gamba, Marcos Garcia, José María García-Miguel, Moa Gärdenfors, Tanja Gaustad, Efe Eren Genç, Fabrício Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Gili Goldin, Xavier Gómez Guinovart, Berta González Saavedra, Mathieu Goux, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Mario Guglielmetti, Bruno Guillaume, Kirian Guiller, Céline Guillot-Barbance, Tunga Güngör, Vladimir Gurevich, Nizar Habash, Hinrik Hafsteinsson, Michael Hahn, Jan Hajič, Jan Hajič jr., Eva Hajičová, Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Naïma Hassert, Dag Haug, Jiří Havelka, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Diana Hoefels, Barbara Hoff, Petter Hohle, Nick Howell, Yidi Huang, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Inessa Iliadou, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Artan Islamaj, Kaoru Ito, Federica Iurescia, Jessica K. Ivani, Sandra Jagodzińska, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Ratanon Jiamsundutsadee, Katharine Jiang, Sylvanus Job, Mayank Jobanputra, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Zhuoxuan Ju, Markus Juutinen, Hüner Kaşıkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóa, Jiří Kárník, Andre Kåsen, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Václava Kettnerová, Lilit Kharatyan, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Petr Kocharov, Arne Köhn, Abdullatif Köksal, Veronika Kolářová, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Nelda Kote, Natalia Kotsyba, Barbara Kovačić, Jolanta Kovalevskaitė, Emmanuelle Kowner, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Lucie Kučová, Adrian Kuqi, Elmurod Kuriyozov, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Kris Kyle, Käbi Laan, Veronika Laippala, Lorenzo Lambertino, Israel Landau, Tatiana Lando, Septina Dian Larasati, Pierre Larrivée, Kusum Lata, Alexei Lavrentiev, John Lee, Phng Lê Hồng, Alessandro Lenci, Wei Qi Leong, Saran Lertpradit, Herman Leung, Lori Levin, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Zoey Liu, Nikola Ljubešić, Irina Lobzhanidze, Olga Loginova, Markéta Lopatková, Lucelene Lopes, Edita Luftiu, Arsenii Lukashevskyi, Stefano Lusito, Anne-Marie Lutgen, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Punyanuch Maitreenukul, Ilya Makarchuk, Aibek Makazhanov, Francesco Mambrini, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček,

Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Cláudia Martins, Arianna Masciolini, Jan Mašek, Sanatbek Matlatipov, Hiroshi Matsuda, Yuji Matsumoto, Caterina Mauri, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Maitrey Mehta, Pierre André Ménard, Gustavo Mendonça, Hilla Merhav, Tatiana Merzhevich, Paul Meurer, Niko Miekka, Marie Mikulová, Emilia Milano, Aleksandra Miletić, Aaron Miller, Junghyun Min, Yael Minerbi, Jiří Mírovský, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Biswakalpita Mohapatra, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Katerina Mouzou, Kadri Muischnek, Robert Munro, Yugo Murawaki, Nikolett Mus, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Minoo Nassajian, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Victor Norrman, Alireza Nourian, Michal Novák, Maria das Graças Volpe Nunes, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Annika Ott, Lilja Øvrelid, Masanori Oya, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Petr Pajas, Thomas Palakapilly, Alessio Palmero Aprosio, Jarmila Panevová, Ludovica Pannitto, Anastasia Panova, Thiago Alexandre Salgueiro Pardo, Shantipriya Parida, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Oggi Peeters, Angelika Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Eva Pettersson, Andrea Peverelli, Jason Phelan, Claudel Pierre-Louis, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Rodrigo Pintucci, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Alistair Plum, Thierry Poibeau, Charin Polpanumas, Larisa Ponomareva, Martin Popel, Clamença Poujade, Rangga Prangwedana Prangwedana, Lauma Pretkalniņa, Rigardt Pretorius, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Christoph Purschke, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Ella Rabinovich, Alexandre Rademaker, Mutee-u Rahman, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Norton Trevisan Roman, Mykhailo Romanenko, Natalia Romanova, Rudolf Rosa, Valentin Roca, Paulette Roulon, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Paolo Ruffolo, Kristján Rúnarsson, Rozana Rushiti, Attapol T. Rutherford, Shoval Sadde, Pegah Safari, Aleksi Sahala, Kalyanamalini Sahoo, Saraswati Sahoo, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Konstantinos Sampanis, Stephanie Samson, Xulia Sánchez-Rodríguez, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Marta Sartor, Albina Sarymsakova, Mitsuya Sasaki, Baiba Saulīte, Agata Savary, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Emmanuel Schang, Robert Schikowski, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Sven Sellmer, Mojgan Seraji, Magda Ševčíková, Petr Sgall, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Gyu-Ho Shin, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Jean Sibille, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Dolores Sollberger, Barbara Sonnenhauser, Shafi Sourov, Nina Speransky, Rachele Sprugnoli, Panyut Sriwirote, Vivian Stamou, Steinþór Steingrímsson, Antonio Stella, Jan Štěpánek, Barbora Štěpánková, Abishek Stephen, Milan Straka, Omer Strass, Emmett Strickland, Jana Strnadová, Sara Stymne, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Jack Sun, Hakyung Sung, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Maria Irena Szawerna, Chihiro Taguchi, Dima Taji, Luigi Talamo, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Nursena Teker, Samson Tella, Isabelle Tellier, Marinella Testori, Santhawat Thanyawong, Guillaume Thomas, Tarık Emre Tıraş, William Chandra Tjhi, Thea Tollersrud, Kamil Tomaszek, Sara Tonelli, Liisi Torga, Lucas Toribio, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Kira Tulchynska, Utku Türk, Francis Tyers, Sveinbjörn Þórðarson, Vilhjálmur Þorsteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Socrates Vakirtzian, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Helena Vaz, Uliana Vedenina, Giulia Venturi, Marianne Vergez-Couret, Annemarie Verkerk, Barbora Vidová Hladká, Eric Villemonte de la Clergerie, Veronika Vincze, Anishka Vissamsetty, Natalia Vlasova, Eleni Vligouridou, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, John Wang, Jonathan North Washington, Leonie Weissweiler, Maximilan Wendt, Paul Widmer, Aleksandra Wieczorek, Shira Wigderson, Sri Hartati Wijono, Vanessa Berwanger Wille, Seyi Williams, Miriam Winkler, Shuly Wintner, Mats Wirén, Christian Wittern, Alena Witzlack-Makarevich, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska,

Qishen Wu, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Qizhen Yang, Xiulin Yang, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Enes Yılandiloğlu, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, Annie Zhang, Larry Zhang, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, Rayan Ziane, Artūrs Znotiņš, and Eleonora Zucchini. 2025. Universal dependencies 2.17. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).