COMPLEX TEXTS

Jabborova Dildora Zokirjanovna
Department of English language theoretical aspects senior teacher PhD
Uzbekistan State World Languages University
dilyacity89@gmail.com
Farangiz Baxromova
Uzbekistan State World Languages university
First English faculty Foreign language and literature
Farangizbaxromova0@gmail.com

Annotatsiya. Lingvistik murakkablik ko‘p darajali konstrukt bo‘lib, matn, gap, so‘z va so‘z osti darajalarida namoyon bo‘ladi hamda turli lingvistik xususiyatlar (masalan, janr, sintaksis, semantika) hamda til o‘zlashtirish, tarjima va ta’lim kabi vazifalar bilan kesishadi. Murakkablikni o‘lchashdagi tillararo tafovutlar tipologik farqlar, madaniy jihatdan singdirilgan janr konvensiyalari va ma’lumotlar to‘plamining o‘ziga xos xususiyatlaridan kelib chiqadi. Ushbu tadqiqotda biz sun’iy neyron tarmoqlardan lingvistik murakkablikni bashorat qilish va ushbu bashoratlarni izohlash uchun foydalanamiz. Neyron modellar yuqori empirik samaradorlikka erishish uchun millionlab parametrlarni optimallashtirsa-da, ular odatda “qora quti” sifatida ishlaydi, ya’ni qarorlariga asos bo‘luvchi lingvistik belgilarni aniq ko‘rsatmaydi. Biz neyron murakkablik baholarini shaffof va izohlanadigan xususiyatlar – jumladan, bog‘lovchilar, diskurs zarralari va ergash gaplar chastotasi bilan bog‘lash usulini namoyish etamiz. Turli janrlarga mansub ingliz va rus tillaridagi matnlardan foydalanib, biz neyron modellarni kam murakkab va yuqori murakkab matnlarni farqlashga o‘rgatamiz. Natijalarimiz shuni ko‘rsatadiki, otlar chastotasi va otli iboralarning tarkibiy murakkabligi matn murakkabligining muhim bashoratchilari hisoblanadi. Yakunida, biz murakkablik va janr o‘rtasidagi bog‘liqlikni tahlil qilib, ayrim xususiyat–murakkablik assotsiatsiyalarining ichki lingvistik qiyinchilikdan ko‘ra janr tafovutlari bilan belgilanishini ochib beramiz.

Kalit so‘zlar: lingvistik murakkablik, neyron tarmoqlar, murakkablikni bashoratlash, interpretatsiyalanuvchanlik, matnlarni tasniflash, janr tahlili, ot birikmalari, ingliz va rus tillari, tillararo variatsiya, diskurs xususiyatlari hamda sintaktik murakkablik tushunchalari qo‘llaniladi.

Аннотация. Лингвистическая сложность представляет собой многоуровневый конструкт, проявляющийся на текстовом, клаузальном, лексическом и сублексическом уровнях и пересекающийся с различными лингвистическими признаками (например, жанр, синтаксис, семантика), а также с такими задачами, как усвоение языка, перевод и обучение. Кросс-лингвистическая вариативность в измерениях сложности д ополнительно обусловлена типологическими различиями, культурно укоренёнными жанровыми конвенциями и свойствами конкретных наборов данных. В данном исследовании мы используем искусственные нейронные сети как для прогнозирования лингвистической сложности, так и для интерпретации этих прогнозов. Хотя нейронные модели оптимизируют миллионы параметров для достижения высокой эмпирической производительности, они обычно функционируют как «чёрные ящики», не предоставляя явного объяснения того, какие лингвистические cues лежат в основе их решений. Мы демонстрируем, как связать нейросетевые оценки сложности с прозрачными, интерпретируемыми признаками, включая частоту союзов, дискурсивных частиц и придаточных предложений. Используя тексты на английском

DOI: 10.5281/zenodo.20377855

и русском языках, взятые из различных жанров, мы обучаем нейронные модели различать менее сложные и более сложные тексты. Наши результаты показывают, что частота существительных и структурная сложность именных групп являются значимыми предикторами текстовой сложности. Наконец, мы анализируем взаимосвязь между сложностью и жанром, выявляя, что некоторые ассоциации между признаками и сложностью обусловлены жанровыми различиями, а не внутренней лингвистической трудностью.

Ключевые слова: лингвистическая сложность, нейронные сети, прогнозирование сложности, интерпретируемость, классификация текстов, анализ жанра, именные словосочетания, английский и русский языки, межъязыковая вариативность, дискурсивные особенности и синтаксическая сложность.

Annotation. Linguistic complexity constitutes a multi-level construct, manifesting across textual, clausal, lexical, and sublexical domains, and intersecting with various linguistic features (e.g., genre, syntax, semantics) as well as tasks such as language acquisition, translation, and instruction. Cross-linguistic variation in complexity measurements further arises from typological differences, culturally embedded genre conventions, and dataset-specific properties. In this study, we employ artificial neural networks both to predict linguistic complexity and to interpret those predictions. Although neural models optimize millions of parameters to achieve high empirical performance, they typically function as black boxes, offering no explicit account of which linguistic cues inform their decisions. We demonstrate how to associate neural complexity estimates with transparent, interpretable features—including the frequency of conjunctions, discourse particles, and subordinate clauses. Using English and Russian texts drawn from multiple genres, we train neural models to discriminate between less complex and more complex texts. Our findings indicate that noun frequency and the structural complexity of noun phrases are significant predictors of textual complexity. Finally, we examine the relationship between complexity and genre, revealing that certain feature–complexity associations are driven by genre differences rather than by intrinsic linguistic difficulty.

Key words: linguistic complexity, neural networks, complexity prediction, interpretability, text classification, genre analysis, noun phrases, English, Russian, cross-linguistic variation, discourse features, syntactic complexity

Introduction

Linguistic complexity is broadly acknowledged as a multifaceted construct. It operates across multiple levels of linguistic organisation, from full texts through sentences and individual words down to subword units. At the same time, complexity is conditioned by a range of linguistic attributes, including genre, syntactic structures, and semantic relations, and it holds significant relevance for several applied domains, such as second language acquisition, translation studies, educational practice, and content adaptation for heterogeneous audiences. In addition, assessments of linguistic complexity are known to differ across languages, owing to typological divergences, culturally grounded genre conventions, and the specific characteristics of individual datasets used in analysis.

Although automatic complexity assessment has attracted growing attention, most conventional approaches depend on manually defined linguistic features and rule-based heuristics. By contrast, artificial neural networks represent a powerful alternative, as they are capable of optimising millions of parameters to develop empirically robust predictive models. Nevertheless, these models typically function as opaque systems, providing no explicit indication of which linguistic factors inform their predictions. This lack of transparency

restricts their utility both for theoretical inquiry in linguistics and for practical applications that demand interpretable explanations.

In the present study, we address this limitation by showing how neural predictions of textual complexity can be mapped onto interpretable linguistic properties, including the frequency of conjunctions, discourse particles, and subordinate clauses. Our investigation centres on neural models trained to differentiate between less complex and more complex texts across a variety of genres in English and Russian. We then examine which linguistic features correlate with the models' predictions and whether such correlations remain stable across languages and genres.

Our findings offer empirical support for the claim that noun frequency and the structural complexity of noun phrases exert a significant influence on predicted text complexity. Furthermore, we demonstrate that certain relationships between linguistic features and complexity are not intrinsic but are instead mediated by genre. These results contribute both to the advancement of interpretable neural models for practical applications and to a deeper theoretical understanding of linguistic complexity as a genre‑sensitive phenomenon.

Conclusion

This study demonstrated that neural network predictions of linguistic complexity can be linked to interpretable linguistic properties, despite the inherent opacity of these models. Using English and Russian texts across multiple genres, we showed that noun frequency and noun phrase complexity are statistically significant predictors of textual complexity. Furthermore, we revealed that certain feature–complexity relationships are not intrinsic but rather mediated by genre. These findings have practical implications for developing interpretable models in educational and translation settings, as well as theoretical implications for understanding complexity as a genre‑sensitive construct. Future research should extend this approach to typologically diverse languages and explore causal mechanisms underlying the observed correlations.

References

1. Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. Journal of Second Language Writing, 26, 66–79.
2. Lau, J. H., & Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. Proceedings of the 1st Workshop on Representation Learning for NLP, 78–86.
3. Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. Proceedings of the 31st International Conference on Machine Learning, 1188–1196.
4. McNamara, D. S., Crossley, S. A., & Roscoe, R. D. (2013). Natural language processing in an intelligent writing strategy tutoring system. Behavior Research Methods, 45(2), 499–515.
5. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26, 3111–3119.
6. Pilán, I., Volodina, E., & Johansson, R. (2014). Rule-based and machine learning approaches for second language sentence-level readability. Proceedings of the 9th Workshop on Innovative Use of NLP for Building Educational Applications, 174–184.
7. Reynolds, R. (2016). The role of noun phrase complexity in L2 reading comprehension. Reading in a Foreign Language, 28(1), 86–105.

8. Solovyev, V., & Solnyshkina, M. (2018). Text complexity evaluation in Russian: Approaches and resources. Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies, 17(24), 607–618.