Complexity in Language

CompiLa2025

4 September 2025 • Siena (Italy)

Satellite of the CCS2025 Conference

Book of Abstracts

Organizers: Els Heinsalu, Marco Patriarca, David Sánchez

Language Models and Language Complexity

Christian Bentz (University of Passau)

The objective of language models is to predict subsequent units in texts. These units can be characters, subwords, or words. Neural language models have become state-of-the-art for this task with the advent of deep feedforward neural networks. The latest transformer-based architectures even outperform humans in terms of predictive capacity for English characters. Hence, neural language models can serve as a tool to estimate the complexity of languages. Namely, the cross-entropy of a language model achieved on a test set can be seen as an approximation to the complexity of the respective language. In this talk, results of preliminary analyses with neural net architectures on several dozen languages of different families, areas and writing systems are reported. While cross-entropies differ for characters and words depending on the typology of the language, the differences in cross-entropies for subwords are negligible. In other words, diverse languages are equally hard to language model when "sweetspot" subword tokenization is applied.

A stochastic model of phonetic variation: Linguistic distances and geography

David Sánchez (Institute for Cross-Disciplinary Physics and Complex Systems IFISC (UIB-CSIC)) • Coauthors: Marius Mavridis, Juan De Gregorio, Raúl Toral

We apply an information-theoretic method to quantify phonetic distances between 67 African, European and Asian languages. This approach has already been carried out in the context of syntactic distances, and has proved insightful to recover language groups and families, while highlighting correlations between geographic and linguistic proximity [1]. We build a large multilingual corpus by transcribing the Bible in the International Phonetic Alphabet (IPA), for each of our 67 languages. After validating the transcriptions using an reference database, we model the sequence of IPA characters (each symbol being mapped to a unique phoneme) as a high-order Markov chain with memory m, meaning that each state or phoneme depends on m previous ones. We estimate the memory using the zeroes of the predictability gain [2], and show that m=2 is the best compromise between statistical significance and computation time, while avoiding undersampling. Next, we compute the pairwise distances between the 3-phone (or sequence of 3 consecutive phonemes) probability distributions of the corpus languages, which in turn are used to identify distinctive clusters of phonetically close languages. These clusters largely reproduce known language groups, with notable exceptions which can be explained by cross-family language contact. We show that phonetic and geographic distances are significantly and positively correlated, meaning that languages spoken in geographically closer regions will tend to share more phonetic similarities. Finally, we use our results to constrain the origin region of the Indo-European family. Interestingly, we find that the region with maximum likelihood is compatible with the Steppe hypothesis for the Proto-Indo-European homeland.

[1] J. De Gregorio, R. Toral, and D. Sánchez, Exploring language relations through syntactic distances and geographic proximity, EPJ Data Science 13, 61 (2024).

[2] J. P. Crutchfield and D. P. Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos 13, 25 (2003).

Noun Phrases in English Scientific Writing: The Diachronic Evolution of Information Density

Isabell Landwehr (Saarland University) • Coauthors: Arianna Bienati

Complex noun phrases (NPs) are a key feature of English scientific writing and typically used for encoding technical and specialized concepts (Halliday 1988; Banks 2008). Over time, scientific writing has moved from an emphasis on clausal structures towards an emphasis on phrasal structures, allowing the transmission of information in a highly compressed way (Biber & Gray 2011). In this way, scientific English has evolved to be an optimized code for expert-to-expert communication (Degaetano-Ortlieb & Teich 2022).

In this study, we analyze these diachronic optimization mechanisms using the information-theoretic notion of Uniform Information Density (UID; Jaeger 2010). The UID hypothesis is based on the possibility of modeling language using surprisal (i.e. predictability in context, Shannon 1948) and predicts that language users prefer structures that are informationally more uniform if several encoding options exist (Jaeger 2010). Evidence for this theory has been found cross-linguistically for many linguistic phenomena, especially syntactic ones (Jaeger 2010; Clark et al. 2023; Liang et al. 2024). We aim to add a diachronic and register-informed perspective, focusing on NPs due to their significance for scientific writing. Our hypothesis is that NPs exhibit increased UID over time, as the register becomes increasingly optimized.

UID has been operationalized in various ways (for an overview, see Meister et al. 2021). Since we assess the information profiles of noun phrases, we choose measures of local variability. Local variability can be conceptualized as the magnitude of information gains and losses in the stream of communication. It is commonly operationalized by computing the delta of adjacent tokens' surprisals and then applying a descriptive statistical summary: Collins (2014) proposes taking the average of the differences and names the resulting measure UIDev, while Bates and Shepard (1993) use the standard deviation, resulting in a measure known as Information Fluctuation Complexity (IFC).

Our dataset is the Royal Society Corpus (Fischer et al. 2020; Menzel et al. 2021), a diachronic corpus of scientific writing. We use a version enriched with Universal Dependencies (De Marneffe et al. 2021) and surprisal annotation, the latter based on a 4-gram language model trained on the corpus. Furthermore, we focus on the articles in the Series A and Series B journals of the corpus, which contain texts from biology, physics and mathematics and span the time from 1887 to 1990. After extracting all subject and direct object NPs, we apply both UIDev and IFC as measures of UID. Using a linear-mixed effects model, we analyze the diachronic development of UIDev and IFC. We expect a positive effect of time on UID, indicating a trend towards increased optimization of the register over time.

References

David Banks. 2008. The Development of Scientific Writing: Linguistic Features and Historical Context. University of Toronto Press.

John E. Bates and Harvey K. Shepard. 1993. Measuring complexity using information fluctuation. Physics Letters A 172(6): 416-425.

Douglas Biber and Bethany Gray. 2011. Grammatical change in the noun phrase: The influence of written language use. English Language & Linguistics, 15(2): 223–250.

Thomas H. 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: 1048–65.

Michael Xavier Collins. 2014. Information density and dependency length as complementary cognitive models. Journal of Psycholinguistic Research, 43(5): 651–681.

Stefania Degaetano-Ortlieb and Elke Teich. 2022. Toward an optimal code for communication: The case of scientific English. Corpus Linguistics and Linguistic Theory, 18(1): 175–207.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre and Daniel Zeman. 2021. Universal Dependencies. Computational Linguistics, 47(2): 255–308.

Stefan Fischer, Jörg Knappen, Katrin Menzel, and Elke Teich. 2020. The Royal Society Corpus 6.0: Providing 300+ years of scientific writing for humanistic study. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 794-802. European Language Resources Association.

Michael A. K. Halliday. 1988. On the language of physical science. In Mohsen Ghadessy (ed.), Registers of written English: Situational factors and linguistic features, pp. 162-177.

Florian T. Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. Cognitive Psychology, 61(1): 23-62.

Yiming Liang, Pascal Amsili, Heather Burnett and Vera Demberg. 2024. Uniform Information Density Explains Subject Doubling in French. Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0).

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, pp. 963–80. Association for Computational Linguistics.

Katrin Menzel, Jörg Knappen, and Elke Teich. 2021. Generating linguistically relevant metadata for the Royal Society Corpus. Research in Corpus Linguistics, 9(1):1–18.

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

An experimental study of iterated learning and language evolution

Conor Houghton (University of Bristol) • Coauthors: Hyoyeon Lee, Seth Bullock

At COMPiLA2024 I presented a novel iterated learning model of language evolution. There is a tendency in iterated learning models for the language to collapse across generations, dwindling to a few signals each encoding a large number of meanings. We discovered that one approach to combatting this collapse is to include auto-encoder learning in the model; a mechanism we claim is an analogue of reflection in language learning (Bunyan et al 2025). We are now testing this idea with an online human participant study which follow the earlier experimental study reported in (Kirby et al 2008). Our experiment includes a test phase that prompts appropriate reflection. This experiment is ongoing but will be complete before the workshop and I propose describing our motivation and the results at COMPiLA2025.

Bunyan, J., Bullock, S., & Houghton, C. (2025). An iterated learning model of language change that mixes supervised and unsupervised learning. PLOS Complex Systems, 2(3), e0000030.

Kirby, S., Cornish H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language, Proc. Natl. Acad. Sci. U.S.A. 105 (31) 10681-10686.

Emergent Syntax in the Left Periphery: A Complexity-Inspired Simulation of Word Order Patterns

Elena Callegari (University of Iceland)

The left periphery (LP) refers to the portion of clause structure that includes elements appearing before the subject, such as topicalized phrases, wh-words, and complementizers. In traditional generative linguistics theories, the LP has been modeled using cartographic templates, which treat it as a fixed sequence of positions. Each type of constituent is assigned a specific structural slot, and the same overall hierarchy is assumed to be present across all languages (Rizzi 1997, Rizzi 2001, Frascarelli 2012, Rizzi & Bocci 2017). While this approach captures many descriptive generalizations, it also assumes that all possible LP elements have pre-assigned structural positions, even when they are not overtly expressed. This means that speakers are taken to build sentences using an underlying syntactic blueprint that includes the full left-peripheral hierarchy, regardless of whether all its parts are needed.

In contrast, this presentation explores whether the observed ordering of LP constituents can emerge as a byproduct of local interactions among syntactic elements, with word order arising dynamically rather than being fixed by a pre-specified hierarchy.
Building on insights from complexity science and emergent systems, I present a computational simulation implemented in NetLogo that models LP word order as a dynamic, self-organizing process. In this framework, syntactic constituents (e.g., topics, foci, interrogatives) are modeled as agents occupying patches in a one-dimensional cellular automaton. Their relative positions are determined by a system of local precedence and compatibility constraints, applied incrementally through pairwise interactions. Crucially, no element is assigned a fixed structural slot: global ordering is an emergent phenomenon, arising through the cumulative effect of local decisions.

Using my simulation, I test whether complex and empirically attested LP configurations, such as those proposed in the cartographic tradition (e.g., Rizzi 1997; Frascarelli 2012), can be reproduced using only local update rules. I show that when a feedback-permissive update strategy is employed (analogous to cascading adjustments in dynamic systems), the model reliably converges on grammatical surface orders. By contrast, a more restrictive update mode (where previously adjusted elements are frozen, akin to strict cyclic derivation) frequently fails to generate licit configurations. This contrast highlights the importance of emergent repair mechanisms in syntactic computation.

These results demonstrate that a small set of local rules, iterated with feedback, is enough to recover the complex patterns usually attributed to an innate template. Word order can therefore be treated as a process that unfolds dynamically rather than as a static architectural plan. Because the simulation code is open-access, researchers can alter the rule set, add language-specific constraints, and derive new predictions, turning the model

into a test bed for interaction-based theories of constituent structure.

More broadly, this presentation illustrates how tools from complexity science and agent-based modelling can be brought to bear on traditional syntactic questions. If the LP (one of the most articulated areas of clause structure) can be modelled without fixed slots, the same methodology may be applied to other ordering domains, encouraging a shift from "syntax as blueprint" to "syntax as emergent organization".

Frascarelli, Mara. 2012. "Discourse-related Features and Functional Projections." In Functional Heads: The Cartography of Syntactic Structures, Vol. 7, ed. C. Cinque & G. Rizzi, 45–64. Oxford: Oxford University Press.

Rizzi, Luigi. 1997. "The Fine Structure of the Left Periphery." In Elements of Grammar, ed. L. Haegeman, 281–337. Dordrecht: Kluwer.

Rizzi, Luigi. 2001. "On the Position 'Int(errogative)' in the Left Periphery of the Clause." In Current Studies in Italian Syntax: Essays Offered to Lorenzo Renzi, ed. G. Cinque & G. Salvi, 287–296. Amsterdam: Elsevier.

Rizzi, Luigi & Giuliano Bocci. 2017. "Left Periphery of the Clause." In The Wiley Blackwell Companion to Syntax, 2nd ed., ed. M. Everaert & H. van Riemsdijk, 1–40. Hoboken, NJ: Wiley-Blackwell.

Scaling limit of the Random Language Model

Eric De Giuli (Toronto Metropolitan University)

Scaling laws characterizing the loss landscape of large language models as a function of training data and parameter size demand the construction of simple models relating properties of texts to their probabilities. Context-free grammars are a broad class of systems generating texts with long-range correlation, and which compactly encode the syntactic structure of human and computer languages. In [1] an ensemble of context-free grammars was proposed and dubbed the Random Language Model. It was shown that by varying the natural temperature-like parameter ε_d of the model one can encounter a transition between a simple 'babbling' regime and a regime in which nontrivial information is carried. Analytical results using a field-theoretic method have been limited to the babbling regime and the transition onset [2,3], and the existence of a true phase transition has been questioned [4].

Here we show that the 'energetic' aspect of the model is governed by a Random Energy Model [5], where the thermodynamic limit of the latter becomes a scaling limit of the language model, in which the number of hidden symbols $N \to \infty$ while $\varepsilon_d \to 0$ at fixed $\varepsilon_d \log N$, as already shown to be the correct scaling for a transition [1,2,3,6]. As a consequence of this mapping the language model has a spectrum of singular temperatures. New theory significantly improves the prediction of Shannon entropy.

[1] Eric De Giuli, Random Language Model, Phys. Rev. Lett. 122 (2019) 128301.

[2] Eric De Giuli, Emergence of order in random languages, J. Phys. A. 52 (2019) 504001.

[3] Eric De Giuli, Corrigendum: emergence of order in random languages, J. Phys. A. 55 (2022) 489501.

[4] Kai Nakaishi and Koji Hukushima, Absence of phase transition in random language model, Phys. Rev. Res. 4 (2022) 023156.

[5] Bernard Derrida, Random-energy model: Limit of a family of disordered models, Phys. Rev. Res. 45 (1980) 79.

[6] Fatemeh Lalegani and Eric De Giuli, Robustness of the Random Language Model, Phys. Rev. E 109 (2024) 054313.

Derivation of the pattern and length distribution of an English sentence using combinatorial categorical grammar

Takehito Suzuki (Takachiho University) • Coauthors: Tsuyoshi Mizuguchi

We examined the distribution of sentence length L (the number of words generating a sentence) using a model based on Combinatory Categorial Grammar (CCG) (Steedman, 2001), which assigns a syntactic category to each word and explains sentence structure through the grammatical combinations of these words. The CCG has three ground categories, nouns (N), noun phrases (NP), and sentences (S), and all the other categories are represented by combinations of the ground categories and the operators / and \. Categories A/B and A\B are defined as follows: The category A is generated if the category B is placed right after A/B or right before A\B. For example, let us consider an intransitive verb. If a noun phrase precedes the intransitive verb, a sentence is generated. Therefore, an intransitive verb can be written as S\NP in terms of CCG. Since combinations of categories via operators such as A/B are also considered categories, there are more than three categories. However, a finite number of categories is sufficient for natural language.

Additionally, we assume functional-composition rules (successive categories including / or \ are transformed into another category) and type-raising rules (extending the scope of the composition rules by extending the category). If the transformation to category S is possible by applying these rules from the beginning of the sentence to the end of the sentence, we consider a grammatical sentence has been generated.

We then define a row vector \bm{W}^i as a vector whose components are categories, where the component W_j^i describes the jth category consisting of combination of words from zeroth to the ith in the sentence. Notably, \bm{W}^i and \bm{W}^{i+1} are found to be related via \bm{W}^{i+1}=\bm{W}^i G, where G is a grammar matrix whose components are categories.

Although CCG is applicable to many languages, in this presentation we restrict ourselves to English. We assume that every sentence begins with an article. Therefore, we cannot deal with some expressions such as imperative sentences, but we will start with a simple assumption. We found that 19 components are sufficient for \bm{W}^i, and that G is a 19×19 matrix. Using G, we first obtained a pattern of the order of the categories. For example, if L=4, there are four patterns, one of which is article + adjective + noun + intransitive verb. By applying G and using the approximation of the random selection from the possible categories, we also obtained the probability that the sentence length is L, p(L), and found that the expected value \sum Lp(L) is about 10, which is slightly smaller than the observed values. This discrepancy can be attributed to our assumption that some expressions, such as parallel representations, are not considered. This study is a basis for the future study related to the phase transition observed in natural languages.

Word Co-occurrence SVN Topic Model: a network approach to analyse textual data

Andrea Simonetti (University of Palermo) • Coauthors: Alessandro Albano, Michele Tumminello, Tiziana Di Matteo

Network science uses word co-occurrence statistics to represent a collection of documents, based on the idea that words with similar meanings tend to appear together. However, analyzing word co-occurrence networks is challenging due to their high link density. To address this challenge, we present a novel methodology, named Word Co-occurrence SVN Topic Model (WCSVNtm). The method is based on Statistically Validated Networks (SVN), that represents a corpus as a bipartite network of words and sentences, as shown in Step 1 of Figure 1. The SVN method uses a statistical test to project only the links among words whose co-occurrences among sentences are statistically significant. The methodology involves four-steps:

1. We create a bipartite network with words and sentences as the two set of nodes. Then, we apply the SVN method obtaining a word-co-occurrence network.
2. We construct a new bipartite network with documents and the word-pairs resulting from the validated network of words in Step 1. Then, we apply again the SVN method to create a network of documents.
3. After constructing the two networks, we apply a community detection algorithm to each one of them. Then, we cluster similar documents together and retrieve communities of words, which are interpreted as topics. By doing so, we automatically identify highly coherent topics based on word co-occurrences.
4. Finally, after identifying the topics, we assess their association with each document by testing the shared words using Fisher's exact test.

We evaluate the performance of the methodology across three datasets, where each document is assigned to a category: 120 Wikipedia articles, the arXiv10 dataset(100,000 scientific abstracts), and a sampled subset of 10,000 documents from arXiv10. To benchmark our results, we compare our approach against the state-of-the-art models in topic modeling and document clustering, such as the hierarchical Stochastic Block Model (hSBM), BERTopic, and Latent Dirichlet Allocation (LDA). The results show that WCSVNtm achieves competitive performance in both document clustering and topic modeling and across datasets of different sizes, showing robustness to stochastic variability and ensuring reproducibility. Moreover, the proposed method addresses several challenges faced by state-of-the-art models across the datasets. In LDA, the number of topics must be selected beforehand, while in hSBM, the hierarchical clustering level needs to be specified. Additionally, BERTopic generates an excessively large number of topics for the largest datasets but only two topics for the smallest dataset, requiring further model optimization. In contrast, the proposed model automatically determines the number of topics and document clusters, avoiding the need for prior knowledge or additional tuning for optimization.

Finally, we demonstrated that applying the SVN method to construct the word co-

occurrence network effectively captures meaningful semantic connections among words while filtering out irrelevant ones. Future advancements in community detection algorithms could further enhance our approach.

Contact helps dispreferred combinations of typological features to survive: Geospatial evidence

Deepthi Gopal (Uppsala university) • Coauthors: Henri Kauhanen, Christopher R. Kitching, Tobias Galla, Ricardo Bermúdez-Otero

Why do certain combinations of typological features -- such as OV order and prepositions -- persist at low but crucially non-zero frequencies in the world's languages? We propose and statistically test two competing hypotheses concerning such 'dispreferred types'. First, the "Goldilocks vertical hypothesis" maintains that low but non-zero frequencies emerge purely from vertical (genetic) transmission with finely-tuned ingress and egress probabilities (cf. Kauhanen et al. 2021): dispreferred typological types arise because they are seldom entered into by languages but are frequently exited. By contrast, the "horizontal support hypothesis" maintains that contact with neighbouring languages helps maintain dispreferred feature combinations: for instance, a language may stay in the dispreferred OV+prepositions state longer than otherwise expected, if horizontally supported by neighbouring OV languages.

Using data from both WALS and Grambank, we examine pairs of binary features and statistically test for the presence of under- and over-attested language types (i.e. cells in the corresponding tetrachoric tables whose frequencies are lower or higher than expected by chance). We then apply Jäger & Wahle's (2021) Bayesian procedure in order to identify those types whose under- or over-attestation can be credibly attributed to feature interaction, rather than to phylogenetic bias: dispreferred types, in this sense, are those whose underattestation is due to interaction, rather than to phylogeny. A novel measure of "neighbourhood entropy" is then introduced to characterize the amount of variation or diversity in a language's immediate geographical neighbourhood. With this measure, we demonstrate that dispreferred language types overwhelmingly have more diverse neighbourhoods than expected at random, whilst the same does not apply either to preferred types or to underattested types whose low frequency is due to phylogeny rather than feature interaction. These results provide strong support for the horizontal support hypothesis over the Goldilocks vertical hypothesis.

In conclusion, we provide one of the first large-scale quantitative demonstrations of the role played by language contact and horizontal transmission in the emergence and maintenance of typological (near-)universals. Possible extensions of our framework, such as the study of non-binary features, will be briefly discussed.