Synthetic linguistic data

Giuseppe Samo (Lab of Biolinguistics and Brain Sciences, Beijing Language and Culture University)

submitted 18/10/2026, accepted 24/11/2026, published 25/11/2026
https://doi.org/10.62408/ai-ling.v6i1.112

samo(at)blcu.edu.cn

Abstract
Synthetic linguistic data have become increasingly prominent in linguistic research, from the analysis of large collections of AI-generated texts to controlled computational experiments investigating grammatical and lexical preferences. This glossary entry extends Bresnan's (2016) distinction between the linguistic garden and the bush by introducing an artificial plant nursery as a metaphor for synthetic data. We distinguish between two complementary forms of synthetic linguistic data according to how they are collected and investigated: synthetic bush-like data, consisting primarily of corpus-like collections of model-generated language, and synthetic garden-like data, in which model outputs are elicited or evaluated under controlled conditions to investigate specific linguistic hypotheses. We argue that synthetic data constitute a distinct object of linguistic investigation; comparing them with natural language data can reveal both shared patterns and divergences, but their interpretation requires careful consideration of the conditions under which they are generated.

Keywords
artificial intelligence, synthetic, large language models, corpora, experimental methods

1 Introduction

We define synthetic linguistic data as linguistic data that do not originate from human linguistic production or observation, but are instead generated through a modelling procedure. What distinguishes these data is their provenance: they result from a model, whether in the form of simulated data, natural-language outputs, or measures designed to evaluate such outputs.

The emergence of these new types of data and their analysis echo a long-standing debate in linguistics concerning the relationship between different approaches to linguistic research. The role of experimental methods and the use of corpora have long been debated in the history of linguistics, particularly with respect to the relationship between controlled linguistic judgements and naturally occurring language data (Gibson and Fedorenko 2013; Sprouse and Almeida 2013). Bresnan's (2016) distinction between the garden and the bush provides a powerful metaphor for capturing this methodological contrast. The garden represents carefully controlled data: sentences and minimal pairs tested through introspective or large-scale judgements, but also curated corpora and grammatical examples designed to describe specific phenomena or test specific hypotheses. By contrast,

CC-BY-NC-SA 4.0

the bush refers to naturally occurring language data, as found in corpora (Gries 2008). This glossary entry argues that the current landscape of linguistic data requires an extension of Bresnan's metaphor to include a third space, an artificial plant nursery, dedicated to synthetic data.

Synthetic and simulated data in general have long been integral to data science and predictive modelling (Nobani et al. 2026 and references therein). They exist with respect to linguistic data, although their boundaries are considerably more complex to define (Samo and Merlo 2019, 2021; Merlo and Samo, to appear).[1]

One form of synthetic data typically associated with the term is the corpus-like output of ConversationalAIs based on large language models (LLMs) (De Cesare 2026), which we present in Section 2. However, synthetic data also include other forms of model-generated material obtained and analysed under other types of carefully controlled conditions. Examples are discussed in Section 3.

In both cases, these types of linguistic data are neither straightforwardly garden data nor direct samples from the bush. Like plants in a nursery, they may resemble natural specimens, while exhibiting properties that distinguish them from both naturally occurring and experimentally elicited data. Importantly, the distinction between synthetic and non-synthetic (natural) data concerns their provenance, whereas the garden/bush distinction concerns how linguistic data are collected and investigated (with possible intersections). Building on the distinction between the garden and the bush, this glossary entry examines two complementary roles of synthetic data in linguistic inquiry: (i) synthetic bush-like data and (ii) synthetic garden-like data.

Before turning to the definitions, an important question is whether synthetic linguistic data, and the outputs of large language models in particular, can provide evidence for theories of language. Piantadosi (2023) argues that the success of LLMs may suggest that these models may constitute viable theories of language. Chesi (2025), while acknowledging their strong empirical performance, questions whether predictive success alone is sufficient to establish explanatory adequacy. Other positions argue that synthetic data may not meaningfully contribute to the study of human language as a cognitive system (Katzir 2023; Moro et al. 2023; Shi et al. 2025). These latter positions raise a broader methodological question concerning the relationship between the empirical evidence from language models

[1] A form of synthetic data also consists of simulated data derived from computational models, such as theory-driven counts or expected frequencies. Such data have long been employed in quantitative computational syntax, particularly in studies comparing competing theoretical accounts (e.g., analyses of Greenberg's Universal 20; Merlo 2015; Merlo and Ouwayda 2018), and are discussed in greater detail in Merlo and Samo, to appear. Moreover, although synthetic linguistic data can also be produced through rule-based procedures, the present entry focuses primarily on data generated through neural language models. This focus reflects a distinction between deterministic approaches, in which outputs are largely determined by explicitly specified rules or templates, and probabilistic approaches, in which outputs are sampled or predicted on the basis of learned model distributions. Other forms of synthetic data generation, including machine translation, therefore fall within the broader scope of the concept, but are not discussed in detail here.

and the evidence required to formulate and evaluate linguistic theories. In this glossary entry, we treat natural data and synthetic data as distinct objects of investigation, building on Rizzi (2025). This distinction allows us to examine what each type of data can reveal about language without assuming that evidence obtained from one can be reduced to the other.

Finally, synthetic data can be used to evaluate computational models through large-scale benchmarks, such as the CALAMITA challenge for Italian (Nissim et al. 2026), which draws on different types of data to assess models' capabilities. Synthetic data can also be used for training models, raising a different set of questions, including whether generated examples reinforce existing biases or limitations (Zhang and Pavlick 2025, Samo and Merlo 2026a).

2 Synthetic bush-like data

One of the most visible forms of synthetic linguistic data is the output of ConversationalAIs based on LLMs, which has brought the Turing test (Turing 1950) and Searle's Chinese room argument (Searle 1980) back into both academic and public debate. Outputs of ConversationalAIs can be collected into corpus-like datasets in order to investigate the properties of generated language (Cegin et al. 2026). Each model can be treated as a source of linguistic data, somewhat analogous to an informant, although not equivalent to a human speaker or a population of speakers. The units of observation in such datasets may include models, model versions, prompts, and sessions, which need to be distinguished when designing and interpreting analyses. The relationship between these outputs and naturally occurring language data, however, remains methodologically complex. Although generated texts can be assembled into collections resembling conventional corpora, they are produced in response to prompts and under specific interactional conditions. In this respect, they can also be understood as elicited data (Di Caro & Samo 2026), whose properties depend not only on the model itself but also on the prompts and other generation parameters (e.g., temperature). Importantly, repeated generations from the same model should not be treated as independent human participants. Similarly, persona prompting may encourage a model to produce linguistic features associated with a given variety or speaker population, but this does not, in itself, establish that its outputs are representative of that population. Such claims require validation against linguistic data from the corresponding population.

Mikros (2025), for instance, investigates GPT-4o's ability to imitate the literary styles of Ernest Hemingway and Mary Shelley. By comparing model-generated texts with the authors' original works using several stylometric methods and three prompting strategies, the study shows that GPT-4o can reproduce some surface-level stylistic features. Although in-context learning improves stylistic alignment, the generated texts retain similarities to more generic outputs generated by the model.

The comparison between human and machine-generated language has become a central concern in both academic research and public debate. An excellent example is the summary in De Cesare (2026) that examines the linguistic, textual, pragmatic, and stylistic properties of *synthetic* Italian, identifying features that distinguish AI-generated texts from naturally produced Italian (De Cesare 2023, 2024a). One aspect of this discussion concerns the pivotal role of English on the linguistic representations of current language models (Wendler et al. 2024). Such findings provide a possible context for understanding the English fingerprints observed in synthetic Italian texts (see De Cesare 2024b), although they do not establish a general causal explanation for English-like features across LLMs. A separate issue concerns the language of prompting: instructions formulated in English can affect the characteristics of outputs generated in other languages, such as Italian (Cicero 2023; Nastase et al. 2025).

The use of generated outputs as linguistic evidence is not restricted to the description of synthetic language itself. Since the emergence of ConversationalAIs based on LLMs, theoretical linguists have increasingly used them to investigate questions traditionally addressed through corpus analysis, experimental elicitation, or grammaticality judgements. Haider (2023), for example, examines ChatGPT's grammatical competence through its responses to questions concerning English and German syntax. Merlo (2023) investigated the capacity of ChatGPT to generalise over linguistic patterns using a task designed to test rule-based generalisation (e.g., subject-verb agreement). Although the model could follow the instructions and identify correspondences between objects and attributes, the study of its output shows that it failed to infer the underlying structure of the task and consistently produced incorrect solutions, even with considerable prompting assistance.

Similarly, Massaro and Samo (2023) investigated ChatGPT's metalinguistic awareness of word-order reorderings in Italian and in less resourced-Italian varieties. The authors found that ChatGPT tended to label topicalisation (clitic left dislocation, Cinque 1990) structures, for example *Il libro, Mario lo ha preso* 'the book, Mario has taken it' as archaic or antiquated, despite evidence that these constructions occur in contemporary Italian (for example, around one fifth of direct objects are fronted in data from Italian extracted from morpho-syntactically annotated treebanks, see Massaro and Samo 2023). The study also identified difficulties in interpreting object clitics (e.g. *lo* in *Mario lo legge*) and suggested that more resourced languages may overshadow minor varieties (e.g. Neapolitan with respect to the variety spoken in San Marco in Lamis, Apulia) in the model's linguistic output.

More broadly synthetic outputs have become an object of systematic evaluation across a wide range of tasks and domains. Beyond linguistic phenomena, research has examined models' ability to represent real-world patterns or scientific knowledge (Wysocka et al. 2024), or detect for cultural artifacts (Ghaboura et al. 2025). A related issue is sycophancy, the tendency of ConversationalAIs to agree with or validate users rather than critically assess their claims. Cheng et al. (2026),

for example, show that this tendency extends to situations in which users describe harmful or questionable behaviour, and that sycophantic responses can influence users' subsequent judgments. Similarly, synthetic outputs can be investigated to explore prototypical semantics. For example, Samo and Ursini (2024) shows that the semantic profile of the Italian geonym *piazza* 'square' in synthetic data differs from the patterns observed in natural-language sources, including general-purpose corpora and specialised sources.

These evaluations serve different purposes: generated outputs may be assessed to probe the capabilities and limitations of the models themselves, or the results may justify the creation of dedicated datasets to redirect models (Samo et al. 2023) so that linguists can harvest more data representative of more linguistic populations.

Finally, application-oriented research can help identify the limitations of synthetic data and caution against their uncritical use in real-world settings. For example, studies on AI-generated facilitator companions for Emergency Medical Teams operating in deep-field settings (Samo and Parker 2025, to appear) examine the use of ConversationalAIs' synthetic outputs in crisis response. In Samo and Parker (to appear), for instance, disaster scenarios generated by DeepSeek-V3.1 and ChatGPT-5 were evaluated in English, Italian, Chinese and Tagalog. These outputs were evaluated in terms of both their linguistic properties and their plausibility and suitability for the intended application by expert practitioners in the field and native speakers of the language under investigation. While the outputs in the first three languages received good evaluations in terms of language and content, those in Tagalog were rated poorly by speakers. Interestingly, a cross-LLM-as-a-judge comparison showed that both models rated the other model's Tagalog outputs extremely high.

Summing up, synthetic bush-like data can be readily collected from ConversationalAIs and analysed using corpus-linguistic methods. However, their apparent ease of creation should not obscure the methodological challenges they raise. To ensure that findings are interpretable and replicable, each dataset should document relevant generation conditions, including the model used, the prompt instructions (content and language), and possibly the date of generation. Systematically documenting these conditions is therefore essential to comparing results across studies and to making generalisations about synthetic language.

3 Synthetic garden-like data

The definition of synthetic garden data is more complex. In contrast to corpus-like collections of generated texts, these data are constructed under controlled conditions to investigate specific linguistic hypotheses. In this respect, some forms of synthetic data can be understood as an extension of experimental methods traditionally used in psycholinguistics.

Early neural approaches to language modelling already reflected a quasi-psycholinguistic approach to linguistic data, using computational models to investigate lexical preferences and grammatical constraints, such as subject–verb agreement, through controlled prediction and evaluation tasks. In autoregressive language modelling, models predict the next token based on the preceding linguistic context. This approach was used to investigate syntactic dependencies in recurrent neural networks (Linzen et al. 2016; Bernardy and Lappin 2017; Gulordava et al. 2018). Related studies also employed supervised classification tasks, such as explicit number prediction and grammaticality judgements, to evaluate models' sensitivity to syntactic constraints (Linzen et al. 2016). A different approach is masked language modelling, in which a token is masked within a linguistic stimulus and the model is asked to predict it based on the surrounding context. Fill-mask procedures allow researchers to investigate lexical preferences and grammatical constraints under these controlled conditions (e.g. Devlin et al. 2019). A parallel can be drawn between the use of synthetic linguistic data and psycholinguistic methods for investigating the expectations and preferences of language models. When carefully designed, synthetic data can illuminate the interactions between grammatical constraints and probabilistic preferences and potentially reveal parallels with language-processing mechanisms: one prominent example is *surprisal*, a measure of the information-theoretic unexpectedness of a linguistic unit given its preceding context. Formally, *surprisal* is calculated as the negative logarithm of the conditional probability of a word given its context: the lower the probability assigned to a word, the higher its *surprisal* (Hale 2016 and references therein; see also Greco et al. 2023). Conversely, lower *surprisal* indicates that a word or continuation is more predictable for the model. It is important to bear in mind, however, that lower *surprisal* does not necessarily correspond to greater grammaticality: it reflects the model's probabilistic expectations, which may or may not align with human judgements and processing patterns, and may depend on the lexical choices of the experimental procedures (Samo and Merlo 2024, Samo and Chen 2026).

Wilcox et al. (2024) provide a prominent example of this approach. They use surprisal-based measures to investigate whether neural language models capture the constraints governing filler-gap dependencies in English, including restrictions associated with syntactic islands. By examining how the presence of a wh-filler affects the predictability of subsequent gaps and potential gap positions, their study tests whether models encode expectations about where gaps can occur. Their results provide evidence that neural language models can acquire aspects of these dependencies and generalise beyond individual examples, while also highlighting differences between model behaviour and human linguistic expectations. More broadly, *surprisal*-based methods have been applied across languages and syntactic phenomena. For instance, Samo and Merlo (2024) shows that surprisal patterns associated with movement in French clefts show similar trends of those observed in human processing (see Chesi and Canal 2016), while Samo and Chen (2026) demonstrates that small-scale transformer models already assign lower ratings (higher *surprisal*) to wh-ex-situ than to wh-in-situ configurations in Chinese.

Fill-mask procedures can be used not only to evaluate models on carefully curated linguistic stimuli, but also to expand existing datasets by generating alternative lexical items or sentence variants. In this way, they provide a relatively straightforward means of automatically augmenting linguistic datasets with model-generated examples, which can subsequently be used for large-scale evaluation and the construction of synthetic benchmarks. The Blackbird Language Matrices (BLM) datasets (e.g., An et al. 2023; Jiang and Merlo 2026) provide examples of how model-generated material can contribute to the construction and expansion of linguistic evaluation datasets (Merlo et al. 2023).

Sentence embeddings can also be analysed to investigate how language models represent abstract linguistic structure and generalise across lexical variation. For example, Samo et al. (2026) examines the representations of lexicalised sentences and their functional counterparts, in which pronouns and adverbs replace more specific nouns or phrases. The results show that the sentences with only pronominal elements and sentences with NPs occupy different regions of the embedding space, and that their shared syntactic structure does not emerge when either type is considered in isolation. Tokenisation strategies provide a further perspective on the limitations of linguistic generalisation. Samo and Merlo (2026b) have shown that subword segmentation can affect the representation of morphological forms in complex verbal paradigms in Turkish and Hebrew, particularly when language-specific morphological patterns do not align with token boundaries. Similarly, Hopton et al. (2025) investigate how subword tokenisation interacts with the representation of the functional lexicon across languages.

Another approach consists in directly querying conversational AI systems for numerical grammaticality or acceptability judgments. Please note that grammaticality and acceptability are distinct notions: the former concerns compatibility with a given grammatical system, whereas the latter refers to an empirical response that may also be influenced by processing, context, plausibility and other factors. Accordingly, prompted numerical ratings should be understood as observable model outputs whose relationship to either construct requires validation. In any case, such outputs do not in themselves provide evidence of the underlying grammatical knowledge, if any, encoded by the model. For example, Samo and Massaro (2025) investigates syntactic configurations in Romansh and Neapolitan by prompting ChatGPT and DeepSeek to rate sentences on a numerical scale. The results reveal asymmetries across models and constructions: while the predictions concerning violations to Verb Second in Romansh are partially reflected in the ConversationalAIs' ratings, topic–subject agreement in Neapolitan remains challenging. For the Romance phenomena investigated, smaller monolingual transformer models even performed better than ConversationalAIs. On the other hand, related work has explored similar approaches for German, comparing LLM-generated grammaticality judgments with those provided by human participants (Catasso 2026). Chen and Huang (2026) compare the acceptability judgments elicited from GPT-5.5 and DeepSeek-V4-Pro with those of native Chinese

speakers, using examples from *The Syntax of Chinese* (Huang et al. 2009). Both models successfully distinguish grammatical from ungrammatical sentences across eight syntactic phenomena, but their judgments show substantial variation in correspondence with human ratings across phenomena.

These findings illustrate both the potential and limitations of numerical ratings elicited from conversational AI systems as a tool for investigating grammatical preferences across languages with different structural properties and levels of representation in language-model training data. Such ratings constitute observable model outputs and should not be taken as direct evidence of grammatical knowledge or human-like grammatical intuitions.

Synthetic garden-like data allow researchers to test specific linguistic hypotheses by asking models to evaluate carefully constructed sentences. Their strength lies in the possibility of comparing model responses with human judgments and the predictions of different linguistic theories using replicable methods. They can also provide a starting point for in-depth research on human language (Samo and Chen 2026).

4 Conclusions

In this glossary article, we introduced the distinction between natural and synthetic linguistic data, focusing on the different forms that synthetic data can take and the methodological implications of their use. The distinction between the synthetic bush and synthetic gardens highlights two complementary approaches: the collection and analysis of corpus-like model outputs, and the controlled generation of linguistic data to investigate specific linguistic phenomena.

Synthetic data offer new opportunities for linguistic research. At the same time, the interpretation of these data requires careful consideration of the conditions under which they are generated: in particular, similarities between model outputs and naturally occurring language, or between model responses and human judgements, should not automatically be taken as evidence that the underlying mechanisms are equivalent.

Rather than treating natural and synthetic data as competing sources of evidence, we follow Rizzi (2025) in considering them as distinct but potentially complementary objects of investigation. Their comparison can help identify patterns shared by human language use and model behaviour, as well as divergences that require further explanation. Future research should therefore ask not only whether synthetic data shows similarities with natural language, but also whether they can open up new questions for linguistic research.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this contribution.

Acknowledgments

I would like to thank the reviewers for their helpful comments. A special thank you to Si Fuzhen. I am also grateful to the audiences for their interesting comments and questions at the event "Quale Italiano nelle IA?" organised by the Centre for Italian Studies at TU Dresden, Germany, and at the public lecture "Synthetic Cartographic Syntax" at the University of Macau, Macao, China.

References

An, Aixiu & Jiang, Chunyang & Rodriguez, Maria A. & Nastase, Vivi & Merlo, Paola. 2023. BLM-AgrF: A new French benchmark to investigate generalization of agreement in neural networks. In Vlachos, Andreas & Augenstein, Isabelle (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 1363–1374. Dubrovnik: Association for Computational Linguistics.

Bernardy, Jean-Philippe & Lappin, Shalom. 2017. Using deep neural networks to learn syntactic agreement. *Linguistic Issues in Language Technology* 15(2). 1–15.

Bresnan, Joan. 2016. Linguistics: The garden and the bush. *Computational Linguistics* 42(4). 599–617. https://aclanthology.org/J16-4001.pdf (last accessed on 19/03/2026)

Catasso, Nicholas. 2026. Benchmarking AI acceptability and grammaticality in German: A study of ChatGPT and human judgments. *AI-Linguistica* 3(1). 1–41. https://doi.org/10.62408/ai-ling.v3i1.35

Cegin, Jan & Pecher, Branislav & Srba, Ivan & Simko, Jakub. 2026. RoSE: Round-robin synthetic data evaluation for selecting LLM generators without human test sets. In Demberg, Vera & Inui, Kentaro & Marquez, Lluís (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5530–5545. Rabat: Association for Computational Linguistics.

Chen, Xu & Huang, C.-T. James. 2026. 大语言模型的语感探究——基于《汉语句法学》的实验研究 [Exploring the linguistic intuition of large language models: An experimental study based on *The syntax of Chinese*]. 语言科学 [*Linguistic Sciences*] 25(6). 561–572. https://doi.org/10.7509/j.linsci.202610.037405

Cheng, Myra & Lee, Cinoo & Khadpe, Pranav & Yu, Sunny & Han, Dyllan & Jurafsky, Dan. 2026. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science* 391(6792), 1348–1356. https://doi.org/10.1126/science.aec8352

Chesi, Cristiano & Canal, Paolo. 2019. Person features and lexical restrictions in Italian clefts. *Frontiers in Psychology* 10. 2105. https://doi.org/10.3389/fpsyg.2019.02105

Chesi, Cristiano. 2025. Is it the end of (generative) linguistics as we know it? *Italian Journal of Linguistics* 37(1). 3–44. https://doi.org/10.26346/1120-2726-237

Cicero, Francesco. 2023. L'italiano delle intelligenze artificiali generative. *Italiano LinguaDue* 15(2), 733–761. https://doi.org/10.54103/2037-3597/21990

Cinque, Guglielmo. 1990. *Types of Ā-dependencies*. Cambridge, MA: MIT Press.

De Cesare, Anna-Maria. 2023. Assessing the quality of ChatGPT's generated output in light of human-written texts: A corpus study based on textual parameters. *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos* 10. 179–210.

De Cesare, Anna-Maria. 2024a. *Giorgia Meloni, Meloni o la Meloni?* La codifica degli antroponimi femminili in biografie generate da ChatGPT e pubblicate su Wikipedia. *Lingue e Culture dei Media* 7(1–2). 1–20. https://doi.org/10.54103/2532-1803/22388

De Cesare, Anna-Maria. 2024b. Nuove dinamiche di contatto linguistico: le 'impronte digitali' dell'inglese nell'italiano generato da LLM. *Lid'O. Lingua Italiana d'Oggi* 21. 67–91. https://doi.org/10.1400/304493

De Cesare, Anna-Maria. 2026. *L'Italiano Sintetico delle Intelligenze Artificiali*. Firenze: Cesati.

Ghaboura, Sara & More, Ketan Pravin & Thawkar, Ritesh & Al Ghallabi, Wafa & Thawakar, Omkar & Khan, Fahad Shahbaz & Cholakkal, Hisham & Khan, Salman & Anwer, Rao Muhammad. 2025. Time Travel: A comprehensive benchmark to evaluate LMMs on historical and cultural artifacts. In Che, Wanxiang & Nabende, Joyce & Shutova, Ekaterina & Taher Pilehvar, Mohammad (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, 23627–23641. Vienna: Association for Computational Linguistics.

Gibson, Edward & Fedorenko, Evelina. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes* 28(1–2). 88–124. https://doi.org/10.1080/01690965.2010.515080

Greco, Matteo & Cometa, Andrea & Artoni, Fiorenzo & Frank, Robert & Moro, Andrea. 2023. False perspectives on human language: Why statistics needs linguistics. *Frontiers in Language Sciences* 2. 1178932. https://doi.org/10.3389/flang.2023.1178932

Gries, Stefan Th. 2008. Corpus-based methods in analyses of second language acquisition data. In Robinson, Peter & Ellis, Nick C. (eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition*, 406–431. London: Routledge.

Gulordava, Kristina & Bojanowski, Piotr & Grave, Edouard & Linzen, Tal & Baroni, Marco. 2018. Colorless green recurrent networks dream hierarchically. In Walker, Marilyn & Ji, Heng & Stent, Amanda (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1195–1205. New Orleans, LA: Association for Computational Linguistics.

Haider, Hubert. 2023. Is Chat-GPT a grammatically competent informant? https://lingbuzz.net/lingbuzz/007285 (last accessed on 23/11/2026)

Hale, John. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass* 10(9). 397–412. https://doi.org/10.1111/lnc3.12196

Huang, C.-T. James & Li, Audrey Y.-H. & Li, Yafei. 2009. *The Syntax of Chinese*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139166935

Jiang, Chunyang & Merlo, Paola. 2026. Analogical structure, minimal contextual cues and contrastive distractors: Input design for sample-efficient linguistic rule induction. In Demberg, Vera & Inui, Kentaro & Marquez, Lluís (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 485–500. Rabat: Association for Computational Linguistics.

Katzir, Roni. 2023. Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. *Biolinguistics* 17. e13153. https://doi.org/10.5964/bioling.13153

Linzen, Tal & Dupoux, Emmanuel & Goldberg, Yoav. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4. 521–535.

Massaro, Angelapia & Samo, Giuseppe. 2023. Prompting metalinguistic awareness in large language models: ChatGPT and bias effects on the grammar of Italian and Italian varieties. *Verbum* 14. 1–11. https://doi.org/10.15388/Verb.42

Merlo, Paola. 2015. Predicting word order universals. *Journal of Language Modelling* 3(2). 317–344. https://doi.org/10.15398/jlm.v3i2.112

Merlo, Paola & Ouwayda, Sarah. 2018. Movement and structure effects on Universal 20 word order frequencies: A quantitative study. *Glossa: A Journal of General Linguistics* 3(1). 84. 1–35. https://doi.org/10.5334/gjgl.149

Merlo, Paola. 2023. Blackbird language matrices (BLM), a new task for rule-like generalization in neural networks: Can large language models pass the test? In Bouamor, Houda & Pino, Juan & Bali, Kalika (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8119–8152. Singapore: Association for Computational Linguistics.

Merlo, Paola & Samo, Giuseppe. To appear. Generative computational modeling. In Leivada, Evelina & Grohmann, Kleanthes K. (eds.), *The Cambridge Handbook of Minimalism and its Applications*. Cambridge: Cambridge University Press. Preprint: https://lingbuzz.net/lingbuzz/007675

Merlo, Paola & Samo, Giuseppe & Nastase, Vivi & Jiang, Chunyang. 2023. Building structured synthetic datasets: The case of Blackbird Language Matrices (BLMs). In Boschetti, Federico & Lebani, Gianluca E. & Magnini, Bernardo & Novielli, Nicole (eds.), *Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023)*, 292–302. Venice: CEUR Workshop Proceedings.

Mikros, George. 2025. Beyond the surface: Stylometric analysis of GPT-4o's capacity for literary style imitation. *Digital Scholarship in the Humanities* 40(2). 587–600. https://doi.org/10.1093/llc/fqaf035

Moro, Andrea & Greco, Matteo & Cappa, Stefano F. 2023. Large languages, impossible languages and human brains. *Cortex* 167. 82–85. https://doi.org/10.1016/j.cortex.2023.07.003

Nastase, Vivi & Samo, Giuseppe & Jiang, Chunyang & Merlo, Paola. 2025. Multilingual vs. monolingual transformer models in encoding linguistic structure and lexical abstraction. In Bosco, Cristina & Jezek, Elisabetta & Polignano, Marco & Sanguinetti, Manuela (eds.), *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, 826–836. Cagliari: CEUR Workshop Proceedings.

Nissim, Malvina & Croce, Danilo & Patti, Viviana & Basile, Pierpaolo & Attanasio, Giuseppe & Musacchio, Elio & Rinaldi, Matteo & Borazio, Federico & Francis, Maria & Gili, Jacopo et al. 2026. Challenging the abilities of large language models in Italian: A community initiative. *Italian Journal of Computational Linguistics* 12(1). 1–79.

Nobani, Navid & Officioso, Giovanni & Pallucchini, Filippo & Sperlì, Giancarlo & Mercorio, Fabio. 2026. Synthetic data generation: A tertiary study. *Information Processing & Management* 63(6). 104715. https://doi.org/10.1016/j.ipm.2026.104715

Piantadosi, Steven T. 2024. Modern language models refute Chomsky's approach to language. In Gibson, Edward & Poliak, Moshe (eds.), *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett*, 353–414. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.12665933

Rizzi, Luigi. 2025. On the complementarity of generative grammar and large language models. *Italian Journal of Linguistics* 37(1). 145–152. https://doi.org/10.26346/1120-2726-244

Samo, Giuseppe & Chen, Xu. 2026. When cartography meets transformers: A study on syntactic locality and the functional lexicon in Chinese wh-questions. In Si, Fuzhen & Rizzi, Luigi (eds.), *Cartographic Representations at the Interfaces*, 187–198. Oxford: Oxford University Press.

Samo, Giuseppe & Massaro, Angelapia. 2025. Syntactic patterns of Italo-Romance CP-layers in transformers, ChatGPT and DeepSeek: Two case studies from Romansh and Neapolitan. *AI-Linguistica* 2(2). 1–22. https://doi.org/10.62408/ai-ling.v2i2.25

Samo, Giuseppe & Merlo, Paola. 2019. Intervention effects in object relatives in English and Italian: A study in quantitative computational syntax. In Chen, Xinying & Ferrer-i-Cancho, Ramon (eds.), *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, 46–56. Paris: Association for Computational Linguistics.

Samo, Giuseppe & Merlo, Paola. 2021. Intervention effects in clefts: A study in quantitative computational syntax. *Glossa: A Journal of General Linguistics* 6(1). 145. https://doi.org/10.16995/glossa.5742

Samo, Giuseppe & Merlo, Paola. 2024. Distributed computational models of intervention effects: A study on cleft structures in French. In Ledgeway, Adam & Bonan, Caterina (eds.), *It-Clefts: Empirical and Theoretical Surveys and Advances*, 157–180. Berlin: De Gruyter.

Samo, Giuseppe & Merlo, Paola. 2026a. Comparing natural and synthetic structured data: A study of the passive verb alternation in French and Italian. In Hinrichs, Erhard & Nivre, Joakim & Osenova, Petya & Pustejovsky, James & Zinn, Claus (eds.), *Proceedings of the Workshop on Structured Linguistic Data and Evaluation (SLiDE)*, 39–51. Palma de Mallorca: European Language Resources Association (ELRA).

Samo, Giuseppe & Merlo, Paola. 2026b. Modelling the morphology of verbal paradigms: A case study in the tokenization of Turkish and Hebrew. In Oflazer, Kemal & Köksal, Abdullatif & Varol, Onur (eds.), *Proceedings of the Second Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2026)*, 82–94. Rabat: Association for Computational Linguistics.

Samo, Giuseppe & Nastase, Vivi & Merlo, Paola. 2026. Generalization through lexical abstraction in transformer models: The case of functional words. arXiv preprint. https://doi.org/10.48550/arXiv.2609.19887

Samo, Giuseppe & Parker, Rich. 2025. Exploring AI-driven decisions in disaster simulation for emergency medical teams. In Mantas, John & Hasman, Arie & Gallos, Parisis & Zoulias, Emmanouil & Karitis, Konstantinos (eds.), *Global Healthcare Transformation in the Era of Artificial Intelligence and Informatics* (Studies in Health Technology and Informatics 328), 151–152. Amsterdam: IOS Press. https://doi.org/10.3233/SHTI250691

Samo, Giuseppe & Parker, Rich. To appear. Evaluating curated multilingual consumer-facing AI-generated synthetic facilitator companions scenarios for training emergency medical teams in deep field access. In *Proceedings of EFMI STC 2026*.

Samo, Giuseppe & Ursini, Francesco-Alessio. 2024. The semantics of generic terms in toponyms: Formal semantics meets multi-source big data. In Hu, Xuke & Purves, Ross & Moncla, Ludovic & Kersten, Jens & Stock, Kristin (eds.), *Proceedings of GeoExT 2024: Second International Workshop on Geographic Information Extraction from Texts at ECIR 2024*, 1–12. Glasgow: CEUR Workshop Proceedings.

Samo, Giuseppe & Ursini, Francesco-Alessio & Caracciolo, Giuliano. 2022. Quantifying formulaic syntax: A quantitative and computational study on temporal complements in Latin subject relatives extracted from the epigraphic collection of the Catacombs in Chiusi. *Annali di Linguistica AION-L* 11. 129–151.

Searle, John R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3). 417–457. https://doi.org/10.1017/S0140525X00005756

Shi, Zhong & Tian, Yinghui & Si, Fuzhen. 2025. 跨越世纪的图灵之问：国际学界关于ChatGPT语言能力的争论与思考 [The Turing question across the century: Debates and reflections on ChatGPT's language abilities in the international academic community]. 语言战略研究 [*Chinese Journal of Language Policy and Planning*] 2025(1). 75–86. https://doi.org/10.19689/j.cnki.cn10-1361/h.20250107

Sprouse, Jon & Almeida, Diogo. 2013. The empirical status of data in syntax: A reply to Gibson and Fedorenko. *Language and Cognitive Processes* 28(3). 222–228. https://doi.org/10.1080/01690965.2012.703782

Turing, Alan M. 1950. Computing machinery and intelligence. *Mind* 59(236). 433–460.

Wendler, Chris & Veselovsky, Veniamin & Monea, Giovanni & West, Robert. 2024. Do Llamas work in English? On the latent language of multilingual transformers. In Ku, Lun-Wei & Martins, Andre & Srikumar, Vivek (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15366–15394. Bangkok: Association for Computational Linguistics.

Wilcox, Ethan Gotlieb & Futrell, Richard & Levy, Roger. 2024. Using computational models to test syntactic learnability. *Linguistic Inquiry* 55(4). 805–848. https://doi.org/10.1162/ling_a_00491

Wysocka, Magdalena & Wysocki, Oskar & Delmas, Maxime & Mutel, Vincent & Freitas, André. 2024. Large language models, scientific knowledge and factuality: A framework to streamline human expert evaluation. *Journal of Biomedical Informatics* 158. 104724.

Zhang, Lingze & Pavlick, Ellie. 2025. Does training on synthetic data make models less robust? arXiv preprint. https://doi.org/10.48550/arXiv.2502.07164

License


This article is licensed under the *Creative Commons Attribution 4.0 International* license, which permits use, sharing, adaptation, distribution, and reproduction in any medium or format (including for commercial purposes) if certain conditions are met. In particular, you must give appropriate credit to the original author and source, provide a link to the Creative Commons license, and indicate if you have made any modifications. Images or other material from third parties included in this publication are also covered by this Creative Commons license.

An easy-to-understand description of this license can be found at https://creativecommons.org/licenses/by/4.0/deed.en. The legally binding, complete text of the license is also linked there.

About the journal

AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses is a Diamond Open Access journal supported by PUBLIA – SLUB Open Publishing.



DOI https://doi.org/10.62408/ai-ling

ISSN 2943-0070

Disclaimer

The statements, opinions and data contained in this paper are solely those of the respective authors and not of *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* and/or the editors and/or the publishing entity. *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* and/or the editors and/or the publishing entity disclaim any responsibility for injury to persons or damage to property resulting from the ideas, methods, instructions or products mentioned in the content.

The authors of this paper are solely responsible for ensuring that all components of this article, including images or other material from third parties, can be published under the CC BY 4.0 license of the article (see above). *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* and/or the editors and/or the publishing entity decline any responsibility for damages resulting from any copyright infringements by the authors. If the paper was written by only one person, the term “authors” also applies to that person.