Large language models as assistants for the parametric comparison method

Dimitar Kazakov*[1], Thamer Aljohani[1], Andari Karina Anom[5], Maria Anastasova[4], David Carrasco Coquillat[10], Rositsa Dekova[4], Ingrid Nascimento Fernandes[6], Sofia Ferroni[14], Marco Longhin[7], Ruslana Margova[8], Lidija Milković[9], Nurul Qomariyah[5], Reshmi Roy[13], Gaia Sorge[7], Jiabao Wang[11], Hediye Yarahmadi[12], Paola Crisma[3], and Giuseppe Longobardi[2]

*Corresponding Author: dlk2@york.ac.uk
[1]Computer Science, University of York, York, UK
[2]Language and Linguistic Sciences, University of York, York, UK
[3]Humanistic Studies, University of Trieste, Trieste, Italy
[4]Plovdiv University "Paisii Hilendarski", Plovdiv, Bulgaria
[5]Bina Nusantara University, Jakarta, Indonesia
[6]Institute of Language Studies, University of Campinas, Campinas, Brazil
[7]University of Modena and Reggio Emilia, Reggio Emilia, Italy
[8]Big Data for Smart Society Institute (GATE), Sofia, Bulgaria
[9]Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
[10]Faculty of Philology, Complutense University of Madrid, Madrid, Spain
[11]School of Arts, English and Languages, Queen's University, Belfast, UK
[12]Cognitive Neuroscience, SISSA, Trieste, Italy
[13]Institute of Applied Linguistics, University of Warsaw, Warsaw, Poland
[14]Department of Linguistic and Literary Studies, University of Padua, Padua, Italy

The Parametric Comparison Method (PCM) offers a principled way to encode syntactic variation across languages in terms of binary parameters that can subsequently be used for phylogenetic reconstruction. Collecting the relevant data, however, requires trained linguists to elicit grammaticality judgements from native speakers—a process that is slow, expensive, and prone to cross-linguistic inconsistency. We explore the use of large language models (LLMs) as preliminary assistants in this workflow, generating examples or counter-examples for parameter questions. Early experiments have shown that such models have the potential to accelerate knowledge elicitation but also to expose ambiguities in the definitions of parameter manifestations, helping PCM evolve into a more explicit and replicable scientific framework. We introduce a prototype software platform that automates this process, as illustrated on Eastern and Western Armenian. This study is designed as a registered report aiming to quantify the potential benefits of such automation by manually evaluating the LLM output accuracy and recording the proportion of parameter manifestation definitions that required editing, if such were encountered.

1. Introduction

The *Parametric Comparison Method* (PCM) (Longobardi, 2003; Longobardi & Guardiano, 2009) is a framework for analysing cross-linguistic syntactic variation through a fixed inventory of binary parameters. Each parameter captures discrete settings representing cross-linguistically recurrent grammatical distinctions, whose value, once set for a language, contributes to a vector of features compara-

ble across languages. These vectors serve as linguistic metadata for reconstructing phylogenetic relationships, in much the same way that shared genetic or lexical traits reveal historical connections among populations and languages.

Over the past two decades PCM has enabled the construction of a dataset that has been used to infer family trees, measure linguistic distances, and test hypotheses about population history and contact. However, obtaining parameter values remains a time-consuming and expertise-heavy process. Setting each parameter involves answering a certain number of carefully formulated yes/no diagnostic questions and providing supporting examples. Linguists must elicit these examples from native speakers, often in languages they do not themselves speak, requiring elaborate instructions and metalinguistic clarification (Schütze, 1996; Gibson & Fedorenko, 2013; Sprouse, Schütze, & Almeida, 2013). These issues hinder the scale and reproducibility of historical linguistics approaches based on comparative syntax and motivate the search for more efficient, standardised methods of data collection.

Recent research has studied how well large language models (LLMs) align with human judgments of grammatical well-formedness (Hu, Mahowald, Lupyan, Ivanova, & Levy, 2024). For instance, Qiu, Duan, and Cai (2024) compare judgments of ChatGPT and Vicuna across 2,400 English sentences and find "substantial alignment…albeit with LLMs often showing more conservative judgments." More recently, Ide et al. (2025) compare methods for extracting acceptability judgments from LLMs—contrastive prompting, such as in-template probability readout, and "Yes/No" probability computing. Tjuatja, Neubig, Linzen, and Hao (2025) show LLM acceptability scores are more sensible to factors like sentence length and unigram frequency when compared to human speakers.

These studies together illustrate both the promise and the caveats of using LLMs as surrogates for human grammaticality judgements: they can accelerate data generation and mirror human intuitions, but must be carefully calibrated and validated when applied in comparative syntax workflows.

2. Background

The parametric theory underlying the PCM (Crisma, Guardiano, & Longobardi, 2020) attributes two key features to parameters. Firstly, each parameter is associated with at least one linguistic structure, its *manifestation*, that constitutes positive evidence for that parameter. Certain parameters have more than one possible manifestation, but languages do not need to instantiate all of them in order to set a parameter to [+]: encountering evidence for one manifestation is sufficient. Secondly, parameters have a default value, coded as [-], that does not need to be associated with overt evidence: the lack of overt evidence for all the possible manifestations for [+] for a given parameter results in its being set to [-]. For some parameters, it is possible to construct overt evidence for the value [-] in the form of structures that are incompatible with the value [+] of the parameter. Overt

evidence for [-] only exists for about half of the parameters in the dataset, and it is arguably ignored by the language acquirer, see (Crisma, Fabbris, Guardiano, & Longobardi, 2025).

3. Method

3.1. *Workflow*

Large Language Models trained on multilingual text offer a new source of linguistic evidence that can be tapped for preliminary hypothesis generation in comparative syntax. Because LLMs have access to broad statistical evidence and can handle technical terminology, they can be prompted to output examples illustrating the presence of specific syntactic features. In the PCM context, this means that a model could answer the diagnostic questions associated with a parameter by generating grammatical sentences in the language in question that correspond to one of the two possible values.

Our proposed workflow is to keep expert linguists in the loop. For a given parameter and target language, the model will answer each canonical PCM question corresponding to a given manifestation and backs up the answer with a pair of examples (see Table 3). The output will then be evaluated by a native speaker, who, importantly, does not need to provide their own examples if they agree with the ones provided by the model. This division of labour transfers much of the example generation effort to the model while preserving the crucial human validation step. The approach is fast, reproducible, and more easily scaled to multiple languages due to its reduced demand on the interviewer's time, and on the metalinguistic competence of the native speaker.

An additional outcome of our preliminary investigations has been that interacting with LLMs sometimes reveals *ambiguities in the phrasing of PCM questions themselves*. In one instance, a diagnostic question that had long circulated within the PCM framework turned out to admit two distinct readings. This observation prompted a revision of the question wording to eliminate the ambiguity. Therefore, our interviewers will also record any cases of such ambiguities with the aim of computing their proportion within the full set of manifestations studied.

This data has highlighted a second advantage of using LLMs in this setup: the models acting as sensitive probes of formulation precision. When an LLM's output diverges from the intended interpretation, the problem will not always lie with the model. Identifying and correcting such weak points will enhance the internal coherence of PCM and strengthens its status as an operationally replicable, data-driven scientific method. Thus, using LLMs aims to turn them not only into data generators but also catalysts for the methodological refinement of syntactic theory.

3.2. *Tools*

To operationalise this workflow we have developed a software platform that automates parameter setting and example generation. The system takes as input a language name and a list of PCM parameters, each linked to a series of diagnostic questions. It queries an LLM for each question, requesting examples, typically a minimal contrasting pair, together with an explicit rationale. The results are stored in a structured format that records both the model's proposed parameter value and the textual evidence supporting it.

In more detail, the LLM used is chatGPT5 to which we connect programmatically, via the OpenAI API. The *system* prompt used to lay the background to the following queries, and one of the *user* prompts, employed to elicit the actual answers from the LLM are listed in Table 1. The platform includes a lightweight

Table 1: System prompt provides context, user prompt triggers the response

System prompt:
The Parametric Comparison Method (PCM) is a framework for representing cross-linguistic syntactic variation through a fixed inventory of binary parameters. Each parameter captures discrete settings representing cross-linguistically recurrent grammatical distinctions, whose value, once set for a language, contributes to a comparable vector of features across languages. These vectors serve as linguistic metadata for reconstructing phylogenetic relationships, in much the same way that shared genetic or lexical traits reveal historical connections among populations. Every parameter is associated with one or more manifestations expressed in the form of existential statements. Finding evidence for one manifestation is sufficient to set the relevant parameter to [+] even if the parameter has multiple manifestations. The same holds for the overt evidence for [−], where available. I want you to perform the following task: given (1) a reference to a specific language or linguistic variety, (2) the name of a syntactic parameter, and (3) a manifestation statement, state whether that statement is true or false, support your statement by two examples where possible, and state how the veracity of the manifestation statement affects the value of the parameter in question.
User prompt:
The language is Western Armenian. The parameter is FGP, "grammaticalised person". The manifestation is: "One finds morphological alternations on the verb that depend on the speech-role of the subject".

interface for human validation, enabling linguists to mark examples as correct, ambiguous, or erroneous. The accumulated feedback can be used to fine-tune prompts

or to train smaller in-domain (specialised) models. A pilot run with Eastern and Western Armenian showed that the LLM can indeed provide the required answers and illustrative examples, shown unedited and without comments in Table 3.

3.3. *Participants*

A summer school with 20+ participants was held in the summer of 2025 with the explicit goal of training graduate students and academics to use the PCM approach. All participants were invited to take part in the experiments described here targeting the same five parameters (FGP, FGK, FGT, CGE and FSP) for a language of their choice, with 14 providing feedback on the ChatGPT5.2 answers for 12 different languages.

4. Results and Evaluation

We have calculated the proportion of answers to the manifestation questions for which chatGPT's answers are accepted by the native speaker without any corrections (see Table 2). Similarly, we report the proportion of examples provided that have been deemed both grammatical and relevant to the manifestation question. The summary of all results through its mean and standard deviation across all studied languages is as follows: $75.95 \pm 16.05\%$ manifestation accuracy, and $66.02 \pm 17.48\%$ example accuracy. The results suggest that the LLM may be capable of answering some manifestation questions even when its own ability to back its answers with accurate examples is limited. One possible explanation is that the LLM has seen grammatical descriptions of those languages in its training data, even if the training corpus of text in that language was limited.

All results are based on the work of evaluators who have close familiarity with the language in question and underwent the two-week summer school training in PCM. Not all of them have a background in linguistics. For example, the ChatGPT score for Indonesian by a PCM-trained linguist with previous experience and access to native speakers was 77.27% (manifestations) and 75.00% (examples).

5. Limitations and Further Work

LLMs can produce large numbers of example pairs across multiple languages in minutes, ensuring consistency of format and facilitating replication. Moreover, their probabilistic nature may approximate frequency-based intuitions about naturalness, offering a complementary perspective to binary grammaticality judgements. Several caveats remain though. LLMs mirror the biases of their training corpora, may over-fit standard written norms, and lack true linguistic competence or semantic grounding. Their apparent fluency should not be mistaken for introspective access to grammar, so rigorous validation by human experts is indispensable. In the longer term, integrating such tools with existing PCM databases may enable large-scale comparative projects encompassing hundreds of languages. If the resulting datasets prove richer and more consistent, it can open new opportu-

nities for quantitative research in historical syntax, typology, and language phylogeny. By treating LLMs as both assistants and diagnostic instruments, we can strengthen PCM's status as a reproducible, empirically grounded methodology.

Table 2: Accuracy per language. The number of Wikipedia articles in that language are shown as a proxy for the language presence on the Internet.

Language (variety)	Wiki code	number of articles	Manifest.n accuracy	Examples accuracy
Spanish (European variety)	es	2,094,806	77.27%	77.27%
Russian	ru	2,086,328	54.55%	79.55%
Italian	it	1,956,958	86.36%	63.64%
Mandarin (Chinese Wikipedia)	zh	1,524,128	86.36%	65.91%
Modern Standard Arabic	ar	1,299,611	81.82%	68.18%
Iranian Persian (*aka* Farsi)	fa	1,068,196	45.45%	72.73%
Indonesian	id	763,978	97.73%	94.32%
Norwegian Bokmål	no	675,193	86.36%	77.27%
Croatian	hr	229,418	72.73%	52.27%
Sanskrit	sa	12,479	86.36%	63.64%
Old Church Slavonic	cu	1,339	81.82%	25.00%
Yaathe	–	0	54.55%	52.27%

Table 3: ChatGPT5.2 answers on Eastern and Western Armenian (sample)

FGP manifestation	**Language**	**Answer**	**Example**
One finds morphological alternations on the verb that depend on the speech-role of the subject	Eastern Armenian	True	1) ես կարդում եմ 'I am reading'; 2) դու կարդում ես 'you are reading'
	Western Armenian	True	1) ես կը խօսիմ 'I speak'; 2) դուն կը խօսիս 'you speak'.
One finds speech-role-sensitive clitics that double the subject of the verb	Eastern Armenian	False	1) Արամը կարդում է 'Aram is reading' (no subject clitic); 2) մենք կարդում ենք 'we are reading' agreement only)
	Western Armenian	False	1) Ես կը սիրէմ 'I love' (no subject clitic); 2) Անոնք կը խօսին 'they speak' (agreement only).

References

Crisma, P., Fabbris, G., Guardiano, C., & Longobardi, G. (2025). What are your values? Default and asymmetry in parameter states. *Journal of Historical Syntax*, *9*, 1–26. (Article 3)

Crisma, P., Guardiano, C., & Longobardi, G. (2020). Syntactic diversity and language learnability. *Studi e Saggi Linguistici*, *58*(2), 99–130.

Gibson, E., & Fedorenko, E. (2013). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, *28*(1-2), 88–124.

Hu, J., Mahowald, K., Lupyan, G., Ivanova, A., & Levy, R. (2024). Language models align with human judgments on key grammatical constructions. *Preprint (arXiv)*. (arXiv:2402.01676)

Ide, Y. I., Nishida, Y., Vasselli, J., Oba, M., Sakai, Y., Kamigaito, H., & Watanabe, T. (2025). How to make the most of LLMs' grammatical knowledge for acceptability judgments. *arXiv*. (arXiv:2408.09639)

Longobardi, G. (2003). Methods in parametric linguistics and cognitive history. *Linguistic Variation Yearbook*, *3*, 101–138.

Longobardi, G., & Guardiano, C. (2009). Evidence for syntax as a signal of historical relatedness. *Lingua*, *119*(11), 1679–1706.

Qiu, Z., Duan, X., & Cai, Z. G. (2024). Evaluating grammatical well-formedness in large language models: A comparative study with human judgments. In *Proceedings of the workshop on cognitive modeling and computational linguistics (cmcl 2024)* (pp. 189–198). Bangkok, Thailand: Association for Computational Linguistics.

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology.* Chicago: University of Chicago Press.

Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, *134*, 219–248.

Tjuatja, L., Neubig, G., Linzen, T., & Hao, S. (2025). What goes into a LM acceptability judgment? rethinking the impact of frequency and length. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human language technologies (Volume 1: Long papers)* (pp. 2173–2186). Albuquerque, New Mexico: Association for Computational Linguistics.