Generation of constructed languages with artificial intelligence-based methods: Theoretical, methodological and ethical aspects

Jorge Antonio Leoni de León
Universidad de Costa Rica

Keywords: Artificial Intelligence, ConLang, Interlinguistics, Language planning, Large Language Models.

The emergence of Large Language Models (LLMs) in multiple applications of Artificial Intelligence (AI) invites us to consider how it can be used in interlinguistics. Like other areas of knowledge, interlinguistics can benefit from a rigorous and methodical application of LLMs that would have a positive impact on material production and investment in terms of time and effort. Thus, our objective is to explore how LLMs can be a useful tool in optimizing the development process, creative or natural, of constructed languages, based on grammatical, ethical and equity considerations in access to international communication.

Our proposal is valid, both for the creation of a new language (GenLangs (Diamond, 2023) or ConLangs Generation (Heyer, 2021), for example) and for the support of existing languages (such as Esperanto). We find that, within the possible linguistic tasks, they require the proposal of a coherent architecture and the establishment of a solid linguistic criteria. These linguistic tasks include all language levels or aspects: phonetics and phonology (design of phonological structures), morphology (inventory and combinatorial rules), syntax (word order, constraints), semantics (semantic fields, meaning, semantic relations) and pragmatics (language use simulation).

Although the grammar of LLMs is not known, we do know that it is quite close to natural languages, a topic under investigation from various perspectives (Diamond, 2023; Wilcox et al., 2023). Hence, the possibilities also extend to the analysis of the internal coherence of a planned language, the generation of resources for learning or comparative evaluation between languages through automatic translation. Human intervention is always necessary at every stage. From a methodological perspective, it is interesting to note that LLMS have the ability to propose lexicographic definitions for non-existent terms, which are contextually induced (Rodríguez Betancourt and Casasola Murillo, 2023), this opens new possibilities for linguistic modeling. LLMs can be fed with user feedback to evaluate perception, something that would undoubtedly have a positive impact on a design that considers better international communication.

Although Esperanto is already a functioning language, several possible tasks that LLMs can perform offer valuable support, for example, in observing linguistic variation, or creating specific terms. However, the application of LLMs also raises various questions that we must answer in terms of ethical aspects, validity of the results, level of human intervention and quality of the resulting products.

References

J. Diamond. ”genlangs” and zipf’s law: Do languages generated by chatgpt statistically look human?, 2023. https://doi.org/10.48550/arXiv.2304.12191

Heyer, Fabian (2021). *Generating Immersive Conlangs* [Unpublished doctoral dissertation]. Christian-Albrecht University of Kiel. https://doi.org/10.13140/RG.2.2.22160.28168

Rodríguez Betancourt, E., & Casasola Murillo, E. (2023, October 30). *Generación de definiciones de palabras usando modelos de lenguaje generativos*, in the *XI Coloquio Costarricense de Lexicografía*, Instituto de Investigaciones Lingüísticas (INIL), University of Costa Rica. October 27th octobre 2023. [Video]. YouTube. https://youtu.be/I8_PEbR_coU

Wilcox Gotlieb, Ethan, Richard Futrell, and Roger Levy; Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry* 2023; doi: https://doi.org/10.1162/ling_a_00491