Issue 3

PARROTS

On Spanish-Speaking Parrots

Gimena del Rio Riande

August 2022

doi:10.5281/zenodo.6567850 (English)
doi:10.5281/zenodo.6567966 (Español)

1. A Parrot Called MarIA

Spanish is the second most widely spoken language in the world as a mother tongue. Official reports, survey-based studies, and Wikipedia confirm it. And Google can predict it.[1] The data speaks for itself: there are more than twenty countries where Spanish is the official language. At United States universities, Spanish is the most popular second language choice, and there are twenty-three Academies of the Spanish Language — Academias de la Lengua Española — all over the world.[2] But Spanish has been losing the geopolitical and symbolic battle against English as the language of science since the last century, and few technological developments have helped to improve this situation.[3]

In recent years, global research on artificial intelligence (AI) and machine learning has grown exponentially, offering computational linguistics and natural language processing (NLP) a place of relevance in the academic curriculum of different disciplines and fields, such as digital humanities. Nonetheless, the North-South imbalance in this field is evident. Most NLP research is still primarily done in English, and it takes a lot of time to make these resources available in Spanish. Even if they do become available, it is often via multilingual versions that are not as accurate as the English alternative.

Among Spanish-speaking countries, Spain has been taking the lead in AI and NLP research with official initiatives like *Aporta*, published in the Ministry of Economic Affairs and Digital Transformation's portal, and their project, "Tecnologías emergentes y datos abiertos: Inteligencia Artificial."[4] Although the Spanish initiative is welcome and reflects a sustained human and technological work, it takes an absolutely technopositivist approach. There is not much critical reflection on the dangers or limitations in the use of open public data to develop and exploit language models, and nothing is said, for instance, about the rare-earth technologies that are part of a strategic industry and a geopolitical asset that only a few countries in the world take profit from.

> *The so-called Global South is suspicious of technology, as an uncomfortable consumer of foreign Northern developments.*

Last July, the Iberian country unveiled the first major project on language and AI from the National Library of Spain (Biblioteca Nacional de España — BNE). MarIA is the first massive AI model of the Spanish language. MarIA, a RoBERTa model, was born from a large amount of data that the BNE ingested in the MareNostrum supercomputer of the Barcelona Supercomputing

Centre. MarIA's data are the files in WARC format resulting from the tracking and archiving of the Spanish website, which, by law, the BNE scrapes and preserves.[5]

Let's talk about numbers: 59 terabytes of the BNE web archive and 6,910,000 hours of the MareNostrum supercomputer were used to build, curate, and compile this corpus. As a result, 201,080,084 clean and duplicate-free documents were obtained, occupying a total of 570 gigabytes. The second step of the training, based on neural network technology, required 184,000 processor hours and more than 18,000 GPU hours.[6] According to the paper published a few months ago, the released models — between 125 million and 355 million parameters respectively — will be expanded using new and different sources, such as scientific publications of the Spanish Higher Council for Scientific Research (Consejo Superior de Investigaciones Científicas — CSIC) and the Spanish Wikipedia.[7] There are also plans to start training models for other Romance languages like Catalan, Galician, Basque, and Portuguese and for much more complex varieties of Spanish, such as what is usually termed Latin American Spanish or *español de América*.

It comes as no surprise that the so-called Global South is suspicious of technology, as an uncomfortable consumer of foreign Northern developments and as a receiver of critiques from the big nations.[8] A primary example is the accusation that Southern nations do little to reduce environmental damage by perpetuating the use of old technologies, a charge that distracts from discussions on the ways that new bitcoin, blockchain, and AI industries are polluting and damaging the Global South.[9]

I must confess that my perspective on language models and AI is one of a Southern researcher who is quite suspicious both of MarIA's Spanish and its capabilities, especially when it adapts its performance, for instance, to the *rioplatense* Spanish in translation apps, subtitling, chatbots, and automatic language prediction or correction.[10] I find this situation to be an example of what Thomas Hervé Mboa Nkoudou defines as the techno-utopian rhetoric that trumpets the benefits of technological innovations but, paradoxically, rarely refers to the risks or drawbacks associated with the adoption of socio-technical infrastructures.[11]

I am convinced that we need more MarIAs. Still, I think these MarIAs should be explained in an open research ecosystem if they really want to overcome the problems of techno-colonialism and become reusable, reproducible models in those more than twenty Spanish-speaking countries. I am not aware of the research performed by the BNE model in terms of linguistic, geographic, or racial bias, but I imagine that the Spanish Language Academies around the world should also take part in this project. When it comes to the release of a model to the Spanish-speaking world for mass adoption, openness and diversity should be mandatory.

2. A Parrot called BERTIN

The same week MarIA was widely announced in the media, BERTIN was born.[12] The aim of this project was to pre-train a RoBERTa-base model from scratch using Common Crawl during the Flax/JAX Community Event, held July 7–14, 2021. I was impressed when, all of a sudden, there were two RoBERTa models available in Spanish. I decided to interview one of its mentors, Dr. Javier de la Rosa, a young and brilliant Spanish researcher and NLP expert. In the following section, part of the conversation we had is transcribed since it can help us reflect on how, why, and which language models are needed for the Spanish language.

2.1. Javier, BERTIN, and Gimena

Gimena (G): *BERTIN is proposed as a collaborative project. What role do humanists or linguists play (or could they play) in creating datasets and curating outputs?*

Javier (J): The creation of BERTIN has been community-oriented from its inception, made open by and for the community. Programmers, engineers, AI researchers, digital humanists, and computational linguists showed keen interest in taking part in BERTIN. Unfortunately, due to a purely practical matter (we only had funding and resources for ten days), not everyone was able to participate. However, we had very interesting and rich conversations about the orientation and goals of the project. In that sense, I think that one of the aspects both humanists and linguists could contribute to in these kinds of projects is the detection and documentation of bias. In our case, this issue emerged as an afterthought. As we analyzed the model, we realized that it preferred words from European or peninsular Spanish ("coche" vs "auto" vs "carro"), and we understood that the model could suffer from certain geographic biases that had to be highlighted and corrected. Unfortunately, it is not easy to decrease bias after the model is trained. Nonetheless, team members with a linguistic background did a very good job documenting it.

> *Not all biases are bad. For example, one way to have a language model that is capable of understanding other varieties of Spanish could be skewing the dataset in order to magnify a less represented variety.*

G: *How should we create linguistic data or datasets in such a way that they do not perpetuate biases and hegemonic norms?*

J: At the moment, AI in general and language models in particular are very active fields. Toolkits are starting to analyze the data before the training begins, for example, to reduce certain known biases or to find patterns that could become problematic when exploiting the models. However, I must say that not all biases are bad. For example, one way to have a language model that is capable of understanding other varieties of Spanish could be skewing the dataset in order to magnify a less represented variety and produce a model capable of adapting better to diverse varieties.

G: *If you had to teach a course on data curation from a humanistic perspective, what methods or practices would you teach? Can we imagine a future where the humanities become important to data experts?*

It is true that humanists have been curating data from time immemorial, but we are lagging behind and moving too slowly compared to the advances of technology and AI. That being said, there are initiatives to make these two worlds a bit more data-savvy. For example, Thomas Padilla's Collections as Data project has opened a new research line for institutions, especially libraries and archives. This can be seen in language models and other AI initiatives such as the one from the National Library of Norway, which used its digital collection to build a language model for Norwegian, including that of the BNE.[13] Therefore, I believe that if I had to teach data curation, I would do it from this hybrid perspective, pointing out the importance of preservation and the need to have data in formats that are not only interoperable but also (re)usable.

G: *Do you see negative consequences of humanists relying on Google either to use or test their models? Is there a chance that we can develop open source resources for this kind of research? Should we continue to trust big tech companies with questionable ethical practices, or start developing smaller language models?*

Business for large companies does not rely on software, but on hardware and infrastructure. Most of the advances in AI are developed using open technologies, and this is why it has progressed so fast during the last 10–15 years. Both the adoption of free software and the release of models and open source libraries are ways of inviting users to work on those platforms. For instance, if you want to train a model for Basque, you can do it in Google Cloud and then leave. But if Google provides you with the code and allows you to integrate it with its platform, and also teaches you how to apply it in thirty other languages, it saves you a lot of work. Of course, you can also try to train your model on a supercomputer like the MareNostrum that has been used for the BNE model, but access to these types of resources is not easy or direct. You could even buy 120 NVIDIA graphic cards and train them, but it is a material investment difficult to justify when there are on-demand solutions that assure an efficient use of money. What we have tried to show with BERTIN is that some costs can be reduced when training a model. We are still a long way from being able to train these massive models on our personal computers, but we must not stop making progress in this regard.

3. Some reflections

From my point of view, an important conclusion that emerges from the BERTIN project is that, although we cannot escape the need for big data and big machines to produce language models with good performance, the expense of training them can be reduced in terms of time and data. That makes it possible for smaller teams to enter a domain that, for the moment, is the exclusive preserve of large Global Northern companies and institutions. It is also a proposal aware of the needs of a more open approach to research and of the environmental impacts of technology. Far from a techno-utopian perspective, BERTIN is not at war with companies or institutions and it is not aligned with open science activism. It somehow tries to take the best of both worlds. Possibly, this middle ground is where global Spanish AI and NLP projects could take place in the future years.

Although the field of AI goes beyond the horizon of humanistic work, it is evident that language models have many applications with an immediate impact on the humanities and digital humanities. They can improve the outputs of systems for optical character recognition (OCR), stylometry, and authorial attribution. They can help generate automated text summaries, find similarities in textual collections, classify works based on theme, genre, or content, link data based only on information contained in a text, etc. Still, it is important to note that the outputs of these systems must also be subjected to study and criticism, aligning, in this way, with open science initiatives. In this regard, another project by Javier De la Rosa I would like to refer to is ALBERTI, a BERT-based multilingual model for poetry. ALBERTI is capable of completing poems automatically. But in a second stage, humans evaluate and determine which ALBERTI substitutions could be considered more poetic.[14] It's a good

> *Language models . . . can help generate automated text summaries, find similarities in textual collections, classify works based on theme, genre, or content, and link data based only on information contained in a text.*

example of how all models must be evaluated and limited by humans, not only in the humanities but in everyday life. As the proverb goes, “Quien mucho habla, mucho yerra” (“He who speaks too much, makes many mistakes”), which seems true both for humans and machines.

[1] The title and content of this essay play with that of the article by Emily M. Bender et al., “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?,” FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (New York: Association for Computing Machinery, 2021), 610–23, https://doi.org/10.1145/3442188.3445922. This essay and “Sobre los loros que hablan español” do not have a direct translation relationship. Initially, when I was invited to participate in the virtual round table “Machine Predictions and Synthetic Text”, which was held on October 26, 2021, and in which I participated together with outstanding specialists such as Lauren Klein, Ted Underwood, Toma Tasovac and two of the authors of the aforementioned article (Angelina McMillan-Major and Margaret Mitchell), I wrote the text in Spanish, in order to organize my ideas. A week before the event, I rewrote the text in English, trying to escape (with medium or little luck) the insurmountable problem of transferring grammatical structures and expressions from Spanish to English. I would like to thank my dear sister, María Cruz del Rio, for correcting the first English draft and Grant Wythoff and David Rivera for final review. ↩

[2] According to a report published by the Cervantes Institute, 528 million people speak Spanish as a native, second or foreign language. Cervantes Institute, “El español, una lengua que hablan 580 millones de personas, 483 millones de ellos nativos,” Oct. 15, 2019, https://www.cervantes.es/sobre_instituto_cervantes/prensa/2019/noticias/presentacion_anuario_madrid.htm. ↩

[3] In this regard, a very interesting multilingual tool for text discoverability that could be adapted to different disciplines, created by an Argentine biologist and an American bioinformatician, is PanLingua. PanLingua allows searches in the user's language on the bioRxiv.org database using Google Translate to provide automatic translations of the query term into different languages. ↩

[4] For an English version of the documentation of this project, see Alejandro Alija, *Emerging Technologies and Open Data: Artificial Intelligence* (Iniciativa aporta, 2020), https://datos.gob.es/en/documentacion/emerging-technologies-and-open-data-artificial-intelligence. ↩

[5] Gutiérrez Fandiño et al. do not mention whether the large amount of literary text digitized by the BNE was used to feed MarIA. Asier Gutiérrez Fandiño et al., “MarIA: Spanish Language Models,” arXiv, updated Apr. 5, 2022, https://arxiv.org/abs/2107.07253, ↩

[6] Gutiérrez Fandiño et al, “MarIA,” https://arxiv.org/abs/2107.07253, ↩

[7] However, they don’t specify whether “the Spanish Wikipedia” points to the Wikipedia of Spain or of the different Spanish-speaking Wikipedias. ↩

[8] For a very interesting initiative on Global South AI, see Ranjit Singh, “Mapping AI in the Global South,” Data and Society: Points (blog), Jan. 26, 2021, https://points.datasociety.net/ai-in-the-global-south-sites-and-vocabularies-e3b67d631508. ↩

[9] Peter Howson, “Climate Crises and Crypto-Colonialism: Conjuring Value on the Blockchain Frontiers of the Global South,” *Frontiers in Blockchain* 3 (May 2020), https://doi.org/10.3389/fbloc.2020.00022. For an explanation of how the bitcoin industry is causing power outages in Argentina, see “Échale la culpa a Bitcoin (por cortes de luz): el Gobierno apunta a empresas de minería ilegales,” iProUP, Dec. 31, 2021, https://www.iproup.com/finanzas/28621-cortes-de-luz-el-gobierno-busca-granjas-de-minado-de-bitcoin. ↩

[10] *Rioplatense* is the variety of Spanish spoken mainly in Argentina and Uruguay. Wikipedia, s.v. “Rioplatense Spanish,” last modified Apr. 30, 2022, 14:12, https://en.wikipedia.org/wiki/Rioplatense_Spanish. ↩

[11] Thomas Hervé Mboa Nkoudou, “Les makerspaces en Afrique francophone, entre développement local durable et technocolonialité: trois études de cas au Burkina Faso, au Cameroun et au Sénégal” (PhD diss., Université Laval, 2020), https://corpus.ulaval.ca/jspui/handle/20.500.11794/67577. ↩

[12] For more about BERTIN, see “Bertin-roberta-base-spanish,” Hugging Face, updated Apr. 28, 2022, https://huggingface.co/bertin-project/bertin-roberta-base-spanish. ↩

[13] Author’s note: For a description of the project, see Kummervold et al., “Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model,” *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, ed. Simon Dobnik and Lilja Øvrelid (Linköping Univ. Electronic Press, 2021), 20–29, https://aclanthology.org/2021.nodalida-main.pdf. ↩

[14] For more about ALBERTI, see “ALBERTI vs BERT,” Hugging Face, accessed May 15, 2022, https://huggingface.co/spaces/flax-community/alberti. ↩