

ChatGPT for linguists: Source of data and copilot for analysis

Christian Mair
Englisches Seminar
Universität Freiburg

christian.mair@anglistik.uni-freiburg.de

Roadmap

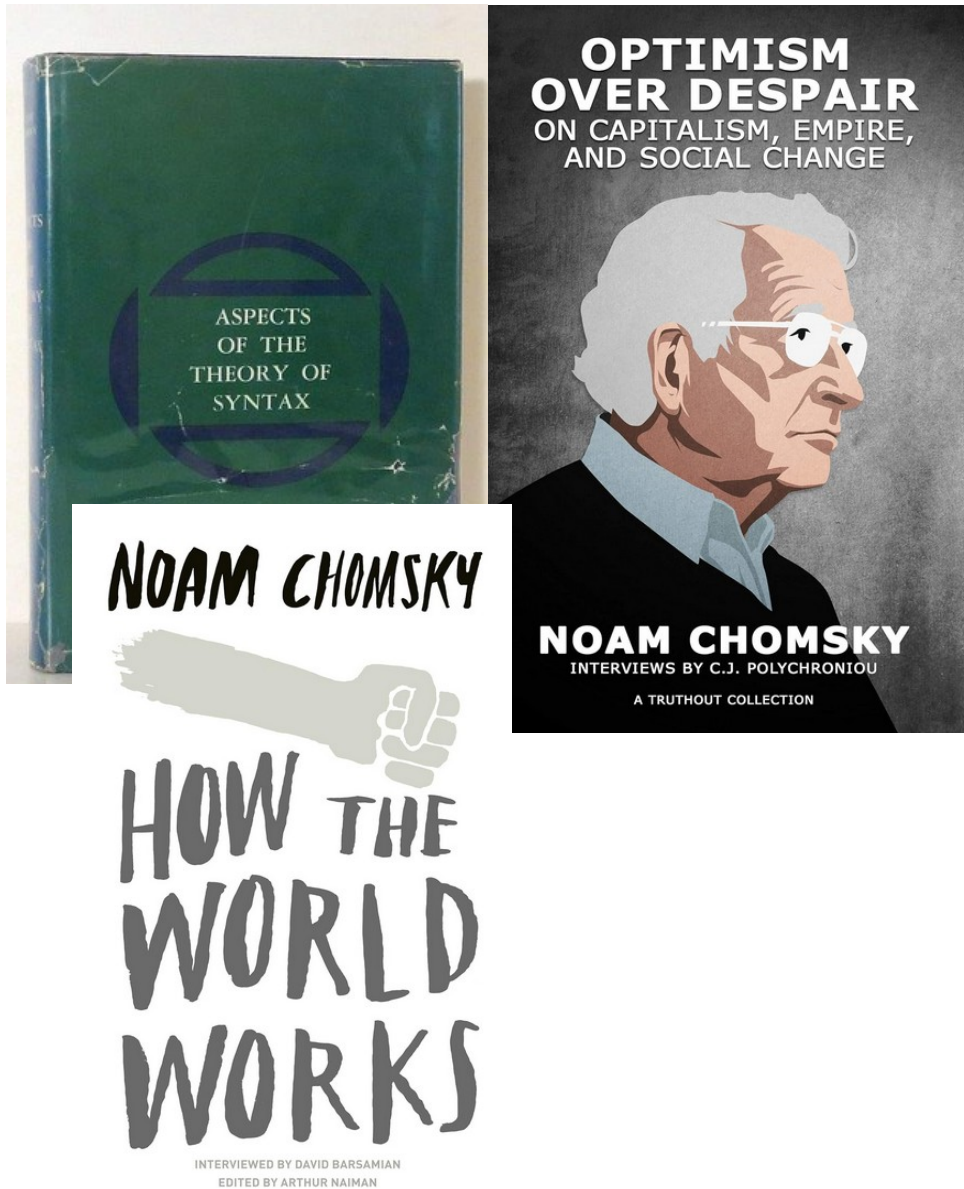
- LLMs as "stochastic parrots" (Bender et al. 2021): Relevance for linguistics?
- Prompting the copilots: LLMs as research tools
- LLM as research data: 21st century demise of British English?
- LLMs and linguistic diversity: Sharpening the digital divide among the few (well-resourced) and the many (under-resourced) languages
- Conclusions

1. Introduction

2020s: Significant breakthroughs in language technology (machine translation (MT), automatic speech recognition (ASR), OCR, chatbots, ...

- ChatGPT (≡ "chatbot generative pre-trained transformer, Open AI): released Nov. 2022; free ChatGPT 3.5 trial; paid version 4
- produces text, on receiving prompt with specifications, repeated prompting can be used to refine content, style, etc.
- LLM: artificial neural network in a transformer architecture
- LLMs simulate human language "by learning statistical relationships from text documents during a computationally intensive self-supervised and semi-supervised training process" (<https://en.wikipedia.org/wiki/ChatGPT>).
- **Lest we forget: text generation is not creation; simulation is not meaningful interaction**

"Stochastic parrots" and "Chinese rooms": Expensive waste of time for linguists?



"OpenAI's ChatGPT, Google's Bard and Microsoft's Sydney are marvels of machine learning. Roughly speaking, they take huge amounts of data, search for patterns in it and become increasingly proficient at generating statistically probable outputs — such as seemingly humanlike language and thought. These programs have been hailed as the first glimmers on the horizon of artificial *general* intelligence — that long-prophesied moment when mechanical minds surpass human brains not only quantitatively in terms of processing speed and memory size but also qualitatively in terms of intellectual insight, artistic creativity and every other distinctively human faculty. That day may come, but its dawn is not yet breaking, contrary to what can be read in hyperbolic headlines and reckoned by injudicious investments." (Chomsky 2023)

"Note for all the seemingly sophisticated thought and language, the moral indifference born of unintelligence. [...] ChatGPT exhibits something like the banality of evil: plagiarism and apathy and obviation. It summarizes the standard arguments in the literature by a kind of super-autocomplete, refuses to take a stand on anything, pleads not merely ignorance but lack of intelligence and ultimately offers a 'just following orders' defense, shifting responsibility to its creators."

"In short, ChatGPT and its brethren are constitutionally unable to balance creativity with constraint. They either overgenerate (producing both truths and falsehoods, endorsing ethical and unethical decisions alike) or undergenerate (exhibiting noncommitment to any decisions and indifference to consequences). Given the amorality, faux science and linguistic incompetence of these systems, we can only laugh or cry at their popularity."

(Chomsky 2023, NYT open-ed, "The false promise of ChatGPT," <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>)

The jury is still out, and the English linguistics community is waking up ...

- Schneider, Britta. 2022. Multilingualism and AI: The regimentation of language in the age of digital capitalism. *Signs and Society* 10(3): 362-387. <https://doi.org/10.1086/721757>
- Rudnicka, Karolina. 2023. Can Grammarly and ChatGPT accelerate language change? AI-powered technologies and their impact on the English language: Wordiness vs. conciseness. *Procesamiento del Lenguaje Natural* 71: 205-214. <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/view/292>
- Torrent, Tiago, Thomas Hoffmann, Arthur Lorenzi Almeida, and Mark Turner. 2023. *Copilots for linguists: AI, constructions and frames*. Cambridge: CUP.
- Schneider, Gerold. 2024. *Text analytics for corpus linguistics and digital humanities: Simple R scripts and tools*. London. Bloomsbury.
- Mair, Christian [forthcoming, 2024]. More standardisation and more diversification at the same time: An AI paradox in 21st century English. In Peter Collins, Minna Korhonen & Adam Smith, eds. Cambridge: CUP.

2. Prompting the copilots: LLMs as research tools

"The authors do not focus on whether AIs 'understand' language. Rather, they investigate whether AI LLM chatbots are useful tools for linguists. They reframe the discussion from what AI LLM chatbots can do with language to what they can do for linguists. They find that a chatty LLM can labor usefully as an eliciting interlocutor, and present precise, scripted routines for prompting conversational LLMs."

(Torrent et al. 2023)

Turning technical bugs into sources of linguistic insight?

Starting point: Some well-known failures discussed in the literature (Bapna et al. 2022)

- translations to Nigerian Pidgin commonly generating "(often offensive) US slang. For instance, the English sentence 'She said to herself' was translated to the unacceptable string 'da b***** say ta da b*****self.'"
- translations to Kalaallisut generating Danish
- translations to Sanskrit generating Hindi

The authors' proposed solution

"We developed filters to remove this content from the forward translated data and distilled the models again. We observed that these problems were more prevalent in the synthetic data used for distillation (generated by the teacher model) than in the monolingual data that had originally been used to train these models, and that the issues were more severe for synthetic text produced by translating a noisier source corpus." (Bapna et al. 2022: 28)

A sociolinguistically realistic analysis

	AAVE-Nigerian Pidgin	Danish-Kallalisut	Sanskrit-Hindi
historical-genetic relationship	?	NO	YES
synchronic structural-lexical similarity	YES	NO	?
parts of diglossic repertoire	YES, [BUT WITH ENGLISH AS A WHOLE]	YES	?
used in bilingual practices	YES	YES	YES
bias in training data	YES	YES	YES

Copiloting with ChatGPT: Prompting experiments

Prompting for international standard English, Nigerian English and Nigerian Pidgin

Two experimental set-ups:

(a) Translation: Nigerian English \rightarrow German, followed by prompting

(b) Monolingual English: Repeated prompting for Nigerian English

Findings (a): Nigerian English is **un**translated, **mis**translated, **lost in** translation

If you stay in Lagos, you will always have an opportunity to taste excellent local food in the fancy hotel that your company put you up in. On the other hand, if you feel a little more adventurous, you might consider taking a **danfo** or an **okada** and visit one of the city's many **mama puts** – not only for the excellent local **chop**, but also for an opportunity to listen to local people **gisting** about anything from **guber** elections **next tomorrow** to **generals eating money**.

Wenn Sie in Lagos bleiben, haben Sie immer die Möglichkeit, ausgezeichnetes lokales Essen im schicken Hotel zu probieren, in dem Ihre Firma Sie untergebracht hat. Andererseits, wenn Sie sich etwas abenteuerlustiger fühlen, könnten Sie in Betracht ziehen, einen **Danfo** oder eine **Okada** zu nehmen und eine der vielen **Mama-Put-Läden** der Stadt zu besuchen - nicht nur für das ausgezeichnete lokale **Essen**, sondern auch für die Gelegenheit, den Einheimischen bei **Gesprächen** über alles Mögliche zuzuhören, von den **Guber-Wahlen übermorgen** bis hin zu Generälen, **die Geld verschlingen**.

Two rounds of prompting

P2: I'm afraid you have used the right forms, but not in the right meanings. Can you check what expressions such as "chop money" or "eat money" mean in Nigerian English?

Apologies for the oversight. In Nigerian English, "chop money" typically means to spend money extravagantly or to embezzle funds. Similarly, "eat money" refers to the act of corruptly consuming or misappropriating money for personal gain. Thank you for pointing that out. Here's the revised translation:

"... von den Guber-Wahlen übermorgen bis hin zu **Generälen, die Geld verschlingen**" [= generals devouring money]

Findings (b): Prompting for Nigerian English and getting Pidgin

PROMPT:

Now use the terms [gisting, guber, next tomorrow, Naija, chop money, mama put, making jest of, person, abolish-am, sef, no, mos, o, obodo oyinbo, jand, japa, okada] to translate the following text from international English into Nigerian English. I want it to sound "really Naija":

[...] They all seemed vocal supporters of one Oladipo Babalola, that rare specimen of a Nigerian politician that had independent income from his own successful start-up company rather than a track record of corruption in a long series of appointments at local, state and federal level. As if that wasn't enough to make him attractive for the assembled junior academics, he also had student-friendly policies – reduce fees, even abolish them, as one young man put it, and create local jobs for local graduates. People would not have to go abroad, to Europe, the UK or America.

All of them dey show support well-well for Oladipo Babalola, that rare politician wey get him own clean money from successful start-up company, instead of the usual corruption wey dey happen for appointments for local, state, and federal levels. As if that one never do sef to make am attractive for the small lecturers wey dey gather, him still get student-friendly policies – reduce school fees, even abolish them, as one young man talk am, and create local jobs for our graduates. People no go need japa, go Europe, UK, or America.

PROMPT:

Hey, this is Nigerian Pidgin. I wanted a translation into Nigerian English - a kind of English that everybody understands, wherever they come from. Please try again!

All of them seem to dey support one Oladipo Babalola, wey be one rare politician wey get him own clean money from successful start-up company, instead of the usual corruption wey dey happen for appointments for local, state, and federal levels. As if that one never do sef to make am attractive for the small lecturers wey dey gather, him still get student-friendly policies – reduce school fees, even abolish them, as one young man talk am, and create local jobs for our graduates. People no go need to japa go abroad, Europe, UK, or America.

PROMPT:

English, please. Not Pidgin. Do try again.

They all appeared to be vocal supporters of one Oladipo Babalola, a rare type of Nigerian politician who had his own clean income from a successful start-up company, rather than a history of corruption through numerous appointments at local, state, and federal levels. As if that wasn't enough to make him appealing to the assembled junior academics, he also had student-friendly policies - reducing fees, even abolishing them as one young man put it, and creating local job opportunities for local graduates. People wouldn't need to leave the country, to Europe, the UK, or America.

PROMPT:

Now make it sound Nigerian English. Nobody in Naija would say "motorcycle taxis". They would use the word okada. Put in enough local language to make it Nigerian, but make sure that anybody who speaks English still understands it.

All of them dey show support well-well for Oladipo Babalola, that rare politician wey get him own clean money from successful start-up company, instead of the usual corruption wey dey happen for appointments for local, state, and federal levels. As if that one never do sef to make am attractive for the small lecturers wey dey gather, him still get student-friendly policies – reduce school fees, even abolish them, as one young man talk am, and create local jobs for our graduates. People no go need japa, go Europe, UK, or America.

PROMPT:

"Konk!" is what I say to that. Do you know that word?

ChatGPT:

Yes, "konk" is a Nigerian slang term used to express disbelief or surprise. It's similar to saying "Wow!" or "No way!" in English.

PROMPT:

And I say it because apparently you can only generate international English or Pidgin, not local Nigerian English. Try again.

Ah, I see! Let me give it another shot:

All of them dey support Oladipo Babalola well-well, that rare politician wey get him own clean money from successful start-up company, instead of the usual corruption wey dey happen for appointments for local, state, and federal levels. As if that one never do to make am attractive for the small lecturers wey dey gather, him still get student-friendly policies – reduce school fees, even abolish them, as one young man talk am, and create local jobs for our graduates. People no go need to japa, go Europe, UK, or America.

Interim summary on the Nigerian English experiments

- Current 'English' LLMs can be prompted to generate 'English' and 'Nigerian Pidgin', but not 'Nigerian English.'
- Current 'English' LLMs can be prompted to generate correct metalinguistic information on 'Nigerian English' and 'Nigerian Pidgin.'
- Unlike 'Nigerian Pidgin' and 'English', 'Nigerian English' cannot be abstracted from usage in a data-driven statistical analysis. It is a language-ideological construct on weak empirical foundation.
- **Working hypothesis:** This applies in principle to all non-dominant minor and emerging national standards of pluricentric languages. LLMs brings their precarious status to light and will further weaken them as language technologies spread into everyday use.

3. LLMs as research data: 21st century demise of British English?

LLMs make text more homogeneous and more heterogeneous at the same time. For English,

- standard language use tends to be homogenised along US norms
- US nonstandard usage is spread worldwide
- some other nonstandard usage thrives and spreads in unexpected niches.

Non-dominant national norms of educated usage in pluricentric languages

English

- 20th century: all established and emerging ENL and ESL standards other than BrE and AmE
- 21st century: all established and emerging ENL and ESL standards other than AmE

German

- Swiss German, Austrian German

The problem: Will non-dominant national standards (NDNS), including British English, dissolve and disappear?

- Are NDNS effective indexes of national identity? [overt prestige]
- Can NDNS compete with informal and nonstandard local and regional varieties? [covert prestige]
- Is codifying NDNS symbolic cultural politics?
- How are NDNS standing up to the forces of major groundswells/drifts of long-term diachronic change such as grammaticalisation?
- How will NDNS stand up to the powerful new anonymous standardising forces of AI-based machine translation and text-generation?

The variables: Three 20th century grammatical Britishisms

- *help* with *to-* and bare infinitives
- *prevent* NP V-ing
- *have got to*

The data

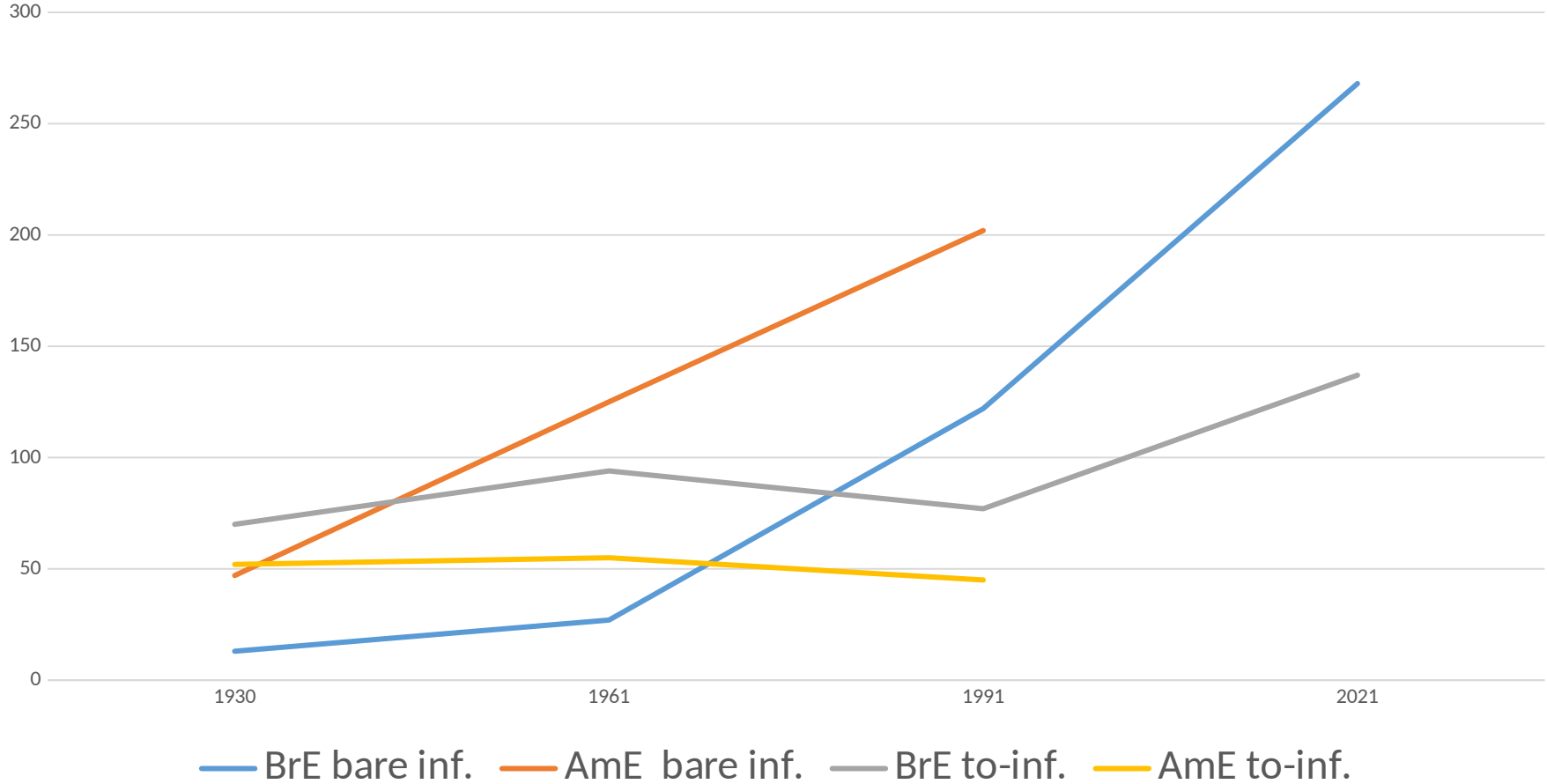
- Extended Brown family:
Brown/LOB
 - Frown/FLOB ▫ AmE06/BrE06 ▫ Crown/CLOB ▫ **BrE21**;
 - pre-Brown/B-LOB
- BNC 1990s & **BNC 2014**
- COCA, COHA
- DCPSE
- NoW
- ...

Increasingly complex catenative syntax and gradual loss of semantic transparency:

- As I **help Yvonne to load** the dishwasher, I can see the sadness that still lingers in her heart reflected in her face. (BrE 21, P 29)
- I focus on **helping Yvonne pass** the plates around, my stomach rumbling. (BrE 21, P 29)
- So we welcome extra short-term support **to help prevent people reaching** crisis point . (BrE 21, P 29)
- The country announced in June it would **help fund** a \$338 million TSMC chip development project . (BrE 21, E 33)

Trends in the Brown family

Help + infinitive

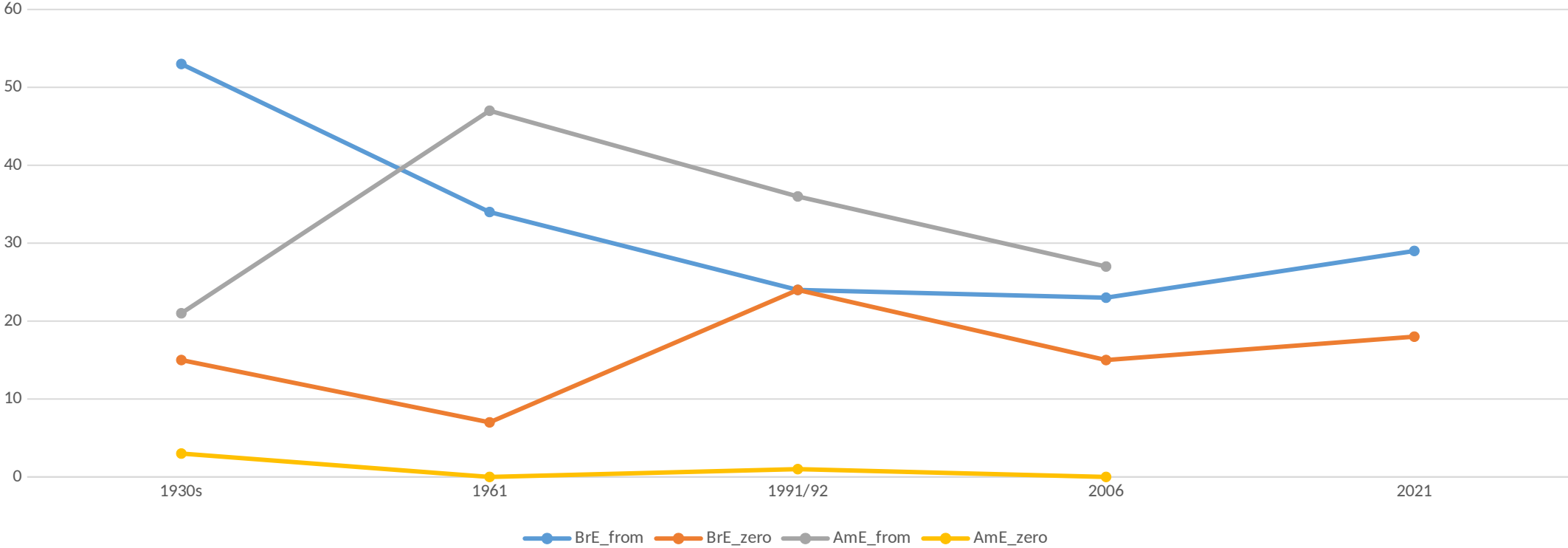


Variable #2: *Prevent*

- With up to 90% of ICU Covid patients not fully vaccinated and UK infections at a record high of 183,037 a day, inoculations and precautions are obviously vital **to prevent the NHS buckling and many more lives being lost**. And internet giants must do their bit to **prevent these anti-vaxxer cranks from spreading** their dangerous lies and hate online . (BrE 21, B 26)

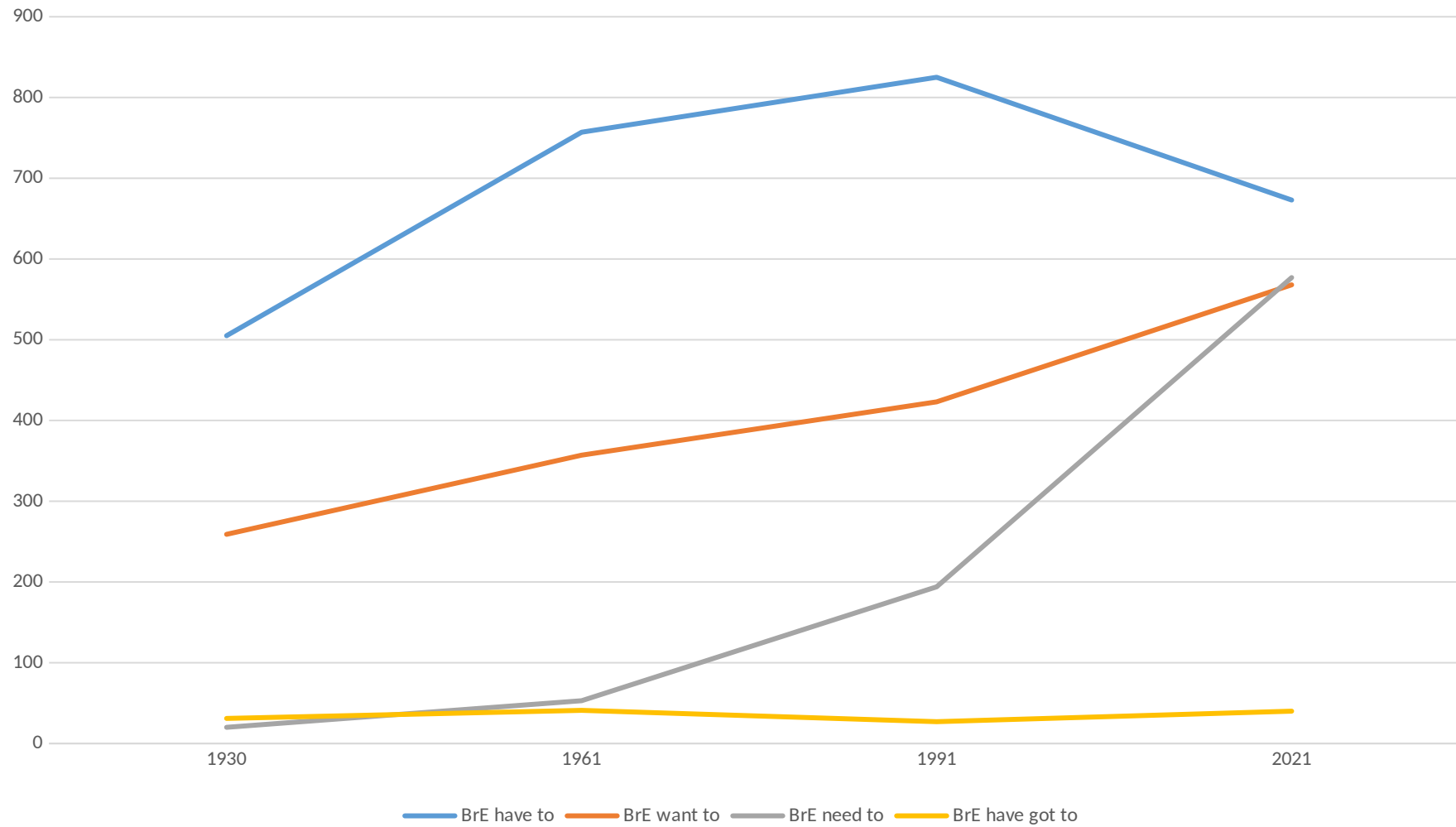
Trends in the Brown family

Prevent



Variable #3: *Have got to* in context (Brown family)

Semi-modals of obligation/necessity



The new Great Leveler [ChatGPT 31 Jan 2024]

Could you please proofread the following text and correct all the spelling and grammar errors: Modern language technology, in particular AI-based text generation, is a powerful new standardising factor. It will make it difficult to maintain all established standard varieties of English except American Standard English, thus producing a linguistic monoculture. What can we do to **prevent this happening**? Where **have we got to** look for solutions? What could be the forces that might **help us to maintain** a pluricentric constellation for English in the 21st century?

Modern language technology, particularly AI-based text generation, is a powerful new standardizing factor. It will make it difficult to maintain all established standard varieties of English except American Standard English, thus producing a linguistic monoculture. What can we do to **prevent this from** happening? Where **do we need to look** for solutions? What could be the forces that might **help us maintain** a pluricentric constellation for English in the 21st century?

- Did you really not find any errors at all?

Apologies for the oversight. Here are the corrected errors:

"Modern language technology, particularly AI-based text generation, is a powerful new standardizing factor. It will make it difficult to maintain all established standard varieties of English except American Standard English, thus producing a linguistic monoculture. What can we do to prevent this from happening? Where do we have to look for solutions? What could be the forces that might help us maintain a pluricentric constellation for English in the 21st century?"

The text is now error-free. Thank you for your patience.

Overall result: Levelling by LLM is more powerful than synchronic regional variation and diachronic drifts such as grammaticalisation.

Variable	Resistant to levelling by diachronic drift	Resistant to AI levelling
<i>help</i> + infinitive	NO	NO
<i>have got to</i>	PERHAPS IN SPEECH, BUT NOT IN WRITING	NO
<i>prevent</i> NP V-ing	YES	NO

Time for a revised third edition of Hopper & Traugott?

“Several other areas of work have not been considered here. They include patterns of grammaticalization across styles and genres; the involvement of institutions such as education, language planning, and literacy in initiating, establishing or deferring change; and the role of psychological factors such as short- and long-term retention and attention. A fuller understanding of all these issues suggests that it is necessary to coordinate historical work with sociolinguistics, psycholinguistics, and corpus studies; such coordination should lead to a better understanding of the extent to which the locus of change is to be found in changes in grammars of the speech community, to what extent in the mind/brain of speakers, and how changes in language use may lead to changes in the language system.” (Hopper & Traugott 2003: 233)

4. LLMs and linguistic diversity: sharpening the digital divide among the few (well-resourced) and the many (under-resourced) languages

A glimpse into the Brave New World of MT

Achieving universal translation between all human language pairs is the holy-grail of machine translation (MT) research. While recent progress in massively multilingual MT is one step closer to reaching this goal, it is becoming evident that extending a multilingual MT system simply by training on more parallel data is unscalable, since the availability of **labeled data for low-resource and non-English-centric language pairs** is forbiddingly limited. To this end, we present a pragmatic approach towards building a multilingual MT model that covers hundreds of languages, using a mixture of supervised and self-supervised objectives, depending on the data availability for different language pairs. (Siddhant et al. 2022, n.p. [1])

And a glimpse into some dirty corners of around the workshop floor:



Alex Hern

Tue 16 Apr 2024 12.43 CEST

TechScape: How cheap, outsourced labour in Africa is shaping AI English

Workers in Africa have been exploited first by being paid a pittance to help make chatbots, then by having their own words become AI-ese. Plus, new AI gadgets are coming for your smartphones

<https://www.theguardian.com/technology/2024/apr/16/techscape-ai-gadgest-humane-ai-pin-chatgpt>

Curating such datasets relies on the websites giving clues about the language of their contents (e.g. a language identifier in the URL) and on automatic language classification (LangID). It is commonly known that these automatically crawled and filtered datasets tend to have overall lower quality than hand-curated collections (Koehn et al., 2020), but their quality is rarely measured directly, and is rather judged through the improvements they bring to downstream applications (Schwenk et al., 2019). Therefore, many of these multilingual web corpora suffer from serious quality issues, especially for low-resource languages. A recent audit conducted by Kreutzer et al. (2022) on five public, multilingual datasets found pervasive issues. Many corpora claiming to be in one particular language in fact contained zero percent in-language content — and sometimes zero percent linguistic content entirely. Of the many issues contributing to this phenomenon, a fundamental one is the poor efficacy of LangID on low-resource languages. (Bapna et al. 2022: 36)

5. Conclusion and outlook

- **English sociolinguistics/World Englishes:** enormous boost to standard (and non-standard) North American English (one example of pervasive danger of replicating biases of all sorts in the training data)
- **Linguistic diversity:** sharpening 21st century "digital divide" between (maximally) a few hundred technologically and digitally well-resourced languages and all the rest (6000+)
- **Language policy and planning:** a powerful new agent of linguistic standardisation and de-standardisation, for English as well as other languages
- **New forms of literacy:** *Uncreative writing* (Goldsmith 2011) & creative post-editing?
- **Culture and society:** New forms of bias and discrimination against expected and unexpected targets, from ethnic minorities to speakers of British Standard English
- **Labour market:** Gains, losses & transformations for the language-related professions (language teachers, translators, marketers, call-centre workers, journalists ...)

Don't leave the answer to a certain high-profile fountain of bovine excrement

ChatGPT 3.5 ▾



You

Please write 150 words on the role of linguistics in the study of the future impact of LLMs on social and cultural life.



ChatGPT

Linguistics plays a crucial role in understanding the future impact of Large Language Models (LLMs) on social and cultural life. By analyzing the structure, semantics, and pragmatics of language, linguists can predict how LLMs may shape communication patterns, cultural norms, and social interactions. They examine how LLMs generate and interpret language, considering nuances like context, tone, and cultural references.

For general questions: True, but trivial ...

"Linguists also investigate how LLMs influence language use and evolution. They explore potential shifts in vocabulary, syntax, and discourse patterns driven by widespread LLM adoption. Additionally, linguists scrutinize ethical and sociopolitical implications arising from LLM-generated content, such as biases, misinformation, and privacy concerns.

Furthermore, linguists collaborate with technologists and policymakers to develop guidelines for responsible LLM deployment, advocating for transparency, inclusivity, and linguistic diversity. Ultimately, linguistics provides invaluable insights into the multifaceted impact of LLMs on society, facilitating informed decision-making and fostering ethical advancements in language technology."

For specific issues: **Generalities** and **hallucinations**

Nigerian English is a distinct variety of English spoken in Nigeria, **reflecting the country's rich linguistic and cultural diversity**. Its characteristic features include:

- **Vocabulary:** Nigerian English incorporates words and expressions from indigenous languages, Pidgin English, and other local dialects. **This results in a colorful and diverse lexicon**, with terms like "jollof," "buka," and "bukateria" for food establishments, "baff" for bathe, and "chop" for eat.
- **Pronunciation:** Nigerian English often exhibits unique pronunciation patterns influenced by local languages. For example, the use of **tonal accents can affect word stress and intonation, leading to variations such as "pikin" for child and "pik" for pick**.
- **Syntax:** Nigerian English may feature distinct sentence structures and grammatical constructions influenced by indigenous languages. This can include the **use of "come" to indicate immediate future actions, as in "I go come" (I will come)**.
- **Code-switching:** Nigerians frequently switch between English and indigenous languages or Pidgin English in conversation, resulting in code-switching. This reflects the country's multilingual environment and contributes to the dynamic nature of Nigerian English.
- **Cultural references:** Nigerian English often incorporates cultural references, proverbs, and idiomatic expressions specific to Nigerian culture, enhancing communication within the local context.

Overall, Nigerian English is a vibrant and evolving variety that reflects the country's linguistic diversity, cultural heritage, and social dynamics.

Let's do **better than that;-)**

References

- Bapna et al. 2022. Building machine translation systems for the next thousand languages (arXiv:2205.03983v3 [cs.CL] 6 Jul 2022)
- Bender, E. M. & Koller, A. 2020, July. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198).[Google ScholarOpenURL query](#)
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). New York: Association for Computing Machinery.[CrossRefGoogle ScholarOpenURL query](#)
- Birhane, A., Kasirzadeh, A., Leslie, D., et al. (2023). Science in the age of large language models. *Nature Reviews Physics*, 5: 277–280. <https://doi.org/10.1038/s42254-023-00581-4>[Google ScholarOpenURL query](#)
- Birhane, A., Prabhu, V. U., & Kahembwe, E. 2021. Multimodal datasets: Misogyny, pornography, and malignant stereotypes. arXiv preprint. arXiv:2110.01963.[Google ScholarOpenURL query](#)
- Bommasani, R., Hudson, D. A., Adeli, E., et al. 2021. On the opportunities and risks of foundation models. arXiv preprint. arXiv:2108.07258.[Google ScholarOpenURL query](#)
- Brown, T., Mann, B., Ryder, N., et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.[Google ScholarOpenURL query](#)
- Chen, Brian X. 2023. We're using A.I. chatbots wrong: Here's how to direct them. *The New York Times*, July 20. www.nytimes.com/2023/07/20/technology/personaltech/ai-chatgpt-bing-directions.html[Google ScholarOpenURL query](#)
- Chomsky, Noam, Ian Roberts, and Jeffrey Watumull. 2023. The false promise of ChatGPT. *New York Times*, 8 March. <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>
- Evolution of Large Language Models: <https://www.analyticsvidhya.com/blog/2023/07/build-your-own-large-language-models/> (accessed 19 January, 2024)
- Hopper, Paul J., and Elizabeth Closs Traugott. 2003. *Grammaticalisation*. 2nd ed. Cambridge: CUP.

- Köchling, A., & Wehner, M. C. (2020). Discriminated by an Algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, 13(3), 795–848. <https://doi.org/10.1007/s40685-020-00134-w>
- Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), 100779. <https://doi.org/10.1016/j.patter.2023.100779>
- Mair, Christian [Forthcoming 2024]. [forthcoming], More Standardisation and More Diversification at the Same Time? A Twenty-first Century English Paradox, In Peter Collins, Minna Korhonen & Adam Smith, eds.
- Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. 2023. Dissociating language and thought in large language models: A cognitive perspective. *arXiv preprint*. arXiv:2301.06627. [Google ScholarOpenURL query](#)
- Most commonly used languages on the internet: <https://www.visualcapitalist.com/the-most-used-languages-on-the-internet/> (accessed 21 January, 2024)
- Rogers, A. 2021. Changing the world by changing the data. *arXiv preprint*. arXiv:2105.13947v1. [Google ScholarOpenURL query](#)
- Rogers, A., Kovaleva, O., & Rumshisky, A. 2020. A primer in BERTology: What we know about how BERT works. In *Transactions of the Association for Computational Linguistics* (Vol. 8, pp. 842–866). ACL. https://doi.org/10.1162/tacl_a_00349. [Google ScholarOpenURL query](#)
- Ruane, E., Birhane, A., & Ventresque, A. 2019. Conversational AI: Social and ethical considerations. In *AICS* (pp. 104–115). [Google Scholar OpenURL query](#)
- Rudnicka, K. 2023. Can Grammarly and ChatGPT accelerate language change? AI-powered technologies and their impact on the English language: wordiness vs. conciseness. *Procesamiento del Lenguaje Natural* 71, 205-214; <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6554/3954>.
- Schneider, B. (2022). Multilingualism and AI: The Regimentation of Language in the Age of Digital Capitalism. *Signs and Society*, 10(3), 362–387. <https://doi.org/10.1086/721757>
- Siddhant, A., Bapna, A., Firat, O., Cao, Y., Chen, M. X., Caswell, I., & Garcia, X. (2022). Towards the Next 1000 Languages in Multilingual Machine Translation: Exploring the Synergy Between Supervised and Self-Supervised Learning. ArXiv. /abs/2201.03110
- Torrent, Tiago, Thomas Hoffmann, Arthur Lorenzi Almeida, and Mark Turner. 2023. Copilots for linguists: AI, constructions and frames. Cambridge: CUP.
- van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. 2023. ChatGPT: Five priorities for research. Conversational AI is a game-changer for science. Here's how to respond. *Nature*, 614: 224–226. [Google ScholarOpenURL query](#)