Linguistic theory, psycholinguistics and large language models

Stela Manova
(Vienna)
manova.stela@gmail.com

Keywords: natural language processing, large language models, linguistic theory, psycholinguistics, phonology/morphology/syntax

Recently much attention has been paid to whether large language models (LLMs) can serve as theories of language (Piantadosi 2023 and replies to other scholars in it). Unfortunately, the discussion has been kept at an abstract level and virtually nothing has been said about how LLMs work technically and what their internal organization means for linguistic theory (LT). My research fills this gap.

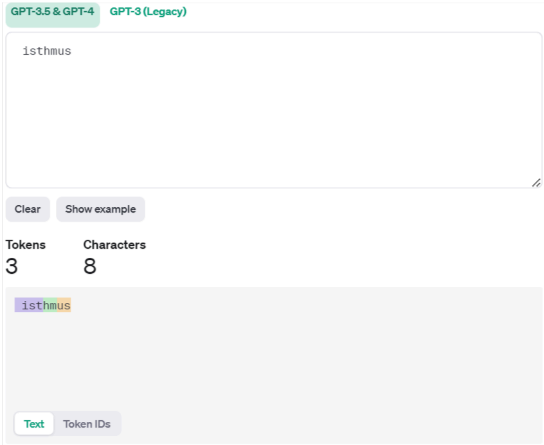
Since algorithms in different LLMs may differ, I focus on ChatGPT. I explain:

(i) the ChatGPT architecture and compare it with claims from A-morphous (cf. Word-Based) Morphology (AM, Aronoff 1976; Anderson 1992; Stump 2001), Distributed Morphology (DM, Halle & Marantz 1993, Bobaljik 2017), Parsability Hypothesis (PH, Hay 2001, 2003, cf. Complexity-Based Ordering, Plag and Baayen 2009), and Chomskyan approach (Chomsky et al. 2023);

(ii) how psycholinguistics can help us check whether humans and LLMs process language in a similar way.

My data come from the ChatGPT tokenizer, https://platform.openai.com/tokenizer. ChatGPT defines an initial set of elements (cf. alphabets in human languages) from which tokens are built. Thus, I start by introducing the ChatGPT tokenization algorithm, Byte Pair Encoding (BPE, Sennrich et al. 2016): https://www.geeksforgeeks.org/byte-pair-encoding-bpe-in-nlp/. ChatGPT has a fixed-size vocabulary, cl100k_base: https://github.com/kaisugi/gpt4_vocab_list/blob/main/cl100k_base_vocab_list.txt, i.e. the vocabulary is limited to 100k and the frequency of the 100k-th token serves as a threshold for vocabulary inclusion. Each token has an ID. Language is a long uninterrupted sequence of tokens in which spaces indicatie word beginnings. Depending on position and frequency of use in this position, a sequence of letters may have different IDs, (1a,b), or may be treated as both non-derived, (2a), and derived, (2b). A smaller ID indicates a higher frequency, e.g., *-ist/*"ist", token ID [380], is more frequent than *ist-/*" ist" in a word-initial position, token ID [6127], (1a,b). Screenshots of the tokenizer search results follow the examples. Note that ChatGPT manipulates not the letters but the token IDs, the second screenshot in each pair.

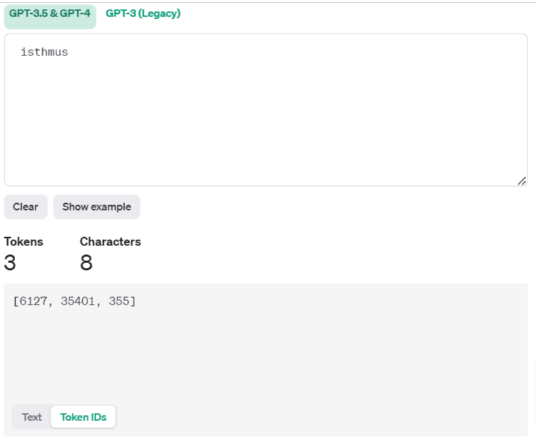
(1) a. Initial position: **" ist",** token ID [**6127**], e.g., as in
" **ist**|hm|us", which consists of three tokens: [**6127**, 35401, 355], i.e.
isthmus is not frequent enough to be included in the vocabulary and is therefore segmented



isthmus

Tokens 3 Characters 8

isthmus



isthmus

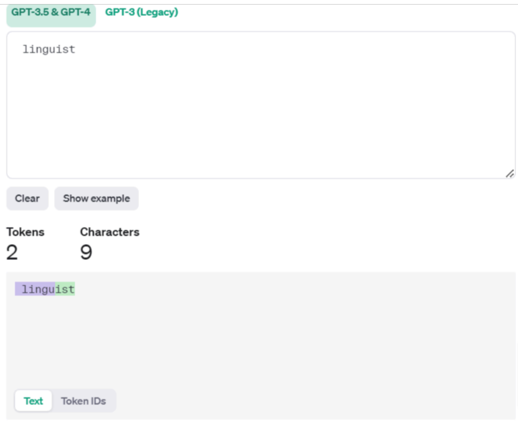
Tokens 3 Characters 8

[6127, 35401, 355]

b. Non-initial position: **"ist",** token ID [**380**], e.g.,

" lingu|**ist**" [39603, **380**]

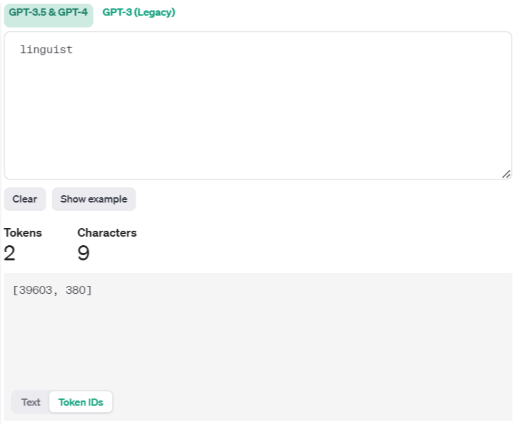
(*linguist* is not frequent enough to be part of the vocabulary but *linguistic* [**65767**] is, (2a)!)



linguist

Tokens 2 Characters 9

linguist



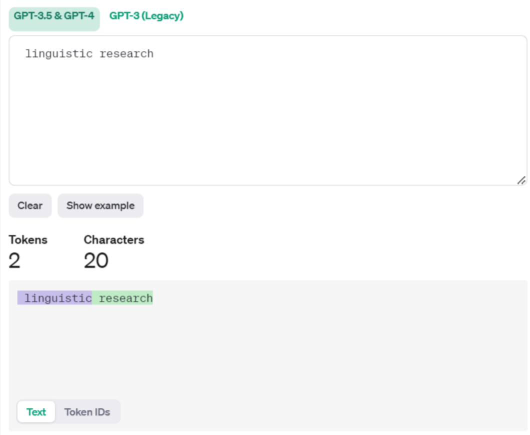
linguist

Tokens 2 Characters 9

[39603, 380]

(2) a. Initial position: **" linguistic"** [**65767**] is in the vocabulary, thus a non-derived item, e.g. in

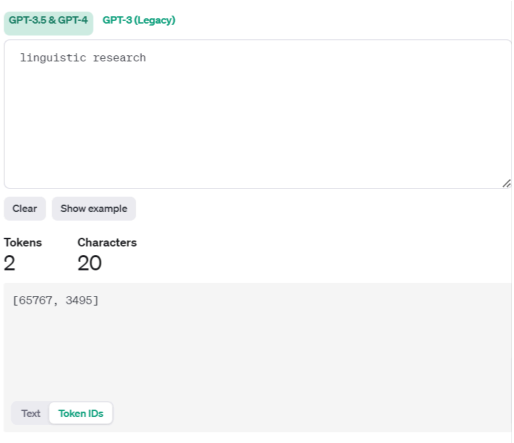
" **linguistic** research" [**65767**, 3495]



linguistic research

Tokens 2 Characters 20

linguistic research

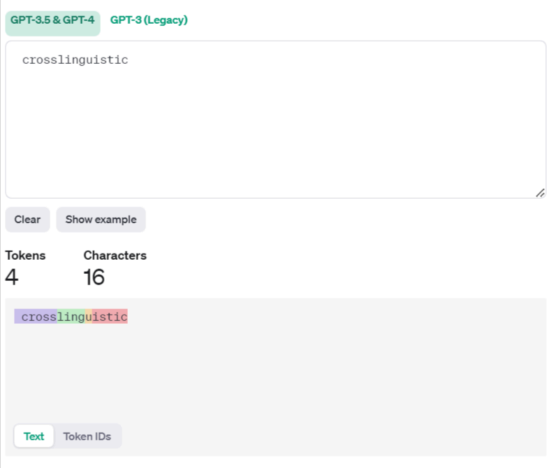


linguistic research

Tokens 2 Characters 20

[65767, 3495]

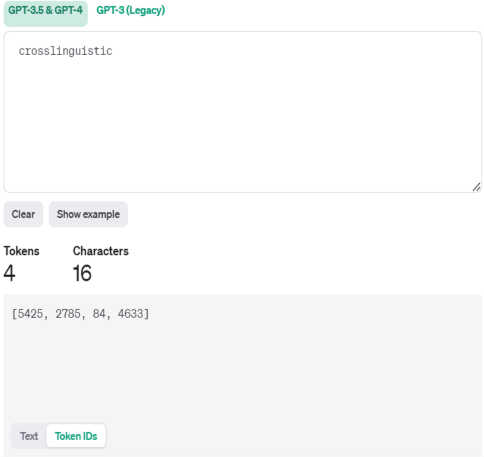
b. Non-initial position: **”ling|u|istic”** [**2785, 84, 4633**] is a derived item, e.g., in
” cross|**ling|u|istic**” [5425, **2785, 84, 4633**]



crosslinguistic

Tokens 4
Characters 16

cross|ling|u|istic



crosslinguistic

Tokens 4
Characters 16

[5425, 2785, 84, 4633]

The PH parses words into morphemes following the same logic, cf. *relative frequency*. Then, (1) and (2) imply that morphemes and words are the same type of unit, which is in accord with DM where the same syntactic rules operate in morphology and syntax. However, DM combines only morphemes, while ChatGPT combines phonemes, morphemes and words simultaneously. Additionally, unlike the meaning-first DM where both morphology and phonology are postsyntactic, ChatGPT is phonology-first: Phonology produces tokens which then form uninterrupted linear sequences. Similar to AM where morphemes are just markings without meaning, ChatGPT tokens are not associated with semantics (Manova et al. 2020). Morphemes without meaning have been evidenced in psycholinguistics (Rastle et al. 2004, Lázaro 2016), e.g., *corn|er* is seen as a derivative, although semantically it is not related to *corn*. Unfortunately, psycholinguistic research on whether the same form in different positions is treated differently by the human brain has been reduced to the trivial distinction between prefixes and suffixes (Crepaldi et al. 2016), and it is still unclear whether the human parser works as in (2). As for why “istic” [4633] is a single token in (2b), the issue is addressed in Manova & Knell (2021) and in Manova (2023). They demonstrate that *-istic/”istic”* is a single unit for the human brain, too.

References

Anderson, Stephen R. (1992), *A-morphous morphology*. Cambridge: Cambridge University Press.

Aronoff, Mark (1976), *Word formation in generative grammar*. Cambridge, MA: MIT Press.

Bobaljik, Jonathan D. 2017. Distributed Morphology. *Oxford Research Encyclopedia in Linguistics*, Oxford: Oxford University Press, https://doi.org/10.1093/acrefore/9780199384655.013.131

Chomsky, Noam, Ian Roberts, and Jeffrey Watumull (2023), Noam Chomsky: The false promise of ChatGPT. *The New York Times*, https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Crepaldi, Davide, Lara Hemsworth, Colin Davis, and Kathleen Rastle (2016), Masked suffix priming and morpheme positional constraints. *The Quarterly Journal of Experimental Psychology,* 69(1), 113–128, https://doi.org/10.1080/17470218.2015.1027713

Halle, Morris and Alec Marantz (1993), Distributed morphology and the pieces of inflection. In K. Hale and S. J. Keyser (eds.), (1993), *The view from building 20*, 111–176. Cambridge, MA: MIT Press.

Hay, Jennifer (2003), *Causes and Consequences of Word Structure*. London: Routledge.

Hay, Jennifer (2001), Lexical Frequency in Morphology: Is Everything Relative? *Linguistics* 39, 1041–1070.

Lázaro, Miguel, Víctor Illera, and Javier Sainz (2016), The suffix priming effect: Further evidence for an early morpho-orthographic segmentation process independent of its semantic content. *Q J Exp Psychol (Hove)* 69(1), 197-208, https://doi.org/10.1080/17470218.2015.1031146

Manova, Stela (2023), ChatGPT, n-grams and the power of subword units: The future of research in morphology, *Proceedings of DeriMo 2023: Resources and Tools for Derivational Morphology*, with slides: lingbuzz/007598

Manova, Stela, Harald Hammarström, Itamar Kastner, and Yining Nie (2020), What is in a morpheme? Theoretical, experimental and computational approaches to the relation of meaning and form in morphology. *Word Structure* 13(1), 1-21.

Manova, Stela and Georgia Knell (2021), Two-suffix combinations in native and non-native English: Novel evidence for morphomic structures. In S. Moradi, M. Haag, J. Rees-Miller, and A. Petrovic (eds.), (2021), *All things morphology*: *Its independence and its interfaces*, 305-323. Current Issues in Linguistic Theory 353. Amsterdam: Benjamins.

Piantadosi, Steven (2023), Modern language models refute Chomsky's approach to language, lingbuzz/007180

Plag, Ingo and Harald Baayen (2009), Suffix ordering and morphological processing. *Language* 85(1), 109–152.

Rastle, Kathleen, Mathew H. Davis, and Boris New (2004), The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review,* 11(6), 1090–1098. https://doi.org/10.3758/BF03196742

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016), Neural Machine Translation of Rare Words with Subword Units, arXiv:1508.07909v5 [cs.CL]

Stump, Gregory T (2001), *Inflectional morphology: A theory of paradigm structure.* Cambridge: Cambridge University Press.