Speaking of Language: Reflections on Metalanguage Research in NLP

Nathan Schneider
Georgetown University
nathan.schneider@georgetown.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Abstract

This work aims to shine a spotlight on the topic of metalanguage. We first define metalanguage, link it to NLP and LLMs, and then discuss our two labs' metalanguage-centered efforts. Finally, we discuss four dimensions of metalanguage and metalinguistic tasks, offering a list of understudied future research directions.

1 Introduction

Language is so powerful that it can be reflected back on itself. All of the following English sentences expressly concern linguistic inventories, structures, and behaviors:

(1) People often confuse the "George" universities.

(2) The expression "kick the bucket" is an idiom meaning "die".

(3) Hebrew has a zero copula in the present tense.

(4) Now read it in an Irish accent.

Sentences such as these may concern a particular instance of language use, or properties of a language or speaker in general; either way, they are **metalinguistic** in making linguistic phenomena (rather than the external world) the subject matter of a linguistic utterance.

Any kind of formal notation that elucidates linguistic properties can also be considered metalanguage, e.g.:

(5) One morning I shot [$_{\text{NP}}$ an elephant in my pajamas].

(6) an$_{\text{DT}}$ elephant$_{\text{NN}}$ in$_{\text{IN}}$ my$_{\text{PRP\$}}$ pajamas$_{\text{NNS}}$

(7) $shot(I, elephant) \wedge in(elephant, pajamas)$

Both kinds of metalanguage enable humans to reflect on linguistic form, meaning, and use, which is why metalanguage is central to fields such as linguistics, language pedagogy, rhetoric, and even law and policy.

The recent advent of fluent multipurpose chatbot tools powered by large language models (LLMs) puts a new focus on metalanguage—or at least we argue that it should. Despite some research on metalanguage and on the metalinguistic abilities of LLMs, the topic remains an understudied one. Yet many real-world tasks in spheres such as language learning and law rely on metalinguistic reasoning rather than simple content understanding. Evaluating whether LLMs can process, generate, and learn from metalanguage therefore provides a crucial test of adaptability and robustness. Zeroing in on the processing of metalanguage, we believe, will ultimately serve applications in important domains where meaning and interpretation are central.

2 Defining "Metalanguage"

The human ability to use language to communicate rests on knowledge that is mostly *implicit*. Linguistics and other fields **communicate explicit conceptualizations of language phenomena** via metalanguage (Berry, 2005). For example, the statement "5-year-old children can productively form regular plurals of nouns" will be incomprehensible to most adults, not to mention the 5-year-olds in question. Not only do the plural-producing 5-year-olds not know the *word* "noun", they presumably have not learned any grammar to the point of comprehending the *concept* of "noun". The metalanguage of a discipline such as linguistics thus reflects many of the concepts at the core of disciplinary expertise.

We provide brief terminological definitions below to setup the stage for the rest of the paper. **Natural metalanguage** is text in a natural language that is interpreted to be about language, grounded in particular utterances or general behavior by a speaker or language community. **Symbolic metalanguage** is formal notation that encodes aspects

Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 107–118

	symbolic	natural
instance-level	tagging, parsing	answering a question about the grammaticality of a sentence
system-level	grammar rule induction	generating a dictionary definition for a word

Table 1: Examples of metalinguistic description tasks (where the metalanguage is language-system–oriented in nature). The metalinguistic inputs and/or outputs in question may be symbolic or natural, and can be formulated at the level of individual instances (tokens) or at the level of an entire system of language (generalizations).

of language (or instances of language use) in a way that explicitly surfaces relationships and patterns within the language system.[1] In general, *a* symbolic metalanguage is a formal system or controlled vocabulary for describing a phenomenon. Finally, we note that quotations (a reference to another speech act or text) are a specific instance of metalanguage that serves (usually) a more narrow/specific communicative purpose.

Broadly, a **metalinguistic task** is one that necessarily processes, leverages, produces, or is defined with metalanguage. Below we discuss kinds of metalinguistic tasks in the context of LLMs (§3.1).

3 Why Study Metalanguage (in NLP)?

Metalinguistic inquiry of one sort or another is common in a wide range of fields and applications. Linguistics, of course, is entirely about the study of language. Consider also: language teaching and learning (e.g., second language learners asking for advice about how to use a word or construction); lexicography; literary studies; and law (the interpretation of legal rules). In these domains, amateur or professional language analysts *consume* instances of language use (in some cases using highly customized corpus search tools), and/or *produce* large quantities of textual metalanguage in textbooks, dictionaries, online discussion forums, legal opinions, and scholarly publications. One impetus, then, for NLP study of metalanguage is to develop tools for metalinguistic inquiry.

Another motivation comes from the inherent goals of modeling language and linguistic meaning. For humans engaged in scientific work, natural as well as formal languages are indispensable when developing theories and making predictions. Within NLP, the tradition of analyzing linguistic grammar and meaning with symbolic structures is one incarnation of metalinguistic NLP (Opitz et al., 2025). (Thus, the study of syntactic parsing, for example, is inherently metalinguistic.)

3.1 Metalanguage and LLMs

In light of the current fascination with large language models (LLMs), it is worth breaking down where metalanguage may come into play in this paradigm, and what studies have or might shed light on its role.

The phenomenon of metalanguage confronts different forms of interaction with an LLM system. We outline these metalinguistic modes below.

Metalinguistic instructions. If a chatbot user explicitly requests a piece of writing—whether it is a summary, translation, homemade pizza recipe, or humorous limerick about cheese—that user is speaking metalinguistically.[2] Thus the practices of instruction tuning and prompting are tied up with metalanguage to some extent. This implies that the presence (or not) and the extent of metalinguistic intent should perhaps inform the evaluation of LLMs in general: do they perform better or worse when the instructions are metalinguistic or not?

Metalinguistic description tasks. By this we mean tasks that center *systematic* aspects of language (or a language). Requesting a definition, grammatical analysis, or explanation of meaning all presuppose a set of conventions constituting a linguistic system, and seek a description that somehow unpacks the conventions or how they apply to a particular instance. Table 1 illustrates tasks that involve *symbolic* or *natural* metalanguage describing linguistic *instances* or *generalizations*. Not all language-manipulation tasks qualify here: machine translation of a sentence, for example, does not in itself reference the organization of either language system (though a translation may be part of a larger

[1] We focus here on metalanguage that is expressed in human-understandable formats, so we will not discuss "style vectors" or "task embeddings" as potential metalanguage.

[2] We consider an instruction metalinguistic if it makes any reference to communication or the linguistic nature of the input or output. Thus "Tell me a joke." and "limerick about cheese" are metalinguistic; "What is the capital of France?" is not.

explanation of linguistic patterns constituting a metalinguistic description).[3]

Metalinguistic interpretation and explanation. To better understand how "black box" models of language operate, one route is to look for correlations between representations or behaviors in the model and their counterparts as described metalinguistically for human language. For example, attempts have been made to localize grammatical knowledge amongst a tangled web of neural network components (e.g., Liu et al., 2019; Tenney et al., 2019; Aoyama and Schneider, 2022; Wang et al., 2022). Other work has investigated model behavior vis-à-vis its generations or probability distributions (e.g., Warstadt et al., 2020; Hu and Levy, 2023). We return to metalinguistic interpretability in §6.

4 A Tale of Two Labs

Research programs within the authors' research groups have prioritized metalinguistic NLP.[4] We give an overview of several such efforts:

- in Anastasopoulos's lab at George Mason University, studies featuring documentary linguistics and low-resource NLP (§4.1 and §4.2);
- in Schneider's lab at Georgetown University, studies motivated by second language learning and legal interpretation (§4.3 and §4.4).

The concluding sections will discuss emerging themes and dichotomies.

4.1 Learning from Reference Grammars

The idea of learning using already-defined grammars is not new; it is in fact one of the first ideas tried out in the early era of symbolic NLP. However, using the text of a reference grammar *as is* to facilitate the creation of language technologies for a given language has only recently come within reach, due to LLMs' capabilities.

Calls to "mobilize the archive", in particular for data-scarce languages, aim to encourage research that leverages linguistic documentation efforts (Bird, 2022). In seminal work, Tanzer et al. (2024) did exactly that, incorporating dictionaries, sentences, and grammar books to perform machine translation using LLMs in a zero-shot setting, i.e., in a language without *any* other data available ("Machine Translation from One Book"). This is perhaps akin to how a documentary linguist or any second-language learner could potentially learn a new language (at least if they did not have access to a teacher or said language's speakers).

In follow-up work, Hus and Anastasopoulos (2024) explored this grammar-based paradigm on 16 languages. Our initial findings were particularly encouraging. For translating extremely low-resource languages like Chuvash, Dogri, and Kalamang into English, providing a combination of dictionary entries and the full grammar book yields almost usable translations (with chrF++ scores between 25–55). Other ongoing work attempts to integrate such approaches into a documentary linguist's workflow—in this case, working on Nepal's Kulung languages (Taguchi et al., 2025).

However, concurrent and followup work has called into question whether the current generation of LLMs can trully understand and leverage metalinguistic content in the form of a reference grammar (Aycock et al., 2025; Marmonier et al., 2025). Regardless, we believe that the potential for reducing data requirements for under-resourced languages of already extremely under-served communities makes this a worthy research direction.

4.2 Inducing and Describing Patterns in Data

Another line of work aims at *generating* metalanguage (symbolic or natural). In particular, we aim at simulating the work of a linguist or a language teacher, producing output that describes a language system, based on raw text samples.

The notion of describing a language "in its own terms" based solely on raw data has an established tradition in descriptive linguistics (Harris, 1951). Early work included discovering morphosyntactic agreement (Chaudhary et al., 2020) or lexical selection preferences (Chaudhary et al., 2021), tying it also to educational applications, by presenting these rules along with selected examples to be used by L2 teachers (Chaudhary et al., 2023) – see an illustration in Figure 1. Earlier work by Howell et al. (2017) aimed to predict the case systems of endangered languages and Zamaraeva (2016) inferred morphotactics from IGT using k-means clustering.

[3] There is probably no bright line that demarcates this category. Between "Proofread this paragraph" and "Indicate the grammatical errors in this paragraph", the latter more overtly invokes a goal of linguistic system–based description, but in practice these serve very similar user needs. An alternative definition going beyond our focus could take into account user intent so that e.g., producing a free translation for an interlinear gloss of an example in a reference grammar could qualify as such a metalinguistic description task.

[4] Supported in part by NSF awards "CAREER: Metalinguistic Natural Language Understanding" (Schneider) and "CAREER: Leveraging Grammar Books to Develop Language Technologies for Data-Scarce Languages" (Anastasopoulos).

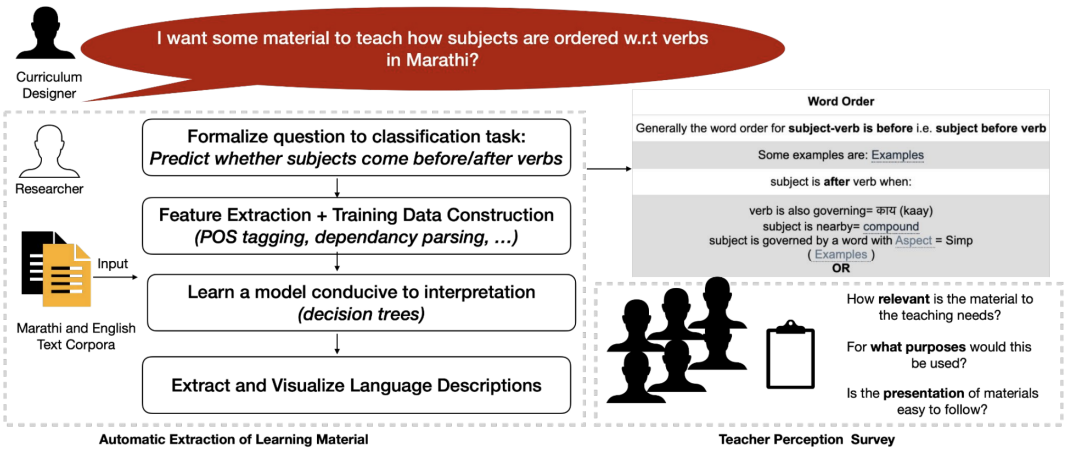


Figure 1: Workflow for the collaboration of NLP researchers and language-learning curriculum designers, to create pedagogical materials (Chaudhary et al., 2023). The input and intermediate and final outputs include metalanguage.

The above line of work is somewhat orthogonal to the works that have typologists as their target audience. Linguists and researchers have long undertaken initiatives to collect linguistic properties in machine-readable formats. *WALS* (Dryer and Haspelmath, 2013) is one such example which can tell us, for instance, that English objects occur after verbs, or that Turkish pronouns have symmetrical case. *Grambank* (Skirgård et al., 2023) is the latest typological database: it covers 2,467 language varieties, capturing a wide range of grammatical phenomena in 195 features, from word order to verbal tense and many other well-studied comparative linguistic variables. Most works on the NLP side aim to fill in the missing entries in such databases, producing a structured typological description of a language based on raw text samples (Daumé and Campbell, 2007; Bjerva et al., 2020, *inter alia*). See Baylor et al. (2023) for additional discussion on the usefulness of such work for NLP in general. More recently, Arčon et al. (2026) converted each entry of the WALS database into a question-answer pair, in order to test LLMs' typological knowledge. Hus and Anastasopoulos (2026) pose similar typological questions but they also assume access to reference grammars for the languages in question.

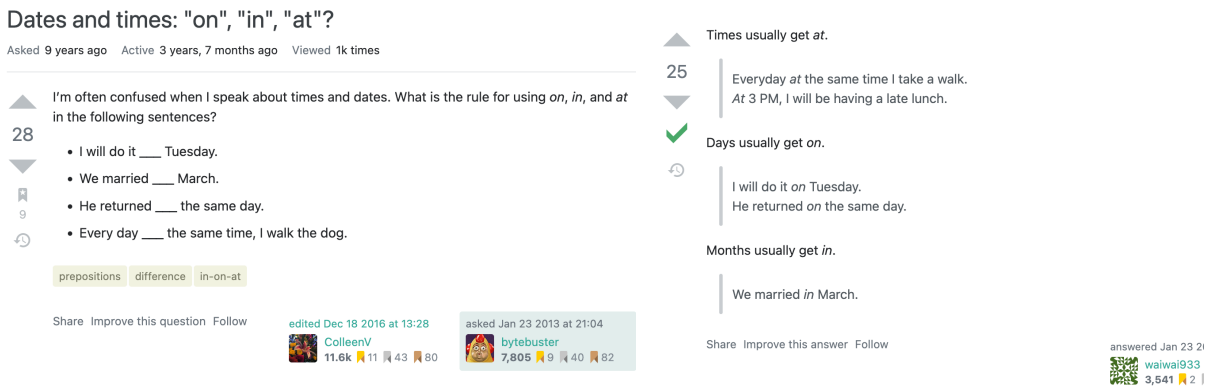
Notably, to our knowledge, there is no substantial progress in integrating LLMs with field linguists' or typologists' workflow. Most works listed above are from the pre-LLM era. One exception is the rather exciting exploration of the metalinguistic reasoning capabilities of LLMs focused on small, artificial problems inspired (or directly taken) from linguistic olympiads, such as the PuzzLing Machines (Şahin et al., 2020) and LingOly (Bean et al., 2024) benchmarks (see also §6). But all such problems with their guarantees of a single correct solution and carefully curated data to reach it barely mimic real-world settings, where incomplete and ambiguous data render the task significantly harder. The very recent work of Yang et al. (2025) that tests whether LLMs can be used to gloss unknown lexical items is one step in the above direction.[5]

4.3 Language Learning Domain

An important domain for metalanguage is the realm of language teaching and learning, whether in a classroom or in a less formal context such as an online forum or chatbot interaction. Complementing the extraction of symbolic rules for educational purposes described above (§4.2), it is valuable to investigate **natural** metalanguage scenarios in the language learning domain.

Here we highlight two studies of LLM processing of natural metalanguage (NML). Behzad et al. (2023) collected data from two online English discussion forums (one designed specifically for L2 English learners) in order to construct a benchmark for English Language Question Answering (ELQA). The metalinguistic questions in this dataset span a range of topics including vocabulary, grammar, and meaning; some, for example, inquire about sentence grammaticality, while others request help with expression or passage interpretation, or asked general questions about linguistic conventions. An example of a question and corresponding answer, both user-generated, appear in

[5] We note, although, that it still operates in a less noisy environment and with more available information as input than one might expect in real-world settings.



(a) Question **(b)** Answer

Figure 2: Screenshots of a page on the English Language Learner Stack Exchange site, which is included in the ELQA dataset (from Behzad et al., 2023). The source page is `https://ell.stackexchange.com/questions/12/dates-and-times-on-in-at`.

Figure 2: the question/answer pair mix a system-level overarching issue (about the rules for prepositions in times and dates) with specific exemplar instances.

Sampling questions and answers from ELQA, Behzad et al. (2023) conducted a human evaluation pitting user responses against responses from LLMs (including GPT-3 with few-shot learning or finetuning). Note that this kind of question answering is an NML-to-NML task. Overall, the best GPT-3 setup was highly fluent across the board, and gave accurate answers for many of the questions, but in some cases underperformed the highest-rated user answer for accuracy. This suggests that LLMs may be helpful interlocutors in answering learners' metalinguistic questions, but at times will make mistakes (with the caveat that today's state-of-the-art models have not yet been evaluated).

In a followup study, Behzad et al. (2024) addressed the *crosslinguistic* dimension of learner QA, speculating that L2 learners using an LLM may frame questions in their native language. This raises the question of whether the pairing of prompt-language and target language matters. This study used a controlled paradigm of minimal pair grammaticality judgment[6] (given an original and a corrected sentence from a grammatical error correction dataset) so that it would be possible to independently manipulate the language of the prompt template and the language of the target sentence. The languages tested were English, German, Russian, Ukrainian, and Korean. The tested models displayed a great deal of sensitivity to the choice of prompt-language, suggesting that the apparent multilinguality of an LLM does not guarantee stable metalinguistic behavior across languages.

[6]Other work has studied metalinguistic prompting for grammaticality judgments (e.g., Hu and Levy, 2023).

4.4 Legal Interpretation

First, a bit of background. **Legal interpretation** is the enterprise of determining the meaning of a rule expressed in natural language (Brannon, 2023). This frequently arises in judicial cases, where a judge must determine the extent of a category in order to decide whether it applies to the facts of the case. (If the rule is "no vehicles are allowed in the park", and "vehicle" is not specifically defined, should it be read to encompass skateboards? Wheelchairs? Ambulances? Are there contextual clues that shed light on the scope of the category?) The vehicle rule is a hypothetical example, but real cases similarly discuss the semantic interpretation of a term, such as the "landscaping" controversy summarized in Figure 3. Other cases implicate grammatical ambiguities in the wording of a statute or contract. Many U.S. judges subscribe to the philosophy that the starting point for interpreting legal language is the general language—they seek to determine the "ordinary meaning" of the text per contemporaneous usage (as members of the public might interpret it today, or when the law was enacted). This "textualist" perspective has engendered scholarly examinations of the nature of meaning in language, and the principles judges have articulated in an attempt to make language analysis rigorous and objective. But the practice of textualism has come under fire from scholars who contest the supposed neutrality of these principles—suggesting that they are poorly formulated, rooted

John is a contractor with insurance that covers property loss, damage, or personal injury claims that arise due to his 'landscaping' work.

John is employed by a family, the Smiths, to install an in-ground trampoline in the family's backyard. A few years after John completes the project, the Smiths successfully sue John for injuries that their daughter sustained while playing on the trampoline. John files a claim with his insurance company to recover losses incurred from the lawsuit.

Considering just how "landscaping" would be understood by ordinary speakers of English, is John covered by the insurance—yes or no?

Figure 3: A legal interpretation scenario represented as a QA task with binary questions. The example is based on the case *Snell v. United Specialty Insurance Co.* and constructed in the style of one of the prompting formats studied by Purushothama et al. (2026a).

in misconceptions about linguistic meaning, or so malleable that they can be manipulated to support any outcome (Eskridge et al., 2023).

Through collaborations with law professor Dr. Kevin Tobia—who has advocated for empirical approaches like survey research to ascertain ordinary meaning (e.g., Tobia, 2020, 2022; Waldon et al., 2025a)—Dr. Schneider and his lab are investigating computational linguistic and NLP tools for textualist inquiry. There are several threads of investigation.

Are LLMs trustworthy as tools for answering interpretive questions? Already judges have begun to entertain the possibility that difficult interpretive questions might be outsourced to LLM chatbots, on the rationale that huge amounts of ordinary usage in training data would entail accurate (perhaps superhuman) metalinguistic conclusions about meaning.[7] Waldon et al. (2025b) push back against this assumption, attributing it to myths about how LLMs work. Further experimentation by Purushothama et al. (2026a) and Petersen et al. (2026) examines prompt sensitivity and alignment with human consensus. (One of the tested prompt formats appears in Figure 3.) The results so far indicate that state-of-the-art platform models are less sensitive to prompt framing than smaller-scale models, and achieve some level of correlation with human judgments, but are not immune to giving an implausible answer when asked for a binary judgment. They are therefore *not* a silver bullet for resolving difficult questions. If they have any utility for interpretive reasoning, it is probably as a brainstorming aid: the system can be asked to generate arguments for and against a position, provided the human judge critically evaluates all claims (Waldon et al., 2025b).

How prevalent are different facets of metalanguage in judicial opinions? Kranzlein et al. (2024); Kranzlein (2024, ch. 5) examined this as a computational social science question by taxonomizing and tagging different facets of metalanguage in a corpus of Supreme Court opinions. An example of a metalinguistic sentence from one of these opinions (Breyer, 2023):

(8) First, the Act defines "pollutant" broadly, including in its definition, for example, any solid waste, incinerator residue, "heat," "discarded equipment,"' or sand (among many other things). §502(6), 86 Stat. 886.

Notably, sentence (8) includes several parts: **metalinguistic cue** words that denote linguistic units or processes ("defines", "definition"); a **focal term** being defined ("pollutant"); portions of a **definition** of the focal term; and a citation to a **legal source**. Kranzlein et al. (2024) envision metalanguage category tagging as an information extraction task, and annotate these categories in the CuRIAM corpus. (The full list of categories appears in Table 2.)

Kranzlein (2024, ch. 5) then trains a tagger for these categories: a **metalanguage identification** task. With automatic tagging, he conducts a content analysis of three decades of Supreme Court opinions. His analysis of metalanguage use over time points to an increase of some of the categories from 1986 to 2018, consistent with the growing popularity of textualism as an interpretive philosophy.

Do judicial canons of construction reflect accurate generalizations about linguistic usage? Over time, textualist judges have developed a suite of heuristics known as **canons of construction**. These assert preferences for resolving certain ambi-

[7] One judge writes: "models train on a mind-bogglingly enormous amount of raw data [across many genres]. Because they cast their nets so widely, LLMs can provide useful statistical predictions about how, in the main, ordinary people ordinarily use words and phrases in ordinary life" (Newsom, 2024). This fails to appreciate the distinction between learning implicit usage patterns, and being able to articulate those patterns metalinguistically in response to a prompt.

Category	Definition
Focal Term (FT)	Word or phrase used metalinguistically and/or whose meaning is under discussion.
Definition (D)	Succinct, reasonably self-contained description of what a word or phrase means. Need not be exhaustive. May also be negative—defining a word by what it's not.
Metalinguistic Cue (MC)	Word or short phrase cueing nearby metalanguage.
Direct Quote (DQ)	Span of text inside quotation marks.
Legal Source (LeS)	Citation or mention appealing to a legal document or authority.
Language Source (LaS)	Citation or mention appealing to an authority on language.
Named Interpretive Rule (NIR)	Mention of a well-established interpretive rule or test used to support an argument about the meaning of a word or phrase.
Example Use (ES)	Intuitive, quoted, or hypothetical examples that demonstrate a word/term can or cannot be used in a certain way.
Appeal to Meaning (ATM)	An explicit argument, implicit value judgment, or other statement indicating how one should go about interpreting meaning (e.g., by appealing to common sense, ordinary meaning, or the language of another statute).

Table 2: Categories of metalanguage annotated in the CuRIAM corpus (from Kranzlein et al., 2024).

guities in the text (Scalia and Garner, 2012; Brannon, 2025).[8] In response to calls for basing canons on stronger empirical foundations (Tobia et al., 2022), we have sought to critically examine the canons via computational linguistic techniques: namely corpus analysis (e.g., compiling judicial opinions referencing a particular canon in order to establish how it tends to be applied), treebanking (e.g., to establish the most frequent resolution of syntactic ambiguity in statutory text; Waldon et al., 2025c), and semantic annotation (Wells et al., 2025). These investigations are ongoing. Our hope is that they will lead to a clearer articulation of the canons (with precise terminology from linguistics), as well as empirical data about the reliability of each canon in practice.

5 Dichotomies in Metalanguage Research

Organizing metalanguage research, even when not taking a very broad view of metalanguage, requires considering multiple axes of analysis. We observe the following notable dichotomies (two of which are highlighted in Table 1):

- *system-level vs. instance-level metalanguage:* system-level metalanguage targets general properties of a linguistic system (e.g., grammatical rules, constructions, or typological facts), at the level of the entire language or subpopulation of the language community; instance-level metalanguage concerns specific linguistic tokens or contexts (e.g., explaining why a given sentence is ambiguous).
- *monolingual vs. multilingual:* the necessary LLM capabilities as well as system requirements and design would likely need to differ for metalinguistic inquiries targeting a single language, contrasting a pair of languages, or if the focus is specifically on second-language settings.
- *symbolic vs. natural metalanguage:* employing formalized notations such as parse trees, symbolic metalanguage enables precision but is less accessible to lay users than using ordinary (natural) language to describe linguistic phenomena.
- *processing vs. generation of metalanguage:* while processing metalanguage as input might largely evaluate models' comprehension, generation reveals whether models externalize linguistic reasoning in useful ways. Of course, many applications such as legal analysis and educational assistance require both capabilities.

By necessity, any metalanguage-related work will occupy a position along many of these axes, as do many of our works discussed in §4.

6 Research Directions

Research connecting metalanguage and LLMs asks how well LLMs fare on different kinds of metalinguistic tasks; why; and to what end. We list a few specific directions below. Some of these have already been subjects of inquiry, for which we give illustrative citations from our labs and others. Many

[8] The Nearest-Reasonable-Referent Canon, for example, makes recommendations about how to disambiguate the syntactic attachment of a modifier (Scalia and Garner, 2012).

directions though are, to the best of our knowledge, heretofore unexplored:[9]

Intrinsic evaluation questions Here we focus on research directions that aim to evaluate the metalinguistic capabilities of LLMs:

1. Can an LLM solve linguistic structure NLP tasks (e.g., Ettinger et al., 2023; Tian et al., 2024), analysis problems in theoretical linguistics (Beguš et al., 2025), or language puzzles (Rozner et al., 2021; Şahin et al., 2020; Bean et al., 2024; Chi et al., 2024; Sánchez et al., 2025; Choudhary et al., 2025)?
2. How well can models understand self-referential language ("This sentence is short.")? (Thrush et al., 2024)
3. How well can models distinguish mentions vs. uses?[10] (Kranzlein, 2024) (discussed in §4.4)
4. How does the choice of the (natural or formal) language in which metalanguage is formulated affect model behavior in relation to the described language? (Behzad et al., 2024) (discussed in §4.3)
5. Are LLMs sensitive to pragmatic phenomena like so-called *metalinguistic negation* (Horn, 1985),[11] where the speaker uses negation to signal disagreement with a choice of words, and quotation, where the speaker is not necessarily committing to the same perspective as the source they are quoting? (Gligorić et al., 2024, focusing on detecting whether hate speech and misinformation reflects the speaker's perspective)

Interpretability questions Next we outline inquiries that are paramount in order to understand *why and how* metalinguistic abilities arise in LLMs:

6. How well-calibrated is *explicit* metalinguistic output with respect to the system's *implicit* linguistic generalizations? (Hu and Levy, 2023; Song et al., 2025)
7. Can metalinguistic distinctions such as use vs. mention be traced to internal model representations?
8. How much of model behavior on metalinguistic tasks can be attributed to metalinguistic text in pretraining data or data provided at inference time (or not, as discussed, e.g., in Aycock et al. (2025) and Marmonier et al. (2025))?
9. To the extent that metalinguistic meaning is grounded in linguistic usage, is distributional learning from form alone (discussed, e.g., in Bender and Koller, 2020; Pavlick, 2023) fundamentally different from learning of non-metalinguistic meaning?

Extrinsic uses Finally, we delineate some underexplored uses of metalanguage for further downstream applications:

10. Can metalinguistic data such as syntax trees or grammar rules/descriptions be leveraged for inductive biases in pipelined systems (Wein and Schneider, 2024), integrated within LM architectures (Prange et al., 2022; Gessler and Schneider, 2023), or via in-context learning (Court and Elsner, 2024; Ginn and Palmer, 2025; Pei et al., 2025; Nakashole, 2026; Purushothama et al., 2026b)?
11. How well can a system perform metalinguistic question answering? (Behzad et al., 2023) (discussed in §4.3)
12. How can NLP shed light on how people use metalanguage? (Kranzlein et al., 2024) (discussed in §4.4)
13. Can we build LLM-powered assistants that deploy metalanguage effectively for language documentation, education, and scholarship? (also discussed in §4.1 and §4.2)

7 Conclusion

Metalanguage is inherently multifaceted. As we outlined in the possible dimensions in §5 above, it spans multiple levels of abstraction, heterogeneous representational formats, and diverse forms of linguistic reasoning. This diversity should be taken into consideration as we devote greater attention to metalinguistic tasks and applications.

We are particularly excited about the interpretability questions around metalanguage. Metalinguistic tasks may be particularly challenging where they require a model to *both* articulate and apply linguistic reasoning.

Metalanguage research also offers a promising pathway to studying learning and generalization. Humans frequently learn from explicit metalinguistic instructions and explanations and are able to

[9] Here we include work with transformer models like BERT and GPT-2, though our main focus in this paper is on contemporary prompt-based LLMs.

[10] *Mentioned language* when text refers explicitly to a linguistic entity like a word or sentence. A thorough definition is given by Wilson (2011, ch. 2), and a study of statistical classifiers is presented by Wilson (2013).

[11] Also known as *frame-rejecting negation*; an example is "John isn't being thrifty, he's just downright stingy" (Fillmore, 1985, p. 243).

apply this new knowledge to new examples. One should expect models to be able to do the same. Advancing research on how models learn from and operationalize metalanguage should dramatically improve the frontier of LLM abilities in general.

Limitations

We do not aim for this work to be a complete survey of metalanguage research. We by design draw heavily from our own work and our own perspective on the field.

Acknowledgments

We thank our collaborators on our metalinguistic journeys, including members of the NERT lab and the George Mason NLP lab, Dr. Amir Zeldes, and Dr. Kevin Tobia. We benefited from the Metalinguistic NLP Bibliography (https://github.com/nert-nlp/metalinguistic-nlp-bib) spearheaded by Abhishek Purushothama. This work was supported in part by NSF awards IIS-2144881 (Schneider) and IIS-2439202 (Anastasopoulos).

References

Tatsuya Aoyama and Nathan Schneider. 2022. Probeless probing of BERT's layer-wise linguistic knowledge with masked word prediction. In *Proc. of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 195–201, Hybrid: Seattle, Washington + Online.

Tjaša Arčon, Matej Klemen, Marko Robnik-Šikonja, and Kaja Dobrovoljc. 2026. Evaluating metalinguistic knowledge in large language models across the world's languages. arXiv:2602.02182 [cs].

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Simaan. 2025. Can LLMs really learn to translate a low-resource language from one grammar book? In *International Conference on Learning Representations*, volume 2025, pages 12334–12357.

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2023. The past, present, and future of typological databases in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1163–1169, Singapore. Association for Computational Linguistics.

Andrew Bean, Simi Hellsten, Harry Mayne, Jabez Magomere, Ethan A., Ryan Chi, Scott A. Hale, and Hannah R. Kirk. 2024. LINGOLY: A benchmark of olympiad-level linguistic reasoning puzzles in low-resource and extinct languages.

Gašper Beguš, Maksymilian Dąbkowski, and Ryan Rhodes. 2025. Large linguistic models: investigating LLMs' metalinguistic abilities. *IEEE Transactions on Artificial Intelligence*, 6(12):3453–3467.

Shabnam Behzad, Keisuke Sakaguchi, Nathan Schneider, and Amir Zeldes. 2023. ELQA: A corpus of metalinguistic questions and answers about English. In *Proc. of ACL*.

Shabnam Behzad, Amir Zeldes, and Nathan Schneider. 2024. To ask LLMs about English grammaticality, prompt them in a different language. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15622–15634, Miami, Florida, USA. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: on meaning, form, and understanding in the age of data. In *Proc. of ACL*, pages 5185–5198, Online.

Roger Berry. 2005. Making the most of metalanguage. *Language Awareness*, 14(1):3–20.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Johannes Bjerva, Elizabeth Salesky, Sabrina J. Mielke, Aditi Chaudhary, Giuseppe G. A. Celano, Edoardo Maria Ponti, Ekaterina Vylomova, Ryan Cotterell, and Isabelle Augenstein. 2020. SIGTYP 2020 shared task: Prediction of typological features. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*, pages 1–11, Online. Association for Computational Linguistics.

Valerie C. Brannon. 2023. Statutory interpretation: theories, tools, and trends. Report R45153, Congressional Research Service.

Valerie C. Brannon. 2025. Canons of construction: a brief overview. In Focus IF12992, Congressional Research Service.

Stephen Breyer. 2023. *County of Maui v. Hawaii Wildlife Fund*. 140 S. Ct. 1462.

Aditi Chaudhary, Antonios Anastasopoulos, Adithya Pratapa, David R. Mortensen, Zaid Sheikh, Yulia Tsvetkov, and Graham Neubig. 2020. Automatic extraction of rules governing morphological agreement. In *Proc. of EMNLP*.

Aditi Chaudhary, Arun Sampath, Ashwin Sheshadri, Antonios Anastasopoulos, and Graham Neubig. 2023. Teacher perception of automatically extracted grammar concepts for L2 language learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3776–3793, Singapore. Association for Computational Linguistics.

Aditi Chaudhary, Kayo Yin, Antonios Anastasopoulos, and Graham Neubig. 2021. When is *wall* a *pared* and when a *muro*?: Extracting rules governing lexical selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6911–6929, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nathan Chi, Teodor Malchev, Riley Kong, Ryan Chi, Lucas Huang, Ethan Chi, R. McCoy, and Dragomir Radev. 2024. ModeLing: a novel dataset for testing linguistic reasoning in language models. In *Proc. of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*.

Mukund Choudhary, KV Aditya Srivatsa, Gaurja Aeron, Antara Raaghavi Bhattacharya, Dang Khoa Dang Dinh, Ikhlasul Akmal Hanif, Daria Kotova, Ekaterina Kochmar, and Monojit Choudhury. 2025. UNVEILING: What makes linguistics olympiad puzzles tricky for LLMs? In *Proc. of COLM*.

Sara Court and Micha Elsner. 2024. Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354, Miami, Florida, USA. Association for Computational Linguistics.

Hal Daumé, III and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *Proc. of ACL*, pages 65–72, Prague, Czech Republic.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

William N. Eskridge, Brian G. Slocum, and Kevin Tobia. 2023. Textualism's defining moment. *Columbia Law Review*, 123(6):1611–1698.

Allyson Ettinger, Jena Hwang, Valentina Pyatkin, Chandra Bhagavatula, and Yejin Choi. 2023. "You are an expert linguistic annotator": Limits of LLMs as analyzers of Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8250–8263, Singapore.

Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2):222–254.

Luke Gessler and Nathan Schneider. 2023. Syntactic inductive bias in transformer language models: especially helpful for low-resource languages? In *Proc. of CoNLL*, pages 238–253, Singapore.

Michael Ginn and Alexis Palmer. 2025. LLM dependency parsing with in-context rules. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 186–196, Vienna, Austria. Association for Computational Linguistics.

Kristina Gligorić, Myra Cheng, Lucia Zheng, Esin Durmus, and Dan Jurafsky. 2024. NLP systems that can't tell use from mention censor counterspeech, but teaching the distinction helps. In *Proc. of NAACL-HLT*, pages 5942–5959, Mexico City, Mexico.

Zellig S. Harris. 1951. *Methods in Structural Linguistics*. University of Chicago Press.

Laurence R. Horn. 1985. Metalinguistic negation and pragmatic ambiguity. *Language*, 61(1):121–174.

Kristen Howell, Emily M Bender, Michel Lockwood, Fei Xia, and Olga Zamaraeva. 2017. Inferring case systems from igt: Enriching the enrichment. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 67–75.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proc. of EMNLP*.

Jonathan Hus and Antonios Anastasopoulos. 2024. Back to school: Translation using grammar books. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20207–20219, Miami, Florida, USA. Association for Computational Linguistics.

Jonathan Hus and Antonios Anastasopoulos. 2026. A rag approach for typological database completion. In *Proceedings of the Eighth Workshop on Computational Research in Linguistic Typology*, Rabbat, Morocco. Association for Computational Linguistics.

Michael Kranzlein. 2024. *Unpacking Meaning with Natural Language Processing: Legal Metalanguage Analysis and Long-Tail Calibration*. Ph.D. dissertation, Georgetown University.

Michael Kranzlein, Nathan Schneider, and Kevin Tobia. 2024. CuRIAM: Corpus Re Interpretation and Metalanguage in U.S. Supreme Court Opinions. In *Proc. of LREC-COLING*.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proc. of NAACL-HLT*, pages 1073–1094, Minneapolis, Minnesota.

Malik Marmonier, Rachel Bawden, and Benoît Sagot. 2025. Explicit learning and the LLM in machine translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31372–31422, Suzhou, China. Association for Computational Linguistics.

Ndapa Nakashole. 2026. Grammar as control: Modular language generation for the long tail. In *Proc. of ACL*, San Diego, California.

Kevin Newsom. 2024. Concurring opinion in *Snell v. United Specialty Insurance Co.* United States Court of Appeals For the Eleventh Circuit, 22-12581.

Juri Opitz, Shira Wein, and Nathan Schneider. 2025. Natural language processing RELIES on linguistics. *Computational Linguistics*, 51(3):1009–1032.

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251):20220041.

Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schuetze. 2025. Understanding in-context machine translation for low-resource languages: A case study on Manchu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8767–8788, Vienna, Austria. Association for Computational Linguistics.

Dawson Petersen, Abhishek Purushothama, and Nathan Schneider. 2026. Sense and sensitivity: "reasoning" models are more robust, but can diverge from human consensus in a legal interpretation task. In *Proc. of CoNLL*, San Diego, California.

Jakob Prange, Nathan Schneider, and Lingpeng Kong. 2022. Linguistic frameworks go toe-to-toe at neuro-symbolic language modeling. In *Proc. of NAACL-HLT*, pages 4375–4391, Seattle, United States.

Abhishek Purushothama, Junghyun Min, Brandon Waldon, and Nathan Schneider. 2026a. Prompting from the bench: Large-scale pretraining is not sufficient to prepare LLMs for ordinary meaning analysis. In *Proc. of the Ninth Annual ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, Montréal, Canada. arXiv preprint: 2510.25356 [cs].

Abhishek Purushothama, Emma Thronson, Alexia Guo, and Amir Zeldes. 2026b. Syntax as a Rosetta Stone: Universal Dependencies for in-context Coptic translation. In *Findings of ACL*. arXiv preprint: 2604.18758 [cs].

Josh Rozner, Christopher Potts, and Kyle Mahowald. 2021. Decrypting cryptic crosswords: semantically complex wordplay puzzles as a target for NLP. In *Advances in Neural Information Processing Systems*, volume 34, pages 11409–11421.

Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020. PuzzLing Machines: a challenge on learning from small data. In *Proc. of ACL*.

Eduardo Sánchez, Belen Alastruey, Christophe Ropers, Arina Turkatenko, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2025. Linguini: A benchmark for language-agnostic linguistic reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Antonin Scalia and Bryan A. Garner. 2012. *Reading law: the interpretation of legal texts*. Thomson/West, St. Paul, MN.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, and 86 others. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

Siyuan Song, Jennifer Hu, and Kyle Mahowald. 2025. Language models fail to introspect about their knowledge of language. In *Proc. of COLM*.

Chihiro Taguchi, J Elizabeth Liebl, Antonios Anastasopoulos, David Chiang, and Géraldine Walther. 2025. Digital documentation for diasporic data: challenges, opportunities, and solutions for working with diaspora communities. In *9th International Conference on Language Documentation & Conservation (ICLDC)*.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. In *International Conference on Learning Representations*, volume 2024, pages 18955–18985.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proc. of ACL*, pages 4593–4601, Florence, Italy.

Tristan Thrush, Jared Moore, Miguel Monares, Christopher Potts, and Douwe Kiela. 2024. I am a strange dataset: metalinguistic tests for language models. In *Proc. of ACL*.

Yuanhe Tian, Fei Xia, and Yan Song. 2024. Large language models are no longer shallow parsers. In *Proc. of ACL*, pages 7131–7142, Bangkok, Thailand.

Kevin Tobia. 2022. Experimental jurisprudence. *The University of Chicago Law Review*, 89(3):735–802.

Kevin Tobia, Brian G. Slocum, and Victoria Nourse. 2022. Statutory Interpretation from the outside. *Columbia Law Review*, 122(1):213–330.

Kevin P. Tobia. 2020. Testing ordinary meaning. *Harvard Law Review*, 134(2):726–806.

Brandon Waldon, Cleo Condoravdi, James Pustejovsky, Nathan Schneider, and Kevin Tobia. 2025a. Reading law with linguistics: the statutory interpretation of artifact nouns. *Harvard Journal on Legislation*, 62(2):415–467.

Brandon Waldon, Nathan Schneider, Ethan Wilcox, Amir Zeldes, and Kevin Tobia. 2025b. Large language models for legal interpretation? Don't take their word for it. *Georgetown Law Journal*, 114(1):115–183.

Brandon Waldon, Micaela Wells, Devika Tiwari, Meru Gopalan, and Nathan Schneider. 2025c. Legal-CGEL: Analyzing legal text in the CGELBank framework. In *Proc. of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025)*, pages 148–153, Ljubljana, Slovenia.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. arXiv:2211.00593 [cs].

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Shira Wein and Nathan Schneider. 2024. Lost in translationese? Reducing translation effect using Abstract Meaning Representation. In *Proc. of EACL*, pages 753–765, St. Julian's, Malta.

Micaela Wells, Brandon Waldon, and Nathan Schneider. 2025. Scope ambiguity resolution of negated connectives in English corpora. In *Proc. of the Annual Meeting of the Cognitive Science Society*, volume 47, page 6608.

Shomir Wilson. 2011. *A computational theory of the use-mention distinction in natural language*. Ph.D. dissertation, University of Maryland, College Park, Maryland.

Shomir Wilson. 2013. Toward automatic processing of English metalanguage. In *Proc. of IJCNLP*.

Changbing Yang, Franklin Ma, Freda Shi, and Jian Zhu. 2025. LingGym: How far are LLMs from thinking like field linguists? In *Proc. of EMNLP*, pages 1314–1340, Suzhou, China.

Olga Zamaraeva. 2016. Inferring morphotactics from interlinear glossed text: Combining clustering and precision grammars. In *Proc. of the SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 141–150, Berlin, Germany. Association for Computational Linguistics.