Last Words

Natural Language Processing RELIES on Linguistics

Juri Opitz[1*,**], Shira Wein[2*], and Nathan Schneider[3*]

[1] University of Zurich, Department of Computational Linguistics
   opitz.sci@gmail.com
[2] Amherst College, Department of Computer Science
   swein@amherst.edu
[3] Georgetown University, Department of Computer Science
   Nathan.Schneider@georgetown.edu

*Large Language Models have become capable of generating highly fluent text in certain languages, without modules specially designed to capture grammar or semantic coherence. What does this mean for the future of linguistic expertise in NLP? We highlight several aspects in which NLP (still) relies on linguistics, or where linguistic thinking can illuminate new directions. We argue our case around the acronym **RELIES**, which encapsulates six major facets where linguistics contributes to NLP: **R**esources, **E**valuation, **L**ow-resource settings, **I**nterpretability, **E**xplanation, and the **S**tudy of language. This list is not exhaustive, nor is linguistics the main point of reference for every effort under these themes; but at a macro level, these facets highlight the enduring importance of studying machine systems vis-à-vis systems of human language.*

1. Introduction

It is 2025. ChatGPT has been an international sensation for over two years, a focal point in the ongoing public hype, enthusiasm, and concern about "generative AI." NLP conferences are awash in papers about prompting Large Language Models that have gobbled zillions of words of text and can produce eerily fluent language (in English, anyway). These models have been refined with corrective feedback on conversational outputs, but (so far as we know) no direct inductive biases about grammaticality or compositional meaning.[1] Like BERT and similar models, an NLP research epoch before—or

---

word embeddings an epoch before that—Large Language Models have elicited both fascination and hand-wringing in scientific circles.

Do these developments spell the end of the relevance of linguistics to NLP?[2] After all, linguistic theories positing elaborate models of syntax are presented as necessary for explaining grammaticality, but are not incorporated into NLP models which nevertheless generate fluent text.

In this piece, we will argue that ideas from linguistics, even if not overtly framed as such, actually continue to underlie much of what we do in NLP. We highlight six facets where linguistics continues to play a major role, and for each facet, point to NLP work which has benefited from (or has been founded on) linguistic knowledge. We hope our survey will resonate with audiences representing a range of expertise—those with technology-oriented perspectives as well as human language perspectives, and their meeting point within the field of NLP.

What do we mean by "linguistics," "NLP," and "CL"? Before developing our argument, we need to define our terms (while acknowledging that our definitions will not be perfectly crisp, as disciplinary boundaries are inherently fuzzy).

Linguistics is the study of systematicity and variation in human communication, as transmitted via speech, sign, and writing. Linguists may specialize in how knowledge of forms is organized and deployed to symbolize meanings; how linguistic behaviors emerge, drawing on human cognitive and social capacities; and how language varies over time (ontogeny, phylogeny), geography, identity group, etc. Linguistics-adjacent fields of language study and practice include language education, translation, rhetoric, communications, and philosophy of language. Below, the term "linguistics" covers these (non-AI) "language fields," broadly defined.

Natural Language Processing (NLP) is the field concerned with developing technology for sophisticated computational processing of text, and especially, computational understanding or generation of individual sentences, documents, or conversations (as opposed to drawing inferences about entire collections).[3] Contemporary NLP research prioritizes empirical evaluation of systems on *tasks*, whether connected with general user applications (like QA, translation, summarization) or more granular and focused on aspects of the language system (like parsing and coreference resolution). In recent years, the dominant concern of NLP has been the design, interpretation, and application of pretrained transformer language models such as BERT and GPT. Pretrained models of considerable scale are often called Large Language Models (LLMs).

Today, **computational linguistics** (CL) has two definitions: a broad one (roughly, "computation and natural language") that includes technologically oriented NLP, and a narrower one that focuses on computational formalization and processing for the end goal of studying how language works. We use "**cL**" for this narrower goal whose chief motivation is to answer questions about language, rather than questions about technology, though methods and sub-questions often overlap between cL and NLP, and consequently both can be found at CL conferences.

Where would NLP be without linguistics? Linguistics is no longer front and center in the way we build NLP systems for practical tasks. Gone are the days when machine translation engines consisted of rules painstakingly crafted by linguists. But there are

---

2 We are hardly the first to ask some version of this question; see §2.

3 Text mining, information retrieval, and speech processing are studied in separate research communities, with some overlap. These areas, in combination with prototypical NLP applications such as machine translation and summarization, constitute **language technologies**. The shifting technological landscape may be opening up opportunities for greater unification of these fields (e.g., Chrupała 2023).

ways expertise in linguistics continues to be essential in NLP—that is to say, there are ways NLP RELIES ON LINGUISTICS.

Imagine what the field would be like *without* linguistics, assuming for the sake of argument that current modeling approaches (word embeddings, transformers, LLMs) had somehow been developed with only a cursory understanding of how language works. Many of the systems might look similar, but the field would look very different. The field would have only the shallowest attention to linguistic analysis (i.e., tasks that focus on representing regularities in linguistic systems beyond the information communicated in a particular application setting). Even for objectives not primarily tied to linguistic analysis, our toolbox would be all the poorer:

- *Resources:* We would not have carefully curated datasets such as lexicons and corpora, with an appreciation for variation between languages, dialects, genres, and styles, etc. We would not have gold standard annotations of language-system phenomena, only of application-oriented phenomena.

- *Evaluation:* Not only would we not have gold standard evaluations for linguistic tasks, but for applied tasks, we would also lack expertise for designing effective human evaluations, interrogating automatic metrics, and characterizing the linguistic phenomena that challenge systems (such as anaphora or dialect variation).

- *Low-resource settings:* We would struggle more to understand why approaches that work well for English or French might not work well for Swahili or Arapaho (due to either the lack of data or the features of the language itself). We would not have the knowledge that would allow us to test linguistic inductive biases in neuro-symbolic models for greater accuracy.

- *Interpretability and Explanation:* It would be harder to develop and test hypotheses about how black box systems such as LLMs process language across domains, and we would lack appropriate metalanguage for describing many observed patterns.

- *Study of language:* Classic cL tasks such as parsing, coreference resolution, and textual entailment would not exist within our community. An NLP devoted purely to commercial technology would also be indifferent to goals of scholarly or community-driven linguistic work, such as documenting endangered languages.

In what follows, we give a sampling of the ways linguistics continues to play a role in NLP, organized under the mnemonic "RELIES:" resources (§3), evaluation (§4), low-resource settings (§5), interpretability/explanation (§6), and study of language (§7). It is worth emphasizing that these categories and the literature review are meant to be illustrative, not exhaustive; that we are speaking at a macro level; and that we are not claiming that linguistics is the only or necessarily the single most important source of expertise for working with language data and systems. We conclude by discussing pertinent challenges for the NLP community which involve linguistic expertise (§8).

2. Background

Today it is not a given that theories and representations from linguistics will form a direct foundation of NLP technologies; it is possible to do research in NLP without traditional training in linguistics. As was the case in the age of the "statistical revolution" of NLP (Johnson 2009; Uszkoreit 2009; Church 2011), deep learning was heralded as a "tsunami" (Manning 2015) and a way to model language "from scratch," rather than via features reflecting linguistic expertise (Collobert et al. 2011). Milestones in the neural era included the introduction of word embeddings (Mikolov et al. 2013) and then pretrained language models (Howard and Ruder 2018), with the well-known inflection points of ELMo and BERT (Peters et al. 2018; Devlin et al. 2019) in 2018/2019 and ChatGPT (Open AI 2022) in 2022/2023. The recent past has seen LLMs with billions of parameters that can be accessed by a wide range of users through a natural language interface ("prompting"). Do these developments prove the "bitter lesson" (Sutton 2019) that language technologies are most effectively developed solely by funneling data on a massive scale into rather general machine learning architectures? What role does linguistics play nowadays in NLP?

We are not the only researchers to reflect on these apparently tectonic shifts. Ignat et al. (2024) collect interesting topics for PhD students, particularly evaluation, and Saphra et al. (2024) hypothesize a cyclical historic model that then would suggest that familiar problems will resurface. We also see that the relationship between Linguistics and NLP (as well as the relationship between Linguistics and AI) has been a topic of discussion for some time (e.g., Lakoff 1978; Raskin 1985; Nirenburg 1986; Baldwin and Kordoni 2009; Xia, Lewis, and Levin 2010). In 2011, an entire collection of papers was devoted to the relationship between linguistics and CL (Baldwin and Kordoni 2011). There it was noted that the field of CL encompasses both scientific and engineering perspectives (Kay 2011; Johnson 2011); and that even with machine learning algorithms, long-tail (rare) phenomena celebrated by linguists are likely to challenge systems (Kay 2011; Levin 2011; Steedman 2011), and claims that NLP solutions will work well for any language need to be carefully evaluated in light of linguistic diversity (Bender 2011) and low-resource conditions in small languages (Bird 2011).

In what follows, we offer a contemporary take on how linguistics has enduring, practical relevance for NLP research, under the mnemonic "RELIES."

3. Resources

The empirical paradigm that dominates NLP today requires language datasets for training and evaluation. Such resources are supported by various degrees of linguistic knowledge—ranging from proficiency in a language to formal training in linguistics.

Resources for General NLP Tasks. By "general NLP tasks" we mean tasks that closely relate to applications in widespread demand, such as machine translation (MT), summarization, and sentiment classification. Empirical study of these tasks requires corpus resources. Even if we could now do without some of these resources in the *training* of NLP systems, they remain highly relevant for testing and studying systems.

The *selection and curation of data can be informed by language expertise*, even if the data is raw text. When developing a dataset, we strive for reflection of language diversity, variation among dialects/speakers, genre variation, code-switching, and (if not the target of a specific research question, probably) reduced social biases or problematic content (e.g., hate speech). Sensitivity to linguistic factors is particularly important

because of the wide variability of human language and the subjectivity inherent in certain labeling problems, like sentiment and toxic language (Sasidharan Nair, Dinkar, and Abercrombie 2024; Abercrombie, Hovy, and Prabhakaran 2023). Specific data selection techniques may involve shallow linguistic heuristics such as frequency counts, or linguistically advanced techniques that increase and validate the diversity and language phenomena coverage of the data (Dryer 1989; Rijkhoff et al. 1993). For example, Ravichander, Gardner, and Marasovic (2022) select data with sufficient *prepositions* and *complementizers*, and Ponti et al. (2020) inform their data selection with a *language typology index* (Littell et al. 2017). The BabyLM challenge (Warstadt et al. 2023; Hu et al. 2024) selects training and testing data according to cognitive linguistic plausibility, addressing fundamental questions in the learning process of LLMs.

In the annotation process, *advanced linguistic knowledge of annotators* has been of value, and people skilled at language analysis can ensure meaningful evaluation of NLP systems. For example, Freitag, Grangier, and Caswell (2020) observe that the quality and diversity of MT references improves when they are paraphrased by linguists. Additionally, only professional translators (rather than crowd-workers) lead to correct MT system rankings (Freitag et al. 2021a). For consistent sentiment annotation, Taboada (2016) highlights the importance of understanding linguistic discourse phenomena. Since not all tasks may be suitable for crowdsourcing without extensive experimentation, linguistically trained annotators can help reduce annotation cost (Gillick and Liu 2010).

Finally, *designing and coordinating* annotation protocols, especially via crowdsourcing, requires sensitivity to linguistic issues in order to craft guidelines and processes that produce a reasonable level of coherence across the annotations, and to recognize where disagreements can be explained by different perspectives (Ambati, Vogel, and Carbonell 2012; Plank 2022; Wein et al. 2023; Prabhakaran et al. 2024).

Resources with Linguistic Annotations. These encompass datasets of various sorts of grammatical and semantic structures. Linguistic knowledge and awareness is, of course, needed for designing analytical frameworks, and also for conducting annotation and verifying model output.

Two examples of linguistically detailed, resource-oriented frameworks prominent in current NLP research are Universal Dependencies (UD), with morphosyntactic annotations for over 150 languages (Nivre et al. 2016, 2020), and Abstract Meaning Representation (AMR), which is a framework for describing semantic graphs that has datasets in several languages (Banarescu et al. 2013; Wein and Schneider 2024). UD and AMR are *applied linguistic theories* (in the sense of Raskin 1985, p. 275). Research in and with linguistic annotation frameworks continues apace, with numerous recent workshops in the ACL community focused on grammar and semantics (Bonial and Tayyar Madabushi 2023; Dakota et al. 2023; Bonn and Xue 2023; Grobol and Tyers 2023; Rambow and Lareau 2023). The AMR framework alone is featured in over 450 papers,[4] and numerous applications (Wein and Opitz 2024; Sadeddine, Opitz, and Suchanek 2024).

Generally, linguistically annotated corpora and NLP models contribute to studying questions about language (§7). However, for some major NLP applications (like MT and QA) in high-resource settings, the future of explicit linguistic representations remains to be seen, with mixed results from neuro-symbolic models (Hamilton et al. 2022; Shwartz 2023). On the other hand, we consider the utility of such formal

---

4 https://nert-nlp.github.io/AMR-Bibliography/.

representations for evaluating and interpreting models (e.g., Xu et al. 2021; Opitz and Frank 2022; Fodor, De Deyne, and Suzuki 2024) to be an exciting path (**E**valuation, §4; **I**nterpretability, §6).

4. Evaluation

We illustrate where linguistics is useful to successfully **evaluate** an NLP system, that is, to describe the degree to which it conforms to our expectations. For tasks that take language input or generate language output, it is important to measure correctness and robustness at the language-system level, for which appropriate evaluations (quantitative and qualitative) are informed by linguistic expertise.[5]

Gold Standard Evaluation. Linguistic resources, particularly so-called *gold standard* annotations, play an important role in the evaluation of NLP technologies. They also contribute to the *benchmarking of LLMs and "AI systems"* (Tenney et al. 2019; Pimentel et al. 2020): Many of the tasks packaged in the *Beyond the Imitation Game Benchmark (BIG-bench)* (Srivastava et al. 2023) are linguistic tasks, such as bridging resolution and detecting common morphemes. Another kind of benchmark tests linguistic inference from just a few examples in puzzles (Şahin et al. 2020; Chi et al. 2024), a format conducive to low-resource languages without established annotation conventions (Bean et al. 2024).

Linguistic resources are also crucial for **system diagnostics** (including error analysis), which is a part of evaluation, but does not necessarily have the aim to produce a "global" ranking of AI systems. Instead, diagnostic methods help find more targeted improvement perspectives, with possible relevance even to broader society, since linguistic phenomena are often contingent with issues of societal interest, like gender biases in coreference (Rudinger et al. 2018) or QA (Parrish et al. 2022).[6] Overall, linguistic diagnostics can come in many forms; here we give a small sample. The "CheckList" (Ribeiro et al. 2020) tests NLP systems in various aspects, using linguistic features (negation, part of speech, etc.) to build templates. Song et al. (2022) assess LLMs on Chinese linguistic phenomena, and Parcalabescu et al. (2022) use linguistic aspects to study (computer) vision and language systems. Moore (2009) discusses examples of challenges that require linguistic knowledge to unpack, such as structural parallelism and translation of "WH" questions. Automated diagnostics can be achieved, e.g., through targeted measurements on parsed semantic structures (Lo and Wu 2011; Opitz and Frank 2021; Fan, Aumiller, and Gertz 2023, calling for work on the intrinsic NLP task of parsing), or stratifying evaluation data by semantic complexity (Antoine et al. 2024). Then there is a large area of probing systems with linguistic structures (Tenney et al. 2019; Rama, Beinborn, and Eger 2020; Pimentel et al. 2020; Starace et al. 2023), also tailored at certain branches of linguistics, e.g., psycholinguistics (Gauthier et al. 2020; Ettinger 2020).

Human Evaluation and Meta-evaluation. The combination of varied *human evaluation* practices with progress towards highly fluent models has led to diminished reliability of human evaluations: Since generated text has become so fluent, it is harder for humans to distinguish their quality. As a result, not all human evaluation studies are useful

---

5 A similar sentiment was expressed by Levin (2011) in the statistical NLP era: "language technologists [ought to] understand the object of study, human language . . . in order to understand where current methods are falling short, we as a field need to understand the data" (p. 19).

6 Raji et al. (2021) generally argue for more targeted evaluation methods to track "AI progress."

measures of model output (Clark et al. 2021; Freitag et al. 2021a), necessitating standardization with regard to design and terminology (Howcroft et al. 2020; Belz, Mille, and Howcroft 2020; van der Lee et al. 2019). Making use of implicit linguistic knowledge held by native speakers (e.g., judging the fluency of system output) or more explicit linguistic training (e.g., assessing the syntactic diversity of system output), can serve a crucial role in reliably assessing the state of the field (Michael et al. 2016; Callison-Burch et al. 2008).

To design approaches for **meta-evaluation** (assessing and comparing metrics), expertise in linguistics can prove highly useful, since the challenging linguistic phenomena for models and metrics need to be understood in order to be identified. In a WMT metric shared task, Freitag et al. (2021b) find that most metrics struggle to handle complex semantic phenomena and do not appropriately punish the reversal of sentiment or negation; this is established with test sets that assess specified linguistic features (here: sentiment or negation reversal).

New Metrics. Linguistic annotations can help build automatic metrics. Some NLG metrics are built from entailment data[7] (Chen and Eger 2023; Xie et al. 2021; Scialom et al. 2021; Steen et al. 2023). Measures for biases, like social biases, have drawn upon the efforts of linguists. For instance, Steen and Markert (2024) propose bias evaluation measures from socio-linguistic lexicons (gender, race) and OntoNotes (Pradhan et al. 2013).

5. Low-resource Settings

In NLP and ML, we aim to *efficiently learn to generalize* to new tasks, domains, and languages, including *global and historic languages*, where feedback and data is sparse or nonexistent (Hedderich et al. 2021).

Processing of Global and Historic Languages. Linguistics is crucial in situations where the amount of language data is limited, e.g., when developing technologies for languages with a limited amount of available recordings or written data. We can improve generalization performance across an array of LLM models by exploiting linguistic features in training (Zhang et al. 2024) or dictionaries for balanced data augmentation (Lu et al. 2024); also with benefits to speech processing (Kim, Jeon, and Lee 2024). Processing particular types of language/particular languages can also pose challenges that require input from linguistics. For examples, *polysynthetic languages* such as those spoken in Canada (Gupta and Boulianne 2020a,b) pose problems for automatic speech recognition (ASR): If "word" is taken as the basic unit for recognition (which makes sense for most languages, but not these ones) the out-of-vocabulary (OOV) rate on new data is extreme.

At first glance, Tanzer et al.'s (2024) study may appear as a potential "linguistics-free" avenue for addressing some extremely low resource-scenarios, since it suggests that LLMs could, to a certain extent, translate a language from one linguist's grammar book instead of a corpus. However, the study shows crucial reliance on linguistics: A linguist is required to assess LLM performance, and even the largest LLMs considerably underperform relative to a linguist (unfamiliar with the language) who has read

---

7 While entailment annotations may come from non-linguists, the notion of textual entailment was inspired by traditions from logic and formal semantics (Dagan, Glickman, and Magnini 2005; Bowman et al. 2015).

the book. Of course, it is also the case that the fundamental resource was created by a linguist.

Even when more data is available, linguistics can be of value: It can inform data selection in multilingual pretraining (Ogunremi, Jurafsky, and Manning 2023), or more language-balanced tokenization (Creutz and Lagus 2002; Limisiewicz et al. 2024; Fusco et al. 2024).[8]

Resources for Endangered Languages. According to some predictions, half of the world's spoken languages might be on the brink of extinction (UNESCO 2023). Thus, there is scope for NLP tooling to document and revitalize endangered[9] languages (e.g., van Esch, Foley, and San 2019; Levow, Ahn, and Bender 2021; San et al. 2022), ideally through collaborations involving community members, linguists, and NLP practitioners. Linguistic expertise is relevant for structurally collecting and documenting language data, as well as helping to distill knowledge about the language into dictionaries and rules. Recent work has brought community members and NLP practitioners together for the purpose of language documentation through provisioning access to speech tools such as ASR (San et al. 2022), building datasets on online platforms (Everson, Honore, and Grimm 2019; Zuckermann et al. 2021), and analysis of phonetic data (Kempton 2017).

Compute Restriction. Systems that require less computation, particularly at inference time, are of great interest. Critically, off-the-shelf conventional parsers/taggers defend their status as the go-to method in some situations, like for certain forms of data analysis (e.g., counting the most common nouns), or affordable processing of huge corpora (e.g., from sentence segmentation to OpenIE [Angeli, Johnson Premkumar, and Manning 2015]).

Effective engineering solutions to efficiency are achieved through quantization (Dettmers et al. 2022) or low-rank adaptation (Hu et al. 2022) of LLMs. However, soft inductive linguistic biases might be complementary, or an alternative in specific situations. Tailored evaluation metrics could help assess any incurred tradeoffs (Zhou et al. 2022).

Linguistically Sensitive Supervision. Developers of technologies for under-resourced languages do not only encounter data scarcity, but probably also scarcity of input from native speakers who could help oversee a system's design and application. Indeed, releasing technology in a top–down fashion can be harmful to local communities (Bird 2022; Doğruöz and Sitaram 2022; Bird and Yibarbuk 2024). Thus, when applied with care, *(field) linguistics* can help become aware of the diverse cultural contexts of local language communication situations, resulting in more sensitive NLP technologies, or the abstention from releasing harmful ones (e.g., when there is no community oversight/feedback).

6. Interpretability and Explanation

How can we efficiently reason about, and explain, observations of observed language phenomena, language systems, and processes, as well as any of their models in the

---

8 Some research also suggests that "simply scaling up the number of languages ... in the pretraining is unhelpful" for generalization (Adelani et al. 2024), and "bigger is not always better" (Wilcox et al. 2024).

9 Terms like "endangered" and "documentation" are widespread in the literature we are surveying, though we also wish to point at a critique of this terminology, suggesting "reclamation" instead (Leonard 2020).

form of NLP systems? We need a *shared terminology and metalanguage*, and develop *binding methods* that relate model internals to human-understandable concepts from said terminology. We argue that linguistics fundamentally helps with both aspects.

As the field concerned with systems of human language, linguistics offers a *metalanguage*—in the form of jargon and formalisms—for precise characterization of linguistic phenomena. The field of NLP draws on this metalanguage, as can be seen in the literature in observations like "phonemic representations exhibit higher similarities between languages compared to orthographic representations" (Jung et al. 2024) and "difficult to score [in MT evaluation] are the transitive past progressive, multiple connectors, and the ditransitive simple future I for English to German, and pseudogapping, contact clauses, and cleft sentences for English to Russian" (Avramidis et al. 2024).

For understanding computational models of language such as LLMs with billions and trillions of parameters, directed **binding methods** are being developed that learn to relate model internals and processes to specific human-understandable descriptions, that often lie within our vocabulary of linguistic metalanguage. We will consider a few examples of such methods that particularly use a linguistic notion of binding: Rassin et al. (2023) bind linguistic concepts in Vision and Language diffusion models with syntax trees and attention maps; Di Marco, Hämmerl, and Fraser (2023) and Jumelet and Zuidema (2023) detect linguistic feature representations in LLMs with prompts; Opitz and Frank (2022) bind parts of neural embedding representations to semantic aspects; Geva et al. (2022) view model decisions as constructed from concepts in a vocabulary. Explanations and binding can also adopt an outside, behavioral view (Beguš, Dąbkowski, and Rhodes 2023; Behzad et al. 2023; Chang and Bergen 2024). For instance, Muñoz-Ortiz, Gómez-Rodríguez, and Vilares (2023) contrast linguistic patterns in human- and LLM-generated text in morphological, syntactic, and sociolinguistic aspects. Since such methods can also be seen as diagnostic methods, the section on **E**valuation (§4) is also relevant.

In the bigger picture, linguistic metalanguage can help us understand, or establish hypotheses, for what happens in complex NLP systems, and how they differ. An example is the hypothesis that a pretrained LM can be viewed through the lens of a classical NLP pipeline, which has received arguments for and against (Tenney, Das, and Pavlick 2019; de Vries, van Cranenburgh, and Nissim 2020; Niu, Lu, and Penn 2022). The learning trajectory of LLMs can be traced linguistically, e.g., noting the point of syntax acquisition (Chen et al. 2024). LLMs have also been interpreted as "models of varieties of language" (a sociolinguistic notion, Grieve et al. 2025).

Building a bridge to **R**esources (§3), we can use *datasets* that elicit linguistic phenomena to "open the black box" of LLMs down to the level of single neurons, as exemplified by Niu et al. (2024), who study neuron-level knowledge with the BLiMP corpus of linguistic minimal pairs (Warstadt et al. 2020). With similar means, Wang, Hu, and Zhao (2024) show that BERT-like models focus on sentence-level features, whereas LLMs such as GPT or Llama are sensitive to conventions, language complexity, and organization. Experiments with complex and simple language structures can illuminate learning environments for LLMs (Qin, Saphra, and Alvarez-Melis 2024).

Linguistically agnostic interpretability methods can be mechanistic, i.e., formalizing the algorithms learned by neural networks (Weiss, Goldberg, and Yahav 2021), or based on attributing a prediction to inputs via integration (Sundararajan, Taly, and Yan 2017), classifiers (Ribeiro, Singh, and Guestrin 2016), or Shapley values (Shapley 1951). However, still, for interpreting and evaluating the results of such methods, we rely on linguistic analysis (Schuff et al. 2022; Feldhus et al. 2023; Moeller, Nikolaev, and Padó 2024) to verify that an interpretability method "outputs meaningful

explanations" (Mardaoui and Garreau 2021) or "emphasize[s] specific types of linguistic compositions" (Kobayashi et al. 2024). This means that linguistic analyses are used to validate interpretability methods whose algorithms do not explicitly incorporate such representations.

Finally, at the thousand-foot view, we see ideas from linguistics and adjacent fields take center stage in debates about how to *interpret* what NLP models are capable of representing—like how to define machine "understanding" (Dunietz et al. 2020; Ray Choudhury, Rogers, and Augenstein 2022), and specifically, whether grounding is required for a model to capture meaning (e.g., Bender and Koller 2020; Merrill et al. 2021; Pavlick 2023).

7. Study of Language

Our last facet highlights linguistics (and related areas) as the application domain. That is, those who study language, but are not necessarily computational linguists themselves, are a user base that can motivate NLP tasks and tools,[10] not NLP models and representations generally. These *language-system-focused* applications can be distinguished from most contemporary NLP applications, which seek to make *content* more accessible through language (for example, by translating it, summarizing it, or reasoning about it).

Classic cL tasks like parsing have applications to corpus linguistics: In order to study a linguistic pattern, it is often important to query not just strings, but also grammatical abstractions like tags, dependencies, and phrases. This was not the original goal of parsing researchers: Parsing was a cL topic before the days of substantial digital corpora, and was studied to examine computational capacity to model natural language grammar, highlighting issues like formal expressive power (Steedman 2011). Now, the tools produced have found a secondary application in corpus linguistics, with parser outputs accessible through corpus search engines (Resnik and Elkiss 2005; Ghodke and Bird 2012; Bender et al. 2012; Hundt, Denison, and Schneider 2012; Krause and Zeldes 2016; Guillaume 2021; Kulick, Ryant, and Santorini 2022).[11] The application to corpus linguistics motivates new angles on parsing research (e.g., syntactically searching a large text collection without having to pre-parse it in its entirety, or in a way that accounts for parser uncertainty).

Documentary and historical linguistics also motivate and contribute to NLP advancement. In documentary and historical linguistics, data may be sparse, fragmented, primarily in image or audio form (without transcriptions), and lacking a standard orthography; the basic grammar of the language may still be a mystery; and the language may have few or no living speakers. (The case of endangered language documentation was discussed above in §5.) The exigencies of such settings call for considerably more noise-tolerance and interactivity than run-of-the-mill NLP tasks. Pure-text tasks like

---

10 Unlike the previous categories, which highlight ways that linguistics contributes to NLP, we view this category as bidirectional: Language study motivates research towards developing NLP tools that can then contribute to the study of language. To be clear, we regard the entirety of cL—spanning formal, descriptive, distributional/statistical, and experimental inquiries—as part of linguistics. (This includes the development and use of frameworks and resources of the sorts discussed in §3.) These inquiries contribute to linguistic understanding by advancing accounts of linguistic diversity and the precision and robustness of theories. Naturally, NLP models and representations are part of the methodology of this research (in particular, the potential relevance of LLMs to broader linguistics and cognitive science has been widely debated: see Futrell and Mahowald [2025] for a review). For purposes of this section, though, we focus on specialized NLP software that can be applied to language studies.

11 Search in parsed corpora may also be motivated by information extraction applications (Shlain et al. 2020).

normalization (Robertson and Goldwater 2018), as well as signal processing capabilities like OCR for historical documents (Berg-Kirkpatrick, Durrett, and Klein 2013) or low-resource speech processing (Duong et al. 2016; Anastasopoulos et al. 2017; Liu, Spence, and Prud'hommeaux 2022), are important here. Algorithms and software infrastructure to assist humans engaging in multiple tasks collaboratively (elicitation, transcription, annotation, discovering grammatical generalizations, and developing usable resources for scholars and community members) in limited-data multimodal conditions are needed, and there is a long way to go before they can be used seamlessly by non-computational linguists (Gessler 2022; Moeller and Arppe 2024).

Language teaching, as studied in the field of Applied Linguistics and practiced in classrooms as well as other modes of instruction and technological support, is another area ripe for further engagement with NLP. Much of the research in the NLP community has been formulated in narrow tasks like grammatical error correction (as a generation task) or essay scoring (as a regression task) (Burstein and Chodorow 1999; Massung and Zhai 2016; Wang et al. 2021; Klebanov and Madnani 2022). Applications with richer contexts—for example, technologies that would monitor student progress over time and deliver adaptive pedagogical experiences—are perhaps more difficult to study, but could enhance student learning more holistically (Cui and Sachan 2023; Qian et al. 2023; Glandorf et al. 2025; Bang et al. 2024).

Finally, we note the potential for NLP to embrace applications facilitating language-focused study in fields beyond linguistics proper, including literature (Jakobson 1987), education, law (Marmor 2014), communications (Miller 1951), translation studies (Gambier and Van Doorslaer 2009), history (Piotrowski 2012), argumentation (Cohen 1987), and lexicography (McArthur 1998). In the NLP community, these fields are associated and developed with active workshops, e.g., LaTeCH-CLfL (CL for Cultural Heritage, Social Sciences, Humanities and Literature—8th iteration in 2024) (Bizzoni et al. 2024); BEA (Building Educational Applications—19th iteration) (Kochmar et al. 2024); Argument Mining (11th iteration) (Ajjour et al. 2024); and NLLP (Natural Legal Language Processing–3rd iteration) (Aletras et al. 2024), among others. Key objectives featured in these workshops include creating structured resources, increasing interpretability, evaluating models and resources, and coping with low-resource settings. This underscores the importance of interdisciplinary work between all of NLP, linguistics, and other fields from humanities and social sciences.

8. Conclusion and Outlook

For a long time, automating analysis into linguistically based representations—classic "cL"—was undertaken in part with the aim of paving the path towards Natural Language Understanding, building modules that capture features of language based on human conceptualizations (realized as manually crafted rules, or latent rules induced through training on annotations). Given the rise of LLMs that leverage huge amounts of data to recognize patterns in text, the question of the relationship between natural language processing and linguistics—in particular, how linguistic knowledge benefits work in natural language processing—is a pressing one. In this article, we have illustrated the role of linguistics when compiling resources and conducting system evaluations (RE-); when building systems in low-resource settings (-L-); when pursuing granular interpretation or control of large-data systems (-IE-); and when connecting NLP to the study of human language (-S). While explicit linguistic knowledge is not necessary to achieve high accuracy on all NLP tasks, in this work we have highlighted the engineering utility of language knowledge.

Beyond continuing to develop cL resources and tools, our review has highlighted *other pertinent challenges for the NLP community*, including: Efforts towards preserving the world's languages via cooperation between linguists, machine learning experts, and language communities (**RLS**); language learning applications and computational models of language acquisition, with potential for both the use and analysis of linguistic phenomena in NLP (**ELIES**); and integration of symbolic representations and linguistic criteria generally into approaches that promote interpretability and explanation (**IE**).

References

Abercrombie, Gavin, Dirk Hovy, and Vinodkumar Prabhakaran. 2023. Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 96–103. `https://doi.org/10.18653/v1/2023.law-1.10`

Adelani, David, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245.

Ajjour, Yamen, Roy Bar-Haim, Roxanne El Baff, Zhexiong Liu, and Gabriella Skitalinskaya, editors. 2024. *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*.

Aletras, Nikolaos, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, Daniel Preoțiuc-Pietro, and Gerasimos Spanakis, editors. 2024. *Proceedings of the Natural Legal Language Processing Workshop 2024*.

Ambati, Vamshi, Stephan Vogel, and Jaime Carbonell. 2012. Collaborative workflow for crowdsourcing translation. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1191–1194. `https://doi.org/10.1145/2145204.2145382`

Anastasopoulos, Antonios, Sameer Bansal, David Chiang, Sharon Goldwater, and Adam Lopez. 2017. Spoken term discovery for language documentation using translations. In *Proceedings of the Workshop on Speech-Centric Natural Language Processing*, pages 53–58. `https://doi.org/10.18653/v1/W17-4607`

Angeli, Gabor, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354. `https://doi.org/10.3115/v1/P15-1034`

Antoine, Elie, Frederic Bechet, Géraldine Damnati, and Philippe Langlais. 2024. A linguistically-motivated evaluation methodology for unraveling model's abilities in reading comprehension tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18376–18392. `https://doi.org/10.18653/v1/2024.emnlp-main.1021`

Avramidis, Eleftherios,, Shushen Manakhimova, Vivien Macketanz, and Sebastian Möller. 2024. Machine translation metrics are better in evaluating linguistic errors on LLMs than on encoder-decoder systems. In *Proceedings of the Ninth Conference on Machine Translation*, pages 517–528. `https://doi.org/10.18653/v1/2024.wmt-1.37`

Baldwin, Timothy and Valia Kordoni, editors. 2009. *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*

Baldwin, Timothy and Valia Kordoni. 2011. The interaction between linguistics and computational linguistics: Virtuous,

vicious, or vacuous? *Linguistic Issues in Language Technology*, 6. `https://doi.org/10.33011/lilt.v6i.1233`

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Bang, Hee Jin, Eric Setoguchi, Alison Mackey, and Akiko Fujii. 2024. L2 learning outcomes of a research-based digital app for Japanese children. *Studies in Second Language Acquisition*, 46(2):504–534. `https://doi.org/10.1017/S0272263124000147`

Bean, Andrew Michael, Simeon Hellsten, Harry Mayne, Jabez Magomere, Ethan A Chi, Ryan Andrew Chi, Scott A. Hale, and Hannah Rose Kirk. 2024. Lingoly: A benchmark of olympiad-level linguistic reasoning puzzles in low resource and extinct languages. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Beguš, Gašper, Maksymilian Dąbkowski, and Ryan Rhodes. 2023. Large linguistic models: Analyzing theoretical linguistic abilities of LLMs. *arXiv preprint arXiv:2305.00948*.

Behzad, Shabnam, Keisuke Sakaguchi, Nathan Schneider, and Amir Zeldes. 2023. ELQA: A corpus of metalinguistic questions and answers about English. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2047. `https://doi.org/10.18653/v1/2023.acl-long.113`

Belz, Anya, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194. `https://doi.org/10.18653/v1/2020.inlg-1.24`

Bender, Emily M. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6. `https://doi.org/10.33011/lilt.v6i.1239`

Bender, Emily M., Sumukh Ghodke, Timothy Baldwin, and Rebecca Dridan. 2012. From

database to treebank: Enhancing hypertext grammars with grammar engineering and treebank search. In Sebastian Nordhoff and Karl-Ludwig G. Poggeman, editors, *Electronic Grammaticography*. University of Hawaii Press, pages 179–206.

Bender, Emily M. and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of ACL*, pages 5185–5198. `https://doi.org/10.18653/v1/2020.acl-main.463`

Berg-Kirkpatrick, Taylor, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 207–217.

Bird, Steven. 2011. Bootstrapping the language archive: New prospects for natural language processing in preserving linguistic heritage. *Linguistic Issues in Language Technology*, 6. `https://doi.org/10.33011/lilt.v6i.1243`

Bird, Steven. 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. `https://doi.org/10.18653/v1/2022.acl-long.539`

Bird, Steven and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839.

Bizzoni, Yuri, Stefania Degaetano-Ortlieb, Anna Kazantseva, and Stan Szpakowicz, editors. 2024. *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*.

Bonial, Claire and Harish Tayyar Madabushi, editors. 2023. *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*.

Bonn, Julia and Nianwen Xue, editors. 2023. *Proceedings of the Fourth International Workshop on Designing Meaning Representations*.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language*

Processing, pages 632–642. `https://doi.org/10.18653/v1/D15-1075`

Burstein, Jill and Martin Chodorow. 1999. Automated essay scoring for nonnative English speakers. In *Computer Mediated Language Assessment and Evaluation in Natural Language Processing*. `https://doi.org/10.3115/1598834.1598847`

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106. `https://doi.org/10.3115/1626394.1626403`

Chang, Tyler A. and Benjamin K. Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350. `https://doi.org/10.1162/coli_a_00492`

Chen, Angelica, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L. Leavitt, and Naomi Saphra. 2024. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In the *Twelfth International Conference on Learning Representations*.

Chen, Yanran and Steffen Eger. 2023. MENLI: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11:804–825. `https://doi.org/10.1162/tacl_a_00576`

Chi, Nathan, 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 *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 113–119.

Chrupała, Grzegorz. 2023. Putting natural in natural language processing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7820–7827. `https://doi.org/10.18653/v1/2023.findings-acl.495`

Church, Kenneth. 2011. A pendulum swung too far. *Linguistic Issues in Language Technology*, 6. `https://doi.org/10.33011/lilt.v6i.1245`

Clark, Elizabeth, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296. `https://doi.org/10.18653/v1/2021.acl-long.565`

Cohen, Robin. 1987. Analyzing the structure of argumentative discourse. *Computational Linguistics*, 13:11–24.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Creutz, Mathias and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30. `https://doi.org/10.3115/1118647.1118650`

Cui, Peng and Mrinmaya Sachan. 2023. Adaptive and personalized exercise generation for online language learning. In *Proceedings of ACL*, pages 10184–10198. `https://doi.org/10.18653/v1/2023.acl-long.567`

Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, pages 177–190. `https://doi.org/10.1007/11736790_9`

Dakota, Daniel, Kilian Evang, Sandra Kübler, and Lori Levin, editors. 2023. *Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023)*.

Dettmers, Tim, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

de Vries, Wietse, Andreas van Cranenburgh, and Malvina Nissim. 2020. What's so special about BERT's layers? A closer look at the NLP pipeline in monolingual and multilingual models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4339–4350.

https://doi.org/10.18653/v1/2020
.findings-emnlp.389

Di Marco, Marion, Katharina Hämmerl, and Alexander Fraser. 2023. A study on accessing linguistic information in pre-trained language models by using prompts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7328–7336. https://doi.org/10.18653/v1/2023
.emnlp-main.454

Doğruöz, A. Seza and Sunayana Sitaram. 2022. Language technologies for low resource languages: Sociolinguistic and multilingual insights. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 92–97.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 13(2):257–292. https://doi.org/10
.1075/sl.13.2.03dry

Dunietz, Jesse, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7839–7859. https://doi.org/10.18653/v1/2020
.acl-main.701

Duong, Long, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 949–959. https://doi.org/10.18653/v1/N16
-1109

Ettinger, Allyson. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48. https://doi.org/10.1162/tacl
_a_00298

Everson, Rebecca, Wolf Honore, and Scott Grimm. 2019. An online platform for community-based language description and documentation. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 1–5. https://doi.org/10.33011/computel
.v1i.259

Fan, Jing, Dennis Aumiller, and Michael Gertz. 2023. Evaluating factual consistency of texts with semantic role labeling. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 89–100. https://doi.org
/10.18653/v1/2023.starsem-1.9

Feldhus, Nils, Leonhard Hennig, Maximilian Dustin Nasert, Christopher Ebert, Robert Schwarzenberg, and Sebastian Möller. 2023. Saliency map verbalization: Comparing feature importance representations from model-free and instruction-based methods. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 30–46. https://doi.org
/10.18653/v1/2023.nlrse-1.4

Fodor, James, Simon De Deyne, and Shinsuke Suzuki. 2024. Compositionality and sentence meaning: Comparing semantic parsing and transformers on a challenging sentence similarity dataset. *Computational Linguistics*, pages 1–52. https://doi.org/10.1162/coli_a
_00536

Freitag, Markus, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474. https://doi.org/10.1162/tacl_a
_00437

Freitag, Markus, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71. https://doi.org/10.18653
/v1/2020.emnlp-main.5

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774.

Fusco, Achille, Matilde Barbini, Maria Letizia Piccini Bianchessi, Veronica Bressan, Sofia Neri, Sarah Rossi, Tommaso Sgrizzi, and Cristiano Chesi. 2024. Recurrent networks are (linguistically) better? An experiment on small-LM training on child-directed speech in Italian. In *CEUR WORKSHOP PROCEEDINGS*, volume 3878.

Futrell, Richard and Kyle Mahowald. 2025.
How linguistics learned to stop worrying
and love the language models. *arXiv
preprint arXiv:2501.17047.*

Gambier, Yves and Luc Van Doorslaer,
editors. 2009. *Metalanguage of Translation.*
John Benjamins Publishing Company.
`https://doi.org/10.1075/bct.20`

Gauthier, Jon, Jennifer Hu, Ethan Wilcox,
Peng Qian, and Roger Levy. 2020.
SyntaxGym: An online platform for
targeted evaluation of language models.
In *Proceedings of the 58th Annual Meeting
of the Association for Computational
Linguistics: System Demonstrations,*
pages 70–76. `https://doi.org/10`
`.18653/v1/2020.acl-demos.10`

Gessler, Luke. 2022. Closing the NLP gap:
Documentary linguistics and NLP need a
shared software infrastructure. In
*Proceedings of the Fifth Workshop on the Use
of Computational Methods in the Study of
Endangered Languages,* pages 119–126.
`https://doi.org/10.18653/v1/2022`
`.computel-1.15`

Geva, Mor, Avi Caciularu, Kevin Wang, and
Yoav Goldberg. 2022. Transformer
feed-forward layers build predictions by
promoting concepts in the vocabulary
space. In *Proceedings of the 2022 Conference
on Empirical Methods in Natural Language
Processing,* pages 30–45. `https://doi.org`
`/10.18653/v1/2022.emnlp-main.3`

Ghodke, Sumukh and Steven Bird. 2012.
Fangorn: A system for querying very
large treebanks. In *Proceedings of
COLING 2012: Demonstration Papers,*
pages 175–182.

Gillick, Dan and Yang Liu. 2010. Non-expert
evaluation of summarization systems is
risky. In *Proceedings of the NAACL HLT
2010 Workshop on Creating Speech and
Language Data with Amazon's Mechanical
Turk,* pages 148–151.

Glandorf, Dominik, Peng Cui, Detmar
Meurers, and Mrinmaya Sachan. 2025.
Grammar control in dialogue response
generation for language learning chatbots.
In *Proceedings of NAACL. ArXiv preprint
arXiv:2502.07544.*

Grieve, Jack, Sara Bartl, Matteo Fuoli, Jason
Grafmiller, Weihang Huang, Alejandro
Jawerbaum, Akira Murakami, Marcus
Perlman, Dana Roemling, and Bodo
Winter. 2025. The sociolinguistic
foundations of language modeling.
Frontiers in Artificial Intelligence, 7:1472411.
`https://doi.org/10.3389/frai`
`.2024.1472411`, PubMed: 39871863

Grobol, Loïc and Francis Tyers, editors. 2023.
*Proceedings of the Sixth Workshop on
Universal Dependencies (UDW,
GURT/SyntaxFest 2023).*

Guillaume, Bruno. 2021. Graph matching
and graph rewriting: GREW tools for
corpus exploration, maintenance and
conversion. In *Proceedings of the 16th
Conference of the European Chapter of the
Association for Computational Linguistics:
System Demonstrations,* pages 168–175.
`https://doi.org/10.18653/v1/2021`
`.eacl-demos.21`

Gupta, Vishwa and Gilles Boulianne. 2020a.
Automatic transcription challenges for
Inuktitut, a low-resource polysynthetic
language. In *Proceedings of the Twelfth
Language Resources and Evaluation
Conference,* pages 2521–2527.

Gupta, Vishwa and Gilles Boulianne. 2020b.
Speech transcription challenges for
resource constrained indigenous language
Cree. In *Proceedings of the 1st Joint Workshop
on Spoken Language Technologies for
Under-resourced languages (SLTU) and
Collaboration and Computing for
Under-Resourced Languages (CCURL),*
pages 362–367.

Hamilton, Kyle, Aparna Nayak, Bojan Božić,
and Luca Longo. 2022. Is neuro-symbolic
AI meeting its promises in natural
language processing? A structured review.
Semantic Web. `https://doi.org/10`
`.3233/SW-223`

Hedderich, Michael A., Lukas Lange, Heike
Adel, Jannik Strötgen, and Dietrich
Klakow. 2021. A survey on recent
approaches for natural language
processing in low-resource scenarios. In
*Proceedings of the 2021 Conference of the
North American Chapter of the Association for
Computational Linguistics: Human Language
Technologies,* pages 2545–2568. `https://`
`doi.org/10.18653/v1/2021.naacl`
`-main.201`

Howard, Jeremy and Sebastian Ruder. 2018.
Universal language model fine-tuning for
text classification. In *Proceedings of the 56th
Annual Meeting of the Association for
Computational Linguistics (Volume 1: Long
Papers),* pages 328–339. `https://doi.org`
`/10.18653/v1/P18-1031`

Howcroft, David M., Anya Belz,
Miruna-Adriana Clinciu, Dimitra Gkatzia,
Sadid A. Hasan, Saad Mahamood, Simon
Mille, Emiel van Miltenburg, Sashank
Santhanam, and Verena Rieser. 2020.
Twenty years of confusion in human
evaluation: NLG needs evaluation sheets

and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182. `https://doi.org/10.18653/v1/2020.inlg-1.23`

Hu, Edward, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hu, Michael Y., Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. *arXiv preprint arXiv:2412.05149*.

Hundt, Marianne, David Denison, and Gerold Schneider. 2012. Retrieving relatives from historical data. *Literary and Linguistic Computing*, 27(1):3–16. `https://doi.org/10.1093/llc/fqr049`

Ignat, Oana, Zhijing Jin, Artem Abzaliev, Laura Biester, Santiago Castro, Naihao Deng, Xinyi Gao, Aylin Gunal, Jacky He, Ashkan Kazemi, Muhammad Khalifa, Namho Koh, Andrew Lee, Siyang Liu, Do June Min, Shinka Mori, Joan Nwatu, Veronica Perez-Rosas, Siqi Shen, Zekun Wang, Winston Wu, and Rada Mihalcea. 2024. Has it all been solved? Open NLP research questions not solved by large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Jakobson, Roman. 1987. *Language in Literature*. Harvard University Press.

Johnson, Mark. 2009. How the statistical revolution changes (computational) linguistics. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 3–11. `https://doi.org/10.3115/1642038.1642041`

Johnson, Mark. 2011. How relevant is linguistics to computational linguistics? *Linguistic Issues in Language Technology*, 6. `https://doi.org/10.33011/lilt.v6i.1249`

Jumelet, Jaap and Willem Zuidema. 2023. Feature interactions reveal linguistic structure in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8697–8712. `https://doi.org/10.18653/v1/2023.findings-acl.554`

Jung, Haeji, Changdae Oh, Jooeon Kang, Jimin Sohn, Kyungwoo Song, Jinkyu Kim, and David R Mortensen. 2024. Mitigating the linguistic gap with phonemic representations for robust cross-lingual transfer. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 200–211. `https://doi.org/10.18653/v1/2024.mrl-1.16`

Kay, Martin. 2011. Zipf's law and l'arbitraire du signe. *Linguistic Issues in Language Technology*, 6. `https://doi.org/10.33011/lilt.v6i.1251`

Kempton, Timothy. 2017. Cross-language forced alignment to assist community-based linguistics for low resource languages. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 165–169. `https://doi.org/10.18653/v1/W17-0122`

Kim, Youngjae, Yejin Jeon, and Gary Lee. 2024. Audio-based linguistic feature extraction for enhancing multi-lingual and low-resource text-to-speech. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13984–13989. `https://doi.org/10.18653/v1/2024.findings-emnlp.817`

Klebanov, Beata Beigman and Nitin Madnani. 2022. *Automated Essay Scoring*. Synthesis Lectures on Human Language Technologies. Springer International Publishing. `https://doi.org/10.1007/978-3-031-02182-4`

Kobayashi, Goro, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2024. Analyzing feed-forward blocks in transformers through the lens of attention maps. In the *Twelfth International Conference on Learning Representations*.

Kochmar, Ekaterina, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan, editors. 2024. *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*.

Krause, Thomas and Amir Zeldes. 2016. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118–139. `https://doi.org/10.1093/llc/fqu057`

Kulick, Seth, Neville Ryant, and Beatrice Santorini. 2022. Parsing early modern

English for linguistic search. *Proceedings of the Society for Computation in Linguistics*, 5:143–157.

Lakoff, George. 1978. Some remarks on AI and linguistics. *Cognitive Science*, 2(3):267–275. https://doi.org/10.1207/s15516709cog0203_4

Leonard, Wesley Y. 2020. Insights from Native American Studies for theorizing race and racism in linguistics (Response to Charity Hudley, Mallinson, and Bucholtz). *Language*, 96(4):e281–e291. https://doi.org/10.1353/lan.2020.0079

Levin, Lori. 2011. Variety, idiosyncracy, and complexity in language and language technologies. *Linguistic Issues in Language Technology*, 6. https://doi.org/10.33011/lilt.v6i.1255

Levow, Gina-Anne, Emily Ahn, and Emily M. Bender. 2021. Developing a shared task for speech processing on endangered languages. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 96–106. https://doi.org/10.33011/computel.v1i.967

Limisiewicz, Tomasz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, and Luke Zettlemoyer. 2024. MYTE: Morphology-driven byte encoding for better and fairer multilingual language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15059–15076. https://doi.org/10.18653/v1/2024.acl-long.804

Littell, Patrick, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14. https://doi.org/10.18653/v1/E17-2002

Liu, Zoey, Justin Spence, and Emily Prud'hommeaux. 2022. Enhancing documentation of Hupa with automatic speech recognition. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 187–192. https://doi.org/10.18653/v1/2022.computel-1.23

Llama Team. 2024. The Llama 3 herd of models. AI @ Meta.

Lo, Chi-kiu and Dekai Wu. 2011. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 220–229.

Lu, Yinquan, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. LLaMAX: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10748–10772. https://doi.org/10.18653/v1/2024.findings-emnlp.631

Manning, Christopher D. 2015. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707. https://doi.org/10.1162/COLI_a_00239

Mardaoui, Dina and Damien Garreau. 2021. An analysis of LIME for text data. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3493–3501.

Marmor, Andrei. 2014. *The Language of Law*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198714538.001.0001

Massung, Sean and Chengxiang Zhai. 2016. Non-native text analysis: A survey. *Natural Language Engineering*, 22(2):163–186. https://doi.org/10.1017/S1351324915000303

McArthur, Tom. 1998. *Living Words: Language, Lexicography, and the Knowledge Revolution*. University of Exeter Press.

Merrill, William, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060. https://doi.org/10.1162/tacl_a_00412

Michael, Erica, Petra Bradley, Paul McNamee, and Matt Post. 2016. Putting the "human" back in HLT: The importance of human evaluation in assessing the quality and potential uses of translation technology. In *Conferences of the Association for Machine Translation in the Americas: MT Users' Track*, pages 453–550.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013.

Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, Curran Associates, Inc.

Miller, George Armitage. 1951. *Language and Communication*. McGraw-Hill. https://doi.org/10.1037/11135-000

Moeller, Lucas, Dmitry Nikolaev, and Sebastian Padó. 2024. Approximate attributions for off-the-shelf Siamese transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2059–2071.

Moeller, Sarah and Antti Arppe. 2024. Machine-in-the-loop with documentary and descriptive linguists. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 27–32.

Moore, Robert C. 2009. What do computational linguists need to know about linguistics? In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 41–42. https://doi.org/10.3115/1642038.1642047

Muñoz-Ortiz, Alberto, Carlos Gómez-Rodríguez, and David Vilares. 2023. Contrasting linguistic patterns in human and LLM-generated text. *arXiv preprint arXiv:2308.09067*. https://doi.org/10.21203/rs.3.rs-4077382/v1

Nirenburg, Sergei. 1986. Linguistics and artificial intelligence: An armchair drama in 3 acts, with a prologue and an epilogue. In Peter C. Bjarkman and Victor Raskin, editors, *The Real-World Linguist: Linguistic Applications in the 1980s*. ABLEX, pages 116–143.

Niu, Jingcheng, Andrew Liu, Zining Zhu, and Gerald Penn. 2024. What does the knowledge neuron thesis have to do with knowledge? In the *Twelfth International Conference on Learning Representations*.

Niu, Jingcheng, Wenjie Lu, and Gerald Penn. 2022. Does BERT rediscover a classical NLP pipeline? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3143–3153.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043.

Ogunremi, Tolulope, Dan Jurafsky, and Christopher Manning. 2023. Mini but mighty: Efficient multilingual pretraining with linguistically-informed data selection. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1251–1266. https://doi.org/10.18653/v1/2023.findings-eacl.93

Open AI. 2022. Introducing ChatGPT.

Opitz, Juri and Anette Frank. 2021. Towards a decomposable metric for explainable evaluation of text generation from AMR. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1504–1518. https://doi.org/10.18653/v1/2021.eacl-main.129

Opitz, Juri and Anette Frank. 2022. SBERT studies meaning representations: Decomposing sentence embeddings into explainable semantic features. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 625–638. https://doi.org/10.18653/v1/2022.aacl-main.48

Parcalabescu, Letitia, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280. https://doi.org/10.18653/v1/2022.acl-long.567

Parrish, Alicia, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*,

pages 2086–2105. `https://doi.org/10 .18653/v1/2022.findings-acl.165`

Pavlick, Ellie. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251):20220041. Special issue. `https://doi.org/10.1098/rsta .2022.0041`, PubMed: 37271171

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. `https://doi.org/10 .18653/v1/N18-1202`

Pimentel, Tiago, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622. `https://doi.org/10.18653/v1/2020 .acl-main.420`

Piotrowski, Michael. 2012. *Natural Language Processing for Historical Texts*. Number 17 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool / Springer. `https://doi.org/10.1007 /978-3-031-02146-6`

Plank, Barbara. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of EMNLP*, pages 10671–10682. `https://doi.org/10.18653/v1/2022 .emnlp-main.731`

Ponti, Edoardo Maria, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376. `https://doi.org/10 .18653/v1/2020.emnlp-main.185`

Prabhakaran, Vinodkumar, Christopher Homan, Lora Aroyo, Aida Mostafazadeh Davani, Alicia Parrish, Alex Taylor, Mark Diaz, Ding Wang, and Gregory Serapio-García. 2024. GRASP: A disagreement analysis framework to assess group associations in perspectives. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies (Volume 1: Long Papers), pages 3473–3492. `https:// doi.org/10.18653/v1/2024.naacl -long.190`

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152.

Qian, Kun, Ryan Shea, Yu Li, Luke Kutszik Fryer, and Zhou Yu. 2023. User adaptive language learning chatbots with a curriculum. In *Proceedings of Artificial Intelligence in Education*, pages 308–313. `https://doi.org/10.1007/978-3-031 -36336-8_48`

Qin, Tian, Naomi Saphra, and David Alvarez-Melis. 2024. Sometimes I am a tree: Data drives fragile hierarchical generalization. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*.

Raji, Inioluwa Deborah, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Rama, Taraka, Lisa Beinborn, and Steffen Eger. 2020. Probing multilingual BERT for genetic and typological signals. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1214–1228. `https://doi.org/10 .18653/v1/2020.coling-main.105`

Rambow, Owen and François Lareau, editors. 2023. *Proceedings of the Seventh International Conference on Dependency Linguistics (Depling, GURT/SyntaxFest 2023)*.

Raskin, Victor. 1985. Linguistics and Natural Language Processing. In *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, pages 268–282.

Rassin, Royi, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. 2023. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ravichander, Abhilasha, Matt Gardner, and Ana Marasovic. 2022. CONDAQA: A contrastive reading comprehension dataset

for reasoning about negation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8729–8755. `https://doi.org/10.18653/v1/2022.emnlp-main.598`

Ray Choudhury, Sagnik, Anna Rogers, and Isabelle Augenstein. 2022. Machine reading, fast and slow: When do models "understand" language? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 78–93.

Resnik, Philip and Aaron Elkiss. 2005. The Linguist's Search Engine: An overview. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 33–36. `https://doi.org/10.3115/1225753.1225762`

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144. `https://doi.org/10.1145/2939672.2939778`

Ribeiro, Marco Tulio, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912. `https://doi.org/10.18653/v1/2020.acl-main.442`

Rijkhoff, Jan, Dik Bakker, Kees Hengeveld, and Peter Kahrel. 1993. A method of language sampling. *Studies in Language*, 17(1):169–203. `https://doi.org/10.1075/sl.17.1.07rij`

Robertson, Alexander and Sharon Goldwater. 2018. Evaluating historical text normalization systems: How well do they generalize? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 720–725. `https://doi.org/10.18653/v1/N18-2113`

Rudinger, Rachel, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14. `https://doi.org/10.18653/v1/N18-2002`

Sadeddine, Zacchary, Juri Opitz, and Fabian Suchanek. 2024. A survey of meaning representations – from theory to practical utility. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2877–2892. `https://doi.org/10.18653/v1/2024.naacl-long.159`

Şahin, Gözde Gül, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020. PuzzLing machines: A challenge on learning from small data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1241–1254. `https://doi.org/10.18653/v1/2020.acl-main.115`

San, Nay, Martijn Bartelds, Tolulope Ogunremi, Alison Mount, Ruben Thompson, Michael Higgins, Roy Barker, Jane Simpson, and Dan Jurafsky. 2022. Automated speech tools for helping communities process restricted-access corpora for language revival efforts. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 41–51. `https://doi.org/10.18653/v1/2022.computel-1.6`

Saphra, Naomi, Eve Fleisig, Kyunghyun Cho, and Adam Lopez. 2024. First tragedy, then parse: History repeats itself in the new era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2310–2326. `https://doi.org/10.18653/v1/2024.naacl-long.128`

Sasidharan Nair, Sachin, Tanvi Dinkar, and Gavin Abercrombie. 2024. Exploring reproducibility of human-labelled data for code-mixed sentiment analysis. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 114–124. `https://doi.org/10.1145/3531146.3533127`

Schuff, Hendrik, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. 2022. Human interpretation of saliency-based explanation over text. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 611–636.

Scialom, Thomas, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari.

2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604. `https://doi.org/10 .18653/v1/2021.emnlp-main.529`

Shapley, Lloyd S. 1951. *Notes on the n-person Game—ii: The Value of an n-person Game*. Rand Corporation.

Shlain, Micah, Hillel Taub-Tabib, Shoval Sadde, and Yoav Goldberg. 2020. Syntactic search by example. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 17–23. `https:// doi.org/10.18653/v1/2020.acl-demos.3`

Shwartz, Vered. 2023. My fruitless endeavours with neuro-symbolic NLP. Keynote Talk at the Negative Results in NLP Workshop, co-located with the 61st Annual Meeting of the Association for Computational Linguistics.

Song, Yixiao, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634. `https://doi.org/10.18653/v1/2022 .emnlp-main.305`

Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. arXiv:2206.04615.

Starace, Giulio, Konstantinos Papakostas, Rochelle Choenni, Apostolos Panagiotopoulos, Matteo Rosati, Alina Leidinger, and Ekaterina Shutova. 2023. Probing LLMs for joint encoding of linguistic categories. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7158–7179. `https://doi.org/10.18653/v1/2023 .findings-emnlp.476`

Steedman, Mark. 2011. Romantics and revolutionaries: What theoretical and computational linguists need to know about each other. *Linguistic Issues in Language Technology*, 6. `https://doi.org /10.33011/lilt.v6i.1257`

Steen, Julius and Katja Markert. 2024. Bias in news summarization: Measures, pitfalls and corpora. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5962–5983. `https://doi.org/10 .18653/v1/2024.findings-acl.356`

Steen, Julius, Juri Opitz, Anette Frank, and Katja Markert. 2023. With a little push, NLI models can robustly and efficiently predict faithfulness. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 914–924. `https://doi.org /10.18653/v1/2023.acl-short.79`

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 3319–3328.

Sutton, Richard. 2019. The bitter lesson. *Incomplete Ideas (blog)*, 13(1).

Taboada, Maite. 2016. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2(1):325–347. `https://doi .org/10.1146/annurev-linguistics -011415-040518`

Tanzer, Garrett, 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 the *Twelfth International Conference on Learning Representations*.

Tenney, Ian, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. `https:// doi.org/10.18653/v1/P19-1452`

Tenney, Ian, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

UNESCO. 2023. A digital future for indigenous languages: Insights from the partnerships forum. Accessed: 2023/11/01.

Uszkoreit, Hans. 2009. Linguistics in computational linguistics: Observations and predictions. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 22–25. `https://doi.org/10.3115/1642038 .1642043`

van der Lee, Chris, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the

human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368. `https://doi.org/10.18653/v1/W19-8643`

van Esch, Daan, Ben Foley, and Nay San. 2019. Future directions in technological support for language documentation. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 14–22. `https://doi.org/10.33011/computel.v1i.341`

Wang, Yu, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. 2021. A comprehensive survey of grammatical error correction. *ACM Transactions on Intelligent Systems and Technology*, 12(5):65:1–65:51. `https://doi.org/10.1145/3474840`

Wang, Yupei, Renfen Hu, and Zhe Zhao. 2024. Beyond agreement: Diagnosing the rationale alignment of automated essay scoring methods based on linguistically-informed counterfactuals. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8906–8925. `https://doi.org/10.18653/v1/2024.findings-emnlp.520`

Warstadt, Alex, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34. `https://doi.org/10.18653/v1/2023.conll-babylm.1`

Warstadt, Alex, 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. `https://doi.org/10.1162/tacl_a_00321`

Wein, Shira, Christopher Homan, Lora Aroyo, and Chris Welty. 2023. Follow the leader(board) with confidence: Estimating p-values from a single test set with item and response variance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3138–3161. `https://doi.org/10.18653/v1/2023.findings-acl.196`

Wein, Shira and Juri Opitz. 2024. A survey of AMR applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6856–6875. `https://doi.org/10.18653/v1/2024.emnlp-main.390`

Wein, Shira and Nathan Schneider. 2024. Assessing the cross-linguistic utility of Abstract Meaning Representation. *Computational Linguistics*, 50(2):419–473. `https://doi.org/10.1162/coli_a_00503`

Weiss, Gail, Yoav Goldberg, and Eran Yahav. 2021. Thinking like transformers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11080–11090.

Wilcox, Ethan Gotlieb, Michael Hu, Aaron Mueller, Tal Linzen, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Ryan Cotterell, and Adina Williams. 2024. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. `https://doi.org/10.31234/osf.io/rfwgd`

Xia, Fei, William Lewis, and Lori Levin, editors. 2010. *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*.

Xie, Yuexiang, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. Factual consistency evaluation for text summarization via counterfactual estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 100–110. `https://doi.org/10.18653/v1/2021.findings-emnlp.10`

Xu, Weiwen, Huihui Zhang, Deng Cai, and Wai Lam. 2021. Dynamic semantic graph construction and reasoning for explainable multi-hop science question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1044–1056. `https://doi.org/10.18653/v1/2021.findings-acl.90`

Zhang, Kexun, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15654–15669. `https://doi.org/10.18653/v1/2024.findings-acl.925`

Zhou, Zachary, Alisha Zachariah, Devin Conathan, and Jeffery Kline. 2022. Assessing resource-performance trade-off of natural language models using data envelopment analysis. In *Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20.

https://doi.org/10.18653/v1/2022
.eval4nlp-1.2
Zuckermann, Ghil'Ad, Sigurður Vigfússon,
Manny Rayner, Neasa Ní Chiaráin,
Nedelina Ivanova, Hanieh Habibi, and
Branislav Bédi. 2021. LARA in the service
of revivalistics and documentary
linguistics: Community engagement and
endangered languages. In *Proceedings of the
4th Workshop on the Use of Computational
Methods in the Study of Endangered
Languages Volume 1 (Papers)*, pages 13–23.
https://doi.org/10.33011/computel
.v1i.953