Date: 10/21/2026

Mike Putnam – Penn State University

AI and the Mental Lexicon: A collaborative manifesto

This 'manifesto' has been written over the course of several months, based on readings, discussions with colleagues and students, and aided by consultation with Google Gemini. This document will be updated frequently to address questions and to incorporate new findings from the literature, including my own research.

At my core, I am an intellectually curious person. I am fascinated and excited to learn more about how dynamic systems emerge, operate, and change over various time scales. At the heart of these inquires is the requirement that we have an appreciation for and an ever-increasing knowledge of the underlying structural architecture and texture of the phenomena that occupy the focus of our research endeavors. Through my training as a formal linguist and my experiences in language and cognitive science throughout the course of my career, I've arrived a finite number of truisms: (1) the idea that a finite number of core, foundational operations are responsible for guiding the acquisition and production of natural language, and (2) that formal linguists have largely done a very poor job of sharing the findings and relevance of our work to colleagues in related disciplines (barring some important exceptions, of course).

My introduction to linguistics led me to syntax, mainly due to the fascinating puzzles that it presents (and continues to present). Over the course of the last decade, my research interest has shifted from the long-standing challenges of syntax to questions centering on the mental lexicon. Although there are certainly elements of syntactic inquiry that play an important role in this research program, the fundamental questions that occupy research on the mental lexicon are unique, and I might argue, of equal importance to syntax. A more nuanced understanding of the mental lexicon requires us, as linguists and language scientists, to come to terms with the tug-of-war existing between computation and storage (Embick et al. 2022), the role of modularity in language systems, and whether/how the 'internal syntax' of smaller units of language are related to larger units (among many other topics). At the heart of these investigations is the quest to define 'lexical items', aka *words*, and the space they occupy in cognition (Mansfield 2021). Without a proper understanding of the function, content, and interactions of the mental lexicon in relation to the other modules of language, arguments directed toward (the) structure-building operation(s), syntactic hierarchy vs dependencies, and the relationship between 'features' and these lexical items are incomplete and potentially built on false premises.

The field of linguistics, especially for those of us trained in any formal/generative tradition, has arrived at a critical inflection point. One of the reasons for my movement away from issues related to (strictly) formal syntax to issues centered on the mental lexicon is that I felt a lack of genuine progress in the realm of new and exciting ideas. Coupled with the decrease

in faculty positions (which simultaneously leads to a lack of interest in students in 'purely' theoretical courses of study), the decrease in external funding for linguistics-related projects, and the lack of significant collaborations between linguistics and other members of the cognitive and language science communities, I embarked on a new journey to learn more about Artificial Intelligence (AI) and Natural Language Processing (NLP) to see how/if these approaches, techniques, and tools could augment my research on the mental lexicon (more on the details of this journey as time progresses). Although it has been refreshing (and challenging!) to be a serious student once again, I have come to realize that AI issues serious challenges to 'traditional' linguists. Contrariwise, many contemporary users of AI (and unfortunately, some leaders in academia) incorrectly conflate science and technology as being one in the same. This essay is motivated by and in response to Grodzinsky's (2026) comparison of 'linguistic theories' and 'language technologies.' In plain language, he demonstrates how Large Language Models (LLMs), although efficient for technological purposes, do not come close to modeling human cognition as they relate to human language. This situation is further reinforced by Riley (2025), who cites Marcus (2022) among a host of other cognitive scientists, that maintain that AI must be able to manipulate symbols (i.e., "neorosymbolic AI") to come close to matching human intelligence. Although I find Grodzinsky's position and supporting arguments quite convincing, the question he raises in the epilogue of his book still haunts and motivates me: *shall we work together?* AI isn't going anywhere, and although it's perfectly fine to maintain a level of healthy skepticism of its usefulness and application in all spheres of the human experience, I feel like simply turning our backs to these developments and continuing to fight headwinds is a poor long-term strategy for the advancement (and survival) of our discipline.

The proposal: A 'constrained' application of AI in linguistic research

The gist of the proposal here is for a 'merger' of AI and computational musings, with the projected outcome that the former can enhance linguistic research in meaningful ways. The fundamental question, of course, is whether they can be reconciled. Over the past year, I've started a journey to become more proficient and active with AI in my research life (more on said journey in due course), with the primary objective being to not only advance my knowledge and understanding on these new technological advances, but also to find a way to promote and graft linguistics into this exciting new frontier of possibilities. The case I hope to make with this 'collaborative manifesto' is that a partnership between 'linguistics' and AI is indeed not only possible, but perhaps also necessary moving forward. With respect to formal approaches to linguistics, I wish to highlight two areas at this point that would benefit from these combined efforts: First, how can AI be used to adjudicate between different proposals about the structure of the mental lexicon? Second, how can this combined

approach potentially amalgamate research programs and models/architectures that heretofore have been regarded to be in opposition with one another?

To address the first point, I illustrate how AI + formal linguistics can be used to '(dis)prove' principal statements in contemporary linguistic theory concerning the structure of the mental lexicon. The goal of these collaborative efforts is not to outsource the heavily lifting of linguistic analysis to AI; rather, the goal is to create some sort of cognitive friction that allows us to make use of AI's predictions to highlight where (aspects of) formal models align or diverge from emerging patterns. As a point of departure, I adopt the desiderata and assumptions associated with Distributed Morphology (DM), summarizing four project proposals below that would advance, refine, and confirm theoretical positions:

1. Project 1: Probing vector spaces for 'acategorial' √roots: A key tenet of DM-desiderata is that √roots are assumed to be 'acategorial', i.e., their labels/categories are 'assigned' in the syntactic computation. AI/NLP tools can test to what extent this theoretical proposal is successful. In NLP, dense vector embeddings (like Word2Vec, FastText, or the hidden states of Transformer models) represent words are coordinates in a massive, multi-dimensional semantic space. We can use Python to isolate the vector representations of morphologically related words (e.g., *destroy, destruction, destructive*) and use vector arithmetic or Semantic Projection to see if they share a common, category-neutral vector core. The key question here is whether we can geometrically isolate the abstract root √destroy from its "functional wrappers" in the form of categorizing heads, e.g., *n, v,* and *a*. If we subtract the "noun-forming" vector space from *destruction*? Does the remaining vector space land in the exact same place as *destructive* minus its "adjective-forming" space?
2. Project 2: Testing the 'natural' class of A(djectives) & P(repositions): An important question in formal linguistics centers on the inventory of 'universal' features and categories. Do adjectives and prepositions form natural classes cross-linguistically? Once again, we could employ AI/NLP tools that measure distributional semantics to potentially test this out. The idea here would be using an open-source model's embedding space (like Hugging Face's transformers library in Python) to analyze the geometric clustering of lexical categories. This would allow us to determine whether spatial prepositions and dimensional adjectives naturally cluster to each other in terms of cosine similarity (as compared with, say, prepositions and transitive verbs). Such a simulation would allow us to see is the "mental lexicon" generated by AI naturally groups "A" and "P" as relational modifiers.
3. Project 3: The "impoverished root" vs. LLM training trajectory: Can the training of how LLMs 'learn' linguistic structure contribute to theoretical musings? Early in training, these models quickly cluster 'words' by frequency and basic syntactic function (the

"functional boxes"). Only later in this process of these models do we observe the development of more nuanced, dense semantic relationships common associated with context words. Once again, upon closer inspection this situation mirrors the theoretical distinction between functional heads and √roots in DM-parlance. One could conceivably design probing tasks to investigate if an AI's internal representation of a √root is initially "impoverished" and (largely) context-blind, only acquiring its context-dependent encyclopedic meaning as the layers deepen.

4. Project 4: Decomposing meaning via 'Linear Discriminative Learning' (LDL): Adopting a popular algorithm used by linguistics such as Harald Baayen, Linear Discriminative Learning (LDL), is a framework that employs linear algebra matrices to map visual/auditory word forms directly to semantic vectors. Continuing with (and expanding) this research program, we can build simple LDL models in Python to simulate how morphological complex words (and pseudowords) are processed. Does a model trained on word-to-meaning mapping automatically segment words into atomic √roots and affixes (which would support a few of decomposition and the texture of the mental lexicon) or does it develop a 'whole-word', discriminative typology?[1]

In the months and semesters to come, I'll augment projects such as these in more ornate detail (especially once I become more familiar and comfortable with handling vectors, matrices, and basic data iteration in Python, scaling up to libraries like scikit-learn, spacy, or pytorch to query these lexical models), but even from their initial description, I hope that the promise of this collaborative approach is evident.

As someone who was trained as a formal linguist who has recently begun to integrate AI into his research program, I, as well as others, view AI as a testing ground for theoretical boundaries that in many instances have predominantly taken place in abstract contexts. The integration of AI into formal linguistic research can contribute to important theoretical debates. For starters, this hybrid approach both compares – and potentially combines – competing approaches. Consider the contrast between Borer's (2005a,b, 2013) Exoskeletal Model (XS) and Paradigm Functional Morphology (PFM) (Stump 2001, 2016; Sims 2015). At first glance, these models fundamental contrast with one another:

[1] A point of minor concern here is involves the conflation of logit vectors (and other linear-algebra constructs) with "meaning", i.e., semantics. These sorts of "meanings" conflict with modules of meaning that represent linguistic competence. Perhaps a more realistic way of thinking about this situation is to view these as compact representations of the statistic distribution of a/the relevant token. (I thank Omer Preminger (p.c.) for bringing this point to my attention and for providing detailed feedback.)

- XS represents the ultra-radical end of Item-and-Arrangement (or structural) approaches, where syntax does the heavy lifting and words are built (exclusively) from syntactic trees
- PFM represents the 'gold standard of Word-and-Paradigm realizational morphological approaches, according to which the word represents the primary unit, and the paradigm grid itself possesses independent, top-down grammatical

So, on the surface, XS and PFM appear to be incompatible with another; however, once we look under the hood, there are aspects of their architectural goals that are shockingly compatible in at least two key traits:

1. Their shared war against 'the morpheme': Both XS and PFM abandon the traditional structuralist definition of the morpheme being a static (hence, stored) 'form-meaning pair'. Roots in XS are blank, inert units, contributing virtually nothing. The syntax is responsible for providing a 'skeleton', and the root simply fleshes this out. In PFM, inflectional markings are not simply elements that one glues onto a stem; rather, they are the result of Paradigm Functions that operate on a stem (based on a cell in a grid). Similarity: *Both frameworks advocate for the position that meaning emerges from systemic properties, not from individual pieces*. Both frameworks attempt to eliminate the "lexical item" as a sovereign, self-contained ontological unit of language.
2. Functors as "realizational" blocks: According to XS, functional items (e.g., tense, aspect, grammatical number, etc.) do not carry lexical meaning; they are open values ([±Past], [±Plural]) that direct the structural interpretations (NB: Nanosyntax appears to take this even further, mapping individual features to specific nodes in a syntactic tree.) This position is incredibly close to PFM's conceptualization of Realizational Rules, whereby a morphological stem and a set of morphological features are fed into a function, outputting the realized word form. *(Potential) similarity*: *The universal functional hierarchy present in XS could potentially be interpreted as a hierarchical/syntactic version of the feature bundles that PFM use to organize its paradigm cells*.

Combined AI/computation musings have the possibly of uniting and reconciling these two distinct approaches. This would require envisioning these two approaches as two distinct levels of description of the same mental phenomenon:

- XS (and potentially also Nanosyntax): This level describes the *online derivation* – the underlying universal structural relations that restrict what can be expressed

- PFM: This level describes the *offline organization* – how the brain clusters, balances, and optimizes those outputs into predictable, systemic networks (paradigms) to save on processing and memory costs.

Viewing these issues through the lens of an AI agent, all that we're left with is an optimization problem: *How does the system map abstract conceptual/functional intent onto a phonological output exerting the least amount of cognitive effort?* When we use neural networks to stimulate lexical access, the boundaries between syntactic trees and paradigmatic grids becomes quite blurry.

Interfacing with neuro- and psycholinguistic research

One major hurdle in the field of neurolinguistics involves the mapping of measured brain activation to abstract concepts and constructs equated to pieces of syntax (i.e., 'phases' or 'chunks') or morphological information (such as information in a paradigmatic cell).

We have collectively moved on from the archaic view of the mental lexicon being nothing more than a static storage space. Rather, most embrace the idea of the mental lexicon being a live network. AI interfaces well with this interpretation, and can enhance this research in several ways:

1. Transition from 'static lists' to 'high-dimension spaces': Traditional psycholinguistic research seeks to measure relationships between words using priming tasks (e.g. how fast do you recognize 'doctor' after hearing 'nurse' or 'hospital'?). Employing AI (via LLMs) to make use of vector embeddings to map words into thousands of geometric dimensions simultaneously provides a more expansive measure and domain of application. This takes on even more significance in attempts to model how multilingual brains shift, bleed, and separate inter-related language networks in real-time.
2. Simulating lexical access and competition: When individuals attempt to retrieve words from their mental lexicon, these elements compete with hundreds of phonological and semantic 'neighbors' for activation. For these tasks, we can use AI to build reinforcement learning environments (or neural decoders) to simulate the exact cognitive cost of lexical selection under a variety of conditions, such as (but not limited to) language-mixing environments, phonological relatedness, etc.
3. Testing real-time structural assembly: If we embrace a non-lexicalist approach to the mental lexicon (as some linguists do), components of lexical items are associated with phonology and semantics. What AI offers us in this respect is a 'sandbox' to see if an AI agent, if forced to optimize for performance-based constraints (e.g., working memory limitations, high communicative clarity, etc.) display patterns (including particular lexical quirks and irregular and defective paradigms) that human language also shows.

It's also possible to use AI has as a computational middle tier to strengthen connections between generative proposals and cognitive neuroscience. Here I highlight three sample projects, once again returning to our previous discussion of XS and PFM approaches to morphology. As a point of departure, let's assume that these two structural models represent the optimal balance between memory and computation:

1. Generative quantifiable "representational spaces" for neuroimaging: One of the major hurdles in contemporary neurolinguistic research concerns the mapping of brain activity to abstract concepts like "syntactic structures" or "paradigm cells". To overcome such hurdles, we could conceivably train an NLP or neural network to process morphology using the structural constraints we've discussed above, enabling us to extract the exact mathematical vectors (i.e., hidden layers of activation) the AI agent uses at different stages of word assembly. As a case in point, researchers could then use Representational Similarity Analysis (RSA) to compare the AI agent's internal geometric space directly against fMRI or EEG data from human subjects performing the same task. If the brain's neural activity patterns match the geometric clustering of an AI agent using a PFM-style network optimization, we would have physical evidence of that paradigmatic architecture (or something close to it) in the human cortex.
2. Simulating the temporal dynamics of morphological processing (EEG/MEG): A classic debate in morphological processing is *when* exactly does decomposition take place. Do we access the full word initially and then break it down further, or do we parse the pieces instantly, e.g., the traditional dual-route vs. single-route debate? We could train an AI agent via Reinforcement Learning or a sequential recurrent network that operates step-by-step. This enables linguists to measure its computational "effort" (i.e., surprisal or gradient descent updates) at every single syllable, feature insertion, or other structural benchmark we wish to include. Translating this into a tangible experiment, this computational effort can be directly mapped onto the temporal waves of human EEG/MEG data. For example, we can test if the AI agent's difficulty spikes mirror the human N400 (semantic/lexical access) or LAN/P600 (syntactic/structural assembly) components during instances of language-mixing or the processing of compound lexical items.
3. A causal "sandbox" for predictable psycholinguistic deficits: In behavioral psycholinguistics, scientists often look at how the grammar system 'breaks down' under increased cognitive load, in (healthy) cognitive aging, or in clinical population (e.g., agrammatical aphasia). Obviously, it's unethical (and thankfully illegal!) to intentionally damage a human brain to test these sorts of theories. If we train an AI framework of the lexicon that is based on specific trade-offs (like memory vs clarity

reward functions), we could hypothetically "break" the AI model in precise ways, for example, we could introduce noise in vector spaces, throttle its memory capacity, or manipulate and "fatigue" underlying syntactic and morphological representations as an attempt to simulate aphasia or cognitive fatigue. Behavioral simulations of this "damaged" AI could be compared with naturalistic data to see if these errors predict similar syncretisms, speech errors, or language-mixing breakdowns observed in experiments with real people. Similarities in these data comparisons would prove that such computational models of the mental lexicon could be considered causally accurate.

a. Simulating aphasic breakdowns via phonological neighborhood density and L_RFG: Under a Lexical-Relational Functional Grammar architecture (L_RFG; Asudeh et al. 2020), lexical access is not merely viewed as the retrieval of lexical items from a static list, but rather as a dynamic, constraint-satisfaction negotiation across a high-dimensional space where **phonological neighborhood density** plays a vital role. Python and PyTorch can manipulate vector spaces and activation thresholds based on neighborhood density metrics when we attempt to model speech errors commonly attributed to aphasic speech (e.g., phonological paraphasias or substitution errors in agrammatical aphasia). By introducing (some degree of) targeted "noise" or throttling activity capacity in models directed and governed by L_RFG principles and desiderata, we can simulate how an impaired system handles lexical competition. Returning to the metaphor of the computational "sandbox", this allows us to observe whether the model predicts the exact patterns of word-retrieval difficulties and error types observed in clinical populations, turning abstract neurolinguistic hypotheses into testable simulations.

Translational impact

Since its modern inception, linguistics as an academic discipline has been quite diverse and adaptable. It continues to not only maintain a level of relevance across most existing disciplines in the humanities and social sciences to varying degrees, but it has also come to play increasingly more pronounced roles in the hard sciences and technological fields. Homing in on linguistics as an academic discipline, a fair and pragmatic concern that prospective students (and their parents) face is the value of pursuing linguistics as a degree. These concerns are concrete and understandable, especially in the face of rising costs of advanced degrees and future career prospects. The collaborative approach I have outlined above is quite flexible and can easily be adapted to industry needs. Here I sketch out three potential applications of this research program with an eye towards translation impact:

Date: 10/21/2026

Mike Putnam – Penn State University

1. Language acquisition: Predictive roadmaps for EdTech (Duolingo): Language-learning apps such as Duolingo rely heavily on adaptive algorithms to "teach" vocabulary and grammar. However, their backends are often driven by generic spaced-repetition algorithms rather than deep cognitive-linguistic models of how a bi/multilingual brain actually manages competition. The translational component of this research would attempt to model something akin to Vocabulary Insertion as a direct competition based on feature bundles and some metric of 'cognitive optimization' (López 2020, Lohndal & Putnam 2024). The potential impact of this research is significant; EdTech companies can use these sorts of ideas and architectures to more accurately predict exactly which morphological forms will likely cause the highest cognitive load or cross-linguistic interference for a learner. Instead of waiting for a student to fail a quiz on irregular verbs or compounding, Duolingo's AI component could use models such as these to preemptively structure a personalized curriculum.
2. Clinic diagnostics: A 'stress test' for language disorders: Currently, diagnosing specific language impairments (SLI) or variations of aphasia relies heavily on behavioral observation and standardized testing. These methods often catch the *symptoms* of a deficit, but not the underlying architectural challenges and failures. The translational component here would be to develop an AI model that simulates that exact trade-off between memory (paradigms) and structural computation (again, think of DM or XS), this can be used as a diagnostic baseline. Once again, the impact of this research would be significant; by running a patient's specific speech errors or language-mixing patterns through the model, clinicians can pinpoint exactly where the system is breaking down. Is it a failure of feature deletion (Impoverishment)? A failure in structural assembly (the syntax)? Or a degradation of the underlying vector space (lexical access)? This moves speech pathology from descriptive diagnostics to precise, predictive modeling.
3. Next-gen tech (Google, Apple): Fixing the "hallucination" and efficiency bottlenecks: Finally, this hybrid research program could benefit tech-giants like Google, who are currently pouring billions into LLMs. A major ecological concern here is that these massive LLMs require unfathomable amounts of data and energy because they process language primarily (if not exclusively!), completing missing the elegant, rule-governed efficiency of human morphology. As highlighted above, humans don't need or require a trillion parameters to learn language, because our mental lexicon(s) use(s) highly optimized architectures and paradigmatic defaults to compress data. The combined approach sketched out here demonstrates how a neural network can successfully implement the constraints of formalisms such as DM, XS, or PFM, linguistically-constrained models of AI provide a blueprint for **neuromorphic and**

ultra-efficient NLP models. These contribute to the call for smaller and smarter models that can run locally on a smartphone, embracing the goal of working toward the goal of developing tools that operate on human-like data efficiency.

Moving from the 'black box' to the 'white box'

Critics of formal approaches to linguistics often point to the infamous notion of the 'black box' responsible for language acquisition. This caricature of formal linguistics harkens back to initial proposals regarding Universal Grammar (Chomsky 1965), and the image of the black box is one in which (i) data goes in, (ii) a result comes out, an (iii) the internal reasoning is/can be quite opaque. The combined research program outlined in this brief manifesto, even though only in its infancy, allows us, as linguists and language scientists, to move to a 'white box' approach, i.e., with a transparent, explainable model that is fundamentally grounded. At this point, two factors come to mind that are mutually beneficial for both linguists and those working in the AI industry: First, the white-box model outlined here makes AI-supported research falsifiable (according to research questions that guide and direct the experiments). Through this we avoid the "large language models"-mistake, i.e., the tendency to treat LLMs as static knowledge bases. Second, AI moves from a "super-predictive toy" into a scientific instrument, where its outputs become "cognitive mirrors", which force us to clarify and stress-test theoretical assumptions about the mental lexicon.

For linguists, the patterns and idiosyncrasies found in natural languages have been an artifact of fascination and amazement for centuries, even millennia. Nothing has changed in this respect, in fact, my hope is that anyone who has read this 'manifesto' to the end shares my optimism that linguistics stands to play a decisive role in these exciting developments. Make no mistake, I'm not here to ride the AI-hype train, but to point out the obvious: As these technological advances continue to permeate all sectors of education and modern society, a collective failure on our part to figure out where the 'human in the loop' resides in this unfolding drama could be cataclysmic for our field. The 'human in the loop' is indeed *the linguist*. Formal research can contribute to these developments immensely and in vital and important ways, however, it is incumbent upon us to embrace these exciting opportunities to use the tools that will ultimately define the infrastructure of the next generation of linguistics and related work in language science.

References

Asudeh, A., P.B. Melchin, & D. Siddiqi. 2020. *DM + LFG = L_RFG*. Presentation handout. Department of Linguistics colloquium – University of Arizona. Tuscon, AZ. December 4.

Borer, H. 2005a. *Structuring sense – Vol 1.: In name only*. Oxford: OUP.

Borer, H. 2005b. *Structuring sense – Vol 2.: The normal course of events*. Oxford: OUP.

Borer, H. 2013. *Structuring sense – Vol. 3.: Taking form*. Oxford: OUP.

Chomsky, N. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Embick, D., A. Creemers, & A. J. Goodwin Davies. 2022. Morphology and the mental lexicon: Three questions about decomposition. In A. Papafragou, J.C. Trueswell, & L.R. Gleitman (eds.), *The Oxford handbook of the mental lexicon* (pp. 77-97). Oxford: OUP.

Grodzinksky, Y. 2026. *How deeply human is language? Chomsky, the brain, and the AI fantasy*. Cambridge, MA: MIT Press.

Lohndal, T., & M.T. Putnam. 2024. The importance of features and exponents: Dissolving Feature Reassembly. *Linguistic Approaches to Bilingualism* 14(1): 1-36.

López, L. 2020. *Bilingual grammar: Toward an integrated model*. Cambridge: CUP.

Marcus, G. 2022. Deep learning is hitting a wall: What would it take for artificial intelligence to make real progress? *Nautilus* https://nautil.us/deep-learning-is-hitting-a-wall-238440 (accessed June 29, 2026)

Mansfield, J. 2021. The word as a unit of internal predictability. *Linguistics* 59(6): 1427-1472.

Riley, B. 2025. Large language mistake: Current AI models are not on the path to artificial general intelligence. *Cognitive Resonance*.

https://buildcognitiveresonance.substack.com/p/large-language-mistake?utm_source=share&utm_medium=android&r=5dlant&triedRedirect=true

(accessed on June 29, 2026).

Sims, A. 2015. *Inflectional defectiveness*. Cambridge: CUP.

Stump, G. 2001. *Inflectional morphology: A theory of paradigm structure*. Cambridge: CUP.

Stump, G. 2016. *Inflectional paradigms: Content and form at the syntax-morphology interface*. Cambridge: CUP.