

What is Knowledge of Grammar?

From Theory to Models and Back

Cem Bozşahin

Boğaziçi Linguistics
LING488
November 23, 2025

- Current LLMs **finitize** a grammar.
- What does that mean?
- A modern LLM: **fixed** (but VERY large) length of:
 - word vectors, input encoding, output decoding, position encoding/decoding, input effect remembrance (attention)
- From ANNs (no retention), RNNs (50 epochs), LSTMs (1000), to encoders-decoders, transformers (layers), attention,...
- All are fixed in a deep and multi-layered NN architecture to facilitate standardized end-to-end modeling and training.
- Restricting dependencies to a finite window (moving window or fixed window) also enables good semantic guesses.

Good enough for a linguist?

- A thought experiment: Indefinite-length mini-English from 4 orthographic words: *I, you, think, like*
 - That's at least 7 words to a morphologist!
 - I like you.*
 - You think I like you.*
 - I think you think I like you.*
 - *I like you like I like you.*
 - *I think you like I think you.*
 - I think you think I think you think I think I like you.*
 - *I think you think I think you think I think I like you think I like you.*
- It seems impossible to finitize the dependencies here. Alas, we can keep going.
- Size does not matter all that much, but what allows and disallows this embedding does. What is about **think** and **like** that does this?

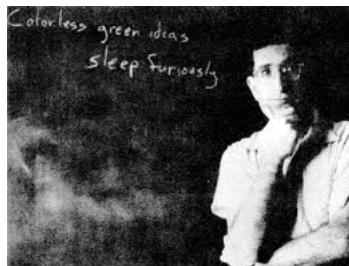
What do we expect from a model?

- Language technology: good, robust, reliable, responsible and accessible **performance**.
- Linguistics: Putting an idea or theory about language to test.
 - Ever since antiquity, that usually means an idea or theory about **grammar**.
- Is it a coincidence that LLMs achieved such levels of performance by rising above the simplistic associationism of ANNs, adding layers and layers of **abstraction**?
 - much like a program or grammar

What is Knowledge of Grammar?

- The answer depends on what we expect from grammar.
 - Minimally, we want it to decide grammaticality (i.e. syntactic configuration).
 - Maximally, we would want it to cover everything under the sun about knowledge of language.
- Is this a matter of personal choice?
 - Every child exposed to NL data in the critical period acquires a natural language.
 - There are systematic exclusions if she is not.
 - Exclusions from what? Where do we show that?
 - Every native speaker has a sense of meaningfulness of an expression which is consequent to it being grammatical.

Colorless green ideas sleep furiously.



- NL grammars are representational spaces. Is this space limited?
- What is a linguistic category? Label? abstract, concrete?
- Anderson (1976): Data from ergativity and mixed systems show that the category space cannot be universally captured morphologically or surface-categorially (e.g. N, V, A, P).
- Categories in a grammar can be abstract objects (see Katz 1985 for a bit of philosophy of linguistics).

Some grammars as conceived by typologists and field linguists:

Ma Manda (Papua New Guinea, Papuan)

Mongsen Ao (Northeastern India, Tibeto-Burman)

North Paiwan (Taiwan, Austronesian)

Two related questions:

- Can we make human grammars (their proxies) determine grammaticality without external means?
- Or do we need extra means? (minimal links, mapping theories for multi-structures, multiple independent computations, movement management, subjacency, synch conditions, reanalysis, recomputation etc.)

The answer is critical: The more intermediaries there are, the more difficult it is to go from theory of grammar to models of grammar.

Scientific modellability is not a personal choice; but, perhaps modeling is.

One consequent question:

Is it enough to determine grammaticality?
(syntactic well-formedness)

The answer will determine what goes in to any NL grammar.

What kind of meaning must enter grammar?

Colorless green ideas sleep furiously.

- Wittgenstein
- Chomsky
- Leśniewski (1929)
- W: Language games can make it meaningful.
- C: It is already meaningful, because it is grammatical. What it lacks is sense.
- L: Categories are semantic in origin; no such thing as meaningless category.

- 1 If something is grammatical and senseful, we can think of world models for meaning.
 - Ungrammatical ones are not meaningful to begin with:
 - a. *Green sleep colorless furiously ideas
 - b. *Hiç okumadığı Mehmet kitapları çok Ahmet'in seviyor Turkish
 - This was (and is) a critique of studying meaning alone, or just probabilistically.
- 2 Where is the feel of meaningfulness coming from for senseless expressions?

- Categorical Grammarians have been drawing attention to a very striking asymmetry for **far more than a century**:¹
- Parts of a clause may require different categories,

But one uniquely determines the clausal structure: **The Verb**

If the verb can determine clause's syntactic structure, it can also determine what makes it meaningful.

- Maybe we can put **grammatical categories**, which are grammar's workhorses, to work on BOTH aspects.

¹Husserl (1900); Sapir (1921); Ajdukiewicz (1935); Montague (1973); Schmerling (2018); Bozsahin (2025)

- Expressions with sense have intuitive world models (decision models, truth-conditional semantics).
- Expressions with no sense may have counter-intuitive but possible world models (possible world semantics).
- But these do not explain the feel of meaningfulness; they eschew the role of **category choice** in grammaticality:

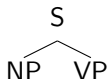
choosing to refer to another event to the extent of affecting grammaticality.

- (1) a. She played the piano for an hour/*in an hour. play
b. She played the sonata *for an hour/in an hour. perform
c. She played the sonata for a year. multiple perf.
d. She played the piano in a year. mastery
e. **She played the piano in an hour.** Genius?

From distributional categories to abstract realities

$$\begin{aligned} S &\rightarrow NP \ VP \\ VP &\rightarrow V_{tv} \ NP \\ V_{tv} &\rightarrow \text{played} \end{aligned}$$

$S \rightarrow NP \ VP$
 $VP \rightarrow V_{tv} \ NP$
 $V_{tv} \rightarrow \text{played}$



Looking top-down, they appear to be relations.
Looking bottom-up, we can see that they are actually **FUNCTIONS**.

This is the first step toward understanding the verb's power.

$$VP = S \backslash NP$$

$$NP = S / VP$$

$$V_{tv} = VP / NP = (S \backslash NP) / NP$$

Focusing on the verb, we get for example:

$$\text{played} = v_{tv} = (S \backslash NP) / NP : \lambda x \lambda y. \text{play}'xy$$

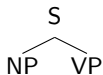
These are structural functions. They determine the tree on the left.

But now we have BOTH syntax and semantics IN A CATEGORY, because they are functionally determined.

$$S \rightarrow NP \ VP$$

$$VP \rightarrow V_{tv} \ NP \quad VP = S \backslash NP$$

$$V_{tv} \rightarrow \text{played} \quad NP = S / VP$$

$$V_{tv} = VP / NP = (S \backslash NP) / NP$$


If the subject is ALSO unique in the main clause, why not start with the subject?

$$NP = S / VP = S / (S \backslash NP)$$

$$\text{Ayşe} = S / (S \backslash NP) : \lambda p. p \text{ ayşe'}$$

Now the predicate is OPAQUE (p). On what basis are we going to choose a category?

This asymmetry is the somewhat neglected discovery of Richard Montague.

- A clause can take many different kinds of constituents with varying categories.
- But one determines what it means to be a clause: The verb
- Since the verb is a predicative element, it can determine the predicate-argument structure (PAS) of the clause.

PAS: Reference Choice for events: Placeholders *then* fill in truth-conditional meanings.

- (2) a. $\text{played} = (S \backslash NP) / NP : \lambda x \lambda y . \text{play}' xy$
b. $\text{played} = (S \backslash NP) / PP_{\text{schdl}} / NP_{\text{score}} : \lambda y \lambda x \lambda z . \text{perform}' (iter' x (play' yz)) z$

- Without that kind of meaning, it is difficult with categories to study grammaticality AND the consequent sense of meaningfulness.

What must syntactic categories do?

- Assemble the PAS as syntactic structure is built.
- Syntax is still autonomous. But now it carries a baggage. The treasure is in the baggage.
- If we don't transmit both structures as we analyze, we would need independent solutions to grammaticality and meaningfulness.
- THAT is not very congenial to modeling. Real data shows why.

Syntactic decomposition does not necessarily mean semantic decomposition. It is verb-controlled.

- (3) a. Zhāngsān shēng qì le
Zhangsan generate air asp
'Zhangsan got angry.'

Chinese; Kao 2024:1

- b. Zhāngsān shēng le hǎodà de qì
Zhangsan generate asp huge nom air
'Zhangsan got very angry.' (lit. 'Zhangsan generated huge air.')

- c. Zhāngsān shēng wán qì le
Zhangsan generate finish air asp
'Zhangsan stop being angry.' (lit. 'Zhangsan finished generating air.')

Semantic idiomaticity does not mean syntactic inertness. That too is verb-sense-controlled.

(4) a. Wǒ yōu le Zhāngsān yí mò
I ASP Z one
'I teased Zhāngsān.'

b. *Wǒ yōu (le) yí mò Zhāngsān

c. *Wǒ yōumò (le) Zhāngsān

To mean 'tease', there has to be an object in between (a–b), light or heavy. Together the verb is not transitive (c).

When we make a category choice, i.e. choice of event reference, we also lay out under what conditions we see grammaticality.

(5) a. She played the piano for an hour/*in an hour.

b. She played the sonata *for an hour/in an hour.

c. She played the sonata for a year.

d. She played the piano in a year.

e. She played the piano in an hour.

(6) a–b. played = $(S \backslash NP) / NP : \lambda x \lambda y. \text{play}' xy$

c. played =

$(S \backslash NP) / PP_{\text{schdl}} / NP_{\text{score}} : \lambda y \lambda x \lambda z. \text{perform}'(\text{iter}'x(\text{play}'yz))z$

d–e. played =

$(S \backslash NP) / PP_{\text{duration}} / NP_{\text{tool}} : \lambda y \lambda x \lambda z. \text{practice}'(\text{iter}'x(\text{play}'yz))z$

Case (e) is not a whole lot different than searching for meaningfulness for the Chomsky example:

Colorless green ideas sleep furiously.

Roman Jakobson's take: "If someone's hatred never slept, why then, cannot someone's ideas fall into sleep?"

sleep = $S \backslash NP : \lambda x. \text{torpid}' x$

Yue Ren Chao's take:

sleep furiously = $S \backslash NP : \lambda x. \text{sleep-on-idea}' x$

- If syntax-semantics is so dependent on the verb,
and verbs tend to be very choosy about the roles of arguments,
- can we address GENERAL problems of syntax-semantics with this way of thinking?

I suggested in (2025) that we can, for example, CASE:

- (7) a. Mary would *to run/*runs/run.
b. Mary wants *run/*runs/to run.
c. Mary *run/*to run/runs.

Grammaticality vs. number of readings

- (8) a. Her çocuk araba-ya bin-di. Turkish
every child-DAT board-PAST
'Every child got in the car.'
b. Araba-ya her çocuk bin-di.

How many cars?

How many in Arabaya her çocuk bindi?

Should grammar say something about the number of readings?

- If we want our grammars to take care of grammaticality and the consequent sense of meaningfulness, we must start with the verb.
- That's where the choices of event reference and meaning are.
- Leaving all that to implicature, pragmatics or discourse would not explain narrow behavior.
- Sandra (1998) warned us about the two extremes (one lexical entry serves all, all senses served by different entries).
- Without narrow behavior, scientific modeling is hopeless.
- Without an independently replicable path to go from theory of grammar to models of grammar, a 'theory' would not be a natural science theory.

The knowledge of grammar must be representable.

- Category choice is an intentional (and intensional) act.
- Only subjects do that; models or computers don't.
- Sapir 1949:17–18 called it *shared psychological spaces*.
- He didn't call it *shared psychological states*.
- We are talking about social semiotics, not psychology.

- We may not be at the mercy of our psychological or biological make-up for categories. At least Schopenhauer (1819); Ryle (1937) didn't think so.
- To understand the role of choice in **linguistic analysis**, we must worry about how and where the relevant knowledge goes in the grammar.
- Unfortunately, it does not mean we can **express** all that can be imagined.
 - There are GAPS in the linguistic data.
 - That is why linguistics is a natural science.
- It is hard to study common timeframe of language acquisition without such gaps.

- There seems to be **invariants** in the structuring of grammaticality and the consequent sense of meaningfulness (e.g. compositionality).
- These invariants may spell the landscape of the **variants**.
 - Cross-linguistic and intra-linguistic typology.
 - We may discover new ideas to go in to any grammar to explore these aspects.

- Mathematics is the time-honored study of invariants.
- If we do not constrain the space of the variants, we would be in the dark about the sufficiency of supposedly necessary mechanisms.

In short:

There seems to be a limit about the space of possible human linguistic categories. Understanding the limit means understanding ourselves.

- Not quite minimally, all and only the knowledge that affects grammaticality and the consequent sense of meaningfulness (not necessarily sensefulness), enters any grammar.
- That doesn't sound to me like everything under the sun.
- Natural Grammars must be modelable, as a consequence of a theory of grammar.
- A scientific model prepares a theory or an idea for experiments.
- For that we need an explicit nomenclature and modeling vocabulary.
 - Either we make everything under the sun testable this way,
 - Or we narrow the knowledge scope of possible NL grammars so that we can do this.

Teşekkürler

Thank you

- Ajdukiewicz, K. (1935). Die syntaktische konnexität. In S. McCall (Ed.), *Polish Logic 1920-1939*, pp. 207–231. Oxford: Oxford University Press. Translated from *Studia Philosophica*, 1, 1-27.
- Anderson, S. R. (1976). On the notion of subject in ergative languages. In C. Li (Ed.), *Subject and Topic*, pp. 1–23. Academic Press.
- Bozşahin, C. (2025). *Connecting Social Semiotics, Grammaticality and Meaningfulness: The Verb*. Newcastle upon Tyne: Cambridge Scholars.
- Husserl, E. (1900). *Logical Investigations*. New York: Humanities Press. 1970 trans. by J. N. Findlay [Original German edition, 1900-1901.].
- Kao, T.-C. (2024). *Word Internal Structure in Chinese: Event Structure, Predicate-Argument Structure and Categories in Separable Verbs*. Ph. D. thesis, Middle East Technical University, Ankara, Türkiye.
- Katz, J. J. (Ed.) (1985). *The Philosophy of Linguistics*. Oxford University Press.
- Leśniewski, S. (1929). Grundzüge eines neuen Systems der Grundlagen der Mathematik [Fundamentals of a new system of basic mathematics]. *Fundamenta Mathematicae* 14, 13–67. Warsaw.
- Montague, R. (1973). The proper treatment of quantification in ordinary English. In J. Hintikka and P. Suppes (Eds.), *Approaches to Natural Language*. Dordrecht: D. Reidel.
- Ryle, G. (1937). Categories. *Proceedings of the Aristotelian Society*. reprinted in ?.
- Sandra, D. (1998). What linguists can and can't tell about the human mind: A reply to Croft. *Cognitive Linguistics* 9, 361–378.
- Sapir, E. (1921). *Language*. New York: Harcourt Brace and Co.

- Sapir, E. (1933/1949). Language. In D. G. Mandelbaum (Ed.), *Selected Writings of Edward Sapir in Language, Culture, and Personality*. Berkeley: University of California Press. Originally published in *Encyclopedia of the Social Sciences* 9: (1933) 155–169, New York: Macmillan.
- Schmerling, S. (2018). *Sound and Grammar: a Neo-Sapirian Theory of Language*. Leiden/Boston: Brill.
- Schopenhauer, A. (1819). *Die Welt als Wille und Vorstellung [The World as Will and Representation]*. Leipzig: Brockhaus.