

# SEMICOMPOSITIONALITY

András Kornai  
SZTAKI Computer Science Research Institute  
and  
Dept of Algebra, Budapest Institute of Technology

NASSLLI, June 23 2025

# OUTLINE

- 1 WHAT IS SEMICOMPOSITIONALITY AND WHY DO WE CARE?
- 2 THE ALGEBRA PART
- 3 PRIMITIVES

# WHO THIS COURSE IS FOR

- 1. Linguists, particularly those interested in AI, KR, NLP  
Semanticists, both “mainstream” and “cognitive”
- The mainstream people, particularly those coming from the MG tradition, will like the formal rigor but possibly hate the data
- Those coming from the cognitive side (Wierzbicka, Fauconnier, Langacker, Talmy, Jackendoff, etc.) will be more sympathetic to the data, but they tend to hate the formal rigor
- Morphologists, lexicographers. We will discuss a fair number of English suffixes, in particular *-er -er -est -est -ing -ist -ize -th -th*. → Wednesday: prefixes, latinate morphology
- 2. Computer science/logic/math people interested in natural lg
- Course perhaps less useful for cognitive neuroscience/psychology people
- Course website at <https://kornai.com> → Teaching → 2025 → NASSLLI

# TODAY

- Prerequisites discussion
- Primitives: what are they
- What do bound morphemes like *-th* mean?
- Claim defended here: they are not any different from free morphemes. The mainstream syntactic view, e.g. (Harley, 2014) distinguishes *roots* from *features* but we see no semantic reason for doing so
- How do we represent lexical knowledge?
- Graphs, automata

# PREREQUISITES DISCUSSION

Who has a background that includes a first (and/or second) course in

- Linguistics
- Syntax, semantics, morphology, phonology, lexicography
- Computer science
- Programming, graphs, automata
- Statistics/probability theory/information theory
- Algebra
- Linear algebra, multilinear algebra, universal algebra, category theory
- English (native), other languages **DLD**

# WHAT IS SEMICOMPOSITIONALITY?

- There is a large body of **compositional** phenomena: most of syntax, inflectional morphology (and also derivational, esp. if sufficiently productive)
- There is an even larger body of **noncompositional** phenomena: root morphemes
- The distribution is bimodal: most phenomena are either fully compositional or fully arbitrary
- But there remain an irritating 15% we can't just put in the lexicon (it's not arbitrary/accidental, there is *grammar* there) and we can't quite derive either (the grammar leaks)
- Examples: Semitic roots and binyanim; phrasal verbs (English, German, Hungarian); multiword expressions (MWEs) like *attractive nuisance*
- They lack full analogical proportions bemegy:kimegy  $\neq$  berúg:kirúg  $\neq$  belát:kilát  $\neq$  betalál:kitalál

# WHAT WOULD A FULLY COMPOSITIONAL ANALYSIS LOOK LIKE?

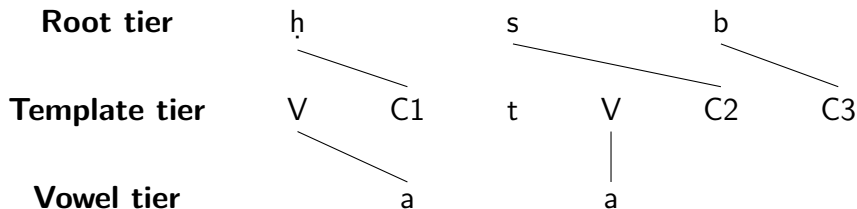
- There are some primitives like *be* 'in'; *ki* 'out'; *rúg* 'kick'; *lát* 'see'; *talál* 'find'
- These give relevant readings such as *kirúg* 'kick out' which can apply to the direct physical action *in soccer, it is normally the goalie who kicks the ball out*
- But these direct physical actions are often overshadowed (in terms of usage frequency) by indirect ones *kirúg* 'get rid of' *the girl kicked out her boyfriend*; 'fire, lay off' *a cég kirúgta* 'the company fired him/her'
- The fully compositional readings must maintain the analogical proportions:  $ki:be::kiX:beX$
- But the semicompositional ones don't. We are in search of a mechanism that is capable of doing this

# ARABIC BINYANIM

	Meaning	Example
I	Basic action	<i>ʔakala</i> (he ate)
II	Causative / Intensive	<i>ʔakkala</i> (he fed)
III	Reciprocal / Associative	<i>ʔākala</i> (he dined with)
IV	Causative (transitive)	<i>ʔaḥfaẓa</i> (he made [someone] keep)
V	Reflexive of II	<i>taʔakkala</i> (he was fed)
VI	Reciprocal of III	<i>taḥāfaẓa</i> (they kept each other)
VII	Passive (inchoative / middle)	<i>inḥafaẓa</i> (it was kept)
VIII	Reflexive / medial modification	<i>iḥtafaẓa</i> (he held onto)
IX	Inchoative for colors or bodily defects	<i>iḥmarra</i> (he became red)
X	Request / seeking / intensive	<i>istaḥfaẓa</i> (he sought protection)



# BINYAN VIII

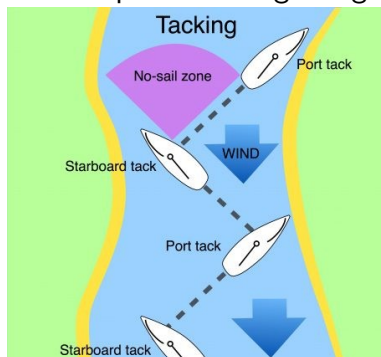


*ḥsb* 'reckon' *iḥtasaba* 'consider'

I	gloss	VIII	gloss	semantic shift
<i>ʕalima</i>	knew	<i>iʕtaʕalama</i>	learned	reflexive (acquired
<i>fahima</i>	understood	<i>iftaḥama</i>	grasped deeply	intensification / ref
<i>jahada</i>	strove	<i>ijtahada</i>	exerted himself	intensification / ref
<i>ḥasiba</i>	reckoned	<i>iḥtasaba</i>	considered	internal/cognitive
<i>kataba</i>	wrote	<i>iktataba</i>	corresponded	reciprocal / middle
<i>ʕadala</i>	was just	<i>iʕtadala</i>	moderated hself	inchoative / reflexiv
<i>ʕatada</i>	prepared	<i>iʕtadda</i>	got ready	reflexive
<i>nadima</i>	regretted	<i>indama</i>	grieved deeply	emotional intensific
<i>ṣabara</i>	was patient	<i>iṣṭabara</i>	persevered	intensification
<i>ṣalaḥa</i>	was sound	<i>iṣṭalaḥa</i>	reconciled	reciprocal

# KEY GOALS

- Do away with ‘taming by naming’: just because we call something *inchoative* or *reciprocal* we haven’t said what it means
- Argue against root/feature distinction
- Offer a technology for the grammarian to deal with the entire compositionality spectrum
- Teach some techniques that are useful beyond morphology
- Offer hope for finding the grammar in the LLM (blackbox AI)



# HOW MUCH ALGEBRA IS REQUIRED?

- This course stays with an abstract/universal algebraic outlook
- There are specific algebraic structures sometimes proposed, such as Hopf Algebras (Marcolli, Chomsky, and Berwick, 2023; Marcolli, Berwick, and Chomsky, 2023a; Marcolli, Berwick, and Chomsky, 2023b; Chomsky et al., 2023) and Frobenius Algebras (Coecke, Sadrzadeh, and Clark, 2010; Kartsaklis, 2014)
- There is an even higher, category theoretical view
- Here we stay at the lower level, with finite state devices in focus
- This is because we want to know how this relates to grammar: what are the primitives, how the rules manipulating them can be formulated, etc.
- Interest with concrete linguistic entities such as words and grammar rules is paramount

# OTHER RELEVANT EXAMPLES

- Much of derivational morphology
- Some of inflectional morphology
- Huge lexical classes like phrasal verbs, named entities
- Good part of this is deferred until Wednesday
- How prevalent is semicompositionality?
- Quite prevalent. Tomorrow we will discuss how to quantify this answer **Hint: we will use information theory**

# A FIRST GLIMPSE

- Compositional: direct product (works well with analogical proportions)
- Semicompositional: subdirect product (must be used when the analogies fail)
- Noncompositional: memorized 'lexicalized'
- Partee's Problem (Partee, 1979; Partee, 2013)
- How big is (linguistic) LTM? Less than a megabyte
- Tomorrow we will see how to estimate **Hint: we will use information theory**

# WHAT IS AN ALGEBRA?

- Narrow view: vector space with a bilinear product (typically with unit) Example: algebra of  $n$  by  $n$  matrices over a field,  $K[x]$
- Of interest to Chomskyans: algebras that are also coalgebras (bialgebras) with antipode: Hopf algebras See <https://nessie.ilab.sztaki.hu/~kornai/2023/Hopf/intr.pdf>
- Broader (more abstract) view: *some* elements (base set  $S$ ); *some* operations (addition, multiplication, ...) with a fixed *signature*; and some *identities*
- A *variety* is a set of algebraic structures so defined. Examples: semigroups, groups, rings, but not fields. Why not?
- Because of the **Birkhoff variety theorem**: bunch of algebras form a variety iff closed under **H**omomorphism, **S**ubalgebra, and **P**roduct
- Can't take direct product of fields, why not?

# THEORIES OF SEMANTICS CLASSIFIED BY MATHEMATICAL APPARATUS

- (1) Logic-based: the Frege–Russell–Tarski–Montague mainstream, henceforth MG, including lineal descendants like DRT, DPL, Inquisitive etc (Ajdukiewicz, 1935; Lambek, 1958; Lambek, 2004)
- (2) Based on (hyper)graphs: Traditional AI/KR (Quillian, 1969; Minsky, 1975; Sondheimer, Weischedel, and Bobrow, 1984; Pereira, 2012), AMR, 4lang
- (3) Based on linear algebra: distributional semantics (CVS)
- (4) Based on automata theory: Finite State models (operational semantics, production systems)
- (5) Based on rejection of formal apparatus: cognitive semantics
- We will touch on all five, in a syncretic and irenic fashion (1) MG and descendants are expected to be known by the students, but have little to offer for semi- and non-compositinality.

Knowledge postulates are a dead dog!



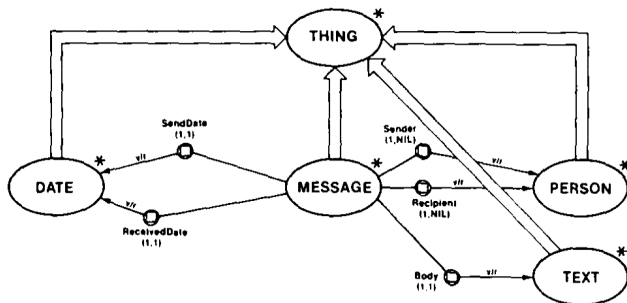
## (5) COGNITIVE SEMANTICS

- Clear linguistic appeal
- Intriguing, but informal, results
- Mainstream formal semantics has nothing to say
- Often insightful, rarely verifiable
- Langacker at the anti-formal extreme, Jackendoff at the formal end, Talmy in between

## (2) KR BASED ON (HYPER)GRAPH REPRESENTATIONS

- Mainstream approach in AI, its popularity moves in tandem with the AI hype cycle
- Linguists always had their own graphs (constituency, dependency, trees/DAGs, LFG diagrams, ...)
- Modern, linguistically inspired versions: AMR (Banarescu et al., 2013); 4lang (Kornai, 2010)
- Now terascale, primary tool in XAI

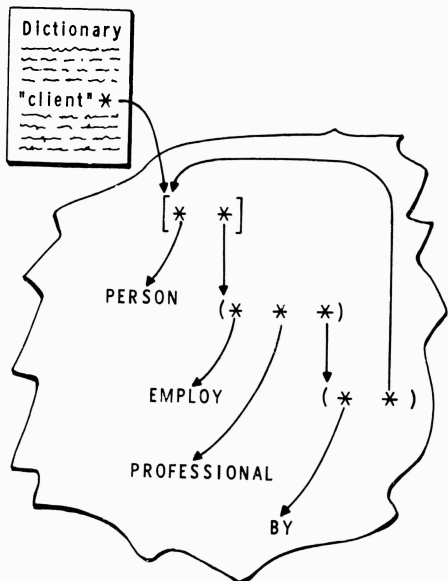
# CLASSIC KR



"A MESSAGE is, among other things, a THING with at least one Sender, all of which are PERSONs, at least one Recipient, all of which are PERSONs, a Body, which is a TEXT, a SendDate, which is a DATE, and a ReceivedDate, which is a DATE."

# QUILLIAN, SCHANK

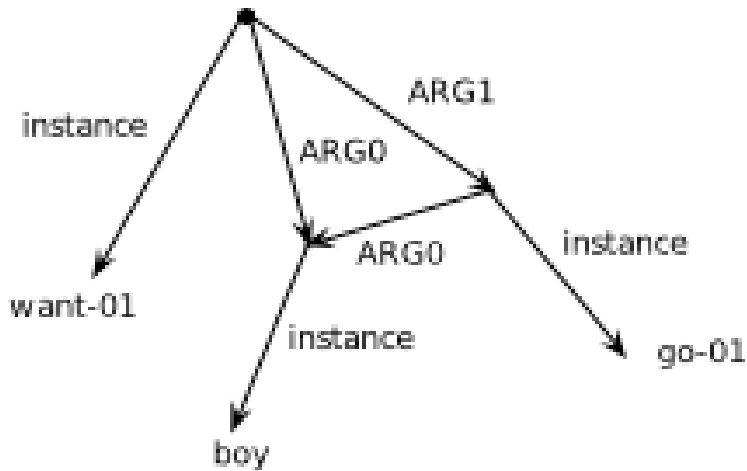
## Semantic Memory



## Conceptual Dependency

John  
‡    ⇒ good  
love  
↑  
one

# AMR GRAPHS

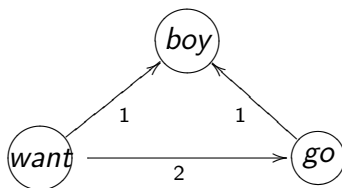


*The boy wants to go*

# AMR GRAPHS *cont'd*

- Rooted, directed, edge- and leaf-labeled graphs
- ~ 100 relations: :accompanier, :age, :beneficiary, :cause, :compared-to, :concession, :condition, :consist-of, :degree, :destination, :direction, :domain, :duration, :employed-by, :example, :extent, :frequency, :instrument, :li, :location, :manner, :medium, :mod, :mode, :name, :part, :path, :polarity, :poss, :purpose, :source, :subevent, :subset, :time, :topic, :value, :quant, :unit, :scale, :day, :month, :year, :weekday, :time, :timezone, :quarter, :dayperiod, :season, :year2, :decade, :century, :calendar, :era
- Neo-Davidsonian graph nodes for entities, events, properties, and states.
- Standardized AMR-parsed corpora (SemBanks) exist for English (60k sentences) and Chinese (5k)

## 4LANG GRAPHS



*The boy wants to go*

- Have two kinds of links: 1 (subject); 2 (object)
- In contrast, Cyc has over 45.000 link types, and contemporary efforts like DBpedia or YAGO have  $10^5 - 10^6$ . The vast majority of these are like *isSpouseOf*, obviously compositional
- 4lang graphs can be built on RDF-like “triple stores”, explicitly addressing known difficulties with these such as **negation**, **quantifier scope**, **nested modals** and relations of seemingly **higher arity** *LA is between San Diego and SF along US101*
- Effort to provide semantics for the entire vocabulary

# PRIMITIVES

- (1) Logic-based: constants, variables, quantifiers, relations/functions
- (2) (Hyper)graphs: (hyper)nodes, (hyper)edges (we will have hypernodes but only ordinary edges)
- (3) Linear algebra: points (vectors), spaces, mappings
- (4) Automata: states, transitions, i/o
- (5) Cognitive: traditional linguistic apparatus: morphemes, words, sentences, ... 'association'
- Neither blender nor full meta-structure (just partial mappings whenever possible)



# WHAT NEEDS TO BE MEMORIZED

## ATTRACTIVE NUISANCE

An *attractive nuisance* is something on a property that is likely to attract children, like a swimming pool, trampoline, or abandoned vehicle, even if they are trespassing, and may cause them harm. The doctrine of attractive nuisance holds the property owner liable for injuries to children who trespass and are injured by the dangerous condition

- *attractive* is not a primitive, it is obviously *attract+ive* where the suffix *-ive* contributes the adjectival role (and not much besides)
- attract =agt cause\_ {=pat want {=pat near =agt}}
- nuisance 'a person, thing, or situation that annoys you or causes problems' annoy, cause\_ problem
- problem situation, difficult, after(solve)
- 'some nuisance that is attractive' captures 90%

# DICTIONARY VERSUS ENCYCLOPEDIA

## What are the laws concerning attractive nuisances in Massachusetts?

According to Massachusetts state law [M.G.L. c. 231, § 85Q](#), a landowner who maintains an artificial condition on their property can be held responsible for physical harm caused to trespassing children under the following conditions:

- The landowner is aware or has reason to believe that children are likely to trespass on the specific area.
- The condition is known or should be known by the landowner, and it presents an unreasonable risk of severe bodily harm or death to the children.
- Due to their young age, the children are unable to discover the condition or comprehend the risks associated with interacting with it.
- The landowner's interest in maintaining the condition is minimal compared to the potential risk it poses to children.
- The landowner fails to take reasonable measures to eliminate the danger or protect the children in any other appropriate manner.

If a visitor or invitee is injured due to an attractive nuisance on their property under these circumstances, then the landowner can be held liable for the accident.

# TERMINOLOGY

- *Lexicon* stores linguistic information “word knowledge”; *encyclopedia* stores “world knowledge”. Cabrera, 2001 distinguishes four views:
- **Strong dualist**: it is feasible to draw a clear-cut distinction between dictionary and encyclopedia
- **Weak dualist**: some distinction between word- and world-information can be made, but dictionary meaning cannot be completely defined prior to implementation in context
- **Strong monist**: there is no dictionary/encyclopedia distinction, either theoretically or functionally, i.e. at the operational level of the actual processes of utterance interpretation
- **Weak monist**: there is no dictionary/encyclopedia distinction, not even in the terms proposed by weak or strong dualistic theories.

# WHAT NEEDS TO BE IN THE LEXICON?

- Universality demands keeping culture-specific assumptions out
- Encyclopedic and lexical knowledge grows at very different rates (2.5PB for the cancer genome database)
- We have an encyclopedia, WP, all we need is a pointer to it
- We have the compositional meaning (approximately) `thing, attracts, is_a problem`. By *salva veritate* substitution of the definition of *attract*, this is `thing cause_ {=pat want {=pat near =agt}}`, `thing cause_ problem`
- What is missing, is that the attracted are not any old `=pat`, they are children.
- This matters, since the whole thing is about legally protecting children from harm they are incapable of anticipating
- Maybe extra stylistic/topical flag 'legalese' → Tue

# SEMANTIC PRIMITIVES







- Are typically definable in terms of the other primitives
- There are only a handful of irreducible ones: and at before between cause\_er\_follow for\_from gen has in ins\_is\_a lack mark\_part\_of under wh but these are insufficient for defining the other primitives or 'semantic primes' (700+ of them)
- What is a primitive under one view may be definable under another! Example: gen (generic quantifier) =  $(\frac{1}{d}, \frac{1}{d}, \dots, \frac{1}{d})$
- We will use a particular set of about 770 primitives available at <https://github.com/kornai/4lang/blob/master/V2/700.tsv>
- '700' set obtained by systematic reduction from larger sets such as the LDV (2,200 entries w/o disambiguation)
- There are smaller sets, in particular Natural Semantic Metalanguage (Wierzbicka, 1992; Wierzbicka, 1996; Goddard, 2002) but these are not known to be truly defining. For an organized version see the Concepticon of (List, Cysouw, and

# Thank you!

Lecture will be made available at

<https://nessie.ilab.sztaki.hu/~kornai/2025/NASSLLI>

Tomorrow: word vectors

-  Ajdukiewicz, Kazimierz (1935). “Die Syntaktische Konnexität”. In: *Studia Philosophica* 1, pp. 1–27.
-  Banarescu, Laura et al. (2013). “Abstract Meaning Representation for Sembanking”. In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 178–186. URL: <https://www.aclweb.org/anthology/W13-2322>.
-  Cabrera, Julio (2001). ““The Lexicon-Encyclopedia Interface” by Bert Peeters (ed.)”. In: *Pragmatics and Cognition* 9.2, pp. 313–327. DOI: 10.1075/pc.9.2.09cab.
-  Chomsky, Noam et al. (2023). *Merge and the Strong Minimalist Thesis*. Cambridge University Press. DOI: 10.1017/9781009343244.
-  Coecke, Bob, Mehrnoosh Sadrzadeh, and Stephen Clark (2010). “Mathematical Foundations for a Compositional Distributional Model of Meaning”. In: *arXiv:1003.4394v1*.
-  Goddard, Cliff (2002). “The search for the shared semantic core of all languages”. In: *Meaning and Universal Grammar – Theory and*

*Empirical Findings*. Ed. by Cliff Goddard and Anna Wierzbicka. Vol. 1. Benjamins, pp. 5–40. DOI: 10.1075/slcs.60.07god.



Harley, Heidi (2014). “On the identity of roots”. In: *Theoretical Linguistics* 40.3/4, pp. 225–276.



Kartsaklis, Dimitrios (2014). “Compositional Distributional Semantics with Compact Closed Categories and Frobenius Algebras”. PhD thesis. Oxford.



Kornai, András (2010). “The algebra of lexical semantics”. In: *Proceedings of the 11th Mathematics of Language Workshop*. Ed. by Christian Ebert, Gerhard Jäger, and Jens Michaelis. LNAI 6149. Springer, pp. 174–199. DOI: 10.5555/1886644.1886658.



Lambek, Joachim (1958). “The mathematics of sentence structure”. In: *American Mathematical Monthly* 65, pp. 154–170.



— (2004). “A computational approach to English grammar”. In: *Syntax*.



List, Johann-Mattis, Michael Cysouw, and Robert Forkel (May 2016). “Concepticon: A Resource for the Linking of Concept Lists”. In: *Proceedings of the Tenth International Conference on*



*Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 2393–2400. URL:

<https://www.aclweb.org/anthology/L16-1379>.



Marcolli, Matilde, Robert Berwick, and Noam Chomsky (2023a).

*Old and New Minimalism: a Hopf algebra comparison*. URL:

<https://lingbuzz.net/lingbuzz/007373>.



— (2023b). *Syntax-semantics interface: an algebraic model*. URL:

<https://ling.auf.net/lingbuzz/007696>.



Marcolli, Matilde, Noam Chomsky, and Robert Berwick (2023).

*Mathematical Structure of Syntactic Merge*. arXiv: 2305.18278 [cs.CL].



Minsky, Marvin (1975). “A framework for representing knowledge”. In: *The Psychology of Computer Vision*. Ed. by P.H. Winston. McGraw-Hill, pp. 211–277.



Partee, Barbara (2013). “Changing Perspectives on the ‘Mathematics or Psychology’ Question”. In: *Philosophy Wkshp on “Semantics – Mathematics or Psychology?”*.



Partee, Barbara H. (1979). “Semantics - mathematics or psychology?” In: *Semantics from Different Points of View*. Ed. by R. Bäuerl, U. Egli, and A. von Stechow. Berlin: Springer-Verlag, pp. 1–14.



Pereira, Fernando (2012). “Low-Pass Semantics”. In: [http://videlectures.net/metaforum2012\\_pereira\\_semantic/](http://videlectures.net/metaforum2012_pereira_semantic/).



Quillian, M. Ross (1969). “The teachable language comprehender”. In: *Communications of the ACM* 12, pp. 459–476. DOI: 10.1145/363196.363214.



Sondheimer, Norman K., Ralph M. Weischedel, and Robert J. Bobrow (1984). “Semantic Interpretation Using KL-ONE”. In: *Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*. Stanford, California, USA: Association for Computational Linguistics, pp. 101–107.



Wierzbicka, Anna (1992). *Semantics, culture, and cognition: Universal human concepts in culture-specific configurations*. Oxford University Press.



Wierzbicka, Anna (1996). *Semantics: Primes and universals*.  
Oxford University Press Oxford.