# Hopf algebra reading seminar

András Kornai

Aug 9 2023 2PM CET

# ORGANIZATION

- 2pm zoom
  https://us02web.zoom.us/j/84045659802?pwd=L3grbWtqREE4OE
- 11pm zoom
  https://us02web.zoom.us/j/89203668566?pwd=M1dRL2ozOWxBT
- Slack https://join.slack.com/t/slack-qyx1689/shared_invite/zt-1xppi4d00-WnJhAvg_ThoSBOw9xH7yIw
- Course webpage
  https://nessie.ilab.sztaki.hu/~kornai/2023/Hopf
  Also reachable as kornai.com → 2023 → Hopf
- Attendance sheet
  https://docs.google.com/spreadsheets/d/17cK-cl3_xdbo73_kHWCIAvwgkd-G6qz44J4D6tyFfAc/edit?usp=sharing

# WHAT IS THE LINGUISTIC CONTENT?

1. Intro to pre-minimalism linguistics (2PM)
2. Intro to minimalism (Avery Andrews, 11PM)

# STRUCTURALIST LINGUISTICS

- Conventionally dated to have started with Ferdinand de Saussure (1879) *Mémoire sur le systme primitif des voyelles dans les langues indo-européennes*
- Major milestones include Bloomfield, 1933, Harris, 1951
- Standard intro textbook Hockett, 1958, still very readable
- Key ideas include the Sausserian *sign*: an association of *form* and *meaning*. Minimal units of form are called *phonemes* roughly corresponding to written letters, minimal (atomic) form/meaning pairs are called *morphemes*. Largest unit considered is typically the utterance (Harris, 1946)
- Descriptive (as opposed to normative), strongly tied to *discovery methods* using native speakers as oracles telling the field linguist whether two utterances have the same form/meaning

# A BIT OF MORPHOLOGY

- Studying the structure of words was the mainstay of classical linguistics (Sanskrit, Greek, Latin, Arabic grammarians going back to 2,500 years)

- The structuralists were no less enthusiastic, because (i) segmentation into words (items that tolerate no pause in the middle) is readily performed by native speakers, even in languages where sentences are often just one word as in Hungarian *Meglátogatnálak* 'I would like to visit you'; (ii) words are often irreducible units of meaning; (iii) there is generally strong *morphotactics* (constraints on how a language puts morphemes together to form words) even in languages where syntactic constraints (how to put words together to form sentences) are rather loose (as in Hungarian).

- Much of phonology is morphophonology (phonological changes driven by putting morphemes next to one another)

# THE TRADITIONAL/STRUCTURALIST PICTURE OF A WORD

- Words are composed of *stems* and *affixes*. The affixes can be *prefixes* (which precede the stem) or *suffixes* (which follow the stem). Infixes and circumfixes also exist

- Typically, the meanings are carried by the stems, the affixes carry grammatical categories (person, number, gender, tense, aspect, case, . . . ).

- We distinguish *inflectional* and *derivational* affixes, the former are important in *paradigms*, the latter for *lexical categories* (parts of speech). For example H. *as/es/os/ös* forms an adjective from the noun stem.

- Often we find a deeper layer of *roots* where meanings are less defined, and derivational affixes are required to form stems

# THE STRUCTURALIST-GENERATIVE OVERLAP

- This is roughly the 1960s, with generative work conventionally starting with Chomsky, 1956; Chomsky, 1957. Chomsky's PhD thesis (1955) is eventually published as Chomsky, 1975 (with some chapters still unpublished until later). For the structuralists:

- Phonemes and morphemes are treated for the most part as *sequential* units, but the structuralists (Trubetskoi, 1939; Jakobson, Fant, and Halle, 1952) already propose *parallel* units called *(distinctive) features* such as ±voiced, ±nasal, ... There is also structuralist work on parallel morphemes (e.g. emphasis in Arabic, see Jakobson, 1957) and on discontinuous ones (Harris, 1951) calls these 'long components'

- Larger structures are built from the sequential units using *immediate constituent analysis* Wells, 1947

# EARLY GENERATIVE GRAMMAR

- Chomsky, 1956 introduces two new mathematical models, CFGs and CSGs (context-free and context-sensitive grammars)
- These model only the form side of signs, which are put together into strings recursively, building trees above the leaves
- Chomsky reconstructs ICA by CFGs. This ignores three major issues:
- Discontinuous constituents
- The labels on the (pre)terminals "lexical categories"
- The labels on internal nodes "X-bar theory"

# Formal grammar refresher

## Formal languages

Let $\Sigma$ be a finite set of symbols (node labels). We divide this in two parts, $N \subset \Sigma$ is called the *nonterminal vocabulary*, which includes a distinguished *start symbol* $S$, and $\Sigma \setminus N$ is called the *terminal vocabulary*. With the concatenation operation, the free monoid over $\Sigma$ is denoted $\Sigma^*$, we call the elements 'strings' or 'words' or 'sentential forms' depending on the phase of the moon. There is also a distinguished *empty string* (unit element) denoted $\lambda$. Sets $L \subset \Sigma^*$ are called *formal languages*

We can specify formal languages by several methods:

- By properties
- By acceptors
- By generators
- By operations on already given languages

# The (extended) Chomsky hierarchy

- $\mathcal{L}_{-1}$: any language (not necessarily recursively enumerable). Over a binary alphabet these correspond to non-computable numbers: we know (e.g. by a cardinality argument) that there are many, but we can't pin them down
- $\mathcal{L}_0$: recursively enumerable. This implies a one-sided decision (e.g. by listing by a TM). Analogous to computable numbers
- $\mathcal{L}_{0.5}$: recursive (iff both $L$ and $\Sigma^* \setminus L$ are r.e.)
- $\mathcal{L}_1$: context-sensitive. Can be decided by TM limited to the tape the size of the input. Can be generated by CSGs
- $\mathcal{L}_{1.5}$: mildly context-sensitive. Can be decided in polynomial time, see Joshi, Vijay-Shanker, and Weir, 1990

# The (extended) Chomsky hierarchy, cont'd

- $\mathcal{L}_2$: context free. Can be generated by CFGs. Can be accpeted by pushdown automata. Analogous to algebraic numbers
- $\mathcal{L}_3$: finite state (regular). Can be accepted by finite automata. Can be generated by regular expressions. Has a finite syntactic monoid. Analogous to rational numbers.
- $\mathcal{L}_4$: counter-free (non-counting). Has an aperiodic syntactic monoid. Star-free regexps. See McNaughton and Papert, 1971; Kornai, 1985
- $\mathcal{L}_{4.5}$: Subregular. See Heinz, 2018
- $\mathcal{L}_5$: Finite

# Persistent issues with formal language theory

- Strings are likely not the most interesting/revealing data structures to consider
- Phonology employs several, partially synchronized strings (autosegmental representations, see Goldsmith, 1990; Kornai, 1995)
- Computational linguists prefer *weighted languages* (weights from a semiring, typically the *tropical semiring* modeled on log probabilities)
- Trees may not be the best descriptor of structures (other descriptors added e.g. in Lexical-Functional Grammar, see Bresnan, 2001)
- This is all about form (phonology, syntax), but what about meaning?

# Early generative work (1956–1973)

- Emphasis on syntax (traditionally hugely underresearched)
- The key technical tool was the *transformation* (tree-to-tree mapping)
- Phonology was handled by CSGs Chomsky and Halle, 1968
- Everything else by Transformational Grammars (TG) Chomsky, 1965
- Morphology (structure of words) and discourse (structure beyond the sentence) pretty much ignored
- Pretend-meaning *Logical Form* by tree structure

# THE PIVOT FROM FORMAL LANGUAGE THEORY

- Salomaa, 1971 and independently Peters and Ritchie, 1973 proves that TG is Turing-powered (every Type 0 language has a TG)
- Chomsky (and most others) want a theory of grammar that is specific to linguistic phenomena. This requires making TG *less* powerful.
- Ross, 1967 starts to work in this direction, by adopting a theory where transformations manipulate meaing representations (LF trees) until they become syntactic trees. Chomsky prefers a different model where the initial (deep structure) trees are manipulated both towards meaning representation (LF) and towards syntactic representation, see Chomsky, 1970; Chomsky, 1971. Thus break out the *linguistic wars*, a story told from Chomsky's side by Harris, 1995 and from the other side by Goldsmith and Huck, 2013

# LESS POWERFUL TRANSFORMATIONS

- Having (in a sociological sense) won the linguistic wars, Chomsky, 1973 also starts working towards less powerful transformations

- There is a whole evolution of theories starting with Government-Binding (Chomsky, 1981) going through early minimalist theory (Chomsky, 1995) and leading to the current (Hopf?) version

- The architecture is based on a revised theory of signs, which have three parts: a meaning (some kind of trees), a syntax (other kinds of trees), and a form (phonology, not discussed in any detail)

- These units are what Harley, 2014 calls *roots*

- Prof. Avery Andrews will take up the story from here

Bloomfield, Leonard (1933). *Language*. London: George Allen and Unwin.

Bresnan, Joan (2001). *Lexical-Functional Syntax*. Oxford, UK: Blackwell.

Chomsky, Noam (1956). "Three models for the description of language". In: *IRE Transactions on Information Theory* 2, pp. 113–124.

— (1957). *Syntactic Structures*. The Hague: Mouton.

— (1965). *Aspects of the Theory of Syntax*. MIT Press.

— (1970). "Remarks on nominalization". In: *Readings in English Transformational Grammar*. Ed. by R. Jacobs and P. Rosenbaum. Waltham, MA: Blaisdell, pp. 184–221.

— (1971). "Deep Structure, Surface Structure, and Semantic Interpretation". In: *Semantics: An Interdisciplinary Reader in Philosophy Linguistics and Psychology*. Ed. by Danny D. Steinberg and Leon A. Jakobovits. Cambridge University Press, pp. 183–216.

Chomsky, Noam (1973). "Conditions on Transformations". In: *A festschrift for Morris Halle*. Ed. by S.R. Anderson and P. Kiparsky. New York: Holt, Rinehart and Winston.

— (1975). *The logical structure of linguistic theory*. Springer Verlag.

— (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

— (1995). *The Minimalist Program*. MIT Press.

Chomsky, Noam and Morris Halle (1968). *The Sound Pattern of English*. New York: Harper and Row.

Goldsmith, John A. (1990). *Autosegmental and Metrical Phonology*. Cambridge, MA: Blackwell.

Goldsmith, John A. and Geoffrey J. Huck (2013). *Ideology and linguistic theory: Noam Chomsky and the deep structure debates*. Routledge.

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

📄 Harris, Randy Allen (1995). *The Linguistics Wars*. Oxford University Press.

📄 Harris, Zellig (1946). "From morpheme to utterance". In: *Language* 22, pp. 161–183.

📄 — (1951). *Methods in Structural Linguistics*. University of Chicago Press.

📄 Heinz, Jeffrey (2018). "The computational nature of phonological generalizations". In: *Phonological Typology*. Ed. by Larry Hyman and Frans Plank. De Gruyter Mouton, pp. 126–195.

📄 Hockett, Charles (1958). *A course in modern linguistics*. New York: Macmillan.

📄 Jakobson, Roman (1957). "Mufaxxama – the emphatic phonemes in Arabic". In: *Studies Presented to Joshua Whatmough*. Ed. by E. Pulgrim. The Hague: Mouton.

📄 Jakobson, Roman, Gunnar Fant, and Morris Halle (1952). *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. MIT Press.

📄 Joshi, Aravind K., K. Vijay-Shanker, and David J. Weir (1990). *The Convergence of Mildly Context-Sensitive Grammar Formalisms*. Tech. rep. University of Pennsylvania.

📄 Kornai, András (1985). "Natural Languages and the Chomsky Hierarchy". In: *Proceedings of the 2nd European ACL Conference*. Ed. by Margaret King, pp. 1–7.

📄 — (1995). *Formal Phonology*. New York: Garland Publishing. ISBN: 0815317301.

📄 McNaughton, Robert and Seymour Papert (1971). *Counter-Free Automata*. MIT Press.

📄 Peters, Stanley and Robert W. Ritchie (1973). "On the generative power of transformational grammars". In: *Information Sciences* 6, pp. 49–83.

📄 Ross, John Robert (1967). *Constraints on Variables in Syntax*. MIT.

📄 Salomaa, A. (1971). "The generative capacity of transformational grammars of Ginsburg and Partee". In: *Information and control* 18.3, pp. 227–232. ISSN: 0019-9958.

📄 Trubetskoi, Nikolai Sergeevich (1939). *Grundzüge der Phonologie*. Göttingen: Vandenhoeck and Ruprecht.

📄 Wells, Roulon S. (1947). "Immediate constituents". In: *Language* 23, pp. 321–343.