Semicompositionality

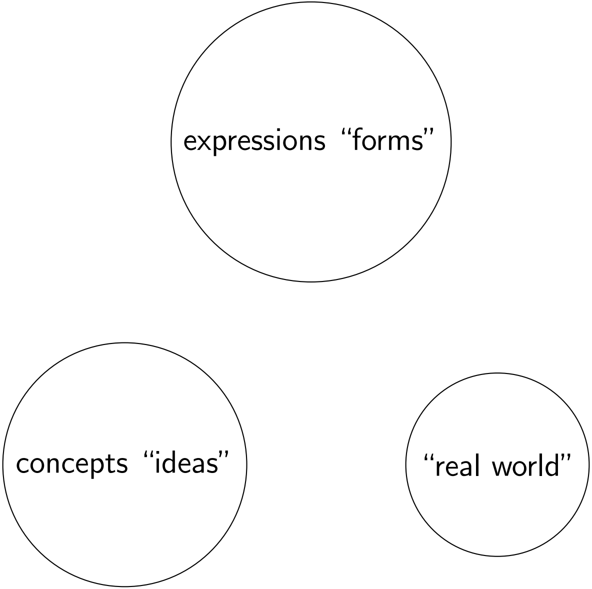
András Kornai
SZTAKI Computer Science Research Institute
and
Dept of Algebra, Budapest Institute of Technology

NASSLLI, June 27 2025

Outline

1. Ontology
2. Schemas
3. Conceptual blending
4. Idioms
5. Negation
6. Bound morphemes
7. Roots and features
8. Morphotactics

The overall model



expressions "forms"

concepts "ideas"

"real world"

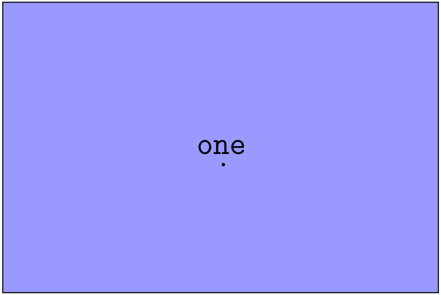
What do we need to cover semicompositional phenomena?

- Some items: *morphemes* composed of *form* and *meaning* parts plus some means of expressing the fact that they jointly participate in a structure
- Further items on the form side (phonemes, features, tiers) plus some means of expressing the fact that they jointly participate in a structure *including temporal relations*
- Further items on the meaning side (schemas, lexemes) plus some means of expressing the fact that they jointly participate in a structure
- Methods for (recursively) combining morphemes/forms/meanings (algebra)
- Methods for (recursively) cutting morphemes/forms/meanings (coalgebra)
- Methods for handling exceptions

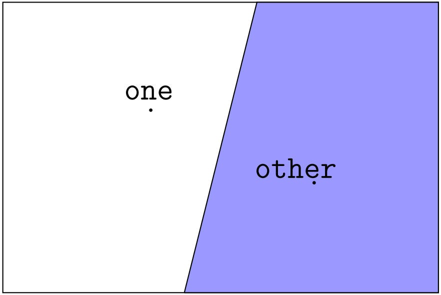
Minimum requirements

- We need some patterns (regexps using typed variables)
- We need finite but fine-grained lexical categories for the typing
- We need some substitution mechanism (equalizers)
- Do we need bracket retention?
- Implementing the above in a rather amorph fabric of (artificial) neurons
- Estimate size requirements based on hard (information-theoretic) bounds
- Plausible acquisition mechanism for LLMs and for humans
- Cover at least the basic grammatical mechanisms such as morphosyntax

The simplest schemas

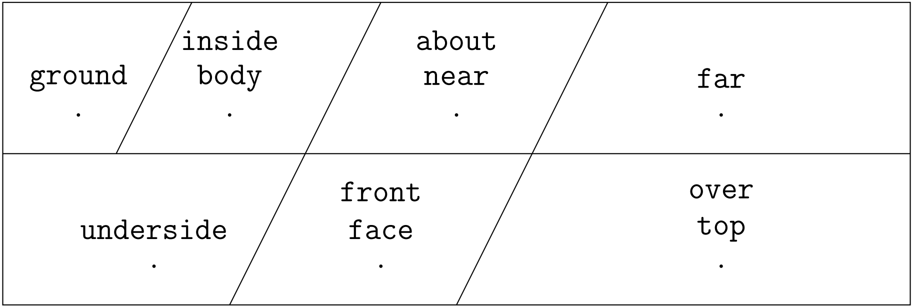


one/all



other

The place_ schema



ground .

inside
body .

about
near .

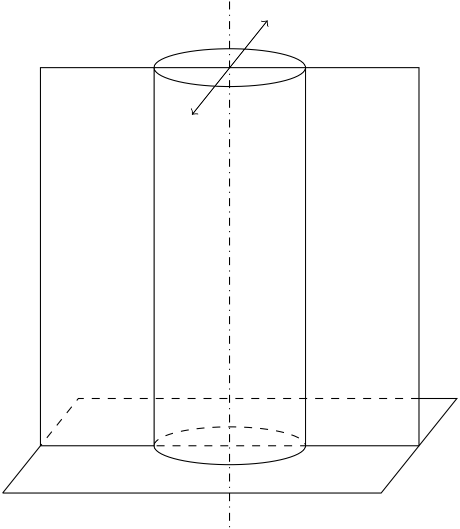
far .

underside .

front
face .

over
top .

Egocentric coordinates



Blending

- Ted Huges: The thought fox
- More an artistic device than part of ordinary communication
- But it is entirely linguistic, so linguistics should be able to say *something* about it
- Prior to generative grammar, linguists felt responsible for this kind of data
- See (Turner, 2021) for more detailed analysis

What moves, what stays fixed

(Kay and Sag, 2014; Jackendoff and Audring, 2020) getting cold feet, taking a Mulligan

- What does *take a Mulligan* mean?
- Does the example *Trump Takes a U.S. Steel Mulligan* (WSJ 5/28) help?
- Because of *take* we know it's a VP
- Inflects like a VP: *took, has taken, will take, . . .*
- compare *take one for the team, take the back seat, take a French leave*
- Entirely fixed expressions: *by and large, right away, first off, all of a sudden, as is*

What is the problem?

- Classic theory: negation is an involution: $\neg\neg = id$ that dualizes conjunction and disjunction
- Great for logic/math, much less so for NLP, where positive and negative statements are quite asymmetric:
  the form and function of negative statements in ordinary language are far from simple and transparent. In particular, the absolute symmetry definable between affirmative and negative propositions in logic is not reflected by a comparable symmetry in language structure and language use. Much of the speculative, theoretical, and empirical work on negation over the last twenty-three centuries has focused on the relatively marked or complex nature of the negative statement vis-a-vis its affirmative counterpart Horn, 1989

What is the solution?

- Dyadic theory: negation is actually a two-variable operator `lack`
- This operates specifically on default values
- *blind* = 'lack sight' meaningful for people, animals, but #blind stone is infelicitous
- Ordinary *no* is analyzed as `gen lack`
- Works well with quantifiers, double negation, imperatives, etc
- For details see Kornai, 2024

Why do we care?

- Small segment of the vocabulary by type (few thousand roots, few hundred affixes)
- Huge by token frequency (typically over 50%) Bound morphemes (if you include them in the frequency count) and ‘function words’ (if you don’t) absolutely dominate the frequency lists
- In this context, roots are less exciting, especially if they yield zero-derived stems, the really bothersome part are the affixes
- They seem to be indispensable for delineating core vocabulary
- LDV suffixes: *-able, -al, -an, -ance, -ar, -ate, -ation, -dom, -en, -ence, -er, -ess, -est, -ful, -hood, -ible, -ic, -ical, -ing, -ion, -ish, -ist, -ity, -ive, -ization, -ize, -less, -like, -ly, -ment, -ness, -or, -ous, -ry, -ship, -th, -ure, -ward, -wards, -work, -y*
- LDV prefixes: *counter-, dis-, en-, fore-, im-, in-, ir-, mid-, mis-, non-, re-, un-, vice-, well-*

Roots

- In linguistics *roots* are generally considered the smallest morphemes (meaning/form pairs)
- Typical examples include Pāṇinian roots (the Dhatupatha lists 1943, Whitney, 1885 lists only 885), triconsonental roots in Semitic languages, etc.
- Sometimes the meanings are very clear, Sanskrit *smi* 'smile', *vadh/badh* 'slay'; Hebrew *t.l.p.n* 'telephone'
- But often the meanings are more hazy, as in Skt *aNh* 'narrow, distressing', English *be* (am, are, is, was, were, would)
- Historical depth just adds to the bleaching of the meaning, consider English *mit* (commit, demit, remit), *pose* (compose, depose, repose) etc
- Harley, 2014 departs from this tradition somewhat, paying little attention to morphological and phonological structure
- Subroot units? "phonestemes", "ideophones", "phonosemantics"

Features

- In linguistics *features* are also minimal (typically bound) morphemes, but they come in contrasting sets, and are seen as dependent on the stems. The consensus list (before Principles and Parameters):
-

1A	Person (1st, 2nd, . . .
1B	Number (singular, dual, . . .
2A	Location (here, there, near, . . .
2B	Direction (to, from, . . .
3A	Gender (feminine, definite, animate, valuable, round-shaped, . . .
3B	Topic (familiar, known, . . .
4A	Tense (past, present, . . .
4B	Aspect (perfect, habitual, . . .
5A	Case (subject, object, . . .
5B	Voice (active, benefactive, . . .
6A	Degree (comparative, superlative, . . .
6B	Mood (interrogative, negative, . . .

Features in grammar

- Long history in modern grammar: features were critical in phonological theory Trubetskoi, 1939; Jakobson, Fant, and Halle, 1952; Chomsky and Halle, 1968
- They were also critical in lexical semantics Katz and Fodor, 1963, anthropology, social science, . . .
- ϕ-features are the ones participating in agreement (Chomsky since 1990s)
- Traditional distinction between inflection and derivation
- As is typical for ling, distinctions are reasonably sharp, but not entirely clear-cut
- This goes both for the inflection/derivation distinction but also for the larger root/feature or syntax/semantics issue

Two endpoints of a scale?

- Typically we see

	roots	**features**
content	yes	specific to itself
in paradigm	no	yes ‘holism’
morphology	lots	little
syntax	little	lots
expression	item	rule
position	head	dependent
information	12 bits	2 bits

- Is reflexive *self* a feature or a root? For a feature it is very contentful (most features just provide one bit, here we get a whole homunculus) and for a root it is very assertive in syntax (most roots only participate in the morphotactics but are not across word boundaries).

Core/simple/basic vocabulary

- Teachers (both L1 and L2) want word lists, readability formulas aimed at "simple" words
- Psychologists, cognitive scientists want "core" vocabulary
- Typologists, lexicographers want "basic" vocabulary
- Main data sources: native speaker judgments (Likert ratings of simplicity); counting measures (length, # of functional parts); developmental order (acquisition age); cross-linguistic (type) frequency; within-language (token) frequency
- For algebra we need a *basic* set (definitional closure)
- For NLP we need a *common* (top-frequency) set

Comparing some candidates

- Corpus: 9.6M token BNC spoken sub-corpus, function words stripped
- Five candidate basic lists evaluated: NSM (78), Swadesh (207), 4lang (732), Ogden Basic English (850), LDV (2,112).
- Coverage metric: probability mass of content tokens; **density** = coverage/ideal Zipf-top of same size.
- Results: small lists capture 13–16 % mass; LDV captures 64 % (79 % of ideal), showing trade-off size vs. coverage.

Coverage of basic vocabularies

list	size	w/o fw	weight	avg wt	density
NSM	78	53	13.3%	0.251%	41.0%
Swadesh	207	185	15.7%	0.085%	30.9%
4lang	732	714	31.2%	0.044%	45.9%
Ogden	850	799	33.4%	0.042%	48.1%
LDV	2190	2112	64.4%	0.030%	78.7%
$\cup$	2390	2310	68.5%	0.030%	82.7%
$\cap_3$	464	428	30.4%	0.071%	50.0%
UG5	1000	913	61.7%	0.068%	86.5%

Weight is the probability mass of content tokens in the BNC spoken section. UG5 is Randall Munroe's Up-Goer Five. Subsequent examples from the UG5 challenge https://kornai.com/VCB/ has the data

Typical problems in defining vocabulary use

- **idiomatic English**. "…interesting because that gives us a real leg up in finding out how the mind works"
- **multiple senses**. *space* 'the area beyond the Earth where the stars and planets are' versus 'the amount of an area, room, container etc that is empty or available to be used'.
- **associative descriptions**. *funny voice air* (helium) *the kind of air that once burned a big sky bag* (hydrogen)
- **nonce compounding**. *train-food* (fuel) vs *idea-paper, air-light, pretend-box, fire rock*
- **circumlocution** *a jumping animal that lives in the water and makes noise* (frog) *the stuff that comes out of the animal with white and black spots*
- **lack of naming** *cold air for burning* (liquid oxygen) *wet and very cold air*

MDL perspective

- Problem with high-rank lists: function words (not very useful for definitions) crowd out more rare but semantically indispensable primitives
- MLD solution: treat word list L as a codebook; overall cost $K(L) + K(D \mid L)$
- This optimises fungibility: complexity of new term = complexity of its definition
- Additional coordination cost c per comma weights syntactic complexity
- Optimal core emerges when incremental cost of adding a word equals cost saved by shorter definitions (*cran* morphemes not worth defining) *scrumpti*ous
- How about *mendicant, mendacious, amend, emend, commend, mend, tremendous*? Which are psychologically related?

Position classes

- Example: Hungarian nouns. Stem + personal possessive (14) + familial (2) + anaphoric (3) + case (17) for a total of 1,428 possibilities
- Can be built inside-out and outside-in. Reduces memorization from 1,428 combinations (and ther harmonic alternants) to 14+2+3+17=36 (2.5%)
- Even if we need some rules in the mix, the savings are irresistible
- Broad concatenative picture: prefixes + stem + suffixes, but order matters. Sometimes it's (pre+stem)+suff, sometimes pre+(stem+suff). How would you argue for one or the other?
- Morphological steps are coupled both to (often semicompositional) semantic and 'cyclic' phonological processes
- Morphophonology, morphotactics

The phonological cycle

- Idea: in the process of word formation, items undergo the same phonological processes again and again
- Example: English stress shift (Chomsky and Halle, 1968)
  atom→ **atomic**→ **atomicity**
- First cycle: **atom**
  - Morphological root: /ǽtəm/
  - Stress: First syllable stressed→ ['æ.təm]
- Second cycle: **atomic**
  - Add suffix *-ic*: /ǽtəm + ɪk/
  - Stress shifts to penultimate syllable: [ə.'tɑ.mɪk]
  - Vowel reduction: [ǽ]→ [ə] in unstressed syllable
- Third cycle: **atomicity**
  - Add suffix *-ity*: [ətɑ’mɪk + ɪ'ti]
  - Stress moves to syllable before *-ity*: [ə.tə.'mɪ.sə.ti]

Cyclic Phonology

- General observation: earlier changes (e.g. in vowel quality) persist and affect output
- Root vowel remains reduced: [æ]→ [ə]
- Output depends on order and structure of affixation
- SPE doesn't discuss, but we will: what does *atomic* mean?
- Well, what does *atom* mean? According to `4lang` , it means `particle, lack part, small`. OK, so what does *particle* mean? `piece, separate, small`
- What does *ic* mean? 'of, like, or related to a particular thing' https://www.merriam-webster.com/dictionary/-ic
- *ity* ‘having the property of X’
- English has 300+ words ending in *-ity*. Problem #1 filter out the false positives like *city, pity* from the real hits like *activity, bestiality*. What about *alacrity*? un-believ-able

Back to the phonology

- The cycle's been researched for 50+ years now!
- 1980s: Lexical phonology (Kiparsky, 1982; Mohanan, 1982; Rubach, 1984) assumes the lexicon has levels (root→ level 1→ level 2 . . .): each level executes *morphological operations* then its own block of *lexical phonology*
- Output of any level is a well-formed *lexical item*; syntax and post-lexical phonology operate only afterwards
- Cyclic/level ordered flavors of Optimality Theory (Prince and Smolensky, 1993)
- 2020s: (**Steriade:2025**)

Level-ordered morphology (Siegel 1974→ Kiparsky 1982)

- **Primary** affixes (level 1) fuse with the stem: + boundary, stress shift, vowel shortening. **Secondary** affixes (level 2) respect word stress, refuse cyclic phonology. Siegel, 1974
- Ordering constraint: *all* primary prefixes lie *outside* secondary prefixes; primary suffixes lie *inside* secondary suffixes (e.g. *nation-al-ism*, never *nation-ism-al).
- Phonology explains morphology: affix order correlates with access to cyclic rules (word-stress, trisyllabic shortening)
- Lexical rules = cyclic, word-bounded, disjunctively ordered, structure-preserving, may have exceptions.
- Post-lexical rules = apply once, phrase-bounded, automatic, non-structure-preserving (aspiration, sandhi).

Bracketing Erasure Convention & cyclicity

- **BEC**: erase all internal brackets at the end of each level→ later rules cannot “peek” inside earlier structure
- Explains Strict Cyclicity: lexical phonology only targets constituents created *this* cycle
- Consequence: syntax or post-lexical phonology don’t reference sub-word constituents
- Zero derivation split across levels
  - Two sources for N/V pairs: level 1 “root-co-category” (re**bélrèbél**)
    level 2 zero affix (pattern→to pattern)
  - Phonology distinguishes them: level 1 verbs shift stress
    level 2 verbs keep nominal stress; only level 1 verbs take primary suffixes (*patternization* vs. **patternal*)

Blocking = "Avoid Synonymy"

- Regular process is *blocked* if it would create a form synonymous with an earlier-level lexical item
- Captures strong-verb pasts (sing → sang blocks *singed) and derivational pairs (*glory* blocks **gloriosity*). Kiparsky, 1982
- Generalized to partial blocking: *cutter* allowed only when no more specific tool name exists

Morphological reanalysis and apparent violations

- Puzzles like *un-grammatic-al-ity* violate level order on the surface.
- Kiparsky's solutionlevel 2 prefix attaches, *then* word is re-bracketed provided every affix's subcat requirements remain satisfied (Projection Principle)
- Only works for non-category-changing level 2 prefixes→ predicts why *de-natural-ity or *ir-resource-ful can't exist

Affix ordering

- Is tricky no matter what
- Distributed Morphology (Halle and Marantz, 1993) uses Root/Stem distinction
- Computational systems use continuation classes
- Analysis of Hungarian uses 168 continuation classes
- What do you do with leg+nagy+bb? (Bobaljik, 2012)

Morphosyntax

- The phenomena: meaning organized by highly abstract schemas
- Examples: active/passive, causative, . . .
- We begin with the earliest such theory
- Fly over modern territory (Gruber, 1965; Fillmore, 1968; Anderson, 1977; Fillmore, 1977; Kiparsky, 1987; Butt, 2006)
- Discuss why we need to brutally simplify

KARTṚ – THE INDEPENDENT DOER

- **Definition** (P. 1.4.54): *svatantraḥ kartā* – the participant who acts autonomously
- **Core case** = nominative; in passives expressed by instrumental
- **Example**

  vipreṇa pacyate — "it is cooked by the brāhmaṇa"

  Here *vipreṇa* (instr.) marks the agent although the verb is passive
- **Modern** Agent (typically nominative)

KARMAN – WHAT IS PRIMARILY AFFECTED

- **Definition** (P. 1.4.49): "that which the agent chiefly intends."
- **Core case** = accusative; surfaces as nominative in passives, genitive in objective compounds, etc.
- **Example chain**(same relation, different forms)
    1. *kumbhān karoti* — "he makes pots" (acc.)
    2. *kumbhāḥ kriyante* — "pots are made" (nom.)
    3. *kumbhāṇāṃ kartā* — "maker of pots" (gen.)
- **Modern** Patient (typically accusative)

KARAṆA – THE EFFICIENT MEANS

- **Definition** (P. 1.4.42): “that which is most efficacious in accomplishing the action.”
- **Core case** = instrumental
- **Example**

  paraśunā vṛkṣaṃ chinatti — “he cuts the tree with an axe”
- **Modern** Instrument

SAṂPRADĀNA – THE INTENDED BENEFICIARY

- **Definition** (P. 1.4.32): the entity *for whom* something is given or done.
- **Core case** = dative
- **Example**

  viprāya gāṃ dadāti — "he gives a cow to the brāhmaṇa"
- **Modern** recipient

APĀDĀNA – THE ABLATIVE SOURCE

- **Definition** (P. 1.4.24): movement *away from* a fixed point.
- **Core case** = ablative
- **Example** *grāmāt āgacchati* — “he comes from the village”
- **Modern** source

- **Definition** (P.1.4.45): the locus where the action is situated
- **Core case** = locative
- **Example** *kaṭe āste* — “he sits on a mat”
- **Modern** locative

Simplify! Simplify!

- Start with the locative. Its utility is compelling. No grammar (especially not that of Hungarian, which has 9 locative cases) can live without it
- Take a simple locative *John is at the office* – what does it mean?
- `John at office`: we assume *John* is the subject, and *office* is the prepositional object of *at*
- OK, but what does *at* mean? `=agt has place, =pat[place], "at _" mark_ =pat` ‘the 2nd argument is a place, the place of the 1st argument’
- Spatial structure determined by egocentric model
- It is clear you need at least two arguments to play this trick (which we use for instruments, datives, ablatives, etc)

Thank you!

Lectures are made available at

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

Anderson, John (1977). *On Case Grammar*. Atlantic Highlands, NJ: Humanities Press.

Bobaljik, Jonathan David (2012). *Universals in Comparative Morphology: Suppletion, Superlatives, and the Structure of Words*. Cambridge, MA: MIT Press. ISBN: 978-0262017596.

Butt, Miriam (2006). *Theories of Case*. Cambridge University Press. DOI: 10.1017/CBO9781139164696.

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

Fillmore, Charles (1968). “The case for case”. In: *Universals in Linguistic Theory*. Ed. by E. Bach and R. Harms. New York: Holt and Rinehart, pp. 1–90.

— (1977). “The case for case reopened”. In: *Grammatical Relations*. Ed. by P. Cole and J.M. Sadock. Academic Press, pp. 59–82.

Gruber, Jeffrey Steven (1965). “Studies in lexical relations”. PhD thesis. Massachusetts Institute of Technology.

- Halle, Morris and Alec Marantz (1993). “Distributed Morphology and the Pieces of Inflection”. In: *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*. Ed. by Kenneth Hale and S. Jay Keyser. Cambridge, MA: MIT Press, pp. 111–176. ISBN: 978-0262581240.
- Harley, Heidi (2014). “On the identity of roots”. In: *Theoretical Linguistics* 40.3/4, pp. 225–276.
- Horn, Larry (1989). *The Natural History of Negation*. Chicago: University of Chicago Press.
- Jackendoff, Ray and Jenny Audring (2020). *The texture of the lexicon*. Oxford University Press.
- Jakobson, Roman, Gunnar Fant, and Morris Halle (1952). *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. MIT Press.
- Katz, Jerrold J. and Jerry A. Fodor (1963). “The structure of a semantic theory”. In: *Language* 39, pp. 170–210. DOI: 10.2307/411200.

- Kay, Paul and Ivan Sag (2014). “A lexical theory of phrasal idioms”. In: *ms.*
- Kiparsky, Paul (1982). “Word-formation and the Lexicon”. In: *Proceedings of the Mid-America Linguistics Conference*. Ed. by F. Ingemann. Lawrence, Kansas.
- — (1987). *Morphosyntax*. Stanford University: ms.
- Kornai, András (2024). “Dyadic negation in natural language”. In: *Acta Linguistica Academica* 71, pp. 235–257. DOI: 10.1556/2062.2024.00656. URL: https://akjournals.com/view/journals/2062/71/1-2/article-p235.xml.
- Mohanan, K.P. (1982). *Lexical Phonology*. MIT.
- Prince, Alan S. and Paul Smolensky (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Piscataway, NJ: Rutgers University Center for Cognitive Science Technical Report 2.
- Rubach, Jerzy (1984). *Cyclic and lexical phonology*. Dordrecht: Foris.

- Siegel, Dorothy (1974). *Topics in English Morphology*. MIT.
- Trubetskoi, Nikolai Sergeevich (1939). *Grundzüge der Phonologie*. Göttingen: Vandenhoeck and Ruprecht.
- Turner, Ross (2021). "Analysis of Ted Hughes's 'The Thought-Fox' using Conceptual Integration Theory (Blending)". In: *Academia Letters* 1571. DOI: https://doi.org/10.20935/AL1571.
- Whitney, William Dwight (1885). "The roots of the Sanskrit language". In: *Transactions of the American Philological Association (1869–1896)* 16, pp. 5–29. DOI: 10.2307/2935779.