Semicompositionality

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

NASSLLI, June 24 2025

Outline

Waltzing, dancing, moving

- (Slides added because of something in Prof. Moss' class yesterday)
- We all know that waltz is a kind of dance, and dance is a kind of motion
- How do we know this? Where is all this information stored?
- Answer: this is something stored in the lexicon, and we learn it in the process of acquiring lexical entries (early on, typically before kindergarten)
- Idea front and center in GOFAI/KR, where the relation is called IsA, as opposed to *is*
- Like many solid ideas, it goes back to Aristotle, who argued that definitions are based on *genus* and *differentia specifica*
- For an early Aristotelian system see the Tree of Porphyry

More formally

- IsA amounts to containment (set theoretical $\subset$) of one definition in another
- To say that `fox IsA animal` is to say that whatever properties animals have are enjoyed by foxes (but foxes will have more, like redness and being clever)
- There is an antitone Galois connection between extensions (domains designated by expressions) and definitions: the shorter a definition (fewer conjuncts it has) the more things fit
- (There will be plenty of discussion of how lexical definitions look like as we go along, there is even a formal grammar for this in Ch. 1.5 of Kornai, 2023)
- You are assuming of any competent speaker of English that they know the lexicon
- But you don't assume they know the encyclopedia

Information content of language

- The word entropy of natural language is about 12–16 bits/word Kornai (2007) Ch. 7.1. Capitalization and punctuation (our best proxies for intonation and related factors) contribute less than 7% (0.12 bits of 1.75 bits per character Brown et al. (1992a).
- Syntax is an information source of its own. There are many formalisms, we just consider binary trees over n words. These contribute *at most* $\log_2 C_n$ bits. C_n is hard to compute exactly, but asymptotically $C_n \sim 4^n/\sqrt{\pi}n^{1.5}$, so encoding a parse tree requires less than 2 bits/word. (The masoretes used 2 bits for parsing the Bible, Aronoff (1985))
- The key takeaway: Information is carried by the words. Logical structure accounts for no more than 12–16% of the information conveyed by a sentence, a number that actually goes down with increased sentence length, and emotive content for even less, perhaps 5–7%.

Kolmogorov complexity

- Shannon information is about maximal compression by prefix-free codes, KC is about maximal compression by algorithm
- Consider binary digits
  01101010000010011110011001100111111100111011110011001000010001011001011111101100010011011001101 1
- You may not immediately recognize this, but these are the first 100 binary digits of $\sqrt{2}$. By the Weyl equidistribution theorem, this sequence cannot be compressed!
- But Andrei knows how to compress it! Just use the channel to transmit the program that knows how to compute $\sqrt{2}$ to arbitrary precision (constant number of bits) plus the desired length of the sequence ($\log n$ bits)

Data compression

- Confers huge evolutionary advantage (the more you can pack in a brain the better off you are)
- All is fair: KC is great if you can get it
- Amortization can/should be assumed in the form of UG/culturally shared memory content
- The lexicon is culturally shared, the encyclopedia isn't
- We will look at major linguistic modalities (speech, sign, writing)
- Holmes 1971

Information in speech

- CD-ROM Audio – 700 kbps (44.1khz, 16 bits per sample)
- MPEG Audio – 112 kbps (44.1khz, 3 bits per sample)
- regular "toll" quality speech – 96 kbps
- ADPCM – 32 kbps (toll quality)
- LPC – 9.6 kbps (near toll quality)
- VQ homomorphic – 0.6 kbps
- symbolic – 0.2 kbps (0.05 kbps)
  https://kornai.com/Drafts/holmes.mp4

Bound morphemes

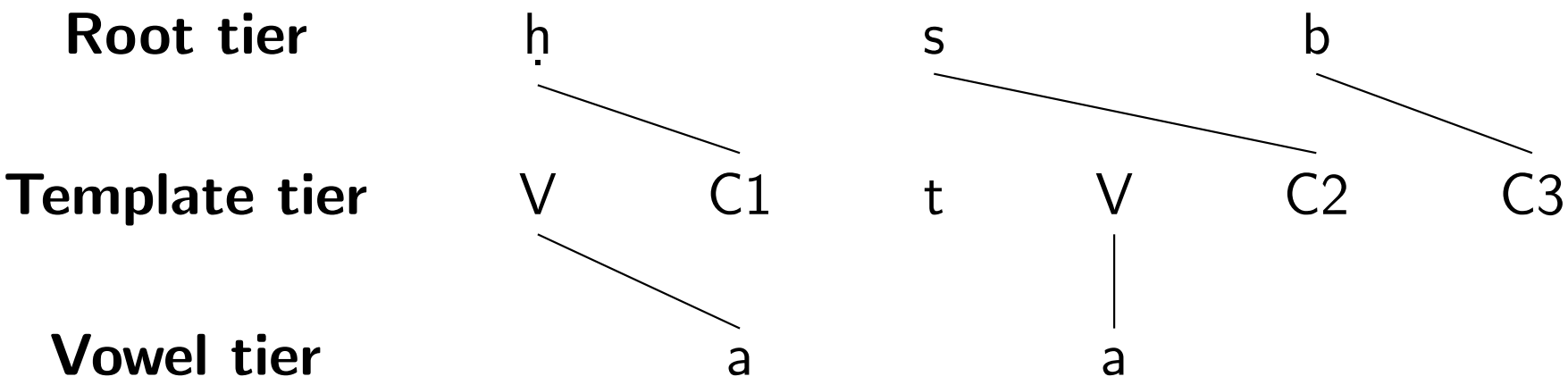
- Example: *-th* "Sixth of the pizza" versus "Finished sixth"
- `part, in whole, before(divide)` versus `position, in sequence`
- If you can't explain the meaning you must assume a primitive, tertium non datur
- What you need is a cross-linguistic inventory or grammatical distinctions
- The single most frequent is SG/PL
- Some languages like Classical Arabic have singular-dual-plural
- Many languages (e.g. Hungarian) have traces of a dual number

Lihir independent pronouns

	Singular	**Dual**	**Trial**	**Plural**
1^{EXCL}	yo	gel	getol	ge
1^{INCL}	—	kito	kitol	giet
2	wa	gol	gotol	go
3	e	dul	dietol	die

- Data from Ross, 1988 via Corbett, 2000
- Many languages have *paucal* 'a few' and *greater paucal* 'several, a bunch' but these are rarely grammaticized

Binyan VIII



- What are the primitives?
- How do we manipulate them?
- What are the relationships between representations, both partial and full?
- We do everything by finite state methods

Automata

- Come in many flavors: Mealy, Moore, Eilenberg, transductive probabilistic, . . .
- We will use Moore with output, aka 'subsequential transducers'
- Defined by state space Q (with initial state q_0)
- Input alphabet Σ, output alphabet O (may have overlap)
- Transition $\delta : Q \times \Sigma \to Q$ (function if deterministic, relation if nondet)
- Output $Q \to O^*$ (function if overt, relation if hidden)
- (Possibly) accepting states, terminating (final) states, reset state

Varieties of automata

- We need to construct direct products, subautomata, define homomorphism
- Formal languages, syntactic congruence, semigroups
- Discuss difference between nondeterministic and probabilistic
- Euclidean automata (continuous input)
- Cluster automata (different timescales)
- Subregular linguistics (Rogers et al., 2013; Yli-Jyrä, 2015; Chandlee and Jardine, 2019; Rawski and Dolatian, 2020; Graf, 2022)

Automata theory looked at from far

- Operational semantics a la Plotkin/Hennessy "small step" will not be discussed
- This has more to do with the limitations of my understanding than with unworthiness of the approach
- FSTs are excellent for morphophonological computation
- Eilenberg machines will not be discussed (but see S19:5.8,6.6)
- Will discuss operational aspects for (hyper)graphs and word vectors as we go along
- These are vaguely analogous to "big step" or "natural" semantics a la Kahn, but the analogy will not be exploited

What is in LTM?

- To produce and understand language certain things need to be memorized, the central questions are *how much* and *what exactly*
- The traditional view (largely defended here) is that you need to learn the words/morphemes
- We collect the words, and whatever ancillary information seems necessary, in the dictionary, what linguists call the *lexicon*, traditionally organized in lexemes, sublexemes, occasionally sub-sublexemes
- Our interest is not so much with printed dictionaries as with the *mental* lexicon: how is it structured? Surely not alphabetically! Is it structured like a databese, with *records* and *keys*?
- We will start by looking at traditional dictionary entries

in a cave

spel·der \'speldər\ *vb* -ED/-ING/-S [obs. E *speld* to split open, spread open (fr. ME *spelden*, prob. alter. — influenced by *spelde* splinter — of *spalden* to split, spread open, fr. MLG, fr. OHG *spaltan* to split) + E *-er* (as in *batter*) — more at SPILL] *vt, chiefly Scot* : SPLIT ~ *vi, chiefly Scot* : STRETCH, SPRAWL

speld·ing \'spel(d)ən\ *n* -S [prob. fr. obs. E *speld* to split open + E *-ing*] *Scot* : STOCKFISH 1a

spel·dring *also* **spel·dron** \'speldrən\ *n* -S [*speldring* prob. fr. *spelder* + *-ing; speldron* alter. of *speldring*] *Scot* : STOCKFISH 1a

spele·o·log·i·cal *also* **spelae·o·log·i·cal** \ˌspēlēəˈläjəkəl, -pel-\ *adj* [*speleological* fr. *speleology* + *-ical; spelaeological* alter. (influenced by *spelaean*) of *speleological*] : of or relating to speleology

spele·ol·o·gist \ˌ==ˈäləjəst\ *n* -S [ISV *speleology* + *-ist*] : a specialist in speleology

spele·ol·o·gy \-jē\ *n* -ES [ISV *speleo-* (fr. L *speleum* cave, fr. Gk *spēlaion*) + *-logy;* akin to Gk *spēlunx* cave, *speos* cave, grotto] : the scientific study or systematic exploration of caves

speleo·them \'==ə,them\ *n* -S [*speleo-* (fr. L *speleum* cave) + *-them* (fr. Gk *thema* something laid down, deposit) — more at THEME] : a cave deposit or formation

spelican *var of* SPILLIKIN

spelk \'spelk\ *n* -S [ME *spelke*, fr. OE *spelc, spilc* splint; akin to ON *spjalk* splint, MD *spalke* chip, W *fflochen* splinter, Arm *p'elk* long piece of wood and prob. to OHG *spaltan* to split — more at SPILL] **1** *chiefly Scot* : SPLINTER **2** *dial Brit* : SPAR 3

¹spell \'spel\ *n* -S [ME, speech, talk, tale, fr. OE; akin to OHG *spel* tale, talk, ON *spjall*, Goth *spill* tale, talk, Gk *apeilē* boast, threat, Latvian *pal'as* rebuke, abuse] **1 a** *obs* : STORY, TALE **b** : a spoken word or set of words believed to have magic power : CHARM, INCANTATION ⟨cause death by muttering ~*s* over the young shoots of a certain tree —W.D.Wallis⟩ **c** : a state of enchantment ⟨it was the voice that cracked the ~ — that pleasant, homely, wheedling voice which brought with it daylight and common sense —John Buchan⟩ **2** : a strong compelling influence or attraction ⟨even . . . enemies were unable to resist the ~ of his presence —Alvin Redman⟩ ⟨writing under the ~ of the slavery controversy —R.A.Billington⟩

²spell \"\ *vt* **spelled** \-ld\ **spelled; spelling; spells** : to put under or as if under a spell : BEWITCH, CHARM ⟨used witchcraft all these years to ~ the ladies —Ray Bradbury⟩

³spell \"\ *vb* **spelled** \-ld,-lt\ *or chiefly Brit* **spelt** \-lt\ **spelled** *or chiefly Brit* **spelt; spelling; spells** [ME *spellen*, fr. OF *espeller*, of Gmc origin; akin to OE *spellian* to relate, talk, MHG *spellen*, ON *spialla* to talk, mention, Goth *spillon* to relate; denominative fr. the root of E ¹*spell*] *vt* **1** : to read slowly and with difficulty ⟨yourselves may ~ it yet in chronicles —Robert Browning⟩ — often used with *out* ⟨laboriously ~ out a newspaper —*Time*⟩ **2 a** : to find out by study or investigation : DISCOVER — often used with *out* ⟨~ out a God

posing a word ⟨where the dictionaries
do —*Time*⟩

spelling bee *n* : a spelling match : SPELLDOWN

spelling book *n* : a book with exercises for teaching how to spell

spelling-bound \'==ˌ=\ *adj* : deaf to or intolerant of a pronunciation because of its discrepancy with its orthographic representation ⟨too *spelling-bound* to realize that the best-educated speakers often say \'seb³m\ for *seven*⟩

spelling pronunciation *n* : pronunciation of a word in which letters or syllables are given their usual sounds in analogous situations rather than the sounds heard among speakers who make greatest use of the word ⟨\'wȯ(r)ˌsestə(r)\ instead of \'wůstə(r)\ for *Worcester*, or \'bōtˌswān\ instead of \'bōs³n\ for *boatswain* are *spelling pronunciations*⟩

spelling reform *n* : a movement to modify conventional spellings so as to lessen or remove the differences between the orthography and the pronunciation of words — compare REFORMED SPELLING

spelling school *n* : a spelling match in rural schools esp. of the 19th century often serving as the occasion for a social event

spell out *vt* : to explain or state explicitly in unmistakable terms ⟨these views will be further *spelled out* in future speeches —*Newsweek*⟩ ⟨in a brief, seemingly unambitious book, without *spelling* anything *out* . . . gets a great deal said —*Time*⟩

spells *pl of* SPELL, *pres 3d sing of* SPELL

¹spelt \'spelt\ *n* -S [ME, fr. OE, fr. LL *spelta*, of Gmc origin; akin to MHG & MD *spelte* split piece of wood, OHG *spaltan* to split; prob. fr. the splitting of the husk during threshing — more at SPILL] : a wheat (*Triticum spelta*) that is of no commercial importance in America but is grown to some extent in Germany and Switzerland and that has lax spikes with spikelets containing two light red kernels — called also *speltz;* compare EMMER

²spelt *chiefly Brit past of* SPELL

¹spel·ter \'speltə(r)\ *n* -S [prob. modif. (influenced by It *peltro* pewter) of MD *speauter* spelter — more at PEWTER] **1** : ZINC; *esp* : zinc cast in slabs for commercial use **2** : SPELTER SOLDER

²spelter \"\ *vt* -ED/-ING/-S : to solder with an alloy high in zinc

spelter solder *n* : a zinc solder (as one of three parts of zinc to four of copper) used in soldering copper, iron, and brass

spelt·oid \'spelˌtȯid\ *n* -S [*¹spelt* + *-oid*] : a variant in wheat having certain characteristics of spelt

speltz \'s(h)pelts\ *n* -ES [G *spelz* spelt, fr. OHG *spelza, spelta*, fr. LL *spelta* — more at SPELT] **1** : SPELT **2** : any of several varieties of emmer

spe·lun·car \spə'ləŋkə(r), (')spē¦l-\ *adj* [L *spelunca* cave + E *-ar*] : of or relating to a cave

spe·lunk·er \"\ *n* -S [obs. E *spelunk* cave (fr. ME, fr. MF or L) + E *-er;* MF *spelunque*, fr. L *spelunca*, fr. Gk *spēlunx* — more at SPELEOLOGY] : one who makes a hobby of exploring and studying caves : CAVER — compare SPELEOLOGIST

spe·lunk·ing \-kiŋ\ *n* -S [obs. E *spelunk* cave + E *-ing*] : the hobby or practice of exploring caves

spence \'spen(t)s\ *n* -S [ME *spence, spense*, fr. MF *despense*

The structure of the lexeme

- Pronunciation (phonology database key)
- Part of speech (syntax db key)
- Definition (semantics db key)
- Bunch of ancillary info: etymology, variants, style, topic, frequency, hyphenation ...
- Headword usually derived via orthography
- Easily extended to bilingual/multilingual
- But what to do with technical vocabulary? Millions of "words" for chemical compounds, animal species, names of people, places, organizations ...

spell 927250 spelling 666868 spells 375175 spelled 237181 spellings 51680 spelt 36573 spellbound 17346 spellbinding 14765 spelen 6823 speller 6687 spellchecker 6539 spellcheck 6059 spel 5062 spellers 4439 spelunking 4089 spellcasting 4058 speling 3722 spellbook 3550 spellcaster 3209 spellbinder 3125 spell's 3030 spellcasters 2970 speleothems 1871 speleology 1455 spelunkers 1345 spellchecking 1313 spellcraft 1126 speleological 1122 spelter 1043 spellcheckers 990 spell&quot 951 spelunker 930 spellwork 766 speleothem 754 spelljamming 683 spellchecked 652 spellen 643 speleologists 641 spellcast 601 spells&quot 598 speleo 558 spellin 550 spelar 548 spell' 486 spela 475 spelvin 432 spelspiel 378 speler 373 spellbind 359 spelende 355 spelta 329 spelling&quot 327 spell&gt 325 spellmasters 322 spelunk 315 spellman 309 spelthorne 291 spelletjes 278 spellyou 264 spellex 252 spelljammer 249 speleologist 248 spellserver 237 spells' 225 spellchk 219 spellworking 217 spellbindingly 213 spelare 209 speltoides 203 spellin' 198 spelling's 195 spelling' 195 spellout 188 speld 185 spello 183 spellbinders 182 spellmaker 180 spellchips 176 spelade 175 spellpoints 172 speleogenesis 172 spellling 169 spelld

Coverage

- Ideally, we'd want the dictionary frequency-ordered
- But high coverage remains elusive, OOV is a big problem
- Common vocabulary often used in L2 instruction → Wed (Kornai, 2021)
- It is less trivial to define than 'most frequent' we need corrected frequency (Thorndike, 1921; Füredi and Kelemen, 1989)
- Our interest is more with basic vocabulary (Ogden, 1944), Simple Wikipedia (Yasseri, Kornai, and Kertész, 2012)
- Everybody tries to build a basic list: `https://concepticon.clld.org` has 450+ sources
- Semantics (Kornai, 2019) and Vector Semantics (Kornai, 2023) discusses how the `4lang` system is built

Speleology

- *speleum* 'cave' + *ology* 'science of' = *speleology* 'science of caves'
- Yes, but what is the '+' and what is the '=' here?
- This will require both morphology/morphophonology/phonology for the '+' and semantics for the '='
- We will not look at the etymology, because the language learner does not have access to it
- But we will look at the frequencies, because the primary linguistic data naturally comes frequency-weighted
- We will also look at other standard parts of lexical entries such as labels for domain *law, medicine, biology, . . .*; for style *taboo, humorous, biblical, . . .*; for geographic distribution *in the speech of the Northerners* (read Kiparsky, 1979 for a better understanding of Pāṇini's labels)
- Syntax also adds significant material (part of speech, subcategorization frame, . . .)

What kind of science is speleology?

- Obviously, there are caves, and we deeply care about them
- But their formation is a matter of geology
- Their flora/fauna (very interesting!) is a matter of biology
- Their population is a matter of archeology
- So we don't have a unified science of speleology, all we have are theories/principles from other, more coherent theories that we try to apply/extend to caves
- Lexicography is not any different

```
irenic    ety     C19: from Greek eir*\_enikos, from eir*\_
eirenic   alt
          head    irenic
          or      eirenical
          syl     ei:ren+ic
          pron    <I1rEnik>, <-1ren->
          pos     adj.
irenic    0.
          or      irenical
          syl     i:ren+ic
          pron    <I1rEnik>, <-1ren->
          pos     adj.
          qual    Chiefly theol.
          def     tending to conciliate or promote peace.
irenically        sub
          head    irenic
          or      eirenically
          syl     i:1ren:i+cal+ly
```

Information

- Measured in **bits** and bytes
- Can be computed by Shannon's formula $H = -\Sigma_i p_i \log_2(p_i)$
- Property of distributions not individual items
- Counts the average number of the best Twenty Questions-style questions it takes to identify a particular item
- If something contains 21 bits of information, there is *no* clever girl who can get to it in 20 questions – entropy is a hard lower bound on how much space we need
- If the distribution is sufficiently uneven, average information content can stay finite even if there are infinitely many choices. Simple example of the 'CoinToss' language discussed in Resources/indra.pdf

An example: geo labels in CED

- We look at geographic labels like *in the U.S, in Canada, in the Caribbean,. . .*
- There are 118 of these. The worst idea: devote one bit to each. This would require a total of $169547 \cdot 118$ bits or 2.385MB
- A sligtly better idea: number the labels 0-118 (reserving 0 to "no geo label") and encode these numbers in 7 bits. Now we are down to $169547 \cdot 7$ bits or $0.142\text{MB} = 145\text{kB}$.
- "The emergence of the unmarked" (in the sense of Trubetskoi, 1939, more narrow than McCarthy and Prince, 1994) – don't assign a label to "no label", leave it unmarked. Now we need $753 * 7$ bits, or 659B
- Can we do better? Yes, by better coding we can bring this down to 428 bytes. Remember, we started with 2.385 megabytes.

Absolute label frequencies in CED

218 in Britain, 105 in the U.S., 68 in India, 33 in England, 30 in South Africa, 26 in Malaysia, 18 in Scotland, 14 in Canada, 13 in Anglo-Saxon England, 11 in medieval England, 11 in Australia, 9 in the U.S. and France, 9 in Britain and Germany, 8 in the Caribbean, 7 in North America, 6 in the British Isles, 6 in Ireland, 5 in the U.S. and Canada, 4 in Pakistan, India, etc., 4 in India and Pakistan, 3 in southern Africa, 3 in some states of the U.S., . . . 1 in India and the East Indies, 1 in India and Africa, 1 in England when the sovereign is male, 1 in England or Scotland, 1 in England and, formerly, Wales, 1 in England and in France before 1789, 1 in England and elsewhere, 1 in England and Wales until 1974, 1 in England and Wales from 1888 to 1974, 1 in East Africa, 1 in E Africa; as modifier, 1 in Commonwealth countries, 1 in Colonial America, 1 in Britain and certain Commonwealth countries, 1 in Britain and Ireland, 1 in Brit?!ain, 1 in Barbados, 1 in Austria, 1 in Aus?!tralia, 1 in Anglo-Saxon Britain, 1 in 19th-century Ireland, 1 in 18th-century

What was that?

- *King's Regulations* 'the code of conduct for members of the armed forces that deals with discipline, aspects of military law, etc.' **Usage**: in Britain and the Commonwealth when the sovereign is male
- *Queen's Regulations* same def, but usage: in Britain and the Commonwealth when the sovereign is female
- By rationalizing the labels, further gains could be made, but we will not go down that path
- There are only 6085 different labels used in CED, and these are unevenly distributed, so
- Total information content of labels in CED is less than 22kB
- Labels contribute only 1.02 bits for a CED entry

Getting some upper bounds

- The information content of a file can be bound (from above) by the size of its compressed version (zip, gzip, xz, . . .)
- Running English text is compressed to about 1/3 of the original file size
- The Collins English Dictionary is 27.9MB uncompressed, 6.2MB compressed
- With low bitrate encoding 1 second of speech is about 120B
- You can say about 6-8 syllables per second, so a word is about 60B
- Compare to the written form, which takes about 1.75 bits/character (Brown et al., 1992b)
- Phonemic, rather than orthograpic, could be even better
- Image format (pdf file) much worse, 80MB

Phonology

- Made easy by the fact that phonology is an advanced theory, with well defined representations (phoneme strings are good enough)
- The statistical properties of phonemes and strings of phonemes are well understood
- It is much easier to look at character entropy than phoneme entropy, since we don't have nearly as much phonemically transcribed speech as orthographically transcribed
- You can do this at home! Take a corpus, and compute the character entropy. For English (lowercased) you will get about 4.5 bits per character
- But if a word is written with 6 letters, you don't need 27 bits!
- Why? Because the character/phoneme string is redundant, knowing the phonotactics helps.

Syntax and other small fry

- The bulk of syntactic information in the lexicon is provided by Part of Speech (POS)
- In CED, this is only 0.85% of the total!
- Compare pronunciation (phonology) which is 5.3%, or syllabification (ortho or phono) which takes 9%
- Etymology (which we continue to ignore) is 8.5%
- Stylistic and other labels 4.4%
- Headword, variants, all other info 13%
- The bulk is in the definitions 48%
- The rough proportions are also evident from visual inspection of the pages

- Aronoff, Mark (1985). "Orthography and Linguistic Theory: The Syntactic Basis of Masoretic Hebrew Punctuation". In: *Language* 61.1, pp. 28–72.
- Brown, P. et al. (1992a). "An Estimate of an Upper Bound for the Entropy of English". In: *Computational Linguistics* 18.1, pp. 31–40.
- Brown, P.F. et al. (1992b). "An estimate of an upper bound for the entropy of English". In: *Computational Linguistics* 18.1, pp. 31–40. URL: https://aclanthology.org/J92-1002.pdf.
- Chandlee, Jane and Adam Jardine (Mar. 2019). "Autosegmental Input Strictly Local Functions". In: *Transactions of the Association for Computational Linguistics* 7, pp. 157–168. DOI: 10.1162/tacl_a_00260. URL: https://aclanthology.org/Q19-1010.
- Corbett, Greville G. (2000). *Number*. Cambridge University Press. ISBN: 0-521-64016-4.
- Füredi, Mihály and József Kelemen (1989). *A magyar nyelv szépprózai gyakorisági szótára*. Akadémiai Kiadó.

- Graf, Thomas (2022). "Subregular linguistics: bridging theoretical linguistics and formal grammar". In: *Theoretical Linguistics* 48.3–4, pp. 145–184. DOI: 10.1515/tl-2022-2037.
- Kiparsky, Paul (1979). *Pāṇini as a Variationist*. Cambridge and Poona: MIT Press and Poona University Press.
- Kornai, András (2007). *Mathematical linguistics*. Springer. ISBN: 978-1-84628-985-9. URL: https://kornai.com/Drafts/ml.pdf.
- — (2019). *Semantics*. Springer Verlag. ISBN: 978-3-319-65644-1. DOI: 10.1007/978-3-319-65645-8. URL: http://kornai.com/Drafts/sem.pdf.
- — (2021). "Vocabulary: Common or Basic?" In: *Frontiers in Psychology*. DOI: 10.3389/fpsyg.2021.730112. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.730112/full.
- — (2023). *Vector semantics*. Springer Verlag. DOI: 10.1007/978-981-19-5607-2. URL: http://kornai.com/Drafts/advsem.pdf.

- McCarthy, John J. and Alan Prince (1994). "The emergence of the unmarked: Optimality in prosodic morphology". In: *Proceedings of the North East Linguistics Society*. Vol. 24. URL: https://scholarworks.umass.edu/linguist_faculty_pubs/18.
- Ogden, C.K. (1944). *Basic English: a general introduction with rules and grammar*. K. Paul, Trench, Trubner.
- Rawski, Jonathan and Hossep Dolatian (2020). "Multi-Input Strict Local Functions for Tonal Phonology". In: *Proceedings of the Society for Computation in Linguistics*. Vol. 3. 25.
- Rogers, James et al. (2013). "Cognitive and sub-regular complexity". In: *Formal Grammar*. Vol. 8036. Lecture Notes in Computer Science. Springer, pp. 90–108.
- Ross, Malcolm D. (1988). *Proto Oceanic and the Austronesian Languages of Western Melanesia*. Pacific Linguistics, series C, no. 98. Department of Linguistics, Research School of Pacific Studies, Australian National University.

- Thorndike, Edward L. (1921). *The teacher's word book*. New York Teachers College, Columbia University.
- Trubetskoi, Nikolai Sergeevich (1939). *Grundzüge der Phonologie*. Göttingen: Vandenhoeck and Ruprecht.
- Yasseri, Taha, András Kornai, and János Kertész (2012). "A practical approach to language complexity: a Wikipedia case study". In: *PLoS ONE* 7.11. DOI: e48386.doi:10.1371/journal.pone.0048386.
- Yli-Jyrä, Anssi (2015). "Three Equivalent Codes for Autosegmental Representations". In: *Proceedings of the 12th International Conference on Finite-State Methods and Natural Language Processing 2015 FSMNLP 2015 Düsseldorf*.