Vector Semantics: Lecture 1

András Kornai
SZTAKI Computer Science Research Institute

13 February 2024

Major themes

- What is semantics?
- Vectors, matrixes
- Ordinary language philosophy V2.0: aims at ordinary (non-technical) language, but heavy use of technical tools
- From linguistics (morphology, lexicography)
- From math (algebra, information theory)
- Logic content: default logic, probabilities, negation, naive implication
- Discussion of big issues: can LLMs have consciousness? are they intelligent? are they safe?

Metainfo

- **Time/Place** Tuesday 12:15-1:45 pm CET/ELTE Logic Dept
- **Course webpage, slack**
  https://nessie.ilab.sztaki.hu/~kornai/2024/VectorSemantics/
  https://vectorsemantics.slack.com
- **Prerequisites** None
- **Requirements** regular attendance (max. 3 classes can be skipped), good knowledge and understanding of the main text, and active participation in discussions. Participants need to write 2 short essays (3-3 pages) responsive to the assigned tasks. By the end of the semester they need to submit another essay (minimum 4 pages) which relies substantively on at least one chapter of the book and one of the assigned readings. The readings will be assigned in consultation with the professor, and will be in English. Over the semester a total of three essays need to be submitted. Grades are based on classroom activity (20%) and the results of these essays (20%+20%+40%).

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 SPRAWL] *vt, chiefly Scot* : SPLIT ~ *vi, chiefly Scot* : STRETCH, SPILL

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*

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

Speleology

- *speleum* 'cave' + *ology* 'science of' = *speleology* 'science of caves'
- Yes, but what is the '+' and what is the '=' here?
- We will not look at the etymology because the ordinary speaker does not have access to it
- But we will look at the frequencies
- We will do a fair bit of linguistic analysis
- Starting with a crash course in morphology

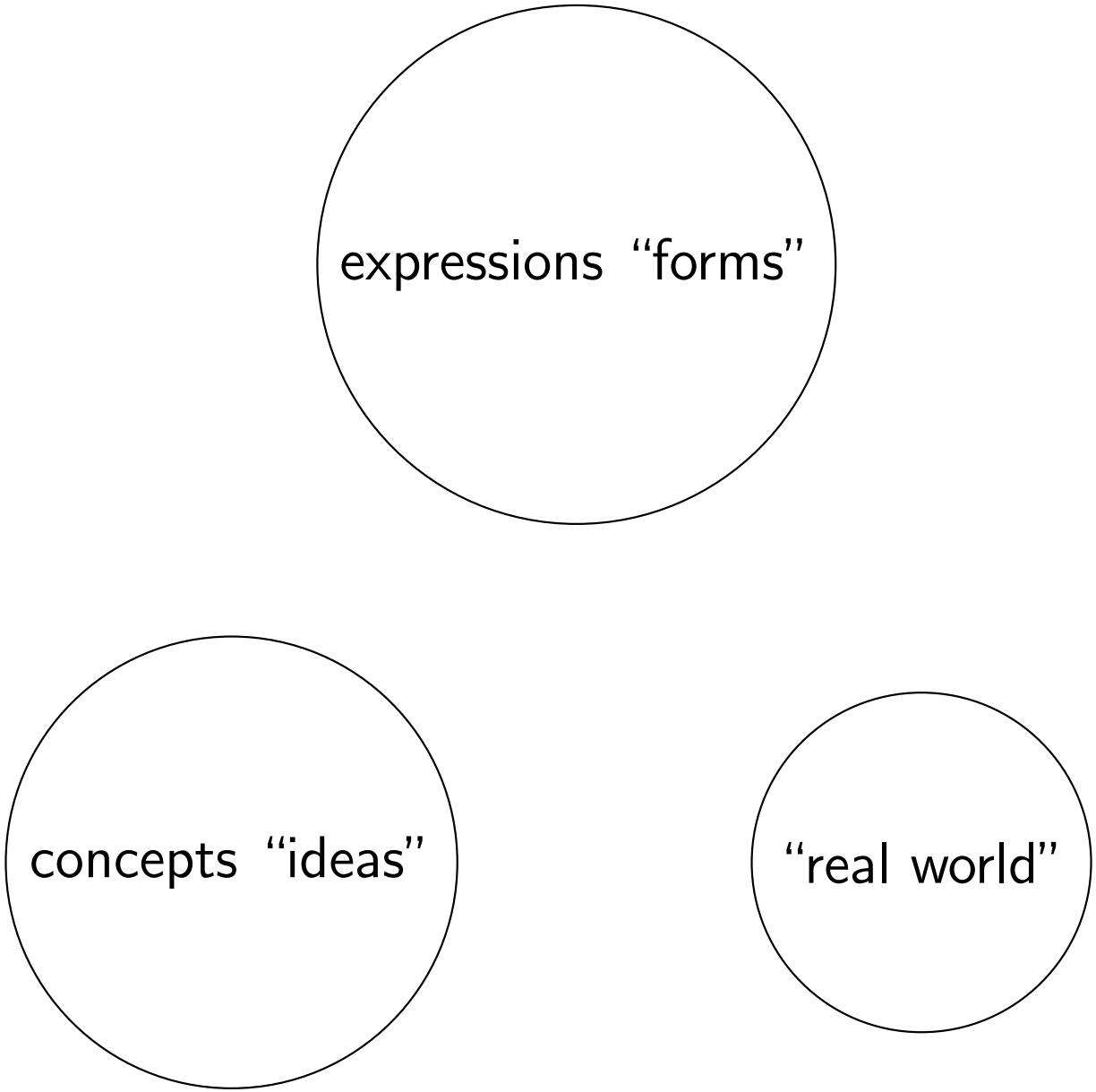
What kind of science is speleology?

- 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
- Semantics is not any different

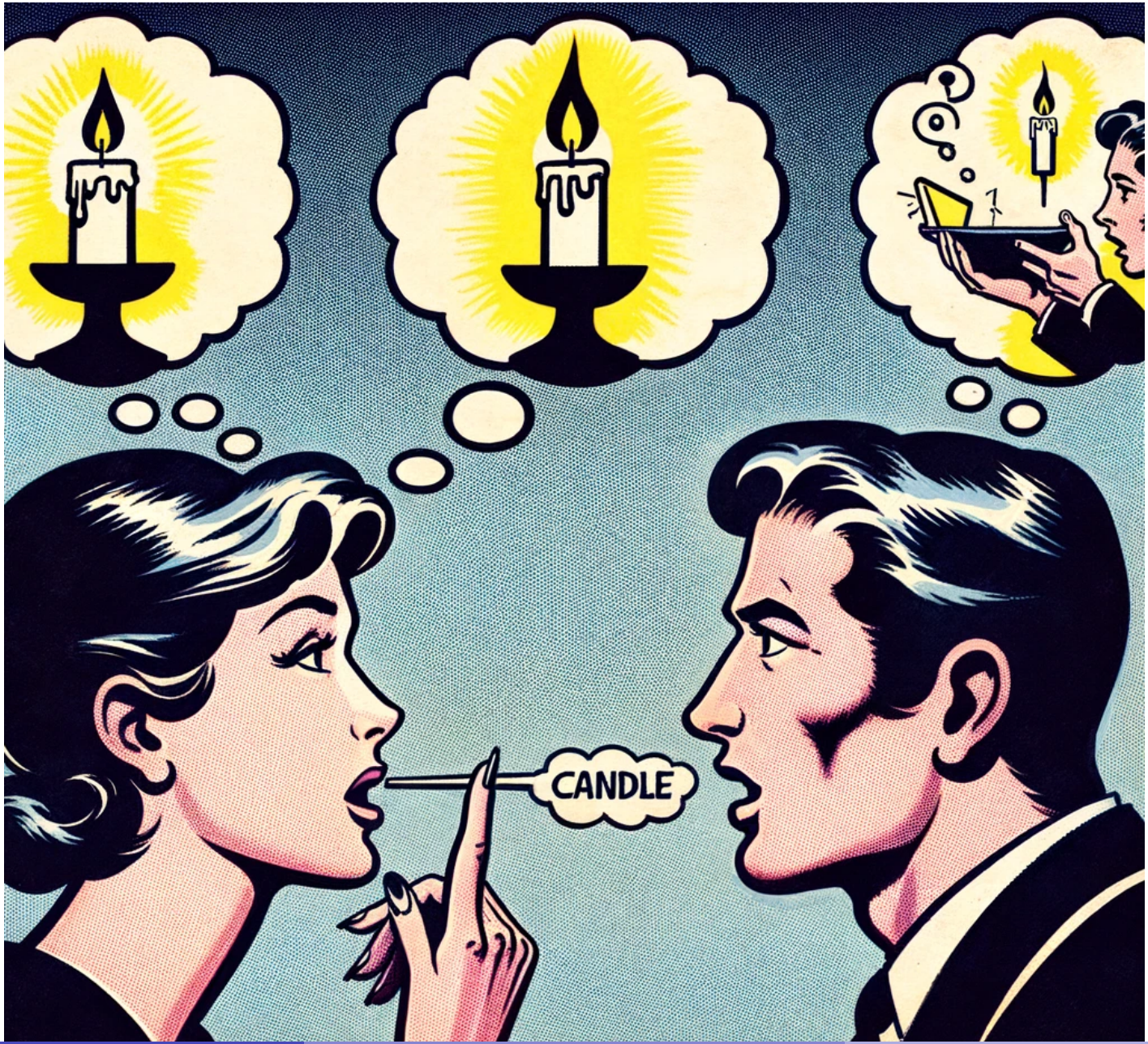
What kind of science is semantics?

- Obviously people talk to each other, and can understand each other well enough to cooperate
- Or go to war when the communication breaks down. The stakes are high!
- We will throw everything at the problem: logic, statistics, math, computer science, linguistics, semiotics, cognitive science, philosophy . . .
- And see what sticks – whatever works, works, the rest goes on the back burner
- The approach taken here is *irenic* and *syncretic*
- It will also be *bottom up* rather than *top down*

The overall model

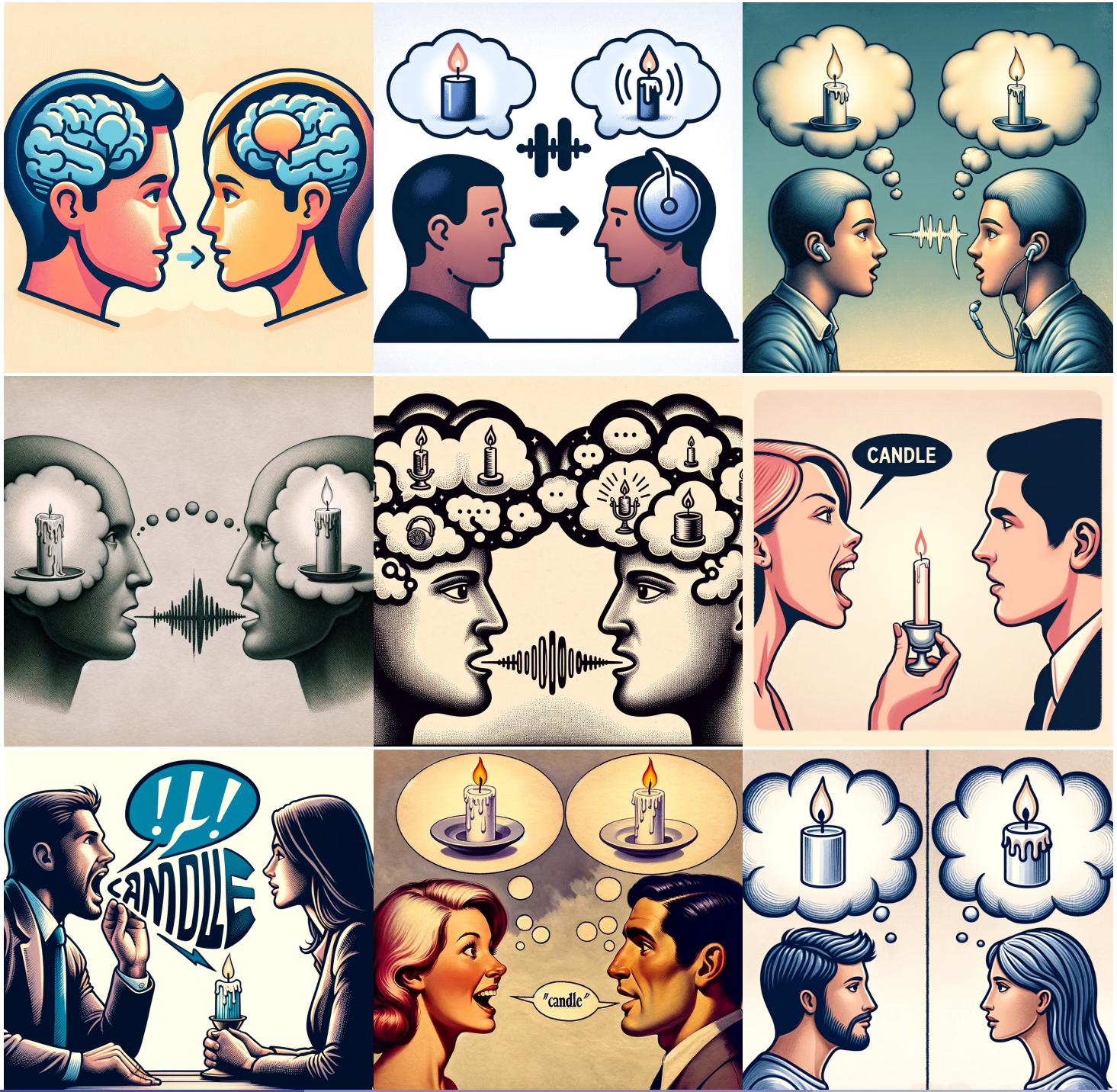


Communication



The prompt

Please create a sequence of three images: the first one should show a woman thinking about a candle. The second should show the woman saying the word "CANDLE" to the man by means of enclosing the text CANDLE inside a text bubble coming from her mouth. The third image should show the man thinking about a different candle.


CANDLE
!!!
CANDLE
"candle"

irenic *or* **ei·renic** \(')ī'renik, -rēn-\ *also* **ireni·cal** \-nəkəl\ *adj* [Gk *eirēnikos*, fr. *eirēnē* peace (prob. of non-IE origin) + *-ikos* -ic, -ical] : conducive to or operating toward peace, moderation, harmony, and conciliation and away from contention and partisanship esp. among disputants ⟨~ measures⟩ ⟨~ without being namby-pamby —*Chicago Theol. Seminary Register*⟩ ⟨the viewpoint is ~ and the author seeks to show the best features of each religion and church in turn —N.K. Burger⟩ **syn** see PACIFIC

ireni·cal·ly \-nək(ə)lē\ *adv* : in an irenic manner : in a way

syn·cret·ic \(')si|n|'kred·ik, sə|n|'k-, |ŋ|\ *adj* [*syncretism* + *-ic*] **1** : characterized or brought about by syncretism : aiming at or making for syncretism : SYNCRETISTIC ⟨~ religious sect⟩ **2** : having absorbed the functions of one or more other grammatical cases ⟨the Latin ablative is a ~ case⟩

syn·cre·tion \sən'krēshən, səŋ'k-\ *n* -s [*syncretic* + *-ion*] : an instance of syncretism : act of syncretizing

syn·cre·tism \'siŋkrə,tizəm, 'sink-\ *n* -s [NL *syncretismus*, fr. Gk *synkrētismos* federation of Cretan cities, fr. *synkrētizein* to unite against a common enemy] **1** : the reconciliation or union of conflicting (as religious) beliefs or an effort intending such; *specif* : a movement of a Lutheran party in the 17th

```
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
```

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

Morphology

- Morphemes: minimal signs
- Words: minimal free forms
- Roots, stems, fully formed words
- Lexemes
- Clitics, phonological v written words
- Word meaning

Information

- Measured in **bits**
- 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

Information content of sentences

Number of possible binary trees over n words is

$$C_n \sim 4^n / n^{1.5} \sqrt{\pi}$$

or $<$ 2 bits per word (see also Aronoff, 1985)
The word entropy H of a language with Zipf constant B is given by

$$H \approx H_k + \frac{1 - P_k}{\log(2)}(B/(B-1) - \log(B-1) + \log(k) - \log(1 - P_k))$$

which yields 12.67 bits for English, 15.41 bits for Hungarian.
Other definitions of logical structure than by parse tree are possible but they do not alter the picture significantly: logical structure accounts for at most 15% of the information.

The linguistic sign

- **The classic model** (Ferdinand de Saussure 1916)
  A two-part structure composed of *form* and *meaning*
- **The modern model** (Kracht, 2003)
  A three-part structure composed of *form*, *category* and *meaning*
- `4lang`

```
mark    jel     nota    znak    shirushi 印     biao1ji4 标记   1182    u       N       sign, visible   %
mark_   jelöl   nota    znak    shirusu 標す    biao1zhi4 标志  3331    u       V       =agt[sign], =pat[meaning], represent    %
```

Coming attractions/essay topics

- Word frequency distributions, Zipf's Law
- n-dimensional space
- Hypernode graphs
- The syntax of `4lang` formulas
- Basic units of meaning: are there any?
- Compositionality, semicompositionality
- Homonymy/polysemy

Aronoff, Mark (1985). "Orthography and Linguistic Theory: The Syntactic Basis of Masoretic Hebrew Punctuation". In: *Language* 61.1, pp. 28–72.

Kracht, Marcus (2003). *The Mathematics of Language*. Berlin: Mouton de Gruyter.