Lexicography for semanticists

András Kornai
SZTAKI Computer Science Research Institute

25 April 2024

The central question

- To produce and understand language certain things need to be memorized
- The central questions are *how much* and *what exactly* needs to be memorized
- The traditional view (largely defended here) is that you need to learn the words
- We collect the words, and whatever ancillary information seems necessary, in the dictionary, what linguists call the *lexicon*
- 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*?

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.

Linguistics and KR

- We will look both at the practice of linguistics/lexicography (dictionaries) and at the practice of AI/Knowledge Representation (databases)
- Three major dictionary types: monolingual, bilingual, frequency-based
- Traditionally organized in lexemes, sublexemes, occasionally sub-sublexemes
- We will start by looking at traditional dictionary entries
- We will assume a simple 'telementation' model of communication whereby speaker has an idea, speaker says something, hearer hears this, understands it, now hearer has the idea

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)ȯn\ *n* -S [prob. fr. obs. E *speld* to split open + E *-ing*] *Scot* : STOCKFISH 1a

spel·dring *also* **spel·dron** \'speldrȯn\ *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 *ffloch*en 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 differ 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 \'seb'm\ 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 \'wu̇stə(r)\ for *Worcester*, or \'bōtˌswān\ instead of \'bōs'n\ 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 [¹*spelt* + *-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 (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, 2023b) 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, . . .)

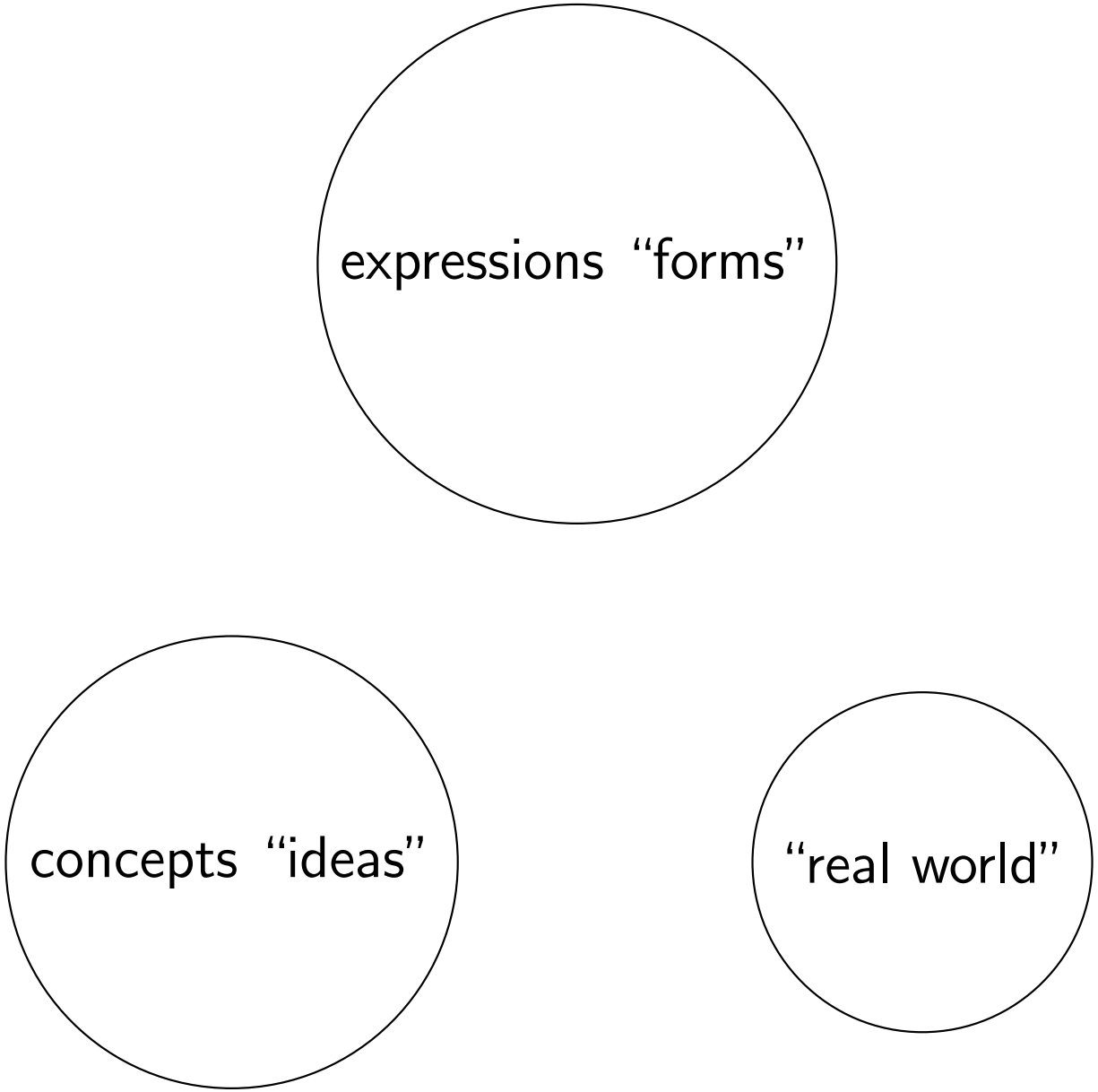
What kind of science is speleology?

- 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

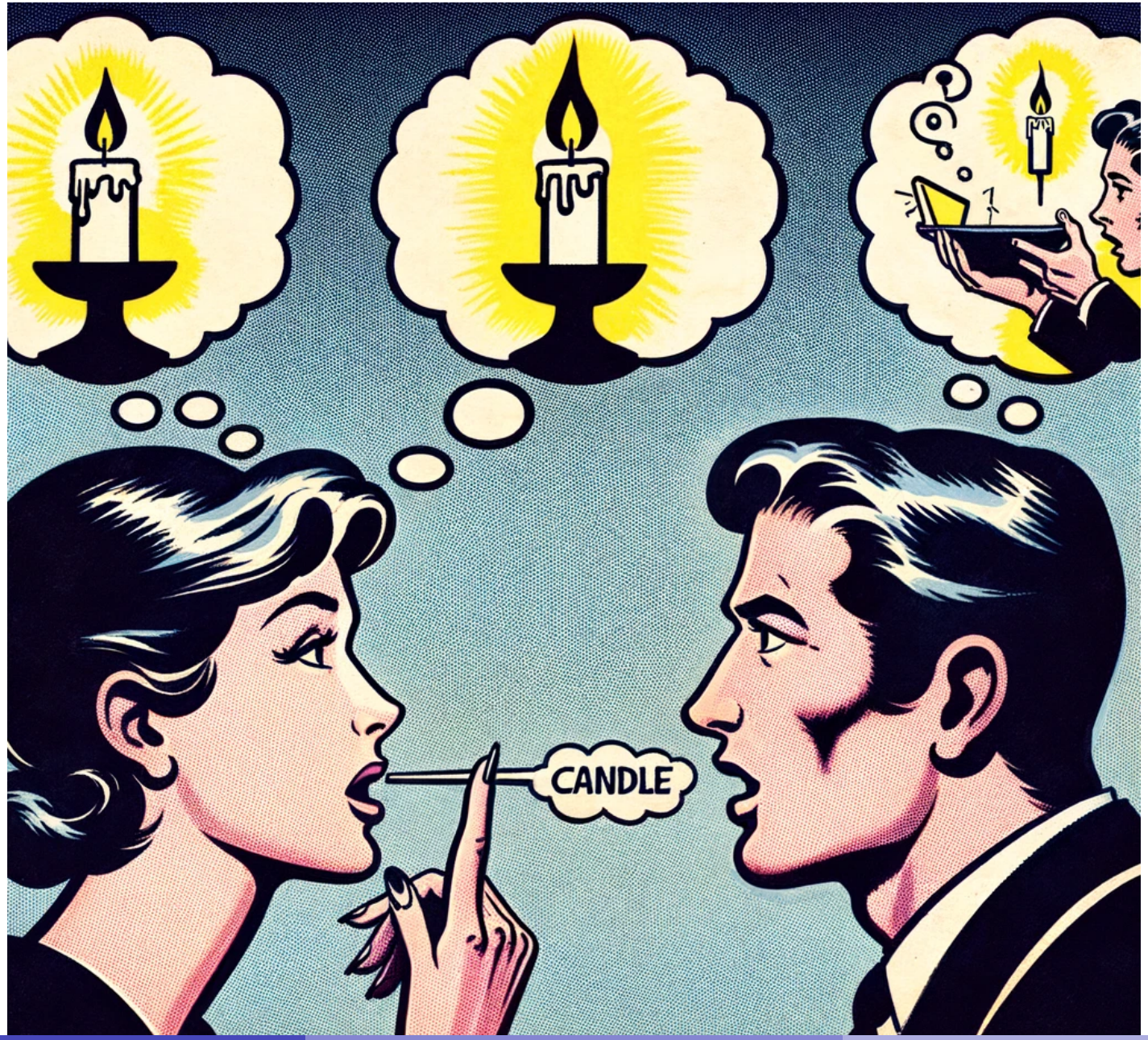
What kind of science is semantics?

- 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

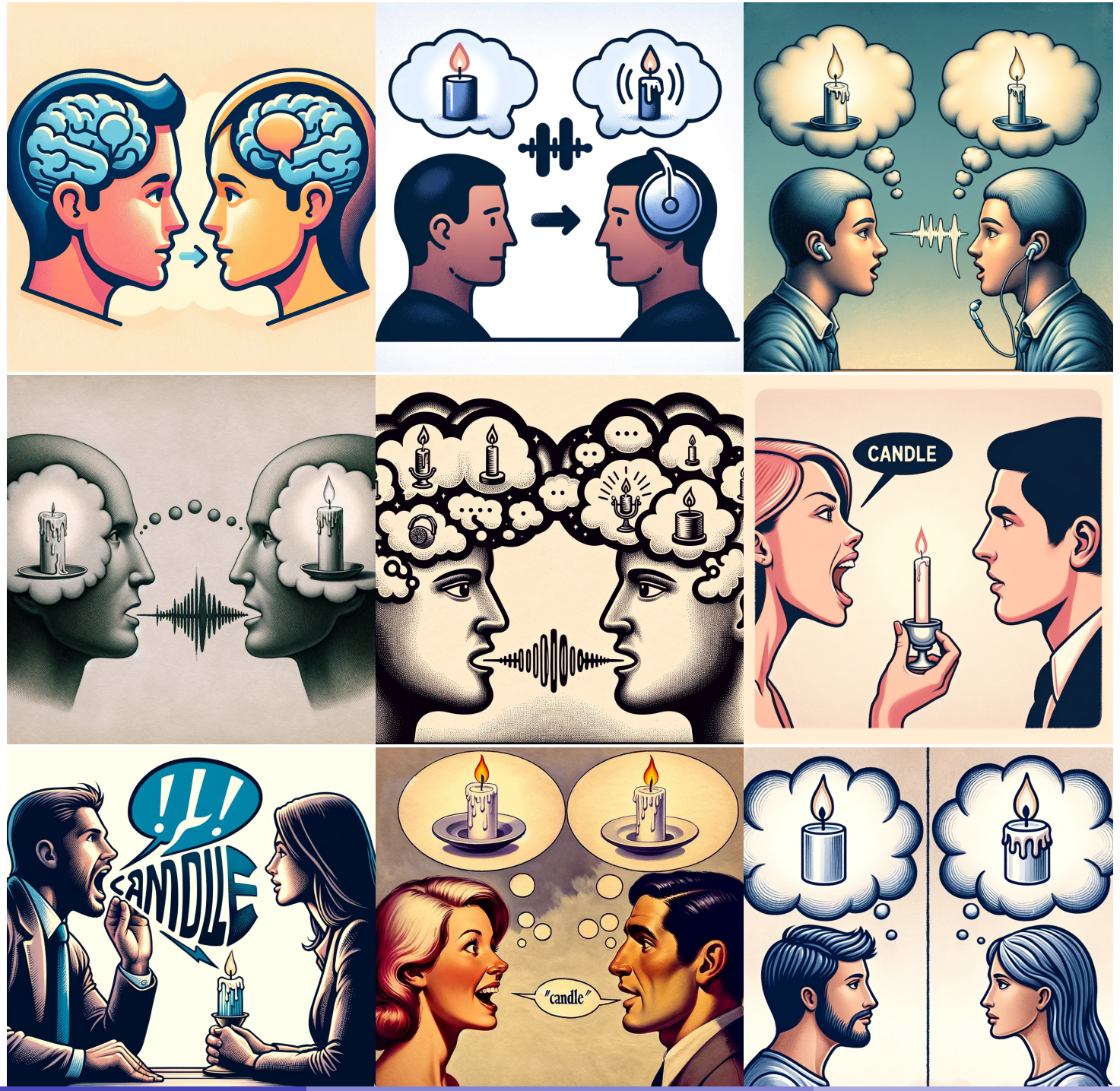


Telementation



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** \(ˈ)ī¦renik, -rēn-\ *also* **ireni·cal** \-nȯkəl\ *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 \-nȯk(ə)lē\ *adv* : in an irenic manner : in a way

syn·cret·ic \(ˈ)si|n|ˈkred·ik, sə|n|ˈk-, |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 \sȯnˈkrēshən, sȯŋˈk-\ *n* -s [*syncretic* + *-ion*] : an instance of syncretism : act of syncretizing

syn·cre·tism \ˈsiŋkrəˌtizəm, ˈsink-\ *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
```

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 0.142MB = 145kB.
- "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.

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 https://nessie.ilab.sztaki.hu/~kornai/2024/VectorSemantics/Resour

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., 1992)
- 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

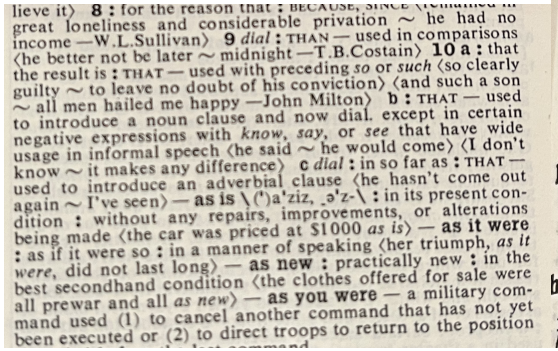
The relative weight of syntax and semantics

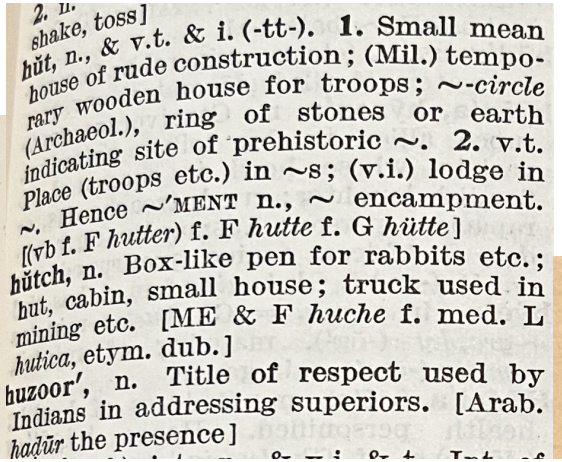
- Straw poll: what is the relative weight of syntax to semantics?
- Based on the amount of information that must be stored, semantics is more than 50 times more important than syntax
- This confirms the habit of traditional (pre-20th century) linguistics of devoting the bulk of the discussion to morphology and putting syntax in a small chapter
- In running text, word entropy is 12-16 bits/word, syntax contributes less than 2 bits/word (see Kornai, 2019 Ch 1.3)
- So syntax is somewhere between 0.8% and 12% of the whole story
- An estimate based on core vocabulary suggests 1.55%

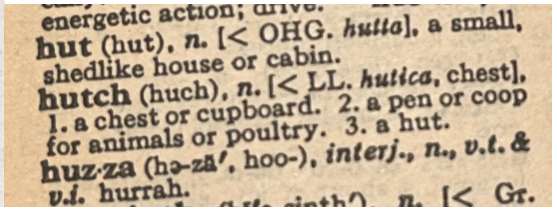
Lexicon or encyclopedia

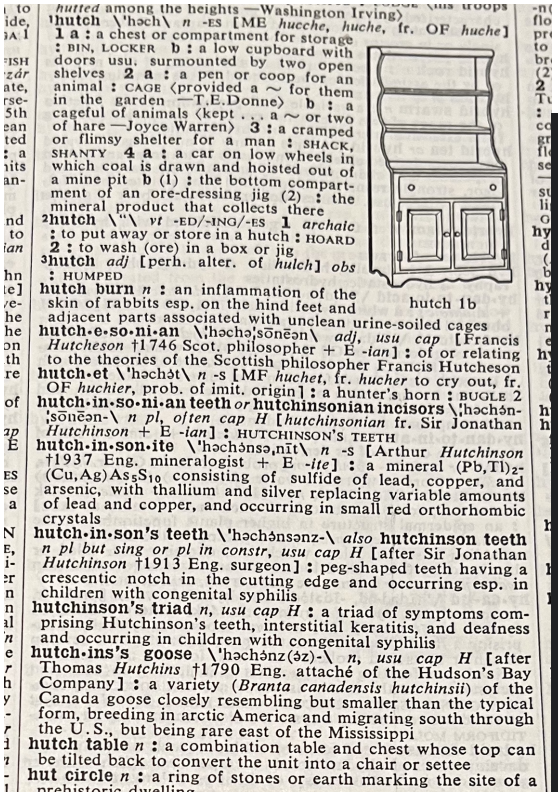
- In many topics, technical vocabulary is key
- Proper names and named entities
- PER, LOC, ORG – hundreds of millions of entries in each category
- *hutch for sale, as is*

HUTCH, AS IS

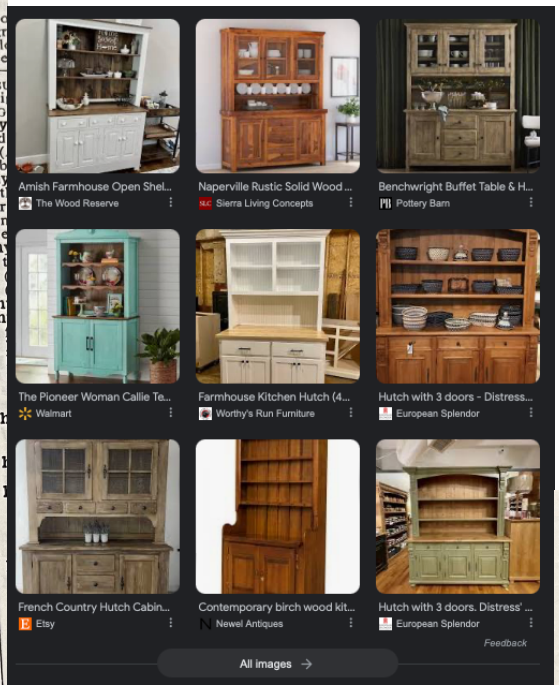

lieve it⟩ **8 :** for the reason that : BECAUSE, SINCE ⟨remained in great loneliness and considerable privation ~ he had no income —W.L.Sullivan⟩ **9** *dial* : THAN — used in comparisons ⟨he better not be later ~ midnight —T.B.Costain⟩ **10 a :** that the result is : THAT — used with preceding *so* or *such* ⟨so clearly guilty ~ to leave no doubt of his conviction⟩ ⟨and such a son ~ all men hailed me happy —John Milton⟩ **b :** THAT — used to introduce a noun clause and now dial. except in certain negative expressions with *know, say,* or *see* that have wide usage in informal speech ⟨he said ~ he would come⟩ ⟨I don't know ~ it makes any difference⟩ **c** *dial* : in so far as : THAT — used to introduce an adverbial clause ⟨he hasn't come out again ~ I've seen⟩ — **as is** \(ˈ)aˈziz, ˌaˈz-\ : in its present condition : without any repairs, improvements, or alterations being made ⟨the car was priced at $1000 *as is*⟩ — **as it were** : as if it were so : in a manner of speaking ⟨her triumph, *as it were,* did not last long⟩ — **as new** : practically new : in the best secondhand condition ⟨the clothes offered for sale were all prewar and all *as new*⟩ — **as you were** — a military command used (1) to cancel another command that has not yet been executed or (2) to direct troops to return to the position

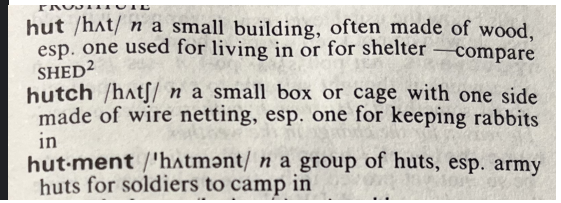

2. n.
shake, toss]
hŭt, n., & v.t. & i. (-tt-). **1.** Small mean house of rude construction; (Mil.) temporary wooden house for troops; ~-*circle* (Archaeol.), ring of stones or earth indicating site of prehistoric ~. **2.** v.t. Place (troops etc.) in ~s; (v.i.) lodge in ~. Hence ~'MENT n., ~ encampment. [(∇b f. F *hutter*) f. F *hutte* f. G *hütte*]
hŭtch, n. Box-like pen for rabbits etc.; hut, cabin, small house; truck used in mining etc. [ME & F *huche* f. med. L *hutica*, etym. dub.]
huzoor', n. Title of respect used by Indians in addressing superiors. [Arab. *hadūr* the presence]


energetic action; drive.
hut (hut), *n.* [< OHG. *hutta*], a small, shedlike house or cabin.
hutch (huch), *n.* [< LL. *hutica*, chest], 1. a chest or cupboard. 2. a pen or coop for animals or poultry. 3. a hut.
huz·za (hə-zä', hoo-), *interj., n., v.t. & v.i.* hurrah.


hutted among the heights —Washington Irving⟩
¹**hutch** \ˈhəch\ *n* -ES [ME *hucche, huche,* fr. OF *huche*] **1 a :** a chest or compartment for storage : BIN, LOCKER **b :** a low cupboard with doors usu. surmounted by two open shelves **2 a :** a pen or coop for an animal : CAGE ⟨provided a ~ for them in the garden —T.E.Donne⟩ **b :** a cageful of animals ⟨kept . . . a ~ or two of hare —Joyce Warren⟩ **3 :** a cramped or flimsy shelter for a man : SHACK, SHANTY **4 a :** a car on low wheels in which coal is drawn and hoisted out of a mine pit **b** (1) : the bottom compartment of an ore-dressing jig (2) : the mineral product that collects there
²**hutch** \"\ *vt* -ED/-ING/-ES **1** *archaic* : to put away or store in a hutch : HOARD **2 :** to wash (ore) in a box or jig
³**hutch** *adj* [perh. alter. of *hulch*] *obs* : HUMPED
hutch burn *n* : an inflammation of the skin of rabbits esp. on the hind feet and adjacent parts associated with unclean urine-soiled cages
hutch·e·so·ni·an \ˌhəchəˈsōnēən\ *adj, usu cap* [Francis *Hutcheson* †1746 Scot. philosopher + E *-ian*] : of or relating to the theories of the Scottish philosopher Francis Hutcheson
hutch·et \ˈhəchət\ *n* -S [MF *huchet,* fr. *hucher* to cry out, fr. OF *huchier,* prob. of imit. origin] : a hunter's horn : BUGLE 2
hutch·in·so·ni·an teeth *or* **hutchinsonian incisors** \ˌhəchən-ˈsōnēən-\ *n pl, often cap H* [*hutchinsonian* fr. Sir Jonathan *Hutchinson* + E *-ian*] : HUTCHINSON'S TEETH
hutch·in·son·ite \ˈhəchənsəˌnīt\ *n* -S [Arthur *Hutchinson* †1937 Eng. mineralogist + E *-ite*] : a mineral $(Pb,Tl)_2(Cu,Ag)As_5S_{10}$ consisting of sulfide of lead, copper, and arsenic, with thallium and silver replacing variable amounts of lead and copper, and occurring in small red orthorhombic crystals
hutch·in·son's teeth \ˈhəchənsənz-\ *also* **hutchinson teeth** *n pl but sing or pl in constr, usu cap H* [after Sir Jonathan *Hutchinson* †1913 Eng. surgeon] : peg-shaped teeth having a crescentic notch in the cutting edge and occurring esp. in children with congenital syphilis
hutchinson's triad *n, usu cap H* : a triad of symptoms comprising Hutchinson's teeth, interstitial keratitis, and deafness and occurring in children with congenital syphilis
hutch·ins's goose \ˈhəchənz(ȥz)-\ *n, usu cap H* [after Thomas *Hutchins* †1790 Eng. attaché of the Hudson's Bay Company] : a variety (*Branta canadensis hutchinsii*) of the Canada goose closely resembling but smaller than the typical form, breeding in arctic America and migrating south through the U.S., but being rare east of the Mississippi
hutch table *n* : a combination table and chest whose top can be tilted back to convert the unit into a chair or settee
hut circle *n* : a ring of stones or earth marking the site of a prehistoric dwelling

hutch 1b




PROSTITUTE
hut /hʌt/ *n* a small building, often made of wood, esp. one used for living in or for shelter —compare SHED²
hutch /hʌtʃ/ *n* a small box or cage with one side made of wire netting, esp. one for keeping rabbits in
hut·ment /ˈhʌtmənt/ *n* a group of huts, esp. army huts for soldiers to camp in

General principles

- Universality – system should work the same for all languages
- Reductivity – can't define the simple by the more (or just equally) complex *speltz* 'any of several varieties of emmer'

  Suppose I make you a gift of a large sum of money saying you can collect it from Titius; Titius sends you to Caius; and Caius, to Maevius; if you continue to be sent like this from one person to another you will never receive anything (Leibniz, quoted in Wierzbicka (1985))

- No encyclopedic knowledge
- OK, but where to draw the line? We keep only *essential* properties

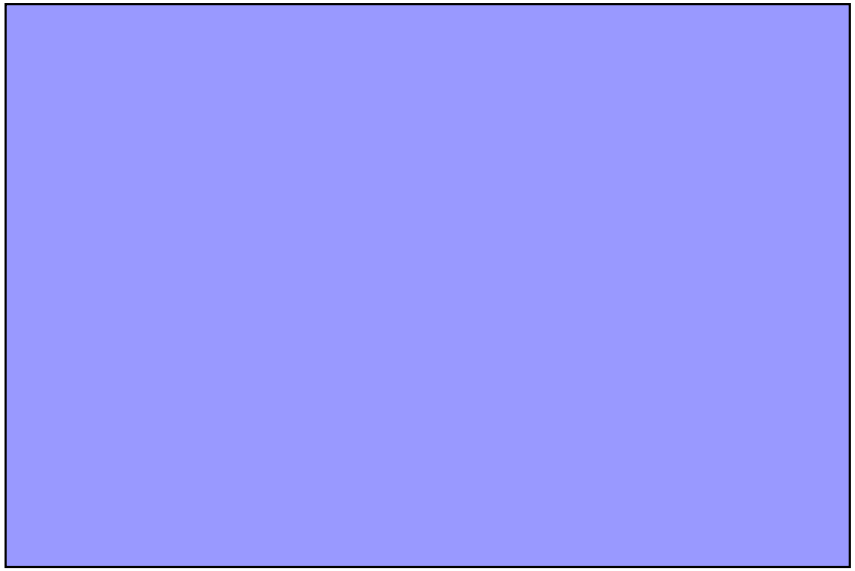
Lexical entries

- There are disjoint lexical entries (for words and morphemes) called *lexemes*
- These overwhelmingly correspond to traditional dictionary entries
- In dictionary databases, these used to be the *records*
- But these are not subdivided into *fields* as in typeset dictionaries or dictionary databases
- Rather, they are associative networks with *spreading activation* (Quillian, 1967; Collins and Loftus, 1975; Carroll, 1983)
- Phonology done by autosegmental representations (Goldsmith, 1976)
- Can be viewed as automata (Eilenberg machines)
- Can also be viewed as vectors

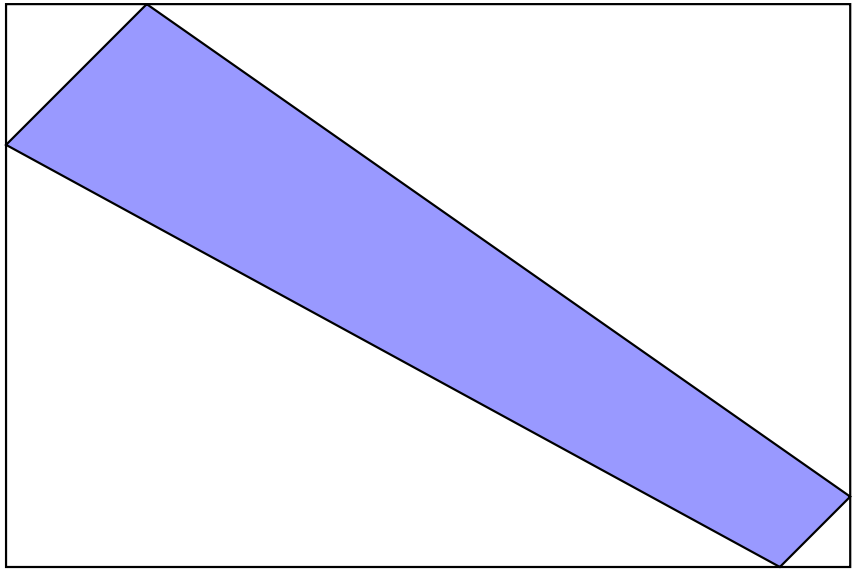
Lexical entries cont'd

- Stylistic and other labels by ultradense subspaces (Rothe, Ebert, and Schütze, 2016; Dufter and Schütze, 2019)
- We have the technology for etymology (diachronic phonological rules are just as easy by automata as synchronic rules) but kids don't have the data
- In addition to traditional lexemes (words, stems) we also have lexical entries for bound morphemes (roots, affixes)
- Morphology has non-compositial semantics, but we can deal with this
- Lexicon also contains *conceptual schemas* (Schank and Abelson, 1977)
- OK, but what about syntax? We use *constructions* (Fillmore and Kay, 1997)
- Traditional concerns of syntacticians are addressed via a sparse system of linkers (thematic roles/deep cases/kārakas) (Kiparsky, 1987; Butt, 2006)

Semicompositionality as subdirect decomposition



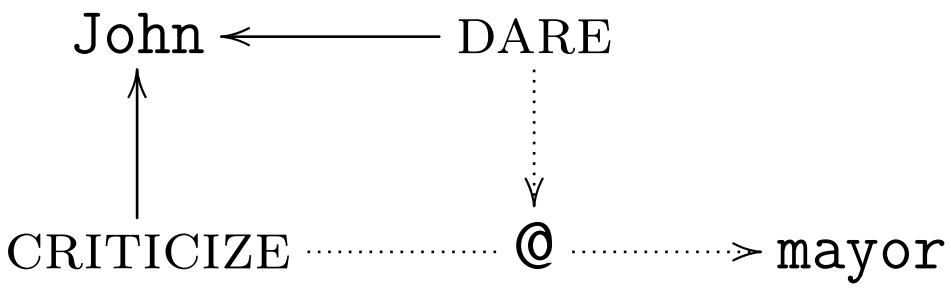
Direct product



Subdirect product

(Figure from Kornai, 2023b Ch. 2.2, but the idea goes back at least to Kiparsky, 1982 on noun-noun compounding)

Hypernode graphs at work

- There are only two linkers: =agt (subject, nominative case) and =pat (object, accusative case) [Doing this right in ergative languages is not hard, but will not be discussed here]
- 

John dared to criticize the mayor
- Hypernodes are S-V-O triples as in RDF – becaue you don't need indirect objects, themes, goals, etc. (Kornai, 2012) you don't need hyperedges
- John DARE {John CRITICIZE mayor}
- Unification operates silently in the background to make sure the two *John*s are the same

The "commercial exchange" schema

buy =agt .	goods =pat .
value<money> .	sell =agt .

Figure: exchange_

Gen 25:29-34

29 And Jacob sod pottage: and Esau came from the field, and he was faint:

30 And Esau said to Jacob, Feed me, I pray thee, with that same red pottage; for I am faint: therefore was his name called Edom.

31 And Jacob said, Sell me this day thy birthright.

32 And Esau said, Behold, I am at the point to die: and what profit shall this birthright do to me?

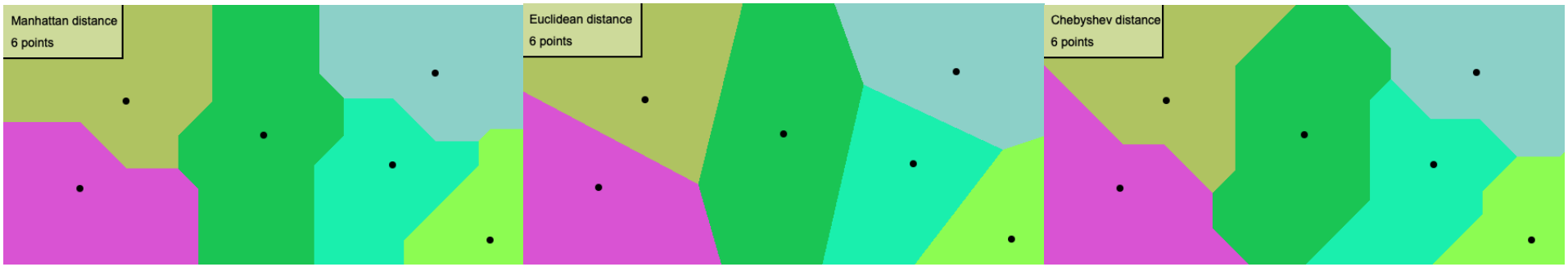
33 And Jacob said, Swear to me this day; and he sware unto him: and he sold his birthright unto Jacob.

34 Then Jacob gave Esau bread and pottage of lentiles; and he did eat and drink, and rose up, and went his way: thus Esau despised his birthright.

The linguistic takeway

- This whole story makes no sense whatsoever unless we interpret it in the `exchange_` frame
- It takes a great deal of work to link up the language with the schema: only in 31, *sell me thy birthright* is the schema triggered
- This already requires sophisticated anaphor resolution: we need to know that Esau will be *seller*, Jacob will be *buyer*. Morphology for nomen agentis *-er/3627* `stem_-er is_a =agt, "_ -er" mark_ stem_`
- *sell* `=agt cause_ buyer has =pat, buyer cause_ =agt has money_, dative_ mark_ buyer`
- We need to figure out the other two slots, that the `goods` are the birthright, and the `thing of value` is the bowl of lentils.
- The mechanism is coercive, it is not that Esau IsA seller, Esau is *the* seller, `Jacob buy birthright` is a valid inference.

Summary

- Aristotle had it largely right: for Knowledge Representation genus/differentia specifica is all you need!
- You can't do full KR with this machinery, in particular you can't do $\nabla \cdot \mathbf{B} = 0$. You need to learn to live with this limitation (easy to say to a linguist, physicists may be upset)
- But you can implement the whole thing in neural networks. All you need are vectors and matrixes, but not higher tensors
- Bonus: you get a decent learning theory
- Special bonus for lexicographers: vector semantics gets you a theory of homonymy versus polysemy (Kornai, 2023a, only in Hungarian)
- 

Brown, P.F. et al. (1992). "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.

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

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.

Carroll, John A. (1983). *An island parsing interpreter for the full augmented transition network formalism*. ACL Proceedings, First European Conference, pp. 101–105.

Collins, A.M. and E.F. Loftus (1975). "A spreading-activation theory of semantic processing". In: *Psychological Review* 82, pp. 407–428. DOI: 10.1037/0033-295X.82.6.407.

Dufter, Philipp and Hinrich Schütze (2019). *Analytical Methods for Interpretable Ultradense Word Embeddings*. arXiv: 1904.08654. URL: https://arxiv.org/pdf/1904.08654.pdf.

Fillmore, Charles and Paul Kay (1997). *Berkeley Construction Grammar*. URL: http://www.icsi.berkeley.edu/%5C~%7B%7Dkay/bcg/ConGram.html.

Füredi, Mihály and József Kelemen (1989). *A magyar nyelv szépprózai gyakoriségi szótára*. Akadémiai Kiadó.

Goldsmith, John A. (1976). *Autosegmental Phonology*. PhD thesis MIT.

Kiparsky, Paul (1979). *Pāṇini as a Variationist*. Cambridge and Poona: MIT Press and Poona University Press.

— (1982). "From cyclic phonology to lexical phonology". In: *The structure of phonological representations, I*. Ed. by H. van der Hulst and N. Smith. Dordrecht: Foris, pp. 131–175.

— (1987). *Morphosyntax*. Stanford University: ms.

Kornai, András (2012). "Eliminating ditransitives". In: *Revised and Selected Papers from the 15th and 16th Formal Grammar Conferences*. Ed. by Ph. de Groote and M-J Nederhof. LNCS 7395. Springer, pp. 243–261. DOI: 10.1007/978-3-642-32024-8_16.

Kornai, András (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.

— (2023a). "Poliszémia politópokkal". In: *Általános Nyelvészeti Tanulmányok* 35. Ed. by Beáta Gyuris, pp. 311–326. ISSN: HU 0569-1338.

— (2023b). *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.

Quillian, M. Ross (1967). "Semantic memory". In: *Semantic information processing*. Ed. by Minsky. Cambridge: MIT Press, pp. 227–270.

Rothe, Sascha, Sebastian Ebert, and Hinrich Schütze (June 2016). "Ultradense Word Embeddings by Orthogonal Transformation". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 767–777. arXiv: 1602.07572 [cs.CL]. URL: http://www.aclweb.org/anthology/N16-1091.

Schank, Roger C. and Robert P. Abelson (1977). *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Hillsdale, NJ: Lawrence Erlbaum.

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.

Wierzbicka, Anna (1985). *Lexicography and conceptual analysis*. Ann Arbor: Karoma.

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.