Vector Semantics: Lecture 3

András Kornai
SZTAKI Computer Science Research Institute

20 March 2024

Main takeaways from studying frequency

- Zipf and Herdan laws work well Kornai, 1999a; Kornai, 2002
- It's all about information Brown et al., 1992
- Entropy measures maximum average compressibility over the wire (Jelinek, 1997) ML Ch 7.1
- Kolmogorov complexity can achive more compression, but only semi-computable (Li and Vitányi, 1997; Vitanyi and Li, 2000) ML Ch 7.3
- Minimum Description Length philosophy used in linguistics (Borbély and Kornai, 2019)
- With a twist: amortizing the universal component (Kornai, Zséder, and Recski, 2013)
- Engineering takeaway (1): do the frequent things first
- Engineering takeaway (2): OOV is a persistent problem

Words

- Minimum free forms (Bloomfield, 1926)
- Phonological words: units between pauses
- (Orthographical words)
- Lexemes (also called lexical entries) can be MWEs like *as is*
- Subentries (Kornai, 2023)
- Subsubentries csinál

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, . . .

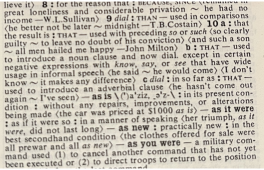
Coverage

- 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' (Thorndike, 1921), corrected frequency
- 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

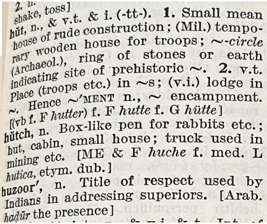
Lexicon or encyclopedia

- 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, ˌəˈz-\ **:** 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

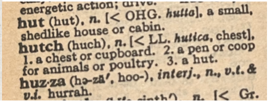


2. n. shake, toss]

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. [(vb 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]

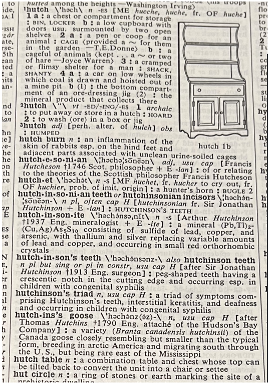


energetic action; drive.

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⟩

¹hutch \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

hutch 1b

²hutch \"\ *vt* -ED/-ING/-ES **1** *archaic* **:** to put away or store in a hutch **:** HOARD **2 :** to wash (ore) in a box or jig

³hutch *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 \ˌhəchəˌsōnēən\ *adj, usu cap* [Francis *Hutcheson* †1746 Scot. philosopher + E *-ian*] **:** of or relating to the theories of the Scottish philosopher Francis Hutcheson

hutch·et \ˈhəchə̇t\ *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** \ˌhəchə̇n-ˌsōnēən-\ *n pl, often cap H* [*hutchinsonian* fr. Sir Jonathan *Hutchinson* + E *-ian*] **:** HUTCHINSON'S TEETH

hutch·in·son·ite \ˈhəchə̇nsəˌnīt\ *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

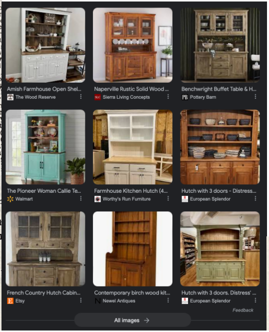
hutch·in·son's teeth \ˈhəchə̇nsənz-\ *also* **hutchinson teeth** *n pl but sing or pl in constr, usu cap H* [after Sir Jonathan *Hutchinson* †1913 Eng. surgeon] **:** peg-shaped teeth having a crescentic notch in the cutting edge and occurring esp. in children with congenital syphilis

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 \ˈhəchə̇nz(ə̇z)-\ *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



Amish Farmhouse Open Shel...
The Wood Reserve

Naperville Rustic Solid Wood ...
Sierra Living Concepts

Benchwright Buffet Table & H...
Pottery Barn

The Pioneer Woman Callie Te...
Walmart

Farmhouse Kitchen Hutch (4...
Worthy's Run Furniture

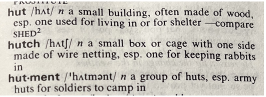
Hutch with 3 doors - Distress...
European Splendor

French Country Hutch Cabin...
Etsy

Contemporary birch wood kit...
Newel Antiques

Hutch with 3 doors. Distress' ...
European Splendor

Feedback
All images →



hut /hʌt/ *n* a small building, often made of wood, esp. one used for living in or for shelter —compare SHED^2

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
- Reductivity
- No encyclopedic knowledge
- Read VS Ch. 1.2

- Bloomfield, Leonard (1926). “A set of postulates for the science of language”. In: *Language* 2, pp. 153–164.
- Borbély, Gábor and András Kornai (June 2019). “Sentence Length”. In: *Proceedings of the 16th Meeting on the Mathematics of Language*. Toronto, Canada: Association for Computational Linguistics, pp. 114–125. URL: https://www.aclweb.org/anthology/W19-5710.
- Brown, P. et al. (1992). “An Estimate of an Upper Bound for the Entropy of English”. In: *Computational Linguistics* **18**/*1*, pp. 31–40.
- Jelinek, Frederick (1997). *Statistical Methods for Speech Recognition*. MIT Press.
- Kornai, András (2002). “How many words are there?” In: *Glottometrics* 2.4, pp. 61–86.
- — (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). “Poliszémia politópokkal”. In: *Általános Nyelvészeti Tanulmányok* 35. Ed. by Beáta Gyuris, pp. 311–326. ISSN: HU 0569-1338.
- — (1999a). “Zipf’s law outside the middle range”. In: *Proceedings of the Sixth Meeting on Mathematics of Language*. Ed. by J. Rogers. University of Central Florida, pp. 347–356.
- Kornai, András, Attila Zséder, and Gábor Recski (2013). “Structure Learning in Weighted Languages”. In: *Proceedings of the 13th Meeting on the Mathematics of Language (MoL 13)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 72–82. URL: `http://www.aclweb.org/anthology/W13-3008`.
- Li, Ming and Paul Vitányi (1997). *An Introduction to Kolmogorov Complexity and Its Applications*. Springer.
- Ogden, C.K. (1944). *Basic English: a general introduction with rules and grammar*. K. Paul, Trench, Trubner.
- Thorndike, Edward L. (1921). *The teacher’s word book*. New York Teachers College, Columbia University.

- Vitanyi, Paul M. B. and Ming Li (2000). "Minimum description length induction, Bayesianism, and Kolmogorov complexity". In: *IEEE Transactions on Information Theory* 46.2, pp. 446–464.
- 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.