Vector Semantics: Lecture 4

András Kornai
SZTAKI Computer Science Research Institute

3 April 2024

Building the 4lang dictionary

- We start with the dictionary, because it stores more information than the grammar
- True at the sentence level: one word contributes 12-16 bits from the lexicon, maybe 2 bits from syntax
- Even more extreme ratio in terms of static information: dictionary at 1MB, universal grammar less than 1kb (Principles and Parameters)
- We will not have much to say about pronunciation (this has no place in a dictionary aiming at universality), we will use language-specific bindings
- Originally 4 (hence the name 4lang) as the goal was to cover major European language families, with one Germanic (English), one Slavic (Polish), one FinnoUgric (Hungarian), and one Romance (Latin) sample
- By now, manual bindings added for Chinese (Huba Bartos) and Japanese (Laszlo Cseresnyesi), and machine-generated for 30+

Syntax in 4lang

- There will be lots of it, but we delay discussion
- No lexical level support (POS) because that would again conflict with universality
- Lexemes are largely root-like (categoryless)
- Syntax is mostly in 'functional structure' rather than 'constituent structure'
- Most relevant precursors are Pāṇini, Tesnière, Fillmore, Bresnan, Perlmutter
- We use `https://universaldependencies.org` quite a bit

Reductivity

- There are limitations in what we can do
- 4lang is not good for technical vocabulary
- Numbers are already a problem (this is a feature, not a bug)
- But we can do ordinary dictionary words
- First, we reduce large dictionaries to smaller ones (good computational project)
- Next we reduce these to a small defining vocabulary (we used LDV, 2,200 words)
- Next we looked for uroboros set in LDV (currently 770 entries)

Have we missed something?

- Input: any word in any language. First find English definition. Start with German *schlagfertig* and find translation *quick-witted* (as opposed to literal translation *ready to hit*)
- Reduce this definition to core vocabulary by repeated substitution *quick-witted* is clearly *quick.wit.ed* (note lack of **witted*) and the morphology will supply 'has quick(-)wit', cf. *triangle-shaped, bite-sized, able-bodied, baby-faced, big-hearted, well-intentioned, ...* (total of 168 candidates in LDOCE).
- In this case we are lucky: LDOCE already has *quick-witted* 'able to think and understand things quickly', but what if we are not so lucky?
- In that case, we have to work on *has*, *quick*, and *wit* separately. Of these, `has` and `quick` are already in 4lang, with definitions '=agt control =pat, =agt has =pat' and 'act in short(time)' respectively
- But *wit* is missing!

Substitution *salva veritate*

- By definition of *has*, we obtain '=agt control {quick wit}, agent has {quick wit}. Substituting the definition of quick, we obtain =agt control {wit, wit act in short(time)}, agent has {wit, wit act in short(time)}
- Unification is automatic (unless blocked by other). But we (a) haven't quite gotten rid of has (and we won't, it's a primitive!) and (b) still need to get away from *wit*.
- less surprisingly than for *quick-witted*, LDOCE also has *wit* (*quickwitted* is #2716299 on the Google frequency list, *wit* is #14661) 'the ability to say things that are clever and amusing'
- So now we substitute this to obtain =agt has {ability to say thing, thing is_a clever, thing is_a amusing, say in short(time)},...
- thing, short, say and time are in 4lang

Substitution con't

- We still need *clever* and *amusing*, but LDOCE has these, and uses only LDV in their definition

```
clever able to learn and understand things quickly
clever able to use your intelligence to get what
you want, especially in a slightly dishonest way
clever skilful at doing a particular thing
clever done or made in an unusual or interesting
way that is very effective
amusing funny and entertaining
```

- So we can go on, getting things defined one by one until everything is in the uroboros core
- People can learn how to produce 4lang definitions surprisingly fast
- Machines have a guarantee of reductivity to the core

Another example

- We can't do numbers in general, but we can do some number names
- Example: *quatre-vingt-trois* is defined as *quatre-vingt* + *trois*. We assume *quatre* is defined as '4', *trois* as '3', and that these are looked up in the lexicon just as *vingt* needs to be.
- Now we have a design choice: either we treat *vingt* as a primitive '20' , or we somehow know that it comes, via L. *viginti*, from PIE "two" **dwóh*$_1$ and "ten" **dékm̥tis*
- In the former case we have '4 20 3', and by knowing 80, 81, 84 etc we know that the second blank is to be replaced by '+'. Now we have '4 20' *quatre-vingt* **plus** *trois*. But what about the first blank?
- How do we know it's multiplication? There is no **trois-vingt*
- 'Elsewhere' principle: the specific entry *soixante* overrides the rule-derived **trois-vingt*. Idea as old as grammar, see Kiparsky, 1973

Is this hard?

- Everybody tries to build a basic list:
  `https://concepticon.clld.org` has 450+ sources
- I don't know of any other one that is actually reductive
- The best of breed is NSM (Natural Semantic Metalanguage)
  only 60+ primitives
- But the syntax is not fully defined, and no reductivity guarantee
- One would need to define all 4lang primitives by NSM primitives
  and they'd be done

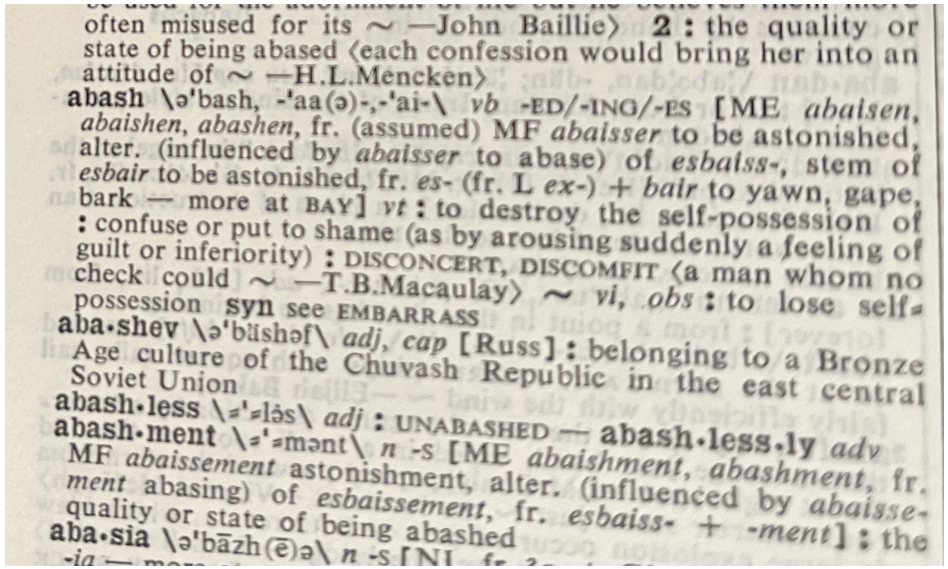
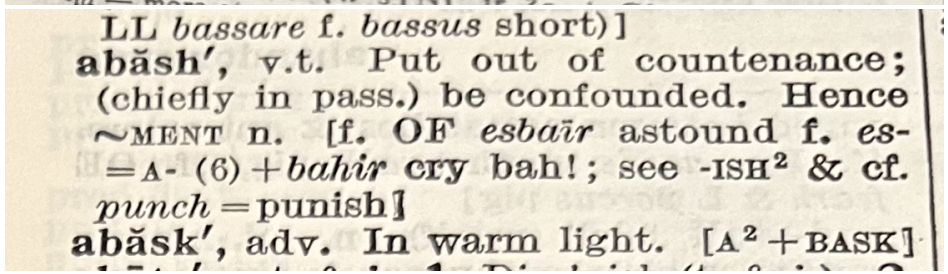
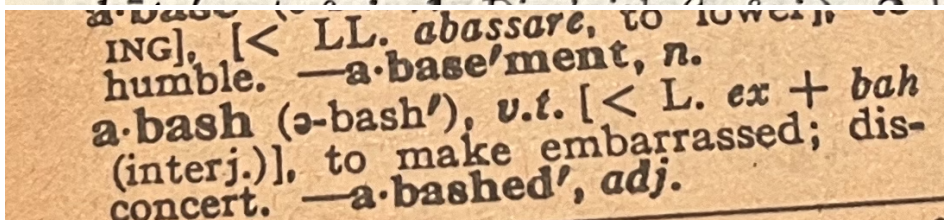
Homework excercise

- Claim: the reduction covers words in the Collins English Dictionary `Resources/ced`
- Brute force: `cut -f1 ced|sort -u|egrep ' ^[a-z]* $' > cedhead` (headwords in CED)
- `comm -23 cedhead longhead >cednotlong` But that's over 60k words!
- Let's check them out. An early instance is *abernethy* 'a crisp unleavened biscuit, C19: perhaps named after Dr. John Abernethy (1764-1831), English surgeon interested in diet'
- What really matters is the defining vocabulary of CED. Since *crisp*, *unleavened* and *biscuit* are there in Longman we are still good

Homework excercise cont'd

- Parse `Resources/ced` to find all and only definitions. Hint: you don't need to write a parser, all you need is a linux one-liner (maybe `awk` helps but it is not required)
- The CED defining vocabulary is <30k words, of which only 13k don't appear as a Longman headword
- This list is dominated by word forms like *abandoning abandonment abandons* .... Clearly, you need to run it through a morphological analyzer
- Download SFST, install EMOR, run it on the 13k words
- Profit!

Pick a word, any word (or MWE)

- /Volumes/114/Language/English/Dic/CED/Stat/uncov
- abash ‘make (someone) feel embarrassed, disconcerted, or ashamed’
- 
often misused for its ~ —John Baillie⟩ **2 :** the quality or state of being abased ⟨each confession would bring her into an attitude of ~ —H.L.Mencken⟩
abash \ə'bash, -'aa(ə)-,-'ai-\ *vb* -ED/-ING/-ES [ME *abaisen, abaishen, abashen,* fr. (assumed) MF *abaisser* to be astonished, alter. (influenced by *abaisser* to abase) of *esbaiss-*, stem of *esbair* to be astonished, fr. *es-* (fr. L *ex-*) + *bair* to yawn, gape, bark — more at BAY] *vt* **:** to destroy the self-possession of **:** confuse or put to shame (as by arousing suddenly a feeling of guilt or inferiority) **:** DISCONCERT, DISCOMFIT ⟨a man whom no check could ~ —T.B.Macaulay⟩ ~ *vi, obs* **:** to lose self-possession **syn** see EMBARRASS
aba·shev \ə'bäshəf\ *adj, cap* [Russ] **:** belonging to a Bronze Age culture of the Chuvash Republic in the east central Soviet Union
abash·less \ə'-ləs\ *adj* **:** UNABASHED — **abash·less·ly** *adv*
abash·ment \ə'-mənt\ *n* -S [ME *abaishment, abashment,* fr. MF *abaissement* astonishment, alter. (influenced by *abaissement* abasing) of *esbaissement,* fr. *esbaiss-* + *-ment*] **:** the quality or state of being abashed
aba·sia \ə'bāzh(ē)ə\ *n* -S [NL, fr.
- 
LL *bassare* f. *bassus* short)]
abăsh′, v.t. Put out of countenance; (chiefly in pass.) be confounded. Hence ~MENT n. [f. OF *esbaīr* astound f. *es-* = A- (6) + *bahir* cry bah!; see -ISH² & cf. *punch* = punish]
abăsk′, adv. In warm light. [A² + BASK]
- 
ING], [< LL. *abassare*, to lower]. —
humble. —**a·base′ment,** *n.*
a·bash (ə-bash′), *v.t.* [< L. *ex* + *bah* (interj.)], to make embarrassed; disconcert. —**a·bashed′,** *adj.*


Kiparsky, Paul (1973). "'Elsewhere' in Phonology". In: *A Festschrift for Morris Halle*. Ed. by Stephen R. Anderson and Paul Kiparsky. New York: Holt, Rinehart, and Winston, pp. 93–106.