Outline

1. **The problem** (wordfinding)
2. **Goal :** enhance an existing electronic dictionary to allow for finding quickly and naturally the elusive word.
3. **Analysis of the problem**
   - speech errors, perception, ...
4. **Solutions**
   - **by others**
   - **my own proposal** (roadmap)

My concern

Language production

- speaking
- writing

Some facts

Spontanenous speech

- **fast** (3-5 words per second)
- quite **robust** and **reliable** (few mistakes)

Underlying process

- remarkably **efficient**
- search in a **huge** lexical data-base (50.-100.000 words), brain

Questions

1° How is this **possible** (online processing), i.e. how does our brain manage?

2° Can we achieve something similar via a computer (off-line processing; dictionary consultation)?

- speed
- accuracy
- success in wordfinding

Questions

3° Why do we have problems?

4° Can we draw on the **mental lexicon** to improve the **electronic dictionaries** of tomorrow?

- If not, why so?
- If yes, on what specific aspects

The 3 principal steps

idea

concepts

form

abstract words/lemma
syntactic category
morphology

sound

phonemes
graphemes

The mice are dancing.

The normal situation
a cascaded flow of information

Yet, consider the following
(too often overlooked) **facts**

It is not because something is **stored** that it can readily be **accessed**

- people (amnesia, anomia, TOT, etc.)
- machines

Can you name these objects?

Navigational instrument

sextant

Instrument used in Asia for eating

chopsticks

Hat of a bishop

mitre

Example : name of a person

Name of **actor**

Film: Silence of the lambs

Role : Hannibal Lecter

Name actor : ???

First name : Anthony

Look for actors whose first name is ‘Anthony’

Anthony Quinn?
Perkins?
Hopkins?

Work on the TOT-phenomenon revealing what people know

Parts of the meaning

- mocha: coffee beverage flavored with milk, sugar, and cocoa

Relations to other concepts or words (associations)

- Mocha : town and port in southern Yemen at the red sea
- Starbucks: place where this beverage is served

Work on the TOT-phenomenon revealing what people know

Information concerning the form of the target word

a) number of syllables
- first and last syllable (bathtub effet)

b) grammatical information
- part of speech
- gender
- colloquial expression

c) origine (eg. Greek, latin)

d) target word: when presented a list containing the target word they will recognize it immediately and unmistakingly.

Hence,

Let's use it, and start from there.

Question : how?

Access should be based on what?

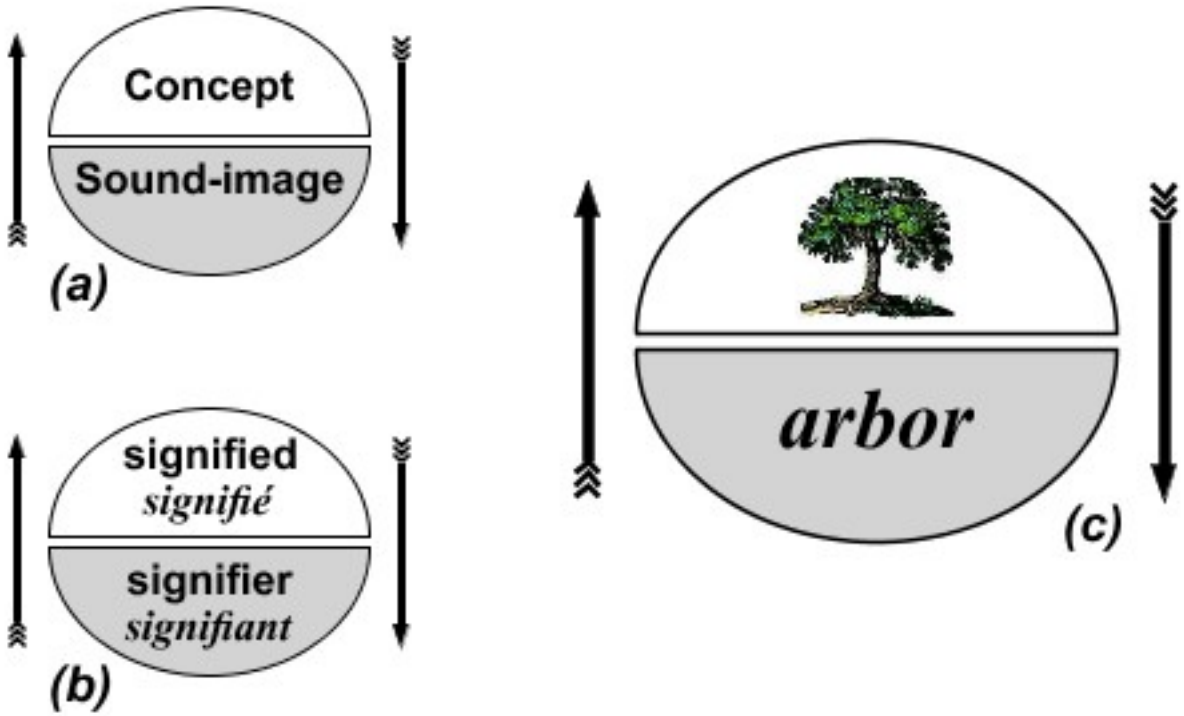
1. meanings (or meaning elements) of the target word
2. concepts or words **related** to the target word

- lexical relations (synonyme, antonyme, hypo/hypernym,…)
- associations

Before doing that...

Let's try to get a clearer picture about the nature of the problem.

Where is the problem?



Saussure's conception of the 'sign'

Why do we have word-access problems, or, what happens when we are in this state?

Words in books and in the brain are fundamentally different.

- in books they exist as tokens (meaning and forms are represented together)
- in the brain they are decomposed. The elements representing meaning, form, sound are distributed over various layers. They need to be activated (not accessed). Yet activation takes time and is error prone. Actually one may question the very fact of symbolic representations in our mind.

We do not have access to all the relevant elements (meaning, form, sound) at the same time or when needed, which might hinder unification.

Analogy: while you may see the *eyes*, *ears* and *nose* you don't see the *entire face*.

Evidence

1. TOT (we do know **fragments** of the word)
2. Speech **errors** at the **different levels**

- semantics : take the first to the left (target: right)
- syntax : I make the kettle on (targets: make some tea + put the kettle on)
- morphology : slicely thinned (target: thinly sliced)
- sound/phonology : histerical (target: historical)

Why do we have word-access problems, or, what happens when we are in this state?

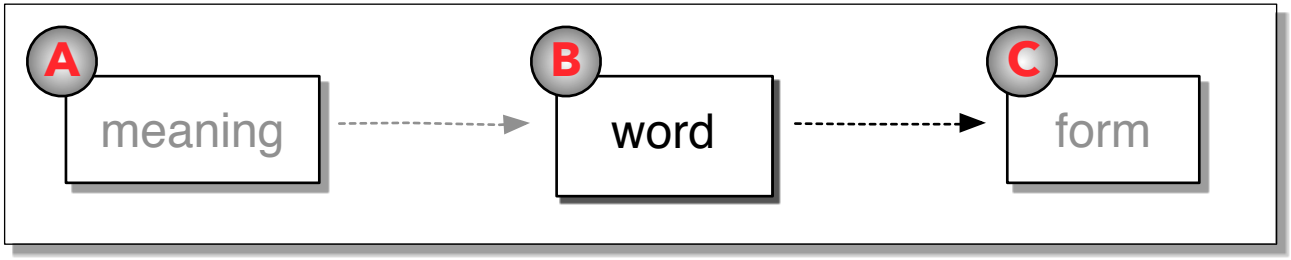
1. Competition between the elements at the various levels
2. Similarity between certain elements, hence potential danger of interference and telescoping of information
3. Activation is gradual and relative rather than absolute (all or nothing). For example, we say: it's on the tip of my tongue

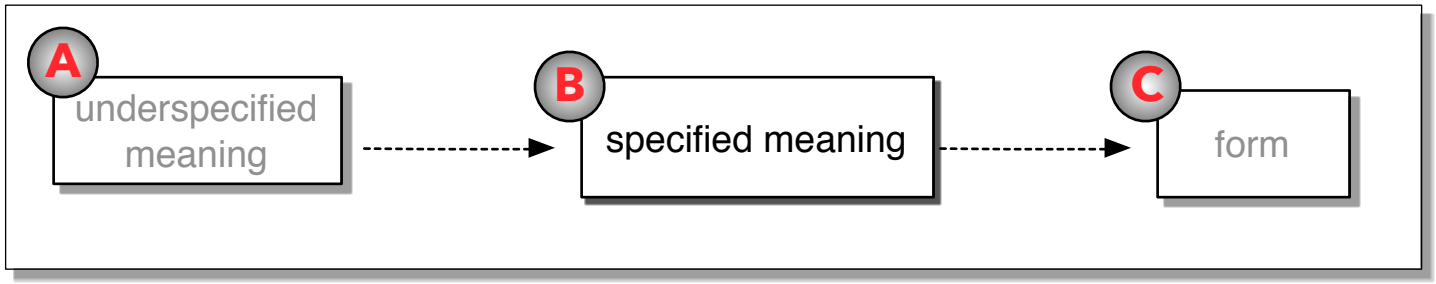
Fragmentary knowledge
puzzles apparently waiting for completion

Important note:

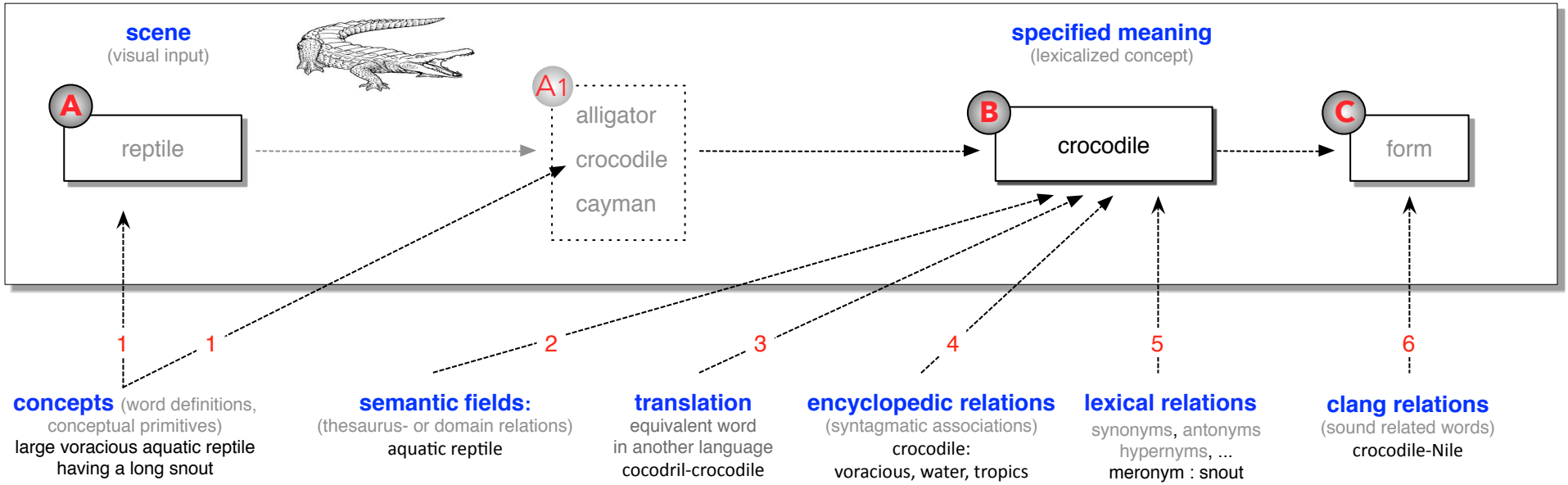
- Words and concepts are fundamentally different.
- We (generally) don't think in terms of words, but rather in terms of concepts.
- If we thought in terms of words we would never experience a word-access problems, we would just use the string representing both our ideas (concepts) and words. Yet, this is not quite what we observe in natural settings (spontaneous speech).

From mind to mouth:
the progressive synthesis of what most of us call a word





From mind to mouth:
the progressive synthesis of what most of us call a word



Idea (intention of communication) - expression

Idea :

request

```
(make drawing_of
make drawing, you
make drawing, for me)
```

Expression : Will you draw me a sheep!

The problem of finding the (rootform) of words

Input

Will you draw me a 

Meaning

woolly usually horned ruminant mammal related to the goat

Semantic candidates

mutton, ram, ewe, lamb, sheep, goat, bovid, ovis

Phonological candidates

cheap, jeep, schliep, seep, sheep, sleep, steep, streep, sweep

Output

/ʃiːp/ - sheep

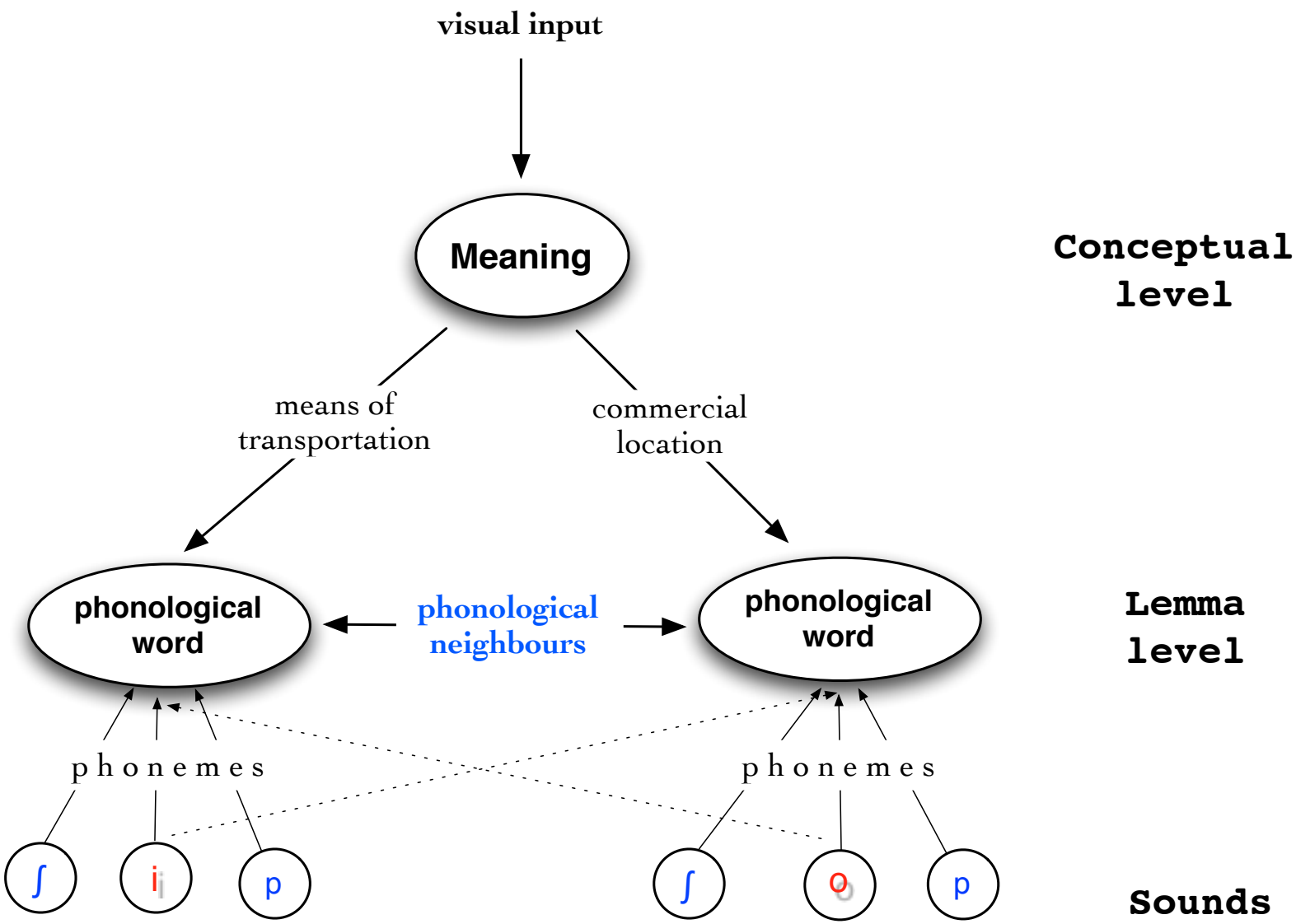
Access vs. activation

(continued)

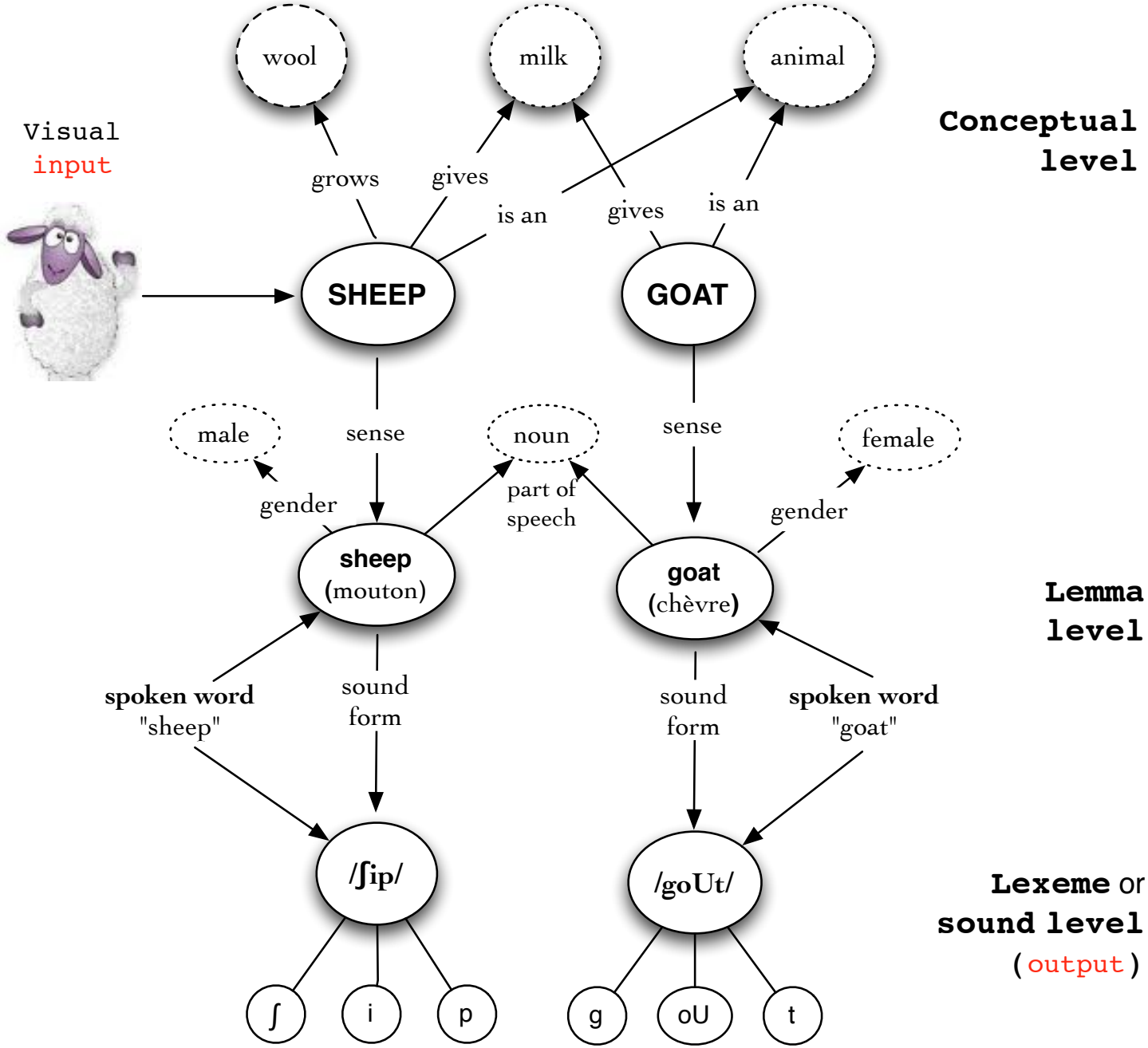
"A potentially counterintuitive idea is that the individual sounds of words are **assembled anew** each time they are spoken rather than **retrieved** as **intact wholes**. Yet, patterns of speech errors and latency data suggest that this is the case. "

Zenzi M. Griffin and Victor S. Ferreira,
Properties of Spoken Language Production, page 35.

In **Handbook of Psycholinguistics**
Traxler, M. and Gernsbacher, M. A. (Eds.), 2006


visual input
Meaning
Conceptual level
means of transportation
commercial location
phonological word
phonological neighbours
phonological word
Lemma level
phonemes
phonemes
ʃ
i
p
ʃ
o
p
Sounds

Levelt's model



Functioning

Activation spreading

Comments

Activation acts blindly: all neighbours are activated equally

==> non-target nodes become activated and remain so for a while

Activation acts in a deterministic fashion

==> we cannot escape it

Can we use this work for dictionary consultation?

- Answer: no
- While computational psycholinguists can tune the weights to have their model mimick human behavior (speed, accuracy), we cannot do the same for dictionary look-up.
- Reason : while we do know the starting node (query, input), we do not know the target (the desired, elusive word). If we did, we wouldn't have bother at all to perform look-up via an external aid, we would simply produce the target word.

Still, functionally speaking there is a way to achieve something equivalent

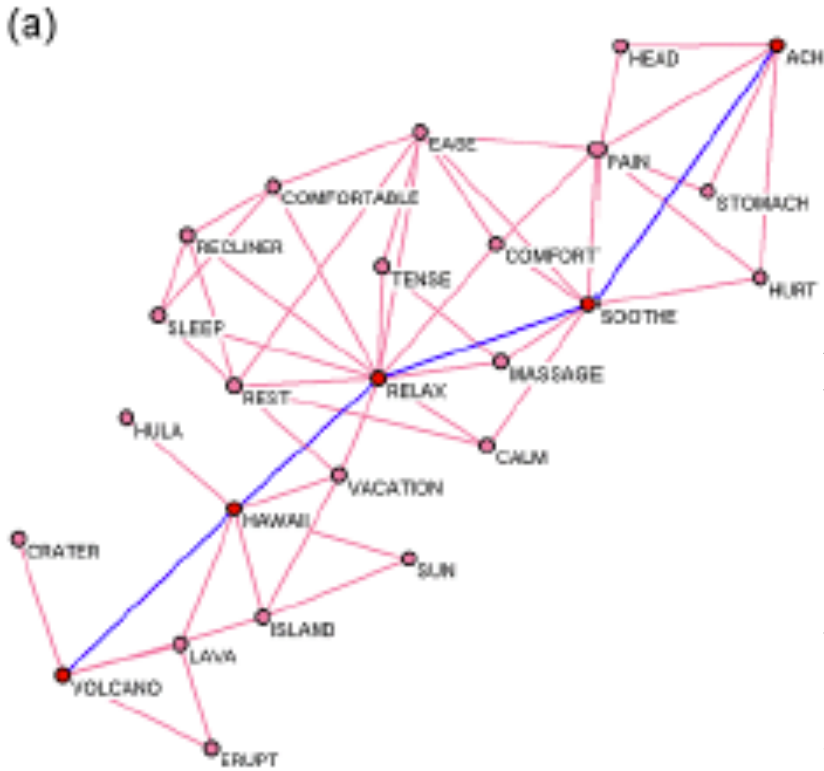
Means

Build an **index**

but, **what kind** of?

Answer : use associations

Navigation in an associative network



Since **search** takes will take place within a semantic network, i.e. a graph where all words (nodes) are related (via a certain kind of association), search consists in entering this network at any point and follow the links to get from the starting point (source word, SW) to the end (target word, TW). This latter may be directly related to the initial input, i.e. SW (direct association/neighbour; distance 1) or not (indirect association).

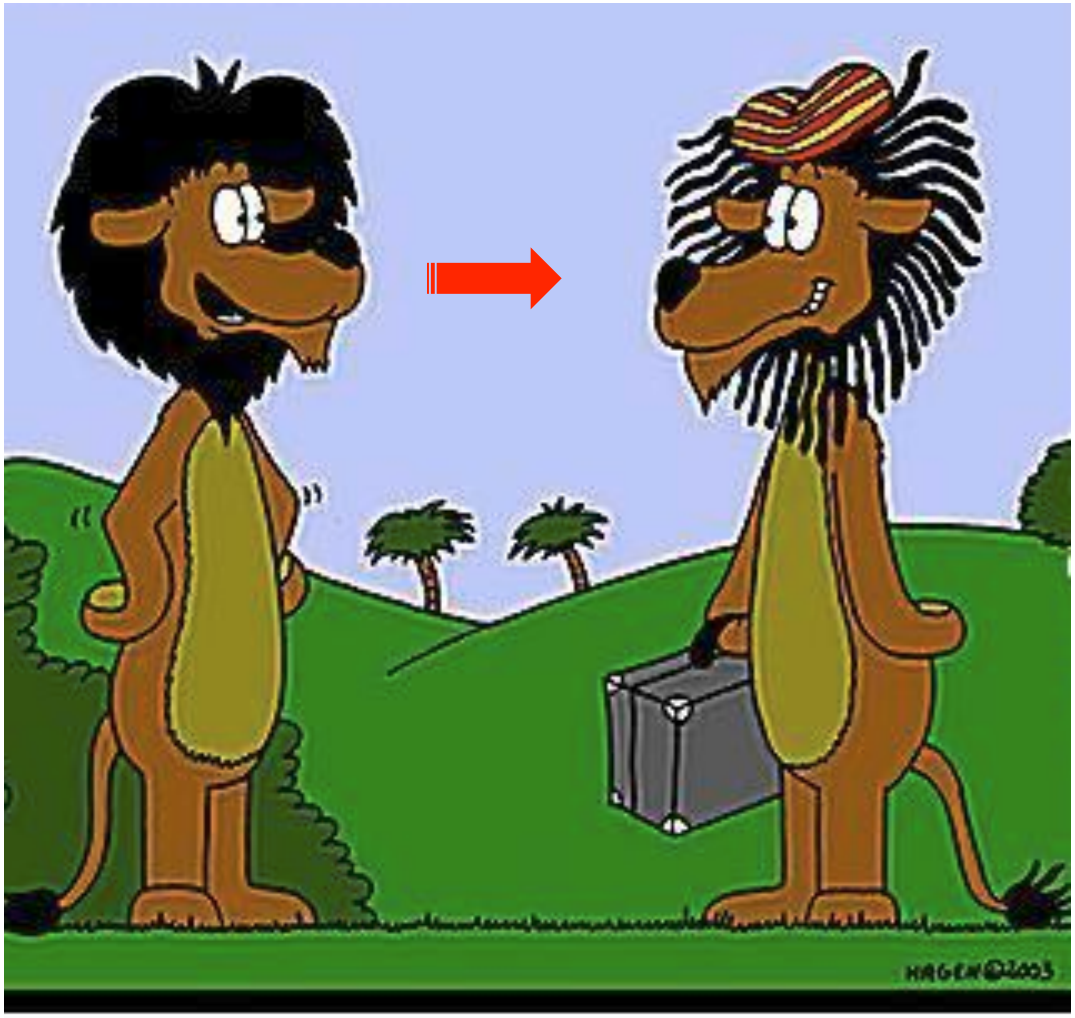
Note that the user knows the starting point, but not the end-point (target).

Evidence of associations

Priming :
activation of information

dreadlocks
Bob Marley
Jamaica

A B



Let me guess:
You went to Jamaica for your holidays...

The crazier the better

Come on, don't do the Fosbury again !

Interpretation

Subjectivity of interpretation of data (*litteral* or *enriched*)

Go **beyond** the information **given**

"Wait! Wait! … Cancel that, I guess it says 'helf.'"

Evidence for associations

What did you read?
What did you understand?
What happened next?

breakfast
primes
bed

Associations

A list of some 20 words is read to the subjects, e.g.

winter, icy, Siberia, warm, cooling, penguin, frozen, flu, chilly, ice, wind, hot, Antarctica,wet, fresh, breezy, igloo, cool, snow, Pole, glacier, frost, sleet

« When trying to remember as many words of the list as possible, people will typically remember the word "cold", even though it is not part of the list. This is because "cold" is strongly associated to all other words. Hence, the brain tends to "fill in" or "induce" the missing piece that it expects to be there. »

Wholes, **parts** and our' natural **tendency** to connect unknown to known, i.e. to impose or restore '**order**'

Aoccdrnig to rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit a porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

Activation (association)

1. By **context** : bread => butter
2. By **meaning** : bread => food
3. Via **form** : bread => red, historical => hysterical
4. Via the **meaning/context** + the **form** :

cat => rat;

DSK => election :

election => erection (phonological neighbour)

Associations are individual and culture specific

Letter for Elise

Associations for the piece called 'Für Elise'

Vienna

Associations for the piece 'Für Elise'

Taipei
Taiwan

Garbage collection

Where to get the associations from?

Association Thesaurus
Corpora (well balanced)

Input: India

http://www.eat.rl.ac.uk/cgi-bin/eat-server

PAKISTAN	12 0.14	FLIES	1 0.01
RUBBER	10 0.12	HIMALAYAS	1 0.01
CHINA	4 0.05	HINDU	1 0.01
FOREIGN	4 0.05	HUNGER	1 0.01
CURRY	3 0.04	IMMIGRANTS	1 0.01
FAMINE	3 0.04	INDIANS	1 0.01
TEA	3 0.04	JAPAN	1 0.01
COUNTRY	2 0.02	KHAKI	1 0.01
GHANDI	2 0.02	MAN	1 0.01
WOGS	2 0.02	MISSIONARY	1 0.01
AFGHANISTAN	1 0.01	MONSOON	1 0.01
AFRICA	1 0.01	PATRIARCH	1 0.01
AIR	1 0.01	PEOPLE	1 0.01
ASIA	1 0.01	PERSIA	1 0.01
BLACK	1 0.01	POOR	1 0.01
BROWN	1 0.01	RIVER	1 0.01
BUS	1 0.01	SARI	1 0.01
CLIVE	1 0.01	STAR	1 0.01
COLONIAL	1 0.01	STARVATION	1 0.01
COMPANY	1 0.01	STARVE	1 0.01
COONS	1 0.01	TEN	1 0.01
COWS	1 0.01	TRIANGLE	1 0.01
EASTERN	1 0.01	TURBANS	1 0.01
EMPIRE	1 0.01	TYRE	1 0.01
FAME	1 0.01	UNDER-DEVELOPED	1 0.01

Frequency and/or **recency?**
weights and **associations** are **not everything**

Output ranked in terms of **frequency**

PAKISTAN	12 0.14	FLIES	1 0.01
RUBBER	10 0.12	HIMALAYAS	1 0.01
CHINA	4 0.05	HINDU	1 0.01
FOREIGN	4 0.05	HUNGER	1 0.01
CURRY	3 0.04	IMMIGRANTS	1 0.01
FAMINE	3 0.04	INDIANS	1 0.01
TEA	3 0.04	JAPAN	1 0.01
COUNTRY	2 0.02	KHAKI	1 0.01
GHANDI	2 0.02	MAN	1 0.01
WOGS	2 0.02	MISSIONARY	1 0.01
AFGHANISTAN	1 0.01	MONSOON	1 0.01
AFRICA	1 0.01	PATRIARCH	1 0.01
AIR	1 0.01	PEOPLE	1 0.01
ASIA	1 0.01	PERSIA	1 0.01
BLACK	1 0.01	POOR	1 0.01
BROWN	1 0.01	RIVER	1 0.01
BUS	1 0.01	SARI	1 0.01
CLIVE	1 0.01	STAR	1 0.01
COLONIAL	1 0.01	STARVATION	1 0.01
COMPANY	1 0.01	STARVE	1 0.01
COONS	1 0.01	TEN	1 0.01
COWS	1 0.01	TRIANGLE	1 0.01
EASTERN	1 0.01	TURBANS	1 0.01
EMPIRE	1 0.01	TYRE	1 0.01
FAME	1 0.01	UNDER-DEVELOPED	1 0.01

Clustering by category

Countries, continents, colors, food, means of transportation, instruments, …

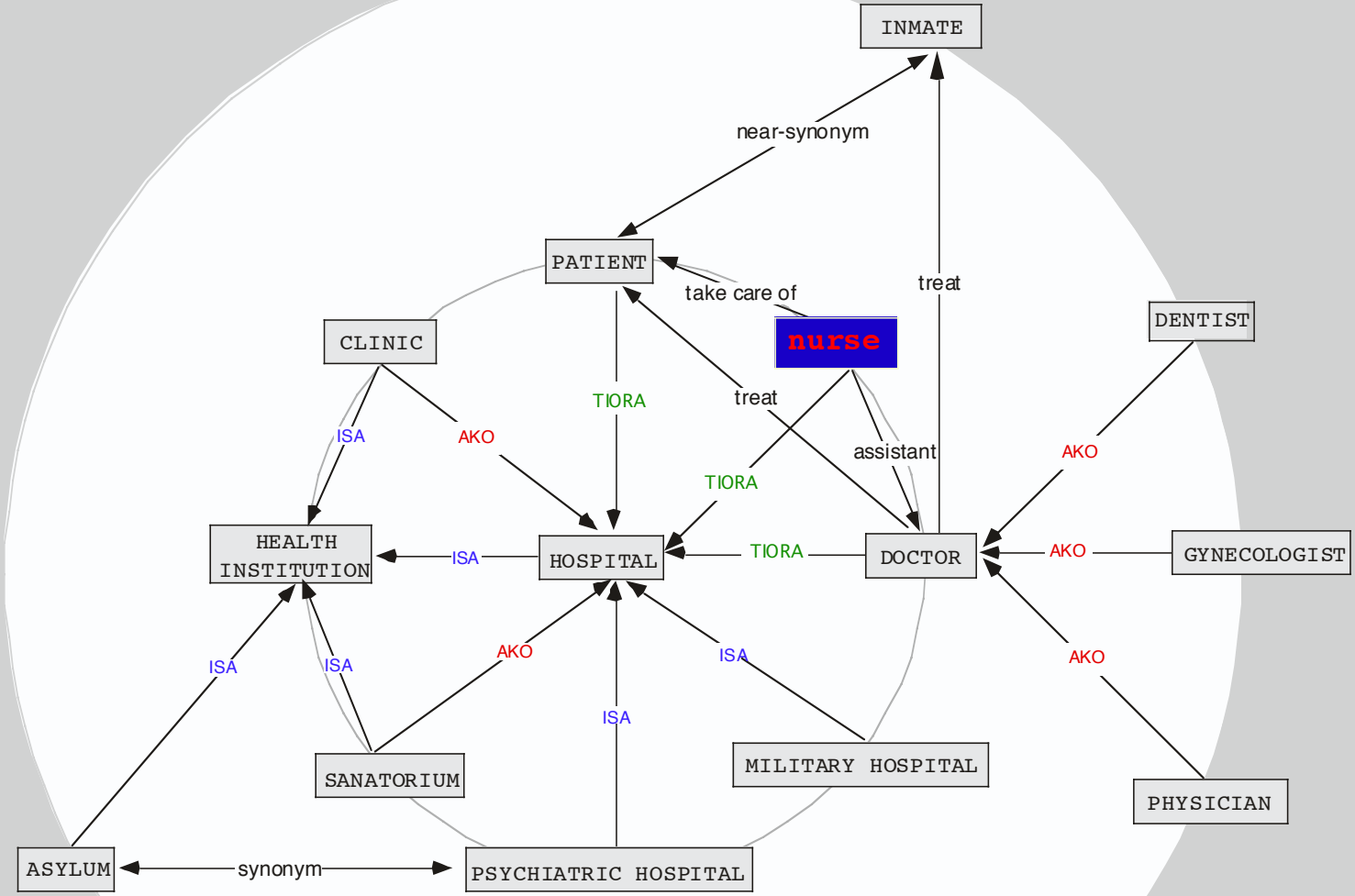
PAKISTAN	12	0.14	FLIES	1	0.01
RUBBER	10	0.12	HIMALAYAS	1	0.01
CHINA	4	0.05	HINDU	1	0.01
FOREIGN	4	0.05	HUNGER	1	0.01
CURRY	3	0.04	IMMIGRANTS	1	0.01
FAMINE	3	0.04	INDIANS	1	0.01
TEA	3	0.04	JAPAN	1	0.01
COUNTRY	2	0.02	KHAKI	1	0.01
GHANDI	2	0.02	MAN	1	0.01
WOGS	2	0.02	MISSIONARY	1	0.01
AFGHANISTAN	1	0.01	MONSOON	1	0.01
AFRICA	1	0.01	PATRIARCH	1	0.01
AIR	1	0.01	PEOPLE	1	0.01
ASIA	1	0.01	PERSIA	1	0.01
BLACK	1	0.01	POOR	1	0.01
BROWN	1	0.01	RIVER	1	0.01
BUS	1	0.01	SARI	1	0.01
CLIVE	1	0.01	STAR	1	0.01
COLONIAL	1	0.01	STARVATION	1	0.01
COMPANY	1	0.01	STARVE	1	0.01
COONS	1	0.01	TEN	1	0.01
COWS	1	0.01	TRIANGLE	1	0.01
EASTERN	1	0.01	TURBANS	1	0.01
EMPIRE	1	0.01	TYRE	1	0.01
FAME	1	0.01	UNDER-DEVELOPED	1	0.01

Search scenario

Let's put this to work and take an example

word you are looking for (**target word**) ➡ **nurse**
word coming to your mind (**source word**) ➡ **hospital**

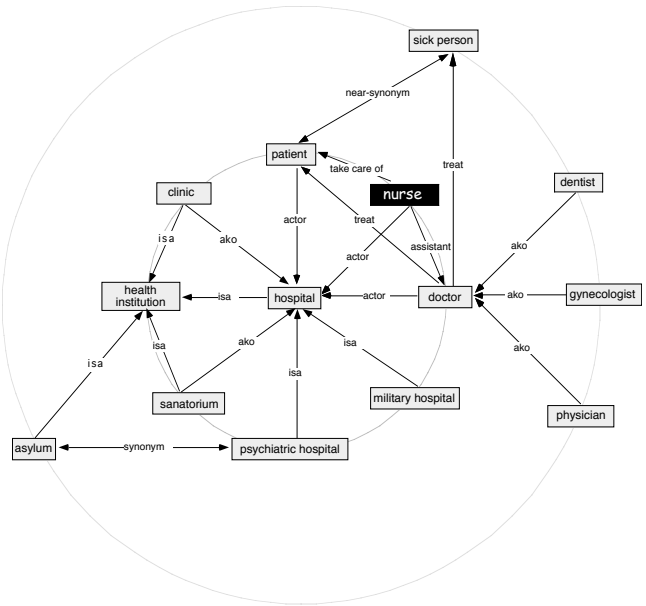
Internal Representation



Links must be interpretable to allow for navigation

Show only what's useful

Internal representation



AKO

- clinic
- sanatorium

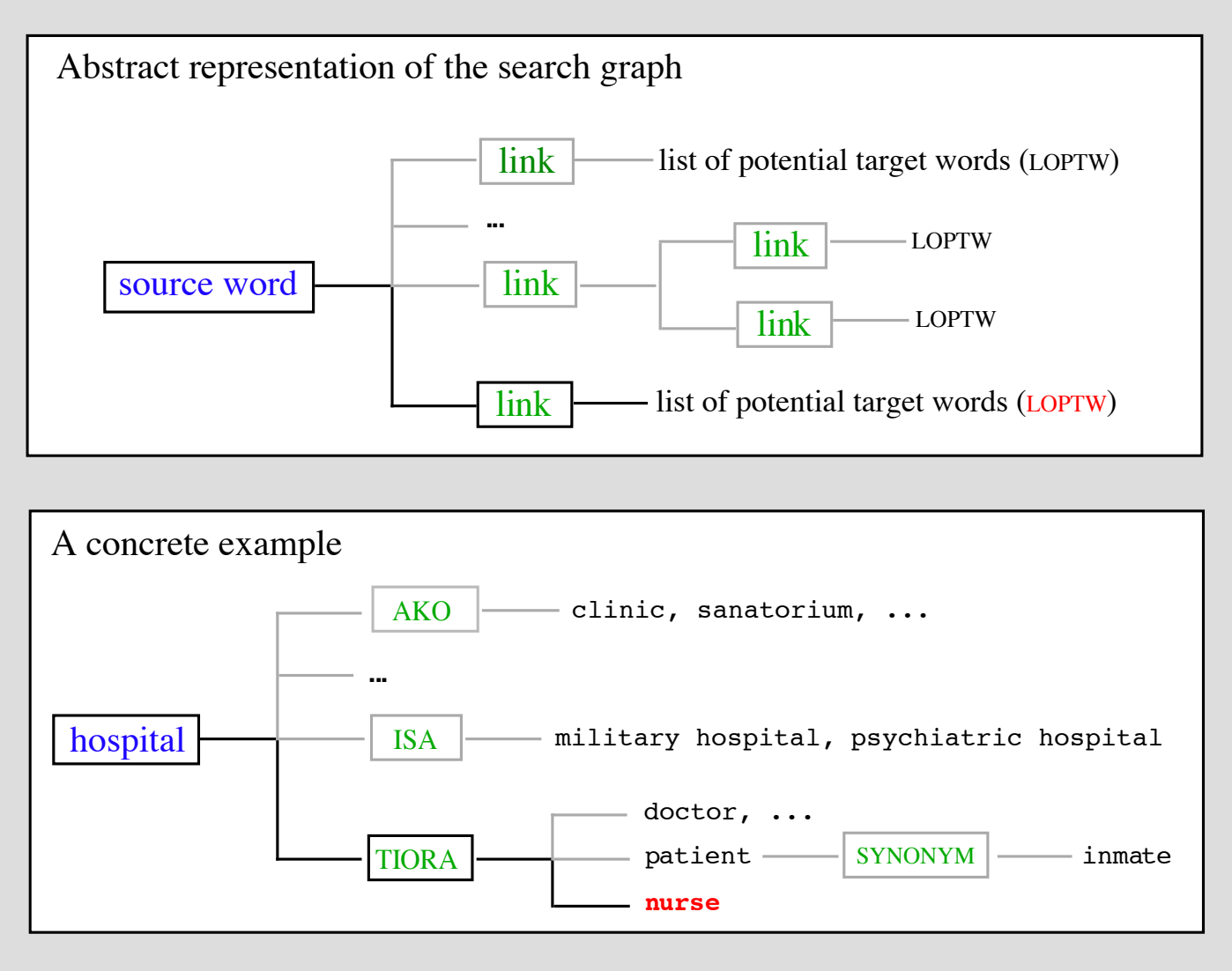
ISA

- military hospital
- psychiatric hospital

ACTOR

- doctor
- patient
- nurse

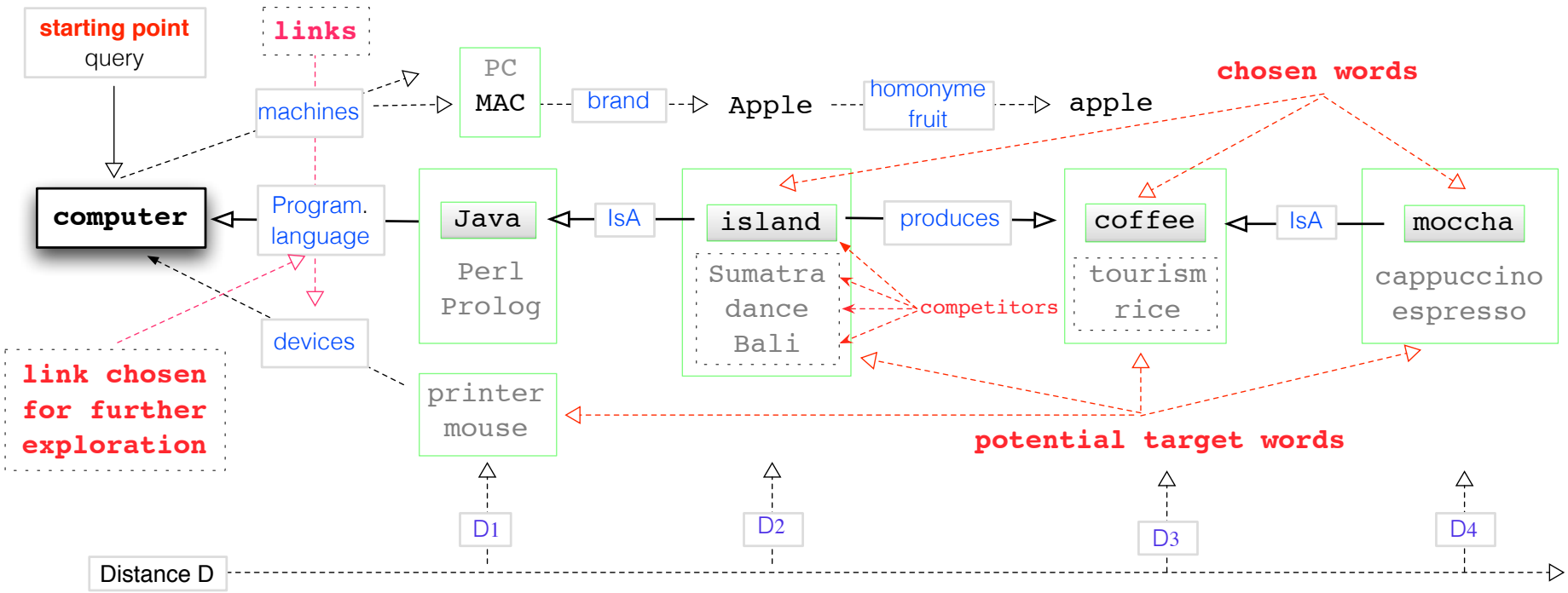
Don't drown the user: ease navigation



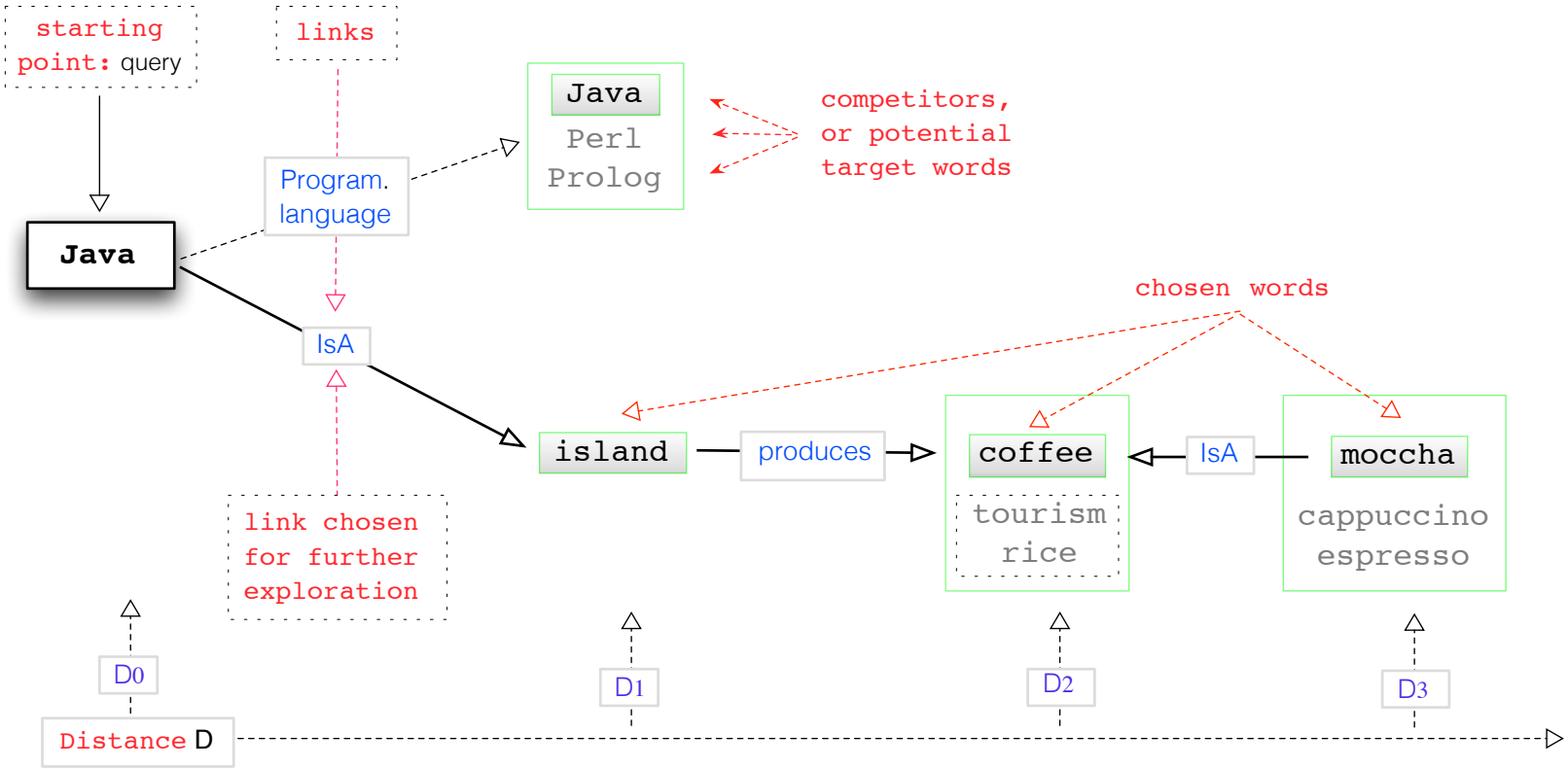
Search scenario

1. Show **not only direct associations**
2. but also **indirectly related words**

Finding a remote item at the distance of four mouse clicks (D_4)

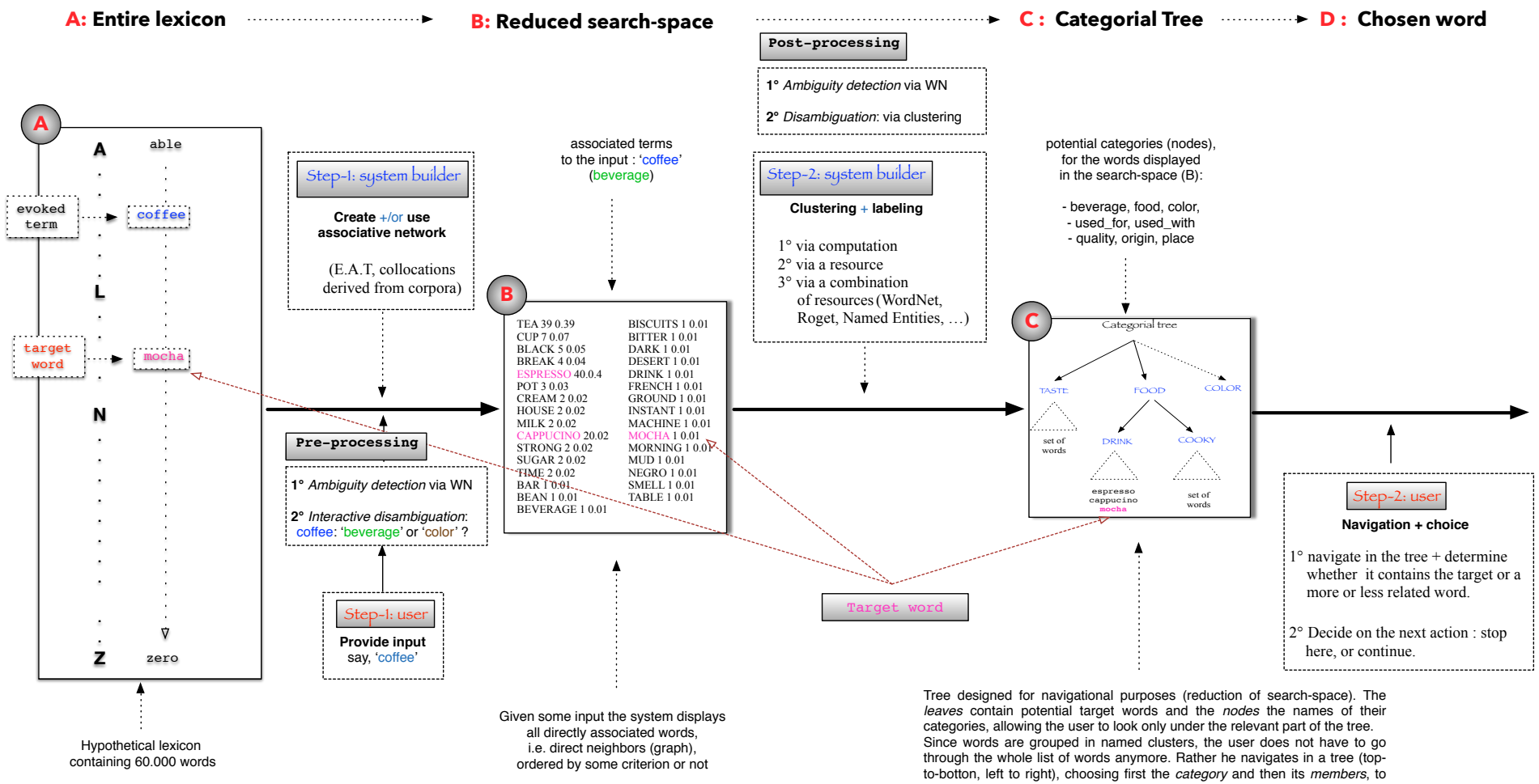


Find a word with a few mouse-clicks



The **nature** of the **problem** of search,
the **framework** of our approach
and its **solution** in a nutshell

How to access the word stuck on the tip of your tongue?



Conclusion

I have presented here some ideas concerning the mental lexicon, trying to see whether some of its functionalities can be used in electronic dictionaries.

I have strongly pleaded for the potential of word associations. While one can certainly rely on the words composing the definition of the target word (meaning, plan A, the normal route), a lot more can be done by using word associations (plan B).

Conclusion

Of course, a lot more work is needed. In particular, we need to

- get the right resources or corpora
- extract the links
- name them and
- build the application allowing to perform the here-described search
- evaluate the tool

Thanks for hanging in!

Dan will tell you now
how to get
all this to work !