My concern

Language production

- speaking
- writing

Grammar and dictionary, two major resources

« Without **grammar** very little can be conveyed,
without **vocabulary**, nothing can be conveyed »

Wilkins, D. (1972). *Linguistics and Language Teaching*.
London: Edward Arnold

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

Some facts (continued)

Performance

- 3-5 words per second
- 300 words per minute
- 18.000 words per hour

Comments

- that makes for a lot of look-ups, especially for talkative people (lecturers, politicians, typical southern-europeans)
- they never seem to get tired
- efficient **search**
- very efficient **organization** of the data-base (the mental lexicon)

Questions

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

2° Can we achieve sth similar on 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?

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	**Meaning**
Will you draw me a 	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

Idea: HAT OF A BISHOP

Meaning-related associations

church, Vatican, abbot, monk, monastery, ceremony, ribbon,

Sound related (rhyming words)

brighter, fighter, lighter, righter, tighter, writer,

possibly leading us to the **target word**: mitre

Believe it or not, even machines can fail

It all depends on the quality of

- the resource
- the query
- the search method

Evaluation of system performance

Critical variables

- type of search algorithm
- nature of the **corpus**

Relative success

- to find the desired target word
- speed
- accuracy

Automatic comparison of output produced by various algorithms

Comparison of the number of steps required by each search method (algorithm) in order to find the target word. We consider a word to be found if it occurs among the top ten.

WordFinder

Welcome to the WORDFINDER webpage

Input harvest wine grapes [send]

Output (found, related words): **23** hits

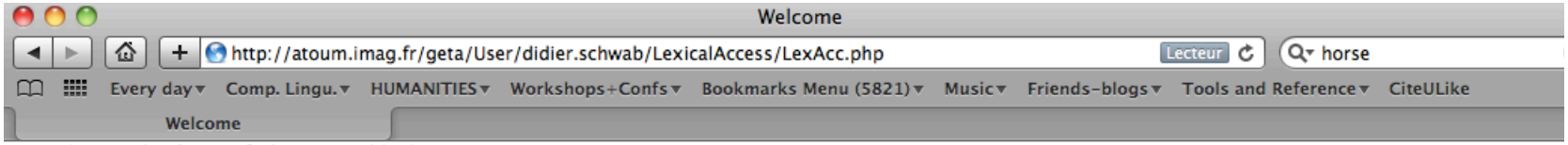
Beaujolais, regions, area, quality, between, vintage, well, usually, vineyards, south, various, year, growing , early, cru, low, north, following, aging, generally, time, potential, very

¶

Automatic comparison of output produced for different 'corpora'

- WordNet
- Wikipedia

Comparison of different search algorithms



Input (successive input of trigger words) : honey

Click here to get target words

Search algorithms : (see paper)

- (•) basic weighting (A1)
- ( ) linear weighting (A2)
- ( ) indirect use with basic weighting(A3)
- ( ) indirect use with linear weighting (A4)

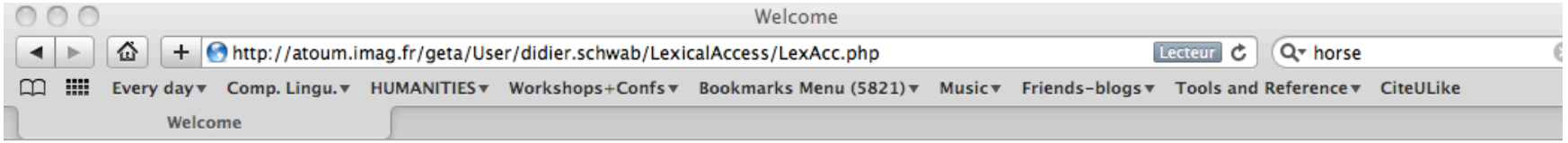
Reset

Target words (honey, 16384) (live, 6720) (bee, 6528) (hive, 5888) (nest, 5120) (refer, 4480)

(domesticate, 4480) (comb, 4480) (structure, 3968) (occupy, 3008) (raise, 2304) (store, 2304) (out, 2304) (enclose, 2240) (extract, 2240) (hollow, 2112) (destroy, 2112) (entrance, 2048) (do, 1856) (only, 1792) (keep, 1792) (find, 1792) (found, 1792) (reinforce, 1792) (beeswax, 1792) (single, 1728) (full, 1728) (describe, 1664) (brood, 1536) (frame, 1536) (smoke, 1344) (remove, 1344) (call, 1344) (give, 1344) (spin, 1344) (empty, 1344) (return, 1344) (estimate, 1344) (reuse, 1344) (increase, 1344) (species, 1280) (tree, 1280) (beehive, 1216) (site, 1152) (smooth, 1152) (require, 1152) (rock, 1024) (parallel, 1024) (uniform, 1024) (volume, 1024) (than, 1024) (ground, 1024) (face, 1024) (layer, 1024) (part, 1024) (lower, 1024) (design, 1024) (size, 1024) (take, 1024) (kill, 960) (include, 960) (set, 960) (number, 896) (issue, 896) (expose, 896) (compose, 896) (show, 896) (grind, 896) (favor, 896) (reference, 896) (bark, 896) (coat, 896) (thin, 896) (harden, 896) (plant, 896) (resin, 896) (attach, 896) (leave, 896) (queen, 896) (build, 896) (edge, 896) (sun, 896) (date, 896) (tomb, 896) (straw, 896) (provide, 896) (must, 896) (piece, 896) (place, 896) (eke, 896) (interest, 896) (spread, 896) (state, 896) (western, 768) (hives, 768) (other, 768) (prefer, 768) (avoid, 768) (small, 768)

(tend, 768) (wall, 768) (top, 768) (depict, 768) (there, 768) (less, 768) (often, 704) (specie, 640)

Comparison of different search algorithms



Input (successive input of trigger words) : honey

Click here to get target words

Search algorithms : (see paper)

- basic weighting (A1)
- linear weighting (A2)
- indirect use with basic weighting(A3)
- indirect use with linear weighting (A4)

Reset

Target words (hive, 99954) (honey, 62248) (brood, 26688) (bee, 25778) (top, 18850) (describe, 16824)

(nest, 14660) (comb, 14612) (box, 11138) (live, 9828) (straw, 7889) (size, 7276) (destroy, 6731) (domesticate, 6713) (refer, 6426) (structure, 6137) (out, 5859) (frame, 5312) (call, 5201) (full, 5130) (hollow, 5068) (smoke, 4648) (single, 4608) (entrance, 4595) (design, 4578) (fix, 4242) (provide, 4028) (find, 3822) (found, 3822) (give, 3784) (occupy, 3531) (volume, 3392) (take, 3375) (extract, 3346) (store, 3324) (harvest, 3065) (enclose, 2968) (parallel, 2832) (beeswax, 2828) (raise, 2815) (site, 2763) (tree, 2762) (interest, 2702) (state, 2702) (do, 2638) (spread, 2604) (reinforce, 2513) (must, 2499) (eke, 2478) (set, 2475) (build, 2373) (rock, 2344) (smooth, 2340) (keep, 2317) (uniform, 2224) (place, 2191) (ground, 2144) (face, 2144) (remove, 2108) (hand, 2072) (include, 2045) (part, 2026) (layer, 1992) (only, 1952) (lower, 1928) (issue, 1883) (there, 1820) (spin, 1799) (empty, 1750) (piece, 1715) (return, 1701) (thin, 1673) (example, 1666) (total, 1648) (attach, 1638) (reuse, 1638) (bark, 1631) (estimate, 1631) (number, 1610) (increase, 1596) (leave, 1575) (end, 1533) (coat, 1519) (winter, 1512) (harden, 1498) (species, 1460) (plant, 1449) (resin, 1449) (advance, 1443) (upright, 1413) (even, 1395) (short, 1377) (discover, 1372) (evidence, 1372) (exist, 1372) (kill, 1354) (say, 1281) (safety, 1281) (quest, 1281)

Comparison of two ressources

Input:	*Output* eXtended WordNet	*Output* Wikipedia
wine	**488 words:** grape, sweet, serve, france, small, fruit, dry, bottle, produce, red, bread, hold...	**3045 words** name, lord characteristics, christian, grape, France, ... vintage (81st), ...
harvest	**30 words** month, fish, grape, revolutionary, calendar, festival, butterfish, dollar, person, make, wine, first,...	**4583 words** agriculture, spirituality, liberate, production, producing, ..., vintage (112th), ...
wine + harvest	**6 words** make, grape, fish, someone, commemorate, person, ...	**353 words** grape, France, vintage (3d), ...

Getting back to people

Observation

We always know **something** about the eluding object (name, word)

- place where we've met
- role s/he played

- meaning
- number of syllables
- origine
- ...

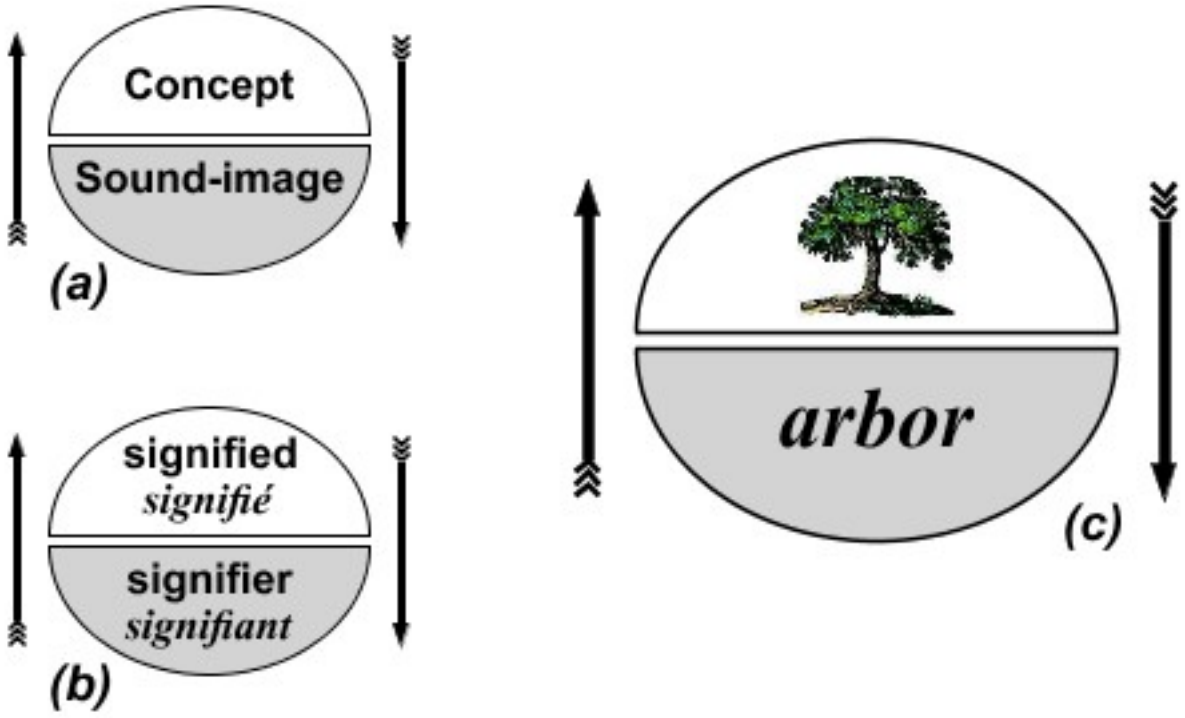
Hence...

Let's use that, and start from there.

Question : how?

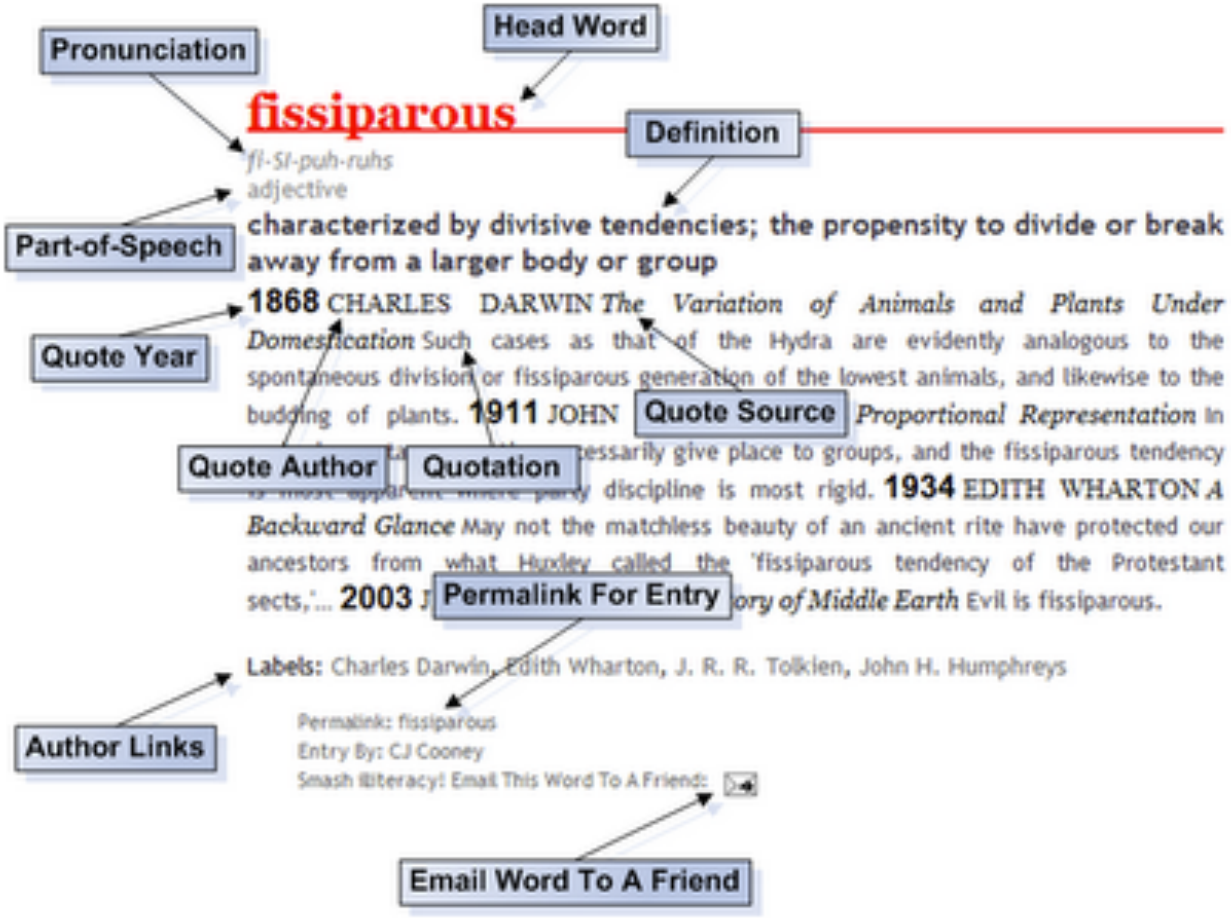
But before, let's try to get a clearer picture about the nature of the problem

Where is the problem?

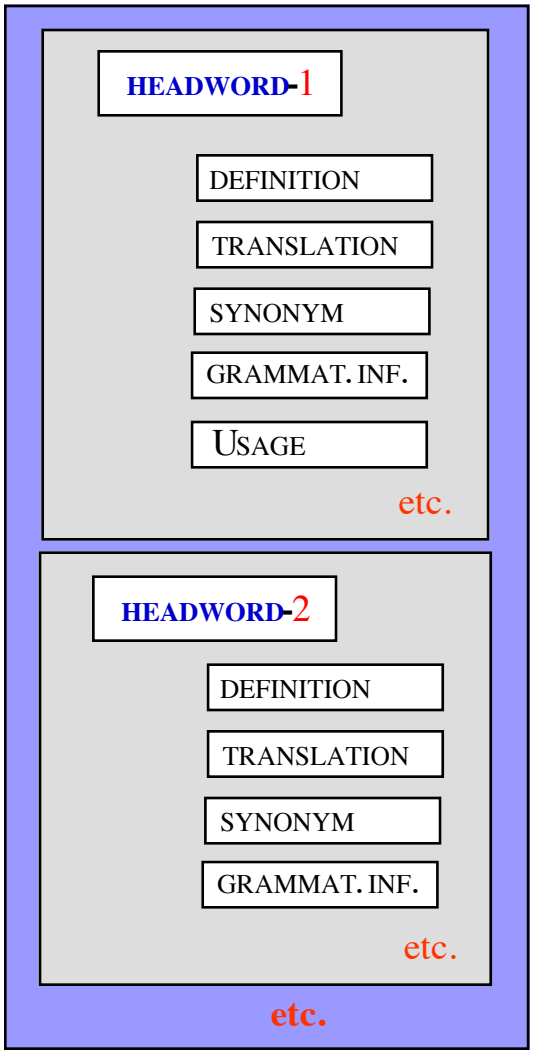


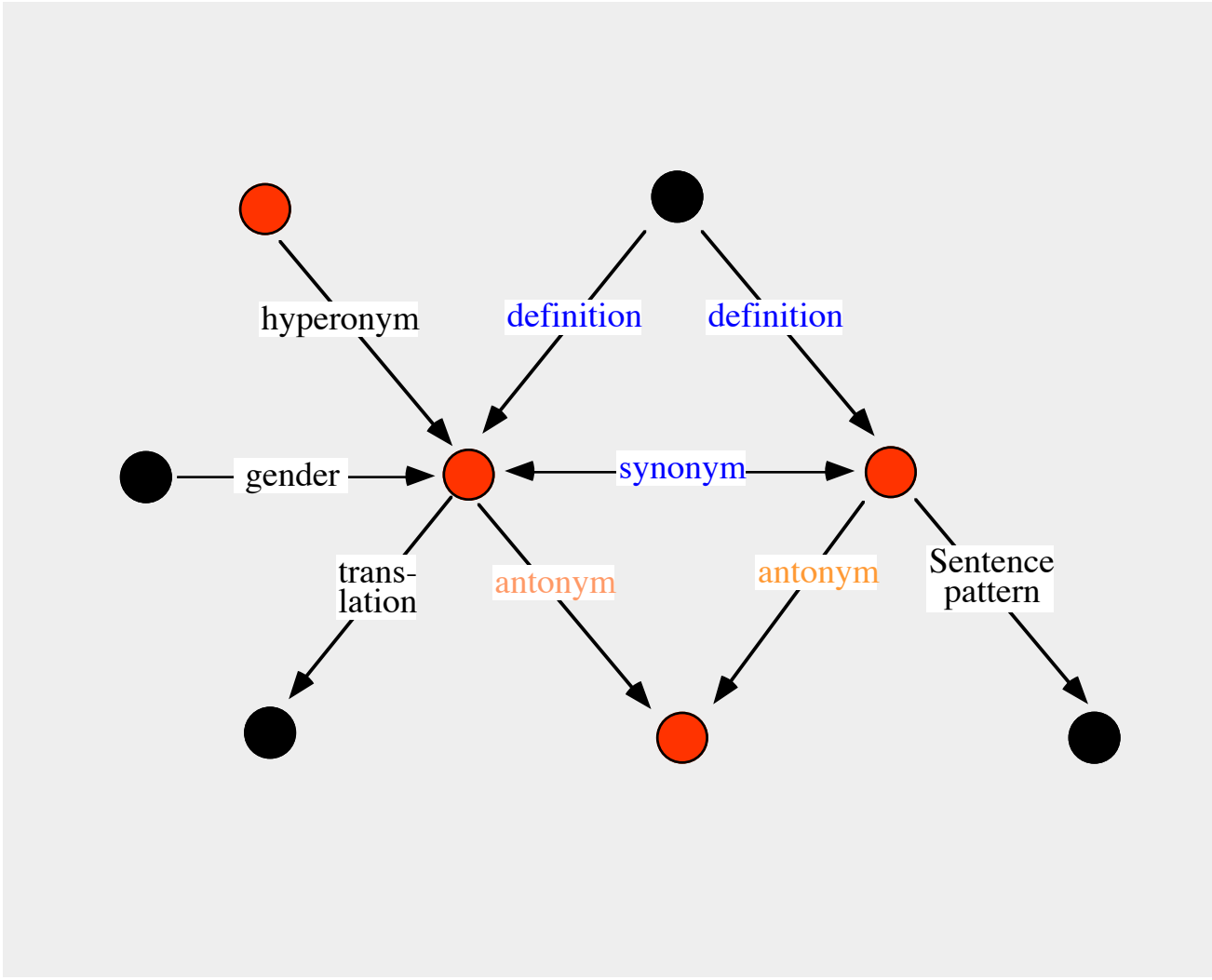
Saussure's conception of the 'sign'

Typical structure of lexical entry



Words as objects vs. words as nodes in a network





Please note

In **computational lexicography** words are viewed as tokens, i.e. *holistic entities*, and this holds both for the DB-view and lexical graphs (WordNet,FrameNet, ...)

In the **human brain**, i.e. mental lexicon, words are decomposed

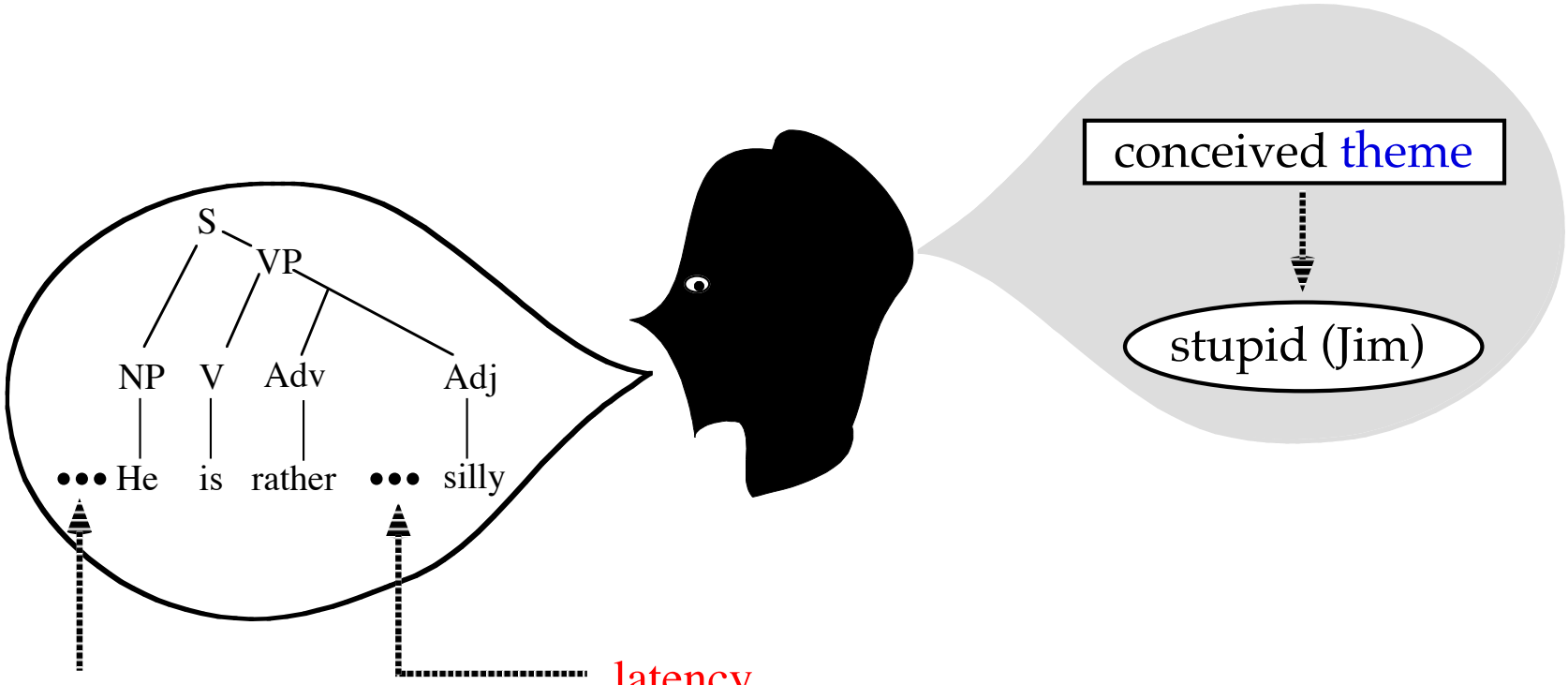
- meaning
- form
- sound

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)

What do you think of Jim ?



latency
to conceive
the theme :
`stupid (Jim)`

latency
to weigh between
"stupid" and "silly"

He is rather stilly.


Cesar


HUMANDESCENT.COM

Lexical access as puzzle completion

Access vs. activation

Storage vs. synthesis

"We do **not** store words at all in our mind, at least not in the sense of the layman's or lexicographer's view who consider word-forms and their meanings as one. If we are right, rather than continue to consider the human mind as a word **store** we could consider it as a word **factory**. Indeed, by looking at some of the work done by psychologists who try to emulate the mental lexicon (...) one gets the impression that words are **synthesized** rather than **located** and read out.

Access vs. activation

(continued)

By propagating **energy** rather than **data** —(as there is no message passing, transformation or cumulation of information, there is only *activation spreading*, that is, changes of energy levels, call it weights, electronic impulses, or whatever), — we propagate **signals**, activating ultimately certain **peripheral organs** (larynx, tongue, mouth, lips, hands) in such a way as to produce **movements** or **sounds**, that, not knowing better, we call words."

Zock, M. & D. Schwab. *Lexical Access Based on Underspecified Input*.
***Cog**nitive **a**spects of the **lex**icon,* 1st **CogALex** workshop,
Coling, Manchester, 2008

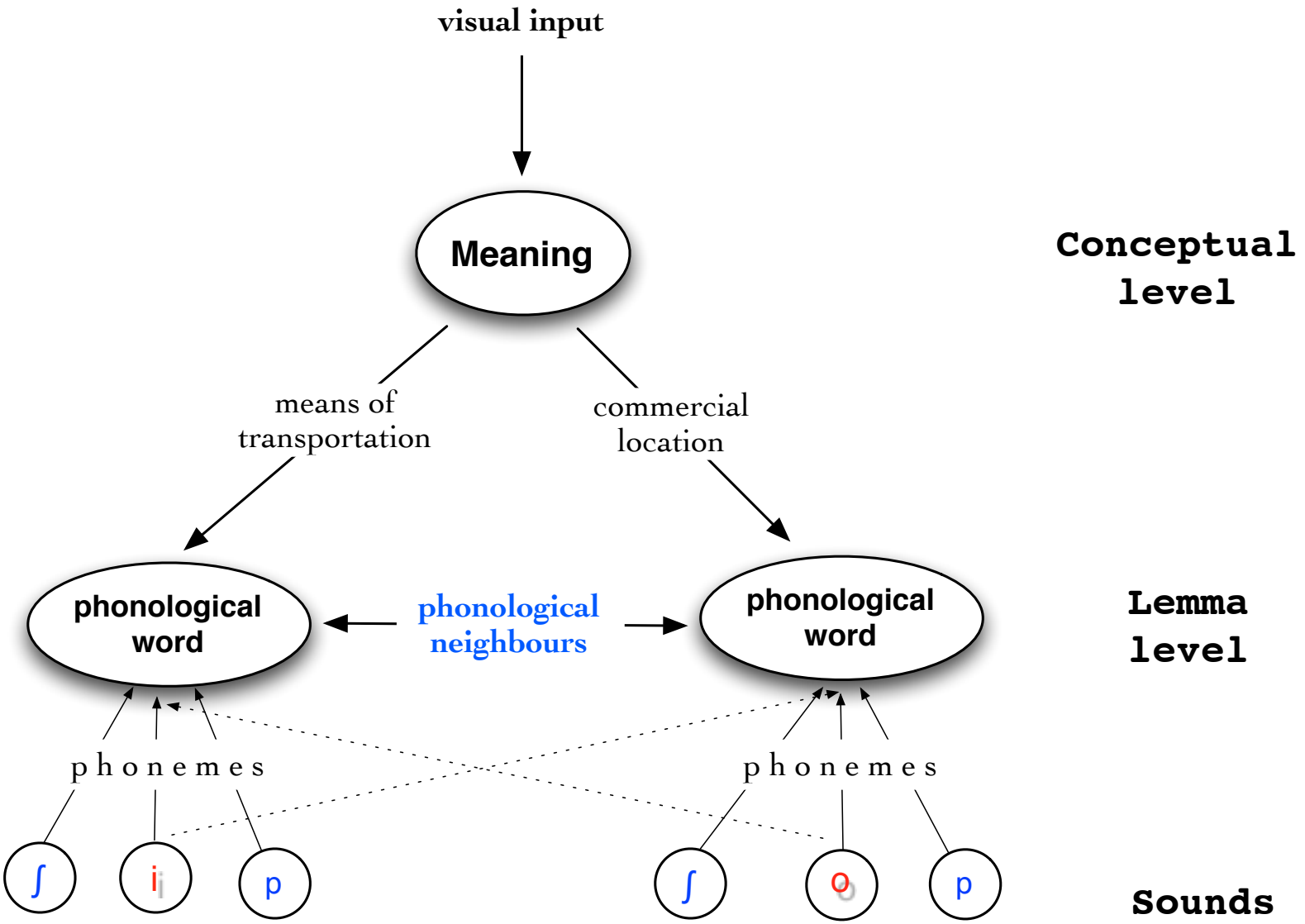
Accès vs. activation

(continued)

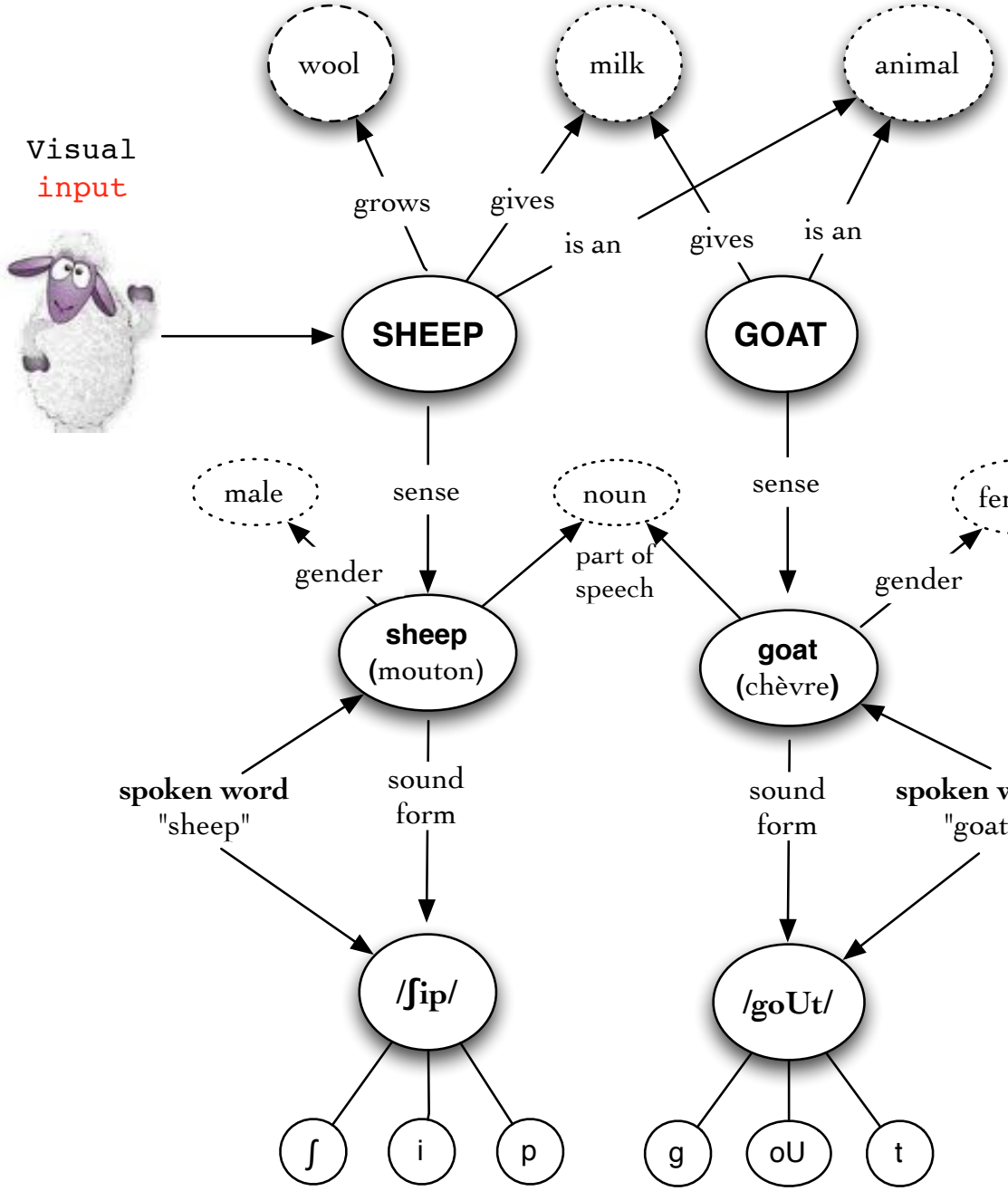
"A potentially counterintuitive idea is that the individual sounds of words are **assembled anew** each time they are spoken rather than **retrieved** as **intact wholes**. Yet, patterns of speech errors and latency data suggest that this is the case. "

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
means of transportation
commercial location
phonological word
phonological neighbours
phonological word
p h o n e m e s
p h o n e m e s
ʃ
i
p
ʃ
o
p
Conceptual level
Lemma level
Sounds

Levelt's model

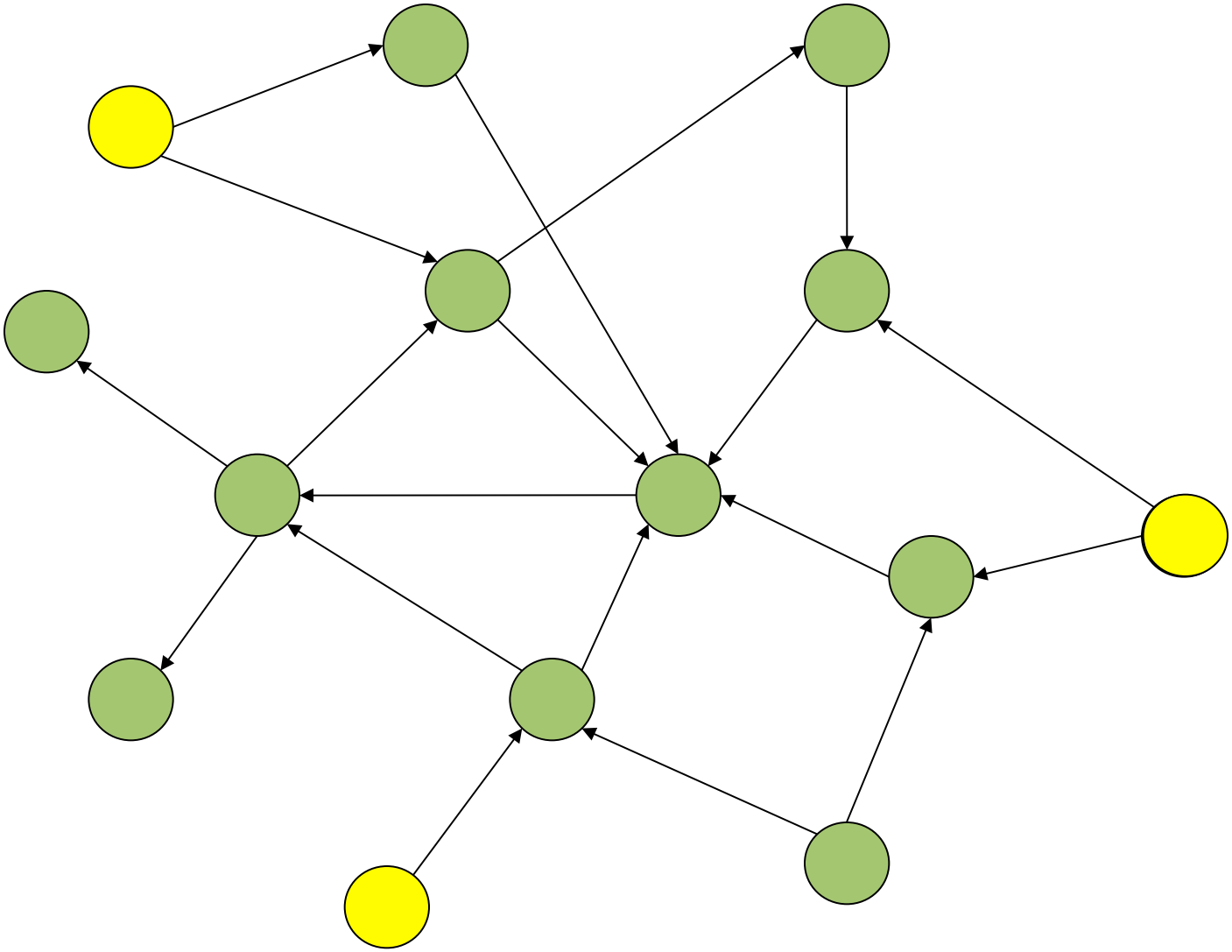


Functioning

Activation spreading

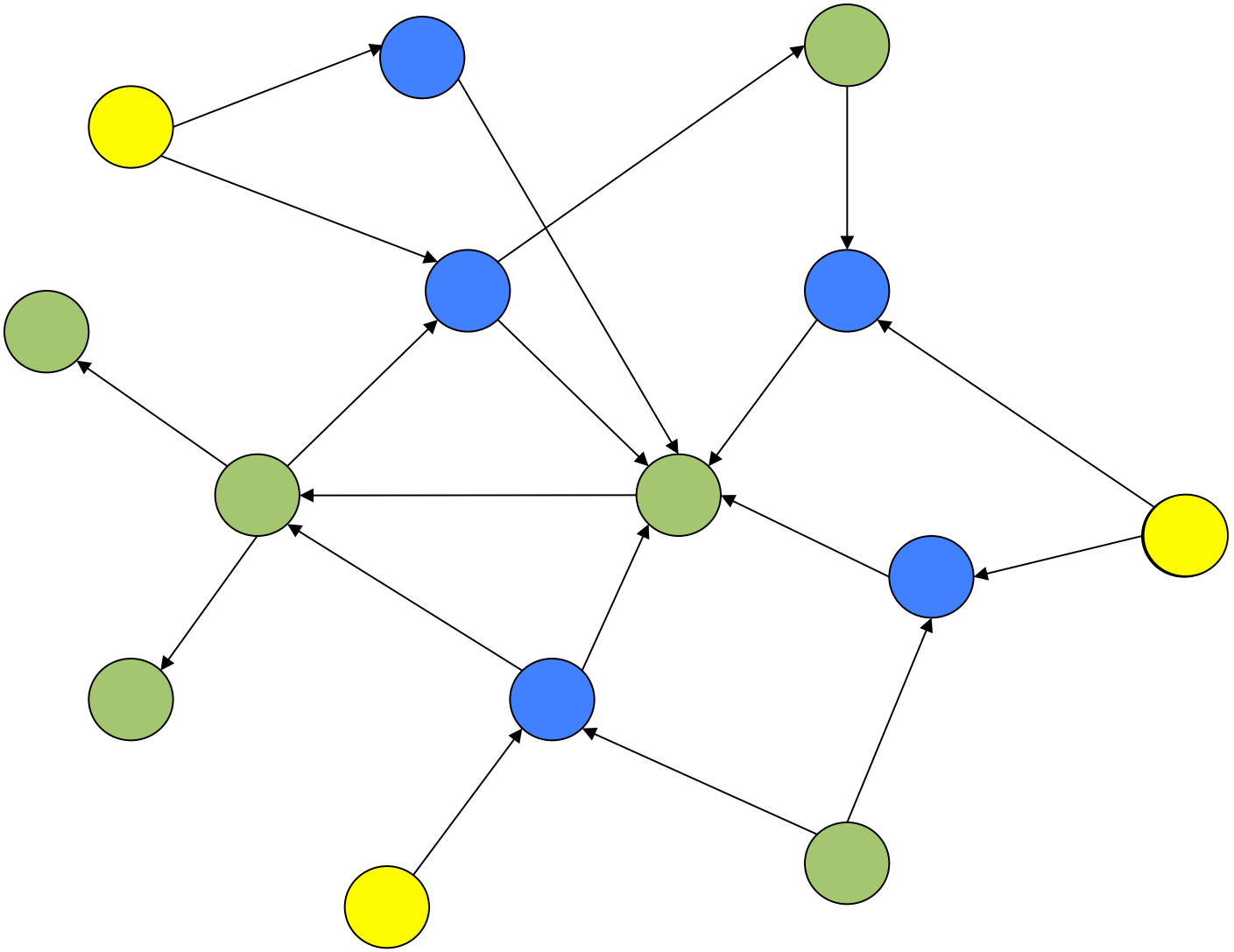
Spreading Activation

Start with an initial set of activated nodes



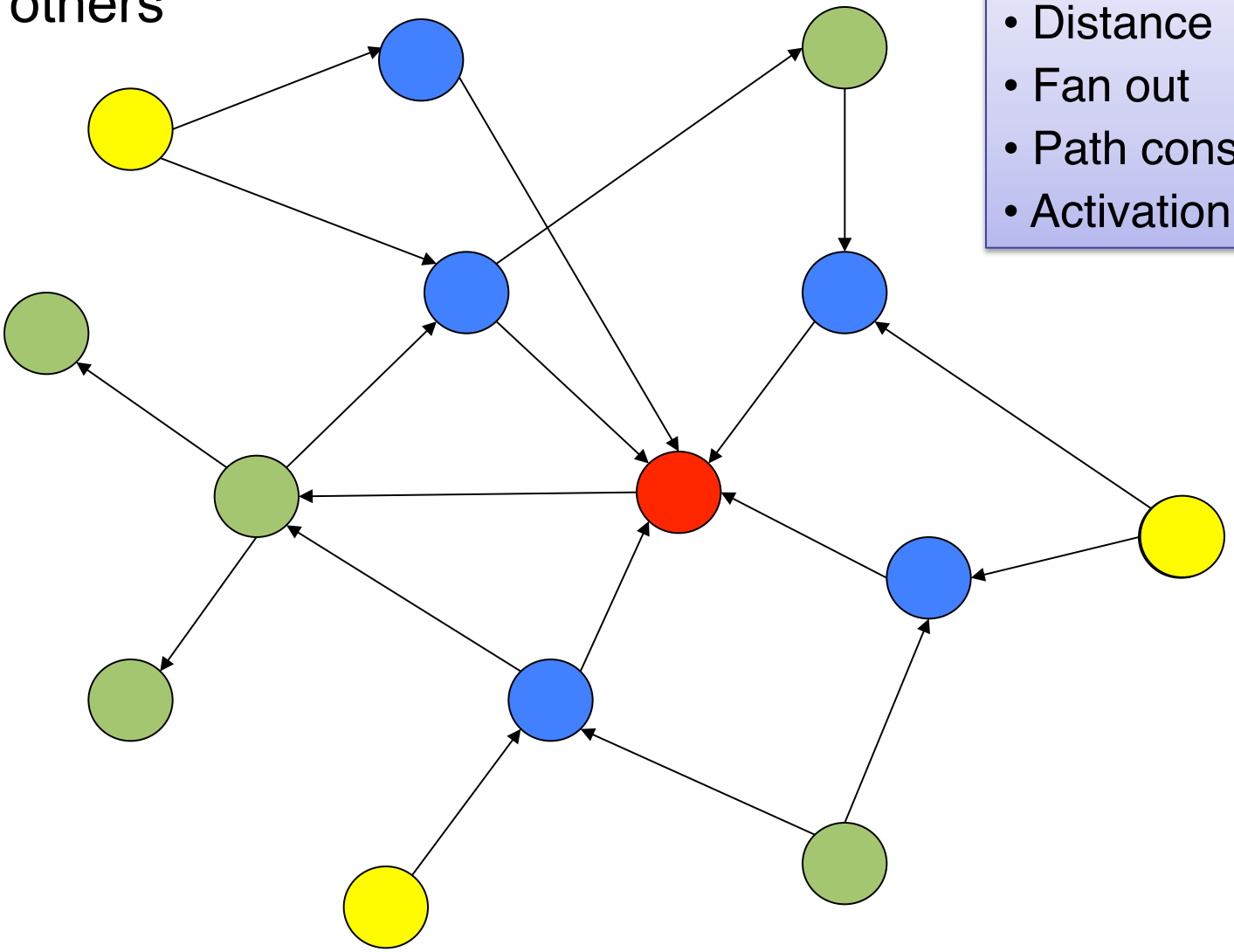
Spreading Activation

At each pulse/iteration, spread activation to adjacent nodes



Spreading Activation

Some nodes will have higher activation than others



Constraints

- Distance
- Fan out
- Path constraints
- Activation threshold

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

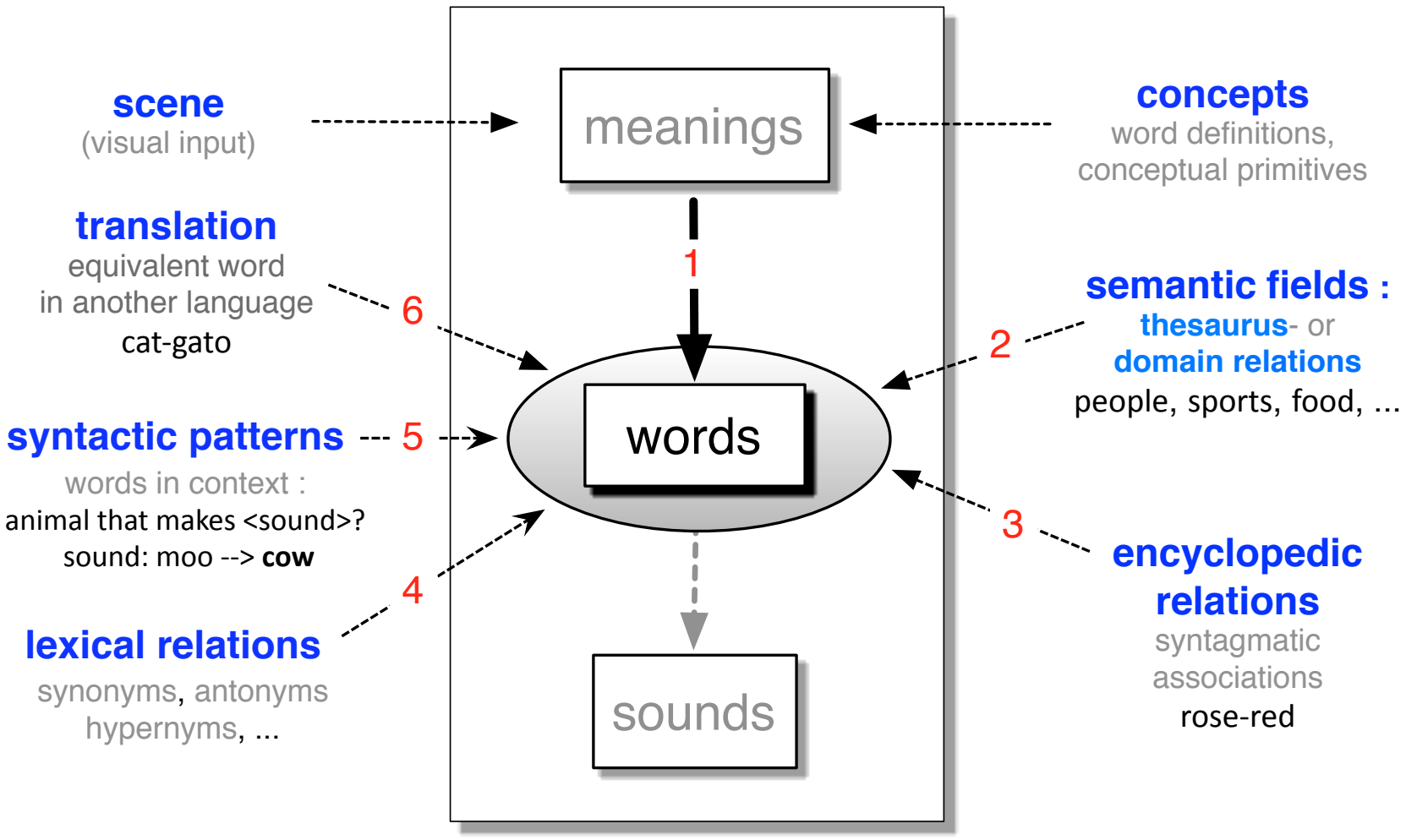
Can we use this for dictionary consultation?

- 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, there is a way to achieve functionally speaking sth equivalent

In addition, there is more than one way, or, the beauty of plan B

Lexical access via different routes



Means

Build an **index**

but, **what kind** of?

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**'

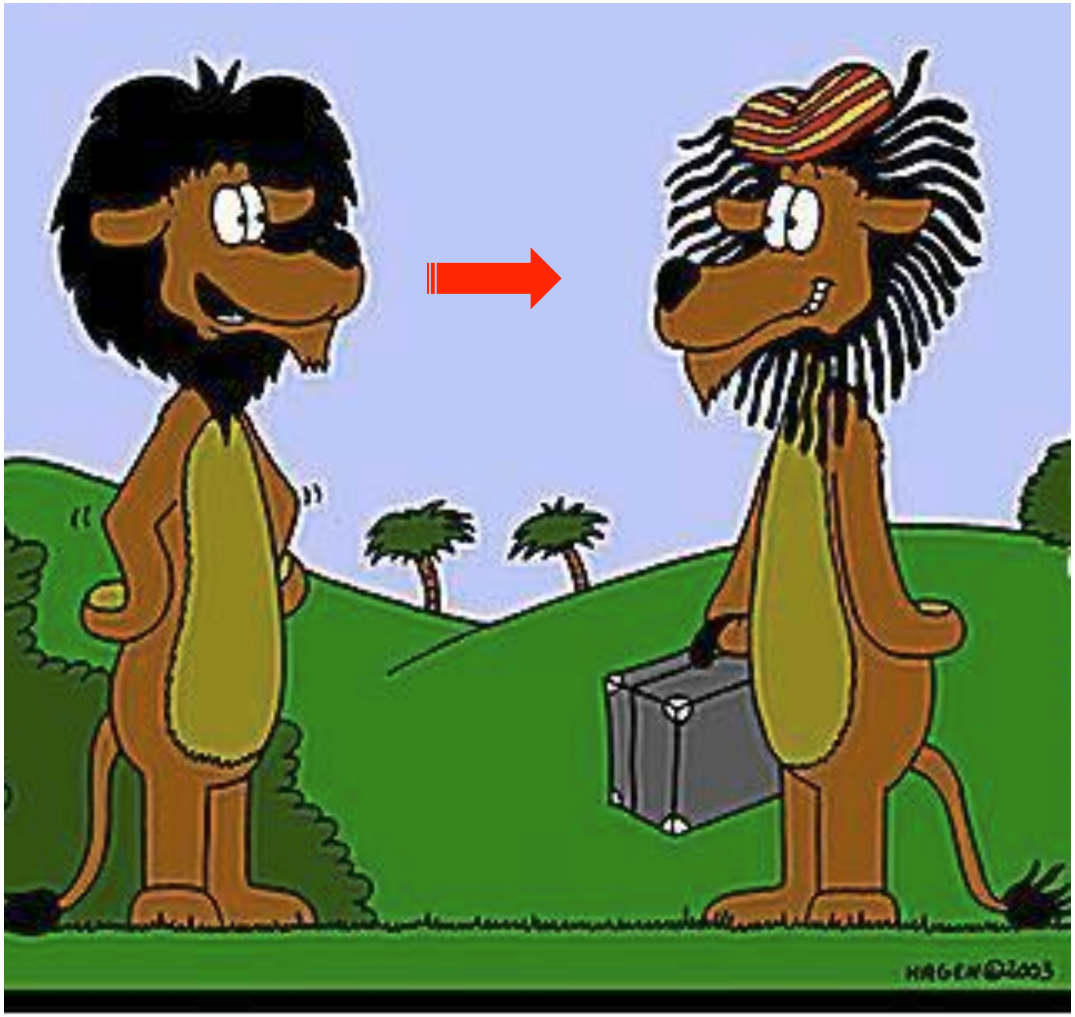
Aoccdrnig to rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit a porbelm. Tihs is bcuseae the huamn mnid deos not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

How to get a nice paycheck with a smile ?

Different ways to get the same message across :
Tiger's smile for Nike

- ✓ **figure** (here Tiger Wood's photo)
- ✓ **symbol** (here Nike AND Tiger's smile)
- ✓ **word** → (eg. Nike)
- ✓ or a **combination**

Priming :
activation of information



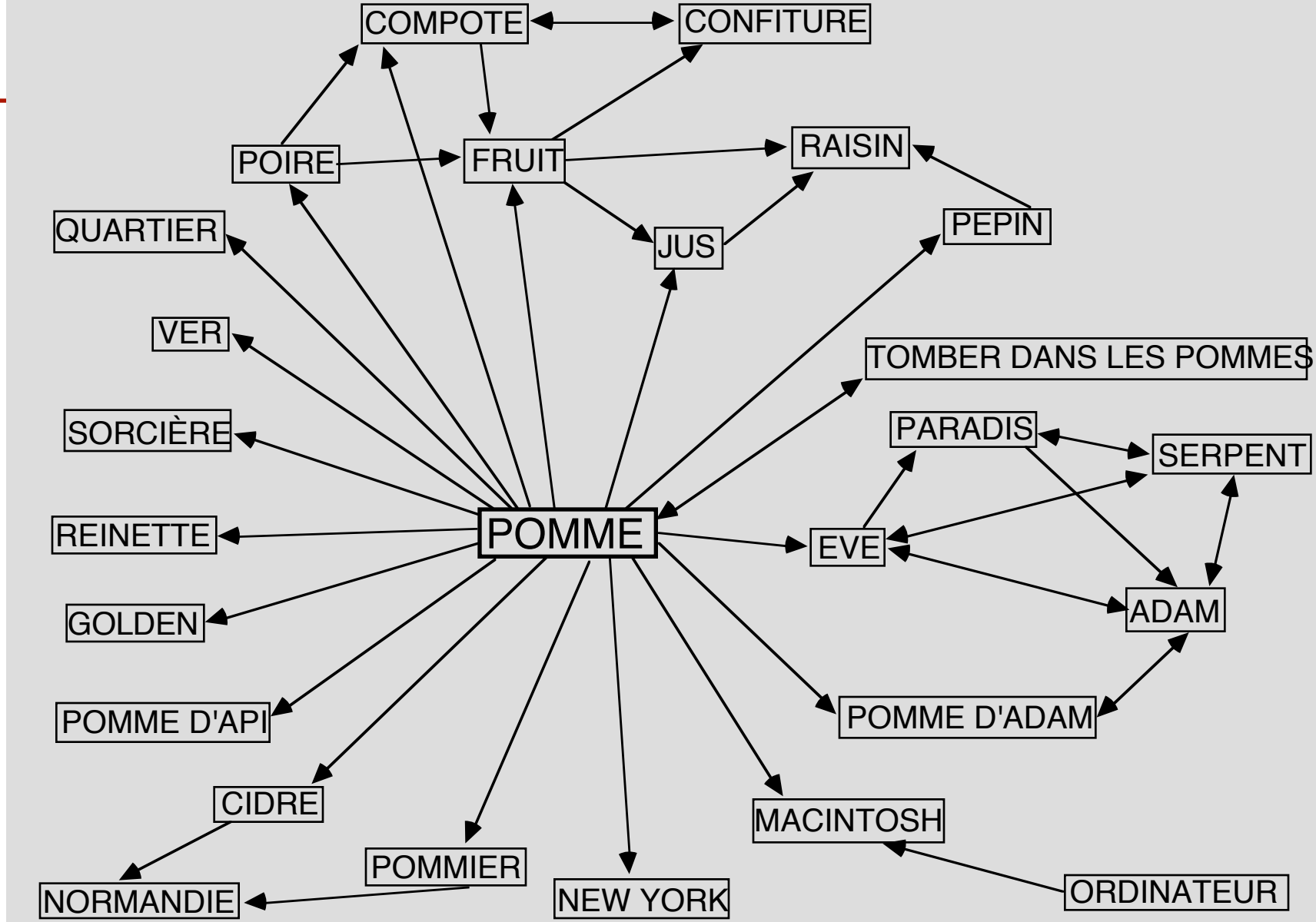
Activation (association)

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)


COMPOTE
CONFITURE
POIRE
FRUIT
RAISIN
QUARTIER
JUS
PEPIN
VER
TOMBER DANS LES POMMES
SORCIÈRE
PARADIS
SERPENT
REINETTE
POMME
EVE
GOLDEN
ADAM
POMME D'API
POMME D'ADAM
CIDRE
MACINTOSH
POMMIER
NORMANDIE
NEW YORK
ORDINATEUR

Claim:

Links have to be made explicit

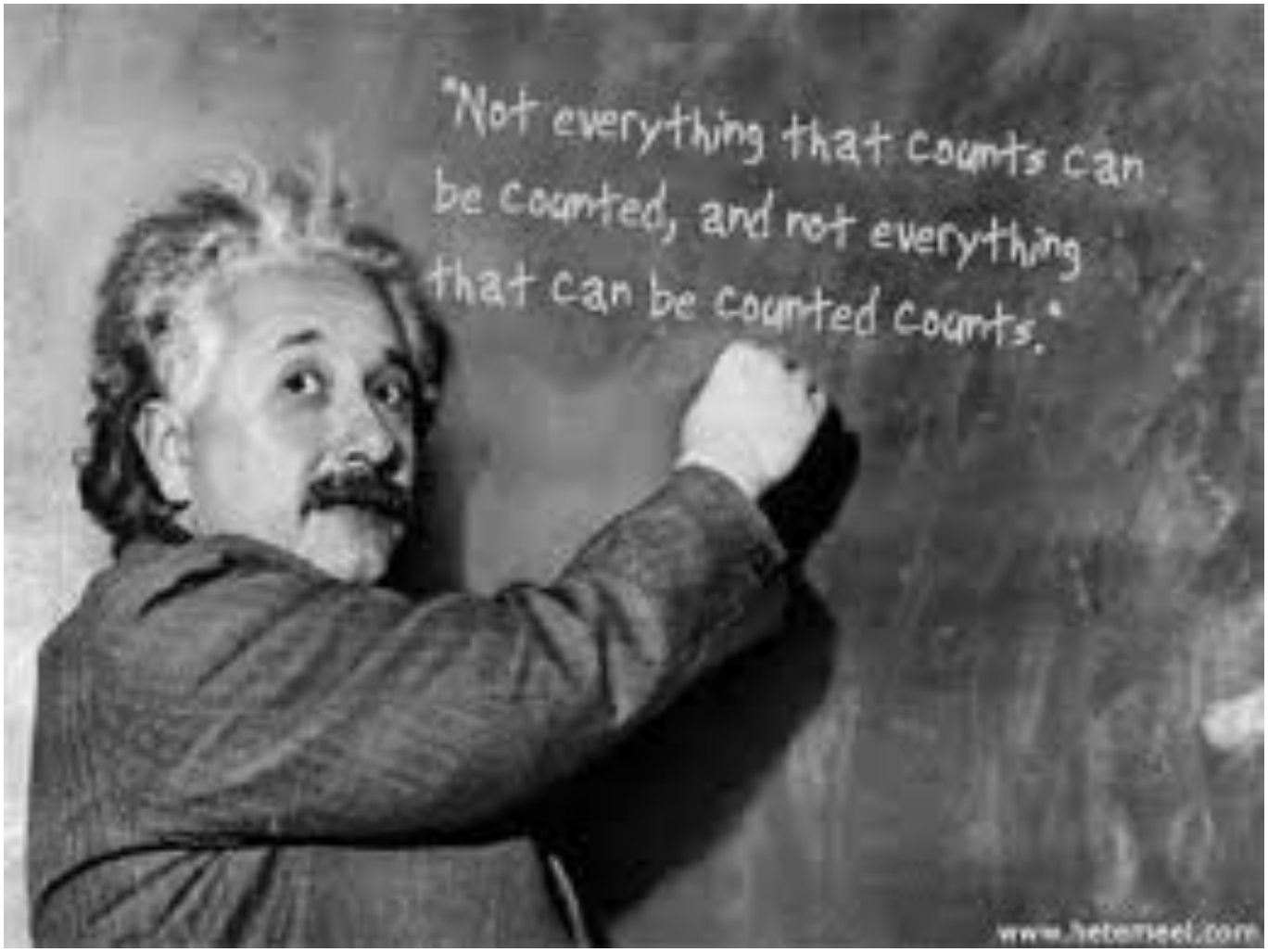
- **Structure** or **organize data** (words) into **clusters**
  - spelling **(alphabetically);**
  - **frequency;**
  - **semantic categories (topically, link type, semantic category, …)**
- **Guide** and **speed up search**
- **Number of associated terms** (great number of associated terms)
- **Directionality of links**
  - Increases the number of items to choose from
  - Different link types
    - rose-flower (hypernym)
    - flower-rose (hyponym)
- **Crossing links**

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

Comment:


"Not everything that counts can
be counted, and not everything
that can be counted counts."

Frequency and/or **recency?** **weights** 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

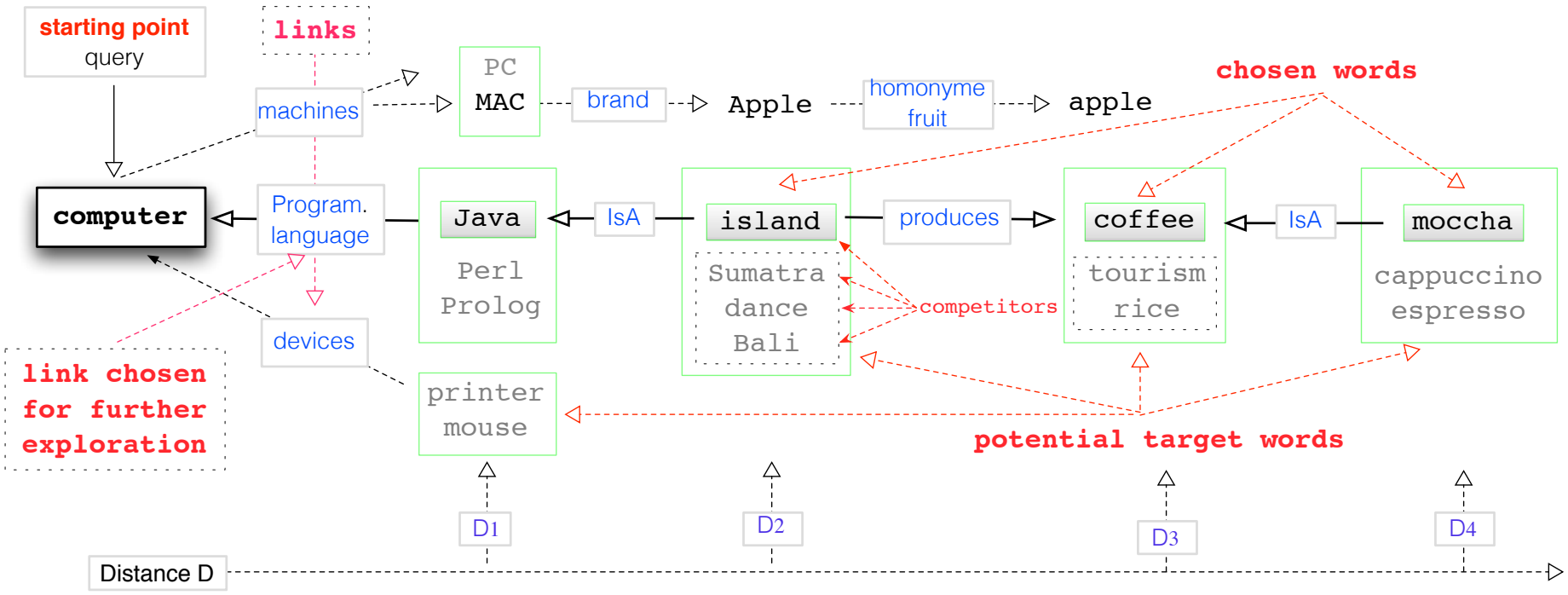
India being the **answer** to the following **stimuli**

SARI	22 0.15	**CAST**	1 0.01	LOOP	1 0.01
CASTE	17 0.12	CATASTROPHY	1 0.01	MEDITATE	1 0.01
NADIR	11 0.08	CHINA	1 0.01	MILDEW	1 0.01
FAMINE	6 0.04	CLUBS	1 0.01	MISSION	1 0.01
AFRICA	5 0.03	COLONIES	1 0.01	NATIVE	1 0.01
PAKISTANI	4 0.03	COMPANY	1 0.01	PLAGUE	1 0.01
STARVING	4 0.03	CONSUMPTION	1 0.01	POVERTY	1 0.01
BEGGING	3 0.02	CONTINENTS	1 0.01	PRIESTESS	1 0.01
CASTS	3 0.02	COTTON	1 0.01	QUININE	1 0.01
NAPALM	3 0.02	COUNTRY	1 0.01	SAVER	1 0.01
STARVATION	3 0.02	EAST	1 0.01	SECT	1 0.01
CHARISMA	2 0.01	ELEPHANT	1 0.01	SERVANT	1 0.01
CURRY	2 0.01	ELEPHANTS	1 0.01	SETTLEMENT	1 0.01
INCENSE	2 0.01	EMPIRE	1 0.01	SHEEPSKIN	1 0.01
KHAKI	2 0.01	FAMISHED	1 0.01	STARVED	1 0.01
PARIAH	2 0.01	FURTHER	1 0.01	SUFFERING	1 0.01
RICE	2 0.01	GHOUL	1 0.01	THUG	1 0.01
SPICE	2 0.01	HEDONISM	1 0.01	THUGS	1 0.01
STARVE	2 0.01	INCA	1 0.01	TIGER	1 0.01
TURBAN	2 0.01	INDIANS	1 0.01	TOGA	1 0.01
ALE	1 0.01	ISLAM	1 0.01	UNCLEAN	1 0.01
AMERICA	1 0.01	LEPER	1 0.01		
BIZARRE	1 0.01	LIFE-SPAN	1 0.01		

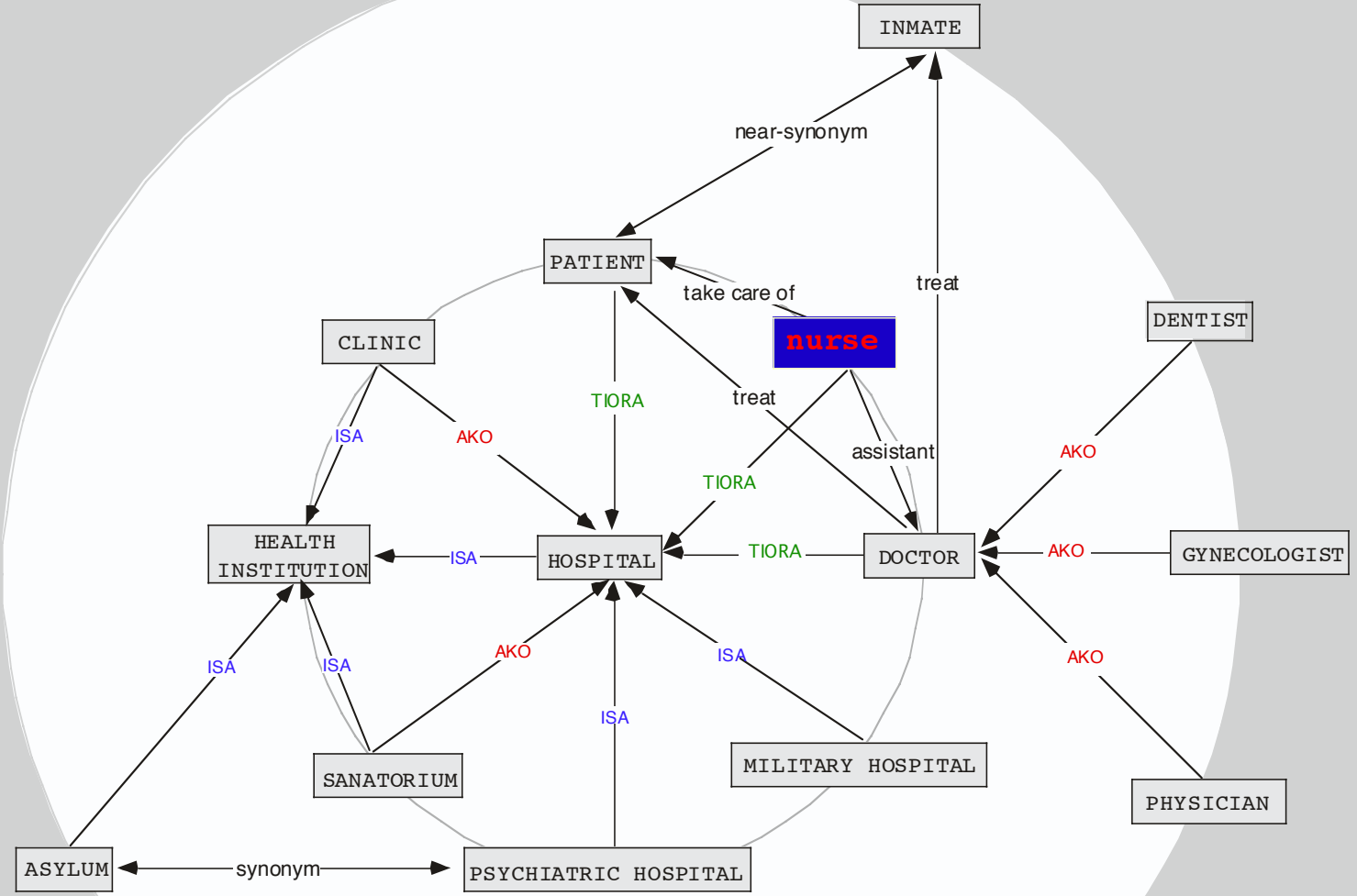
Search scenario

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

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

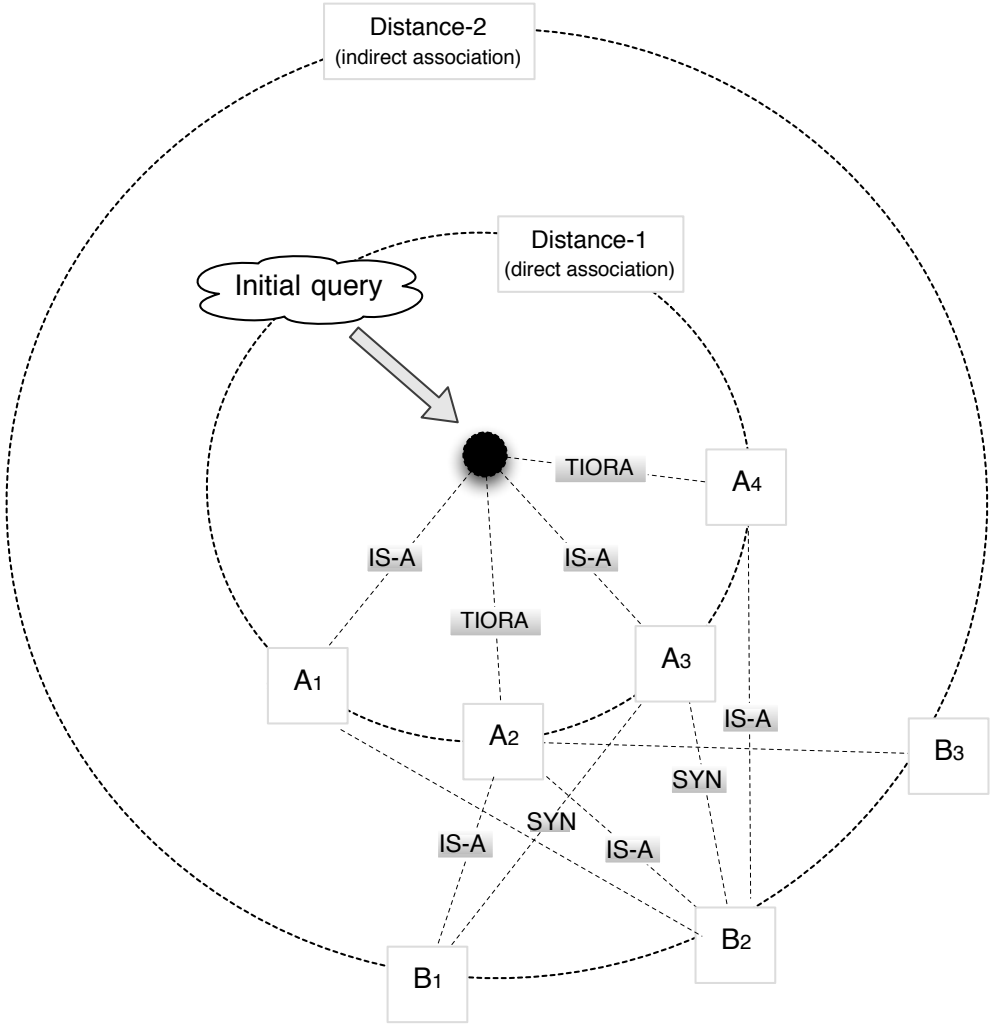


Internal Representation



Links must be interpretable to allow for navigation

Potential problems with graphs: crossing links with indirect neighbours. IS-A (subtype); TIORA (Typically Involved Object, Relation or Actor: tools, employees)



Interface for search : 3 islands of clustered words in response to the input *computer*

Links- or cluster names : ISA, used_for, accessory

ISA (subtype, instance_of)

- Mac
- PC
- else

accessory

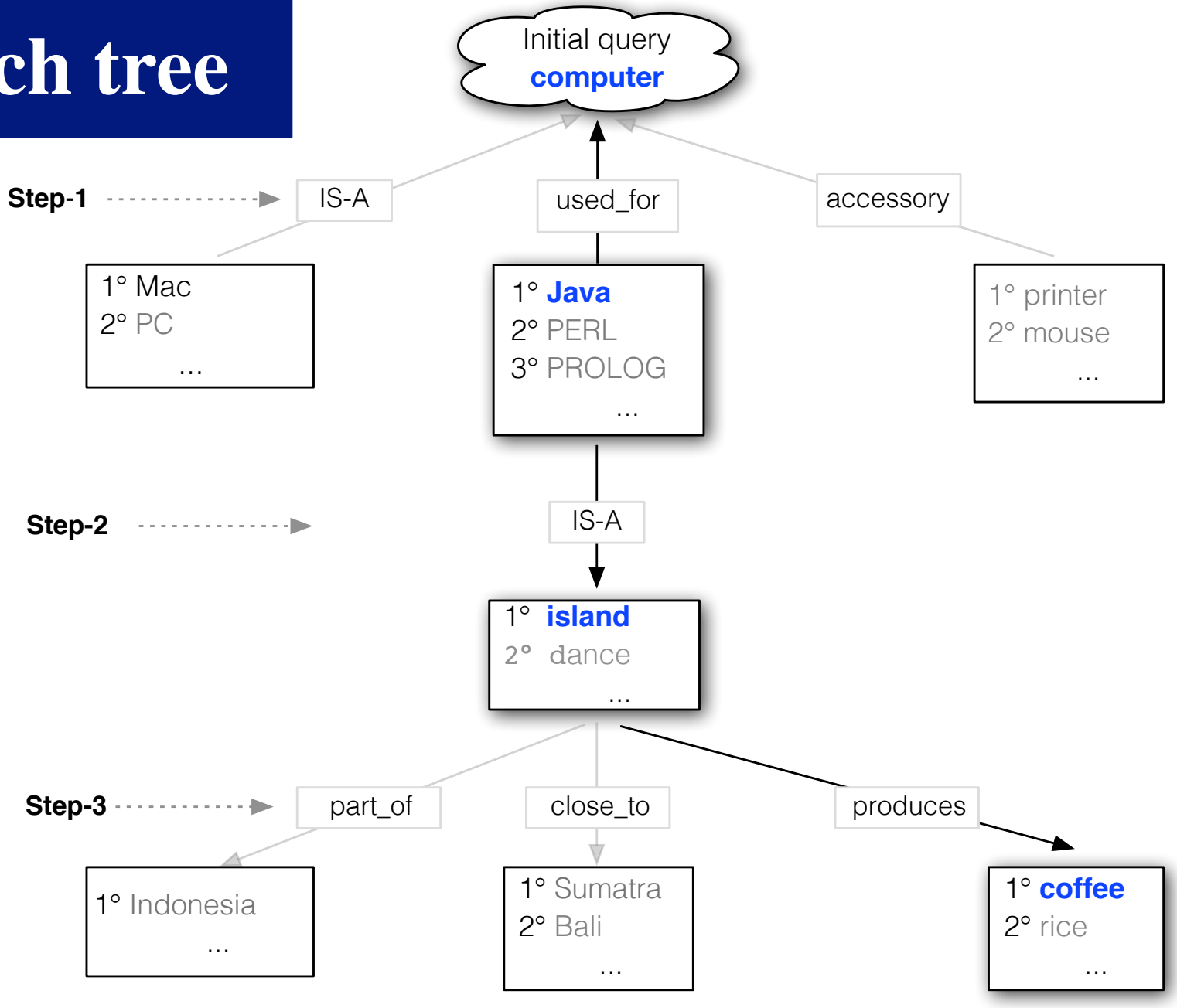
- mouse
- printer
- else

used_for (tool)

- JAVA
- PERL
- PROLOG
- Else

} Potential target words

Search tree


Initial query
computer
Step-1
IS-A
used_for
accessory
1° Mac
2° PC
...
1° Java
2° PERL
3° PROLOG
...
1° printer
2° mouse
...
Step-2
IS-A
1° island
2° dance
...
Step-3
part_of
close_to
produces
1° Indonesia
...
1° Sumatra
2° Bali
...
1° coffee
2° rice
...

Why do we need a well balanced corpus?

Corpus should be well balanced in order to represent our world-knowledge (encyclopedic and episodic knowledge)

- **encyclopedic knowledge**
  - ⇢ New Delhi – capital_of – India

    (stable knowledge, shared by many people)
- **episodic knowledge**
  - ⇢ Nadal – winner_of - French Open

    (fact likely to change over time, shared by a smaller group of people)

Goal

Extract automatically **relational** information from corpora

- corpora are an externalized version of our brain, as they contain episodic + encyclopedic knowledge
  - Paris – capital of – France (stable)
  - Nadal – winner of - French Open (recent event, subject to change)
- knowledge, i.e. corpora change dynamically
  - fast updating
  - data mining
  - index creation
  - define search patterns

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.

While it is probably difficult to do much better than to rely on the words composing the definition (meaning) of the target word (plan A, the normal route), a lot can be done to help the user to find the target via an associated concept or word (plan B).

Conclusion

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

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

Thanks for hanging in!

Just one more talk before hanging out!