# WordNets and Beyond:

## the Case of Lexical Access in Speaking or Writing

**Michael Zock** & **Didier Schwab**

Aix-Marseille university (LIF-CNRS)
Campus de Luminy, **Marseille**, France
michael.zock@lif.univ-mrs.fr

University Grenoble Alpes (LIG – GETALP)
Campus de **Grenoble** / France
didier.schwab@imag.fr

# Knowledge is power



# Knowledge concerning words

1. **Lexical competency**, i.e. knowledge of words (word forms).
   - You can only find what is there. The target must be in the knowledge-base.
   - Information associated to words : meaning, grammar, ...
2. **Metaknowledge**
   - existence of words (*tomorrance* → *tomorrow*)
   - organization (topology: direct neighbors; moving window/flashlight on the map of the mental lexico)
   - direction to go in order to continue search
1. **Cognitive states**:
   - *unpredictable*;
   - *variable* : different from person to person + moment to moment;
   - *fragmentary* (Tip of the Tongue)

# Knowledge is power

Provided that you know how to *organize*, *access* and *use* (control) it. Knowledge is a prerequisite to

- **thinking** (conceptualization, feeding thought)
- **communication** (allowing to express the *process* and *products* of thought)

# **Knowledge** is **power,**
# **provided** that you **can access it.**

Our ***focus***: *lexical access* by a human dictionary user being in the *production mode* (speaker/writer).

- **Challenge**: find the target among a large number of words, i.e. find the *needle* in a *haystack* (time + effort).
- **Difference :** *storage* vs. *access*

**Needles** in a **haystack** and **how** to **find** them?

Building a **resource** to help **authors** (speakers/writers)

to **overcome** the  **problem**

# The importance of concepts and words

**Words** are to communication what **concepts** are to thought : they are neither the *process* nor the *product*, but they are the fuel allowing the realization of both.

**Words** are to language what **bricks** are to houses. They are neither the *building* nor the *method* for creating it, they are 'only' the building blocks allowing us to build the house. Hence words are important.

# The importance of the words' organization

**Searching** for a word in a dictionary *without* a good index is like **searching** for a location on an island *without* a decent map.

# Problem

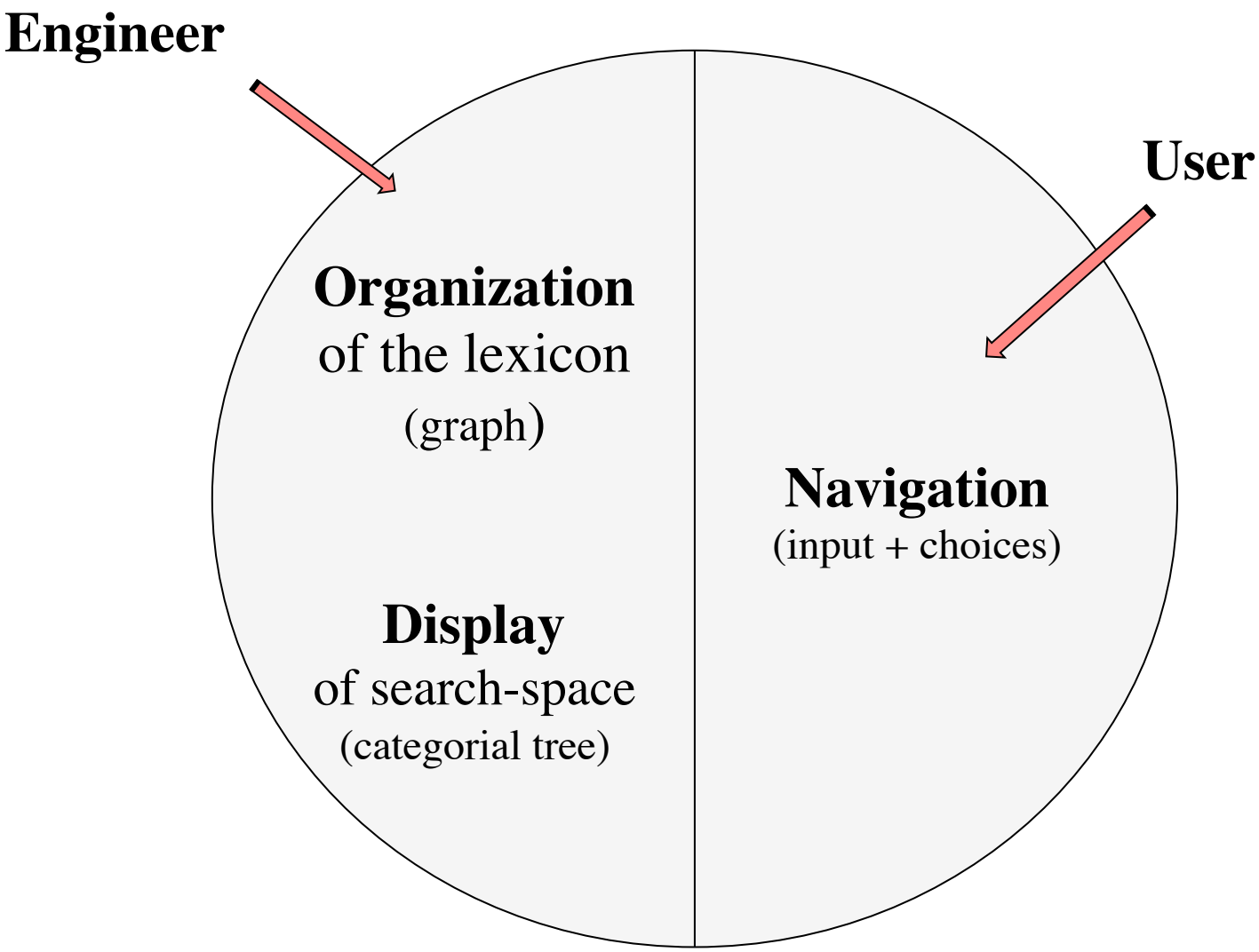
Lexical access, i.e. **wordfinding**
when *speaking* or *writing*
(deliberate, off-line)

Three perspectives :
man, machine, man-machine

# Some questions addressed

1. What does it mean to **access** a word?
2. **Where** and **how** to search ?
3. What kind of **tools** do we need (map/compass)?
4. How must the lexicon be **organized** to allow for **quick access**?
5. What are the tricks, i.e. magic short-cuts, allowing us to reach basically all words via very **few** (2-4) mouseclicks?

# Major points

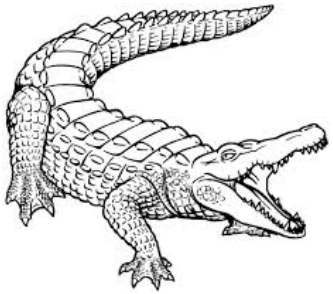


# So, our **problem**: lexical access

But what does that mean?

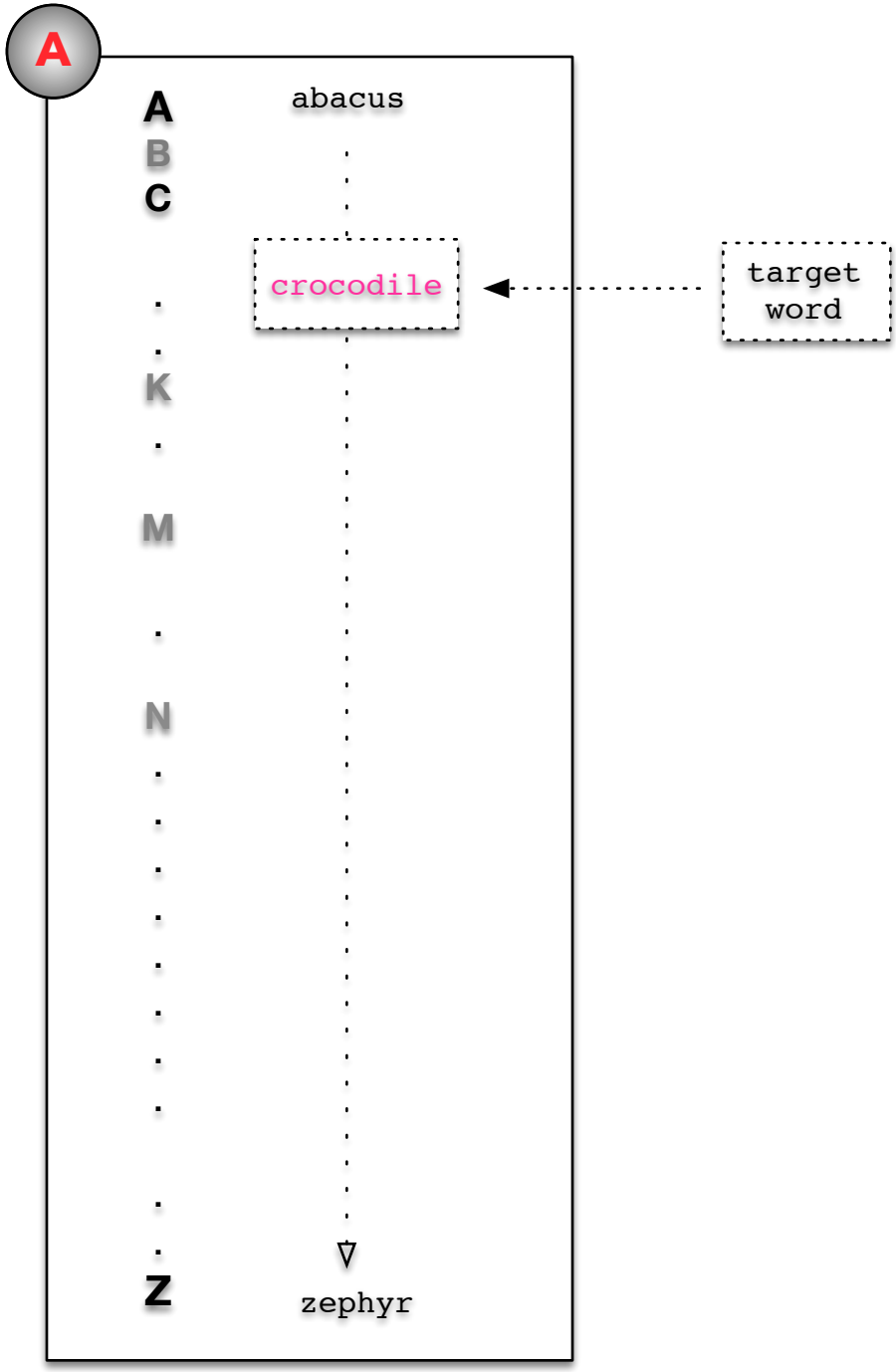
**Find** in the lexicon the target,
i.e. **reduce** the *entire* lexicon to 1, the **elusive** word.

# **Idea** to express



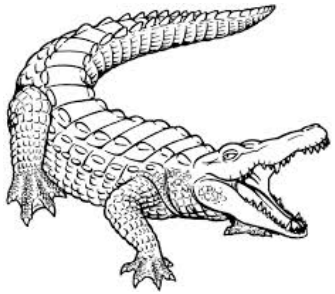
**Search entire lexicon**
i.e. reduce the
whole set to 1

Hypothetical alphabetically
organized lexicon
containing 60.000 words



**Input**

**Output**

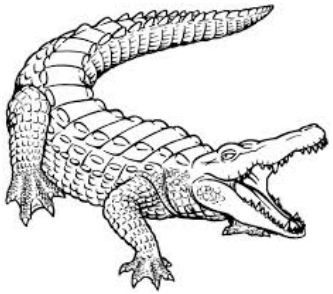


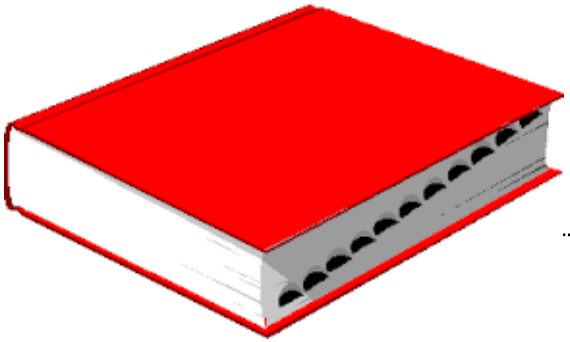
crocodile

....

**Input**

**Output**





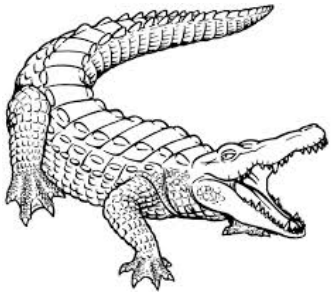
- alligator
- caiman
- crocodile

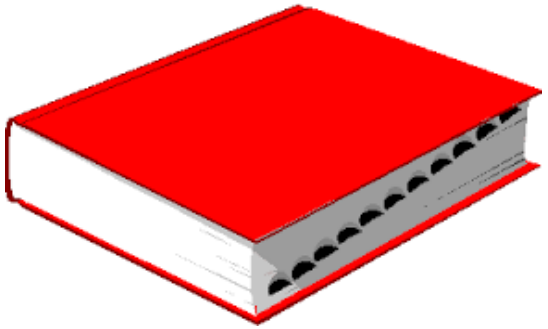
….

**Mediator**

**Visual Input**

**Output**





- alligator
- caiman
- crocodile

….

reptile

**Linguistic** input
(underspecified)

Mediator

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

It is **not** because something is **stored** that it can (always/ readily) be **accessed**

- people (amnesia, anomia, TOT, etc.)
- machines (quality of query, organization, etc.)

# Problem

**`Input`**:

What was his name again?

**`Think_of`**: umbrella

**`Target`** : Mandela

# Problem

Gosh! Where did I put my <object> ?

object:
keys, glasses,
passport, …

# Problem

**Input**:

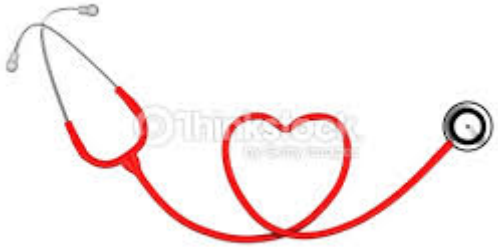
What is the **word** for this nocturnal mammal with long ears and a snout which feeds on termites and inhabits the grasslands of Africa?

**Target** : aardvark

# Problem

"The doctor listened to her chest with his *periscope*"

**`Target`** :



stethoscope

# Consequences

1. **Disruption** (hesitation, silence)
2. **Nuisance** in social life: how come that he doesn't **remember me** ?

**Problems**, problem, problems, ….

# Any solutions on the horizon ?

We believe that the answer is **YES**, and we will try to show how this could be done by improving an existing electronic dictionary.

But before that, a few preliminary comments.

# Major search scenarios

| Known | Goal | Type of search |
|---|---|---|
| word form | meaning,<br>spelling,<br>Grammar | semasiological |
| meaning | word form | onomasiological |

# Major search scenarios

| Known | Goal | Type of search |
|---|---|---|
| shamrock |  | semasiological |
|  | word form<br>shamrock | onomasiological |

# Different kinds of accessible information

| concept | sound | related information |
| --- | --- | --- |
| small plant with 3 round green leaves on each stem | sham - jam<br>rock – knock<br>shamrock - Sherlock | happy<br>good luck<br>Ireland<br>Holmes |

# Consider the following

- Unlike in *reading* (semasiological search) **word look-up** in *speaking* or writing (onomasiological search) is **indirect**. Since we don't know the word (it is our goal) we can access it only via another word.
- Search is **not** performed in the *entire* lexicon, but only in a part of it (called 'search space'), which is built **dynamically** on the basis of the information currently available in the user's mind. This can be of any sort and vary from moment to moment (imagine a major event: 9/11, 2001), and person to person.
- This being so it is important to determine properly the search-space (content, size, relevance).

# Build the search-space

**Based** on **what**?

- definitions (bag of words)
- co-occurrences (typical associations) – our current focus !!!
- topics (Roget's Thesaurus)

# Three problems

Build the **search space**

- *size* (not too big, not too small)
- *content*: contain potentially relevant data (density)

**Organization** + **presentation**

- graph => categorial tree
- *scope*: direct neighbor (limitations of screen size, danger to drown the user;)

**Metalanguage** (understandable by ordinary user)

# Characteristics

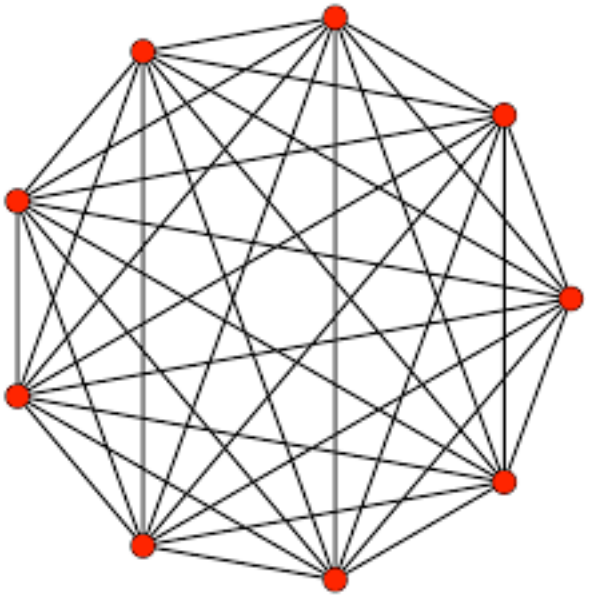
- **Fully connected graph** (all words are reachable, regardless of the entry point).
- **Content** (size, content: contain potentially relevant data)
- **Based** on **user's current knowledge state**
- **Redundancies** (flexibility: the same word can be reached via different routes) – hence, our network is different from a traditional dictionary which contains a given lemma with a specific sense only once (e.g. 'rose').

# The problem of organizing the lexicon or the search space

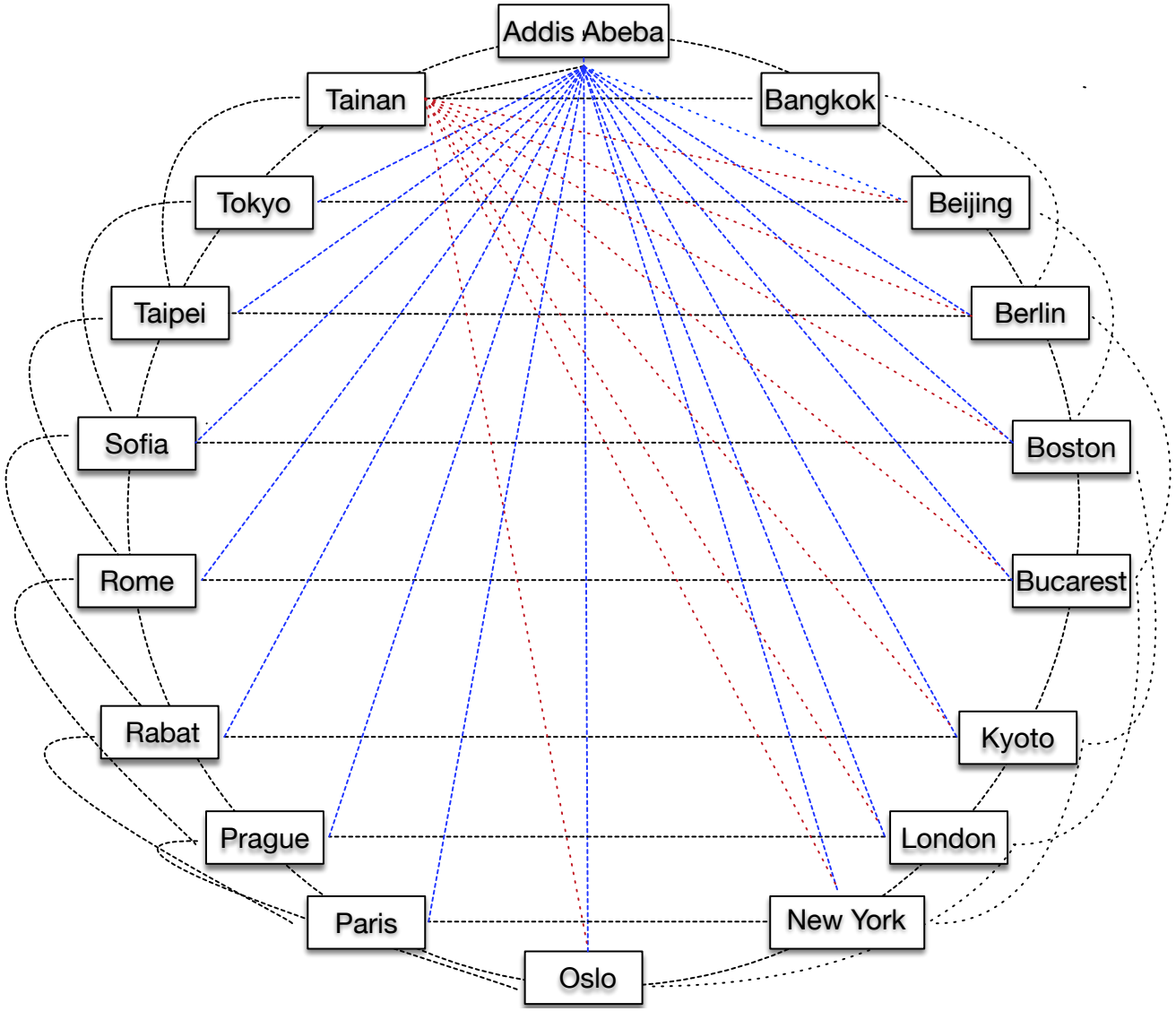
Alphabetically, topically, statistically

# Fully connected graph

Everything
is accessible
from anywhere



# Fully connected graph



**Goal**

Tōkyō

**Input**

Kyōto

1. Addis Ababa
2. Bangkok
3. Beijing
4. Berlin
5. Boston
6. Bucarest
7. Kyōto
8. London
9. New York
10. Oslo
11. Paris
12. Prague
13. Rabat
14. Rome
15. Sofia
16. Taipei
17. Tōkyō
18. Tainan

**Various ways of organizing the data**:
continent, country, size, alphabetic, … (mixed form)

# The problem with statistics

# **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 |

# **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** 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 |

# It is not because a word is stored that we can access it under all circumstances

This holds not only for *humans* (TOT-problem), but also for *machines*;

Much depends on the **quality** of the *query* and the way the *resource* is built

# Believe it or not, even machines can fail

It all depends on the quality of

- the resource
- the query
- the search method

# Automatic comparison of output produced for different 'corpora'

- eXtended WordNet
- Wikipedia

# WordNet : 2 ways of using it

- via *machine* (write some algorithm to make WN comply with it)

  "WordNet is an online lexical database designed for use under program control." (Miller, 1995, p. 39).

- via the *GUI* (this is the case *we are concerned with here*) :

  http://wordnetweb.princeton.edu/perl/webwn

# Evaluation of system performance

Critical variables

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

Relative success

- to find the desired target word
- speed
- accuracy

# **Target:** vintage


# WordFinder

**Welcome** to the **WORDFINDER** Webpage

**Input** harvest wine grapes | send

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

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

# **Comparison** of **query terms** and **ressources**

| **Input** (query) | **eXtended WN** | **Wikipedia** |
|---|---|---|
| wine | **488 words:** grape, sweet, serve, France, small, fruit, dry, bottle, produce, red,... | **3045 words** name, christian, grape, France, ... vintage (81st), ... |
| harvest | **30 words** month, fish, grape, revolutionary, calendar, festival, 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) |

**Comparison** of **query**:
wine better than harvest; wine + harvest together better than either of them alone.

**Comparison** of **resource**: Wikipedia better than WN in all three cases.

# Under what conditions can WN be used for consultation?

1° The *author* ***knows*** the *link* holding between the source word (input) and the target, e.g.

([dog]+*synonym* = [?] → [bitch]);
([dog]+*hypernym* = [?] → [canine]);

2° The *input* (source word) and the *target* are ***direct neighbors*** in the resource. For example,

[seat]-[leg] (*meronym*);
[talk]-[whisper] (*troponym*), ...

3° The *link* is ***part*** **of** WN's database :

'more specific', 'more general',... are part of it,
'better than; famous_for' are not.

# Under what conditions is WN not really good for consultation?

1° The source (input) and the target are only *indirectly* related, the distance between the two being greater than 1. This would be the case when the target ('Steffi Graf') cannot be found directly in reponse to some input ('tennis player'), but only via an additional step, say, 'tennis pro' :

([tennis player] → [tennis pro]);

input at next cycle:

([tennis pro] → [Steffi Graf])

2° The input ('play') and the target ('tennis') belong to different parts of speech. This if often referred to as the 'tennis problem' (Chafe in Fellbaum, 1998);

# Under what conditions is WN not really good for consultation?

3° The prime and the target are linked via a *syntagmatic association* ('smoke'-'cigar'). Since the majority of relations used by WN connect words from the same part of speech, word access is difficult if the output (target) belongs to a different part of speech than the input (prime);

4° The user ignores the link, he cannot name it, or the link is not part of WN's repertory. This holds true (at least) for nearly all syntagmatic associations;

# Beware

Efforts have been made though to add *domain* information and *syntagmatic* links, but they are not integrated in the version that is accessible via theweb interface. Yet this is the one accessed by the ordinary language user.


- Bentivogli, L. & Pianta, E. (2004). *Extending WordNet with Syntagmatic Information*. Sojka, P., Pala, K., Smrz, P., Fellbaum, C. & Vossen, P. (Eds.): Global Wor(l)dNet Conference, Proceedings, pp. 47-53. Masaryk University, Brno
- Bentivogli, L., Forner, P., Magnini, B., & Pianta, E. (2004). Revising the wordnet domains hierarchy: semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Ressources* (pp. 101-108). Association for Computational Linguistics.

# Some more references


- Boyd-Graber, J., Fellbaum, C., Osherson, D. & Schapire, R. (2006). *Adding Dense, Weighted, Connections to WordNet*. Proceedings of the Global WordNet Conference.
- Gliozzo, A. & Strapparava, C. (2008). Semantic domains in computational linguistics. Springer.
- Nikolova, S., Tremaine, M., & Cook, P. R. (2010). Click on bake to get cookies: guiding word-finding with semantic associations. In *Proc. of the 12th int. ACM SIGACCESS conference on Computers and accessibility* (pp. 155-162).

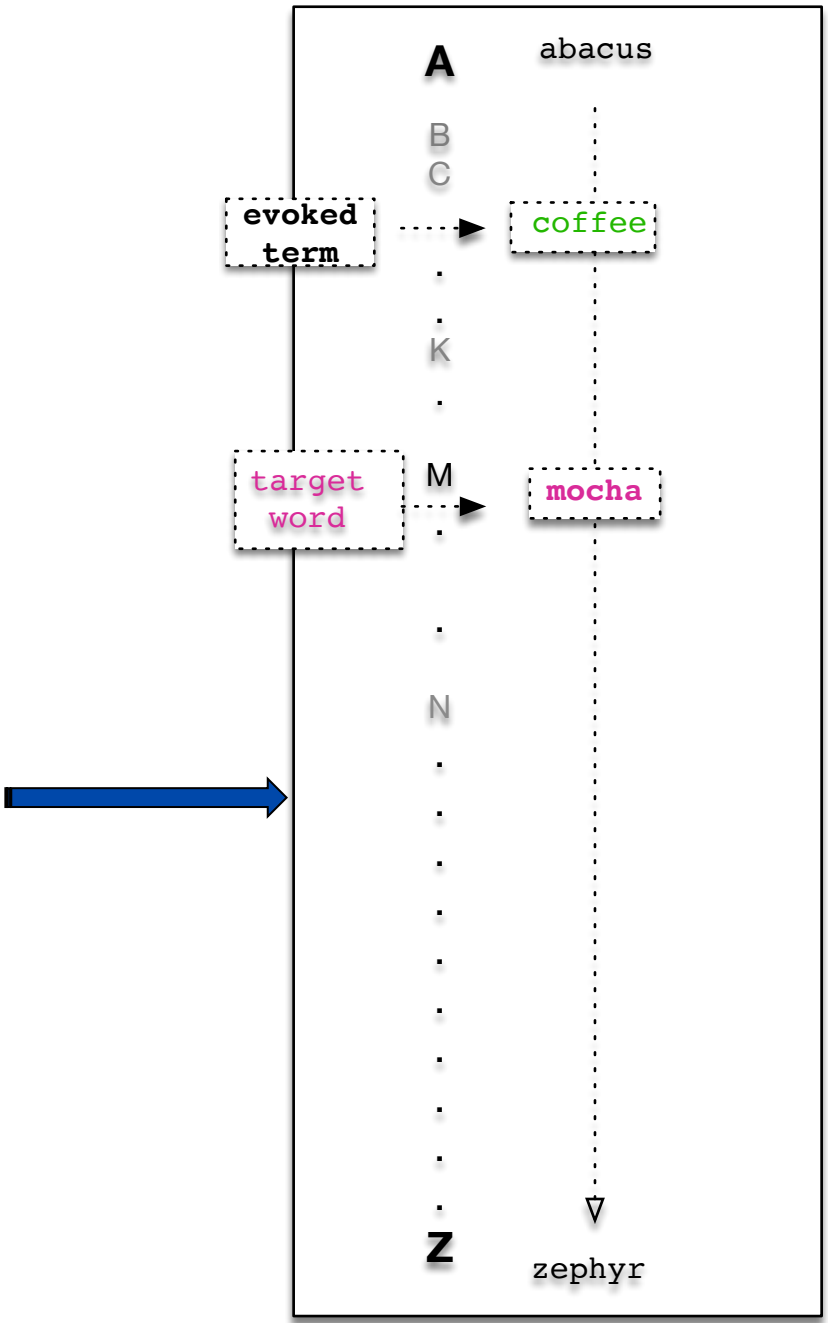
# WN and beyond

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

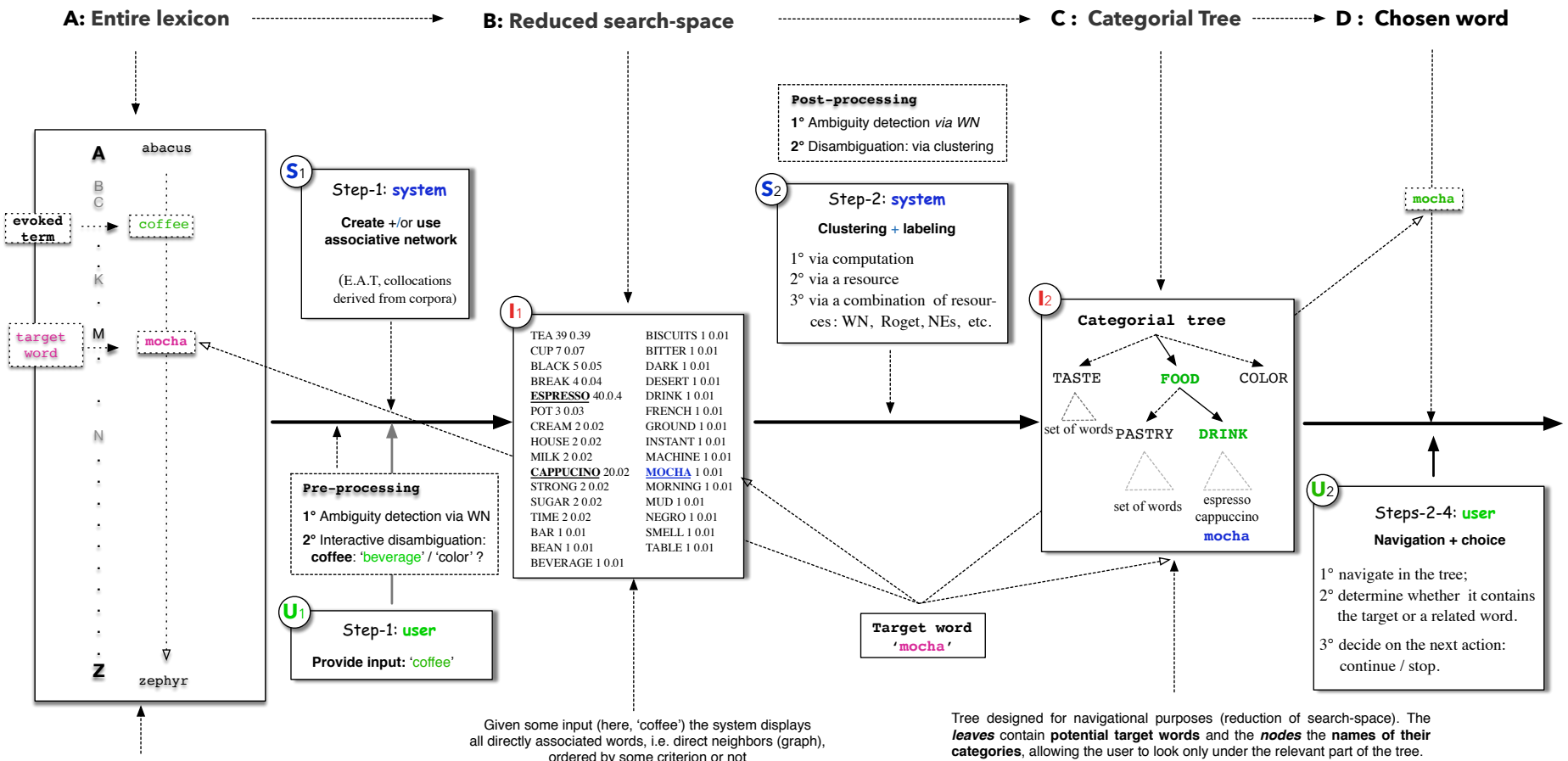
**Idea** to express

## Entire lexicon

Hypothetical size:
60.000 words



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




Hypothetical lexicon containing 60.000 words

Given some input (here, 'coffee') the system displays all directly associated words, i.e. direct neighbors (graph), ordered by some criterion or not

Tree designed for navigational purposes (reduction of search-space). The ***leaves*** contain **potential target words** and the ***nodes*** the **names of their categories**, allowing the user to look only under the relevant part of the tree. Since words are grouped in named clusters, the user does not have to go through the whole list of words anymore. Rather he navigates in a tree (top-to-bottom, left to right), choosing first the *category* and then its *members*, to check whether any of them corresponds to the desired target word.

**S**: *system*, **U**: *user*, **I**: *interaction*

Lexical access as a three-step process

(provide input, navigate and choose then among the possible outputs)

# Conclusion

We have presented here some ideas of how to build a resource likely to help authors to overcome the TOT-problem.

We 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

**Lexical access** was conceived as a three-step process. *Engineers* have to organize words (association network), determine the space within search takes place given some user input, and present them then the result in a manageable way (categorial tree). The *user* has to provide an input, which will determine the space within which search takes place (direct neighbors of the input), determine the category in which to look for the target and decide then whether to stop or to continue search.

So far we have only presented a *roadmap*. The next step should be, of course, to build the resource. To this end one can combine and experiment with existing resources (BabelNet, ConceptNet, word definitions, topic maps,…), and while building the *semantic net* within which search takes place seems feasable, the second step (clustering and naming the clusters, i.e. building the categorial tree) is quite a bit more of a challenge.

# Thanks for hanging in!

Please enjoy the next talk(s)