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 »

Some facts

Spontaneous speech

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

Underlying process

- remarkably efficient
- search in a huge lexical data-base (50,000-100,000 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

form

sound

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?

<table>
<thead>
<tr>
<th>Navigational instrument</th>
<th>sextant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instrument used in Asia for eating</td>
<td>chopsticks</td>
</tr>
<tr>
<td>Hat of a bishop</td>
<td>mitre</td>
</tr>
</tbody>
</table>
Example : name of a person
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

<table>
<thead>
<tr>
<th>Input</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Will you draw me a</strong></td>
<td>woolly usually horned ruminant mammal related to the goat</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Semantic candidates</th>
<th>mutton, ram, ewe, lamb, sheep, goat, bovid, ovis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonological candidates</td>
<td>cheap, jeep, schliep, seep, sheep, sleep, steep, streep, sweep</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Output</th>
<th>/ʃiːp/ - sheep</th>
</tr>
</thead>
</table>
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.
Welcome to the WORDFINDER webpage

Input

harvest wine grapes

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)

Comparison of different search algorithms

Input (successive input of trigger words): honey

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

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)
- guest (1281)
# Comparison of two resources

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

- **Head Word**
- **Pronunciation**
- **Part-of-Speech**
- **Quote Year**
- **Quote Author**
- **Quote Source**
- **Quotation**
- **Permalink For Entry**
- **Author Links**
- **Email Word To A Friend**

- **Definition**

```
characterized by divisive tendencies; the propensity to divide or break
away from a larger body or group

1868 CHARLES DARWIN The Variation of Animals and Plants Under
Domestication Such cases as that of the Hydra are evidently analogous to the
spontaneous division or fissiparous generation of the lowest animals, and likewise to the
budding of plants.

1911 JOHN PROPORTIONAL REPRESENTATION I

1934 EDITH WHARTON A

2003 MIDDLE EARTH Evil is fissiparous.
```

Labels: Charles Darwin, Edith Wharton, J. R. R. Tolkien, John H. Humphreys

Email this word to a friend:
Words as objects vs. words as nodes in a network

HEADWORD-1
- DEFINITION
- TRANSLATION
- SYNONYM
- GRAMMAT. INF.
- USAGE
  - etc.

HEADWORD-2
- DEFINITION
- TRANSLATION
- SYNONYM
- GRAMMAT. INF.
  - etc.

etc.
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
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?

He is rather silly.

He is rather **stilly**.
Lexical access as puzzle completion
"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."

"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*
Meaning

visual input

- means of transportation
  - phonological word
    - phonemes
      - \[\text{ʃ} \quad \text{i} \quad \text{p}\]

- commercial location
  - phonological word
    - phonemes
      - \[\text{ʃ} \quad \text{ɒ} \quad \text{p}\]
Levelt’s model

Visual input

Conceptual level

wool grows milk gives animal

SHEEP

goes gives

is an

animal

SHEEP

GOAT

milk

SHEEP

GOAT

noun

sense

gender

part of speech

male

female

sheep (mouton)

gender

male

female

noun

sense

sheep (mouton)

gender

male

female

noun

sense

goat (chèvre)

spoken word "sheep"

spoken word "goat"

Lexeme or sound level (output)

spoken word "sheep" spoken word "goat"

Lexeme or sound level (output)

Levelt’s model

Visual input

Conceptual level

wool grows milk gives animal

SHEEP

goes gives

is an

animal

SHEEP

GOAT

milk

SHEEP

GOAT

noun

sense

gender

part of speech

male

female

sheep (mouton)

gender

male

female

noun

sense

goat (chèvre)

spoken word "sheep"

spoken word "goat"

Lexeme or sound level (output)

spoken word "sheep" spoken word "goat"
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
Activation acts blindly: all neighbours are activated equally

$$\implies$$ non-target nodes become activated and remain so for a while

Activation acts in a deterministic fashion

$$\implies$$ we cannot escape it
Can we use this for dictionary consultation?

• Answer: no
• While computational psycholinguists can tune the weights to have their model mimic 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

- **scene** (visual input)
- **translation** equivalent word in another language: *cat-gato*
- **syntactic patterns** words in context: *animal that makes <sound>?* sound: moo --> **cow**
- **lexical relations** synonyms, antonyms, hypernyms, ...
- **sounds**
- **meanings**
- **concepts** word definitions, conceptual primitives
- **semantic fields**: thesaurus- or domain relations: people, sports, food, ...
- **encyclopedic relations** syntagmatic associations: rose-red
Build an index
but, what kind of?
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 Unievrtisy, it deosn't mttae 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**
Let me guess:
You went to Jamaica for your holidays...
Activation (association)

1. By context: \( \text{bread} \Rightarrow \text{butter} \)
2. By meaning: \( \text{bread} \Rightarrow \text{food} \)
3. Via form: \( \text{bread} \Rightarrow \text{red}, \text{historical} \Rightarrow \text{hysterical} \)
4. Via the meaning/context + the form:
   \( \text{cat} \Rightarrow \text{rat}; \)
   \( \text{DSK} \Rightarrow \text{election} : \)
   \( \text{election} \Rightarrow \text{erection} \) (phonological neighbour)
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 *(hyponym)*
    - flower-rose *(hyponym)*

- **Crossing links**
Input: India

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

<table>
<thead>
<tr>
<th>Country</th>
<th>Score</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pakistan</td>
<td>12</td>
<td>0.14</td>
</tr>
<tr>
<td>Rubber</td>
<td>10</td>
<td>0.12</td>
</tr>
<tr>
<td>China</td>
<td>4</td>
<td>0.05</td>
</tr>
<tr>
<td>Foreign</td>
<td>4</td>
<td>0.05</td>
</tr>
<tr>
<td>Curry</td>
<td>3</td>
<td>0.04</td>
</tr>
<tr>
<td>Famine</td>
<td>3</td>
<td>0.04</td>
</tr>
<tr>
<td>Tea</td>
<td>3</td>
<td>0.04</td>
</tr>
<tr>
<td>Country</td>
<td>2</td>
<td>0.02</td>
</tr>
<tr>
<td>Ghandi</td>
<td>2</td>
<td>0.02</td>
</tr>
<tr>
<td>Wogs</td>
<td>2</td>
<td>0.02</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Africa</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Air</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Asia</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Black</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Brown</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Bus</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Clive</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Colonial</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Company</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Coons</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Cows</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Eastern</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Empire</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Fame</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Flies</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Himalayas</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Hindu</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Hunger</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Immigrants</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Indians</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Japan</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Khaki</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Man</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Missionary</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Monsoon</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Patriarch</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>People</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Persia</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Poor</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>River</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Sari</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Star</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Starvation</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Starve</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Ten</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Triangle</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Turbans</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Tyre</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Under-developed</td>
<td></td>
<td>0.01</td>
</tr>
</tbody>
</table>
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**

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

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

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAKISTAN</td>
<td>12.01</td>
<td>Flies</td>
</tr>
<tr>
<td>RUBBER</td>
<td>10.12</td>
<td>Himalayas</td>
</tr>
<tr>
<td>CHINA</td>
<td>4.05</td>
<td>Hindu</td>
</tr>
<tr>
<td>FOREIGN</td>
<td>4.05</td>
<td>Hunger</td>
</tr>
<tr>
<td>CURRY</td>
<td>3.04</td>
<td>Immigrants</td>
</tr>
<tr>
<td>FAMINE</td>
<td>3.04</td>
<td>Indians</td>
</tr>
<tr>
<td>TEA</td>
<td>3.04</td>
<td>Japan</td>
</tr>
<tr>
<td>COUNTRY</td>
<td>2.02</td>
<td>Khaki</td>
</tr>
<tr>
<td>GHANDI</td>
<td>2.02</td>
<td>Man</td>
</tr>
<tr>
<td>WOGS</td>
<td>2.02</td>
<td>Missionary</td>
</tr>
<tr>
<td>AFGHANISTAN</td>
<td>1.01</td>
<td>Monsoon</td>
</tr>
<tr>
<td>AFRICA</td>
<td>1.01</td>
<td>Persia</td>
</tr>
<tr>
<td>AIR</td>
<td>1.01</td>
<td>Poor</td>
</tr>
<tr>
<td>ASIA</td>
<td>1.01</td>
<td>River</td>
</tr>
<tr>
<td>BLACK</td>
<td>1.01</td>
<td>Sari</td>
</tr>
<tr>
<td>BROWN</td>
<td>1.01</td>
<td>Star</td>
</tr>
<tr>
<td>BUS</td>
<td>1.01</td>
<td>Starvation</td>
</tr>
<tr>
<td>CLIVE</td>
<td>1.01</td>
<td>Starve</td>
</tr>
<tr>
<td>COLONIAL</td>
<td>1.01</td>
<td>Ten</td>
</tr>
<tr>
<td>COMPANY</td>
<td>1.01</td>
<td>Triangle</td>
</tr>
<tr>
<td>COONS</td>
<td>1.01</td>
<td>Turbans</td>
</tr>
<tr>
<td>COWS</td>
<td>1.01</td>
<td>Tyre</td>
</tr>
<tr>
<td>EASTERN</td>
<td>1.01</td>
<td>Under-developed</td>
</tr>
<tr>
<td>EMPIRE</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td>FAME</td>
<td>1.01</td>
<td></td>
</tr>
</tbody>
</table>
India being the **answer** to the following **stimuli**

<table>
<thead>
<tr>
<th>SARI</th>
<th>22 0.15</th>
<th>CAST</th>
<th>1 0.01</th>
<th>LOOP</th>
<th>1 0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASTE</td>
<td>17 0.12</td>
<td>CATASTROPY</td>
<td>1 0.01</td>
<td>MEDITATE</td>
<td>1 0.01</td>
</tr>
<tr>
<td>NADIR</td>
<td>11 0.08</td>
<td>CHINA</td>
<td>1 0.01</td>
<td>MILDEW</td>
<td>1 0.01</td>
</tr>
<tr>
<td>FAMINE</td>
<td>6 0.04</td>
<td>CLUBS</td>
<td>1 0.01</td>
<td>MISSION</td>
<td>1 0.01</td>
</tr>
<tr>
<td>AFRICA</td>
<td>5 0.03</td>
<td>COLONIES</td>
<td>1 0.01</td>
<td>NATIVE</td>
<td>1 0.01</td>
</tr>
<tr>
<td>PAKISTANI</td>
<td>4 0.03</td>
<td>COMPANY</td>
<td>1 0.01</td>
<td>PLAGUE</td>
<td>1 0.01</td>
</tr>
<tr>
<td>STARVING</td>
<td>4 0.03</td>
<td>CONSUMPTION</td>
<td>1 0.01</td>
<td>POVERTY</td>
<td>1 0.01</td>
</tr>
<tr>
<td>BEGGING</td>
<td>3 0.02</td>
<td>CONTINENTS</td>
<td>1 0.01</td>
<td>PRIESTESS</td>
<td>1 0.01</td>
</tr>
<tr>
<td>CASTS</td>
<td>3 0.02</td>
<td>COTTON</td>
<td>1 0.01</td>
<td>QUININE</td>
<td>1 0.01</td>
</tr>
<tr>
<td>NAPALM</td>
<td>3 0.02</td>
<td>COUNTRY</td>
<td>1 0.01</td>
<td>SAVER</td>
<td>1 0.01</td>
</tr>
<tr>
<td>STARVATION</td>
<td>3 0.02</td>
<td>EAST</td>
<td>1 0.01</td>
<td>SECT</td>
<td>1 0.01</td>
</tr>
<tr>
<td>CHARISMA</td>
<td>2 0.01</td>
<td>ELEPHANT</td>
<td>1 0.01</td>
<td>SERVANT</td>
<td>1 0.01</td>
</tr>
<tr>
<td>CURRY</td>
<td>2 0.01</td>
<td>ELEPHANTS</td>
<td>1 0.01</td>
<td>SETTLEMENT</td>
<td>1 0.01</td>
</tr>
<tr>
<td>INCENSE</td>
<td>2 0.01</td>
<td>EMPIRE</td>
<td>1 0.01</td>
<td>SHEEPSKIN</td>
<td>1 0.01</td>
</tr>
<tr>
<td>KHAKI</td>
<td>2 0.01</td>
<td>FAMISHED</td>
<td>1 0.01</td>
<td>STARVED</td>
<td>1 0.01</td>
</tr>
<tr>
<td>PARIAH</td>
<td>2 0.01</td>
<td>FURTHER</td>
<td>1 0.01</td>
<td>SUFFERING</td>
<td>1 0.01</td>
</tr>
<tr>
<td>RICE</td>
<td>2 0.01</td>
<td>GHoul</td>
<td>1 0.01</td>
<td>THUG</td>
<td>1 0.01</td>
</tr>
<tr>
<td>SPICE</td>
<td>2 0.01</td>
<td>Hedonism</td>
<td>1 0.01</td>
<td>THUGS</td>
<td>1 0.01</td>
</tr>
<tr>
<td>STARVE</td>
<td>2 0.01</td>
<td>INCA</td>
<td>1 0.01</td>
<td>TIGER</td>
<td>1 0.01</td>
</tr>
<tr>
<td>TURBAN</td>
<td>2 0.01</td>
<td>INDIANS</td>
<td>1 0.01</td>
<td>TOGA</td>
<td>1 0.01</td>
</tr>
<tr>
<td>ALE</td>
<td>1 0.01</td>
<td>ISLAM</td>
<td>1 0.01</td>
<td>UNCLEAN</td>
<td>1 0.01</td>
</tr>
<tr>
<td>AMERICA</td>
<td>1 0.01</td>
<td>LEPER</td>
<td>1 0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIZARRE</td>
<td>1 0.01</td>
<td>LIFE-SPAN</td>
<td>1 0.01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. Show not only direct associations
2. but also indirectly related words
Finding a remote item at the distance $D$ of four mouse clicks ($D_4$)

**starting point** query

- **links**
  - machines
    - **Program. language**
      - Java
        - Perl
        - Prolog
      - printer
      - mouse
    - **PC**
      - brand
    - **MAC**
  - **Apple**
    - homonyme
    - fruit
    - apple
  - **coffee**
    - tourism
    - rice
  - **moccha**
    - cappuccino
    - espresso

**link chosen for further exploration**

**chosen words**

**potential target words**

**Distance $D$**

$D_1$, $D_2$, $D_3$, $D_4$
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
  - 1° Mac
  - 2° PC
  - ...

Step-2
- IS-A
  - 1° Java
  - 2° PERL
  - 3° PROLOG
  - ...

Step-3
- part_of
  - 1° Indonesia
  - ...

- close_to
  - 1° Sumatra
  - 2° Bali
  - ...

- produces
  - 1° coffee
  - 2° rice
  - ...

- accessory
  - 1° printer
  - 2° mouse
  - ...

- used_for
  - 1° printer
  - 2° mouse
  - ...

(IS-A)
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)
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
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).
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!