Motivations for a new task group devoted to lexical access

Michael Zock[1] & Dan Cristea[2,3]

[1] Aix-Marseille Université, CNRS
michael.zock@lif.univ-mrs.fr

[2] Alexandru Ioan Cuza, University of Iași
[3] Institute of Computer Science, Romanian Academy
dcristea@info.uaic.ro

Knowledge is power

Knowledge is power

Provided that you **can** access and use it (control)

We will focus here only on the former, dealing with lexical access for language production

- storage *vs*. access
- finding the needle in a haystack

Some concerns

Keep things under control

Find what you are looking for

Do so within a reasonable amount of time

Danger of getting overwhelmed

We know many words

Stay on top of the wave

Don't get drowned!

Get into the driver's seat

Take control: direction, speed

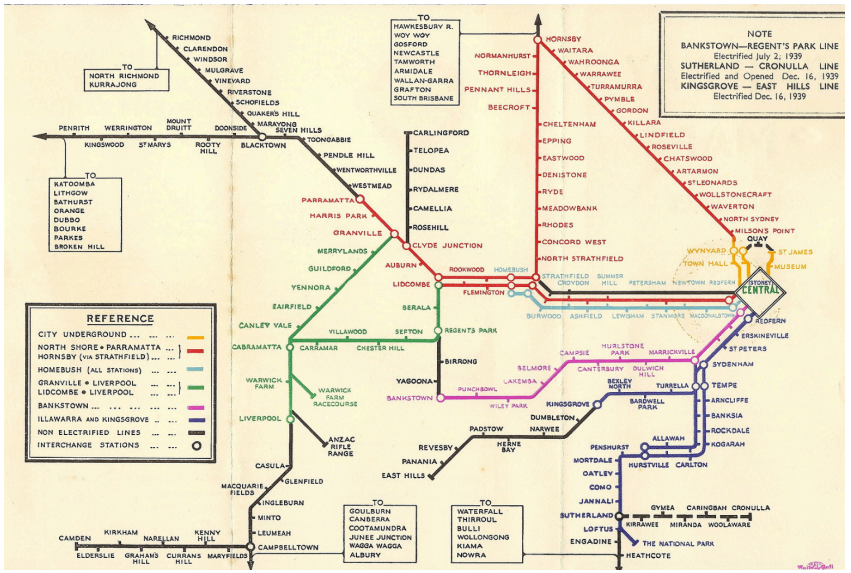
For all this we need
the right kind of tools

Tools for orientation

maps

compass

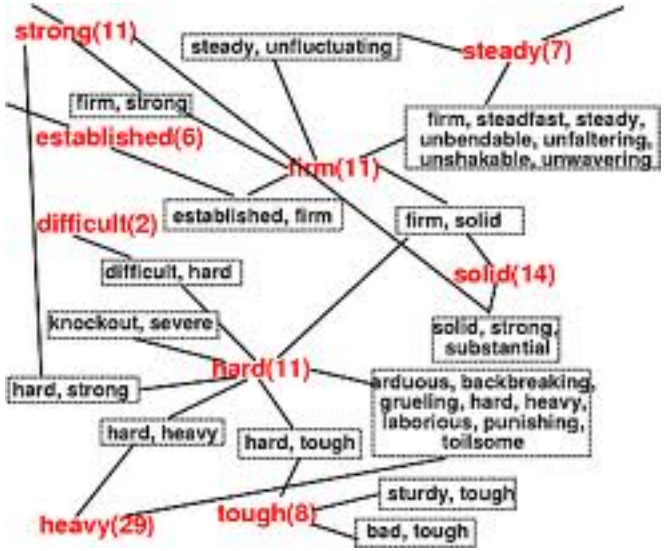
Map out the territory



See how things are related.
Where am I?
How can I get from A to B?

There are maps for many things:
cities, subways, galaxies, words

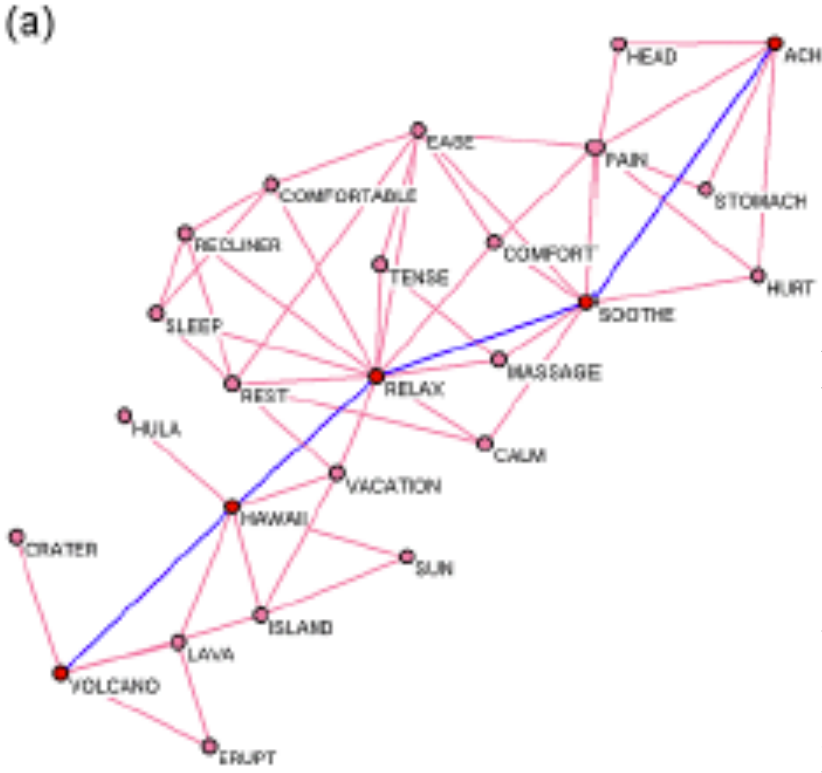
Semantic maps



See how words are related.

Association thesaurus
WordNet

Navigation in an associative network



Since **search** takes 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).

Compass

Where are we now ?
How can we reach from here the goal?

What about the Compass?

The **compass** is in **peoples' minds**.

While we have to provide them with the semantic map and the signposts (orientational guidelines; categorial tree), the decision where to go is left to the user, as he is the only one to know the target. Even if he is not able to name it, he is still able to recognize it in a list. Hence we have to present this list (in our case, the direct neighbors of the input, query word).

In the case of **word access** authors generally know fairly well where in the map is located the target word and what are its direct neighbors, as this is *generally* the one they will use as input.

Where to search

1. Dictionary
2. Index (semantic network). You start by looking at the index to find the corresponding item in the DB (typical example : Roget's Thesaurus)

Organizing words
by **topics** or **domains**

Thesaurus (Roget)

Table 2.6 Roget's system

class	section	given code
1. abstract relations	existence	1–8
	relation	9–24
	quantity	25–57
	order	58–83
	number	84–105
	time	106–139
	change	140–152
	causation	153–179
2. space	space in general	180–191
	dimensions	192–239
	form	240–263
	motion	264–315
3. matter	matter in general	316–320
	inorganic matter	321–356
	organic matter	357–449
4. intellect (the exercise of the mind)		
(1) formation of ideas	general	450–454
	precursory conditions and operations	455–466
	materials for reasoning	467–475
	reasoning processes	476–479
	results of reasoning	480–504
	extension of thought	505–513
	creative thought	514–515
(2) communication of ideas	nature of ideas communicated	516–524
	models of communication	525–549
	means of communicating ideas	550–599
5. volition (the exercise of the will)		
(1) individual volition	volition in general	600–619
	prospective volition	620–679
	voluntary action	680–703
	antagonism	704–728
	results of action	729–736
(2) social volition	general social volition	737–759
	special social volition	760–767
	conditional social volition	768–774
	possessive relations	775–819
6. emotion, religion, and morality		
	general	820–826
	personal emotion	837–887
	interpersonal emotion	888–921
	morality	922–975
	religion	976–1000

What else?

1. **Completeness** : named entities, terminology
2. Sekine's *Extended Named-Entity* Classification

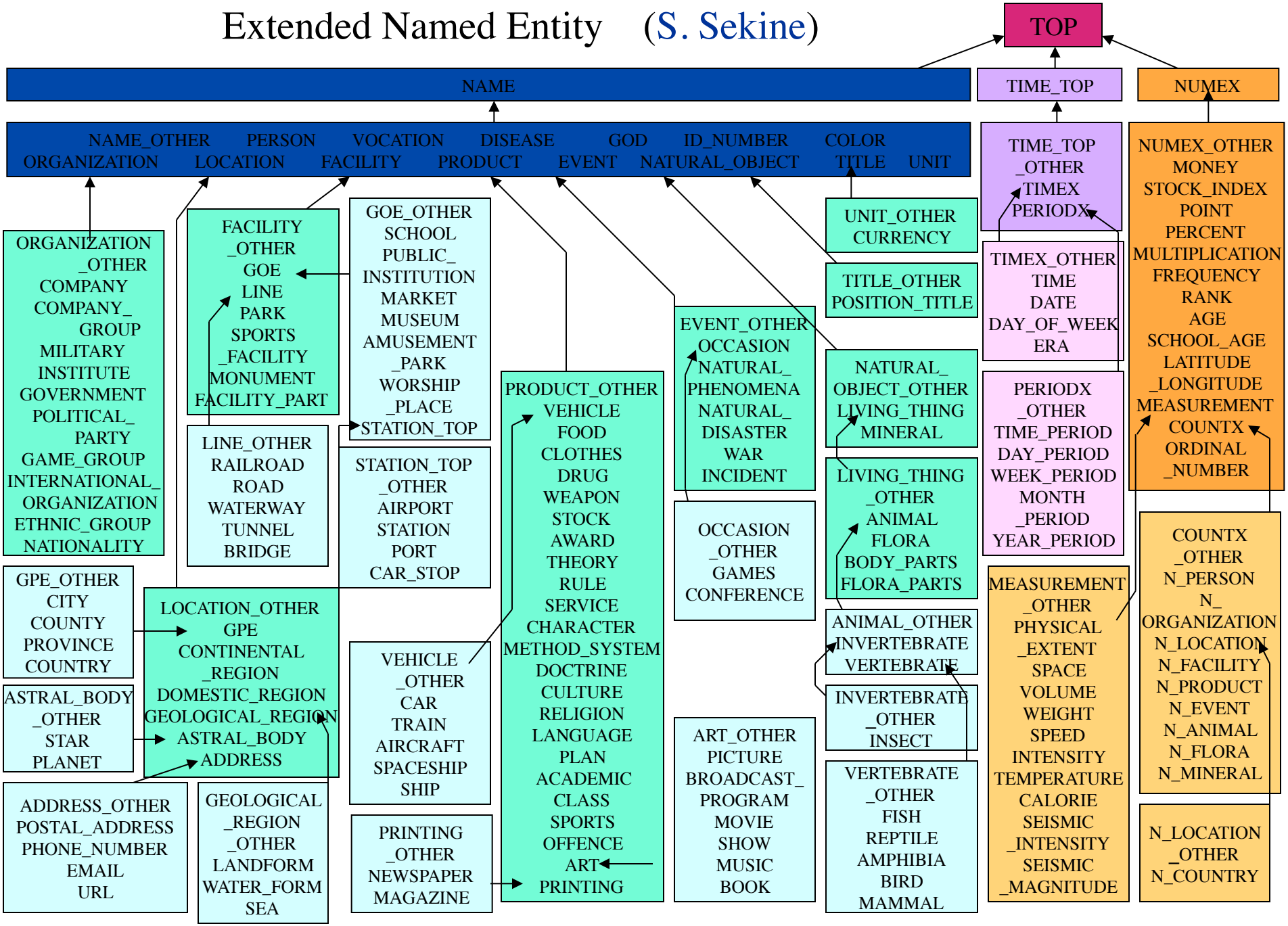
http://nlp.cs.nyu.edu/ene/

http://nlp.cs.nyu.edu/ene/version7_1_0Beng.html

Name

ENE			Examples
Name_Other			Barbaro, Bubbles, Max, Maggie
Person			Bush, Michael Jackson, Elizabeth II, LeBron Raymone James
God			Zeus, Indra, Danu, Ra
Organization	Organization_Other		the Capone Family, Department of Computer Science, CS Dept., general affairs department
	International_Organization		UN, League of Nations, Pacific Island Forum, SEATO
	Show_Organization		The Cleveland Orchestra, The Beatles, the Bolshoi Ballet troupe, Sex Pistols
	Family		The House of Hamilton, Clan Henderson, Tokugawa clan, Koga family
	Ethnic_Group	Ethnic_Group_Other	White people, Jew, Slavic peoples, Mongoloid race, Japanese Diaspora
		Nationality	Japanese, Israeli, American, American people
	Sports_Organization	Sports_Organization_Other	the Breen Gym, UCLA Bruins, Ma family army, Shinagawa Jogging Club
		Pro_Sports_Organization	New York Yankees, Seattle, NYY, Manchester United
		Sports_League	NFL, National Basketball Association, Atlantic Coast Conference, National League West
	Corporation	Corporation_Other	Association for Computational Linguistics, National Rifle Association, NHK, BBC
		Company	Toyota, SONY, CNN, Microsoft
		Company_Group	Tata Group, JR, the Big Three, Big Four auditors
	Political_Organization	Political_Organization_Other	Palestine Liberation Organization, Clinton Regime, Tokugawa shogunate, Ayyubid dynasty
		Government	National Security Council, Ministry of Finance, the United States Senate, USTR
		Political_Party	Democratic Party, Bharatiya Janata Party, Conservative Party, LDP
		Cabinet	Thatcher's Cabinet, Major's Cabinet, Tanaka's Cabinet, Koizumi's Cabinet
		Military	Self-Defense Forces, US Air Force, Royal Navy, UN forces
Location	Location_Other		Times Square, Ground Zero, Three Views of Japan, Garden of Eden
	Spa		Hakone Spa, Fukuchi Spa, Hakuba Spa, Yunoyama Spa
	GPE	GPE_Other	Taiwan, Hong Kong, Puerto Rico, French Polynesia, Macau
		City	New York City, Brooklyn, Sydney, Rio de Janeiro
		County	West Chester County, Madison County, Orange County, Shima District
		Province	Osaka Prefecture, NY, Kansas, Nova Scotia, Nagorno-Karabakh
		Country	the United States, Japan, UK, Vatican City
	Region	Region_Other	
		Continental_Region	North America, Asia, the Caribbean area, NIES
		Domestic_Region	New England, East Coast, the South, Upper New York
	Geological_Region	Geological_Region_Other	Grand Canyon, Altamira Cave, Great Barrier Reef, Ayers Rock
		Mountain	Mount Everest, K2, Mt. Fuji, Alps
		Island	Florida Keys, Key West, Gilbert Islands, Iriomote
		River	Mississippi River, Hudson River, Yangtze River, Danube
		Lake	Lake Michigan, Lake Baikal, Dead Sea, Great Lakes
		Sea	Pacific Ocean, Sea of Japan, Sunda Strait, English Channel
		Bay	Bay of Bengal, Delaware Bay, Persian Gulf, Gulf of Guinea
	Astral_Body	Astral_Body_Other	Andromeda Galaxy, Solar System, Halley's Comet, Callisto
		Star	Antares, Sirius, North Star, Barnard's Star
		Planet	Sun, Earth, Moon, Icarus
		Constellation	Taurus, Cassiopeia, Argo Navis, Lepus

Extended Named Entity (S. Sekine)



What else?

1. We do need **more** than just a ***semantic network***, even if the nodes and links are weighted. Weights are not everything. Frequency and/or recency?

2. How to (re)**present** the data (flat lists, graphs, trees)?

Two important problems

1. how to specify the input and
2. how to present the output to the user?

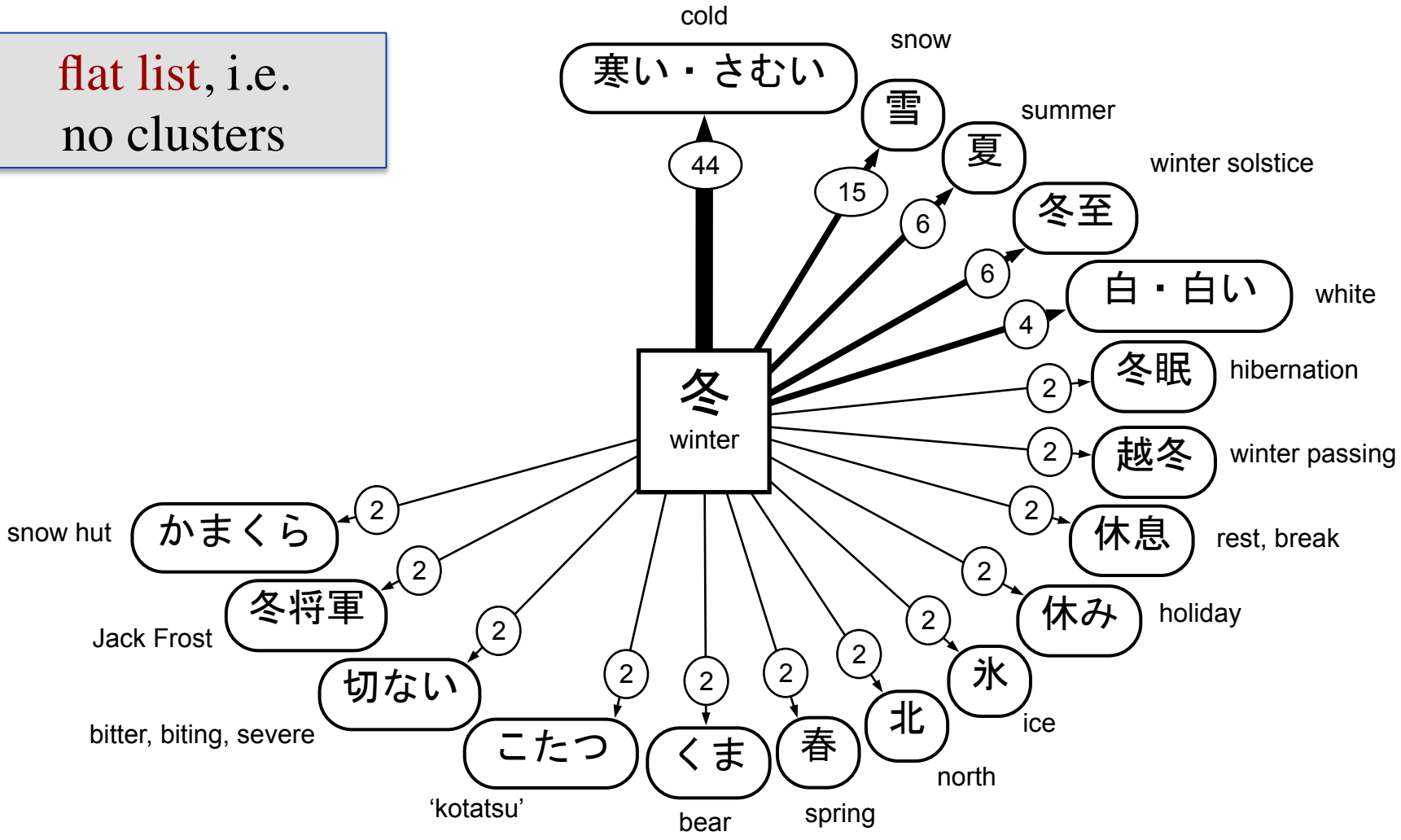
In both cases we will use words. Let's forget here about the input, but what about the output?

How to present the output to the user?

- unorganized (or alphabetically organized) lists of words
- word clusters
- categorial trees: clusters labeled by category (food, animal, plant, ...)

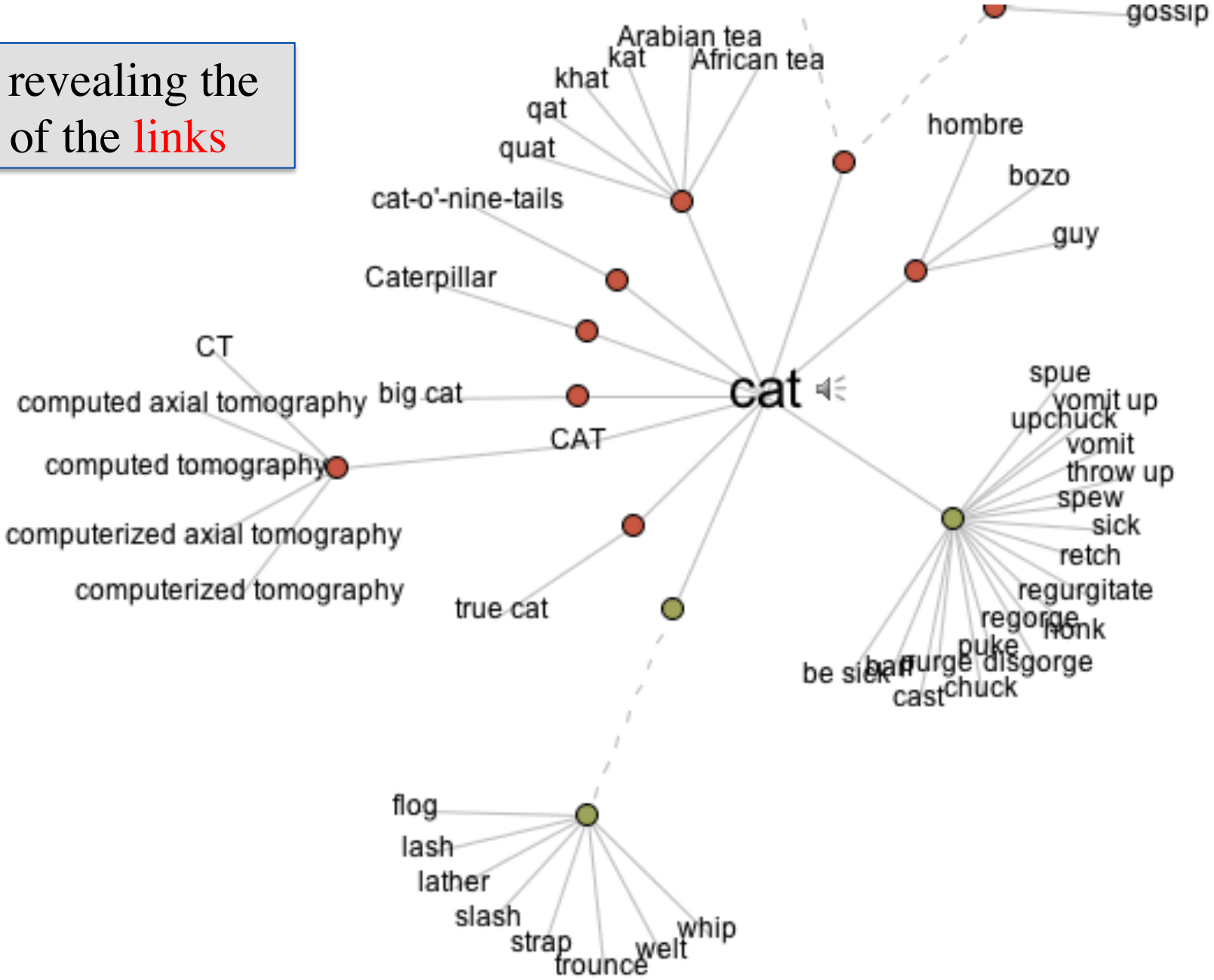
Association network for the word 'winter'

flat list, i.e.
no clusters



Visual Thesaurus: Words as clusters

without revealing the name of the links



The way how E.A.T. (Edinburgh Association Thesaurus) presents the output to the input 'India'

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

flat list, i.e. no clusters

PAKISTAN	12 0.14	FLIES	1 0.
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…

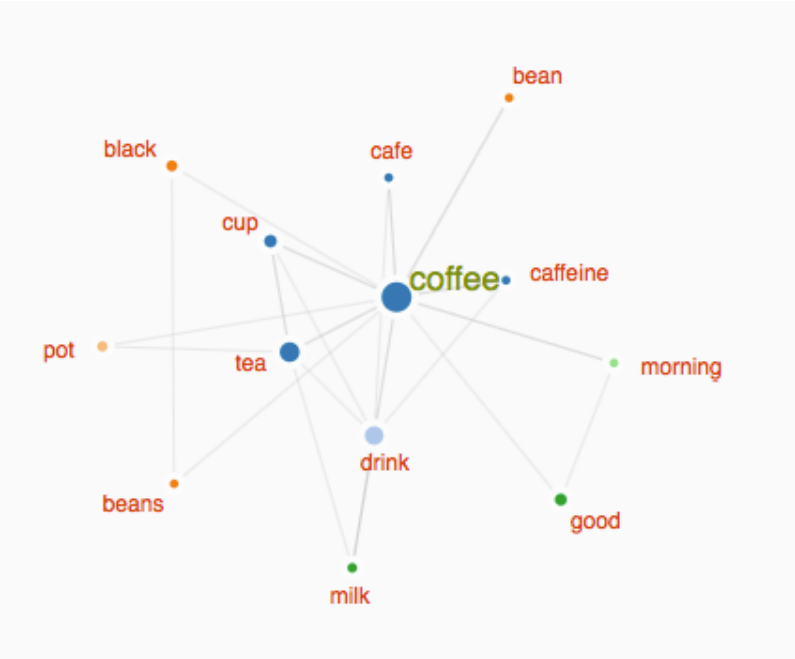
PAKISTAN	12 0.14	FLIES	1 0.01
RUBBER	10 0.12	HIMALAYAS	1 0.01
CHINA	4 0.05	HINDU	1 0.01
FOREIGN	4 0.05	HUNGER	1 0.01
CURRY	3 0.04	IMMIGRANTS	1 0.01
FAMINE	3 0.04	INDIANS	1 0.01
TEA	3 0.04	JAPAN	1 0.01
COUNTRY	2 0.02	KHAKI	1 0.01
GHANDI	2 0.02	MAN	1 0.01
WOGS	2 0.02	MISSIONARY	1 0.01
AFGHANISTAN	1 0.01	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

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

Roadmap

Step-1

- given some input, *source word*, build a graph with all direct neighbors ==> lexical graph or association network);



TEA 39 0.39	BISCUITS 1 0.01
CUP 7 0.07	BITTER 1 0.01
BLACK 5 0.05	DARK 1 0.01
BREAK 4 0.04	DESERT 1 0.01
ESPRESSO 40.0.4	DRINK 1 0.01
POT 3 0.03	FRENCH 1 0.01
CREAM 2 0.02	GROUND 1 0.01
HOUSE 2 0.02	INSTANT 1 0.01
MILK 2 0.02	MACHINE 1 0.01
CAPPUCINO 20.02	MOCHA 1 0.01
STRONG 2 0.02	MORNING 1 0.01
SUGAR 2 0.02	MUD 1 0.01
TIME 2 0.02	NEGRO 1 0.01
BAR 1 0.01	SMELL 1 0.01
BEAN 1 0.01	TABLE 1 0.01
BEVERAGE 1 0.01	

http://www.smallworldofwords.com/new/visualize/#

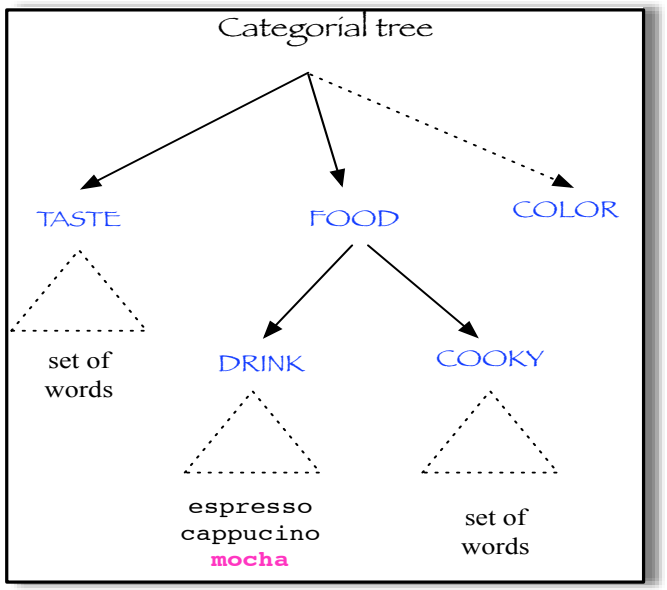
http://www.eat.rl.ac.uk/cgi-bin/eat-server

Roadmap

Step-2

- cluster and label the words produced in response to some input (build a categorial tree). Suppose the input to be 'coffee', with the target word being 'mocha'.

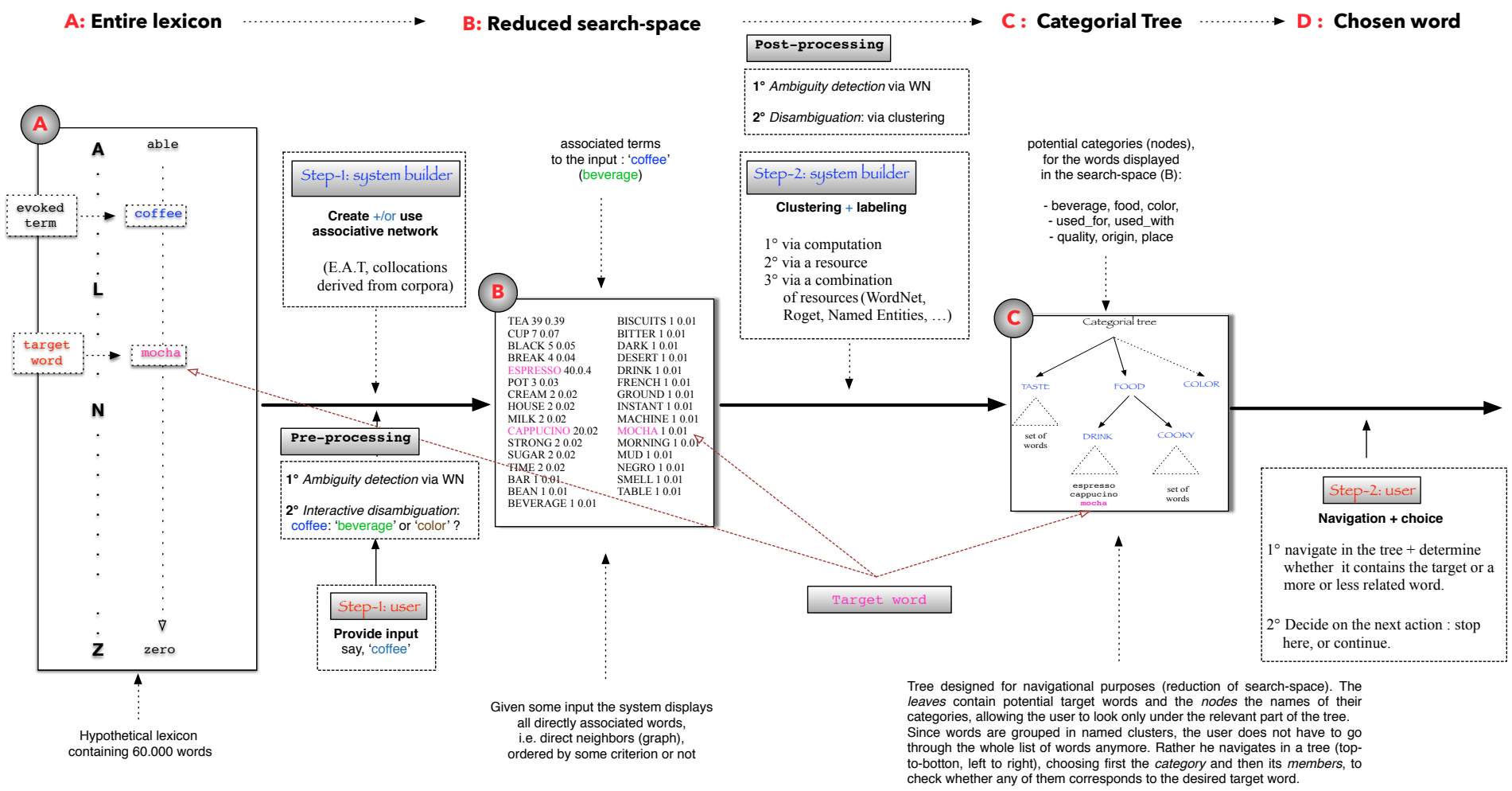
TEA 39 0.39	BISCUITS 1 0.01
CUP 7 0.07	BITTER 1 0.01
BLACK 5 0.05	DARK 1 0.01
BREAK 4 0.04	DESERT 1 0.01
ESPRESSO 40.0.4	DRINK 1 0.01
POT 3 0.03	FRENCH 1 0.01
CREAM 2 0.02	GROUND 1 0.01
HOUSE 2 0.02	INSTANT 1 0.01
MILK 2 0.02	MACHINE 1 0.01
CAPPUCINO 20.02	MOCHA 1 0.01
STRONG 2 0.02	MORNING 1 0.01
SUGAR 2 0.02	MUD 1 0.01
TIME 2 0.02	NEGRO 1 0.01
BAR 1 0.01	SMELL 1 0.01
BEAN 1 0.01	TABLE 1 0.01
BEVERAGE 1 0.01	



Direct neighbors of "coffee'
http://www.eat.rl.ac.uk/cgi-bin/eat-server

Categorial tree corresponding to the output (direct neighbors)
Produced in response to the input (source-word : ocoffee)

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



Some Resources we may want to consider:

- WordNet
- ConceptNet
- BabelNet
- Roget's Thesaurus
- Yago
- UBY

Conclusion + future work

What have we achieved with respect to word-access?

- define a framework within which lies the solution

What still needs to be done?

- find the best resources (Roget, E.A.T., WordNet)
- combine them properly
- find a good algorithm for clustering (check whether Roget or a similar resource is well suited for this task)
- find a way to determine the adequate labels ('*hypernym*' vs. '*more general term*')

Dan

will tell you now /one day
how to get
all this to work !

Interconnecting lexicographic resources.
A first attempt to check Michael's model

Dan Cristea and Andrei-Liviu Scutelnicu
"Alexandru Ioan Cuza" University of Iași
Institute of Computer Science of the Romanian Academy
dcristea@info.uaic.ro
liviu.scutelnicu@info.uaic.ro

Investigation

- **What to do when one resource does not give us the expected result**?
  - mix resources of different types
- **Could a mix of resources be more successful**?
  - Let's see…

Investigation

What kind of resources could we mix?

- *for the time being*: WN and an explanatory dictionary
- *in the future*: a corpus (access to contexts), Wikipedia/ DBpedia, a collection of dictionaries, etc.

At what price?

- ad-hoc programming right now => expensive
- big expectation for a nice generalisation => cheep

How is WN connected with the Dictionary?

A **tight** alignment: WN lexicals in synsets aligned with the senses of entries of the explanatory dictionary (ExDi)

- a WN synset:

  pos (def, ex, w_1^{s1} … w_k^{sk} … w_n^{sn})

- an explanatory dictionary entry:

  w_k, pos, <w_k^{s1},def_1,ex_1>… <w_k^{sk},def_k,ex_k>… <w_k^{sm},def_m,ex_m>

How is the WN connected with the Dictionary?

- In the actual implementation, a **light** alignment: WN *lexicals* in synsets aligned with *title words* in ExDi
  - a WN synset:

    pos (def, ex, w_1 …w_k … w_n)

  - an explanatory dictionary entry:

    w_k, pos, <w_k^{s1},def_1,ex_1>… <w_k^{sk},def_k,ex_k>… <w_k^{sm},def_m,ex_m>

Algorithm

Given a source word *w*

=> extract from ExDi definitions of *w* : *defs*

=> search *w* in **WN** and get its domain: *d*

=> filter the words belonging to *defs* and *d*: *f*

=> merge all literals belonging to the synsets of *f* in **WN**: *m*

=> cluster *m*: replace the list *m* with their hypernyms

=> let the user choose among these clusters: *c*

=> display the cluster *c*

=> iterate if the target word is not in the list *c*

An example…

- **Input** (Source word): *abate* (EN: *superior* – religion)
- **DEX-online**: collection of definitions in various dictionaries yielding 7 entries for *abate*
- **Output** (end of Step 1): 601 words, the majority of them belonging to the domain of *religion*, the target word *vicar* being among them.

Further work

- Still to be done
  - *clusterisation*: take the hypernyms of all these words => no. hypernyms – smaller than the initial list.
  - Move up the hierarchy until a hypernym covers the set of considered words. For example, the set 'child, man, Obama' should yield the category "human".
  - the intersection of all the categorys' hyponyms with the expanded initial list (see previous slide) should drastically reduce this list and ideally contain the target word.
- **Question**: up to what level shall we go to choose the right hypernym? Is this merely a quantitative issue?