From Semantics to… Semantics

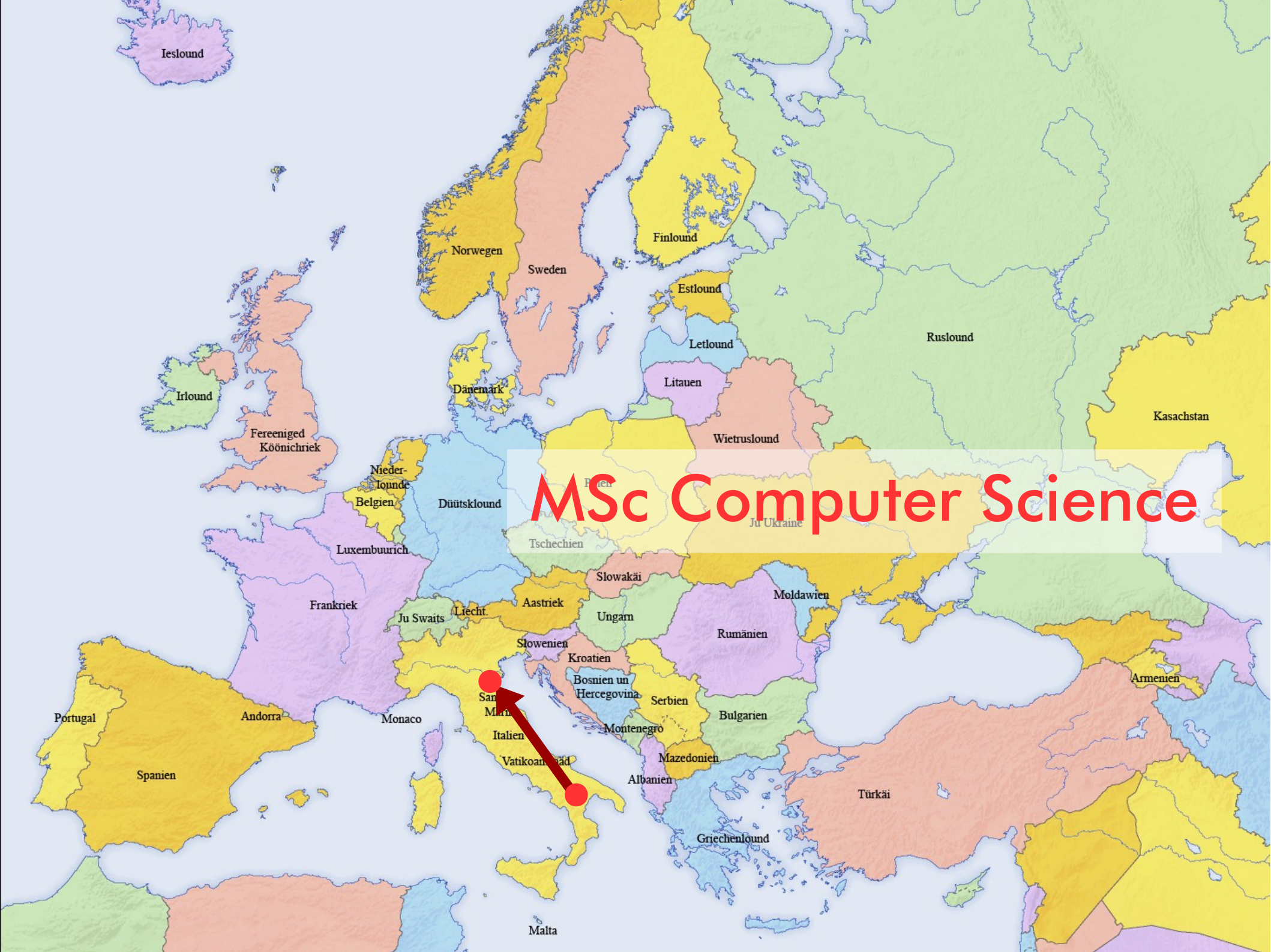
Adventures in Meaning Banking and other stories

Valerio Basile

2/10/2015


Ieslound
Norwegen
Sweden
Finlound
Estlound
Ruslound
Letlound
Litauen
Irlound
Dänemark
Kasachstan
Fereeniged Köönichriek
Nieder-lounde
Belgien
Düütsklound
Polen
Luxembuurich
Tschechien
Slowakäi
Frankriek
Ju Swaits
Liecht.
Aastriek
Ungarn
Slowenien
Kroatien
Bosnien un Hercegovina
San Marino
Portugal
Andorra
Monaco
Italien
Vatikoanstääd
Mazedonien
Albanien
Spanien
Türkäi
Griechenlound
Malta


Ieslound
Norwegen
Sweden
Finlound
Estlound
Irlound
Fereeniged
Köönichriek
Dänemark
Nieder-
lounde
Belgien
Düütsklound
Polen
Luxembuurich
Tschechien
Frankriek
Ju Swaits
Liecht.
Aastriek
Ungarn
Slowenien
Kroatien
Bosnien un
Hercegovin
San
Marino
Monaco
Portugal
Andorra
Italien
Vatikoanstääd
Spanien
Albanien
Türkäi
Griechenlound
Malta



MSc Computer Science


PhD (defense pending)
Ieslound
Norwegen
Sweden
Finlound
Estlound
Letlound
Litauen
Ruslound
Kasachstan
Irlound
Fereeniged Köönichriek
Dänemark
Wietruslound
Nieder-lound
Belgien
Düütsklound
Pole
Ju Ukraine
Tschechien
Luxembuurich
Frankriek
Ju Swaits
Slowakäi
Aastriek
Ungarn
Moldawien
Rumänien
Slowenien
Kroatien
Bosnien un Hercegovina
Serbien
Bulgarien
Armenien
Portugal
Andorra
Monaco
Italien
Montenegro
Mazedonien
Albanien
Spanien
Türkäi
Griechenlound
Malta


Post-doc at Wimmics
Ieslound
Norwegen
Sweden
Finlound
Estlound
Ruslound
Letlound
Litauen
Dänemark
Irlound
Fereeniged Köönichriek
Wietruslound
Kasachstan
Belgien
Düütsklound
Tschechien
Slowakäi
Luxembuurich
Frankriek
Aastriek
Ungarn
Moldawien
Rumänien
Slowenien
Kroatien
Bosnien un Hercegovina
Serbien
Bulgarien
Armenien
Portugal
Andorra
Monaco
Italien
Montenegro
Mazedonien
Albanien
Spanien
Türkäi
Griechenlound
Malta

Groningen, January 1st 2011.

A quest to build a semantically annotated free corpus.

A Meaning Bank.

The Meaning Bankers

What's in a Corpus

lots of texts

+

layers of annotation

What's in a Corpus

raw text

- English
- Multiple sources
- Public domain
- Full documents
- Not too long

What's in a Corpus

multi-layer annotation

- Word boundaries
- Part of speech
- Lemmas
- Named entities
- Word senses
- Animacy
- Syntax
- Semantics

Sources of text

- Voices of America
- CIA World Factbook
- MASC
- Tatoeba
- Fables
- Jokes

Splitting words

Tokenization is a solved problem(?)

The dog bit John. He bleeds.

⬇

```
The dog bit John .
He bleeds .
```

Splitting words

```
The dog bit John. He bleeds.
SIIOTIIOTIIOTIIITOSIOTIIIIIT
```

elephant

Evang et al. 2013

Elephant: Sequence Labeling for Word and Sentence Segmentation

Tagging words

Token	The	dog	bit	John

Tagging words

Token	The	dog	bit	John
POS	det.	noun	t. verb	p. noun

Tagging words

Token	The	dog	bit	John
POS	det.	noun	t. verb	p. noun
lemma	the	dog	bite	John

Tagging words

Token	The	dog	bit	John
POS	det.	noun	t. verb	p. noun
lemma	the	dog	bite	John
namex				person

Tagging words

Token	The	dog	bit	John
POS	det.	noun	t. verb	p. noun
lemma	the	dog	bite	John
namex				person
sense		dog.n.1	bite.v.1	

Tagging words

Token	The	dog	bit	John
POS	det.	noun	t. verb	p. noun
lemma	the	dog	bite	John
namex				person
sense		dog.n.1	bite.v.1	
animacy		animate		human

Combinatory Categorial Grammar

- Lexical items have types.
  Primitive types: S, N, NP, ...
  Complex types: X/Y, X\Y

- Parsing: reducing to S by means of combinators

Combinatory Categorial Grammar

Combinators

Application

$$\frac{\alpha{:}X/Y \qquad \beta{:}Y}{\alpha\beta{:}X}\,{>} \qquad\qquad \frac{\beta{:}Y \qquad \alpha{:}X\backslash Y}{\beta\alpha{:}X}\,{<}$$

Composition

$$\frac{\alpha{:}X/Y \quad \beta{:}Y/Z}{\alpha\beta{:}X/Z}\,B_{>} \qquad\qquad \frac{\beta{:}Y\backslash Z \quad \alpha{:}X\backslash Y}{\beta\alpha{:}X\backslash Z}\,B_{<}$$

Type-raising

$$\frac{\alpha{:}X}{\alpha{:}T/(T\backslash X)}\,T_{>} \qquad\qquad \frac{\alpha{:}X}{\alpha{:}T\backslash(T/X)}\,T_{<}$$

CCG - Example

the dog bit John

CCG - Example

the	dog	bit	John
NP/N	N	(S\NP)/NP	NP

CCG - Example

the	dog	bit	John
NP/N	N	(S\NP)/NP	NP

$$\frac{\text{NP/N} \qquad \text{N}}{\text{NP}}\,>$$

CCG - Example

the	dog	bit	John
NP/N	N	(S\NP)/NP	NP
NP >		S\NP >	

CCG - Example

the	dog	bit	John
NP/N	N	(S\NP)/NP	NP

NP/N N >

(S\NP)/NP NP >

NP S\NP <

S

C&C Tools

By James Curran and Stephen Clark

C&C Tools

By James Curran and Stephen Clark

The dog bit John

Part-of-speech tagger

The|DT dog|NN bit|VBT John|NNP

C&C Tools

By James Curran and Stephen Clark

The dog bit John

Part-of-speech tagger

The|DT dog|NN bit|VBT John|NNP

Named entity tagger

The|DT|O dog|NN|O bit|VBT|O John|NNP|PER

C&C Tools

By James Curran and Stephen Clark

The dog bit John

Part-of-speech tagger

The|DT dog|NN bit|VBT John|NNP

Named entity tagger

The|DT|O dog|NN|O bit|VBT|O John|NNP|PER

Supertagger

The|DT|O|NP/N dog|NN|O|N bit|VBT|O|(S\NP)/NP John|NNP|PER|NP

C&C Tools

By James Curran and Stephen Clark

The dog bit John

Part-of-speech tagger

The|DT dog|NN bit|VBT John|NNP

Named entity tagger

The|DT|O dog|NN|O bit|VBT|O John|NNP|PER

Supertagger

The|DT|O|NP/N dog|NN|O|N bit|VBT|O|(S\NP)/NP John|NNP|PER|NP

CCG Parser

Discourse Representation Theory

- Logic-based
- Scope of discourse referents
- (neo)Davidsonian

x_1 e_1 x_2
$dog(x_1)$
$bite(e_1)$
$john(x_2)$
$agent(e_1,x_1)$
$theme(e_1,x_2)$

Discourse Representation Theory

- Logic-based
- Scope of discourse referents
- (neo)Davidsonian
- Recursive
- Extensible

$x_1\ e_1$

$dog(x_1)$

$wants(e_1)$

$agent(e_1,x_1)$

$theme(e_1,p_1)$

p_1:

$x_2\ e_2$

$bite(e_2)$

$john(x_2)$

$agent(e_2,x_1)$

$theme(e_2,x_2)$

Boxer

By Johan Bos

- Rule-based
- CCG as input
- Prolog or XML output
- Part of C&C tools

```
vbasile@yamanaka:~$ curl -d"The dog bites John" 'http://gingerbeard.alwaysdata.n
et/candcapi/proxy.php/raw/pipeline?box=true&integrate=true&resolve=true&semantic
s=drs&theory=drt&instantiate=true'
%%% This output was generated by the following command:
%%% /home/gingerbeard/candc/bin/boxer --stdin --box true --resolve true --theory
 drt --instantiate true --integrate true --semantics drs

:- multifile     sem/3, id/2.
:- discontiguous sem/3, id/2.
:- dynamic       sem/3, id/2.
%%% The dog bites John
id(1,1).
sem(1,[1001:[tok:'The',pos:'DT',lemma:the,namex:'O'],1002:[tok:dog,pos:'NN',lemm
a:dog,namex:'O'],1003:[tok:bites,pos:'VBZ',lemma:bite,namex:'O'],1004:[tok:'John
',pos:'NNP',lemma:'John',namex:'I-PER']],merge(drs([[]:x2,[1001]:x1],[[1004]:nam
ed(x2,john,per,nam),[1002]:pred(x1,dog,n,0)]),drs([[]:e1],[[]:rel(e1,x2,theme,0)
,[]:rel(e1,x1,agent,0),[1003]:pred(e1,bite,v,0)]))).
%%%   ___________________   _____________
%%%  |x2 x1              | |e1           |
%%%  |...................| |.............|
%%% (|named(x2,john,per)|+|theme(e1,x2)|)
%%%  |dog(x1)            | |agent(e1,x1)|
%%%  |___________________| |bite(e1)     |
%%%                        |_____________|
```

Open API
`http://gingerbeard.alwaysdata.net/candcapi/`

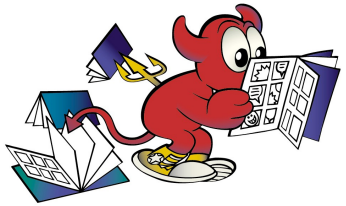
The Pipeline

GNU Make

elephant

C&C Tools

Boxer



GRUGSBUD

Groningen MEANING BANK

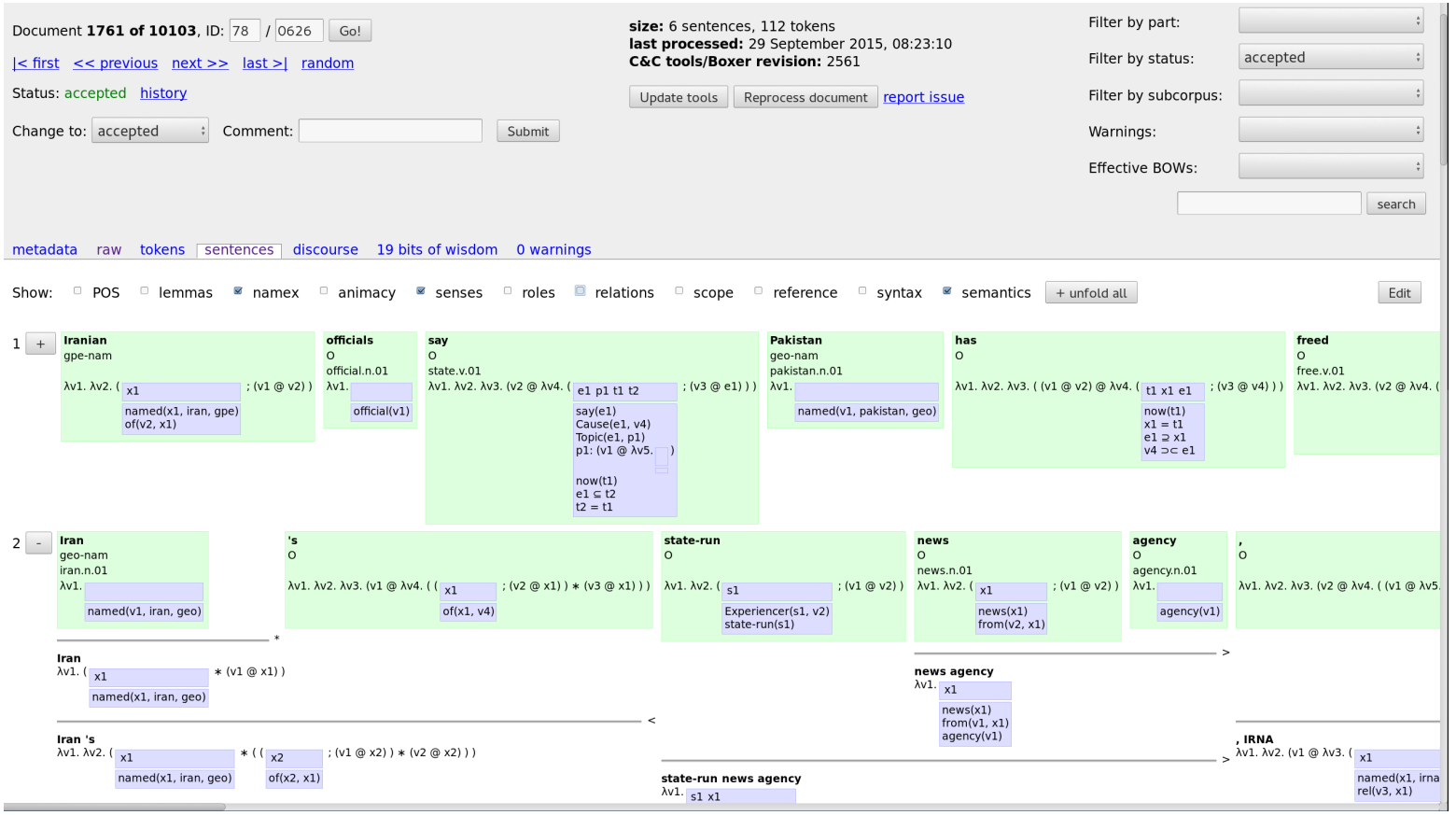
33,106	Documents
88,058	Sentences
1,629,969	Tokens
5	Releases

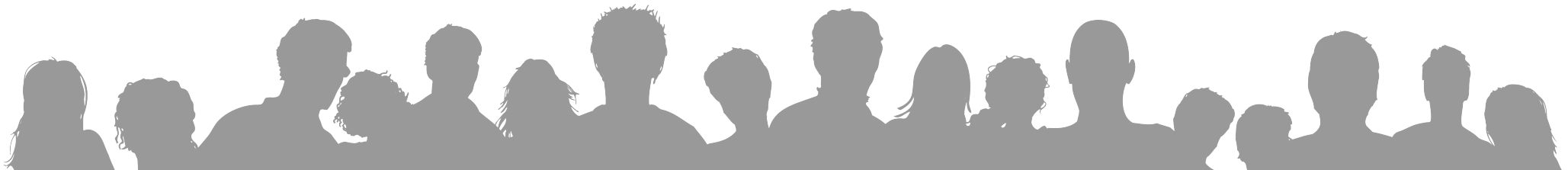
available for download

`http://gmb.let.rug.nl`

Crowdsourcing

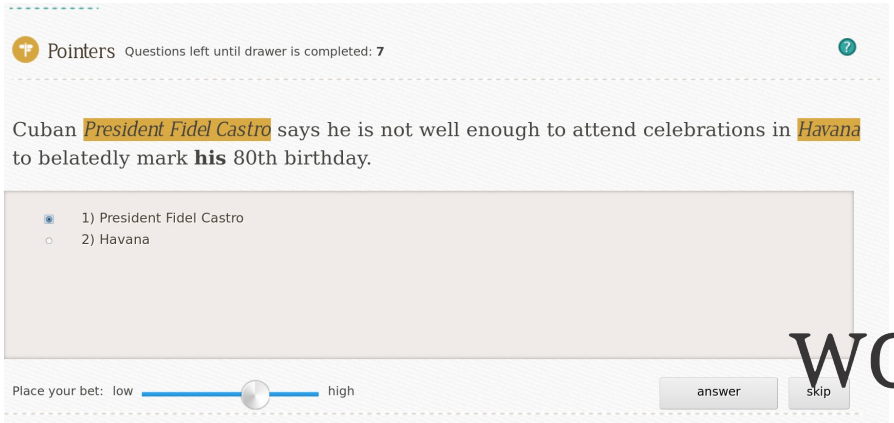
The GMB Explorer

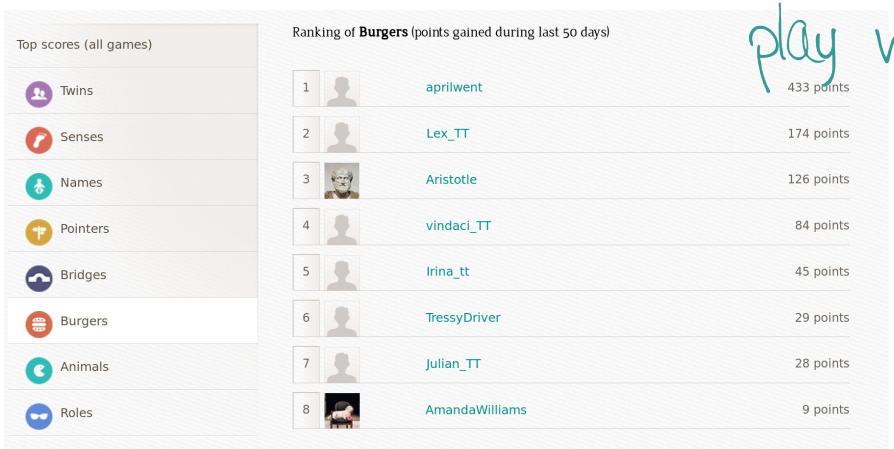




Crowdsourcing

Gamification

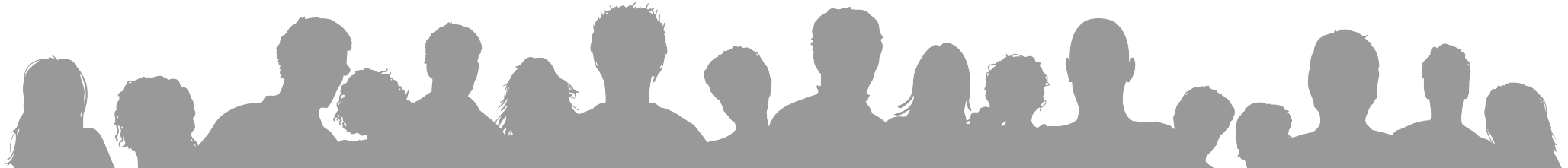




?

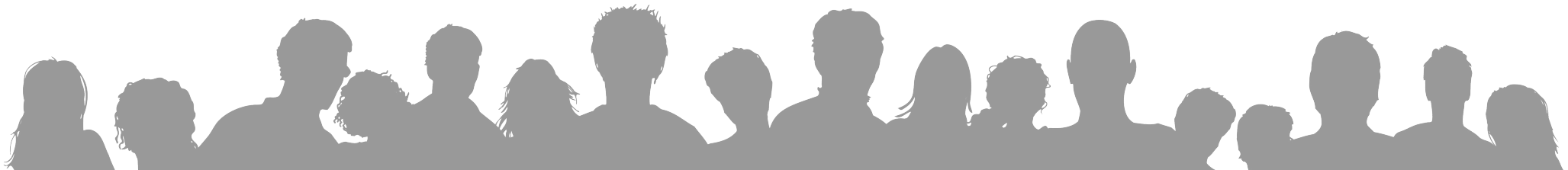
wordrobe

play what you mean



Crowdsourcing

173,173	"Bits of Wisdom"
32	Explorer users
59,413	Answers
13	Games
1,732	Players
2	Datasets



Conclusion

A new, large resource of linguistic knowledge.

A pipeline of tools for linguistic analysis.

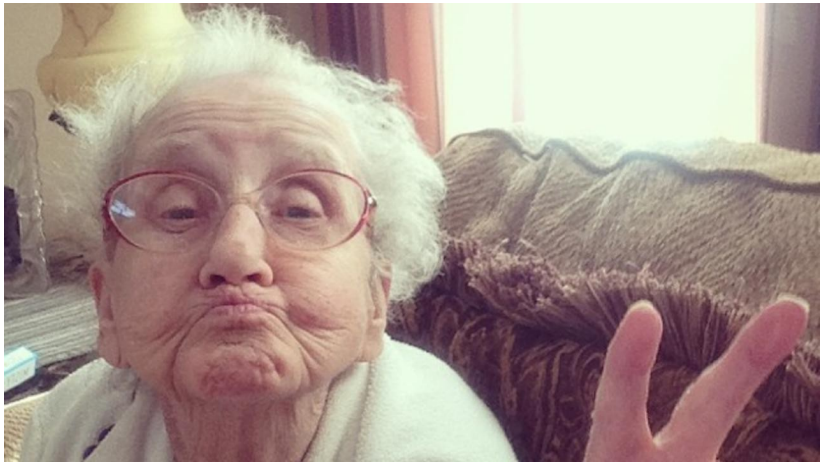
Sophia Antipolis, August 1st 2015.

ALOOF

Autonomous Learning of the Meaning of Objects

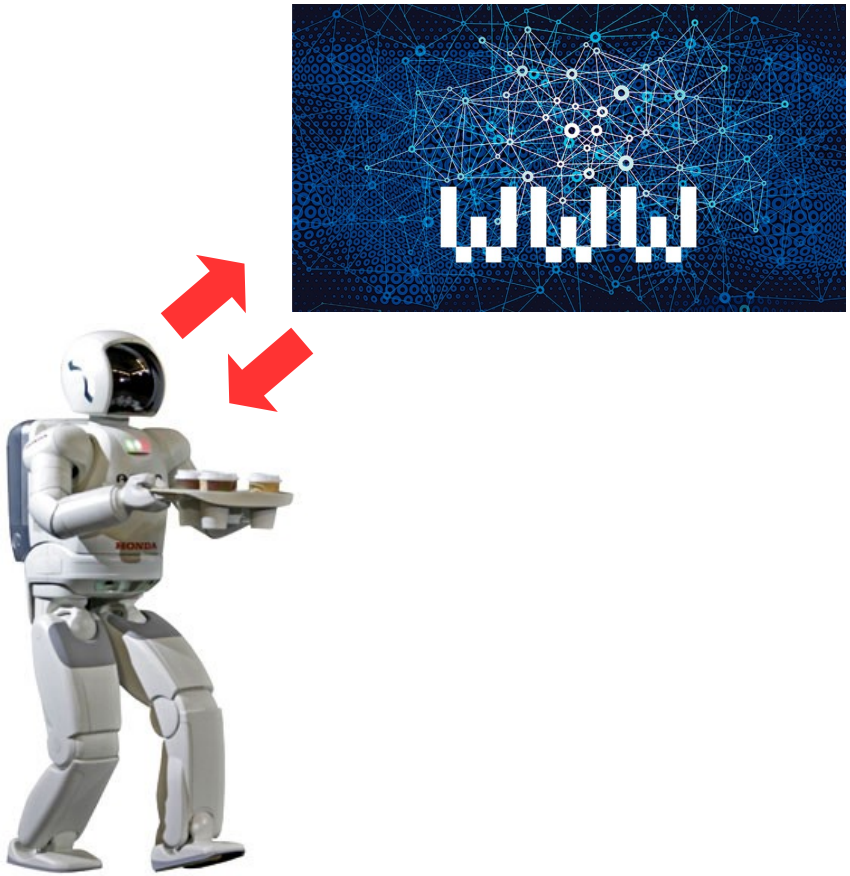
ALOOF

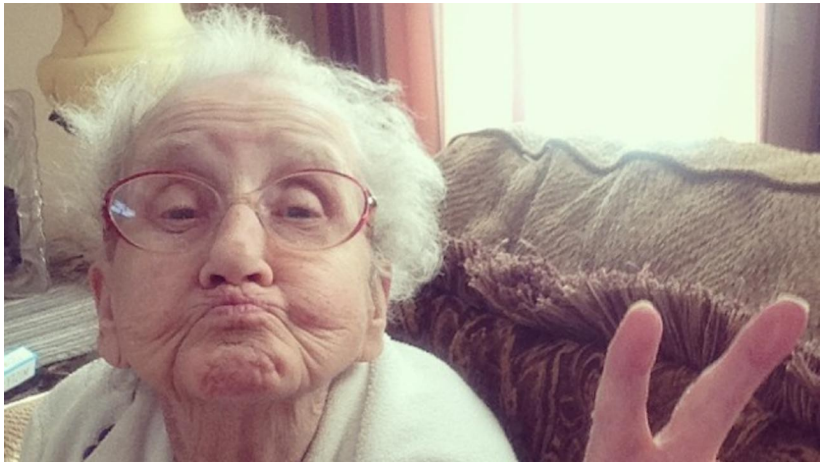
Autonomous Learning of the Meaning of Objects



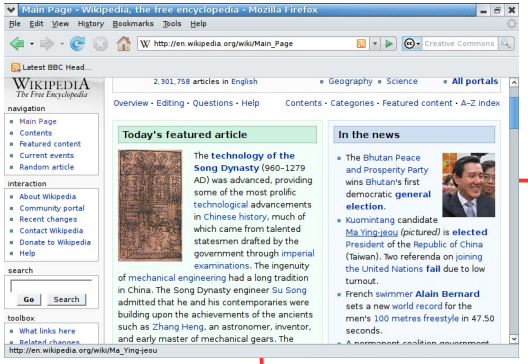
ALOOF

Autonomous Learning of the Meaning of Objects





Learning by Reading

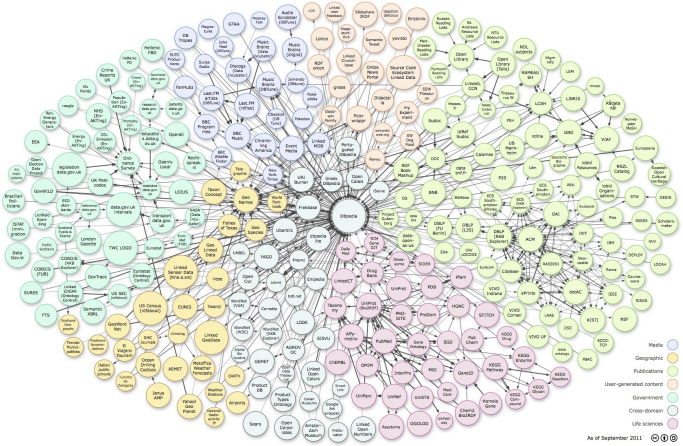


Entities

Entity relations

Entitiy categorization

New Knowledge



Learning by Reading

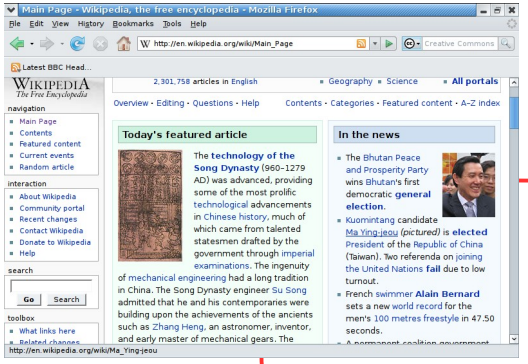
Pilot Experiment

Based on the co-mention hypothesis:

Two entities that are often mentioned together are semantically related.

Focus on tools and rooms

Learning by Reading

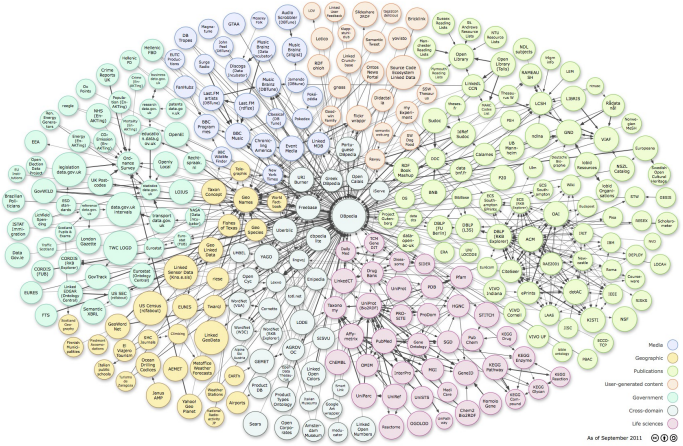


Entity co-mentions

Entities

Rooms and tools

Tool Locations



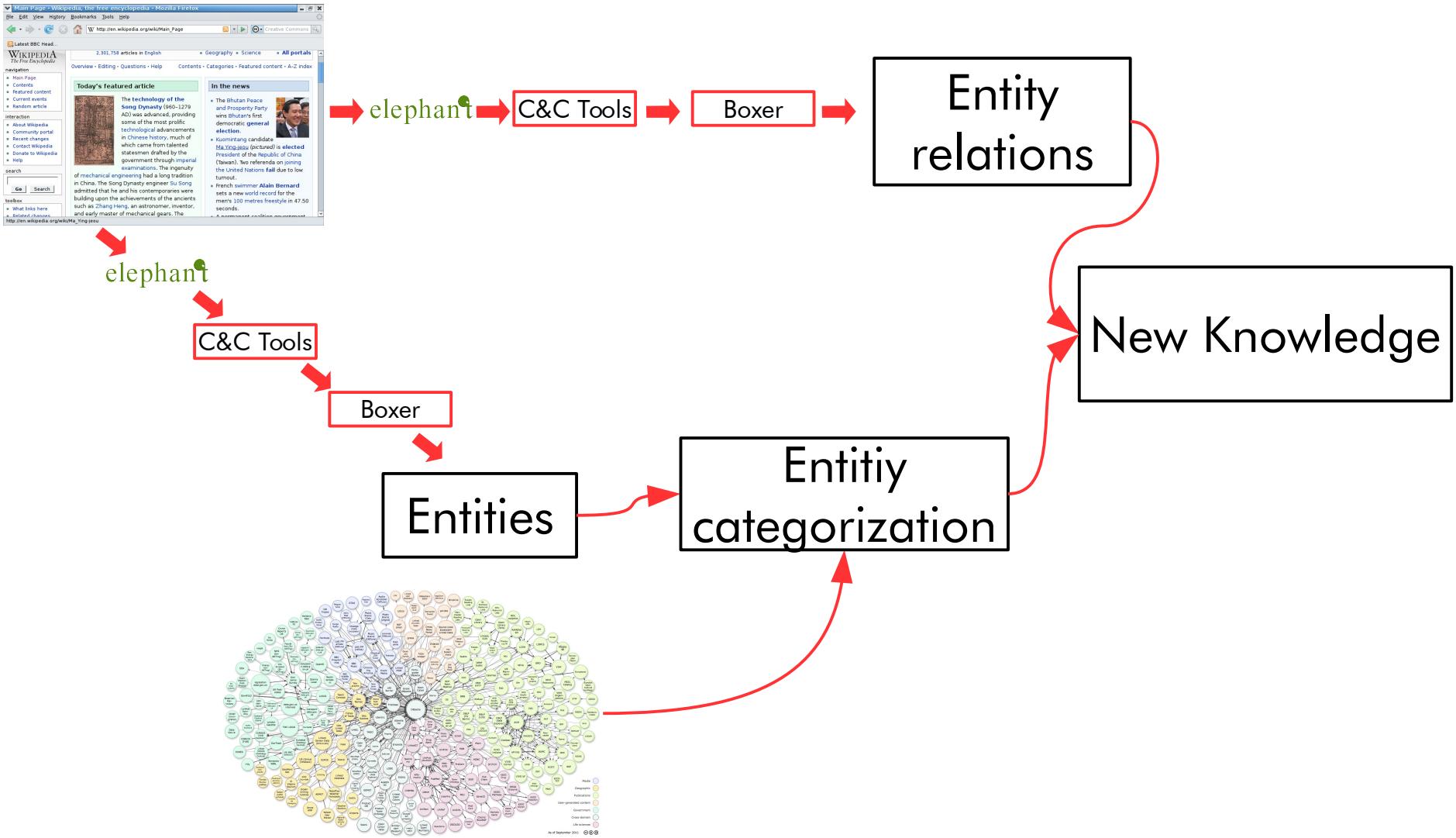
Learning by Reading

6	Task descriptions
911	911 Words
19	Rooms
82	Tools
53	Location inferences

Learning by Reading

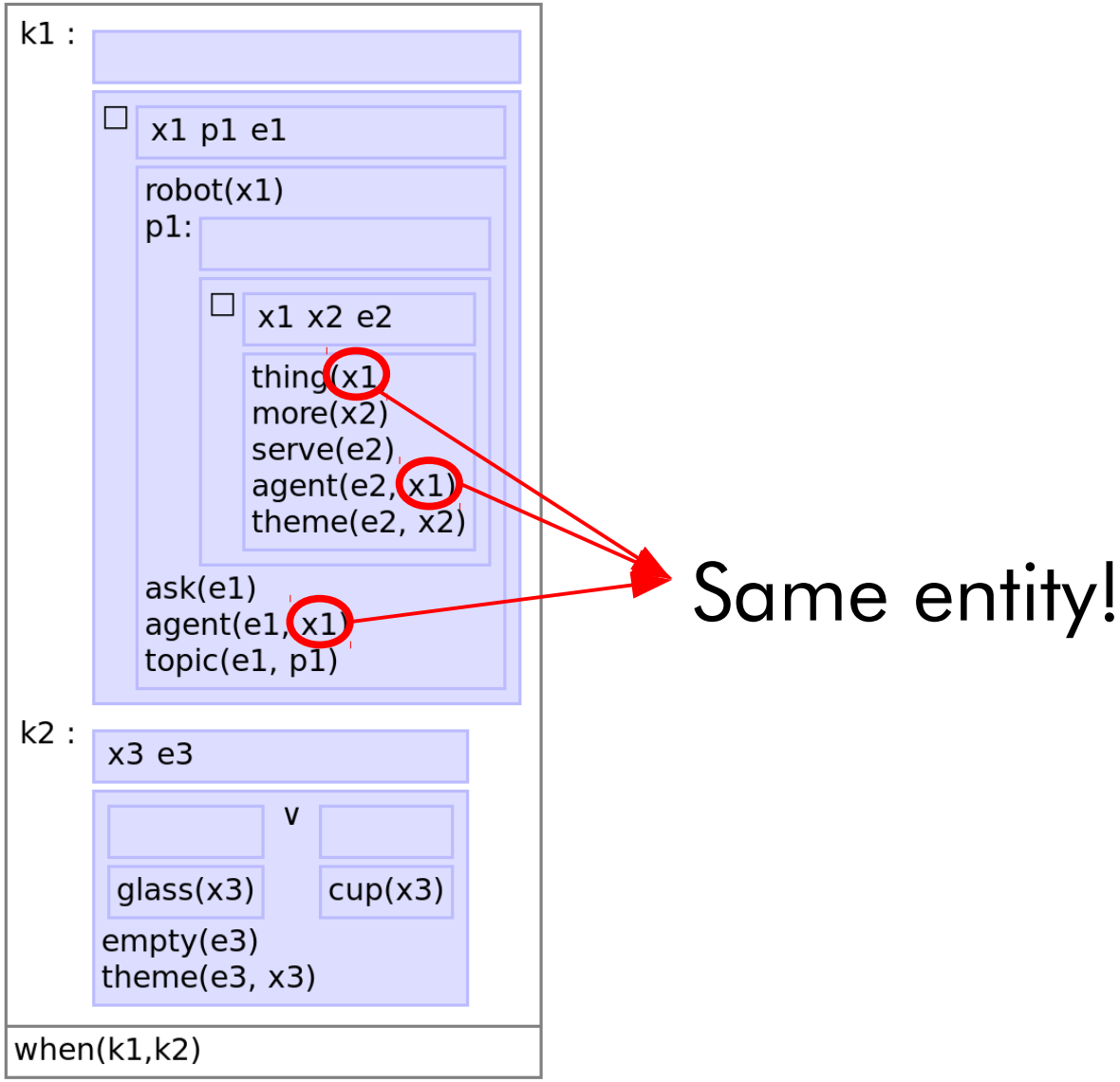
dbr:Dishwasher aloof:likelyLocation dbr:Kitchen
dbr:Oven aloof:likelyLocation dbr:Kitchen
dbr:Kitchen_cabinet aloof:likelyLocation dbr:Kitchen
dbr:Refrigerator aloof:likelyLocation dbr:Kitchen
dbr:Container aloof:likelyLocation dbr:Bathroom
dbr:Bathtub aloof:likelyLocation dbr:Bathroom
dbr:Napkin aloof:likelyLocation dbr:Parlour
dbr:Cup aloof:likelyLocation dbr:Parlour
dbr:Bathtub aloof:likelyLocation dbr:Bathroom
dbr:Bathtub aloof:likelyLocation dbr:Room
dbr:Bathtub aloof:likelyLocation dbr:Public_toilet
dbr:Fork aloof:likelyLocation dbr:Kitchen
dbr:Tin_can aloof:likelyLocation dbr:Kitchen
dbr:Cutlery aloof:likelyLocation dbr:Kitchen
...

Learning by Deep Reading



Learning by Deep Reading

"When a glass or cup is emptied, The robot will ask if it should serve more"



Conclusion

A pipeline of tools for semantic analysis.

A new, large resource of ontological knowledge.