

# From Semantics to... Semantics

Adventures in Meaning Banking and other stories

Valerio Basile

2/10/2015





Ieslound

Norwegen

Sweden

Finlound

Estlound

Irlound

Fereeniged  
Köönichriek

Nieder-  
Iounde

Belgien

Düütsklound

Polen

Tschechien

Slovak

Ungarn

Slowenien

Kroatien

Bosnien un  
Hercegovin

Portugal

Andorra

Spanien

Monaco

San  
Marino

Italien

Vatikoanstääd

Albanien

Türkäi

Griechenlound

Malta

Luxembuurch

Frankriek

Ju Swaits

Liecht.

Aastriek

Slowenien

Kroatien

Bosnien un  
Hercegovin

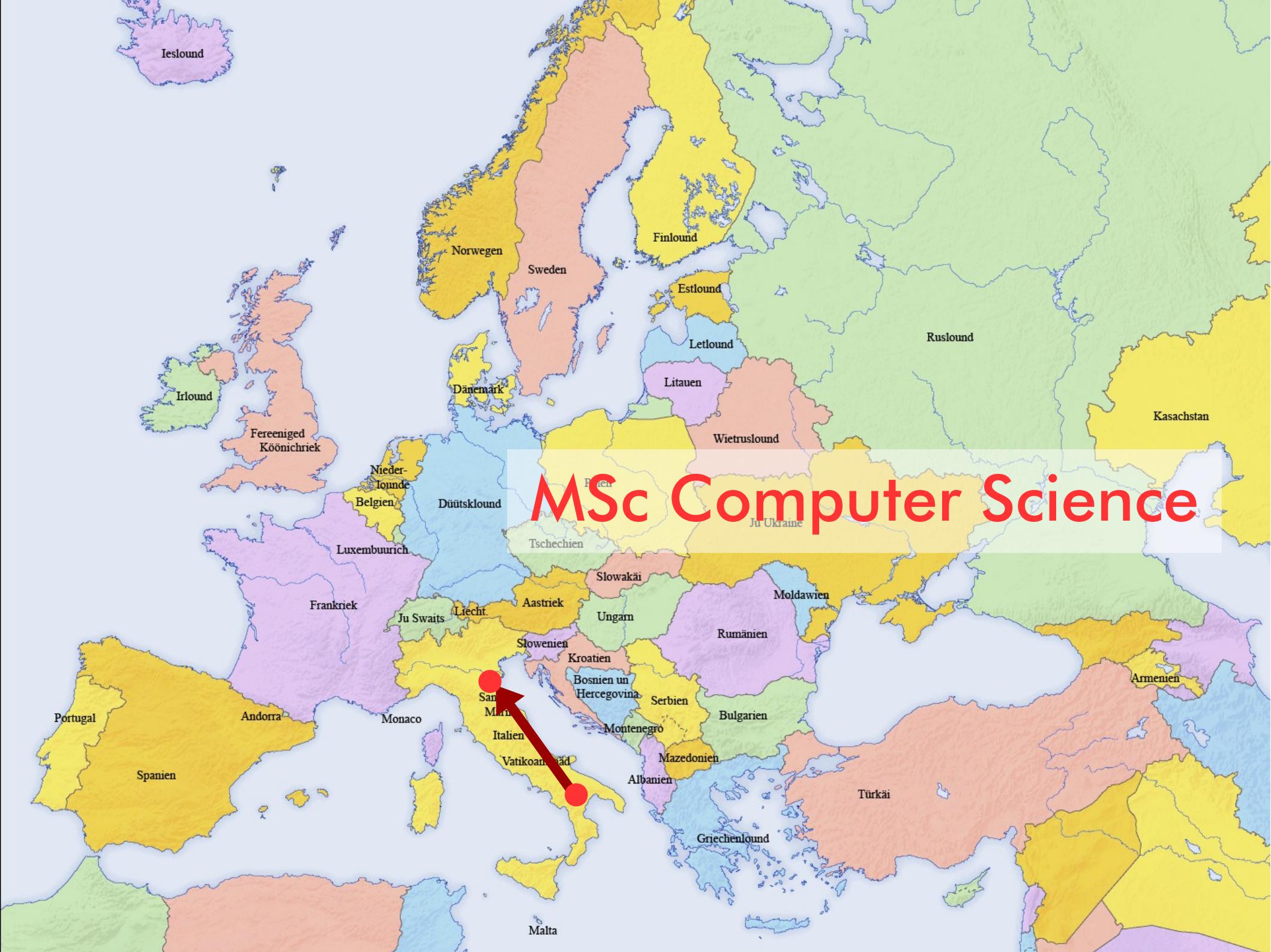
Monte

Albanien

Türkäi

Griechenlound

Malta



# MSc Computer Science



PhD (defense pending)



Groningen, January 1<sup>st</sup> 2011.



A quest to build a  
semantically annotated  
free corpus.

A Meaning Bank.



# The Meaning Bankers

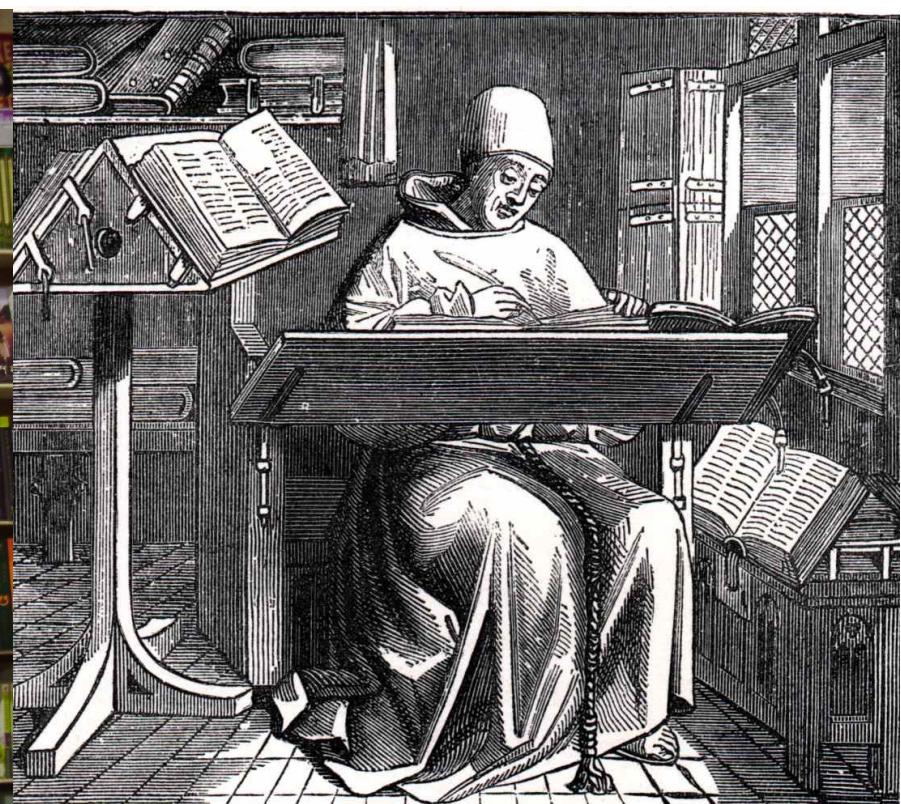


# What's in a Corpus

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.

SIIOTIIOTIIOTIIIITOSIOTIIIIIT

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

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

Composition

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

Type-raising

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

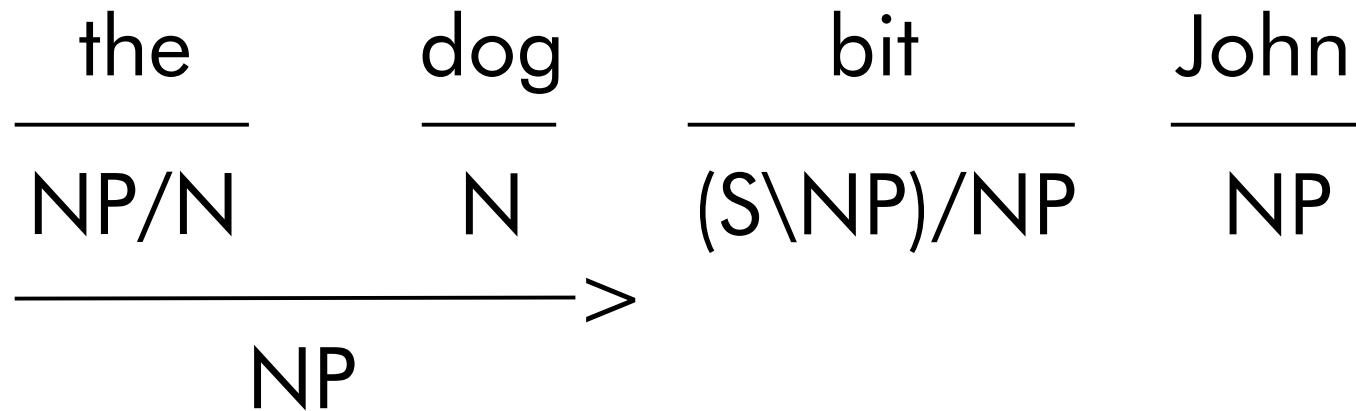
# CCG - Example

the            dog            bit            John

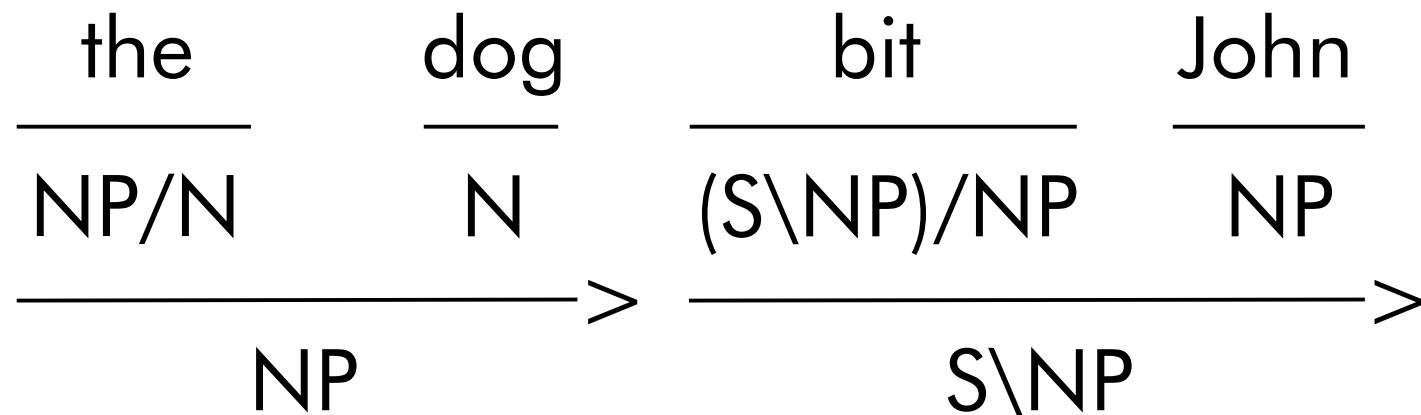
# CCG - Example

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

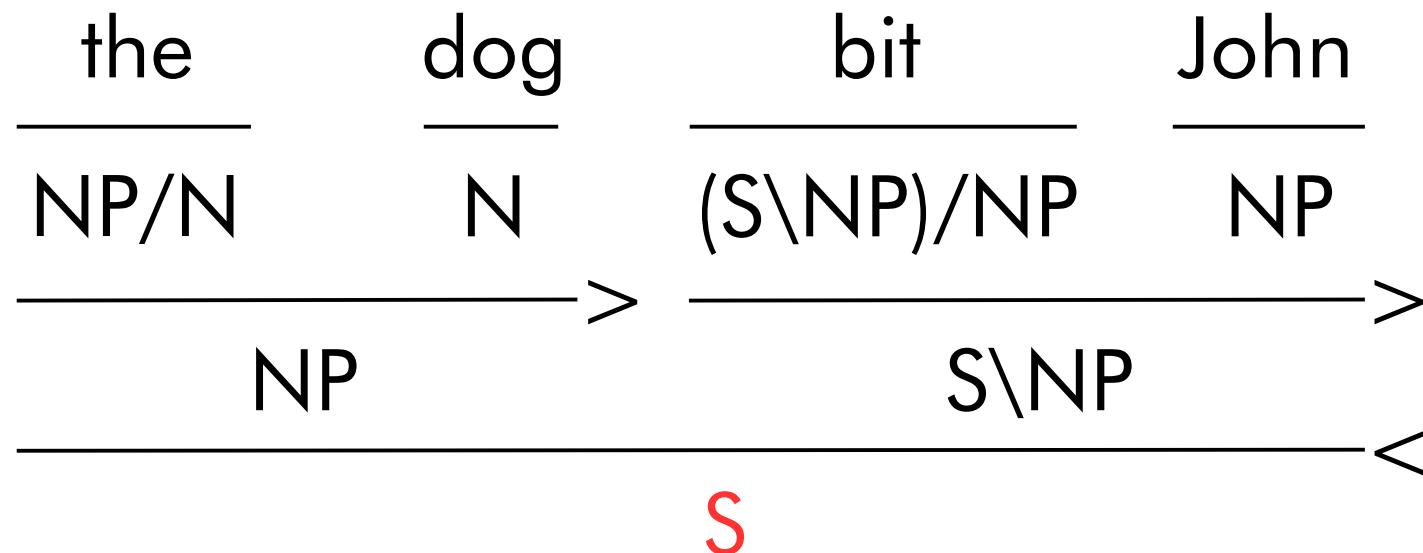
# CCG - Example



# CCG - Example



# CCG - Example



# 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**

$x_1 e_1$

$\text{dog}(x_1)$

$\text{wants}(e_1)$

$\text{agent}(e_1, x_1)$

$\text{theme}(e_1, p_1)$

$x_2 e_2$

$p_1:$

$\text{bite}(e_2)$

$\text{john}(x_2)$

$\text{agent}(e_2, x_1)$

$\text{theme}(e_2, x_2)$

# 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.net/candcapi/proxy.php/raw/pipeline?box=true&integrate=true&resolve=true&semantics=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:'0'],1002:[tok:dog,pos:'NN',lemma:'dog',namex:'0'],1003:[tok:bites,pos:'VBZ',lemma:'bite',namex:'0'],1004:[tok:'John',pos:'NNP',lemma:'John',namex:'I-PER']],merge(drs([[]:x2,[1001]:x1],[[1004]:named(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) |
%%% |-----|-----|-----|
%%% |-----|-----|-----|
%%% |-----|-----|-----|
```

Open API

<http://gingerbeard.alwaysdata.net/candcapi/>

# The Pipeline



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

Document 1761 of 10103, ID: 78 / 0626 Go!

|< first << previous next >> last >| random

Status: accepted history

Change to: accepted Comment:  Submit

size: 6 sentences, 112 tokens  
 last processed: 29 September 2015, 08:23:10  
 C&C tools/Boxer revision: 2561

Filter by part:   
 Filter by status: accepted  
 Filter by subcorpus:   
 Warnings:   
 Effective BOWs:

Update tools Reprocess document report issue

metadata raw tokens sentences discourse 19 bits of wisdom 0 warnings

Show:  POS  lemmas  namex  animacy  senses  roles  relations  scope  reference  syntax  semantics + unfold all Edit

**1** + Iranian gpe-nam  
 $\lambda v1. \lambda v2. (x1 ; (v1 @ v2))$   
 named(x1, iran, gpe)  
 of(v2, x1)

officials O official.n.01  
 $\lambda v1. official(v1)$

say O state.v.01  
 $\lambda v1. \lambda v2. \lambda v3. (v2 @ \lambda v4. (e1 p1 t1 t2 ; (v3 @ e1)))$   
 say(e1)  
 Cause(e1, v4)  
 Topic(e1, p1)  
 p1: (v1 @ \lambda v5.)  
 now(t1)  
 e1 ⊆ t2  
 t2 = t1

Pakistan geo-nam pakistan.n.01  
 $\lambda v1. named(v1, pakistan, geo)$

has O  
 $\lambda v1. \lambda v2. \lambda v3. ((v1 @ v2) @ \lambda v4. (t1 x1 e1 ; (v3 @ v4)))$   
 now(t1)  
 x1 = t1  
 e1 ⊇ x1  
 v4 ⊇ e1

freed O free.v.01  
 $\lambda v1. \lambda v2. \lambda v3. (v2 @ \lambda v4. (t1 x1 e1 ; (v3 @ v4)))$   
 now(t1)  
 x1 = t1  
 e1 ⊇ x1  
 v4 ⊇ e1

**2** - Iran geo-nam iran.n.01  
 $\lambda v1. named(v1, iran, geo)$

's O  
 $\lambda v1. \lambda v2. \lambda v3. ((x1 ; (v2 @ x1)) * (v3 @ x1))$   
 of(x1, v4)

state-run O  
 $\lambda v1. \lambda v2. (s1 ; (v1 @ v2))$   
 Experiencer(s1, v2)  
 state-run(s1)

news O news.n.01  
 $\lambda v1. \lambda v2. (x1 ; (v1 @ v2))$   
 news(x1)  
 from(v2, x1)

agency O agency.n.01  
 $\lambda v1. agency(v1)$

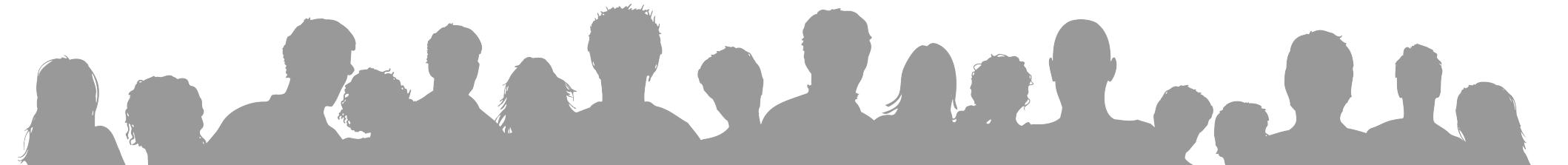
, O  
 $\lambda v1. \lambda v2. \lambda v3. ((v1 @ \lambda v4. (x1 ; (v3 @ v4)))$   
 now(t1)  
 x1 = t1  
 e1 ⊇ x1  
 v4 ⊇ e1)

Iran  
 $\lambda v1. (x1 * (v1 @ x1))$   
 named(x1, iran, geo)

Iran's  
 $\lambda v1. \lambda v2. (x1 * ((x2 ; (v1 @ x2)) * (v2 @ x2)))$   
 named(x1, iran, geo)  
 of(x2, x1)

news agency  
 $\lambda v1. x1$   
 news(x1)  
 from(v1, x1)  
 agency(v1)

, IRNA  
 $\lambda v1. \lambda v2. (v1 @ \lambda v3. (x1 ; (v3 @ v4)))$   
 named(x1, irna)  
 rel(v3, x1)



# Crowdsourcing

## Gamification

Pointers Questions left until drawer is completed: 7

Cuban *President Fidel Castro* says he is not well enough to attend celebrations in *Havana* to belatedly mark **his** 80th birthday.

1) President Fidel Castro  
2) Havana

Place your bet: low —————— high

answer skip

Top scores (all games)

Ranking of Burgers (points gained during last 50 days)
1 aprilwent 433 points
2 Lex_TT 174 points
3 Aristotle 126 points
4 vindaci_TT 84 points
5 Irina_tt 45 points
6 TressyDriver 29 points
7 Julian_TT 28 points
8 AmandaWilliams 9 points

? wardrobe

play what you mean

PLAY TWINS      PLAY SENSES      PLAY NAMES

Noun or verb? Good game for beginners. Double the trouble, twice the fun!

Identify the correct sense of a word. Quite a challenging game. Does it make sense?

Find out what type of proper name is used. A game for starters. What's in a name?

BSenthilKumar | 489 points   vindaci\_TT | 396 points   christ1ne | 372 points

PLAY POINTERS      PLAY BRIDGES      PLAY BURGERS

Select the referent of a pronoun. Not for produce beginners! Point of no return?

Find out if a referent is new, or already introduced in the text. Not a bridge too far!

Choose the topping of your noun burger. Which preposition fits best?

Maikel\_CD | 797 points   aprilwent | 189 points   aprilwent | 433 points

PLAY ANIMALS      PLAY ROLES

Dead or alive? Classify the level of animacy of phrases.

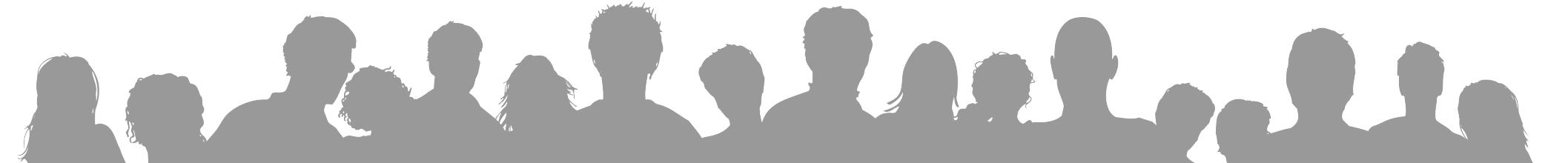
Find out what thematic roles are played in texts. A game full of drama! Are you a role model?

aprilwent | 328 points   vindaci\_TT | 368 points

# 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 1<sup>st</sup> 2015.**





Autonomous Learning of the **Meaning** of Objects

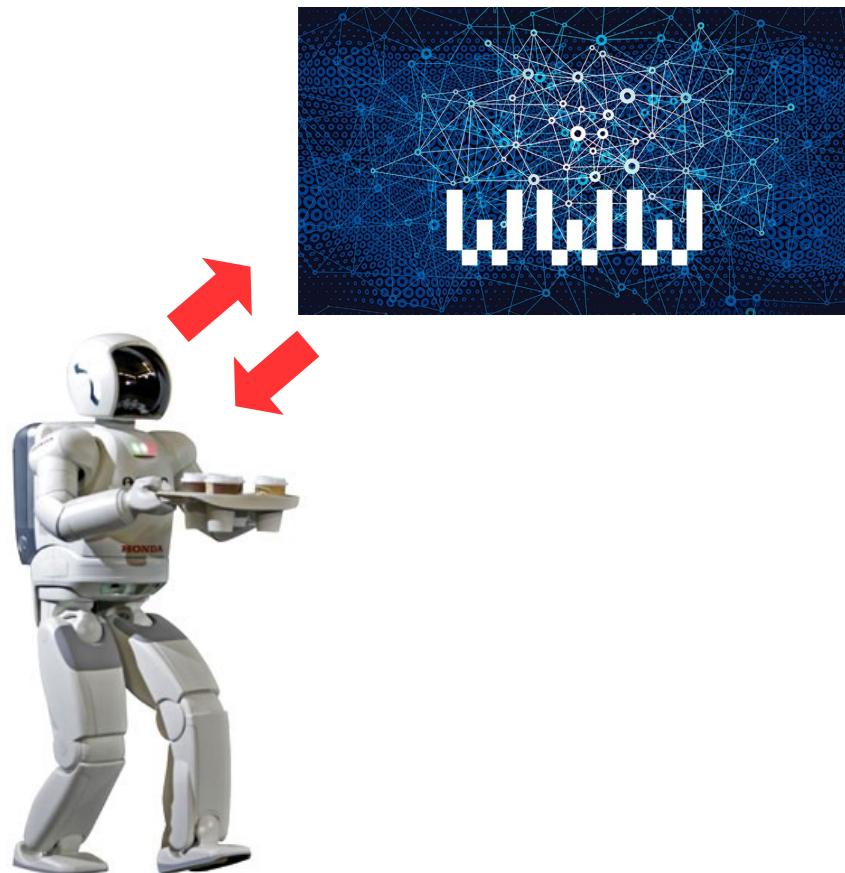
# ALOOF

Autonomous Learning of the **Meaning** of Objects

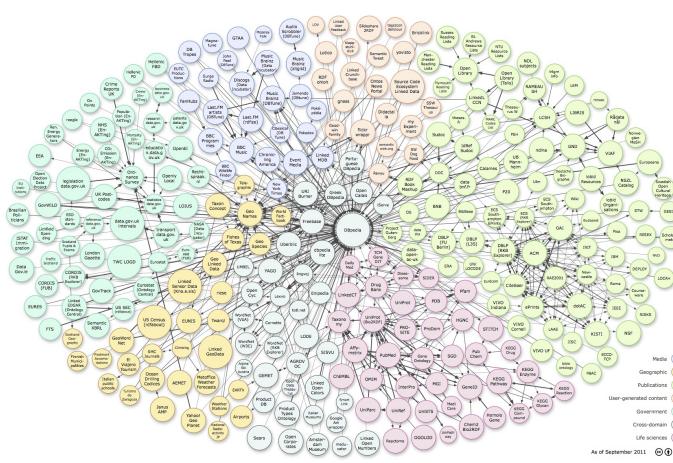
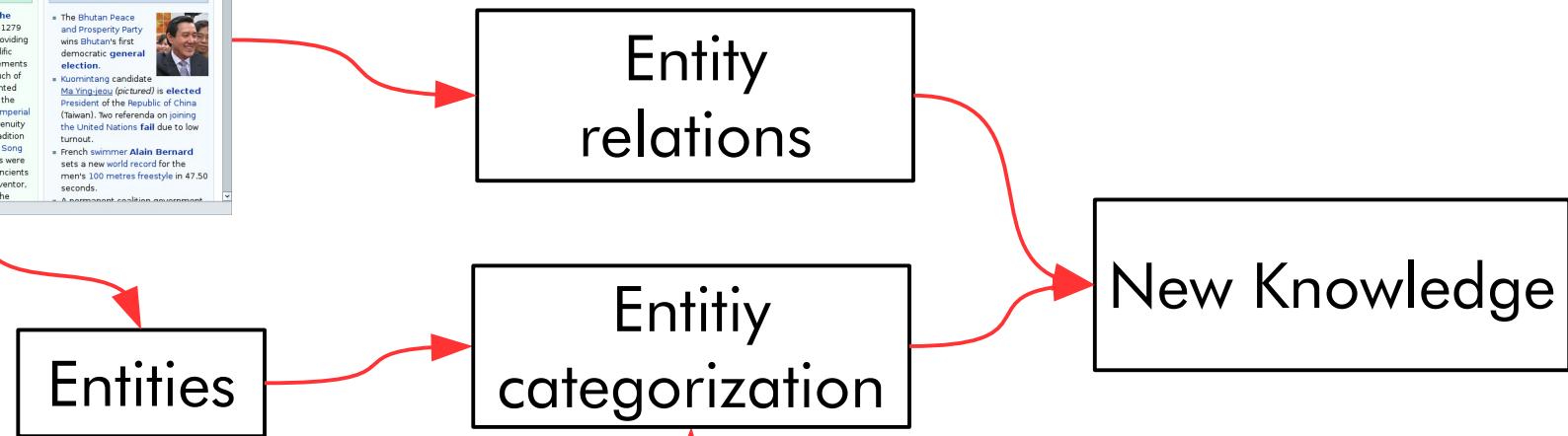
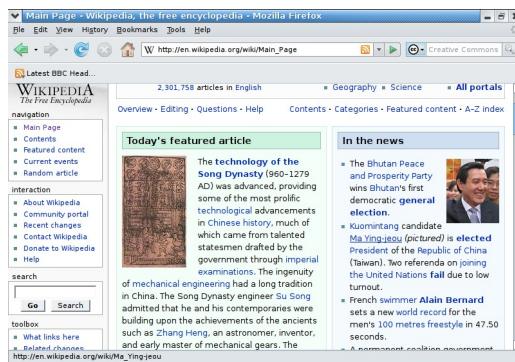


# ALOOF

Autonomous Learning of the **Meaning** of Objects



# Learning by Reading



# Learning by Reading

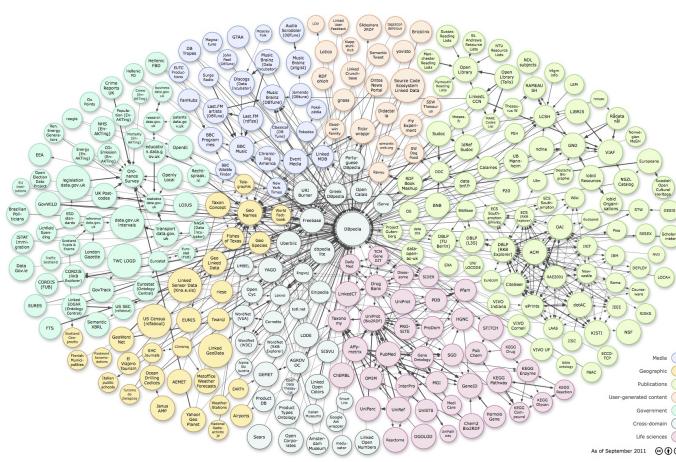
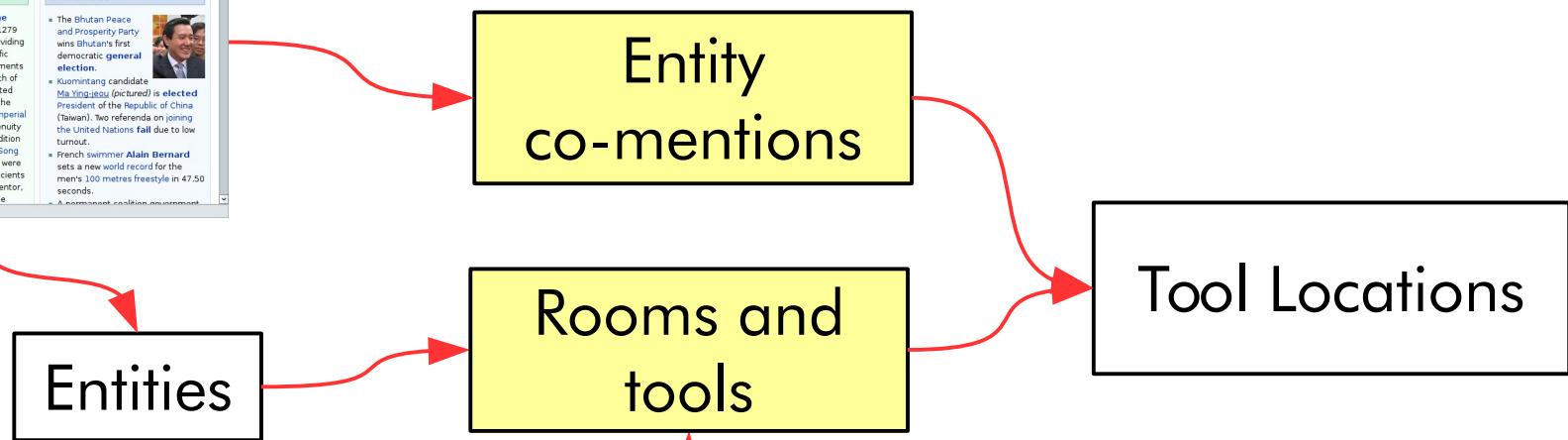
## Pilot Experiment

Based on the **co-mention** hypothesis:

*Two entities that are often mentioned together are semantically related.*

Focus on **tools** and **rooms**

# Learning by Reading



# 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



elephant → C&C Tools → Boxer

Entity relations

elephant

C&C Tools

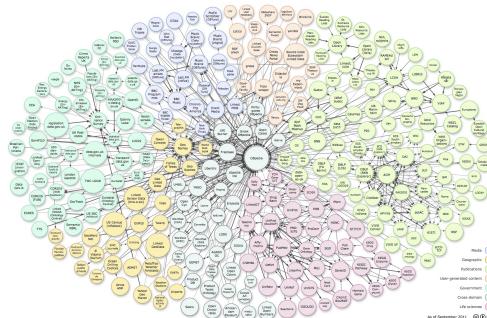
Boxer

Entities

Entity relations

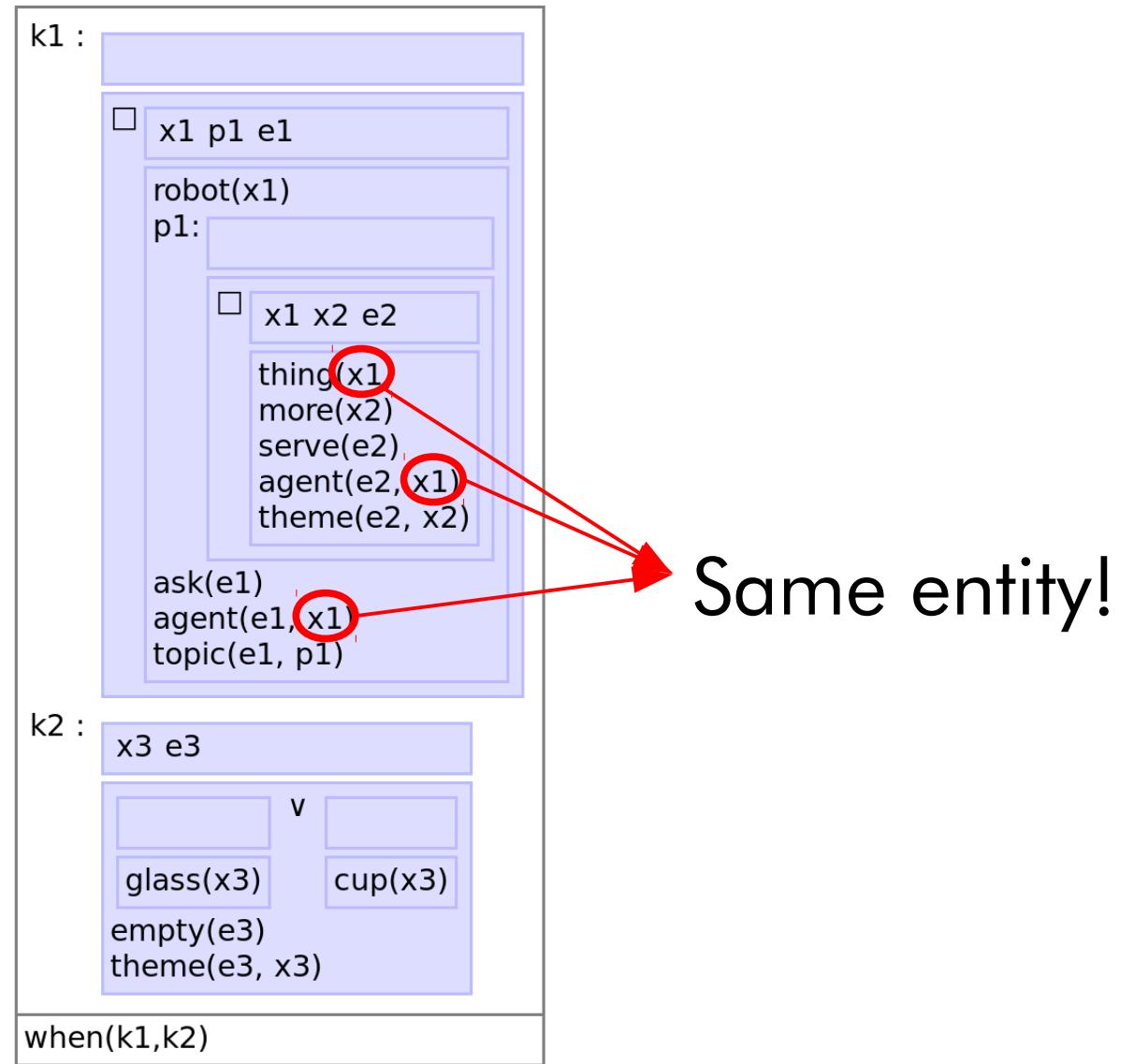
New Knowledge

Entity categorization



# 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.