

# Computational Semantics and Knowledge Engineering

**Johan Bos**

University of Rome "La Sapienza"



# A text understanding system

---

*User:*

Mary has a dog.  
Does Mary have an animal?

*System:*

# A text understanding system

---

*User:*

Mary has a dog.

Does Mary have an animal?

*System:*

Yes! (dogs are animals)

# A text understanding system

---

*User:*

Vincent is not married.

Vincent is married.

*System:*

# A text understanding system

---

*User:*

Vincent is not married.

Vincent is married.

*System:*

Hey! That's inconsistent.

# A text understanding system

---

*User:*

Florence is the cultural capital of Italy.  
Is Florence the capital of Italy?

*System:*

# A text understanding system

---

*User:*

Florence is the cultural capital of Italy.  
Is Florence the capital of Italy?

*System:*

Yes! (a cultural capital is a capital)

# A text understanding system

---

*User:*

Mia's husband is a gangster.  
Is Mia married?

*System:*

# A text understanding system

---

*User:*

Mia's husband is a gangster.  
Is Mia married?

*System:*

Uhh, don't know...

# Knowledge acquisition

---

- You need knowledge?
- You want a machine to get it?
- Why don't you get it from ... *text...*?
  - Wikipedia
  - Definitions
  - Dictionary glosses

# Catch 22

- To build an intelligent NLP system we need background knowledge
- To acquire background knowledge automatically we need NLP



# This talk

---

- Background: computational semantics
- Building a text understanding system
  - Syntax-semantics interface
  - Semantic representation and inference
  - Knowledge
- Open-domain question answering
  - Why you need semantics
  - Why you need knowledge
- A case study

# Computational Semantics

---

- **Semantics**

The branch of logic and linguistics concerned with meaning

- **Computational Semantics**

Using a computer to build meaning representations, and reason with the result (inference)

# Applications

- Spoken Dialogue Systems
- Question-Answering Systems
- Textual Inference Systems



# Reasoning and Natural Language

---

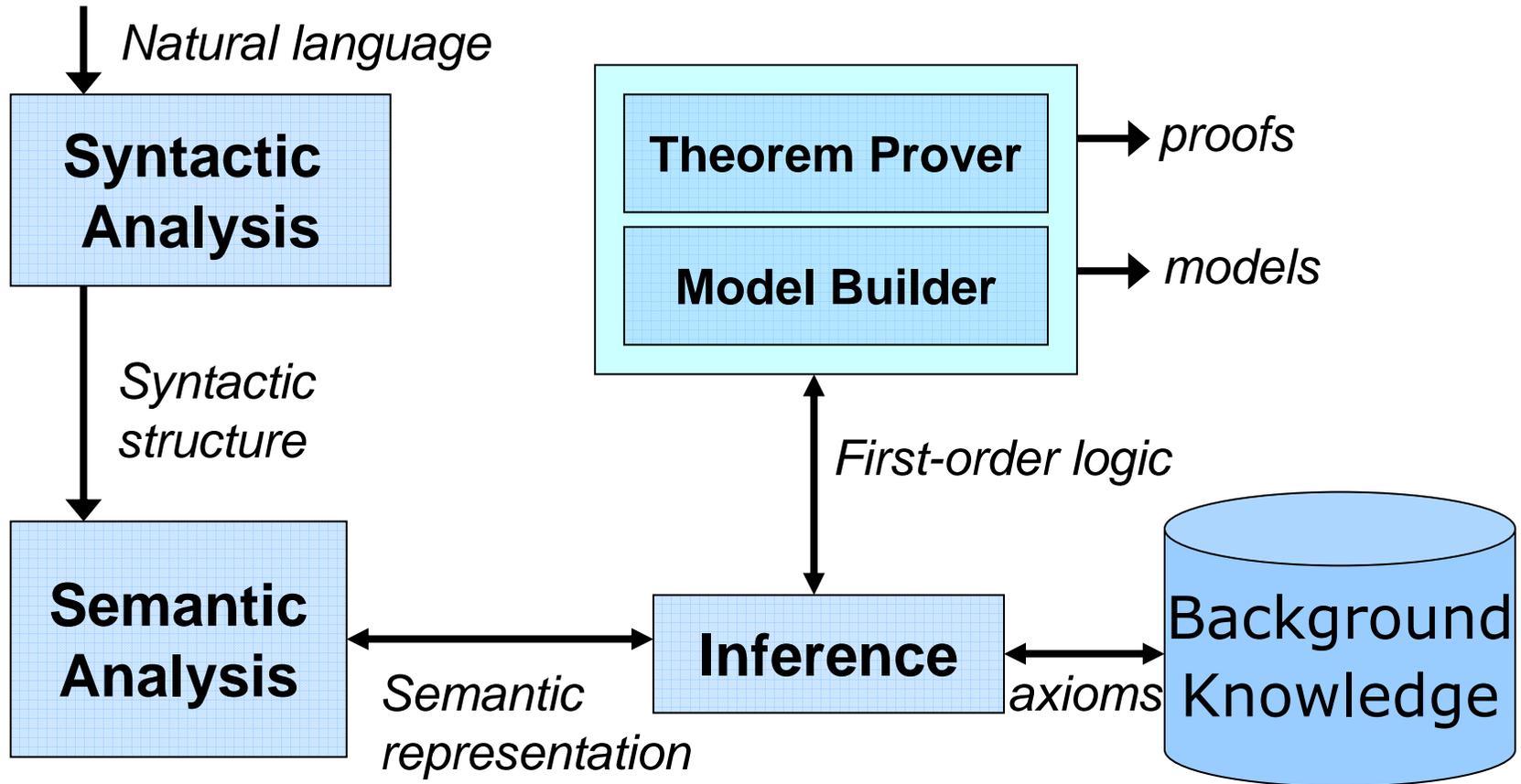
- To reason you need
  - A logical formalism
  - A reasoning engine for your logic
  - Something that maps language into logic
  - Supporting background knowledge

# Reasoning and Natural Language

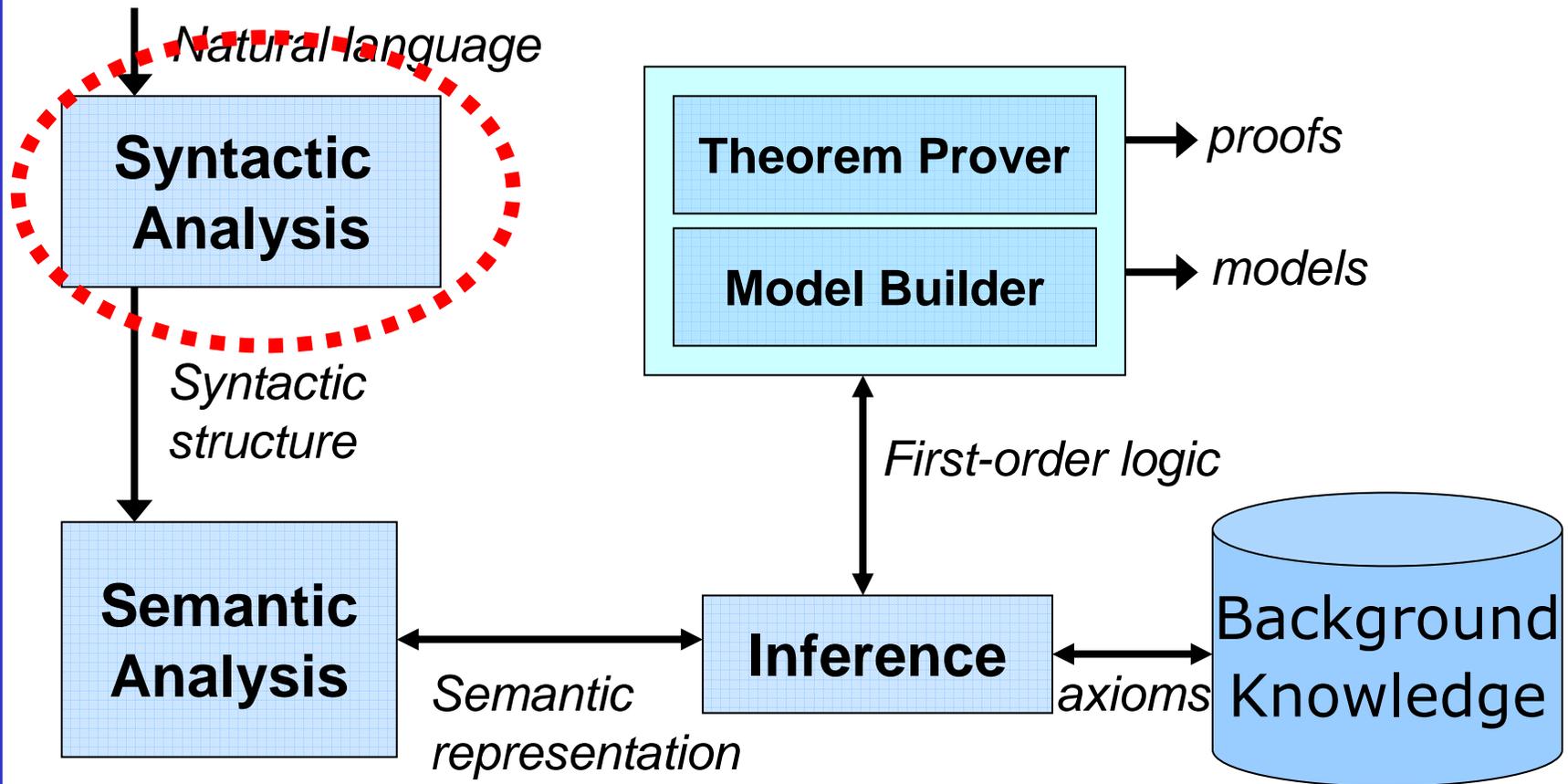
---

- To map language into logic you need
  - A parser that produces syntactic structure
  - A syntax-semantics interface
  - A semantic formalism

# The bigger picture



# The bigger picture

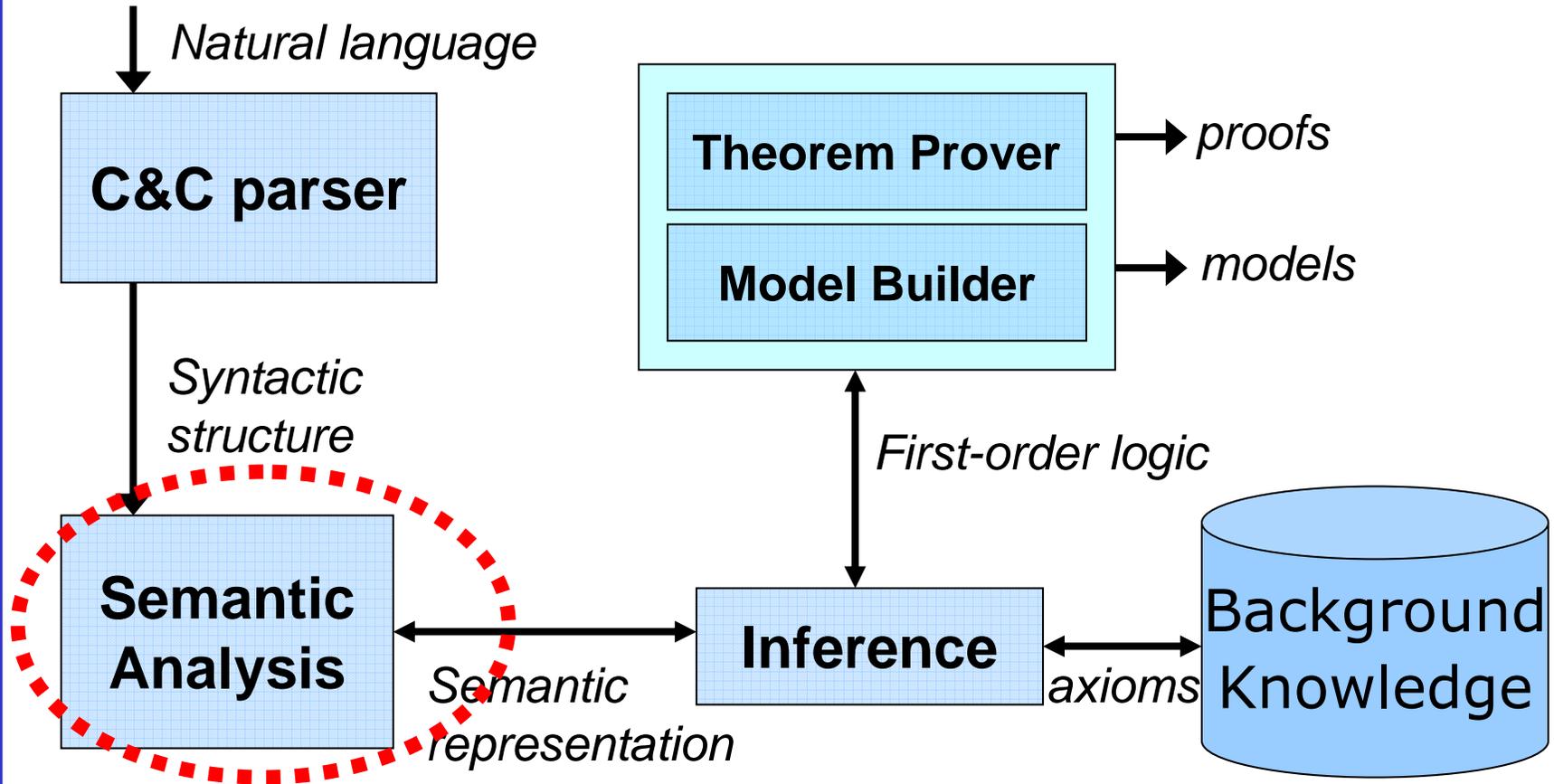


# Syntactic analysis: C&C

---

- Creation of tree-banks
- Stochastic parsers trained on such tree-banks
- C&C parser (Clark & Curran)
  - Combinatory Categorical Grammar
  - Efficient and robust

# The bigger picture



# Semantic Analysis: Boxer

---

- The recent developments in parsing has lead to developing wide-coverage semantic analysis components
- One of such systems is Boxer, developed over the last five years

# Boxer

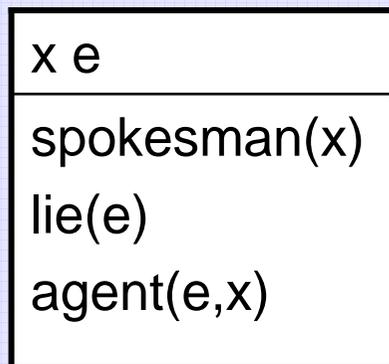
---

- Follows the principles of Hans Kamp's **Discourse Representation Theory**
- Produces formal semantic representations
- Translation to first-order logic
- Systematic syntax-semantic interface, using lambda calculus
- Pronoun and presupposition resolution

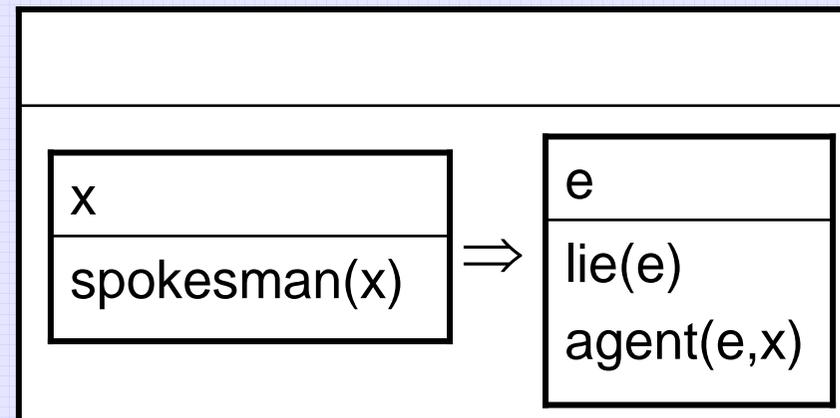
# Discourse Representation Theory

- Box-like structures as semantic representation
- Structure plays role in pronoun resolution
- Neo-Davidsonian event semantics [VerbNet]

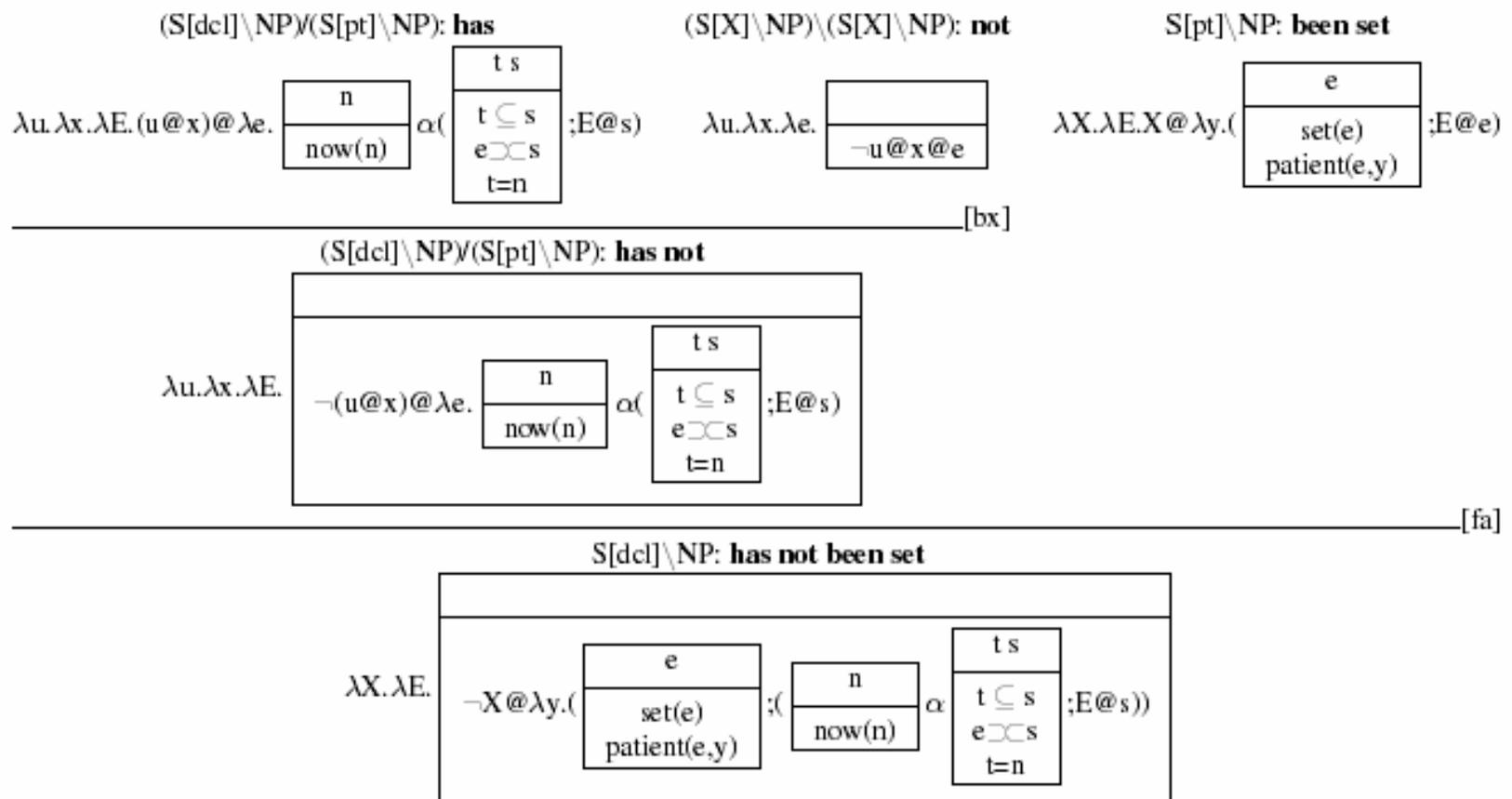
*A spokesman lied.*



*Every spokesman lied.*



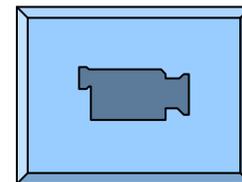
# Syntax-semantics interface



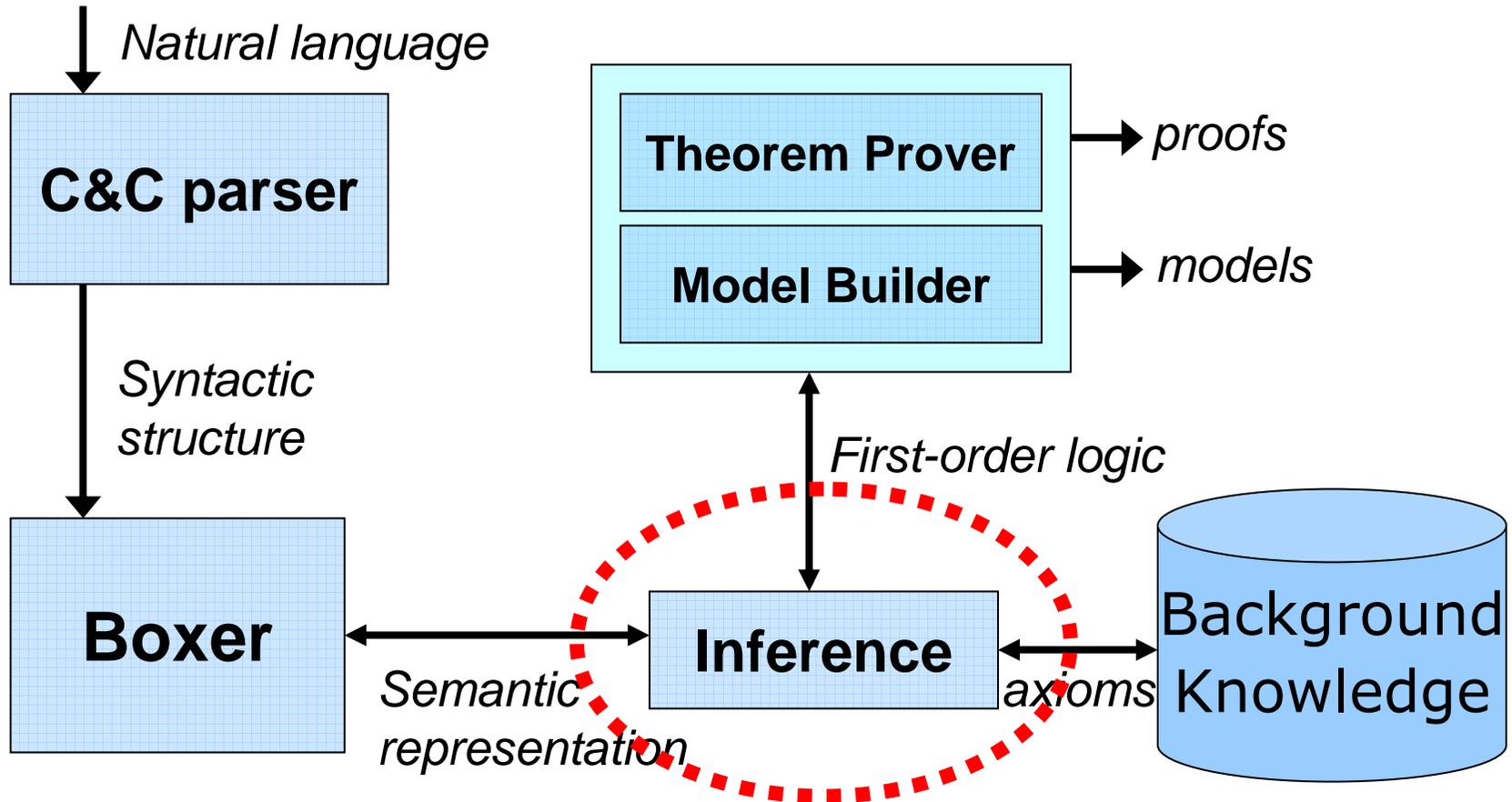
# Breakthrough

---

- Why C&C and Boxer make a difference
  - Broad-coverage
  - Reasonably efficient
  - Clean syntax-semantics interface
  - Interpretable structures



# The bigger picture

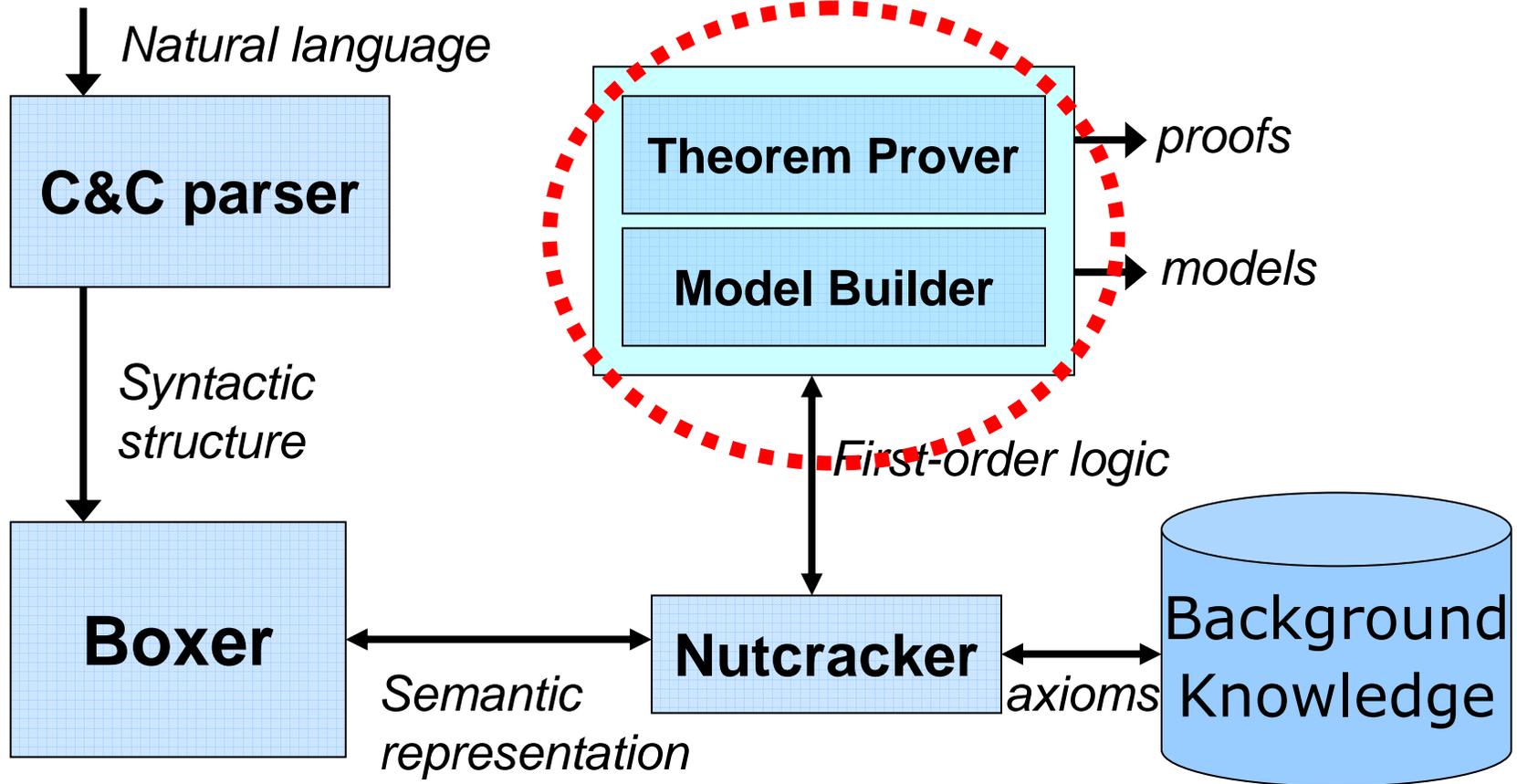


# Inference: Nutcracker



- Selects appropriate background knowledge for inference
- Acts as a mediator between Boxer and inference engines for first-order logic
  - Theorem prover
  - (Finite) Model builders
- Reports results back to Boxer or front-end application

# The bigger picture



# Inference engines

---

- Which inference engines? Off-the-shelf!
- How do we know which are the best?
  - CADE world cup automated deduction
  - Theorem proving: **vampire**
  - Model building: **paradox**

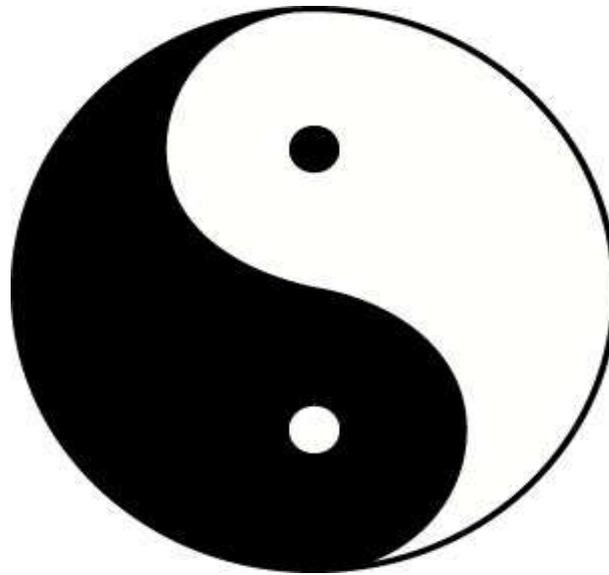
# Some concerns

---

- Isn't first-order logic undecidable?
- Why do we need both a theorem prover and a model builder?

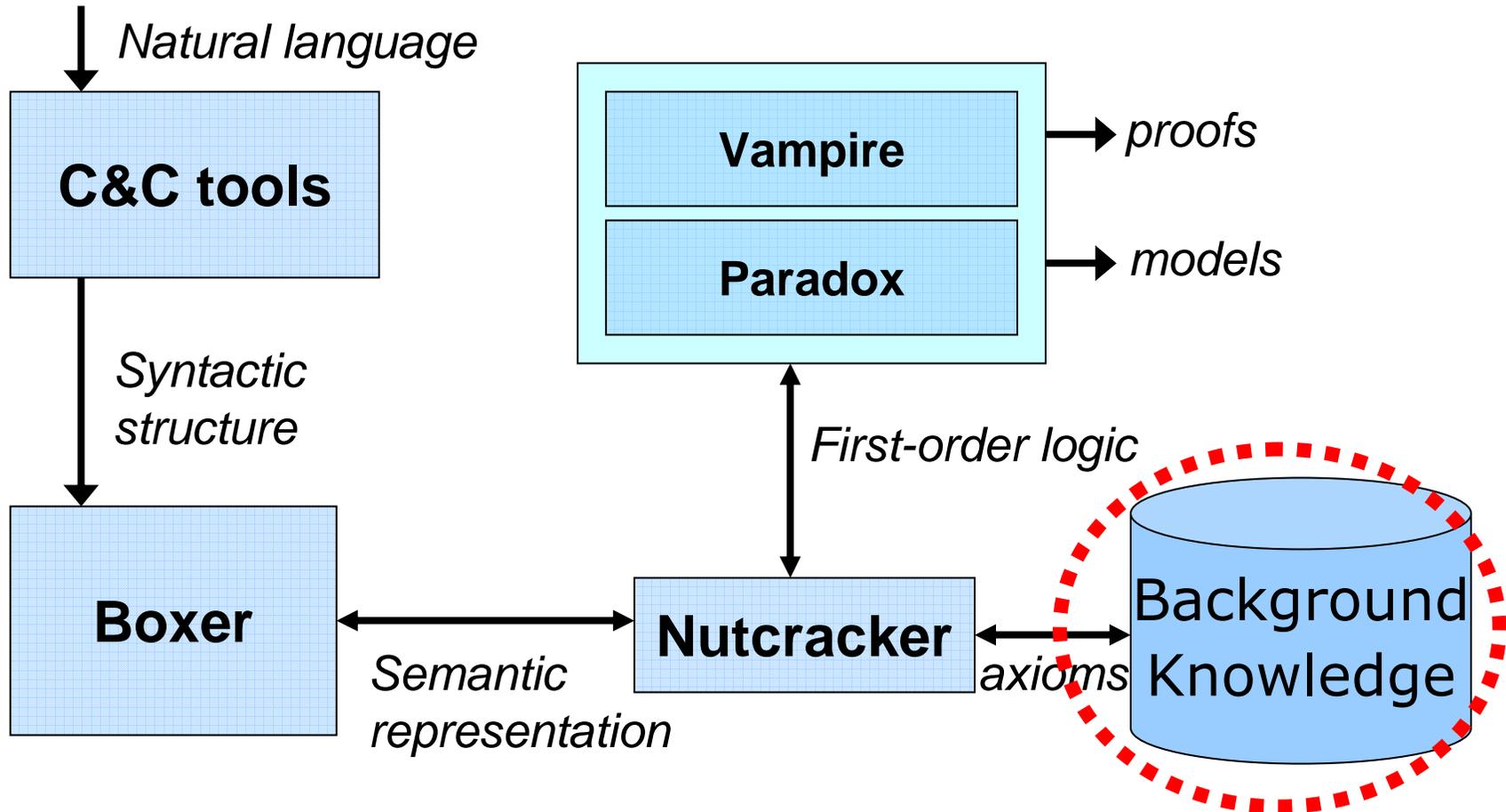
# Yin and Yang of Inference

---



- **Theorem Proving and Model Building** function as opposite forces

# The bigger picture



# Background Knowledge

---

- Format:  
first-order logic
- Amount:  
ideally as few axioms as possible

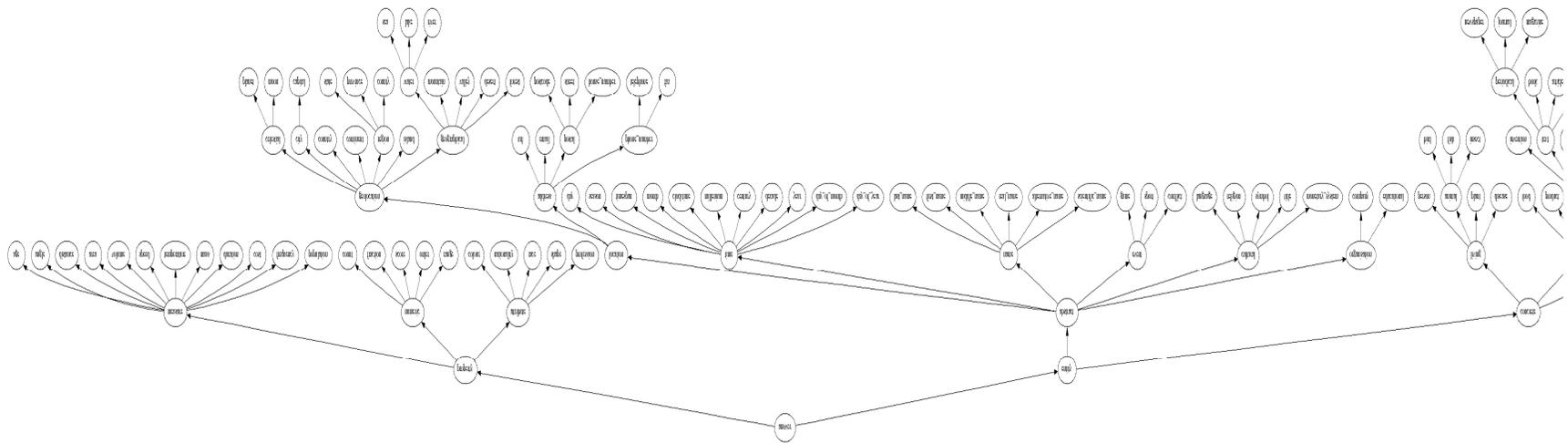
# Man or Machine?

- Manually constructed
  - WordNet
  - NomLex
  - CIA Factbook
  - CYC/OpenCYC
- Automatically constructed
  - Linguistic patterns [e.g. Hearst]
  - Paraphrases [e.g. Lin & Pantel]



# WordNet

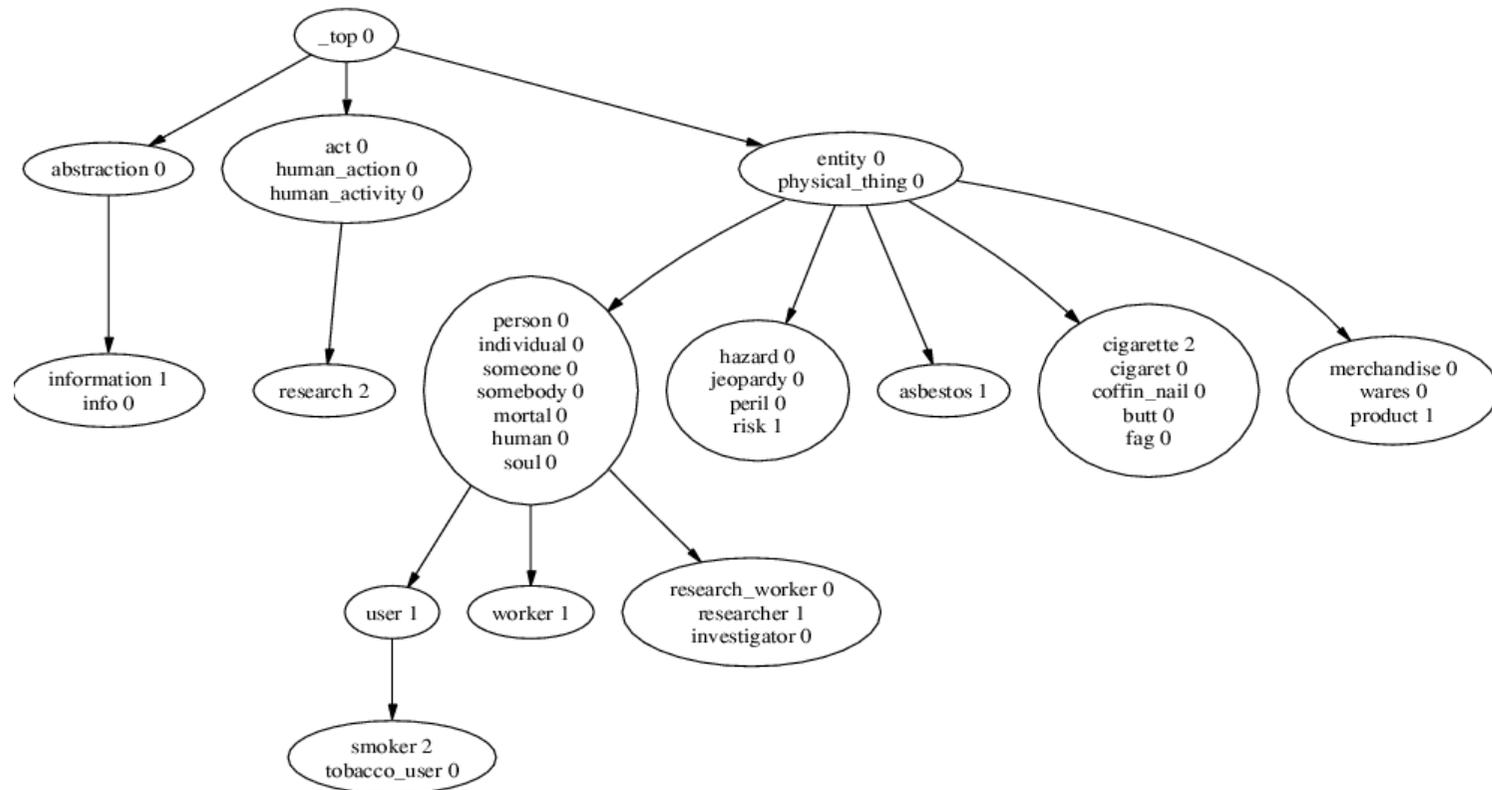
- Use hypernym relations from **WordNet** to build an isa-hierarchy



- Create **MiniWordNets** for small texts
  - Convert these into first-order axioms

# MiniWordNet

There is no asbestos in our products now.  
Neither Lorillard nor the researchers who studied the workers  
were aware of any research on smokers of the Kent cigarettes.

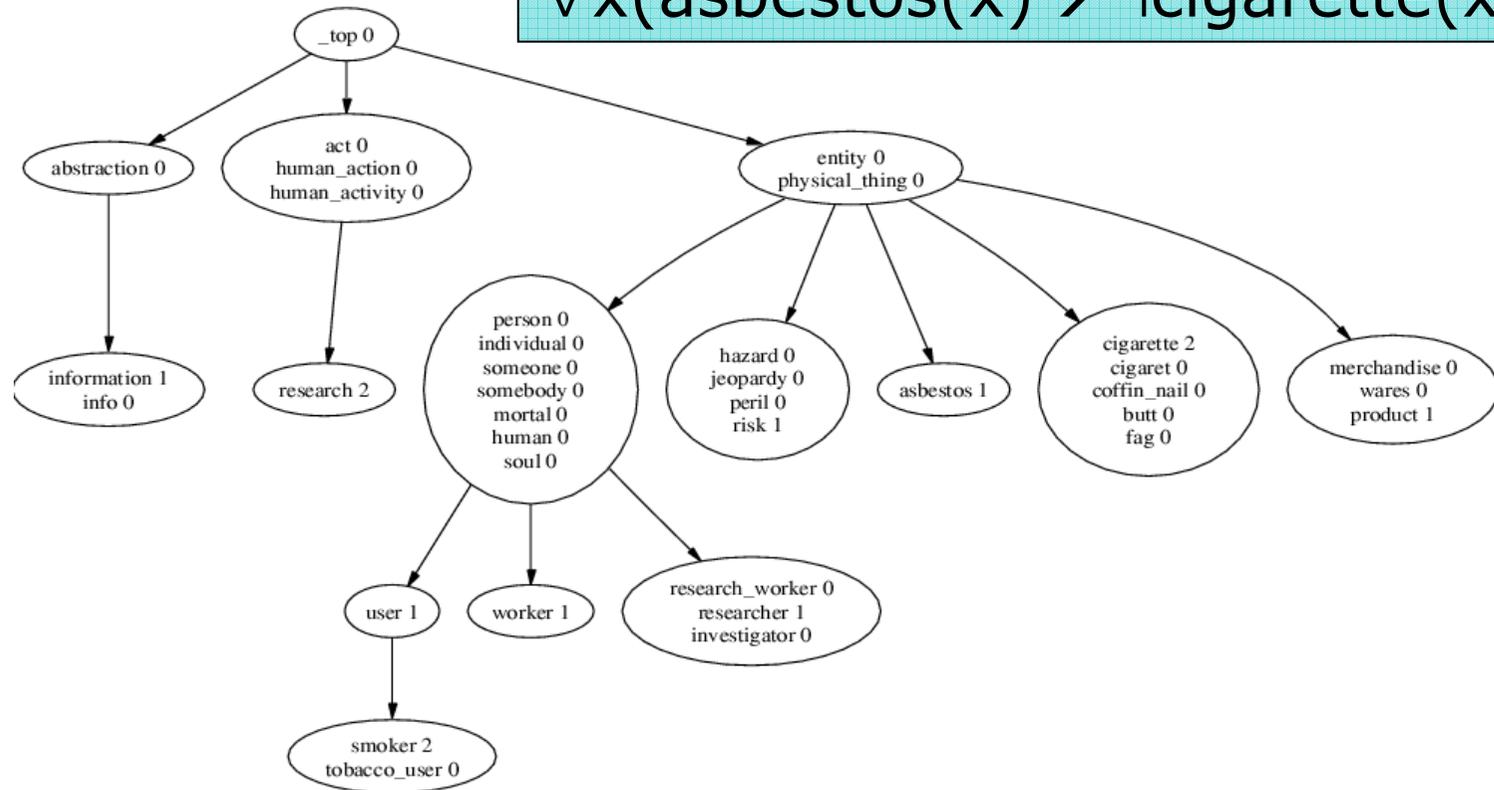


# MiniWordNet

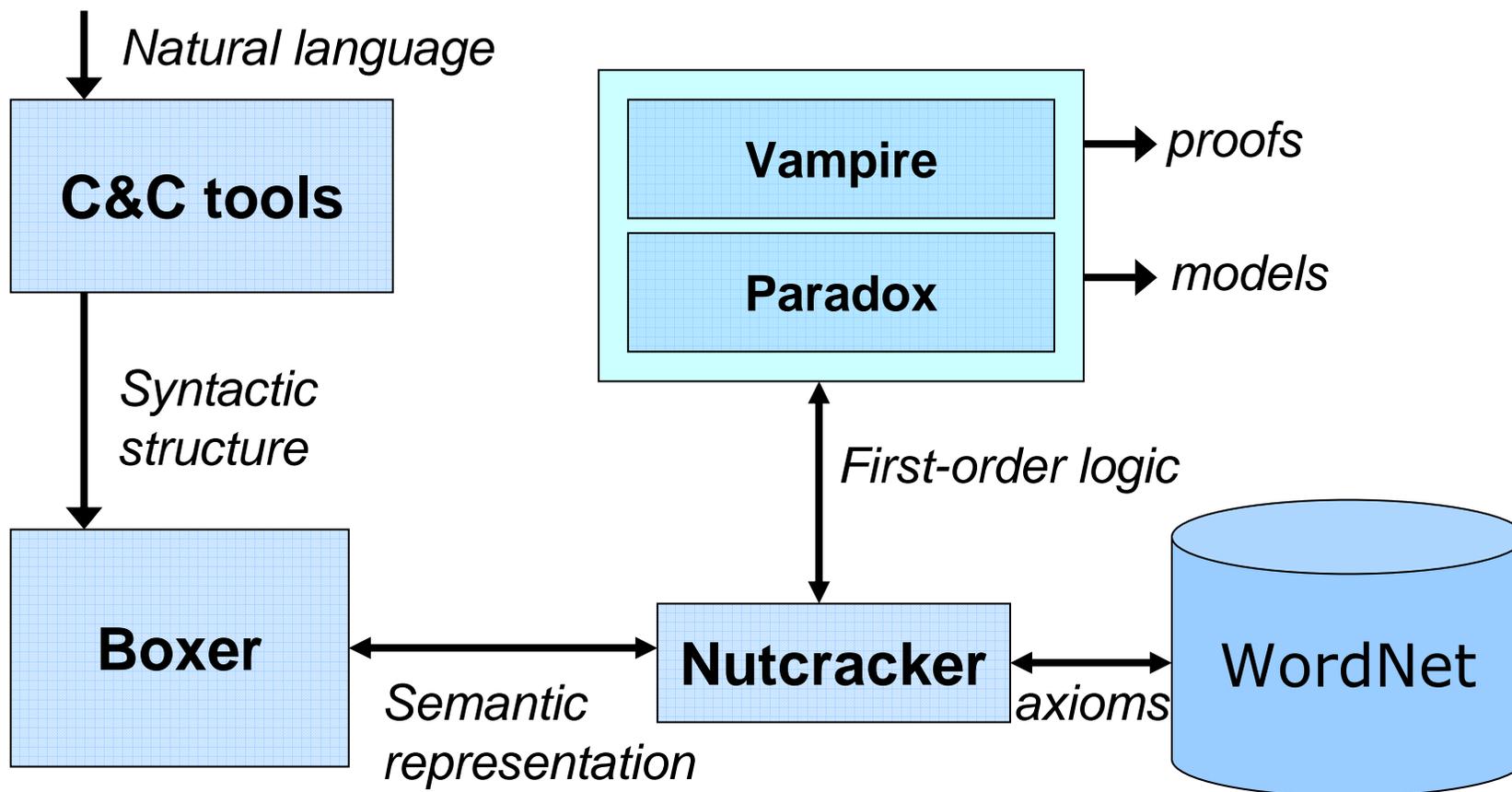
$\forall x(\text{asbestos}(x) \rightarrow \text{entity}(x))$

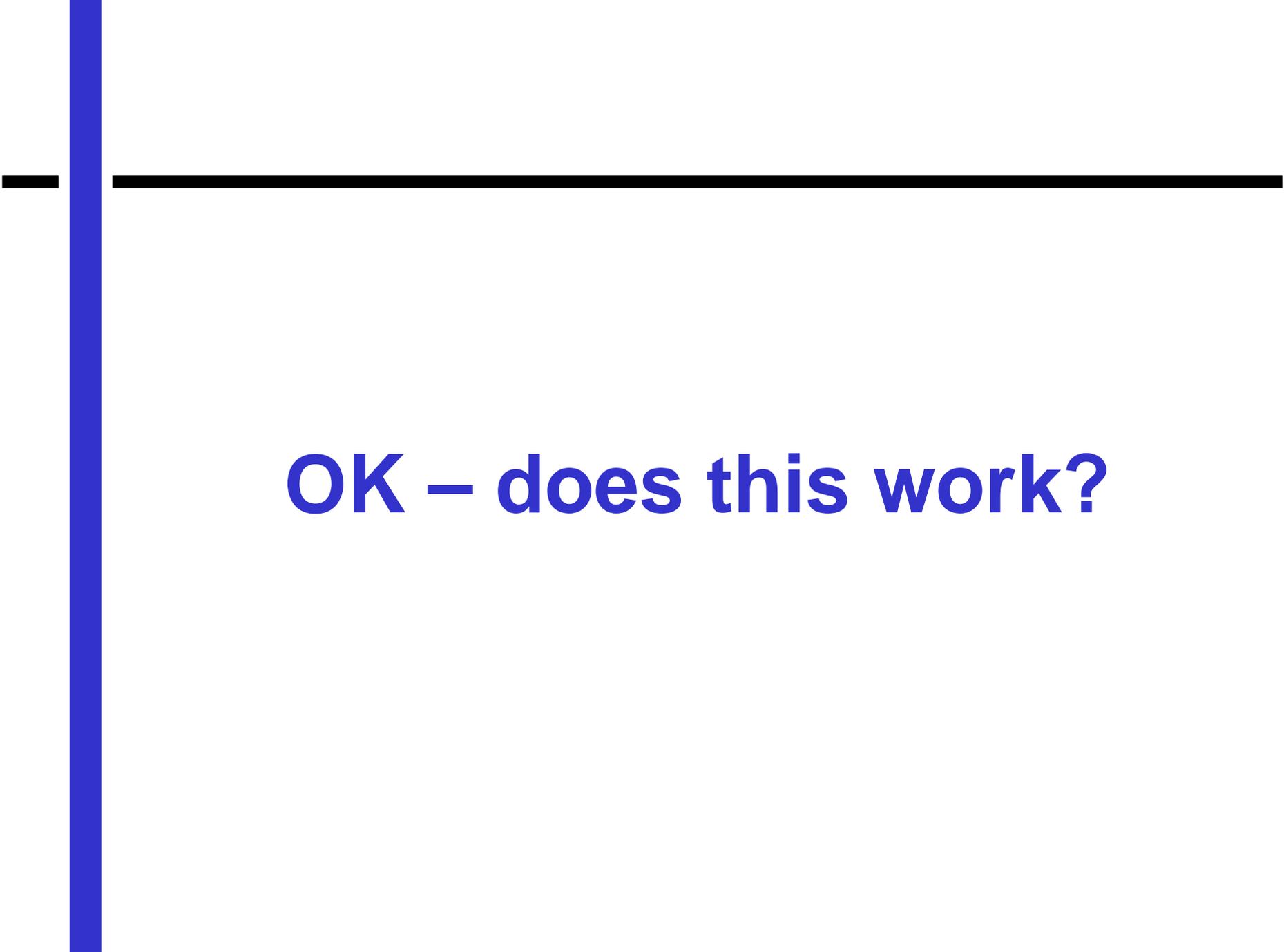
$\forall x(\text{cigarette}(x) \rightarrow \text{entity}(x))$

$\forall x(\text{asbestos}(x) \rightarrow \neg \text{cigarette}(x))$



# The bigger picture





**OK – does this work?**

# It does work for small domains...

Examples:

- Talking robots
- Smart houses



# But what about larger domains?

---

- Semantic interpretation of e.g.
  - newspaper texts
  - web pages
  - wikipedia
- Let's look at a real-world application in which we used C&C and Boxer:
  - Open-domain question answering
  - NIST/TREC campaigns
  - Pronto QA system

# What is Question Answering?

---

- Questions, no queries!
- Answers, not documents!

**Q:** Where did Olof Palme die?

**A:**

# What is Question Answering?

---

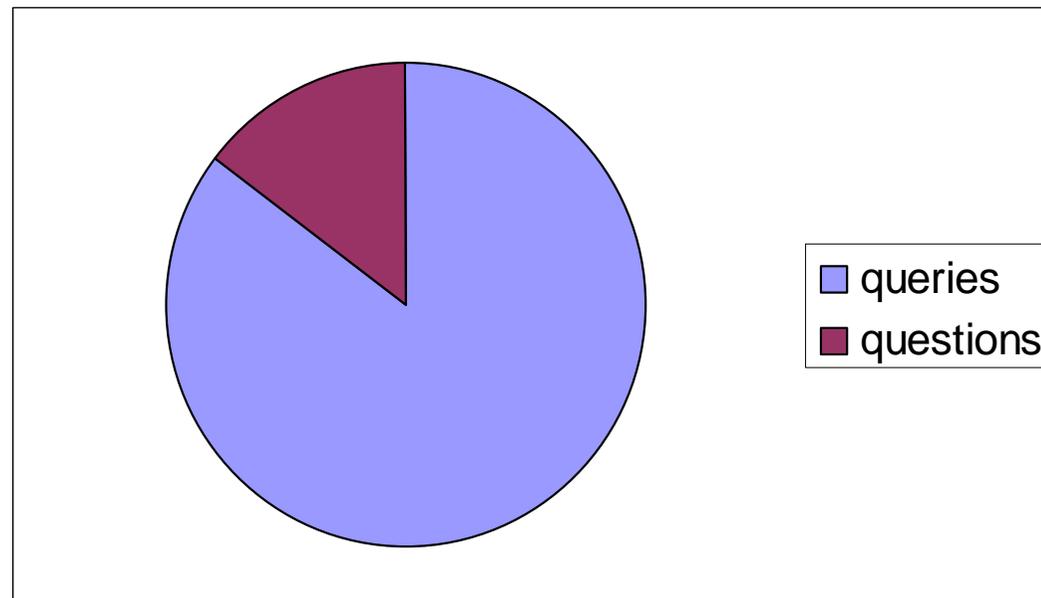
- Questions, no queries!
- Answers, not documents!

**Q:** Where did Olof Palme die?

**A:** In Stockholm.

# People ask questions

- Excite Search Engine Log with 2,477,283 queries
- Collected at 20 December 1999
- Around 15% natural language questions



090037 34A03113DB00014

10 arrowsmith

# Excite Query Log

090037 A2545DBA2067C914

10 Wrestlemania 2000

090037 E0879EB63392B1FA

090037 B5A2C0D05990DEA6

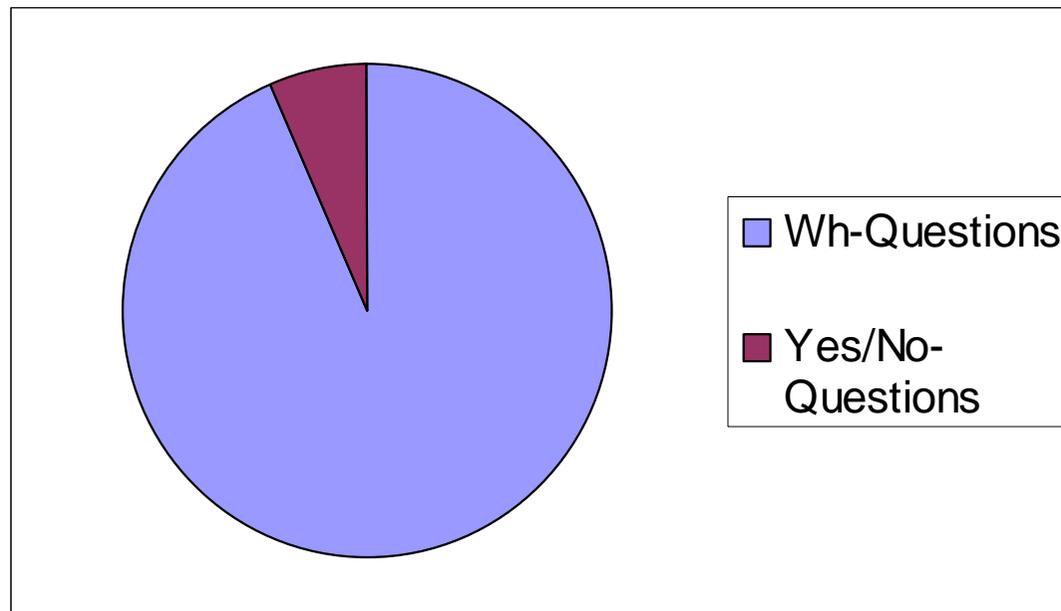
0 What is the average life span of a human male?

090039	82F30745D2BF5C2C	golf
090039	9252B344DB1B7047	www.Odcgov/cia/Publication/Fatbook/index
090039	E6D3DCFA0A39BA3E	www.ufallfahrzeug.de
090039	AFCDD0FD4A65D34C3	montogery ward
090039	F54BC573FF3C94B1	What is the website for the newspaper in Scranton, PA?
090039	3DF4E9B0AFF6B808	Universities in Wellington
090039	009CD46686D4A1FD	jc pantyfreek
090039	F26E5D31D804ED98	carman ministries
090039	D87CE5C149126B4B	Where can I find information about the television show Felicity
090039	91255EEF958730CE	how can i convert liquid audio files to wav files?
090039	03D685B3F001A858	"CD duplication software"
090039	F54BC573FF3C94B1	What does the FDA say is its mandate?
090039	DAF1CD4E35BB6D98	ebony ayres+homepage
090039	BDF5F884F42DB05F	hacking VP Instant Message readers messenger
090039	1246BA311FD5AF65	thiazolidisediones
090039	9F0FF534C84D7873	chicago newspapers
090039	E00F9C2509761702	yahoo
090039	70B9F6D94EB08EB7	New Century Professions, Inc.
090039	5AE9D0A42F05FDE1	What presidents were born in texas

20 hollywood celebrity residences

# Question Types

- Wh-Questions: 342,075
- Yes/No-Questions: 23,882



# Question Types

How is the U.N. funded?

Was Don lying about the shooting?

Who invented the rangefinder?

what exactly is a hydraulic system?

When was Fisher Price started?

Where can i find exams on the internet about database management systems course with sol

Where can I find yesterday's weather?

What is the probability of extraterrestrial life existing in the Universe?

What vacation rentals are available in the pacific Northwest?

Who was president Grant's private secretary?

Where can I get a recipe for sloppy joes?

What are the times of sunrises and sunsets in various cities in the US?

pros and cons of professional managers in government?

Where can I find information about FLSA?

where can I find details of property in Scotland?

Who sings the song "Ice, Ice Baby"?

what percentage of schools have internet access?

When were gingerbread houses invented?

how do I find out the purchase price of a house that sold in California?

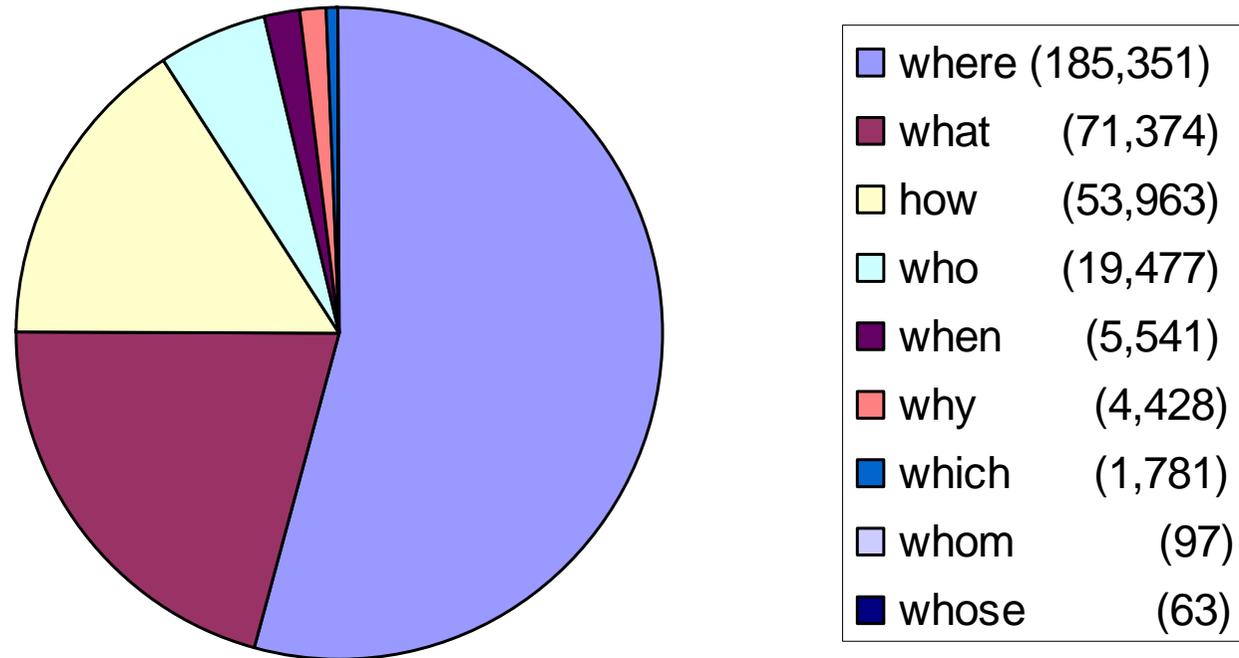
Are computer chips made from sand?

what is lupus?

where can i find how to write a resume?

WHERE CAN I FIND ON-LINE GREETING CARDS?

# Wh-Question Types



# Natural search

---

- Google gets about 200 million searches per day
- That would mean about 30 million natural language searches

# Trying to guess

---

**Q:** What country is Berlin in?

**A:** Ties between **Germany** and Iran strained after a Berlin court verdict ruled on April 10 that Iran's top leaders were behind the assassination of four exiled Kurdish opposition leaders in a Berlin restaurant.

# Trying to guess

---

**Q:** What country is Berlin in?

**A:** Ties between Germany and **Iran** strained after a Berlin court verdict ruled on April 10 that Iran's top leaders were behind the assassination of four exiled Kurdish opposition leaders in a Berlin restaurant.

# Another guessing example

**Q:** Port Arthur Massacre.  
What was the killer's nationality?

**A:** A nation asks why, the portrait of a lone gunman, Martin Bryant, should have no reason to be a killer. The man arrested after the Port Arthur massacre should be a wealthy man, only four years ago he inherited more than 500,000 **Australian** dollars 375,000 US dollars from one of the heirs of George Adams's great Tattersalls fortune.

# Trying to google

---

**Q:** What is the state bird of Alaska?

**A:** Google search:  
"The state bird of Alaska is the \* "

# Trying to google

---

**Q:** What is the state bird of Alaska?

**A:** Google results:  
The state bird of Alaska is the **ptarmigan**.  
The state bird of Alaska is the **willow ptarmigan**.

# Trying to google

**Q:** What is the state bird of Alaska?

**A:** Google results:  
The state bird of Alaska is the **ptarmigan**.  
The state bird of Alaska is the **willow ptarmigan**.  
The state bird of Alaska is the **mosquito**.  
The state bird of Alaska is the **Mosquito**.

# Using Semantics and Inference

---

**Q:** Where did Olof Palme die?

**A:** Stockholm.

# Inference, internally

**Q:** Where did Olof Palme die?

**A:** .....  
given clause #13: (wt=2) 69 [hyper,47,16] vehicle(\$c3).  
given clause #14: (wt=2) 71 [hyper,50,16] vehicle(\$c4).  
given clause #15: (wt=2) 74 [hyper,52,18] building(\$c5).  
given clause #16: (wt=3) 51 [] have(vincent,\$c4).  
given clause #17: (wt=2) 76 [hyper,55,18] building(\$c6).  
given clause #18: (wt=2) 78 [hyper,61,16] vehicle(\$c6).  
given clause #19: (wt=2) 80 [hyper,65,10] organism(\$c1).  
given clause #20: (wt=2) 84 [hyper,67,10] organism(\$c2).  
given clause #21: (wt=3) 53 [] die(palme,\$c5).  
given clause #22: (wt=2) 86 [hyper,69,15] instrument(\$c3).  
given clause #23: (wt=2) 88 [hyper,71,15] instrument(\$c4).  
given clause #24: (wt=2) 90 [hyper,74,7] artifact(\$c5).  
given clause #25: (wt=2) 94 [hyper,76,7] artifact(\$c6).  
given clause #26: (wt=3) 56 [] \$c7=\$c6.  
given clause #27: (wt=2) 96 [hyper,78,15] stockholm(\$c6).  
----> EMPTY CLAUSE at 0.01 sec ----> 113 [hyper,96,24,76] \$F.

# Inference, externally

**Q:** Where did Olof Palme die?

- A:**
1. You are looking for location.
  2. I know that Stockholm is a city.
  3. Every city is a location.
  4. If x is shot to death then x died.
  5. I found the following evidence in document APW20000227.0124:  
"In 1986, Swedish Prime Minister Olof Palme was shot to death in central Stockholm. "
  6. Hence **Stockholm** is the answer.

# Why bother?

---

- Why Semantics?
- Consider some examples  
[not made up!]

# Predicate Argument Structure

---

**Q:** When was NATO established?

**A:** NATO launched its first attack against Yugoslavia on **March 24**.

WordNet: launch=establish

# Logical operators

**Q:** Where did Ricky Williams, American football player, grow up?

**A:** Texas running back Ricky Williams is from California. Those who follow the Longhorns don't like Williams any less because he didn't grow up in Texas.

# Word Sense Disambiguation

**Q:** Bing Crosby.  
What was his profession?

**A:** Crosby sang her first song, Starlight. She scribbled notes for the song on the back of a menu at a New York jazz **bar** in 1931.

WordNet: **bar** hyponym of **profession**

# Proper Name Reference

---

**Q:** Where is the Taj Mahal?

**A:** The Taj Mahal is a mausoleum located in **Agra, India**, that was built under Mughal Emperor Shah Jahan in memory of his favorite wife, Mumtaz Mahal.

# Proper Name Reference

---

**Q:** Where is the Taj Mahal?

**A:** The Taj Mahal is a mausoleum located in **Agra, India**, that was built under Mughal Emperor Shah Jahan in memory of his favorite wife, Mumtaz Mahal.

**A:** The Taj Mahal first opened its door to a receptive but cautious public back in 1964. Being the first restaurant of its kind in **Stevenage**, our main problems were initially to get people to try foods that they had never tried before.

# Granularity

**Q:** Where did Franz Kafka die?

**A:**

# Granularity

---

**Q:** Where did Franz Kafka die?

**A:** In his bed

# Granularity

---

**Q:** Where did Franz Kafka die?

**A:** In his bed

**A:** In a sanatorium

# Granularity

**Q:** Where did Franz Kafka die?

**A:** In his bed

**A:** In a sanatorium

**A:** In Kierling

# Granularity

---

**Q:** Where did Franz Kafka die?

**A:** In his bed

**A:** In a sanatorium

**A:** In Kierling

**A:** Near Vienna

# Granularity

---

**Q:** Where did Franz Kafka die?

**A:** In his bed

**A:** In a sanatorium

**A:** In Kierling

**A:** Near Vienna

**A:** In Austria

# Moral

---

- Perhaps semantics and inference will get you quite far, but there is still this need for knowledge

# Knowledge acquisition

---

- What kind of knowledge do we need for QA applications (apart from WordNet)
- Can we produce this on the fly?
- Let's have a look at some examples
  - Linguistic patterns
  - Paraphrase collections

# Failing instances

## **TREC 20.2 (Concorde)**

What airlines have Concorde in their fleets?

- WordNet has no instances of airlines.

- Note: DBpedia does. So does CYC.

# Find instances with patterns

## TREC 20.2 (Concorde)

What airlines have Concorde in their fleets?

- Search for linguistic patterns in corpora (Hearst 1992)
- Example pattern: “X such as Y and”
- Text: ... said that airlines such as Continental and United now fly...

# Find instances with patterns

## TREC 20.2 (Concorde)

What airlines have Concorde in their fleets?

- Search for linguistic patterns in corpora (Hearst 1992, Aguado de Cea et al. 2008)
- Example pattern: “X such as Y and”
- Text: ... said that airlines such as Continental and United now fly...

# Pattern result

## TREC 20.2 (Concorde)

What airlines have Concorde in their fleets?

- Knowledge (Acquaint corpus):

Air Asia, Air Canada, Air France, Air Mandalay, Air Zimbabwe, Alaska, Aloha, American Airlines, Angel Airlines, Ansett, Asiana, Bangkok Airways, Belgian Carrier Sabena, British Airways, Canadian, Cathay Pacific, China Eastern Airlines, China Xinhua Airlines, Continental, Garuda, Japan Airlines, Korean Air, Lai, Lao Aviation, Lufthansa, Malaysia Airlines, Maylasian Airlines, Midway, Northwest, Orient Thai Airlines, Qantas, Seage Air, Shanghai Airlines, Singapore Airlines, Skymark Airlines Co., South Africa, Swiss Air, US Airways, United, Virgin, Yangon Airways

# Paraphrases

## TREC 4.2 (James Dean)

When did James Dean die?

----

APW19990929.0165: In **1955**, actor James Dean was killed in a two-car collision near Cholame, Calif.

- DIRT database (Lin & Pantel):

"X was killed in Y" == "X died in Y"

# Paraphrases

## TREC 4.2 (James Dean)

When did James Dean die?

----

APW19990929.0165: In **1955**, actor James Dean was killed in a two-car collision near Cholame, Calif.

- In first-order logic:

$$\forall x \forall t (\exists e (\text{kill}(e) \& \text{theme}(e, x) \& \text{in}(e, t)) \rightarrow \exists e' (\text{die}(e') \& \text{agent}(e', x) \& \text{in}(e', t)))$$

# General Knowledge

## TREC 14.4 (Horus)

Who was his father?

-----

XIE19990713.004: It also hosted statues of Amon's wife, Mut, the goddess Isis, her husband, **Osiris**, and their son Horus.

- Knowledge (manually coded?):

$$\forall x(\text{husband}(x) \rightarrow \text{male}(x))$$
$$\forall x\forall y((\text{son}(x,y)\&\text{male}(y))\rightarrow \text{father}(y,x))$$

# Knowledge in QA

---

- We need knowledge for inference-based QA system
- This is knowledge not explicitly expressed in the text or question
- Current background knowledge resources are not sufficient
- Automatically extracted knowledge
  - Ideally word sense disambiguated
  - Deal with more complex relations

# Automatically harvested knowledge

---

- Good at simple relations
  - hyponyms, instances, synonyms
  - typically high precision and low recall
- Not so good at paraphrases
  - DIRT paraphrases [Lin and Pantel]
  - no improvement on QA and RTE
  - typically low precision, high recall
- General knowledge is hard to get, but of course we can use Boxer itself!

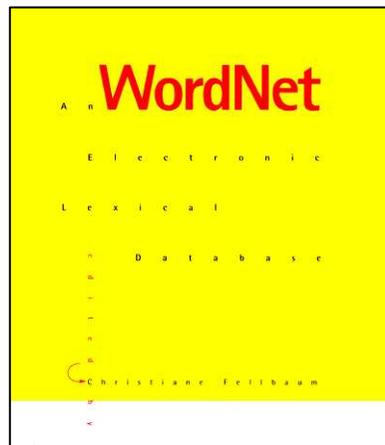
**There is only one catch...**



# A small case study

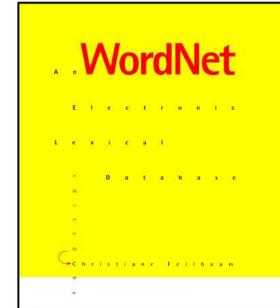
- Let's try to derive knowledge from texts
- Let's use Boxer to analyse definitions
- Pilot study:

Find axioms for **wife** and **husband**



# Case study: *wife* and *husband*

- WordNet 3.0:

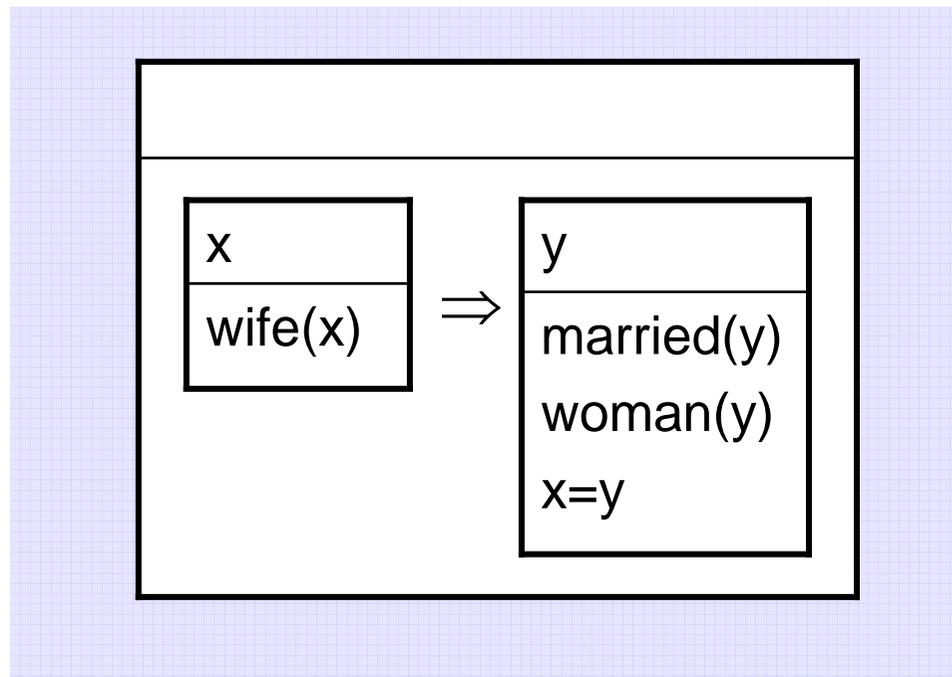


**wife** (a married woman; a man's partner in marriage)

**husband** (a married man; a woman's partner in marriage)

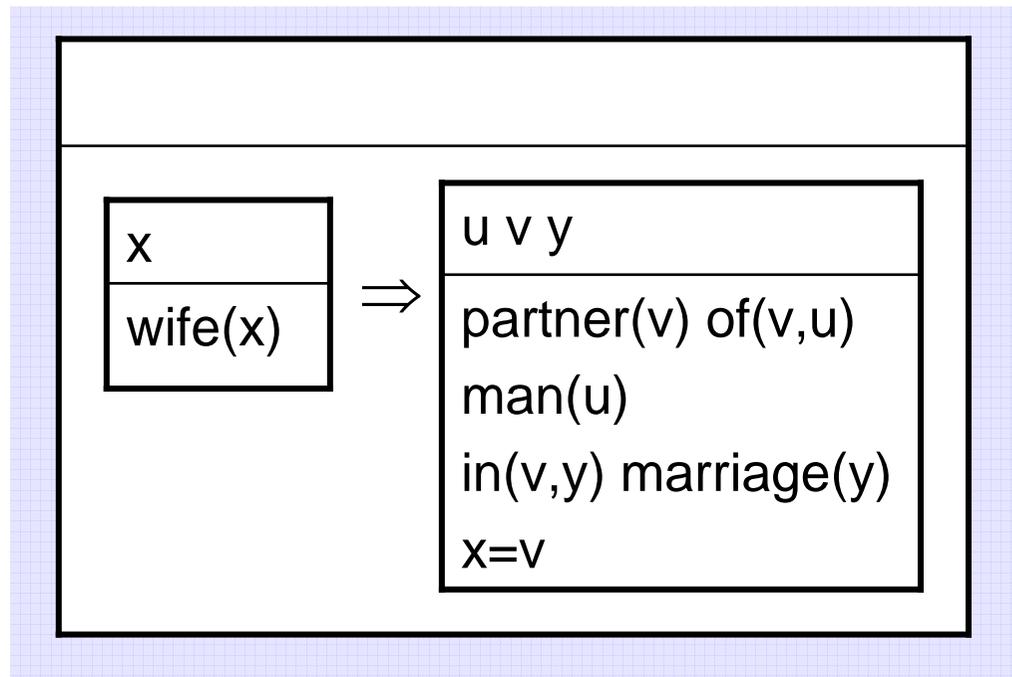
# Boxing WordNet glosses

- Original:  
wife (a married woman; a man's partner in marriage)
- Rephrased (1):  
Every wife is a married woman.



# Boxing WordNet glosses

- Original:  
wife (a married woman; a man's partner in marriage)
- Rephrased (2)  
Every wife is a man's partner in marriage.



# Case study: *wife and husband*

---

- Wikipedia (Oct 1, 2008):

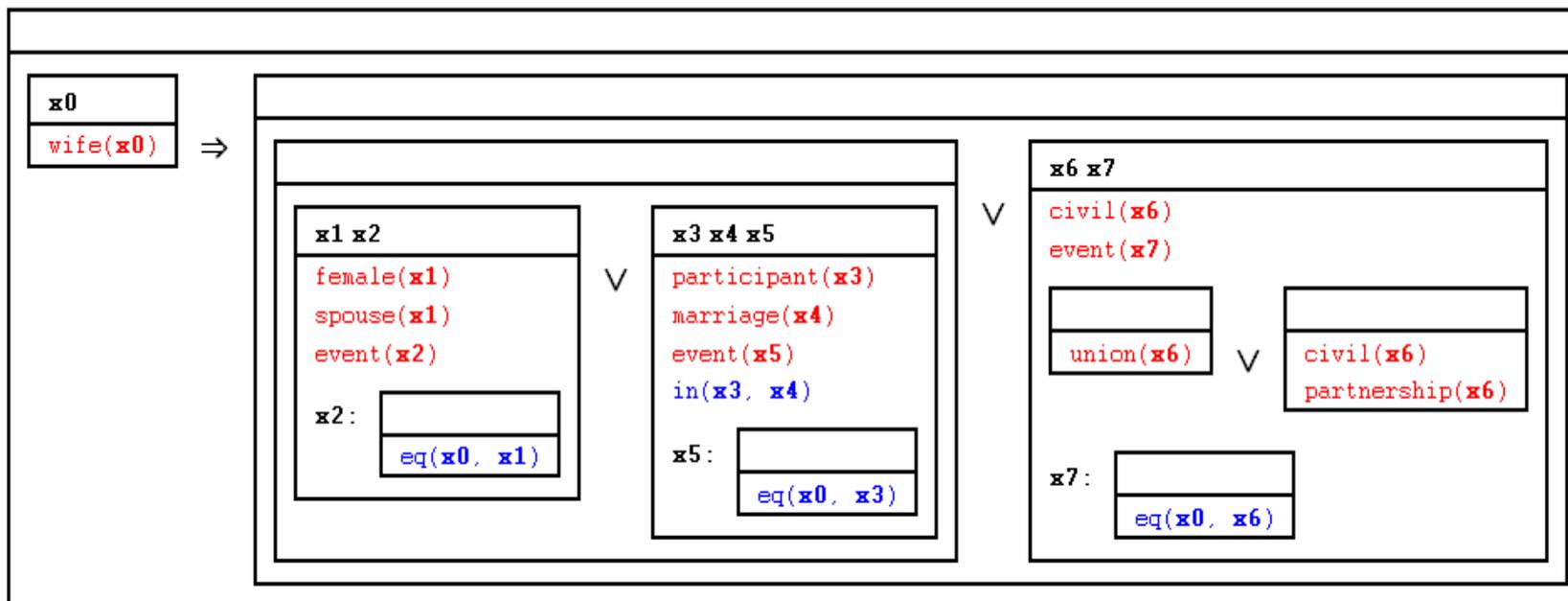


A **wife** is a female spouse, or participant in a marriage, or civil union or civil partnership.

A **husband** is a male spouse (participant) in a marriage, civil union or civil partnership.

# Boxing Wikipedia definitions

- Every **wife** is a female spouse, or participant in a marriage, or civil union or civil partnership.



# Boxing glosses and definitions

---

- Not straightforward to get good results
- (Manual?) reformulation required
- Several issues
  - implication in one or two directions?
  - Word sense disambiguation
  - Modifier attachments
  - Scope of disjunction
  - Interpretation of disjunction
  - Strict or default rules?
- But: inference seems to play no major role

# Ideas for definition analysis

---

- Develop a controlled natural language for definition templates in sources like Wikipedia
- Train new parsing models on annotated definitions

# Conclusions

---

- We can build semantic representations with high coverage and reasonable accuracy for open-domain natural texts
- We also have reasonable inference engines at our disposals that work well enough for small texts
- To use these in real-world applications inference tasks we need additional background knowledge

# More Conclusions

---

- Some of the background knowledge can be derived from existing resources and ontologies
- The word sense problem makes symbol grounding in ontologies hard
- Deriving general knowledge from natural language definitions sounds like a nice idea but is not without stumbles

# And finally...

- The C&C tools, Boxer and Nutcracker are freely available for research
  - <http://svn.ask.it.usyd.edu.au/trac/candc/wiki>
  - There is also an online demo

