# Developing and Applying
# An Integrated Semantic Framework
## for Natural Language Understanding

Le Tuan Anh
Supervised by: Francis Bond

Division of Linguistics and Multilingual Studies
School of Humanities and Social Sciences
Nanyang Technological University

10th July 2015

# Acknowledgement

My sincere thanks to my committee members:

Supervisor Francis Bond

Co-supervisor Jung-jae Kim

Adviser Randy John La Polla

Adviser Frantisek Kratochvil

and to my colleagues at NTU Computational Linguistics Lab, especially Giulia Bonansinga,

for many helpful critiques, advices, guidance and encouragement upon my works.

# Table of Contents

# Problems

## Observation

We have a lot of text.

# Problems

## Observation

We have a lot of text.

## Facts

- At least 129 million books
- At least 30 million digitalised (Google Books, 2010)
- At least 4.65 billion pages indexed (`http://http://www.worldwidewebsize.com/`)

# Problems

## Observation

We have a lot of text.

## Facts

- At least 129 million books
- At least 30 million digitalised (Google Books, 2010)
- At least 4.65 billion pages indexed (`http://http://www.worldwidewebsize.com/`)

## Challenges

How to make more efficient use of information?

# Problems

## Observation

We have a lot of text.

## Facts

- At least 129 million books
- At least 30 million digitalised (Google Books, 2010)
- At least 4.65 billion pages indexed (`http://http://www.worldwidewebsize.com/`)

## Challenges

How to make more efficient use of information?

## Solutions

Information Systems

# Branches of Research

- Information Retrieval Systems (e.g. Search engines, e-dictionaries, etc.)

# Branches of Research

- Information Retrieval Systems (e.g. Search engines, e-dictionaries, etc.)
- Speech recognizer/synthesizer

# Branches of Research

- Information Retrieval Systems (e.g. Search engines, e-dictionaries, etc.)
- Speech recognizer/synthesizer
- Optical Character Recognition

# Branches of Research

- Information Retrieval Systems (e.g. Search engines, e-dictionaries, etc.)
- Speech recognizer/synthesizer
- Optical Character Recognition
- Machine Translation

# Branches of Research

- Information Retrieval Systems (e.g. Search engines, e-dictionaries, etc.)
- Speech recognizer/synthesizer
- Optical Character Recognition
- Machine Translation
- Automatic Summarisation

# Branches of Research

- Information Retrieval Systems (e.g. Search engines, e-dictionaries, etc.)
- Speech recognizer/synthesizer
- Optical Character Recognition
- Machine Translation
- Automatic Summarisation
- Question Answering

# Branches of Research

- Information Retrieval Systems (e.g. Search engines, e-dictionaries, etc.)
- Speech recognizer/synthesizer
- Optical Character Recognition
- Machine Translation
- Automatic Summarisation
- Question Answering

# Branches of Research

- Information Retrieval Systems (e.g. Search engines, e-dictionaries, etc.)
- Speech recognizer/synthesizer
- Optical Character Recognition
- Machine Translation
- Automatic Summarisation
- Question Answering

# Branches of Research

- Information Retrieval Systems (e.g. Search engines, e-dictionaries, etc.)
- Speech recognizer/synthesizer
- Optical Character Recognition
- Machine Translation
- Automatic Summarisation
- Question Answering

# Branches of Research

- Information Retrieval Systems (e.g. Search engines, e-dictionaries, etc.)
- Speech recognizer/synthesizer
- Optical Character Recognition
- Machine Translation
- Automatic Summarisation
- Question Answering

# Popular Systems

## Successful Stories

- Question Answering (e.g. Wolfram Alpha, Siri)
- Machine Translation (e.g. Google Translate)
- Automatic Summarisation (e.g. IBM Watson)

# Popular Systems

## Successful Stories

- Question Answering (e.g. Wolfram Alpha, Siri)
- Machine Translation (e.g. Google Translate)
- Automatic Summarisation (e.g. IBM Watson)

## Approaches

- Question classification & answer ranking
- Statistical phrase mapping
- Deep linguistic analysis

# My Research

## Research Gap

- How deep is deep linguistic analysis?
- Pipeline: Tokenizer $\rightarrow$ Syntactic Parser $\rightarrow$ Semantic Parser $\rightarrow$ Named Entity Recognition, Co-reference Resolution, etc.
- Statistical methods? Manual analyses?

# My Research

## Research Gap

- How deep is deep linguistic analysis?
- Pipeline: Tokenizer $\rightarrow$ Syntactic Parser $\rightarrow$ Semantic Parser $\rightarrow$ Named Entity Recognition, Co-reference Resolution, etc.
- Statistical methods? Manual analyses?

## Integrated Semantic Framework

- Construction Grammar
- Integrated Semantic Representation
- Hybrid model (statistical and manual analyses)

# Table of Contents

# Linguistic Analysis

## Main aspects of linguistic analysis considered

- Morphosyntactic theory
- Semantic theory

These two are perspectives for looking at a single phenomenon rather than independent components.

# Linguistic Analysis

## Main aspects of linguistic analysis considered

- Morphosyntactic theory
- Semantic theory

These two are perspectives for looking at a single phenomenon rather than independent components.

## Implementation

- A grammar (actual analysis of a language)
- Semantic Representation (i.e. how to "capture" and "store" semantics)

# Morphosyntactic Analysis

## Aspects

- We do not remember all possible sentences
- Underlying system is more complex than that (some sentences, some patterns, etc.)

# Morphosyntactic Analysis

## Aspects

- We do not remember all possible sentences
- Underlying system is more complex than that (some sentences, some patterns, etc.)
- What are the basic elements?
- How do we form sentences?

# Morphosyntactic Analysis

## Aspects

- We do not remember all possible sentences
- Underlying system is more complex than that (some sentences, some patterns, etc.)
- What are the basic elements?
- How do we form sentences?

## Construction Grammar

- Basic units are grammatical constructions
- A construction is a form-function pair
- Constructions can be used to form more complex constructions
- Chosen formalism: Sign-Based Construction Grammar (Boas & Sag, 2012)

# Sign Representation

$$\begin{bmatrix} \text{sign} & \\ \text{PHON/ORTH} & \text{phon/orth-obj} \\ \text{FORM} & \text{morph-obj} \\ \text{SYN} & \text{syn-obj} \\ \text{SEM} & \text{sem-obj} \\ \text{CNTXT} & \text{context-obj} \end{bmatrix}$$

# Example *Dogs sleep.*: Atomic constructions

$$
\begin{bmatrix}
\text{dog-1} & \\
\text{PHON} & [\ldots] \\
\text{FORM} & \begin{bmatrix} \text{LEMMA} & \text{dog} \\ \text{FORM} & \text{dogs} \\ \ldots & \end{bmatrix} \\
\text{SYN} & \begin{bmatrix} \text{noun} & \\ \text{NUM} & \text{pl} \\ \text{COUNTABLE} & + \\ \ldots & \end{bmatrix} \\
\text{SEM} & \begin{bmatrix} \text{IND} & x \\ \text{DOG} & x \\ \ldots & \end{bmatrix} \\
\text{CNTXT} & [\ldots]
\end{bmatrix}
$$

$$
\begin{bmatrix}
\text{sleep-1} & \\
\text{PHON} & [\ldots] \\
\text{FORM} & \begin{bmatrix} \text{LEMMA} & \text{sleep} \\ \text{FORM} & \text{sleep} \\ \ldots & \end{bmatrix} \\
\text{SYN} & \begin{bmatrix} \text{verb-i} & \\ \text{NUM} & \text{non-3sg} \\ \text{SUBJ} & \textbf{subj-constraint} \end{bmatrix} \\
\text{SEM} & \begin{bmatrix} \text{IND} & x \\ \text{SLEEP-ER} & x \\ \ldots & \end{bmatrix} \\
\text{CNTXT} & [\ldots]
\end{bmatrix}
$$

# Example *Dogs sleep.*: Sentence construction

$$
\begin{bmatrix}
\text{dog-sleep-1} & \\
\text{PHON} & [\ldots] \\
\text{FORM} & \begin{bmatrix} \text{FORM} & \text{“dogs sleep”} \\ \ldots & \end{bmatrix} \\
\text{SYN} & \begin{bmatrix} \text{sentence} \\ \\ \ldots \end{bmatrix} \\
\text{SEM} & \begin{bmatrix} \text{IND} & x \\ \text{DOG} & x \\ \text{SLEEP-ER} & x \\ \ldots & \end{bmatrix} \\
\text{CNTXT} & [\ldots]
\end{bmatrix}
$$

# Aspects of meanings

## What is meaning?

Linguistic meaning is the relation between linguistic signs and the things being denoted.

# Aspects of meanings

## What is meaning?

Linguistic meaning is the relation between linguistic signs and the things being denoted.

## Difficulties

**Circularity in definition** What is the foundation of meanings? Words?

**Exact definition** Is *whale* a *mammal* or a *fish*?

**Meaning in context** Pragmatic aspect of meaning

# Formal Semantics

## Main Ideas

- Objectivist semantics: meaning is independent from human mental processes.
- Main theoretician: Richard Montague (1970)

# Formal Semantics

## Main Ideas

- Objectivist semantics: meaning is independent from human mental processes.
- Main theoretician: Richard Montague (1970)

## Formalism

- SYN = { whale, dog, cat, . . . }
- SEM = { WHALE, DOG, CAT, . . . }
- $f = [\![]\!] = \{$ whale $\rightarrow$ WHALE; dog $\rightarrow$ DOG; cat $\rightarrow$ CAT, . . . $\}$

# Formal Semantics

## Main Ideas

- Objectivist semantics: meaning is independent from human mental processes.
- Main theoretician: Richard Montague (1970)

## Formalism

- SYN = { whale, dog, cat, . . . }
- SEM = { WHALE, DOG, CAT, . . . }
- $f = [\![\,]\!] = \{$ whale $\rightarrow$ WHALE; dog $\rightarrow$ DOG; cat $\rightarrow$ CAT, . . . $\}$

## Example

- $[\![whale]\!]$ = WHALE
- $[\![S]\!] = [\![NP \backslash dog]\!]\ [\![VP \backslash bark]\!] = [\![BARK(DOG)]\!] = 1$
- How about $[\![bank]\!]$ = BANK-1 or BANK-2 ???

# Cognitive Semantics

## Main Ideas

- Subjective Semantics: Meanings are created in the human mind
- Guessing game: Speakers try to constrain possible interpretations while hearers try to guess

# Cognitive Semantics

## Main Ideas

- Subjective Semantics: Meanings are created in the human mind
- Guessing game: Speakers try to constrain possible interpretations while hearers try to guess

## Critiques

Subjectivism only captures idiolects (does not capture generalisations)

# Cognitive Semantics

## Main Ideas

- Subjective Semantics: Meanings are created in the human mind
- Guessing game: Speakers try to constrain possible interpretations while hearers try to guess

## Critiques

Subjectivism only captures idiolects (does not capture generalisations)

## Arguments

- An idiolect is a generalised version of language.
- Everything is subjective, even corpus linguistics.

# Frame Semantics

## Main Ideas

Grammatical constructions such as words, phrases, sentences, etc. have meaning because they are linked into conceptual structures in minds. (C. J. Fillmore, 2006)

# Frame Semantics

## Main Ideas

Grammatical constructions such as words, phrases, sentences, etc. have meaning because they are linked into conceptual structures in minds. (C. J. Fillmore, 2006)

## Motivation

1. *Jim's only spent two hours on land today.*
2. *Jim's only spent two hours on the ground today.*
3. *Jim's only spent a few years on this Earth.*

# "Research" Frame

A $\left[Researcher_{(INHERITANCE)}\right]$ (an individual a group or institution) attempts to answer a $\left[Question_{(SUBFRAME)}\right]$ by means of consulting literature, observation, or conducting experiments in a particular $\left[Field_{(SUBFRAME)}\right]$ pertinent to the $\left[Question_{(SUBFRAME)}\right]$. The $\left[Question_{(SUBFRAME)}\right]$ may be underspecified in the form of a $\left[Topic_{(SEE-ALSO)}\right]$.

# Abstract Meaning Representation

## Main Ideas

- Paraphrases have the same semantics

# Abstract Meaning Representation

## Main Ideas

- Paraphrases have the same semantics

## Example

- *The man described the mission as a disaster.*
- *The man's description of the mission: disaster.*
- *As the man described it, the mission was a disaster.*

# Abstract Meaning Representation

## Main Ideas

- Paraphrases have the same semantics

## Example

- *The man described the mission as a disaster.*
- *The man's description of the mission: disaster.*
- *As the man described it, the mission was a disaster.*

## Outputs

```
( d / describe-01
:  arg0 ( m / man )
:  arg1 ( m2 / mission )
:  arg2 ( d / disaster ) )
```

## Problems

Too vague, too abstract, not consistent, ...

# Minimal Recursion Semantics

## Main Ideas

Primary units are **elementary predications** (EP) which are relations and associated arguments.

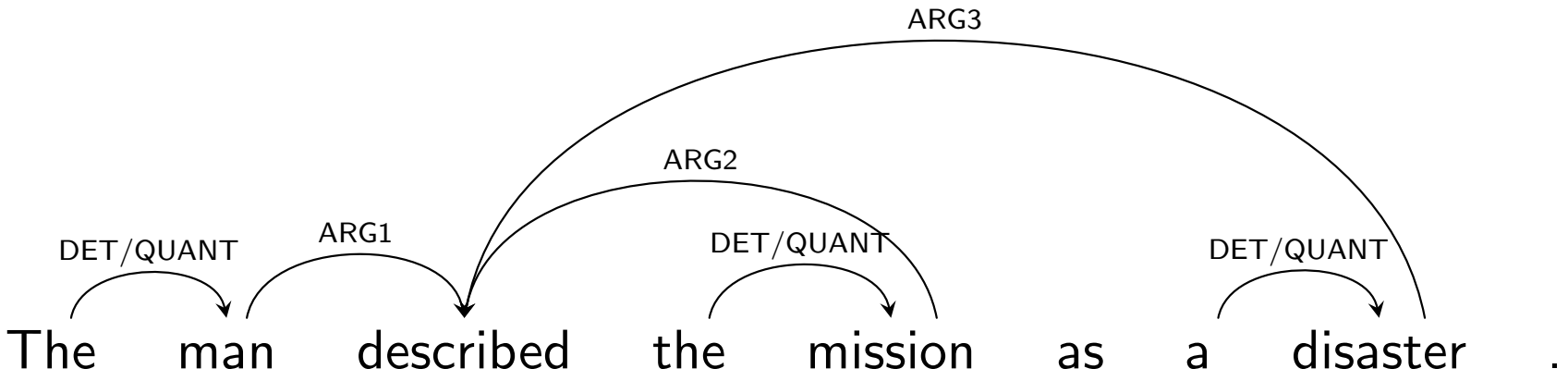
## Principles

**Expressive Adequacy** Linguistic meanings can be expressed correctly

**Grammatical Compatibility** Meanings must be linked to other grammatical information (e.g. Syntax)

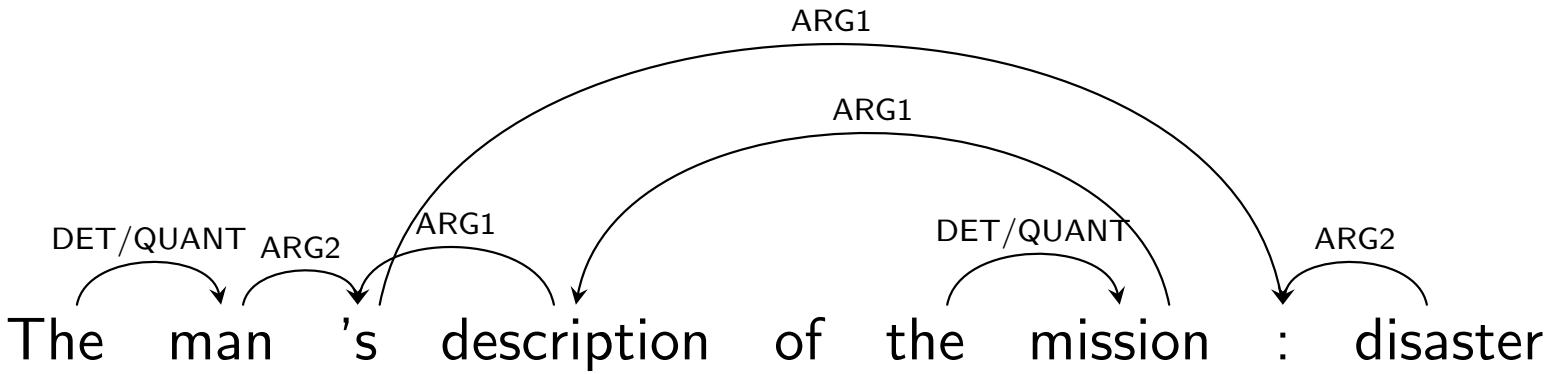
**Computational Tractability** Meanings can be processed and checked for semantic equivalence efficiently

**Underspecifiability** Underspecification should be allowed (leaving semantic distinctions unresolved).




Predicate view

```
 _1:  the_q[x6]
x6:  man_n_1[]
describe_v_as[ARG1 x6, ARG2 x9, ARG3, x10]
_2:  the_q[x9]
x9:mission_n_1[]
_3:  a_q[x10]
x10:disaster_n_1[]
```




Predicate view

```
 _1:_the_q[x6]
x6:_man_n_1[]
_2:def_explicit_q[x11]
e14:poss[ARG1 x11, ARG2 x6]
x11:_description_n_of[ARG1 x15]
_3:_the_q[x15]
x15:_mission_n_1[]
e3:_colon_v_id[ARG1 x11, ARG2 x20]
_4:udef_q[x20]
x20:_disaster_n_1[]
```

# WordNet

## Main Ideas

- Documenting lexical semantics (i.e. words' meanings)
- Traditional method: collections of word-meaning pairs that are ordered by lexicographical order (i.e. dictionaries)
- WordNet: meanings are documented as synonym sets (synsets - collections of near synonyms) with relations among them

# WordNet

## Synsets for *pen*

- Synonym sets

  01698271-v : to produce a literary work
  03906997-n : a writing implement with a point from which ink flows

- Word forms

  01698271-v → write, compose, pen, indite
  03906997-n → pen

# WordNet

## Synsets for *pen*

- Synonym sets

  01698271-v : to produce a literary work
  03906997-n : a writing implement with a point from which ink flows

- Word forms

  01698271-v → write, compose, pen, indite
  03906997-n → pen

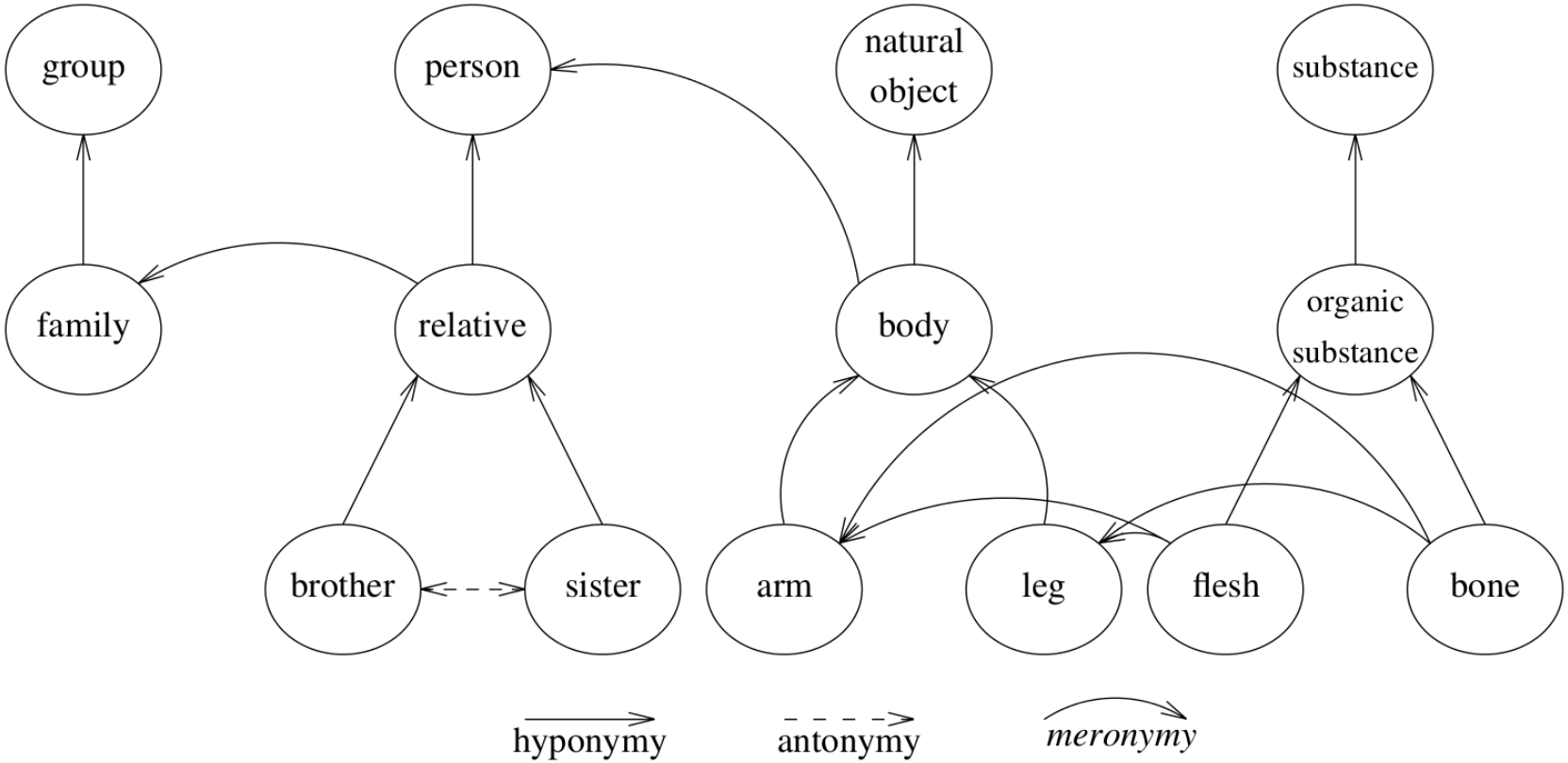
## Main relations

Synonymy/Antonymy Contrastive relations

Hyponymy/Hypenymy Hierarchical relations

Meronymy/Holonymy Part-whole relations

# Synset Relations Example



# Table of Contents

# Pipeline Architecture

## Main Ideas

Pipeline architecture is the trending method. I will discuss its basic design and propose my alternative one in the next section.

## Raw sentence

*I saw a girl with a telescope.*

# Pipeline Architecture

## Main Ideas

Pipeline architecture is the trending method. I will discuss its basic design and propose my alternative one in the next section.

## Raw sentence

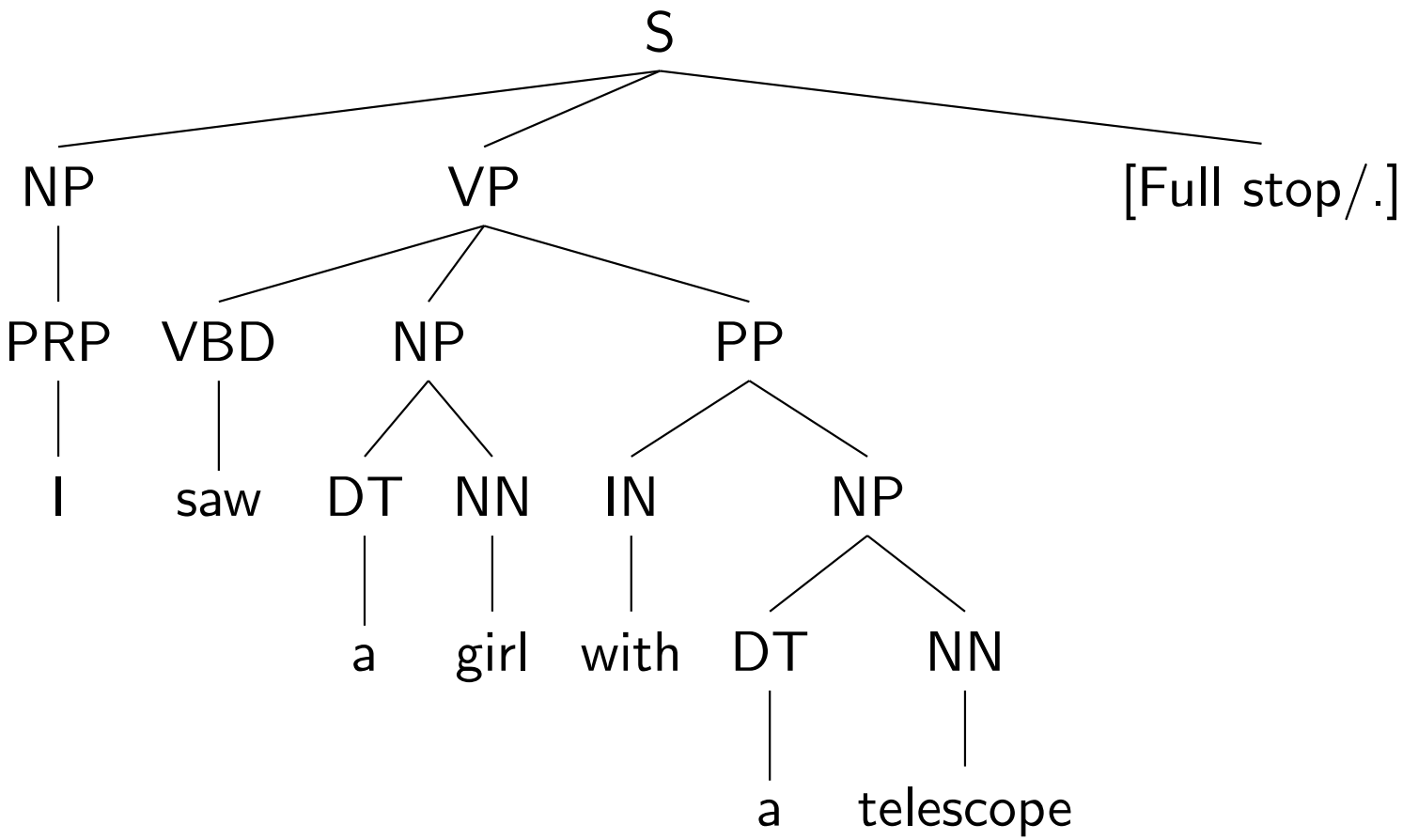
*I saw a girl with a telescope.*

## Tokenization

| *I* | *saw* | *a* | *girl* | *with* | *a* | *telescope* | *.* |
|---|---|---|---|---|---|---|---|
| PRP | VBD | DT | NN | IN | DT | NN | . |

# Pipeline Architecture (Cont.)

## Syntactic Tree



# Pipeline Architecture (Cont.)

## Semantic Role Labelling

| Relation | Word |
|---|---|
| See-er | "I" |
| See-ee | "girl" |
| Instrument | "telescope" |

# Pipeline Architecture (Cont.)

## Semantic Role Labelling

| Relation | Word |
|---|---|
| See-er | "I" |
| See-ee | "girl" |
| Instrument | "telescope" |

## Word-Sense Disambiguation

| Relation | Word | Sense |
|---|---|---|
| See-er | "I" | |
| See-ee | "girl" | girl_n_1 |
| Instrument | "telescope" | telescope_n_1 |

| Sense | Synset ID | Meaning |
|---|---|---|
| girl_n_1 | 10084295-n | a youthful female person |
| telescope_n_1 | 04403638-n | a magnifier of images of distant objects |

# Problems

- Error accumulation
  - Tokenization's errors
  - Part-of-speech tagger's errors
  - Tree parser's errors
  - Semantic parser's errors
- Lack of underspecification

# Integrated Semantic Framework

## Main Ideas

- No pipeline!
- Use a single integrated semantic representation to represent structural and lexical semantics: *paraphrases have similar semantics*.
- Underspecify information when it's not possible to resolve ambiguities in current context (not completed sentence, lack of references, etc.) as the discourse is carried further.

## Powered by

| | |
|---|---|
| Theory | Sign-Based Construction Grammar |
| Grammar | (enhanced) English Resource Grammar |
| Knowledge base | (enhanced) WordNet |
| Representation | **Integrated Semantic Representation** |

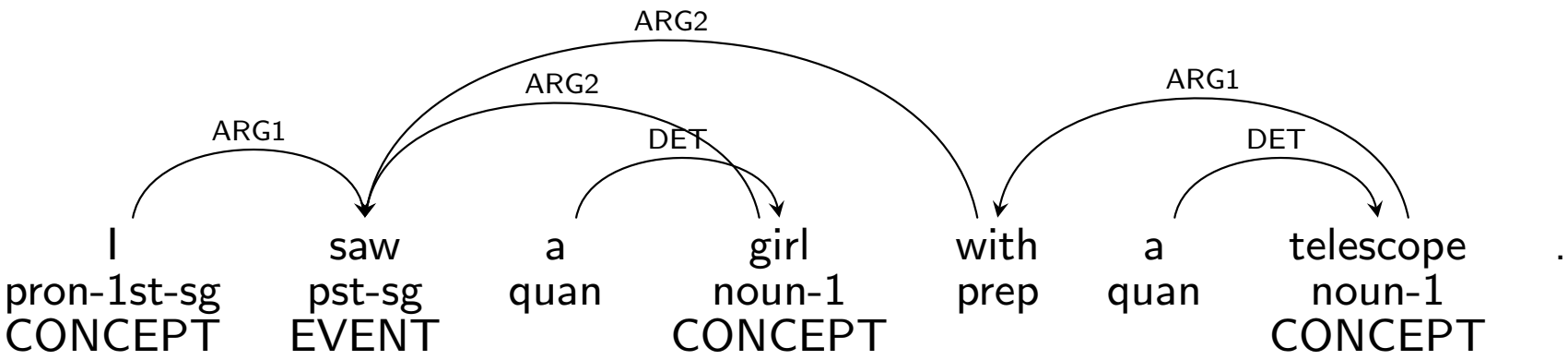
# Example (SBCG)

Word Construction

$$\begin{bmatrix} \text{see-1-le} & \\ \text{ORTH} & \begin{bmatrix} \text{Orthography} & \text{see} \\ \text{Form} & \text{saw} \end{bmatrix} \\ \text{SYN} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \text{ARG} & \begin{bmatrix} \text{PER} & \text{1st} \\ \text{NUM} & \text{sg} \\ \text{TENSE} & \text{pst} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{SEM} & \begin{bmatrix} \ldots \end{bmatrix} \\ \ldots & \end{bmatrix}$$

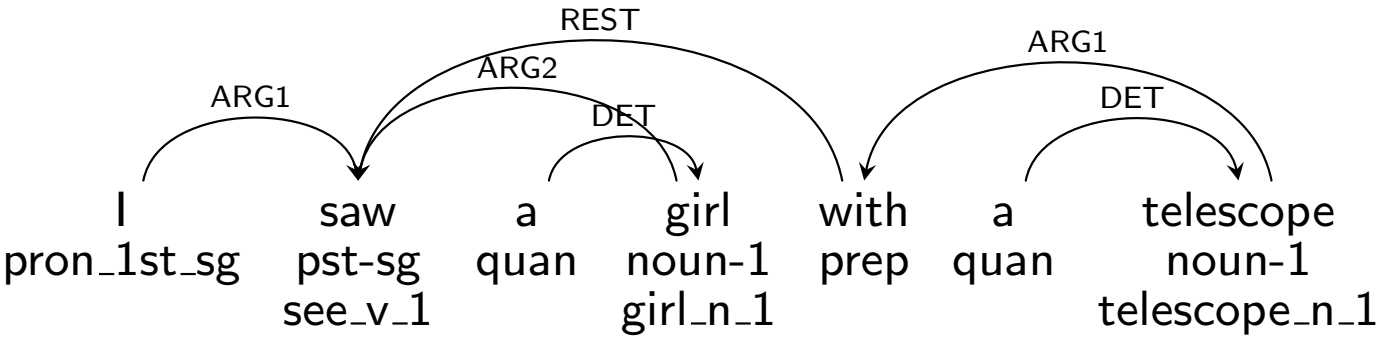
# Example (SBCG+ERG)

Structures with slots for sense information



# Example (SBCG+ERG+WordNet)

With sense information




| Sense | Synset ID | Meaning |
|---|---|---|
| girl_n_1 | 10084295-n | a youthful female person |
| telescope_n_1 | 04403638-n | a magnifier of images of distant objects |

# Discussion

## Main Contributions

- ERG doesn't provide frame semantics for many constructions in English. My final semantic analysis will be much richer than what ERG can give us now. This requires a lot of work.
- WordNet is rich, but it is not perfect. After I chose and studied my reference corpus, based on my analysis for the phenomena that I found, I will propose changes to WordNet.
- Linking the resources is non-trivial, many interesting issues of representation are still unresolved.
- Uniform Semantic Representation

# Table of Contents

# Word-Sense Disambiguation

## Main Ideas

LESK is a family of dictionary-based WSD algorithms which determine the correct senses of a word by calculating the score as intersection between its context (e.g. other words in the same sentence) and the gloss information of each sense available for that particular word.

## Simplified LESK

```
for each sense s_i of word w_j:
score_i = overlap(definition of s_i, context of w_j)
if score_i > max_score:
max_score = score_i
best_sense = s_i
```

## Example

```
word      = 'bank'
context = ['I', 'go', 'to', 'the', 'bank', 'to', 'withdraw'
           , 'money']
bank_n_1 = ['a', 'financial', 'institution', 'that', 'accepts',
           'deposits', 'and', 'channels', 'the', 'money',
           'into', 'lending', 'activities']
bank_n_2 = ['sloping', 'land', 'especially', 'the', 'slope',
           'beside', 'a', 'body', 'of', 'water']
score_1 = overlap(bank_n_1, context) # ['money', 'the']
score_2 = overlap(bank_n_2, context) # ['the']
# So the best sense for 'bank' is bank_n_1
```

# Extended Simplified LESK

## Main Ideas

Use hypernyms and hyponyms to extend token list.

```
bank_n_1 = ['a', 'financial', 'institution', 'that', 'accepts',
            'deposits', 'and', 'channels', 'the', 'money',
            'into', 'lending', 'activities']
hype(bank_n_1) = [ financial-institution_n_1 ]
hypo(bank_n_1) = [ agent-bank_n_1, state-bank_n_1 ]
financial-institution_n_1 = ['an', 'institution', 'public', 'or',
    'private', 'that', 'collects', 'funds', 'from', 'the',
    'public', 'or', 'other', 'institutions', 'and', 'invests',
    'them', 'in', 'financial', 'assets']
agent-bank_n_1 = ['a', 'bank', 'that', 'acts', 'as', 'an',
    'agent', 'for', 'a', 'foreign', 'bank']
state-bank_n_1 = ['a', 'bank', 'chartered', 'by', 'a', 'state',
    'rather', 'than', 'by', 'the', 'federal', 'government']
final_tokens = bank_n_1 + hype(bank_n_1) + hypo(bank_n_1)
```

# Le's Lesk

## Main Ideas

Use sense-disambiguated Princeton WordNet Gloss Corpus (PWGC) to extend token list for Extended Simplified LESK.

# PWGC Example

| Word | | Synset ID | Annotated Gloss |
|---|---|---|---|
| a financial institution | → | 08054721-n | an institution (public or private) ... |
| that accepts | → | 02236124-v | receive willingly something ... |
| deposits and channels | → | 01933900-v | direct the flow of |
| the money | → | 13384557-n | the most common medium of exchange ... |
| into lending activities | → | 02324182-v | give temporarily ... |

# Evaluation

## MFS vs Le's Lesk

Most-Frequent Sense (MFS) is the baseline for Word-Sense Disambiguation. Word-senses are provided based on sense-count learnt from an annotated corpus. It's very difficult to beat.

| Information | Instances | | Types | |
|---|---|---|---|---|
| | Le's Lesk | MFS | Le's Lesk | MFS |
| Correct sense ranked the first | 2544 | 2808 | 1143 | 1042 |
| Correct sense ranked the 2nd or 3rd | 1328 | 1203 | 844 | 682 |
| Wrong | 1325 | 1186 | 868 | 756 |
| No Sense | 122 | 122 | 54 | 54 |
| Total Sense | 5319 | 5319 | N/A | |

# Table of Contents

# Research Plans

- Study the phenomena in an English text from our corpus and provide linguistic analysis for all sentences
- Design the actual integrated semantic representation and implement the computational model
- Evaluate my analysis against a human annotated gold profile (most likely will be NTU-MC)

# Research Timeline

| Year | Quarter | Tasks |
|---|---|---|
| 2015 | 3 | Study English constructions in NTU-MC |
| | 4 | Provide analysis for documented constructions. Write the chapter on integrated semantic framework. |
| 2016 | 1 | Uniform semantic representation & write the chapter. |
| | 2 & 3 | Fine-tune the system performance |
| | 4 | New way to do WSD (based on constructions, concepts, etc.) |
| 2017 | 1 & 2 | Final design & software. Add more constructions, lexical items and idioms to the knowledge base. Experiment with Machine learning techniques. |
| | 3 & 4 | Complete thesis. (Optional) Experiment with other languages (Japanese, Chinese, Indonesian and Vietnamese). |

# Integrated Semantic Framework

Thank you