(Starting) Deep Grammar Development for Mandarin Chinese

Yi Zhang

yzhang@coli.uni-sb.de
Computational Linguistics
Saarland University, Saarbrücken

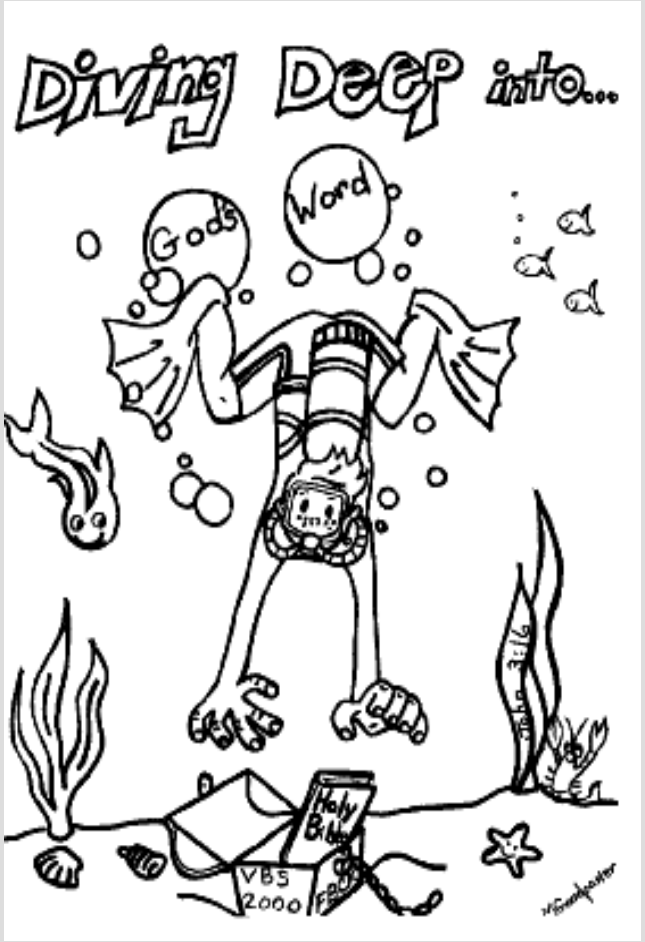
Outline

- Introduction & Motivation (Survey)
- Chinese Syntax
- Semantics with MRS
- Conclusion & Future Work

Introduction & Motivation

Objective

- To develop a deep linguistic HPSG resource grammar for Mandarin Chinese, to ...
  - Fill in a gap in Chinese deep processing;
  - Testify the applicability of HPSG formalism to Chinese;
  - For application purpose.

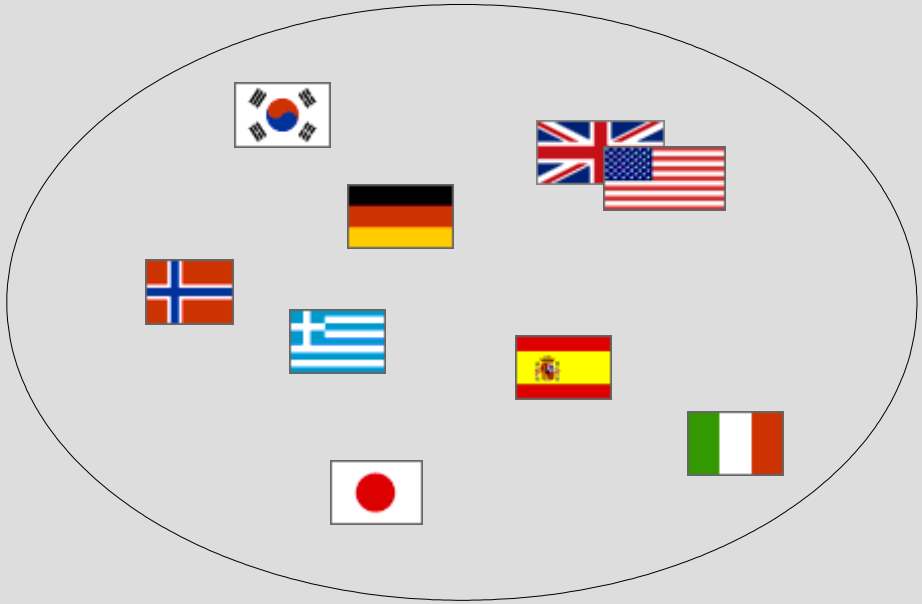


Situation

- Very few reported systematic deep grammar development for Chinese
- Local linguistic theories are nice, though not formalized
- HPSG is NOT adopted by most of Chinese linguists (for some or other reasons).
  - *“... Just as you have mentioned, researchers in mainland China don't show much interest on HPSG. They(We) know "a little" about HPSG but can not understand it thoroughly. I think it's a great pity for CL in China. ... ”*

What Follows

- Chinese see themselves outside the international linguistics community.



What Follows

- Deep processing of Chinese is far lagging behind.
- Linguistic theories without formalism are not able to help the development of application.
- Cross-lingual application becomes extremely difficult, if not impossible.

Motivation

- There are matured systems for grammar engineering and efficient deep processing (LKB, PET, [incr tsdb ()], ...).
- Large scale deep grammar engineering has been carried out for a lot of languages.
- The experience gain from large scale grammar development enables quick starting of new grammar development(LinGO Grammar Matrix).

Motivation

- With a deep grammar, we can:
  - Parsing
  - Generation
  - Semantic analysis together with syntax
  - Treebanking
  - ... ...

Theoretical Framework

- Syntactic theory for Chinese (Zhu, 1982) & (Zhu 1985).
  - Pure syntax
  - Phrase based analysis
- HPSG (Pollard & Sag, 1994)
  - Typed Feature Structure
  - Unification based
  - Constraint based
  - Lexicalist
- MRS (Copestake et al., 1999) & (Copestake et al., 2001)

Platform & Resource

- LKB System
- LinGO Matrix Grammar (version 0.6).
- [incr tsdb()]
- Lexicon: ``*The grammatical knowledge-base of contemporary Chinese*'', ICL of PKU. Public edition with about 10,000 word entries.

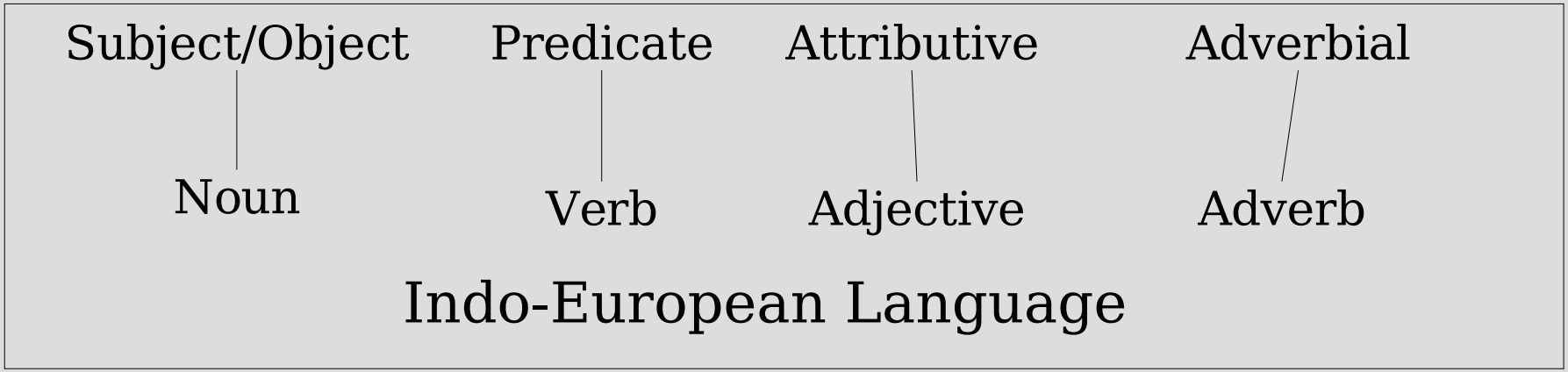
Chinese Syntax

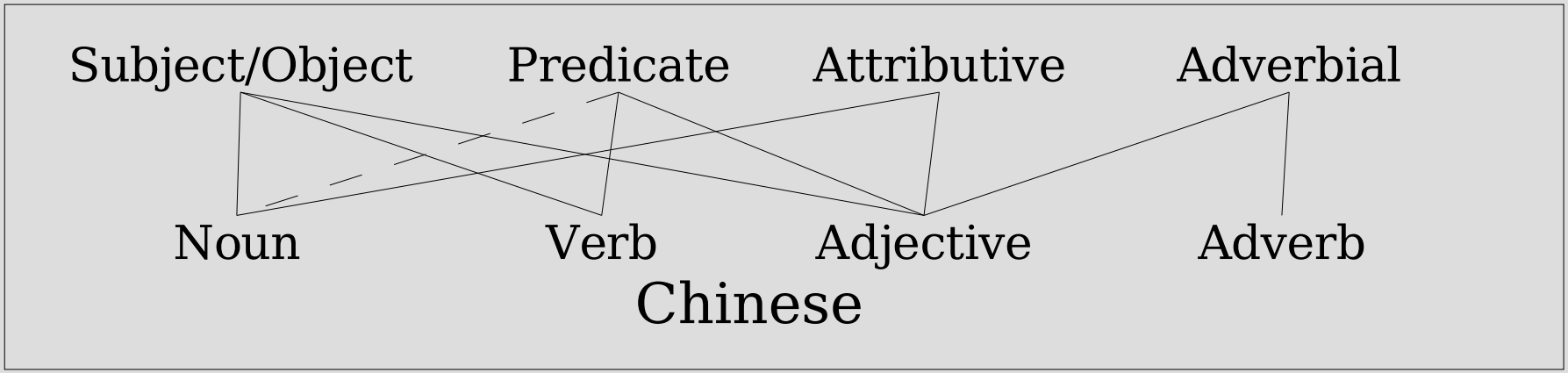
Phenomena

- No morphology
  - ta kai che.
    he drive car
    `He drives a car.'
  - he conglai mei kai guo che.
    he always not drive ASP car
    `He has never driven a car.'
  - kai che bu rongyi.
    drive car not easy
    `Driving a car is not easy.'
  - ta xihuan kai che.
    he love drive car
    `He likes to drive the car.'
- More complex syntax

Phenomena

- Complex relation between syntax units and word categories





Phenomena

- 0~N verbs in a sentence
  - zhe ge ren piqi hao.
    this CL person temper good
    `This person has good temper.'
  - wo kan bao.
    I read newspaper
    `I am reading the newspaper.'
  - wo mai bao kan.
    I buy newspaper read
    `I bought the newspaper and read.'
  - wo xiang mai bao kan.
    I want buy newspaper read
    `I want to buy some newspaper to read.'
  - wo xiang qu mai bao kan.
    I want go buy newspaper read
    `I want to go to buy some newspaper to read.'

Approach

- (Zhu, 1982) & (Zhu, 1985) provided a thorough and consistent analysis of Chinese syntax, though not formalized.
- Settling the syntax theory in HPSG framework is a good choice.

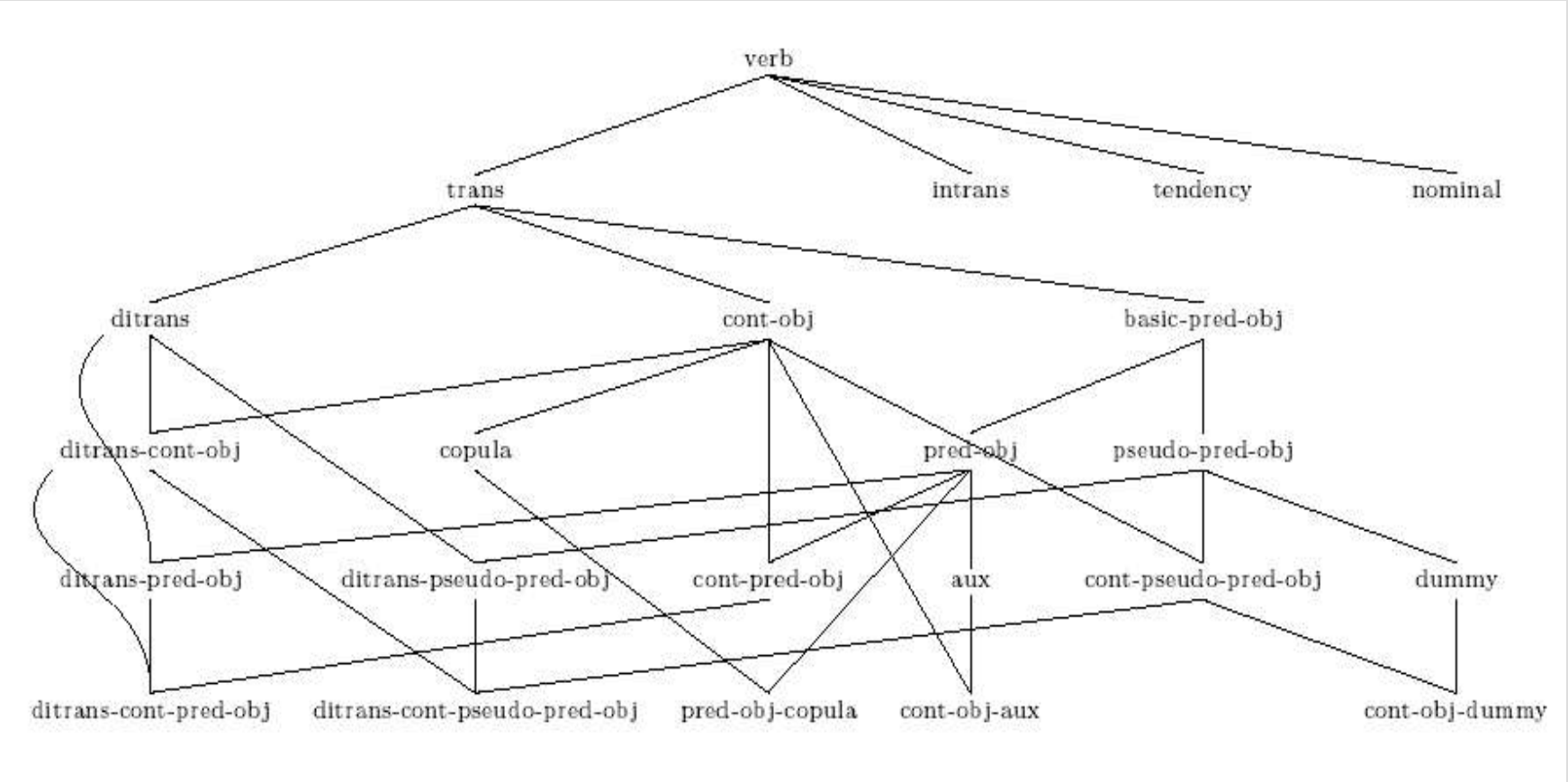
Basic Word Categories

basic word					extra word		
substantive			functional		small unit	large unit	
content	predicate						
noun temporal spacial direction number classifier pronoun	verb adjective situation pronoun	diff. adverb	preposition conjunction auxiliary modal	onoma. excl.	prefix suffix morpheme non-morph.	idiom locution abbre.	punc.

(Zhu, 1982) & (Yu, et al. 1998)

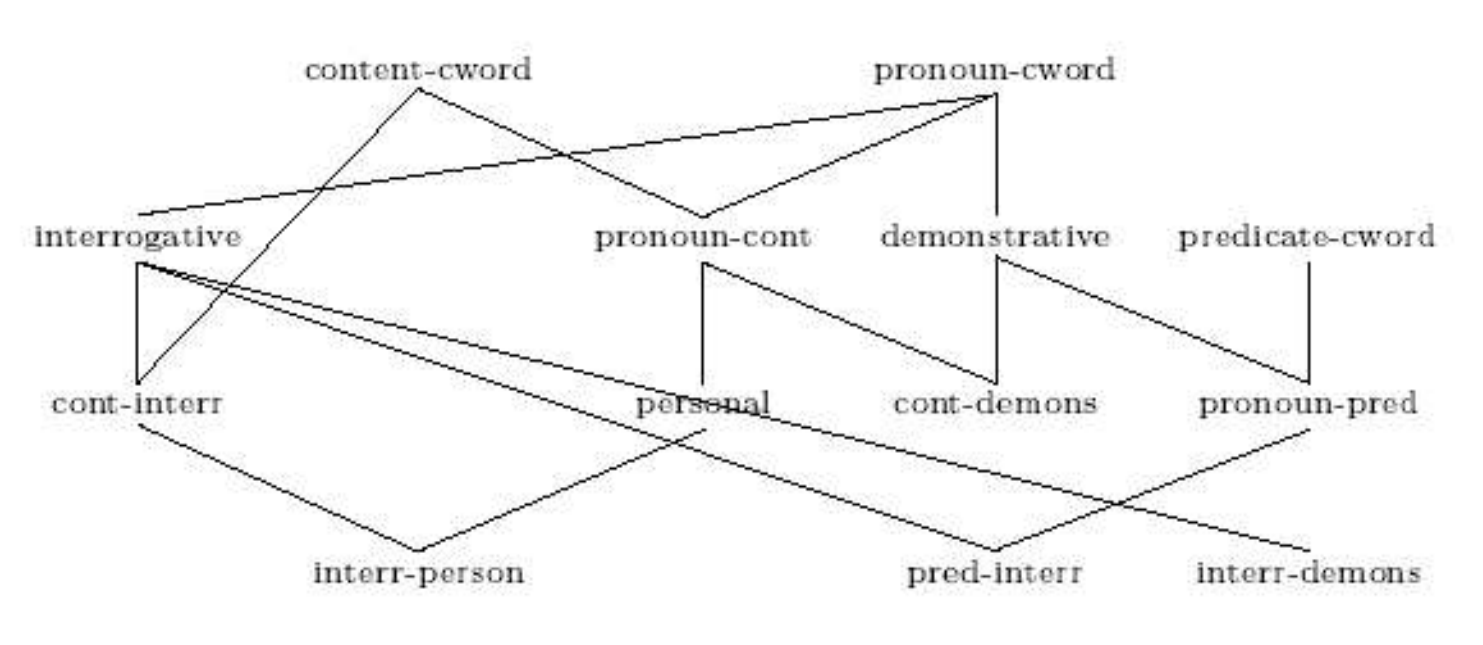
Lexical Types

- Verb



Lexical Types

- Pronoun



Lexical Types

- Classifier
  - cl-unit-cword: unit classifier
  - cl-mass-cword: massive classifier
  - cl-meas-cword: measurement classifier
  - cl-volm-cword: volume classifier
  - cl-type-cword: type classifier
  - cl-shape-cword: shape classifier
  - cl-undet-cword: undetermined classifier
  - cl-vq-cword: verbal quantity classifier
  - cl-tq-cword: temporal quantity classifier

HEAD Feature

- For orthogonal features, rather than creating subtypes, I used features in ***SYNSEM.LOCAL.CAT.HEAD***.

	ZAI-V	V-ZHE	V-LE
děngdài (wait)	+	+	+
jìdù (envy)	+	+	−
rùchǎng (enter)	+	−	+
xiězuò (write)	+	−	−
xiàngzhēng (resemble)	−	+	+
kěwàng (desire)	−	+	−
dàodá (arrive)	−	−	+
qǐfú (fluctuate)	−	−	−

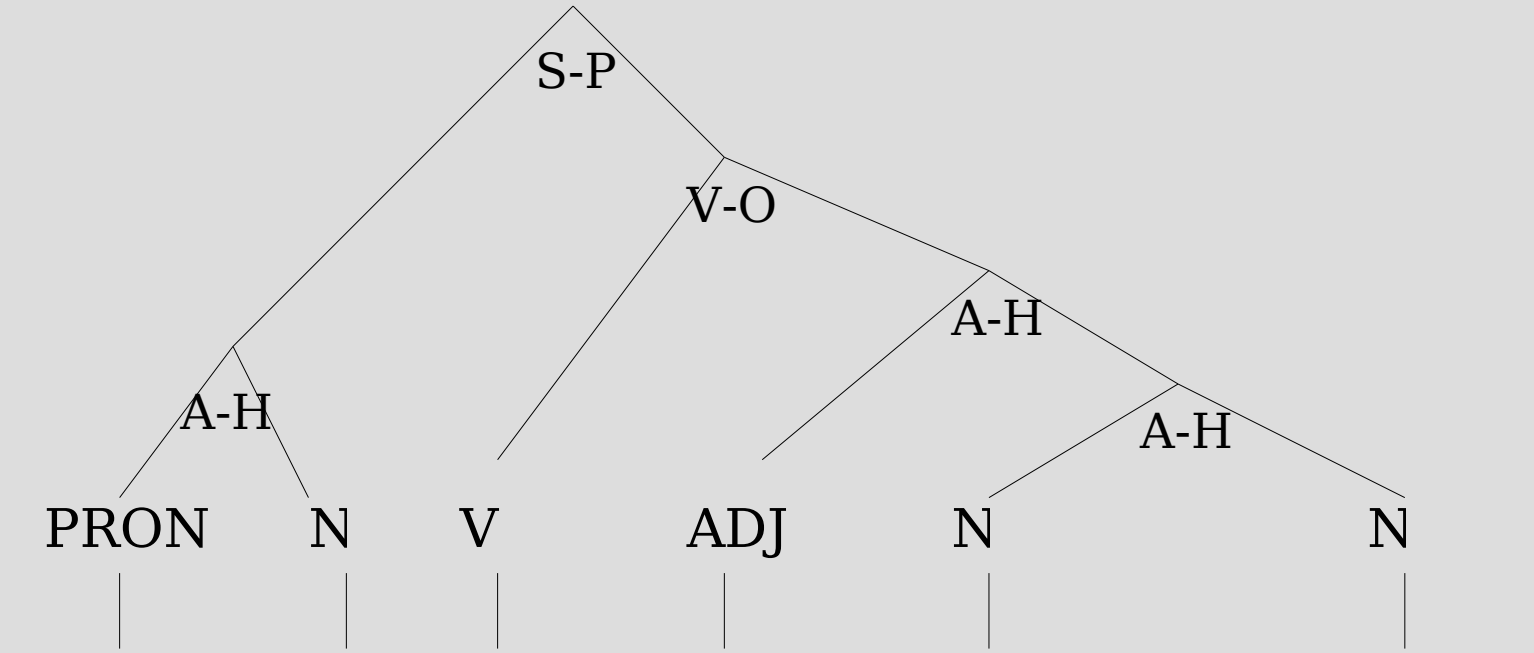
Valence **Feature**

- c-valence := valence &
  [SUBJ list, <-- subject
  OBJS list, <-- real objects
  POBJS list, <-- pseudo objects
  CCOMP list, <-- ``complement''
  SPR list]. <-- specifiers
- Corresponding schemata
  - head-subj-phrase
  - head-obj-phrase
  - head-pobj-phrase
  - head-comp-phrase
  - head-spec-phrase

Phrase Structure Rule Types

- Subject-Predicate
- Verbal-Object
- *Verbal-Complement(Post-verb modifier)*
- Adjunct-Head
  - Adjunct-Content
  - Adjunct-Predicate
- *Serial Verb*
- *Pivotal*

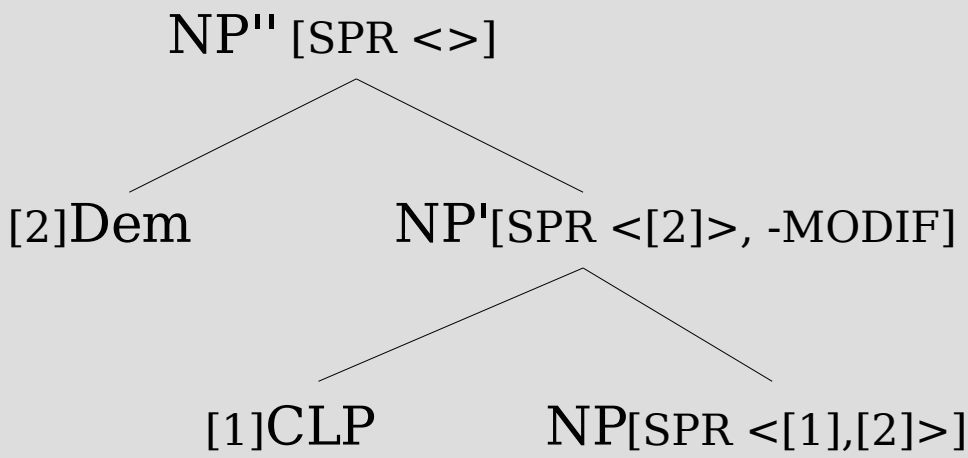
An Example



`There are a lot of foreign students in our class.'

Nominal Phrases

- Double-specifier account for Chinese NP (Say Kiat Ng, 1997)
- Some modifications to allow “Dem + Noun” construction.



Semantics with MRS

MRS Basic

- Minimum Recursion Semantics (Copestake et al., 1999) & (Copestake et al., 2001)
  - Flat semantic representation
  - Elementary Predication (EP)
    - a handle
    - a relation
    - a list of variable arguments
    - a list of scope arguments
  - Top handle
  - Constraints on scope relations (qeq condition)

MRS Basic

- the dog sleeps
  - <h0, <h1:_det(x,h2,h3), h4:dog(x),h5:sleep(e,x)>, {h0 qeq h5, h2 qeq h4}>
  - the(x,dog(x),sleep(e,x))

MRS Basic

- every dog probably chases some white cat
  - <h0, {h1:every(x,h2,h3),h4:dog(x),h5:probably(h6),h7:chase(x,y),h8:some (y,h9,h10),h11:white(y),h11:cat(y)},{h0 qeq h5, h2 qeq h4, h6 qeq h7, h9 qeq h11}>
  - probably(every(x, dog(x), some(y, white(y) ^ cat(y), chase(x, y))))
    every(x, dog(x), probably(some(y, white(y) ^ cat(y), chase(x, y))))
    every(x, dog(x), some(y, white(y) ^ cat(y), probably(chase(x, y))))
    probably(some(y, white(y) ^ cat(y), every(x, dog(x), chase(x, y))))
    some(y, white(y) ^ cat(y), probably(every(x, dog(x), chase(x, y))))
    some(y, white(y) ^ cat(y), every(x, dog(x), probably(chase(x, y))))

Problems with Chinese

- The syntax theory of (Zhu, 1982) & (Zhu, 1985) doesn't count for semantics. Semantic composition would be more difficult.

Problems with Chinese

- Subject vs. ARG1
  - women qu beijing.
    we go Beijing
    We go to Beijing.
    <h0,{h1:women_p(x1),h2:qu_v(e,x1,x2),h3:beijing_n(x2)},{h0 qeq h2}>
  - mingtian qu beijing.
    tomorrow go Beijing
    Somebody will go to Beijing tomorrow.
    <h0,{h1:mingtian_t(e),h2:qu_v(e,x1,x2),h3:beijing_n(x2)},{h0 qeq h2}>

Solution

- Further subcategorizing phrase structure types.
- Argument binding both in lexicon and in construction.

```
sp-pron-pred-phrase := subj-pred-phrase & head-subj-phrase &
                       [ NON-HEAD-DTR pronoun-cont-cword ].
sp-tempo-pred-phrase := subj-pred-phrase &
                       [ SYNSEM.LOCAL.CAT.VAL #val,
                         HEAD-DTR.SYNSEM.LOCAL  [  CAT.VAL #val,
                                                   CONT.HOOK.INDEX #event ],
                         NON-HEAD-DTR temporal-cword &
                                                [ SYNSEM.LOCAL.CONT.HOOK.INDEX #event ] ].
```

Conclusion & Future Work

Conclusion

- Syntax:
  - Basic word categories and phrase structure rules implemented.
- Semantics:
  - Semantics composition for basic phrase structures implemented.

Statistics

- Starting day: May 10th, 2004
- Lexical Types: 108
- Phrase Structure Rules: 43
  - Unary Rules: 5
  - Binary Rules: 38
- Lexicon: 10,069 entries
  - Noun: 3571
  - Verb: 2094
  - Adjective: 1471
  - Adverb: 719
  - Idiom: 552
- Lines of Grammar: 2,100 (excluding Matrix & lexicon entries).

Remaining Work

- Serial verb phrase
- Pivotal phrase
- Coordination phrase
- Other special constructions, including “ba” (disposal) construction and “bei” (passive) construction.

Remaining Work

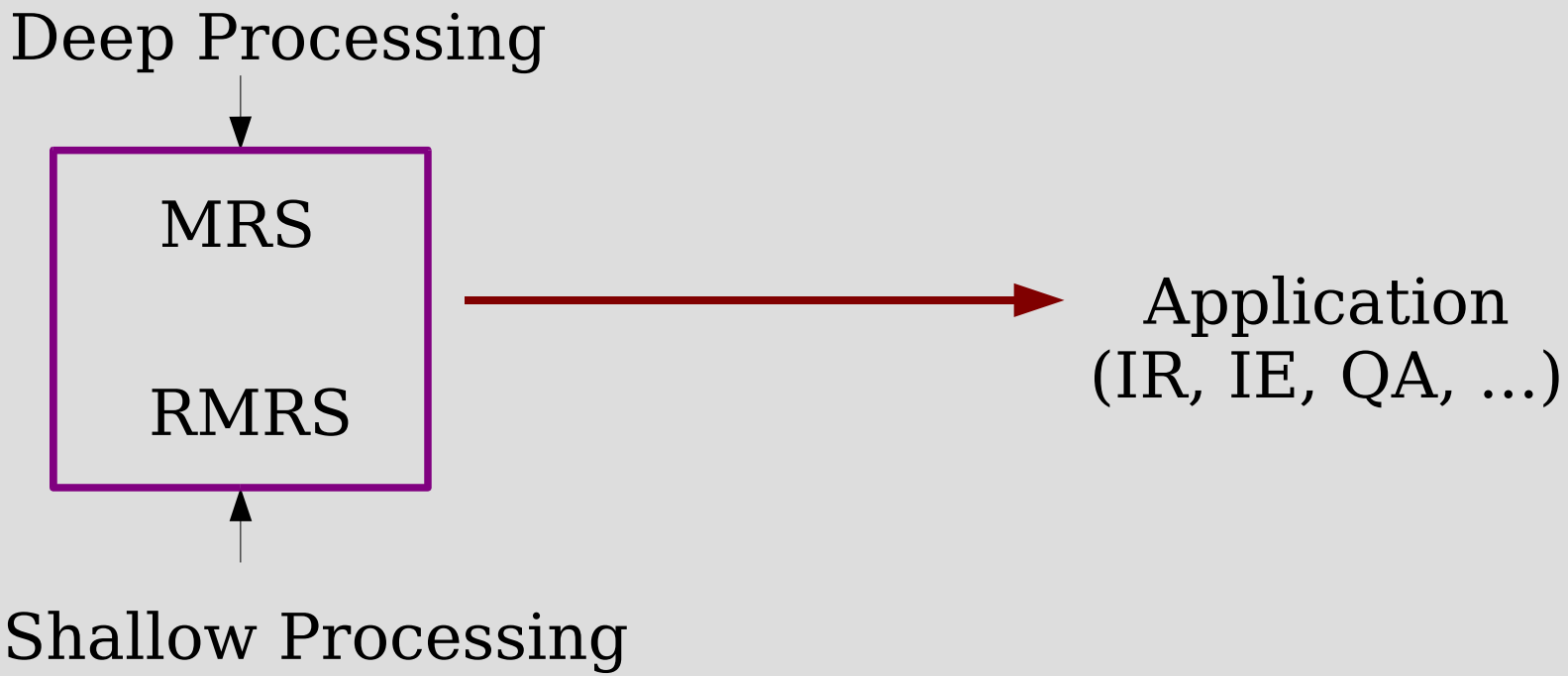
- A larger test corpus.
- More comprehensive evaluation of grammar coverage.

Beyond Grammar Engineering

- Problem with Deep Processing
  - Efficiency
    - Much larger search space than shallow methods
  - Robustness
    - Heavily depends on grammar coverage
  - Ambiguity & Specificity
    - Too many analysis results

Beyond Grammar Engineering

- Combination of shallow and deep processing



Thank you!