

Korean/English MT Project, and Korean/English Parallel Treebanks

<http://www.cis.upenn.edu/~xtag/koreantag/>

Chung-hye Han

Feb. 6, 2001

TIDES SITE VISIT

Outline of the Talk

- Linguistic issues.
- System overview;
Deep Syntactic Structure (DSyntS).
- Parser output conversion.
- Handling structural divergences.
- Dropped argument recovery.
- Treebank.
- Future work.

Linguistic Issues in Korean/English MT

Word Order

SOURCE: chuka kongkwupmul-eul 103 ceonwiciweontaetae-eke saryeongpu-ka ponayssta.

GLOSS: additional supply-Acc 103rd forward support battalion-Dat headquarters-Nom sent

OUTPUT: Headquarters sent an additional supply to a 103rd forward support battalion

Linguistic Issues in Korean/English MT

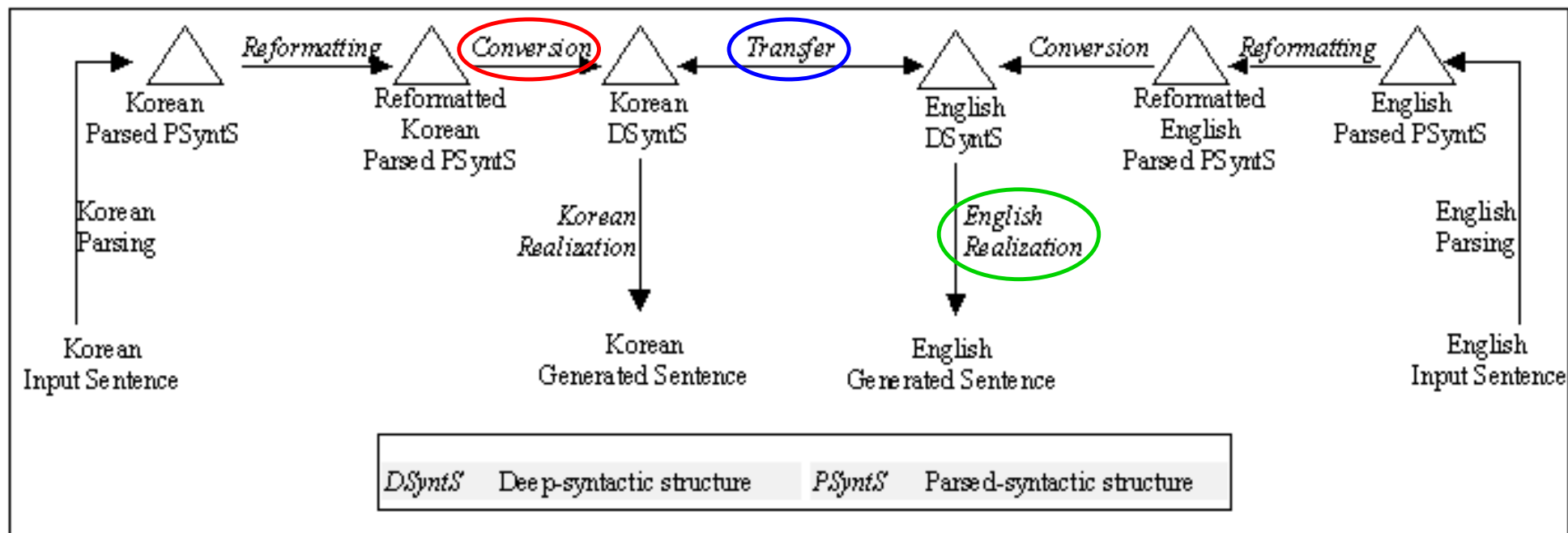
Dropped arguments and Morphology

SOURCE: IBP hwail-eul keomsaekhaci moshhaess-tamyeon cikeum tasi ponaekessta.

GLOSS: IBP file-Acc retrieve could_not-if now again will_send

OUTPUT: If one can not retrieve an IBP file, one will send it again now.

Overview of the System

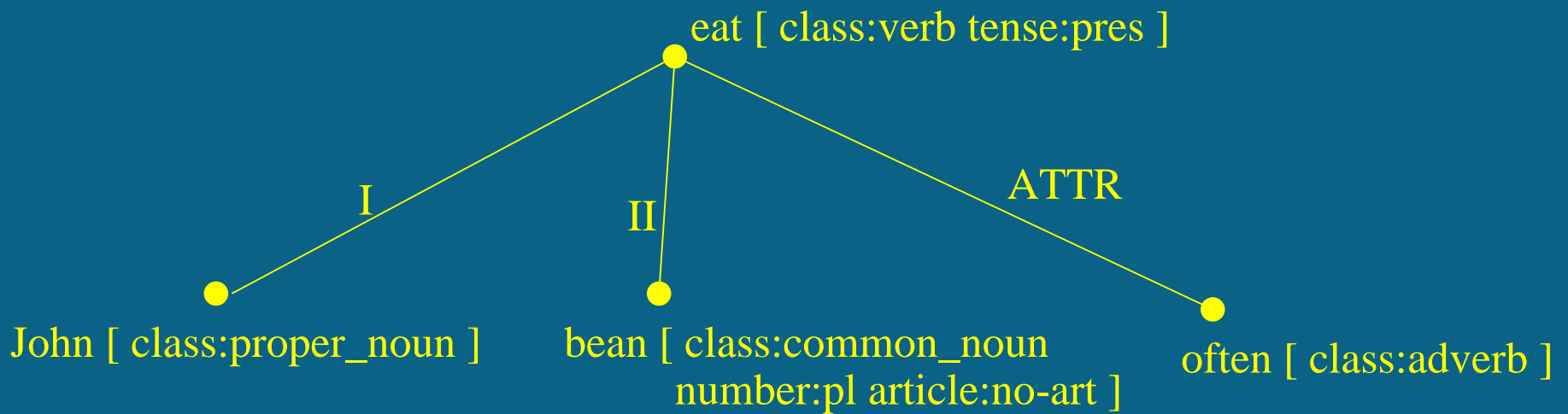
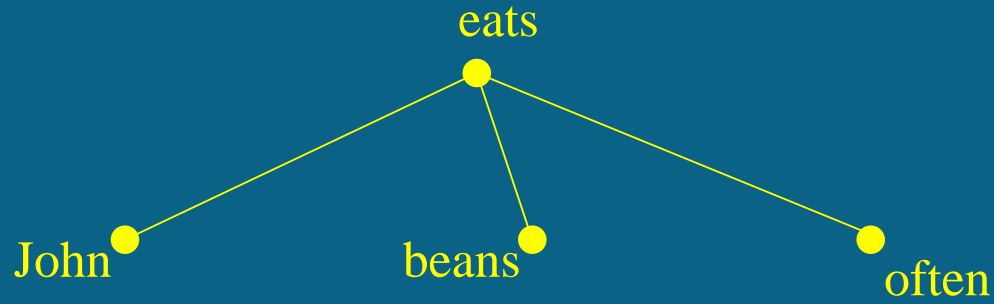


Deep Syntactic Structure

- Enriched Dependency structure.
- Directed arcs with dependency relation labels: I, II, III, ATTR.
- Grammatical information is represented as features on the node labels.
- Well suited to MT:

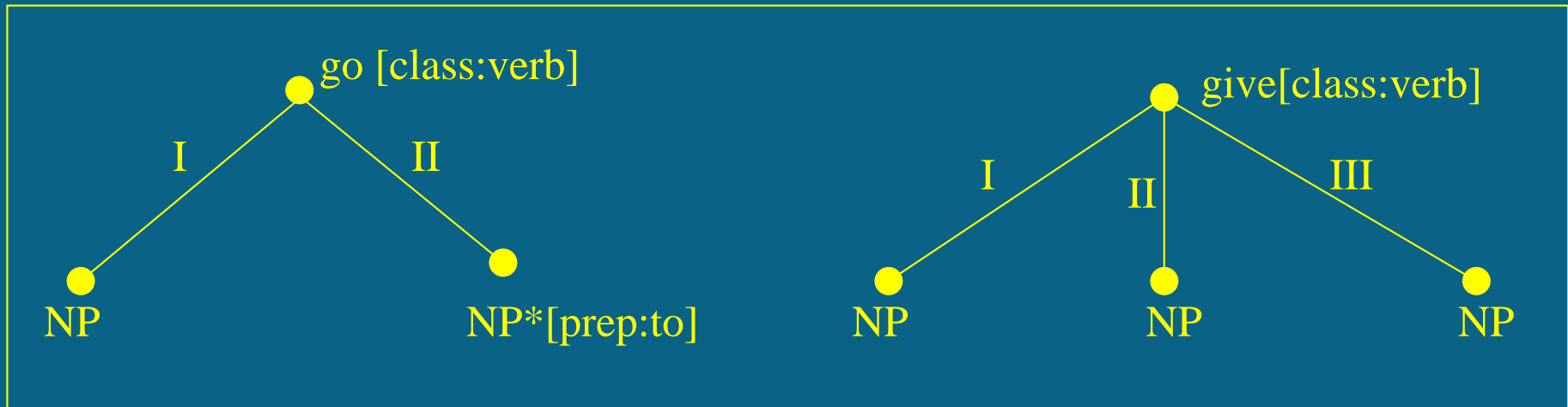
Abstracts away from superficial grammatical differences between languages, such as linear order and the usage of function words.

'John often eats beans.'



Predicate-Argument Lexicon: English

Subcategorization information for verbs and adjectives.



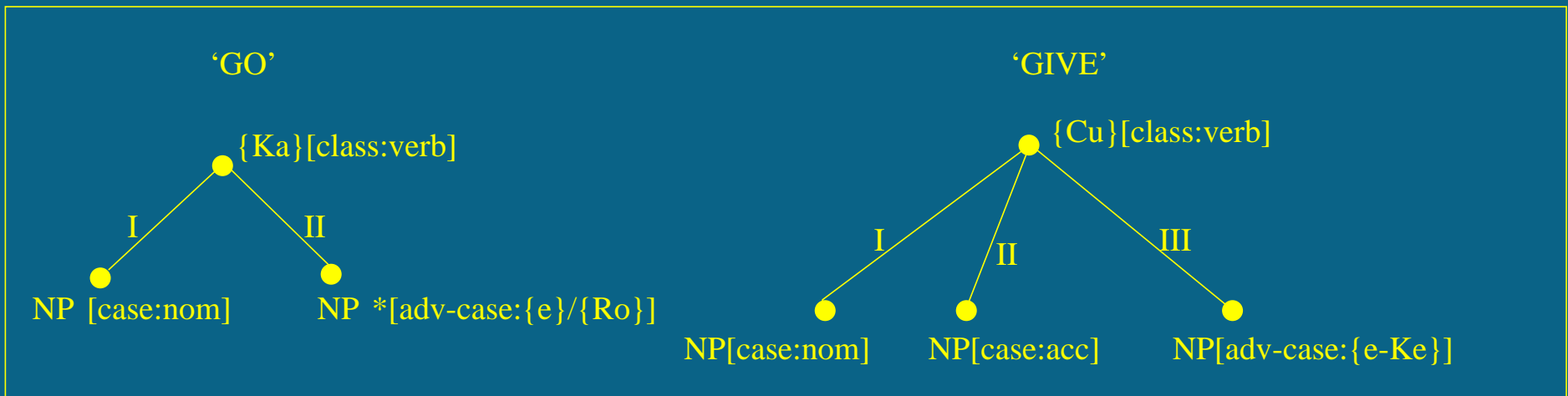
Critical for recovery of dropped arguments!!!

Predicate-Argument Lexicon: Korean

Arguments are listed with case or adverbial postpositions.

case postpositions: nominative, accusative.

adverbial postpositions: {e-Ke} ('to'), {Ro} ('to'), {e-Seo} ('from').



Critical for conversion!!!

Conversion

Generic dependency structure (Yoon et. al. 1997) \Rightarrow MTT-based DSyntS

STEP 1: Rewriting feature labels.

STEP 2: Making dependency relationships more explicit.

Korean predicate-argument lexicon is used as a guide.

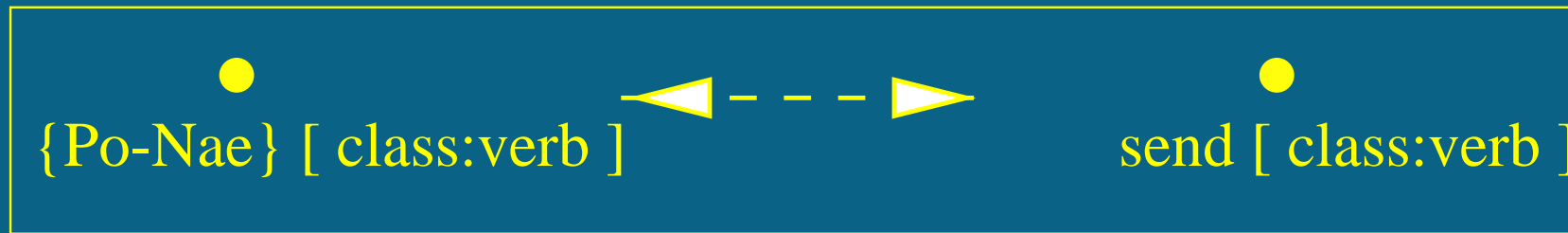
STEP 3: Promoting features to lexemes and vice versa.

Transfer

- Based on DSyntS grammars that are independently motivated by source and target languages.
- Map source DSyntS subtrees to target DSyntS subtrees.
- Use of variables allows generalization of rule application.
- Features on DSyntS nodes constrain rule application.

Transfer

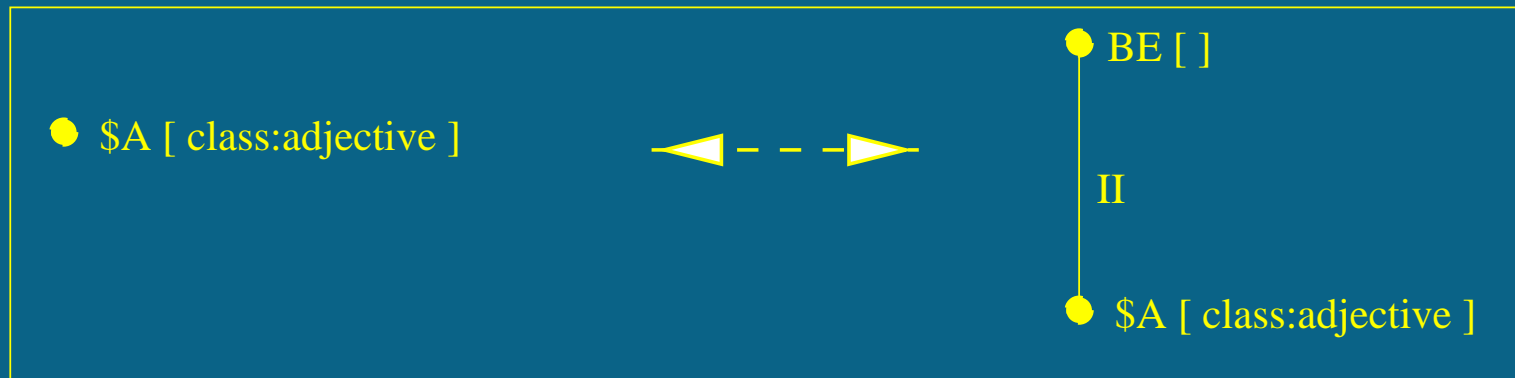
- Simplest case: The related subtrees are reduced to a single node.



- Structural divergence is represented in the transfer lexicon by including contextual information in the related subtrees.

Multi-word Transfer

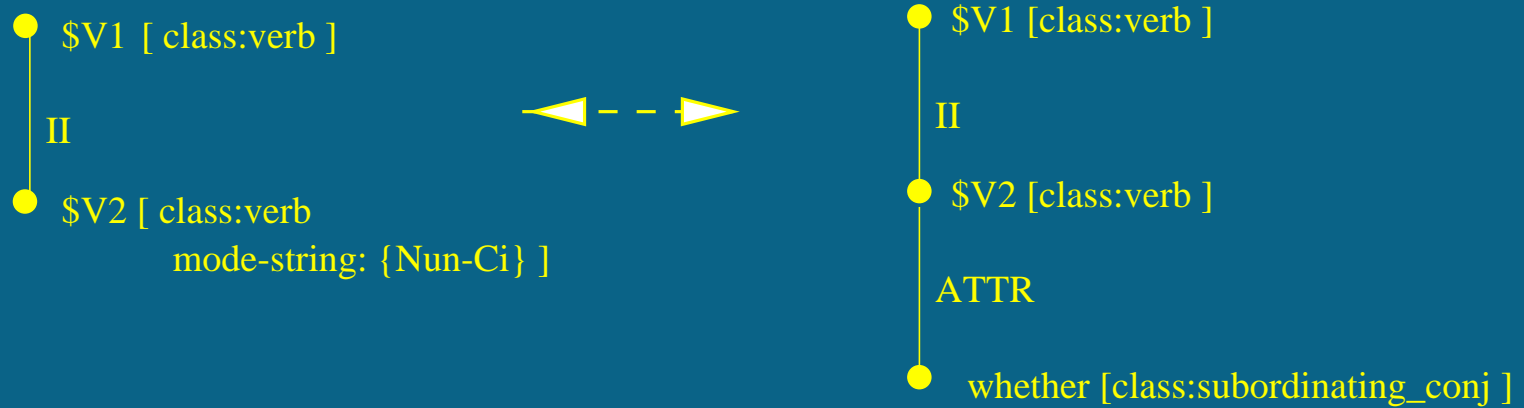
Transfer of predicative adjectives



Transfer of modificational adjectives

`$A [class:adjective mode-string:{N}%{eun}]`  `$A [class:adjective]`

Transfer from Inflection to a Lexeme

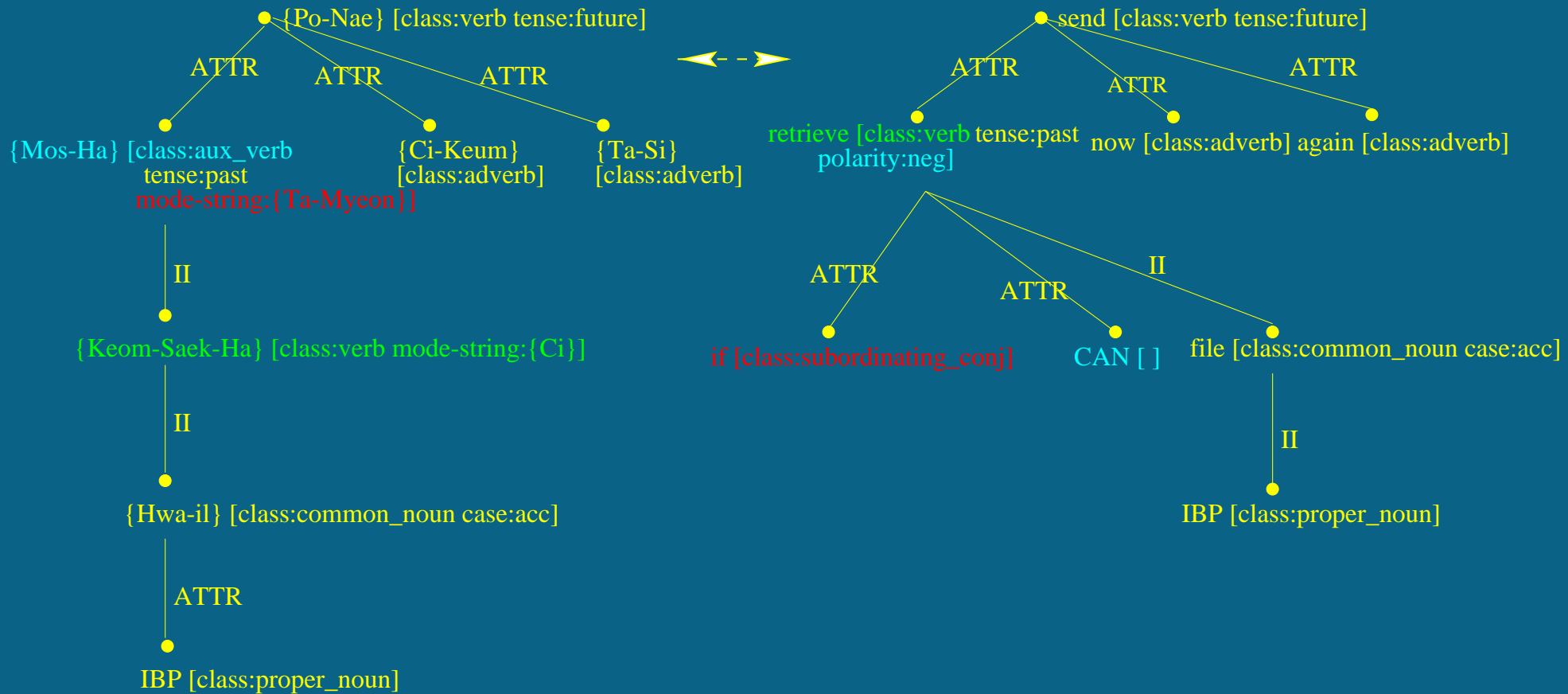


Source: {Pyeon-Sok} {Cang-Chi-e-To} {Ca-Cu} {Mun-Ce-Ka}) {iss-Nun-Ci} {Mwul-eoss-Seup-Ni-Ta}.

Gloss: transmission device-at-also frequently problem-Nom eixt-Comp ask-Past-Decl

Output: One also asked whether there is a frequent problem in a transmission.

Transfer from Korean DSyntS to English DSyntS



Argument Recovery

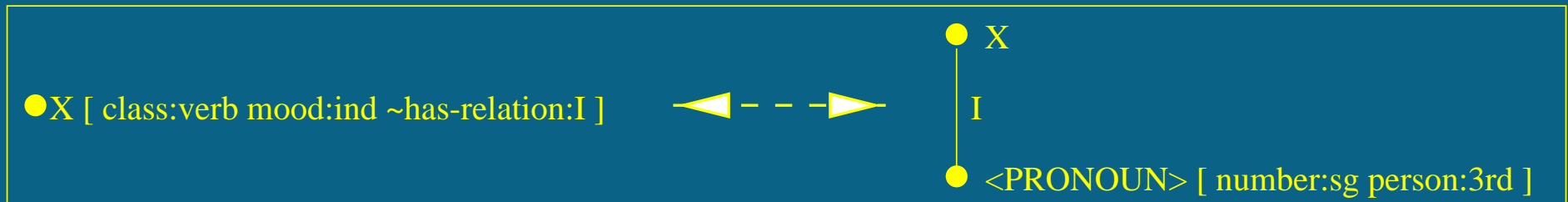
- Dropped arguments must be recovered in order to obtain grammatical English sentences.
- Add default pronouns for missing arguments using grammatical and lexical knowledge.

English predicate-argument lexicon is critical.

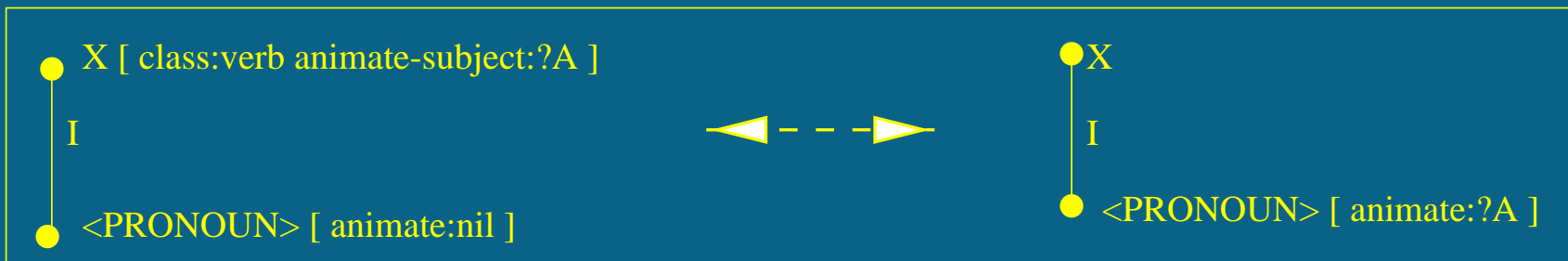
- This is performed just before English realization, by preprocessing the English DSyntS obtained from transfer.

Argument Recovery Rules

Insertion of Missing Actant I:

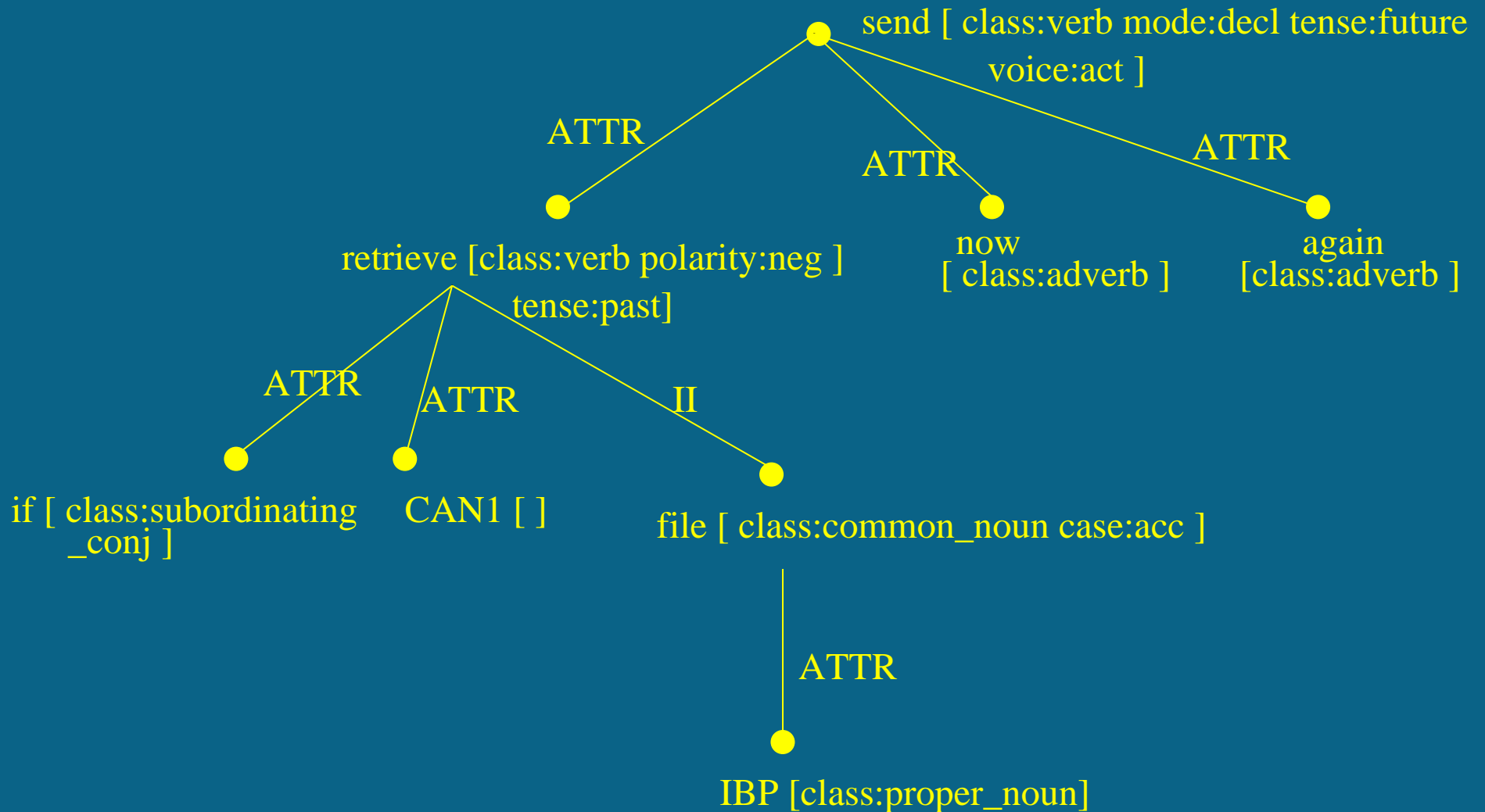


Determining whether pronouns are animate or not:



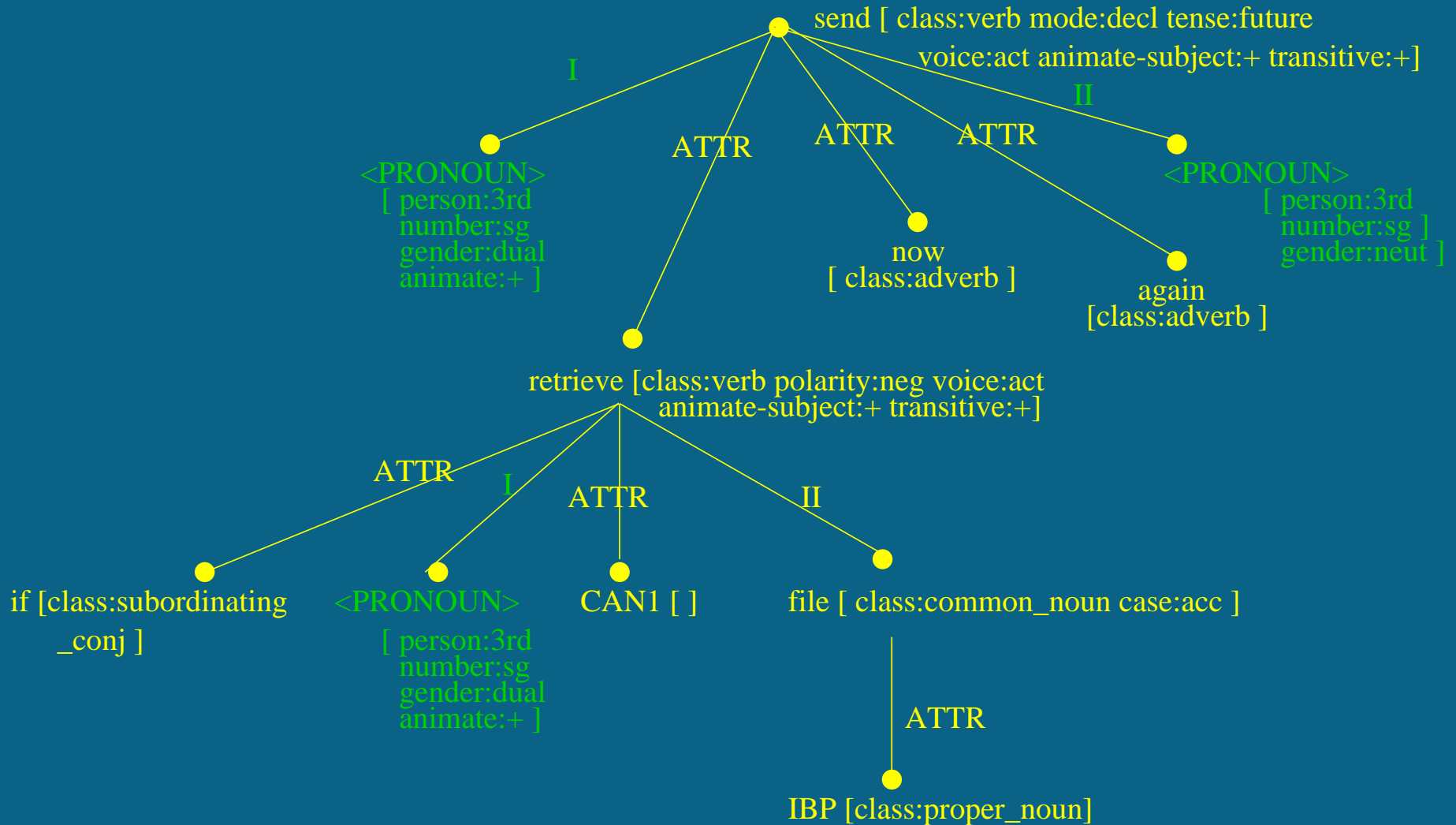
Before Argument Recovery

'If (NP1) could not retrieve IBP file, (NP2) will send (NP3) again now.'



After Argument Recovery

'If one cannot retrieve an IBP file, one will send it again now.'



Current Status

- Parallel corpus: military language training manual
50,000 word tokens, 3800 word types, 5000 sentences.
- Predicate-argument lexicon
Korean: 1000 entries, English: 19,000 entries.
- Transfer lexicon
4000 entries (+24,187).
- Grammatical analysis
simple clause (declaratives, imperatives, interrogatives), complex clause (subordination, coordination), scrambling, empty argument, adjective phrase, noun phrase (compound nouns, NP modifiers, relative clauses, complex noun phrases), verb phrase (auxiliary verbs, light verbs, compound verbs), negation, copular sentence, adverb modification, etc.

Korean Treebank

- Phrase structure annotation:
 - Head/phrasal node distinction;
 - Empty argument;
 - Argument/adjunct distinction;
 - Movement and trace.
- Purpose of the treebank:
 - Train and evaluate parsers;
 - Train supertagger: 80.42%;
 - Train part-of-speech tagger: 96.07%;
 - Extract LTAG grammar;
 - Annotated resource for other applications.

Korean Treebank

- Time Table

6/99 - 8/99: part-of-speech tagging and morphological analysis.

9/99 - 12/99: syntactic bracketing guidelines.

1/00 - 6/00: first-pass syntactic annotation.

7/00 - 8/00: updated the bracketing guidelines.

9/00 - 11/00: second-pass syntactic annotation.

12/00 - 1/01: third-pass syntactic annotation.

2/01: finishing up the bracketing guidelines.

- Quality of the Treebank

	Between annotators	Annotator 1	Annotator 2
Bracketing Recall	96.60	97.69	98.84
Bracketing Precision	97.97	98.89	98.84
Tagging Accuracy	99.72	99.99	99.77

Future Work

- Using the Korean-English parallel Treebank for automatic extraction of transfer rules.
- Using Explicit annotation of empty arguments in the Korean Treebank to incorporate a discourse model for a more principled recovery of implicit arguments.
- Train a Korean statistical parser using the Treebank.
- Parse additional 50K newswire corpora with a dependency parser, convert to phrase structure using LexTract, and retrain parser.
- Apply computational morphology techniques being developed by Mark Liberman's group and train a Korean morphological analyzer using the Treebank.