# Statistical Morphological Tagging and Parsing of Korean with an LTAG Grammar


Anoop Sarkar and Chung-hye Han

University of Pennsylvania
anoop@cis.upenn.edu

Simon Fraser University
chunghye@sfu.ca



TAG+ 6, May 2002 — Venice, Italy

## Overview

- Introduction to Supervised Statistical Parsing with LTAG
- LTAG grammar extracted from the Penn Korean Treebank
- Morphological Tagging: Motivation and Experiments
- Statistical parsing of Korean using a Morphological Tagger

## Parsing as a machine learning problem

- $S$ = a sentence
  $T$ = a parse tree
  A statistical parsing model defines $P(T \mid S)$
- Find best parse: $\arg\max_{T} P(T \mid S)$
- $P(T \mid S) = \frac{P(T,S)}{P(S)} = P(T,S)$
- Best parse: $\arg\max_{T} P(T,S)$
- e.g. for PCFGs: $P(T,S) = \prod_{i=1\ldots n} P(\mathrm{RHS}_i \mid \mathrm{LHS}_i)$

## Parsing as a machine learning problem

- Training data for English: the Penn WSJ Treebank (Marcus et al. 1993)
- Convert Treebank into LTAG derivations using LexTract (Xia 2001)
- Train statistical LTAG parser from these events
- Evaluate accuracy on test data
- A standard evaluation:
  Train on 40,000 sentences
  Test on 2,300 sentences

## Parsing as a machine learning problem

- Training data for Korean: the Penn Korean Treebank (Han et al. 2002)
- Train statistical morphological tagger and statistical LTAG parser
- Evaluate accuracy on test data
- Our evaluation:
  Train on 4,653 sentences (49,473 words)
  Test on 425 sentences (3,717 words)

## Statistical Parsing with Tree Adjoining Grammars

- Substitution: $\sum_{\alpha} P_s(t, \eta \rightarrow \alpha) = 1$

- Adjunction: $P_a(t, \eta \rightarrow \textsf{NA}) + \sum_{\beta} P_a(t, \eta \rightarrow \beta) = 1$

- Multiple adjunctions at a node (Schabes and Shieber 1994):

$$
\begin{aligned}
P_{la}(\tau, \eta \rightarrow \textsf{NA}_l) + \sum_{\tau'} P_{la}(\tau, \eta \rightarrow \tau') &= 1 \\
P_{ra}(\tau, \eta \rightarrow \textsf{NA}_r) + \sum_{\tau'} P_{ra}(\tau, \eta \rightarrow \tau') &= 1
\end{aligned}
$$

Statistical Parsing with Tree Adjoining Grammars

- Start of a derivation: $\sum_{\alpha} P_i(\alpha) = 1$

- Probability of a derivation:

$$
\begin{aligned}
Pr(\mathcal{D}, w_0 \ldots w_n) = \\
& P_i(\alpha, w_i) \times \prod_p P_s(\tau, \eta, w \rightarrow \alpha, w') \times \\
& \prod_q P_a(\tau, \eta, w \rightarrow \beta, w') \times \prod_r P_a(\tau, \eta, w \rightarrow \textsc{na})
\end{aligned}
$$

## Overview

- Introduction to Supervised Statistical Parsing with LTAG
- LTAG grammar extracted from the Penn Korean Treebank
- Morphological Tagging: Motivation and Experiments
- Statistical parsing of Korean using a Morphological Tagger

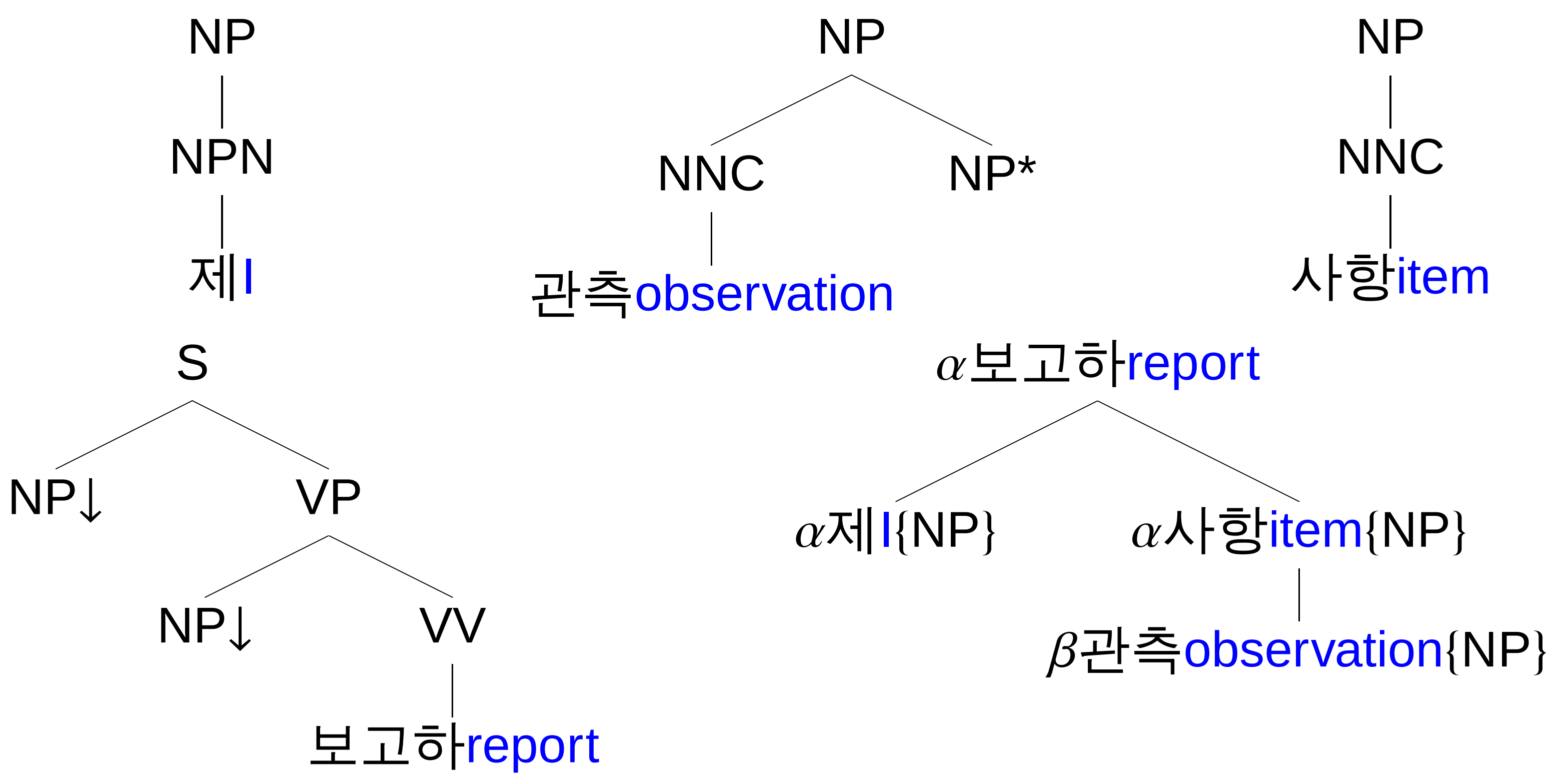
Korean Treebank

(S (NP-SBJ 제I/NPN+가nom/PCA)
    (VP (NP-OBJ 관측observation/NNC
                    사항item/NNC+을acc/PCA)
        보고하report/VV+었past/EPF+습니다decl/EFN)
    ./SFN)

→ I-Nom observation item-Acc report-Past-Decl

→ ‘I reported the overvation items.’

LTAG Grammar and Derivation Tree using LexTract (Xia 2001)



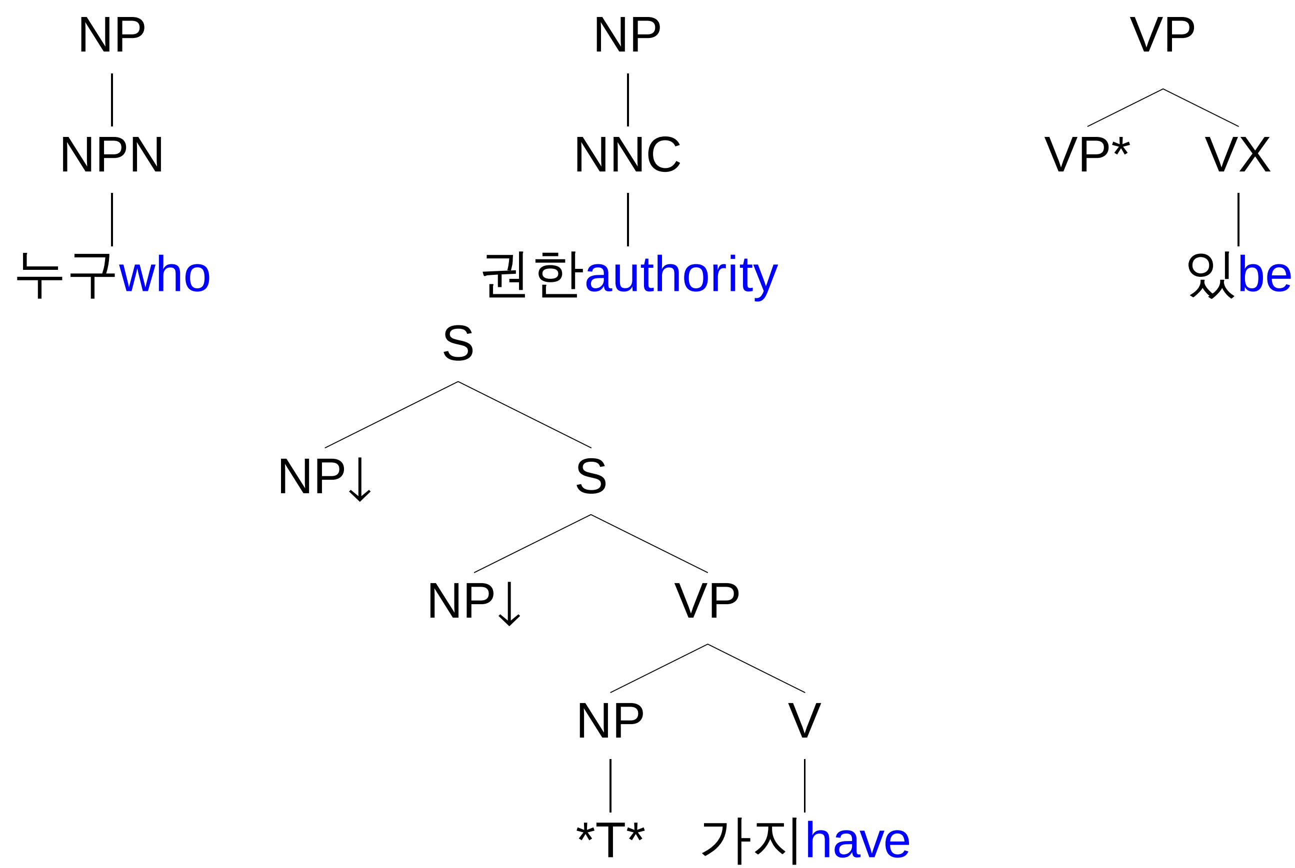
Korean Treebank

```
(S (NP-OBJ-1 권한authority/NNC+을acc/PCA)
    (S (NP-SBJ 누구who/NPN+가nom/PCA)
        (VP (VP (NP-OBJ *T*-1)
                    가지have/VV+고aux/EAU)
              있be/VX+지int/EFN))
    ?/SFN)
```

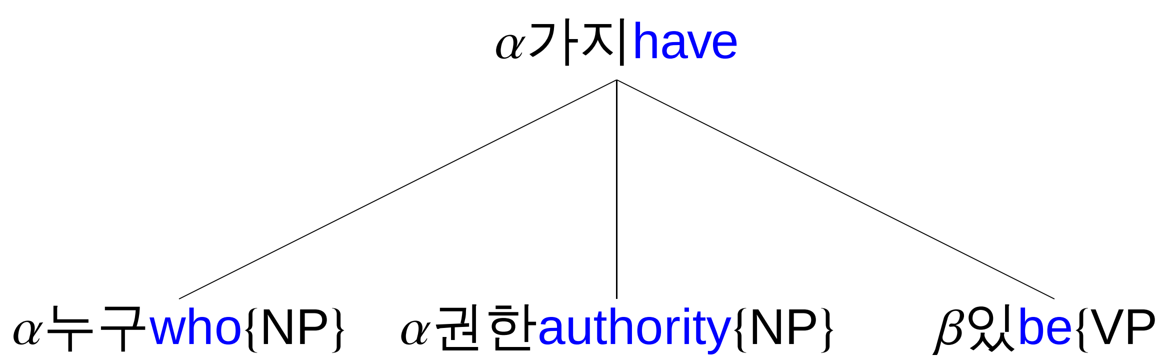
→ authority-Acc who-Nom have-AuxConnective be-Int

→ ‘Who has the authority?’

# LTAG Grammar for Korean using LexTract



LTAG Derivation Tree



## Overview

- Introduction to Supervised Statistical Parsing with LTAG
- LTAG grammar extracted from the Penn Korean Treebank
- Morphological Tagging: Motivation and Experiments
- Statistical parsing of Korean using a Morphological Tagger

## Motivation for Morphological Tagging

- Each substitution, adjunction is a relation between a pair of words
- Korean is an agglutinative language with a very productive inflectional system
- A fully inflected word seen in the training data will rarely occur in the unseen (test) data
- Sparse data problem is much worse than in English: the part-of-speech tags for inflected word forms are complex and can be novel in unseen data

## Motivation for Morphological Tagging

- The morphological tagger provides lemma splitting plus part-of-speech tagging
- Instead of multiplying ambiguity in the parser, we choose to implement a statistical morphological tagger
(provides a single-best analysis of the input sentence)
- Both lemma splitting and tagging are trained using the Penn Korean Treebank (same training/test split as in the parser)
- Lexical stem and suffix information as well as part-of-speech information from the morphological tagger is used in the statistical parser

Example input and output from the morphological tagging phase

**Input:** 제가␣관측␣사항을␣보고했습니다␣.

**Output:** 제/NPN+가/PCA␣관측/NNC␣사항/NNC+을/PCA␣
보고하/VV+었/EPF+습니다/EFN ␣./SFN

The part-of-speech tags for inflected word forms are complex and can be novel in unseen data

## Evaluation of the Morphological Analyzer/Tagger

| | unseen test data (3,717 words) precision/recall (%) |
|---|---|
| Treebank trained | 95.78/95.39 |
| Off-the-Shelf | 29.42/31.25 |

# Overview

- Introduction to Supervised Statistical Parsing with LTAG
- LTAG grammar extracted from the Penn Korean Treebank
- Morphological Tagging: Motivation and Experiments
- Statistical parsing of Korean using a Morphological Tagger

## Morphological Analysis Incorporated into the Statistical Model

In each probability model used in the parser where inflected word forms are used we incorporate the output of the morph tagger as a backoff level

For example, take the probability model for adjunction:

$$
\begin{aligned}
P_a(t, \eta \rightarrow t') &= Pr(t', p', w' \mid \eta, t, w, p) && (1) \\
&= Pr(t' \mid \eta, t, w, p) \times && (2) \\
&\quad\; Pr(p' \mid t', \eta, t, w, p) \times \\
&\quad\; Pr(w' \mid p', t', \eta, t, w, p)
\end{aligned}
$$

## Morphological Analysis Incorporated into the Statistical Model

- $e_1$ = lexicalized model using stems;
  $e_2$ = part-of-speech tags from the morphological tagger:

$$
\begin{aligned}
Pr_{e_1} &= Pr(t' \mid \eta, t, w, p) \\
Pr_{e_2} &= Pr(t' \mid \eta, t, p)
\end{aligned}
$$

- The backoff model is computed as follows:

$$\lambda(c) \times Pr_{e_1} + (1 - \lambda(c)) \times Pr_{e_2}$$

## Parsing Experiment: Training and Test Data

- Training data for Korean: the Penn Korean Treebank (Han et al. 2002)
- Train statistical morphological tagger and statistical LTAG parser
- Evaluate accuracy on test data
- Our evaluation:
  Train on 4,653 sentences (49,473 words)
  Test on 425 sentences (3,717 words)

## Example derivation reported by the statistical parser

| Index | Word | Gloss | POS tag (morph) | Elem Tree | Node Address | Subst/ Adjoin |
|---|---|---|---|---|---|---|
| 0 | 모든 | every | DAN | $\beta$NP*=1 | root | 2 |
| 1 | 호출 | call | NNC | $\beta$NP*=1 | root | 2 |
| 2 | 대호+는 | sign-topic | NNC+PAU | $\alpha$NP=0 | 0 | 6 |
| 3 | 매일 | everyday | ADV | $\beta$VP*=25 | 1 | 6 |
| 4 | 24 | | NNU | $\beta$NP*=1 | 0 | 5 |
| 5 | 시+에 | hour-at | NNX+PAD | $\beta$VP*=17 | 1 | 6 |
| 6 | 바뀌+게 | switch-aux | VV+ECS | $\alpha$S-NPs=23 | - | TOP |
| 7 | 되+지요 | be-decl | VX+EFN | $\beta$VP*=13 | 1 | 6 |
| 8 | . | | SFN | - | - | - |

Parser evaluation results

| | On training data | On unseen test data (425 sents) |
|---|---|---|
| Current Work | 97.58 | 75.7 |
| (Yoon et al. 1997) | – | 52.29/51.95 P/R |

## Summary

- First LTAG-based parsing system for Korean.
- LTAG-based statistical parsing is feasible for a language with free word order and complex morphology.
- Our system has been successfully incorporated into a Korean/English machine translation system as source language analysis component.

## Summary

- The tagger/analyzer obtained the correctly disambiguated morphological analysis with 95.78/95.39%

- The statistical parser obtained a dependency accuracy of 75.7%

- These performance results are better than an existing off-the-shelf Korean morphological analyzer and parser run on the same data.

Grazie . . .

## Experiments with and without the Morphological Tagger

- Even the part-of-speech tags are often unseen in the test data
- When we lexicalize trees we use words from the training data and for unknown words the output of a part-of-speech tagger
- Without a morphological tagger the lexicalization step becomes infeasible (We can annotate the Treebank with a new smaller tagset, but the number of trees for unknown words explodes)
- Thus, we could not easily compare parsing with and without a morphological tagger