Tamil Dependency Parsing: Results using Rule Based and Corpus Based Approaches

Loganathan Ramasamy – Zdeněk Žabokrtský

Charles University in Prague

Feb 21, 2011

Outline

Motivation & Objectives

- Resource poor (?)
- Morphologically Rich

- Develop a Treebank and Parser for Tamil
- Identify issues in Treebank developement
- Test Rule based (RB) and Corpus based (CB) parsers

General Aspects of Tamil Language

- Morphologically rich
  - Agglutinative
  - Compound word constructions
- Head final & Relatively free word order
  - strictly head final
  - within clause word order freedom
- Subject–Verb agreement
  - Subject agrees with verb in person–number–gender

Annotation Scheme

- Developed a small treebank (approx. 3000 words)
- Based on Prague Dependency Treebank PDT 2.0
- PDT 2.0 uses 3 levels of annotation.
- Ours uses only the first 2 layers: *morphological* and *analytical*
- There are 19 analytical functions (or dependency relations) defined for the Tamil treebank.
- Morphological layer contains $\approx$ 460 unique tags
- Rule based parser under TectoMT framework

Annotation Scheme - Annotation of sentence fragments

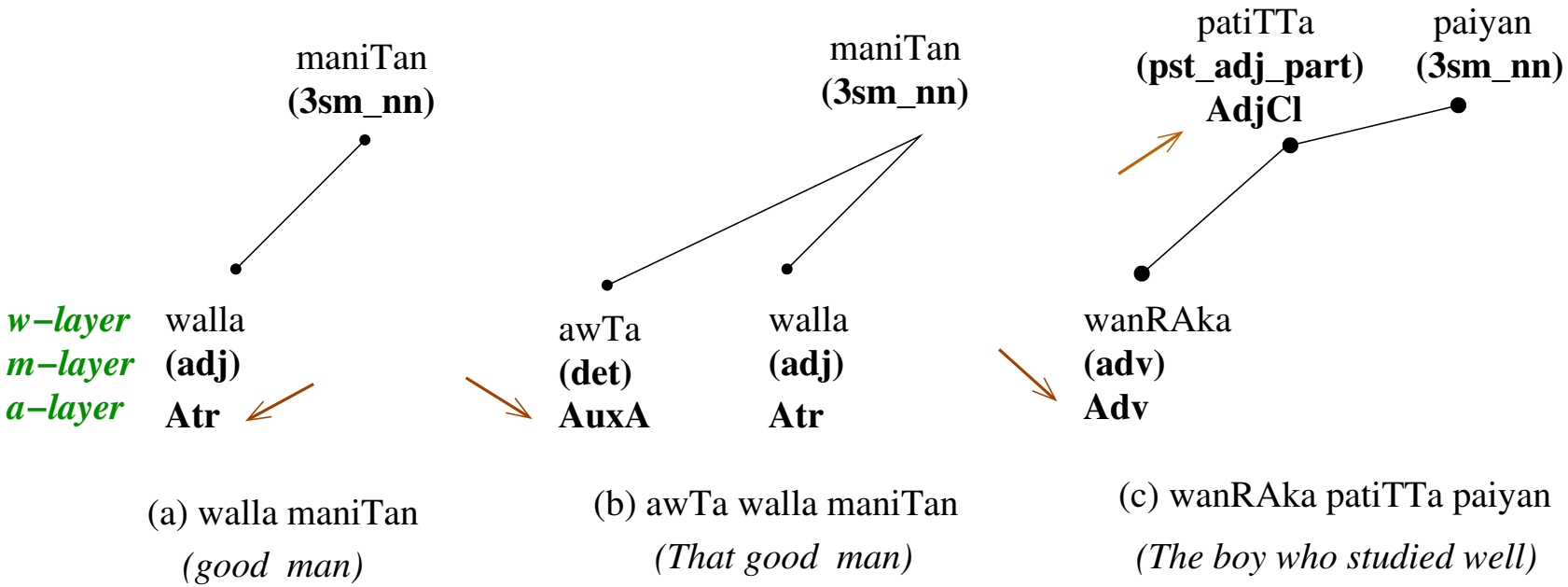
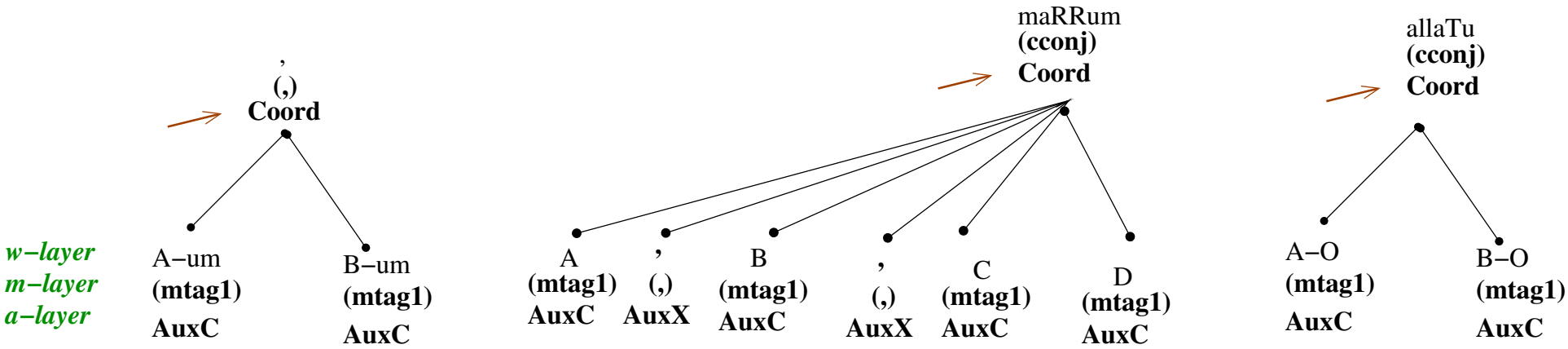


Figure: Illustration of Atr, Adv, AuxA, AdjCl dependencies

Annotation Scheme - Annotation of sentence fragments



(a) A–um , B–um
(A , B)

(b) A, B, C maRRum D
(A, B, C and D)

(c) A–O allaTu B–O
(A or B)

Figure: Illustration of coordination conjunction

Annotation Scheme - Full annotation example

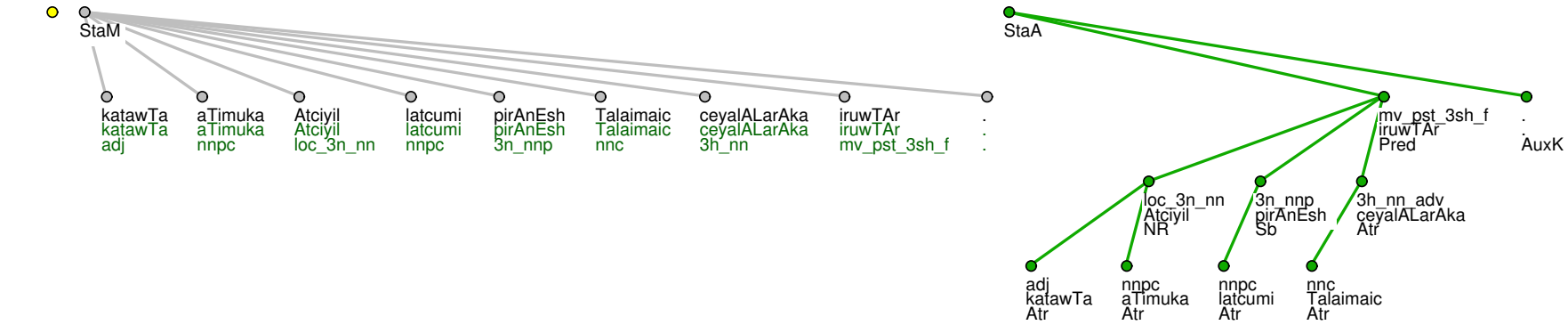


Figure: Annotation using TrEd tool

Rule Based Parser for Tamil

- Uses tagger and simple morphological & syntactic rules to build unlabeled and labeled dependency trees.

Algorithm

1. Tag the input sentence. We used TnT tagger.
2. Build the unlabeled dependency tree by calling
   - *Identify_main_predicate()*
   - *Resolve_coordination()*
   - *Identify_trivial_parents()*
   - *Process_complements()*
3. Assign labels to the edges.

Parsing Example

How to parse the following Tamil sentence using RB parser?

katawTa	aTimuka	Atciyil	latcumi	pirAnEsh	Talaimaic	ceyalALarAka	iruwTAr	.
adj	nnpc	loc_3n_nn	nnpc	nnp	nnc	3h_nn_adv	mv_pst_3sh_f.	Morphological Tag
last	ADMK	in_the_rule	Lakshmi	Pranesh	chief	as_secretariat	was	English gloss

Lakshmi Pranesh was the chief secretariat in the last ADMK rule — English Translation

Figure: A Tamil sentence

Parsing Example

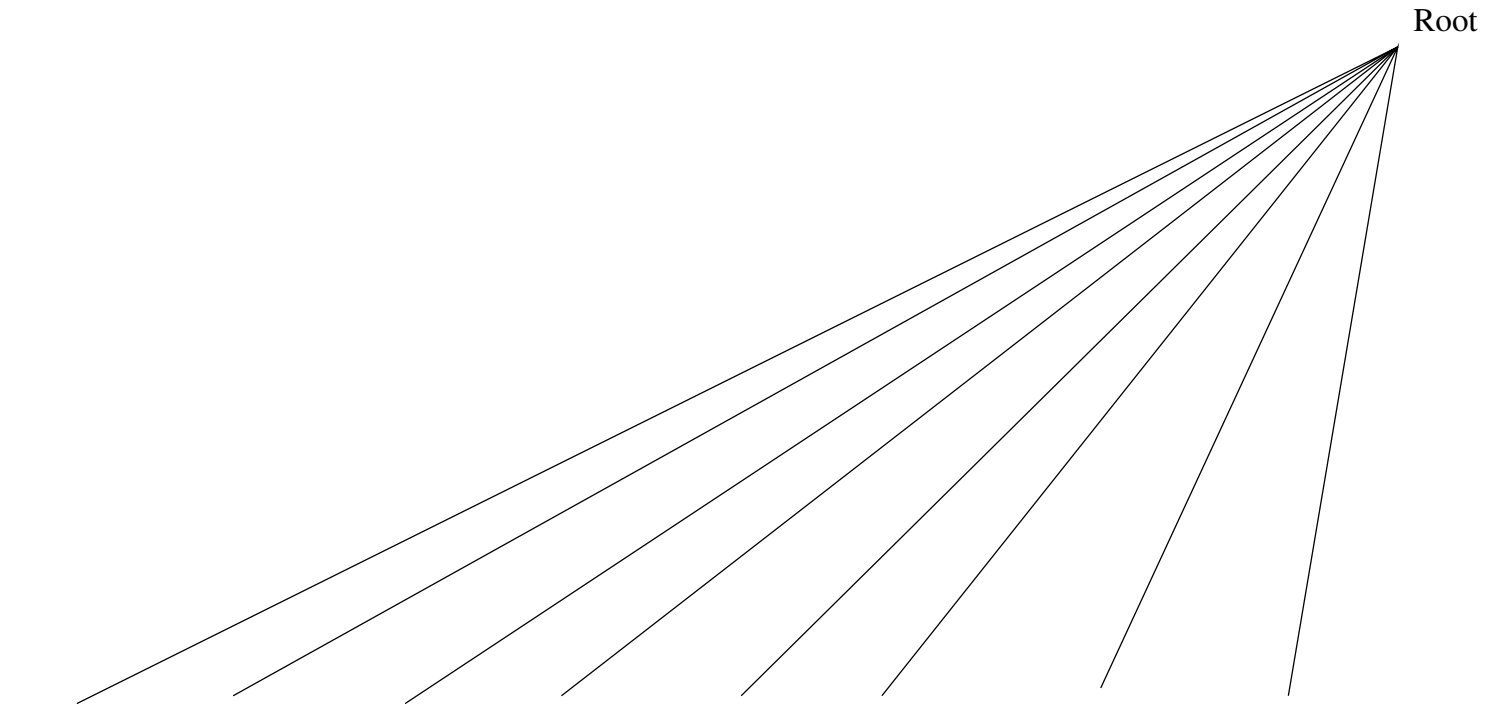


Figure: Initial flat tree

Parsing Example

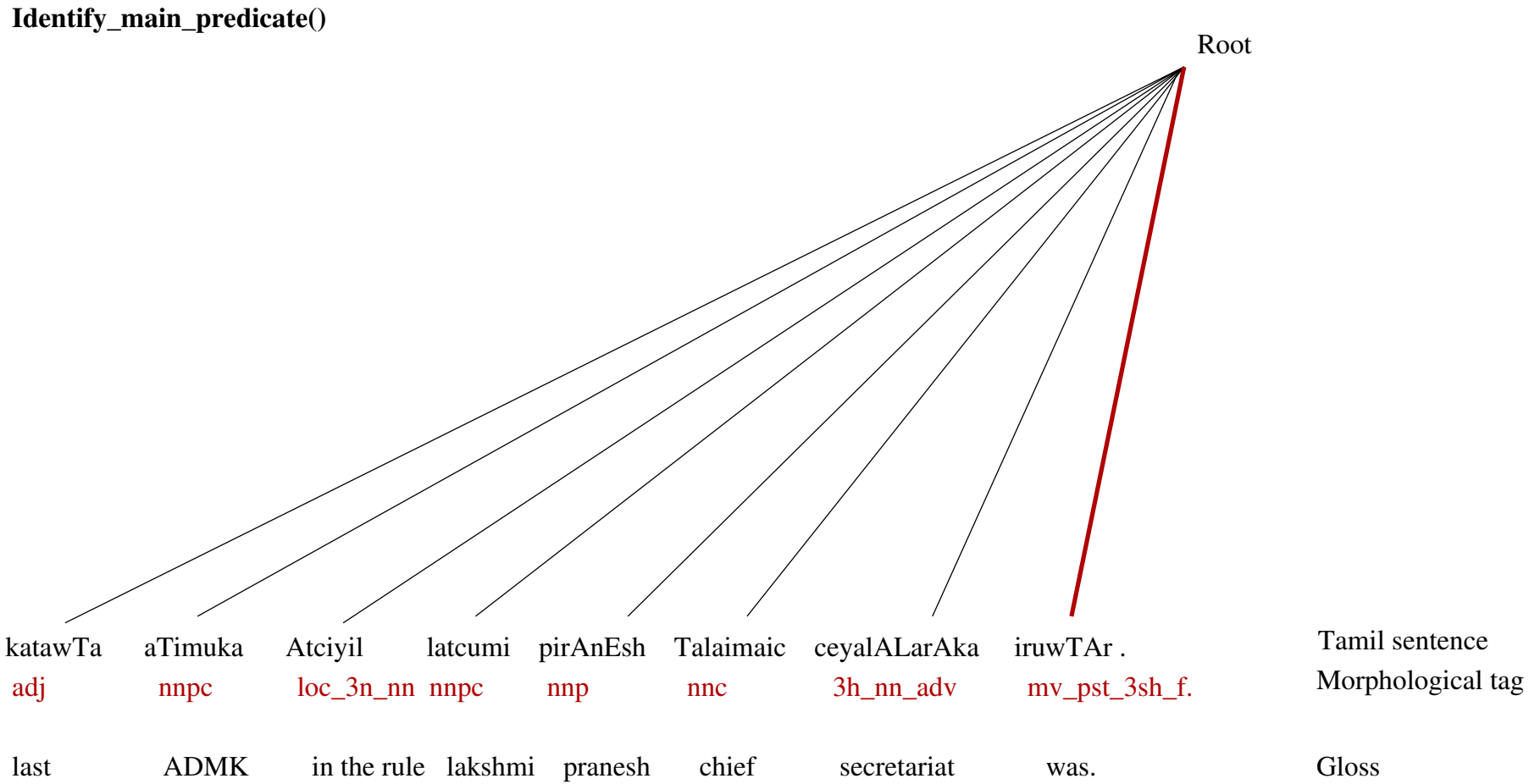


Figure: Identify predicate of the sentence

Parsing Example

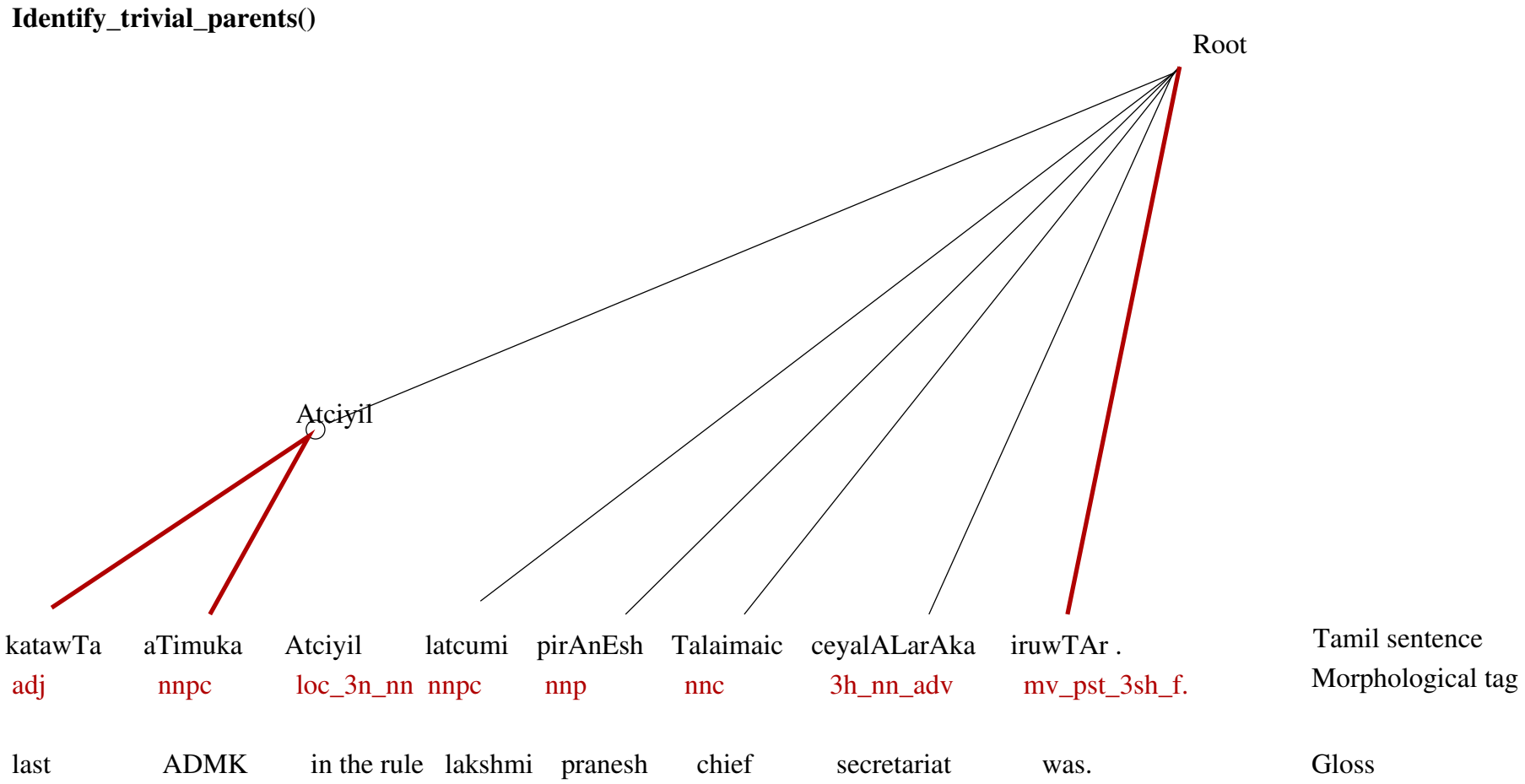


Figure: Attach modifiers to phrasal heads

Parsing Example

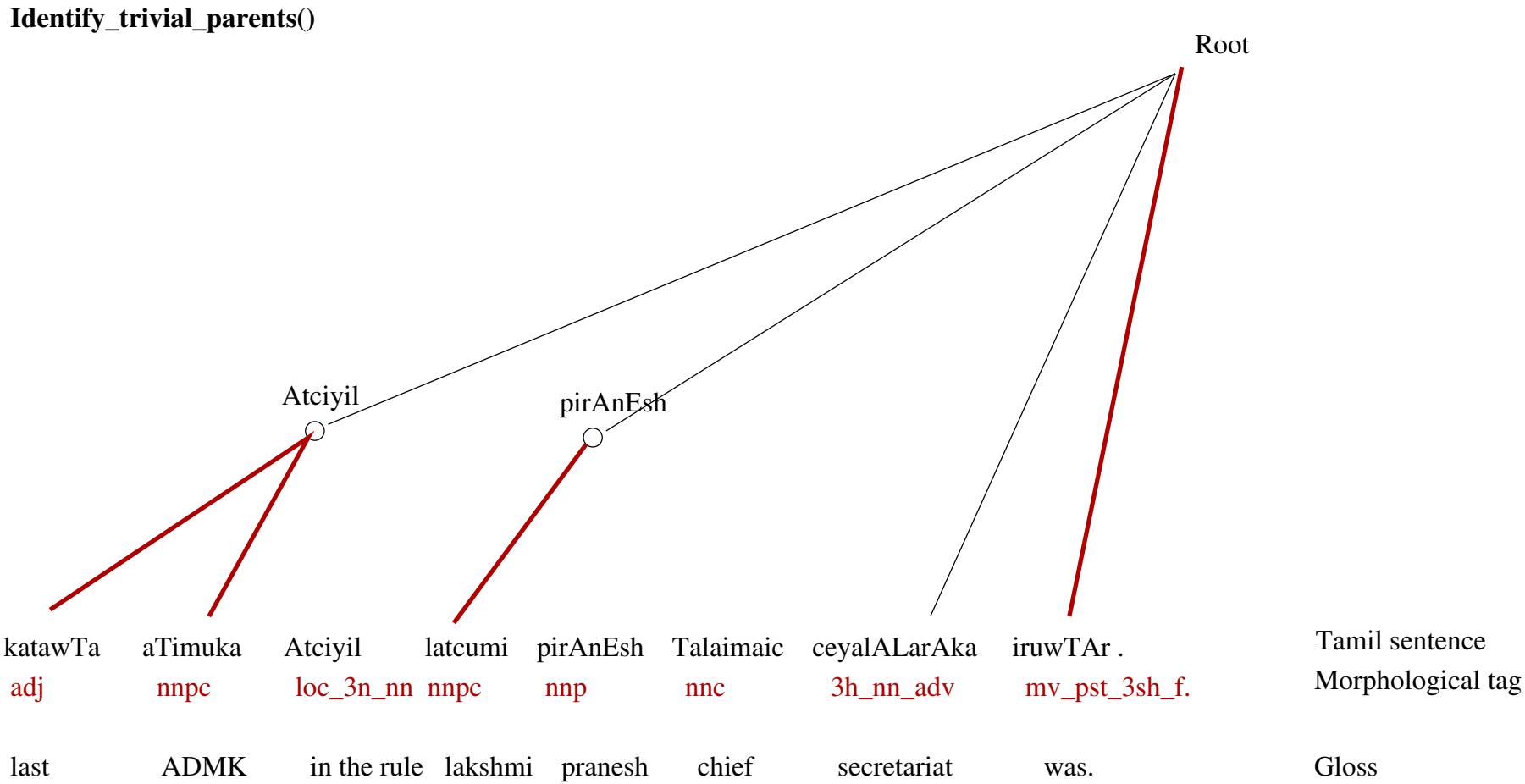


Figure: Attach modifiers to phrasal heads

Parsing Example

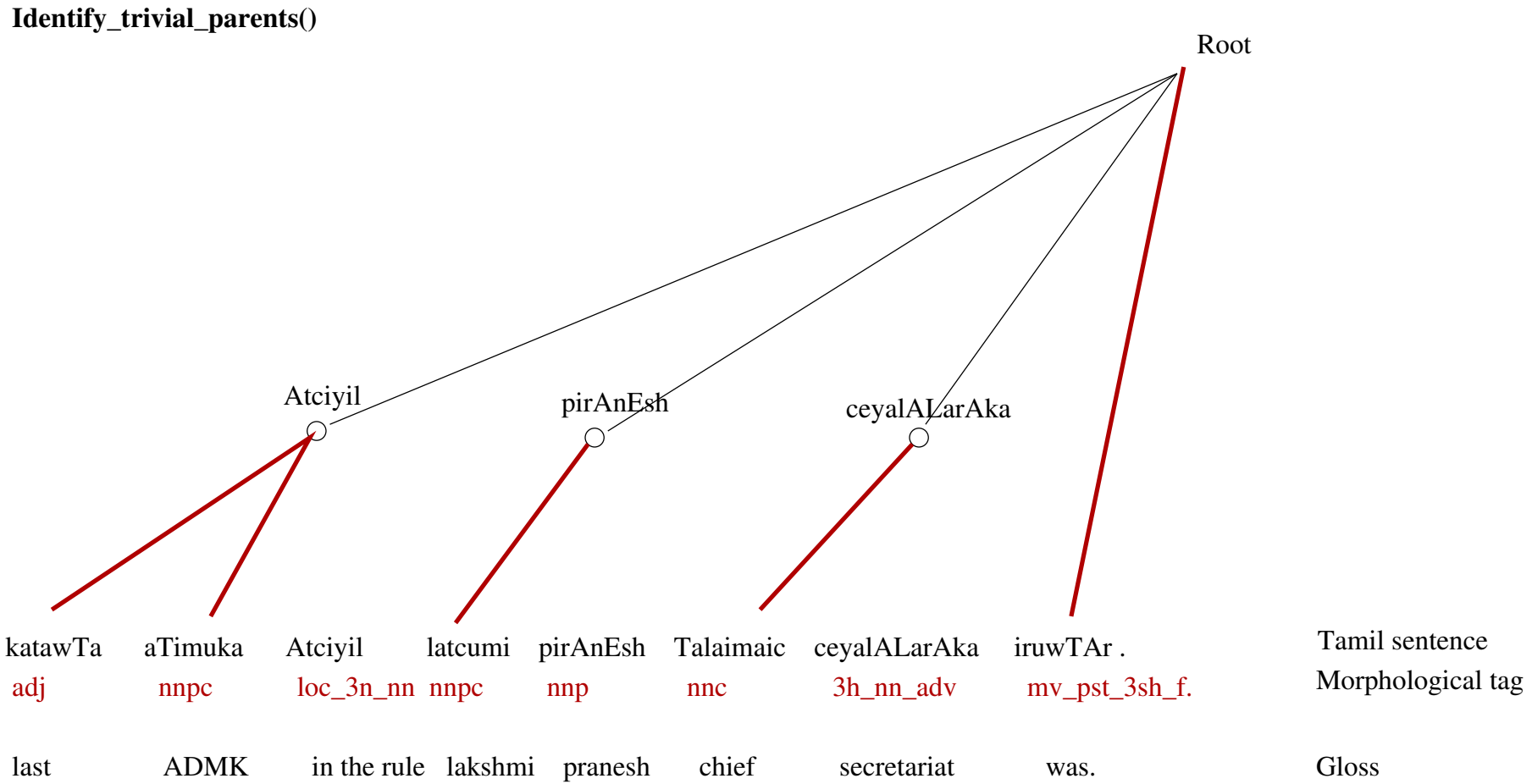


Figure: Attach modifiers to phrasal heads

Parsing Example

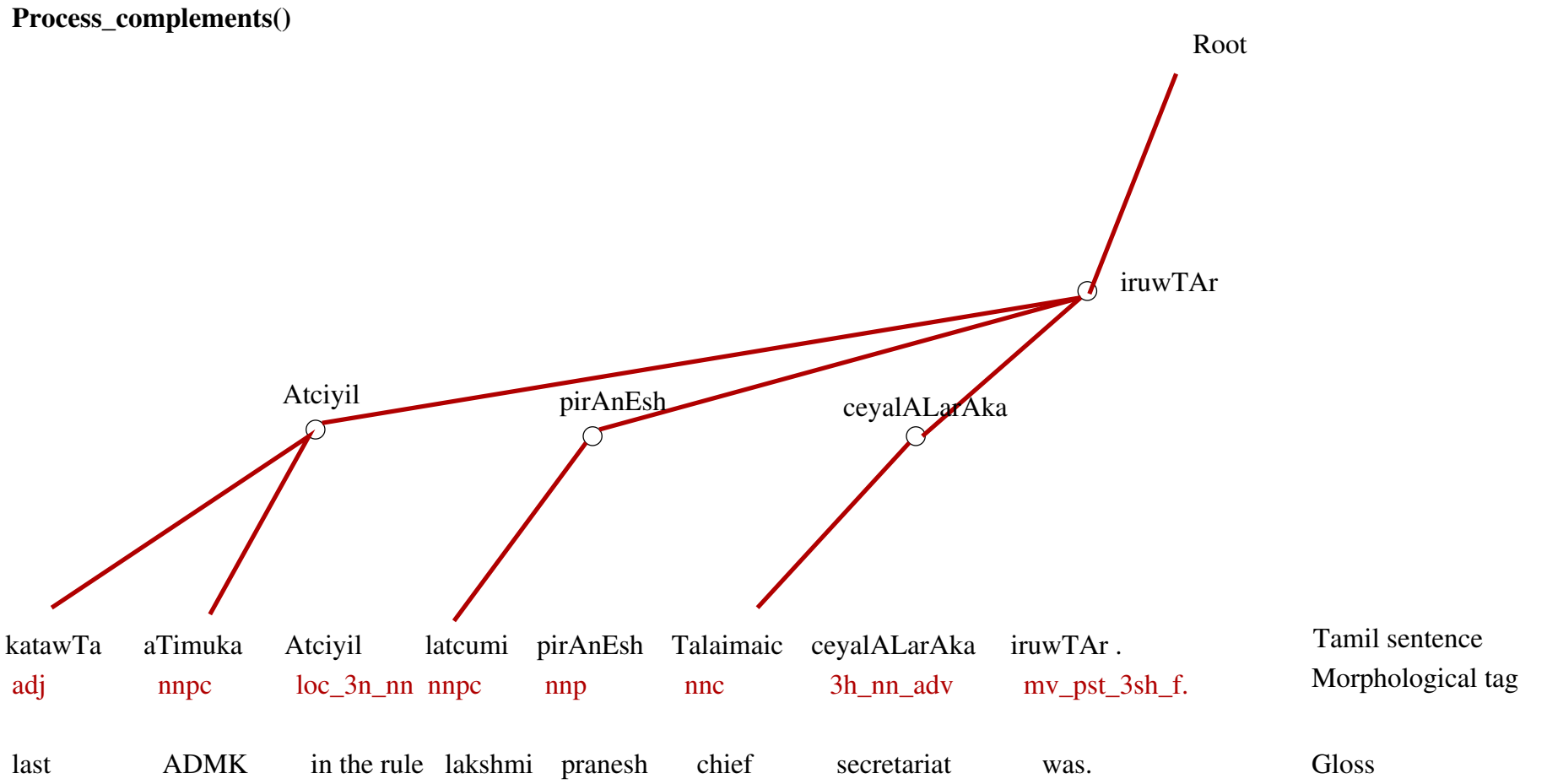


Figure: Process complements, attach arguments to clausal predicates

Parsing Example

Labeling the tree

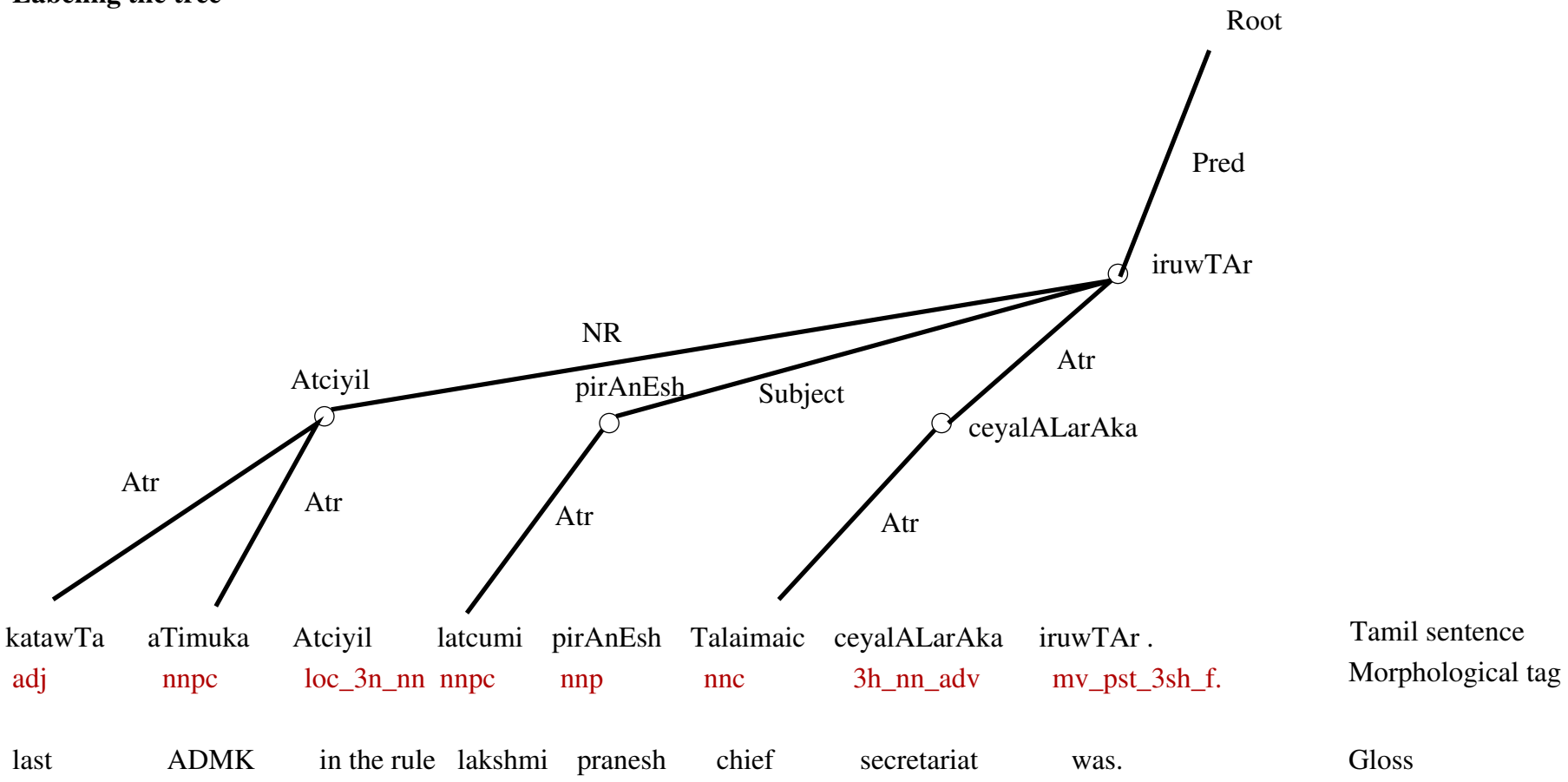


Figure: Labeling of the dependency tree

Experiments and Results: Data

- Corpus 1 is morphologically and syntactically annotated.
- Corpus 2 is only morphologically tagged.

	Corpus 1	**Corpus 2**
Tagset size	296	459
Lexical verb tags	120	194
Auxliary verb tags	31	44
# of words	2961	8421
Unique tokens	1634	3747
1 tag count	1534/(93.88%)	3427/(91.46%)
2 tag count	92/(05.63%)	284/(07.58%)
3 tag count	8/(00.49%)	33/(00.88%)
4 tag count	0/(00.00%)	3/(00.08%)

Table: Corpus statistics

Experiments and Results: Accuracy of RB & CB Parsers

- Rule based parser is tested against the whole 2961 tokens.
- Manual POS experiment uses gold standard tagged data.
- In Auto POS eperiment, tagging was done by TnT tagger.
- MST and Malt parsers are trained on 2008 word tokens and tested against 953 tokens.

	Auto POS	Manual POS
Unlabeled	71.94	84.73
Labeled	61.70	79.13

Table: Rule Based parser accuracy

	MaltParser	MST Parser
Unlabeled	75.03	74.92
Labeled	65.69	65.69

Table: Corpus Based parser accuracy

Conclusion & Future work

- Certain dependency relations are easy to tackle using RB parser.
- Accurate tagging is required for RB approach.
- Prediction of relations such as Subject and Coordination is difficult in both RB and CB parsers.

- More data will be annotated.