An overview of the Natural Language Toolkit

Steven Bird, Ewan Klein, Edward Loper

nltk.org

Summary

- NLTK is a suite of open source Python modules, data sets and tutorials
- supporting research and development in natural language processing
- Download NLTK from nltk.org

Components of NLTK

1. **Code**: corpus readers, tokenizers, stemmers, taggers, chunkers, parsers, wordnet, ... (50k lines of code)
2. **Corpora**: >30 annotated data sets widely used in natural language processing (>300Mb data)
3. **Documentation**: a 400-page book, articles, reviews, API documentation

1. Code

- corpus readers
- tokenizers
- stemmers
- taggers
- parsers
- wordnet
- semantic interpretation
- clusterers
- evaluation metrics
- …

2. Corpora

- Brown Corpus
- Carnegie Mellon Pronouncing Dictionary
- CoNLL 2000 Chunking Corpus
- Project Gutenberg Selections
- NIST 1999 Information Extraction: Entity Recognition Corpus
- US Presidential Inaugural Address Corpus
- Indian Language POS-Tagged Corpus
- Floresta Portuguese Treebank
- Prepositional Phrase Attachment Corpus
- SENSEVAL 2 Corpus
- Sinica Treebank Corpus Sample
- Universal Declaration of Human Rights Corpus
- Stopwords Corpus
- TIMIT Corpus Sample
- Treebank Corpus Sample
- …

NLTK Modules

- **corpora:** a package containing modules of example text
- **tokenize:** functions to separate text strings
- **probability**: for modeling frequency distributions and probabilistic systems
- **stem** – package of functions to stem words of text
- **wordnet** – interface to the WordNet lexical resource
- **chunk** – identify short non-nested phrases in text
- **etree**: for hierarchical structure over text
- **tag**: tagging each word with part-of-speech, sense, etc.
- **parse**: building trees over text
  - recursive descent, shift-reduce, probabilistic, etc.
- **cluster**: clustering algorithms
- **draw**: visualize NLP structures and processes
- **contrib:** various pieces of software from outside contributors

(Some) Modules in NLTK

Language Processing Task	NLTK module	Some functionalities
Accessing corpora	Nltk.corpus	Standardized interfaces to corpora and lexicons
String processing	Nltk.tokenize	Sentence and word tokenizers
	Nltk.stem	Stemmers
Part-of-speech tagging	nltk.tag	Various part-of-speech taggers
Classification	Nltk.classify	Decision tree, maximum entropy
	Nltk.cluster	K-means
Chunking	Nltk.chunk	Regular expressions, named entity tagging

Getting Started: Corpora

- **Task:** Accessing corpora
- **NLTK module:** nltk.corpus
- **Functionality:** standardized interfaces to corpora and lexicons
- Example:

```
>>> from nltk.corpus import gutenberg

>>> gutenberg.fileids()

>>> hamlet = gutenberg.words('shakespeare-hamlet.txt')

>>> hamlet[1:100]
```

- Also: Brown, Reuters, chats, reviews, etc.

Getting Started: String Processing

- **Task:** string processing
- **Modules:** nltk.tokenize, nltk.stem
- **Functionality:** word tokenizers, sentence tokenizers, stemmers
- Example:

```
>>> text = nltk.word_tokenize("The quick brown fox jumps over the lazy
dog")

>>> text = nltk.sent_tokenize("The quick brown fox jumps over the lazy dog.
What a lazy dog!")

>>> from nltk.stem.wordnet import WordNetLemmatizer

>>> WordNetLemmatizer().lemmatize(’dogs’,’n’)

>>> WordNetLemmatizer().lemmatize(’jumps’,’v’)
```

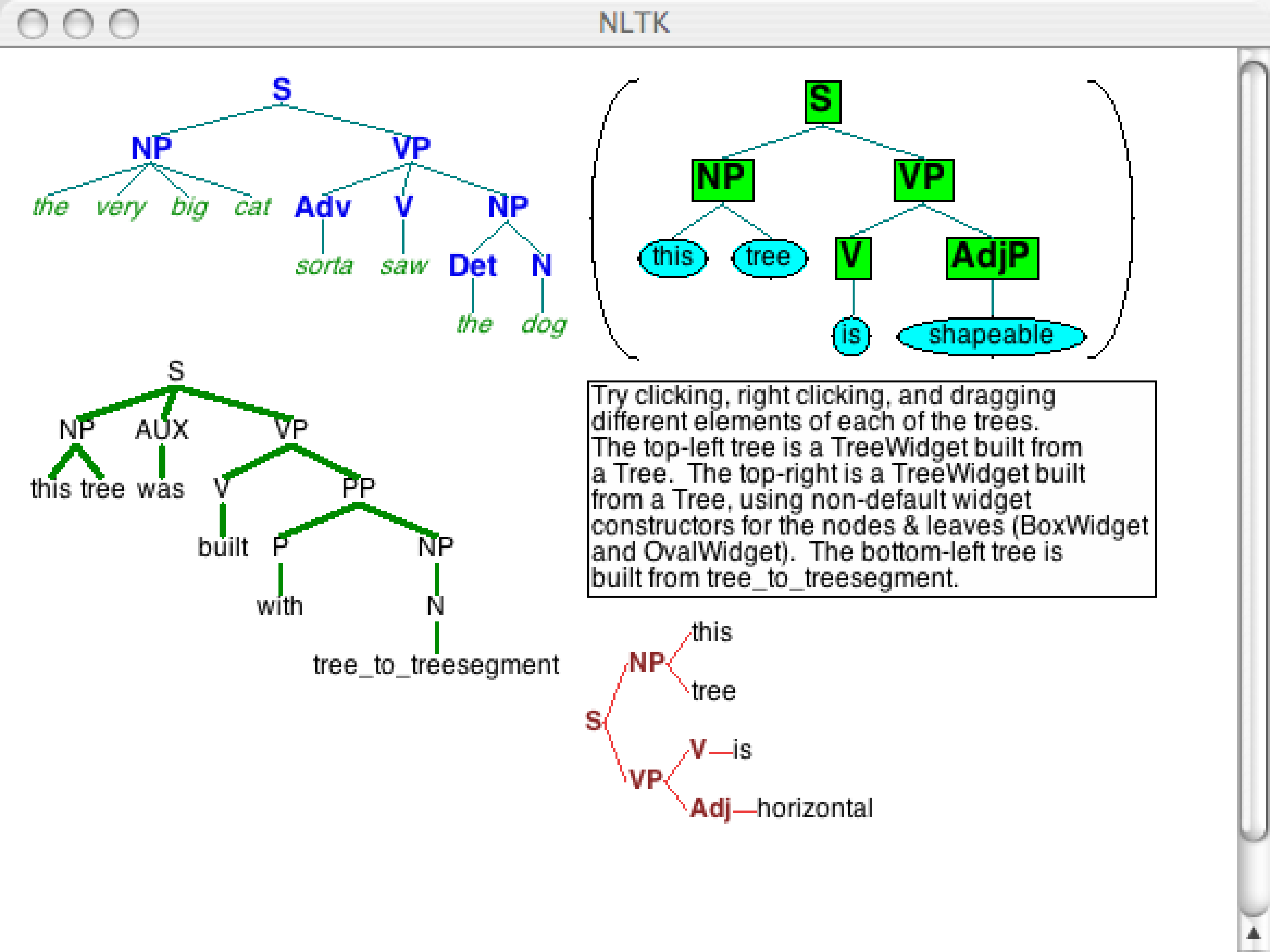
Getting Started: Part-of-Speech Tagging

- **Task:** Part-of-speech tagging
- **Module:** nltk.tag
- **Functionality:** Brill, HMM, TnT taggers
- Example:

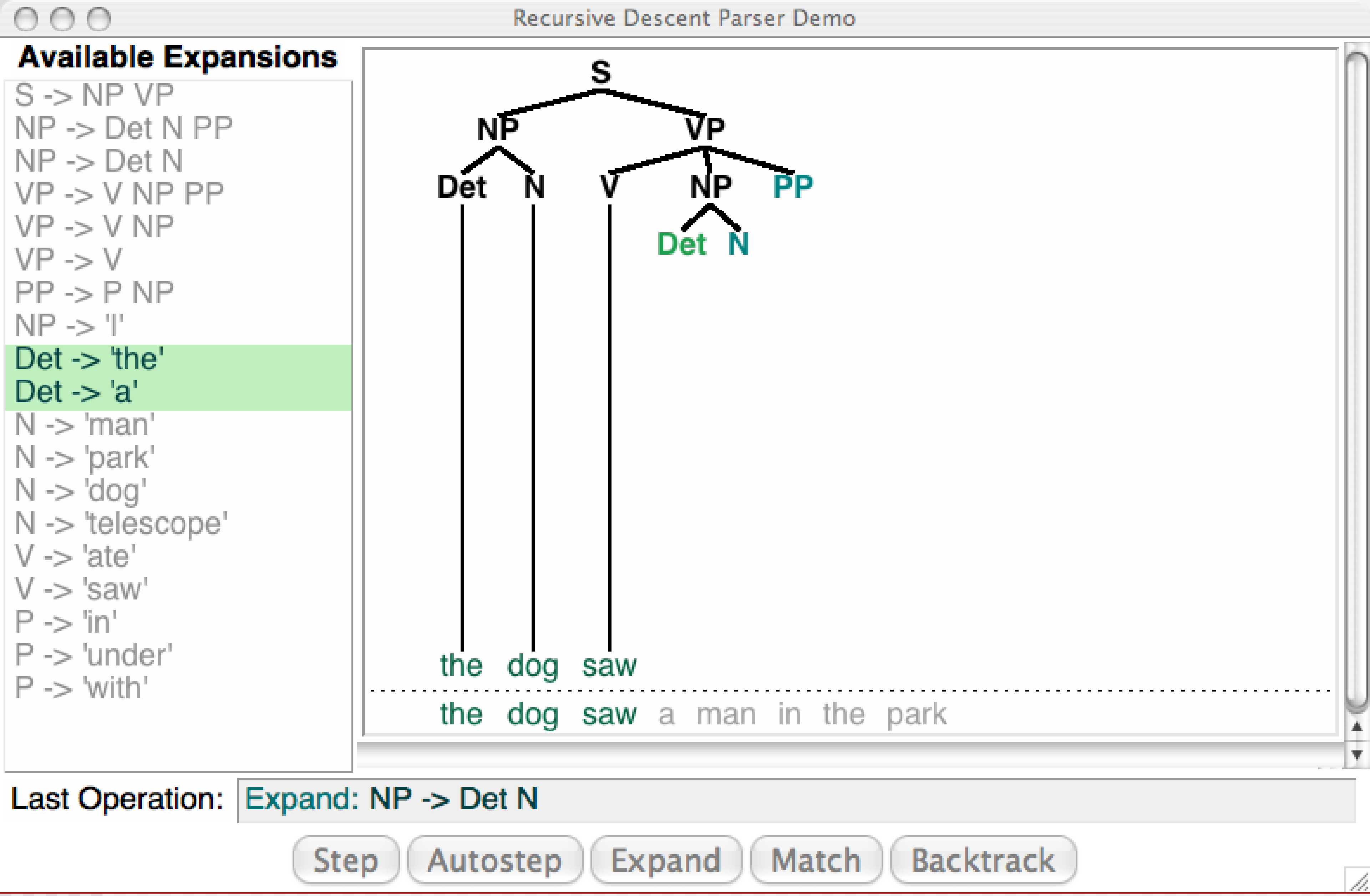
```
 >>> text = nltk.word_tokenize(“It was the best of times, it
was the worst of times.”)

>>> nltk.pos_tag(text)
```

(Penn Treebank tag set: http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html)



Try clicking, right clicking, and dragging different elements of each of the trees. The top-left tree is a TreeWidget built from a Tree. The top-right is a TreeWidget built from a Tree, using non-default widget constructors for the nodes & leaves (BoxWidget and OvalWidget). The bottom-left tree is built from tree_to_treesegment.

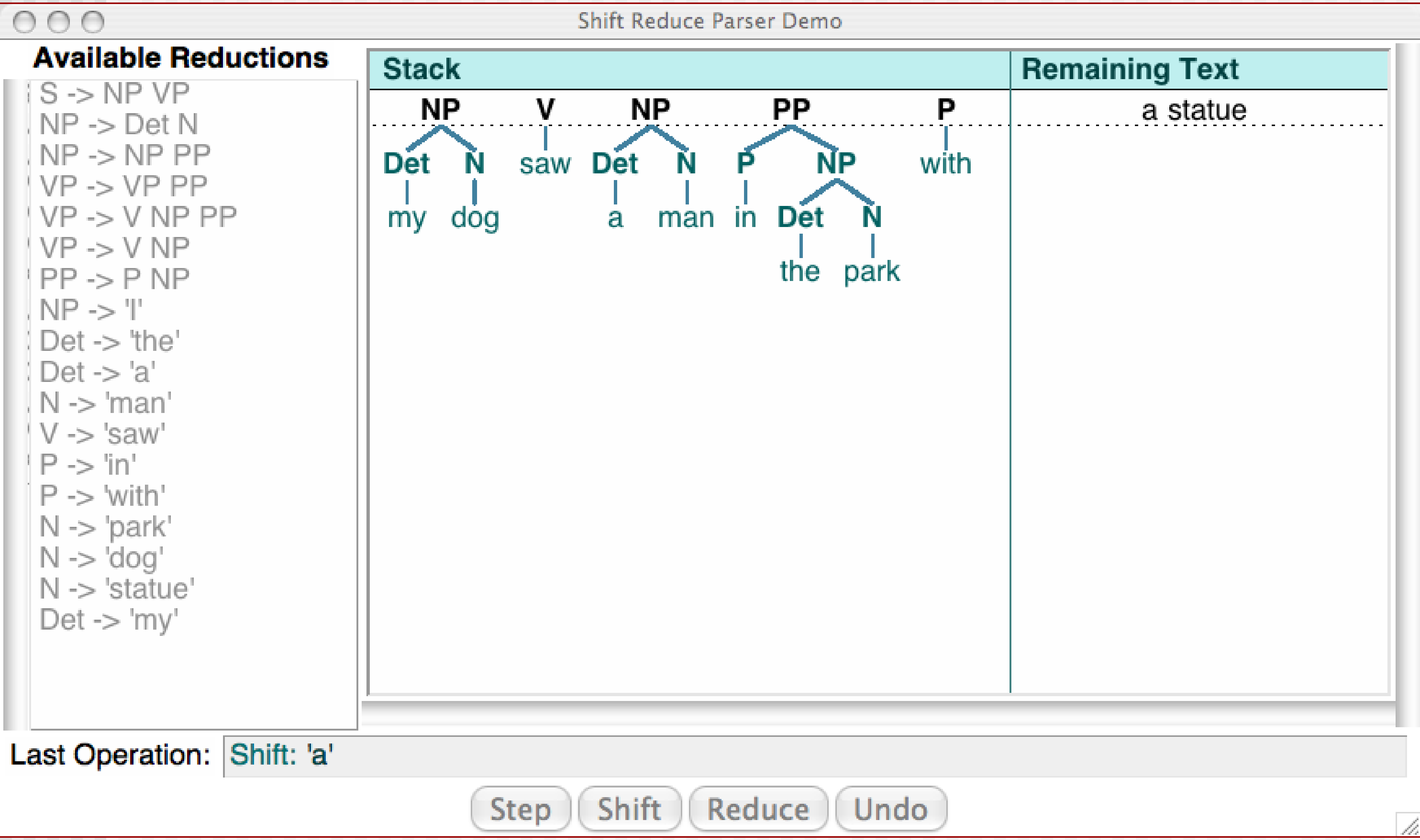


Recursive Descent Parser Demo

Available Expansions

S -> NP VP
NP -> Det N PP
NP -> Det N
VP -> V NP PP
VP -> V NP
VP -> V
PP -> P NP
NP -> 'I'
Det -> 'the'
Det -> 'a'
N -> 'man'
N -> 'park'
N -> 'dog'
N -> 'telescope'
V -> 'ate'
V -> 'saw'
P -> 'in'
P -> 'under'
P -> 'with'

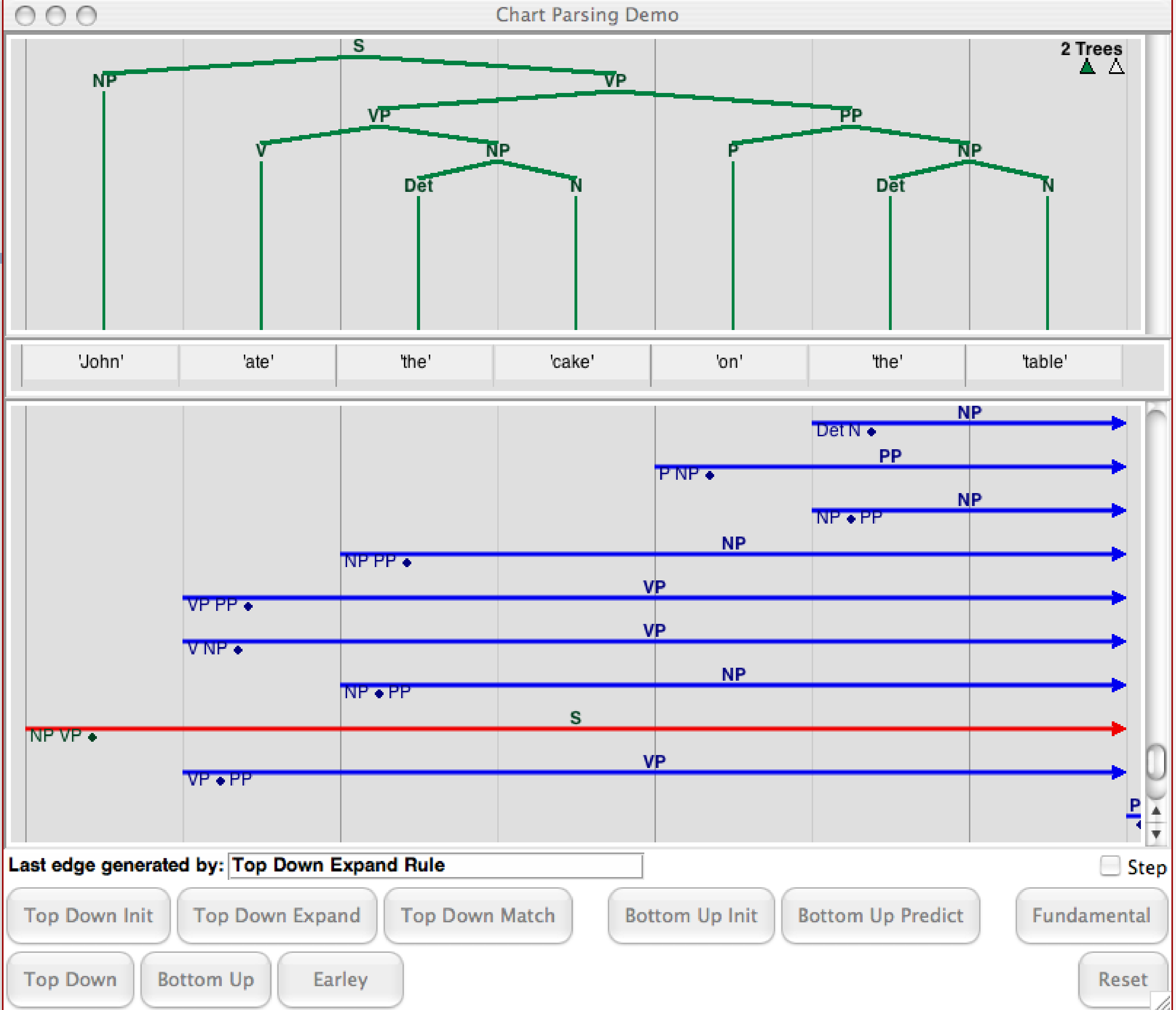
Last Operation: Expand: NP -> Det N



Available Reductions

S -> NP VP
NP -> Det N
NP -> NP PP
VP -> VP PP
VP -> V NP PP
VP -> V NP
PP -> P NP
NP -> 'I'
Det -> 'the'
Det -> 'a'
N -> 'man'
V -> 'saw'
P -> 'in'
P -> 'with'
N -> 'park'
N -> 'dog'
N -> 'statue'
Det -> 'my'

Last Operation: Shift: 'a'



Last edge generated by: Top Down Expand Rule

Adoption in NLP courses

Amsterdam, Ben-Gurion, Brown, Bryn Mawr, CDAC-Mumbai, Coruña, Edinburgh, Erlangen, Georgetown, Helsinki, IIT-Bombay, Iowa State, Konstanz, MIT, Macquarie, Magdeburg, Malta, Marquette, Melbourne, Nancy, Naval Postgraduate School, Northeastern, Ohio State, Pitt, San Diego State, Simon Fraser, Stanford, Syracuse University, Tsuda College, U Colorado, UC Berkeley, UMass Amherst, UNAM, U Penn, UT Austin, Warsaw

Tutorials for Python and NLTK

- Python

  http://docs.python.org/tut/tut.html, the classic by Guido van Rossum

- NLTK is a SourceForge project at: http://www.nltk.org

  documentation: http://www.nltk.org/documentation, including

  book: http://www.nltk.org/book

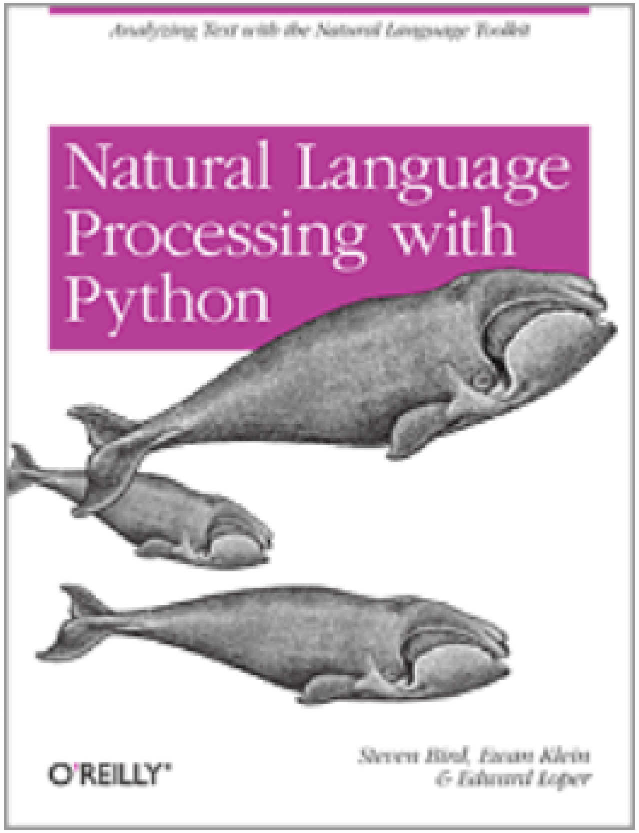
  API: http://nltk.googlecode.com/svn/trunk/doc/api/index.html

3. Documentation

- a 400-page book about natural language processing in Python and NLTK
  - teaches Python and NLP
  - provides numerous examples and exercises
- installation instructions
- presentation slides for some of the book chapters
- API Documentation: describes every module, interface, class, and method

NLTK Book

- Very useful resource
- Can buy a physical copy (~$45 amazon.ca)
- Also available for free online: http://nltk.org/book/


Analyzing Text with the Natural Language Toolkit
Natural Language Processing with Python
O'REILLY
Steven Bird, Ewan Klein & Edward Loper

Contribute…

- NLTK is an open source project
- all code, data, documentation is free
- dozens of people have contributed over the past 6 years
- please visit the website for project ideas
- sign up on the NLTK-Announce mailing list to hear about new releases