

Book Proposal: *Beginning Python Programming for Language Research* by Stuart Robinson and Harald Baayen

August 31, 2006

1 Overview

Beginning Python Programming for Language Research will provide an introduction to the Python programming language for language researchers, assuming that the average reader will have little to no previous programming experience. The introduction to Python is reasonably general and could profitably be read by any beginner wishing to learn Python from scratch but will focus primarily on its use in text processing. Although it focuses on text processing, the book will provide a comprehensive introduction that provides the foundation for the development of general programming skills. It can be used either for self-study or in the classroom as a textbook.

2 Reason for Writing

Language researchers typically lack the programming skills required to automate the various tasks that they perform as part of their everyday working life, such as those listed in (1), and could profit enormously by learning the basics of a scripting language such as Python.

- (1)
 - text mining/corpus analysis (counting words, tags, phrases)
 - querying data bases (Celex, Shoebox, etc)
 - querying web pages
 - manipulating data to fit the requirements of experimental software
 - working with XML/SGML encoded text

In addition, there is a growing trend in linguistics in favor of quantitative techniques (see, for example, Bod et al. (2003)), which has emerged partially out of dissatisfaction with the rationalist approach of mainstream generative linguistics and partially in response to industry needs driven by the growth of the worldwide web. In many cases, there do not exist ready-made tools to pursue these quantitative techniques, and researchers must therefore be able to create their own, which requires at least basic programming skills.

3 Contents

At present, the book consists of two parts. The first part introduces the basics of Python programming.

First program Provides a simple introductory program that is elaborated upon through the book.

Statements Statements and expressions.

Data types Simple data types: strings, integers, floats.

Data structures Data structures: lists, tuples, and dictionaries.

Data flow If statements, while loops, for loops.

Functions The basics of functions: arguments, return values, and issues of scope.

Exceptions Error-handling: try/except blocks, built-in vs. custom exceptions.

IO File input and output, command-line arguments, command-line options, and directories.

Modules and packages Modules and packages, import statements.

Strings Strings covered in considerable depth, a review of the string object and its methods.

The second part introduces more advanced topics relevant to language research.

Object-Oriented Programming An introduction to the basic concepts in object-oriented programming.

Classes Detailed coverage of how object-oriented programming works in Python: class definitions, instance vs. class variables, public vs. private variables.

An Introduction to Regular Expressions An in-depth introduction to regular expressions in the abstract.

Regular Expressions in Python How regular expressions are handled in Python.

XML The basics of XML and how they are handled in Python.

Data Persistence Using files, pickling, and SQL for data persistence.

The Web Web programming basics (querying a web page and extracting its contents).

All of the code used in the book can be made available separately as an accompany CD and/or on a companion web site.

A manuscript of the book already exists which consists of 270 pages. We estimate that the book will grow to approximately 350 pages in its final form.

4 Readership

The readership of this book is fairly broad and falls into three groups:

Humanities Researchers in the humanities could use the book to learn the basics of text manipulation, which is increasingly relevant for research in the humanities thanks to the growth of electronic texts available on the WWW.

Language Sciences Linguists and psychologists who do linguistic research increasingly need to learn how to manipulate data, do modelling, perform analysis, process corpora, and the like using a programming language. As a scripting language with good string processing capabilities, Python is ideally suited to the task.

Computational Linguistics Programming skills are a must for computational linguistics and Python is increasingly the language of choice in the field. There is increasing interest in text processing due to the growth of the WWW and this book should cater well to this audience.

The book will serve as a textbook in introductory computational linguistics courses. For commentary on the existing resources for, see <http://www.ai.uga.edu/mc/PythonForNewbieLinguists.html>. Quite a number of introductory computational linguistics courses use Python as their language of choice for instruction, as can be seen from the following partial list:

- <http://www-rohan.sdsu.edu/~malouf/ling571.html>
- <http://www.cog.brown.edu/~mj/classes/cg136/>
- http://www.cis.upenn.edu/~cis530_fall2001/
- http://www.indiana.edu/~deanfac/blfal03/ling/ling_1545_3320.html
- <http://uts.cc.utexas.edu/~jbaldrid/courses.html>
- <http://www.cs.umass.edu/~mccallum/courses/cl2006/syllabus.html>
- <http://www9.georgetown.edu/faculty/mad87/06/420/syllabus.html>
- <http://www.cs.mu.oz.au/460/>
- <http://www.inf.ed.ac.uk/teaching/years/ug3/CourseGuide/node53.html>
- <http://www.cs.brandeis.edu/~cs114/>

5 Competing Titles

There are a few titles currently in publication that cover text processing for linguistics, but most of these are for other programming languages, address a different audience, or have somewhat different goals than the proposed book.

There are two books aimed at linguists wishing to learn how to program: Michael Hammond's books *Programming for Linguists: Java for Language Researchers* and *Programming for Linguists: Perl for Language Researchers* (Hammond, 2002, 2003). These books do not address Python and suffer from other shortcomings, as well (see Manning (2005)'s review in *Language*).

There are other introductory Python books, but none of them is appropriate for the intended audience since they are pitched at the wrong level and/or they focus on areas of little interest to the intended audience. For example, Mertz (2003) provides in-depth treatment of text processing in Python but is too advanced for the intended audience.

Mark Johnson's introductory textbook on Python for corpus analysis, entitled *Essential Python for Corpus Analysis*, is scheduled for release in August (Johnson, 2006), but, as its title suggests, it is less ambitious in scope than the proposed book.

6 Other Relevant Information

A rough draft of the textbook was used to teach a two-week intensive Python course at the Max Planck Institute during May 2006.

There are no special requirements as far as its printing is concerned—i.e., there are no foldouts or colored text or graphics. Other than screenshots, there should be no photo reproductions.

Finally, the book is written in L^AT_EX and its formatting can therefore be easily changed.

7 About the Authors

Both authors are affiliated with the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands (MPI). The first author is a Ph.D student with a background in linguistics and software engineering. He is actively involved in the development of the Natural Language Toolkit for Python (see `nltk.sf.net` for more info) (Loper and Bird, 2002, 2004). The second author is professor of quantitative linguistics at the Radboud University Nijmegen and research associate of the MPI. For more information, see attached CVs.

References

- Rens Bod, Jennifer Hay, and Stefanie Jannedy, editors. Cambridge, MA, MIT Press, 2003.
- Michael Hammond. *Programming for Linguists: Java Technology for Language Researchers*. Blackwell, New York, 2002.
- Michael Hammond. *Programming for Linguists: Perl for Language Researchers*. Blackwell, New York, 2003.
- Mark Johnson. *Essential Python for Corpus Analysis*. Blackwell, New York, 2006.
- Edward Loper and Steven Bird. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Philadelphia, July 2002, Association for Computational Linguistics*, 2002. URL <http://arxiv.org/abs/cs/0205028>.
- Edward Loper and Steven Bird. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Demonstration Session, Barcelona, July 2004*, 2004. URL <http://www.ldc.upenn.edu/sb/home/papers/nltk.pdf>.
- Chris Manning. Review of programming for linguists: Java technology for language researchers. *Language*, 81(3):740–742, 2005.
- David Mertz. *Text Processing in Python*. Addison-Wesley, Reading, MA, USA, 2003.