A Finite State Network for Phonetic Text Processing

Edward John Garrett

Eastern Michigan University
egarrett@emich.edu

Abstract. In the past, phonetic transcriptions were made using a wide variety of fonts and formats, which hampered the development of phonetic text processing tools. Today, however, the increasing number of language documentation projects making their data freely available over the Web, combined with the adoption of the Unicode Standard by linguists as "best practice" character encoding, present linguistic software developers with an unprecedented opportunity to develop powerful tools for the analysis of phonetic text. This paper describes the generation of a finite state transducer that converts text represented in the International Phonetic Alphabet into phonetic feature sets.

1 Introduction

In response to the crisis of language endangerment, the number and variety of language documentation projects has risen dramatically in recent years. Once content to produce printed grammars, dictionaries and texts, however, such projects are now taking advantage of developments in Internet technologies and multilingual computing to make diverse multimodal linguistic data freely available over the Web, for language researchers and community members alike.

This expansion of linguistic data on the Web presents both challenges and opportunities. The challenges include the problem of insuring that linguistic data is archived in accessible and flexible formats. Under the rubric of "best practices" in language documentation, much progress has already been made in this area [1]. Simultaneously, the adoption by linguists of best practices in language documentation is also creating opportunities for new computational methods and tools for linguistic processing. This paper explores one such opportunity: phonetic text processing.

1.1 Phonetic Text Processing and the Unicode Standard

Until recently, phonetic data represented by the International Phonetic Alphabet (IPA) had to be stored in a variety of fonts and encodings, determined by platform, software, or convenience. With the emergence of the Unicode Standard [2], however, this situation has changed: now, linguists are told to encode phonetic text as Unicode [3]. Since Unicode has the virtue of assigning characters from nearly all written scripts, including IPA, with uniform and invariant code points, IPA text encoded as Unicode can