CHAPTER SEVEN

CONCLUSION

7.1 INTRODUCTION

As initially discussed in Section 1.4, the aim of this thesis was two-fold: (a) to obtain a mechanism for pronunciation modelling that is well suited to bootstrapping; and (b) to analyse the bootstrapping of pronunciation models from a theoretical and a practical perspective, as a case study in the bootstrapping of HLT resources. In this chapter we evaluate the extent in which we were able to reach these goals. We summarise the contribution of this thesis, and discuss future work.

7.2 SUMMARY OF CONTRIBUTION

This thesis was able to demonstrate conclusively that the proposed bootstrapping approach is a practical and cost-efficient way to develop pronunciation dictionaries in new languages. The specific contributions made in the course of this research are the following:

- A demonstration of a fully interactive (on-line) bootstrapping approach to the development of pronunciation dictionaries, in Section 6.5 [82].

- Development and evaluation of a practical system that allows users (without specialist linguistic expertise) to develop such pronunciation dictionaries, and an analysis of the factors that influence this process, in Section 6.3 [83, 84].

- The development of *Default&Refine*, a new algorithm for grapheme-to-phoneme prediction, in Section 4.6 [85]. This algorithm has a number of desirable features, including language independence, rapid generalisation from small data sets, good asymptotic accuracy, robustness to human error, and the production of compact rule sets.

- A number of algorithmic refinements to ensure a practical bootstrapping system, including optimised alignment and an incremental (on-line learning) version of the g-to-p algorithm used during bootstrapping, in Sections 4.4.2 and 4.6.4 [84, 86].
- The development and evaluation of a novel error-detection tool that can assist in the verification of pronunciation dictionaries – both during bootstrapping and in support of alternative dictionary development approaches, in Section 6.4 [86].
- Definition of a conceptual framework that can be used to describe the bootstrapping process in general, and the bootstrapping of pronunciation dictionaries in particular, in Chapter 3.
- Development of usable pronunciation dictionaries in a number of South African languages (isiZulu, Sepedi, Afrikaans and Setswana), and the integration of these dictionaries in actual speech technology (speech recognition and speech synthesis) systems, in Section 6.6.
- The development of *minimal representation graphs*: a theoretical framework that supports the rigorous analysis of instance-based learning of rewrite rule sets, in Section 5. This framework aims to derive the smallest possible rule set describing a given set of discrete training data.

7.3 FURTHER APPLICATION AND FUTURE WORK

The current thesis forms the basis for three main directions of future research, related to (1) the process of bootstrapping pronunciation dictionaries, (2) grapheme-to-phoneme conversion, and (3) further refinement of the *minimal representation graph* framework.

The current bootstrapping process provides an effective platform for the development of pronunciation dictionaries but further gains are likely to arise from future improvements. Specific issues that we would like to address in future include:

- Active learning during bootstrapping: determining optimal ways in which to choose the next instance or set of instances to utilise during bootstrapping.
- An evaluation of the implications of different initialisation mechanisms, for example when a limited rule set is known prior to dictionary creation, or when a pronunciation dictionary exists in a phonologically similar language.
- Further analysis of the ways in which algorithmic requirements change for different phases of the bootstrapping process.
- Practical support for phone set manipulation during bootstrapping, including re-bootstrapping of appropriate sections of the dictionary after phone set manipulation.
- Support for the bootstrapping of other linguistic entities such as intonation, stress or hyphenation.

G-to-p conversion algorithms in general have been well studied, especially with regard to asymptotic accuracy and computational complexity. However, little work has been published to date in evaluating and improving initial learning efficiency (accuracy when trained on very small data sets) and robustness to noise (transcription errors occurring in the training dictionaries) – two aspects that are of importance during bootstrapping. We are interested whether further improvements may be obtained from the following sources:

- Adapting the algorithm (or its parameters) according to the specific grapheme extracted. As all rule extraction and rule application occurs on a per-grapheme basis, it should be possible to introduce further algorithmic refinements suitable to the characteristics of the specific grapheme being considered. We would like to analyse the current graphemic behaviour in further detail.

- Utilising this algorithm within a framework that includes additional data sources (such as part-of-speech tags).

- Learning from and predicting multiple pronunciations (incorporating word-level pronunciation variants).

- Incorporating class-based learning in the current algorithm: combining graphemes according to predictive behaviour in such a way that learning is accelerated.

- Investigating the threshold for valuable exceptions. In Section 6.4 it was clearly shown that the effect of errors in the training data tend to accumulate in the last $10 - 20\%$ of rules extracted. For *Default&Refine* specifically (and for noisy training sets) all exceptions may not contribute to predictive accuracy.

Some of the above questions related to grapheme-to-phoneme conversion may be better analysed in terms of the *minimal representation graph* framework. The current framework provides a theoretical basis for understanding the task of instance-based learning of rewrite rules. Further work related to this framework specifically include:

- Further development of the set of allowed operators utilising the framework, as well as a rigorous analysis of the legality and optimality of the set of operators.

- The application of established techniques related to the solution of constraint satisfaction problem, in order to improve the computational tractability of the current graph solution process. This will be required before a rigorous evaluation of the extracted rule sets on larger training dictionaries will become possible.

Additionally, the *Default&Refine* algorithm provides an interesting perspective on the grapheme-to-phoneme conversion task, viewing pronunciation as a hierarchy of regularity – with systematic instances and exceptions occuring in a continuum of regularity. We are interested in applying the same algorithm to other natural language processing tasks that exhibit similar behaviour.

7.4 CONCLUSION

This thesis has developed a number of tools in support of the bootstrapping process, and has demonstrated the value of this approach for the practical and cost-effective development of pronunciation dictionaries. Human language technologies have great potential value in the developing world, and bootstrapping will undoubtedly play a significant role in accelerating the development of such technologies. We therefore hope that theoretical interest and practical importance will continue to drive developments in this area.