Chapter 1

INTRODUCTION

People strive toward the future. People strive to find out about themselves. This thesis is hopefully a small step in those two directions. It is a work on natural language modeling and processing with connectionist models. The first goal is situated within language processing, the importance of which everybody would agree on in the digital era, where “communication” is of great importance. Connectionism, on the other hand, is a computational paradigm inspired by the way the humans think, and which is developed to discover the secrets of human intelligence.

Complex symbolic natural languages are one of the things most peculiar to people. Human languages have hierarchical structures, they can be represented in many ways, and they can ultimately express anything possible to imagine. Yet, all this would mean nothing to us if we did not produce and comprehend this language; without production and comprehension capabilities, it would be like the unlimited number of other representational systems, meaningless to us. People can produce and comprehend almost any expression in the language they speak, even though it may contain very complex meaning. Our nervous system does this in a very reliable way. Directly inspired by the structure of this system is a computational paradigm known variously as Neural Networks, Parallel Distributed Processing, or connectionism (Rumelhart, Rumelhart & McClelland 1986).

Yet, natural language modeling is mostly practiced with other, classical approaches which find their foundations in the natural languages themselves – symbolic methods. Those approaches usually perform very well when doing this job, but they do not explain the way humans process their languages; they provide no link to the subsymbolic architectures of the human brain. All they can suggest in that respect is the basic assumption that a neuronal
device capable of complex language processing and containing inherent linguistic knowledge – for example, Universal Grammar – has been developed during evolution and is now exploited. Connectionism goes further than this. Connectionist models only suppose a general neuronal substrate without specific (linguistic) bias, which can learn from the environment, that is, can adjust its memory in order to improve the human’s behaviour in that environment.

The main point in this thesis is to show that natural languages can be modelled and processed with connectionist models, at least at a lexical level. For this purpose, two basic lexical problems will be concerned. The first one is modeling the sound structure of the words, also known as lexical phonotactics, studied in Chapter 4. The other one is translating words from one representational modality (written forms) into another one (phonological forms), which is explored in Chapter 5. Those two problems do not exhaust at all the space of problems related to lexical language modeling. Yet, successful work on them contributes to our understanding of what kind of connectionist models might be used for lexical modeling, how do they do it, and what connectionist models might be used for other linguistic problems.

No doubt, studying complex language expressions is even more challenging work. To do this, however, proper connectionist tools are needed. Recognising this and the limitations of some of the existing models, Chapter 6 presents a model aimed at developing representations of sequential structures to be used for further complex processing. This model is called “Recurrent Autoassociative Networks” (RAN). It exploits two basic observations: Firstly, external data is mostly dynamic (varying temporally) but internal processing benefits in speed if it uses static representations of sequential data. The other observation is that most of the learning is based on repetition. This suggests the use of sequential autoassociation – repeating the just observed data in order to develop unique static representations of the input sequences. This approach results in one very useful side-effect: having both an encoder and decoder in the same computational device. Further, I suggest two directions of using the representations developed by this model. The first one is modeling hierarchy in languages, which is explicaded by a two-level lexical model. The other, very promising direction is holistic modeling, which is a step toward high-level connectionist symbolic modeling. Chapter 7 suggests some ideas how this could be done and provides some basic holistic operators.

This thesis targets two main groups of readers: (1) people with connectionist backgrounds aiming at cognitive language modeling, and (2) people with linguistic backgrounds willing to explore cognitively plausible methods.
for linguistic modeling. For that reason, the thesis will start by presenting background material in two introductory chapters: Chapter 2 will outline some issues related to natural languages and their modeling, and Chapter 3 will focus on connectionism and will present in detail the main connectionist model used in this thesis – the Simple Recurrent Networks, developed by Elman (1988). In addition, since this is a bridging work between different sciences, the main chapters will also provide extended explanations for easier comprehension.

I will further provide detailed introduction to the different problems outlined above locally in each chapter, since those problems have their own specificity and it is not necessary to reiterate this information.

Finally, I want to finish this introduction by remarking that, first, all the models presented here were implemented in my own programs, written in the programming language C++ under Unix, and next, that most of the work in this thesis has already been published in seven conference papers and one chapter of a book on Recurrent Neural Networks. Four are coauthored with my supervisor Prof. Dr. Nerbonne and/or Dr. Laurie Stowe (Stoianov, Nerbonne & Bouna 1998, Stoianov & Nerbonne 2000b, Stoianov, Stowe & Nerbonne 1999, Stoianov & Nerbonne 2000a), the others were written independently (Stoianov 1998, Stoianov 1999, Stoianov 2000b, Stoianov 2000a). This thesis presents that work systematically and contains some new material as well.
Introduction Recap:

- The main perspective is the Cognitive Science approach toward human cognition and in particular the language capacity. Theories of language will be judged with regard to their fitness to the brain structure and capacity.

- Main NL problems to be exploited: lexical processing; (1) lexical learning, (2) mapping from orthography to phonology, (3) developing representations of sequences, and (4) holistic computations.

- Main connectionist model: Simple Recurrent Network (Elman 1988), as an universal sequential connectionist device. I will use this model throughout the thesis, exploiting the principle of reusing existing models for different functions, as outlined by Reilly (1997).

- Main claim: Lexical learning with general connectionist models is possible (Chapters 4, 5).

- Main achievements: (1) Lexical learning with Simple Recurrent Networks (Chapter 4). (2) Modeling Grapheme-to-Phoneme Conversion of Dutch words (Chapter 5). (3) A connectionist model for general language modeling and holistic computations – Recurrent Autoassociative Networks (Chapters 6, 7).