Chapter 8

CONCLUSIONS

The chronology of the work presented in this thesis approximately follows the structure of the thesis. It all started in April 1997 with learning the phonotactics of the Dutch language with Simple Recurrent Networks (SRN), aiming at replicating and improving earlier unsuccessful experiments on the same problem, and having the prospective idea of using alternative models, but since SRNs were successful at this task, it was worth challenging them with even more complex lexical problems, going through GPC modeling, which in turn inspired the development of the RAN model, which develops holistic representations of words and sequences, in general. There remains the even more challenging idea of modeling syntax with SRNs, using those representations. The focus on lexical problems, however, was constant throughout the project.

Given that short introduction to the work presented in this thesis, in this final Chapter I will summarise the contributions claimed in the four central chapters 4 - 7, also discussing possible further extensions.

Phonotactics learning

One of the most important tools in computational linguistics is grammar. Grammars have rich representational power; they can model natural languages to some extent, and therefore they are widely used in linguistics to represent linguistic structures. Computational linguistics works with grammars by using high-level symbolic processors, such as Lisp, Prolog, etc. Humans process natural languages, and therefore they should also have some neurobiological devices which deal with languages. Natural languages, however, are very complex and therefore some linguists, such as Chomsky, claim that the neuronal module(s) that process languages should have in some form
a strong, genetically encoded, linguistic bias – so-called Universal Grammar – that facilitates the acquisition of languages, that is, language learning. This hypothesis is bold since currently no neurobiological evidence is available for the existence of such a complex bias. Therefore, it is an important question whether connectionist models – currently the closest models of the neurobiological substrate – can learn natural languages with no, or very little built-in knowledge. Given the rich representational power of the neural networks, and the experimental and theoretical works on their learnability, the general working hypothesis in this thesis is that NNs can indeed learn the structure of natural languages. This was the problem explored in Chapter 4, although limited to the lexical level only, that is, learning phonotactics, i.e., lexical grammar.

To test the hypothesis whether phonotactics is learnable by connectionist models, I first defined two general dynamic connectionist devices for language learning – neural transducers and neural predictors. When implemented with SRNs (Elman 1988) – a connectionist model proven able to represent regular languages (Kremer 1995, Alquezar & Sanfeliu 1995) – I conjectured that they are also able to learn regular languages, given two earlier works on the learnability of SRNs (Kuan et al. 1994, Arai & Nakano 2000). This result is of particular importance, given the claim that phonology and morphology can be described with regular languages (Kaplan & Kay 1994). Given all that, a neural predictor implemented with SRNs was trained here to learn the phonotactics of monosyllabic Dutch words, which display most of the phonotactic relationships of that language.

In spite of the above theoretical work on SRN learnability, during the experiments practical limitations hindered the networks from learning the lexical grammars perfectly. Hence, the second part of Chapter 4 focused on methods for evaluation of the trained networks, systematically exploiting variety of aspects of the knowledge learned by the network – context-dependent predictions, phonotactics, and lexical recognition. Similarly to other earlier works on phonotactics learning (Tjong Kim Sang 1995, Tjong Kim Sang 1998), it was found that in order properly to extract the abstract, “symbolic” knowledge of the network in the latter two problems, a specific signal – a threshold that distinguishes those phonemes which are allowed to combine with the previous sequence of phonemes – needs to be found and used. In contrast to these earlier works, however, which concluded that learning phonotactics is not feasible with SRNs because there is no such a threshold, here a specific value of the threshold – an optimal threshold – was found by optimising the predictions made so that it best distinguishes permitted successors from other phonemes. In particular, for the given mono-
syllabic Dutch dataset, a threshold of 0.0175 distinguished correctly 90% of the phonemes in the phonotactics task and at a threshold of 0.0160, the network on average recognised 95% of the words from the testing non-words. The optimal threshold was found here in a non-connectionist manner, but I suggested that it can also be tuned interactively, so as to minimise incorrect predictions.

Analysing the knowledge encoded in the network was also fruitful. A very interesting finding was, for example, that the network discovered without any prior knowledge some of the categories (features) that the linguists use to describe the phonemes. At a syllabic level there were also structural “findings” made by the network. The network displayed a very sharp “break” of the syllables at the point between the onset and the nucleus, which might be interpreted as relative independence of the onset from the nucleus, as compared to the relatively more predictable coda-consonants. This is to say that the syllable has a non-linear structure as given in (4.19), the same structure which Kessler & Treiman (1997) postulate for the English language.

Grapheme to Phoneme Conversion

The basic external representation of natural languages is speech. Language in humans, however, is grounded, or associated with all other cognitive modalities – visual, effectual, etc. Possibly this associative cognitive capacity is the basis of the human capacity to process other external language representations, such as written languages and sign languages. Starting from the basic hypothesis that general associative mechanisms should be able to convert such alternative representations into the “standard” phonological representations, Chapter 5 explored the possibility that SRNs – the same general connectionist mechanism that was earlier used to learn phonotactics – can also learn one of the basic steps in perceiving written languages, namely Grapheme-to-Phoneme Conversion (GPC). I choose SRNs since they were shown capable of representing any FSA (the so-called Mealy or Moore machines), plus the works on their learnability, as commented earlier. In addition, this model was also used for other sequential tasks. Therefore, embracing the idea of reusing existing models for other functions, with slight modifications, I suggested that SRNs should be able to learn this particular transformation, too.

Learning full-scale GPC of the Dutch language turned out to be a very challenging task, especially with a limited-size SRN. Starting from the classical approach at connectionist learning of any given task, which for the
GPC task resulted in some 90% score at a word level, and going through a number of steps analysing the difficulties of this problem, I ended with a model that could learn to transform all but 1% of Dutch monosyllabic words correctly. The same model, with a slightly larger hidden layer, could also learn to transform correctly some 93% of a dataset containing 10,000 polysyllabic words, which I also consider a significant success.

These results may have practical implications. In addition, better learning algorithms and larger networks might allow almost perfect learning of even larger polysyllabic corpora. But this is not the end of the story. I also claim that the models presented here have cognitive significance in terms of a plausible model of our capacity to transform incoming stream of abstract graphemic signs into a stream of abstract phonological representations which can further be transformed into a series of articulatory commands, or be associated with other abstract representations bearing semantics.

To support this claim, an analysis of the model’s performance with respect to various factors was presented, showing patterns of performance close to those shown by humans. However, this is just one part of the complex language faculty, and hence precise emulation to humans is still out of reach. There are simply too many other factors to take account of. Nevertheless, some very basic patterns, such as worse performance for infrequent words, or words with inconsistent pronunciation was very reliably shown. In addition, the performance of the model depended on dynamic parameters of the data, such as word length and phonemic position (offset).

Finally, the basic computational principles of distributed processing nicely modelled how one specific type of acquired reading impairment – so-called acquired surface dyslexia – might arise. This was shown by multiple damaging of a trained network with two types of damage – noise addition and neuron removal. These cases leaded only to mixed damages and surface dyslexia, which I explained with the basic transformation knowledge which is stably learned in a distributed manner, and more shallowly represented transformation of exceptional patterns, which are overlaid over the basic GPC patterns. I suggested that the other types of dyslexic performance can be modelled by including a semantic system in the model.

**Recurrent Autoassociative Memory and Holistic computations**

Natural languages are undoubtedly very complex and so is the system capable of processing them. The organisation of the languages across the world varies a lot, but people from any region can learn the language of any other region. Therefore, the human language processing system must
be quite flexible in order to be able to do this. These facts have led 
Chomsky to postulate very specific linguistic bias in human learning – 
Universal Grammar, which is claimed to be genetically encoded and tuned 
during early language acquisition. Alternatively to having such a complex bias, 
one might suggest using simpler models, rich in their representational ca-
pacity and capable of learning language and other complex signals just from 
experience. This is the global idea behind the model presented in Chapters 
6 and 7. It is an hierarchical model, based on modules that process se-
quential information of gradually increasing complexity. These models are 
sequential auto-associators – Recurrent Autoassociative Networks (RAN) – 
that develop static distributed representations of sequences by exploring one 
very basic learning mechanism – repetition – that is found across almost all 
species with nervous system. Actually, the more direct motivation for the 
development of the RANs was to build a model that develops static dis-
tributed representations of words, which may be used for further syntactic 
processing.

I implemented RANs with SRNs, although other models with global 
contextual memory can be used, too. This again demonstrates that ex-
isting models can be reused for other functions with slight modifications.

To develop a unique static distributed representation (DR) of an incoming 
stream of tokens, RANs initially process it sequentially, without having 
specific targets at the output layer. At the end of the input sequence, a 
specific distinct item (end-of-word) symbol appears at the input layer and 
is processed. This is the moment when a unique DR of the input sequence 
is developed at the hidden (and context) layer. Then, having built a DR of 
the input sequence, the network repeats this sequence in order at the 
output layer. During training, the network is only trained to do this. After 
that, however, the network indeed develops unique representations of every 
input sequence. This is due to the specific learning task, repetition of the 
input sequence, which necessitates unique DRs of the input sequences. Ex-
perimental work showed that the network is capable of learning this task 
quite successfully, for a variety of input sequences: datasets of Dutch words 
and a set of numbers represented as strings containing up to four digits. 
The network was also found to generalise well, provided rich training data. 
In particular, the experiments on the complete set of numbers showed that 
training on 20-30% of the whole data guarantees excellent generalisation. In 
addition, it was shown that the distributed representations are developed 
in a very systematic way, especially when more training data is used. This 
systematicity in turn allows further complex processing.

Having RANs as basic building blocks, and static representations de-
veloped in a systematic way, one can build a complex cascade system that receives at its lowest level sensoric input and gradually re-represents it at different levels. The RAN at each level in this cascade receives as input sequences of static representations developed by the RAN at the immediate lower level and provides input to the RAN at the immediate higher level. Such a cascade system works like the system of hands in a watch: in order the hour-hand to pass one hour, the minute-hand has to pass 60 minutes, and in turn each minute requires that the second-hand passes 60 seconds. This way, having a cascade of RANs, input sensoric data gradually produces even more (time-) complex representations at each level, the highest level producing most (time-) complex representations. On the other hand, if a given DR is set to the hidden layer of the RAN at a certain level, it can be decoded to perceptual/effectual level.

This is not everything, however. It is a very interesting possibility to use these representations for complex static processing – different types of transformations, which in Chapter 7 are called holistic computations. Holistic computations work on holistic static distributed representations of structures/sequences without prior decoding and consequent recoding. Holistic representations are distributed representations that belong to a holistic system and which are systematically developed. They represent the data as a whole, and therefore it is not necessary to decode them for processing the encoded data.

RANs produce a set of holistic representations – a holistic system that consists of DRs of all encodable and decodable sequences. Due to the systematic way of developing DRs, a RAN can produce a holistic system without experiencing (being trained on) all sequences corresponding to the DRs of the holistic system, which was shown with the experiments on learning DRs of numbers.

Holistic representations can be used for complex associations across or between different modalities. Development of such a complex system is beyond the goal of this thesis, and therefore in Chapter 7 only a few simple computations were investigated, demonstrating the capacity for larger processing.

**Directions for further development**

It was difficult to stop working on this project. Each of the main subjects in this thesis can further be explored.

Phonotactics was studied here from a computational perspective rather than from a linguistic one, and the linguistic analysis here was rather shal-
low. Hence, linguists can further analyse the knowledge developed by networks trained on phonotactics and discover structures and features 'developed' by the network. More theoretical and practical work on SRN learn-ability is also necessary. The BP/BPTT learning algorithms are endangered from falling into local minima and alternative approaches are welcomed. This is necessary for the more challenging task of syntax learning, possibly combined with the RAN model.

Grapheme-to-Phoneme Conversion also needs further explorations, especially learning polysyllables. From a cognitive point of view, there is also a lot of work to be done on matching human behaviour to the way the networks perform. In particular, psycholinguistic data for the Dutch language is necessary for precise matching.

I expect that most of the interest will be directed toward RANs, the RAN-cascade, holistic computations, and cognitive modeling with them. There are numerous possible extensions, and a lot of work for justification and understanding the RANs. Analysing the DRs developed is also very important with respect to improving the RAN so that it develops even better DRs, e.g., more discrete DRs. Other basic holistic computations and more challenging tasks, such as unification, holistic GPC, and holistic syntax parsing should be in the agenda for exploring in the years to come.

Perhaps the reader has also already pictured a number of practical applications of the work presented here. Phonotactics can immediately be used for lexical decision task in speech recognition, spell-checking, etc. GPC modeling is also interesting for automatic speech production systems and experiments with speech synthesis systems would be intriguing. The author would be gratified if the work presented here stimulated others to test and further develop neuro-computational linguistic models.