Antti Järvelin

Applying Machine Learning Methods to Aphasic Data

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Information Sciences of the University of Tampere, for public discussion in the B1097 Auditorium of the University on June 16th, 2008, at 12 noon.

DEPARTMENT OF COMPUTER SCIENCES
UNIVERSITY OF TAMPERE

A-2008-4
TAMPERE 2008

Supervisors: Professor Martti Juhola, Ph.D.
Department of Computer Sciences
University of Tampere
Finland

Professor Matti Laine, Ph.D.
Department of Psychology
Åbo Akademi University
Finland

Opponent: Docent Timo Honkela, Ph.D.
Laboratory of Computer and Information Science
Helsinki University of Technology
Finland

Reviewers: Docent Tapio Grönfors, Ph.D.
Department of Computer Science
University of Kuopio
Finland

Associate Professor Nadine Martin, Ph.D.
Department of Communication Sciences
Temple University
PA, USA

Department of Computer Sciences
FIN-33014 UNIVERSITY OF TAMPERE
Finland

ISBN 978-951-44-7348-7
ISSN 1459-6903

Tampereen yliopistopaino Oy
Tampere 2008

Electronic dissertation
Acta Electronica Universitatis Tamperensis 738
ISBN 978-951-44-7364-7 (pdf)
ISSN 1456-954X
http://acta.uta.fi

Abstract

This thesis aimed to study the inner dynamics of both normal and disordered word production using machine learning methods. A set of experiments where machine learning methods were applied to naming data was performed. The data was produced by aphasic and non-aphasic speakers in various aphasia tests. In this thesis two different approaches on applying these methods on aphasic data have been taken. In the first part, the efforts are concentrated on developing a computational model for simulating the actual cognitive naming process, i.e., lexicalization. Modeling lexicalization has both theoretical and practical benefits, as the models might provide new insight to the process of lexicalization and serve as a guide for treating aphasia. The latter part of this thesis explores the possibilities of applying machine learning classifiers to classify aphasic and non-aphasic speakers into groups based on their aphasia test results. This way, relationships between clinical aphasia syndromes could be identified from the classification results. Inconsistencies in the currently used aphasia classification system could also be revealed. On the other hand, these classifiers could be used as a basis for a decision support system to be utilized by clinicians diagnosing aphasic patients. Based on the results, it can be concluded that, when correctly applied, machine learning methods provide new insight to the spoken word production of aphasic and non-aphasic speakers. However, both application areas would greatly benefit from larger aphasia data sets available. This would enable more reliable evaluation of the models of lexicalization and classifiers developed for the data.

Keywords: Machine learning · Neural networks · Classification · Multi-layer perceptrons · Aphasia

Acknowledgments

I wish to thank my supervisors Professors Martti Juhola, Ph.D., and Matti Laine, Ph.D., for their support and guidance during the preparation of my thesis. Their work is greatly appreciated, and this thesis would not have been accomplished without their efforts.

I am also greateful to the reviewers of this thesis, Docent Tapio Grönfors, Ph.D., and Associate Professor Nadine Martin, Ph.D., who provided constructive and prompt comments on the manuscript.

My coworkers at the Department of Computer Sciences and at the Data Analysis Research Group have provided me a pleasant and inspiring working environment. Especially Jorma Laurikkala, Ph.D., and Kati Iltanen, Ph.D., have offered their advise whenever I have needed them. The department's administration has also been very helpful.

I wish to thank my friends for their support during this project. Particularly Rami Saarinen, M.Sc., Toni Vanhala, M.Sc., and Mr. Ossi Anttalainen have contributed to the completion of this thesis propably more than they know. The almost daily discussions with them on the most profound questions have provided me hours and hours of inspiration. For instance, I now know how to solve a difficult problem of serving coffee for four when the only tools available are one stone, scissors, and a paper – not enough for all, as a careful reader might notice.

I wish to thank my parents and sisters, who have offered their unconditional encouragement and support during the preparation of my thesis. This is the case also with my *suoceri*.

My studies were funded by Tampere Graduate School in Information Science and Engineering (TISE) and Academy of Finland (under grant # 78676), which are greatfully acknowledged.

Finally, I wish to thank Liina for her presence, love, and patience during this once in a lifetime introspective human experiment.

This thesis is dedicated to the memory of Ilmi Järvelin.

Tampere, May 2008
Antti Järvelin

Contents

List of Abbreviations

Abbreviation	**Description**
α_L	The noise parameter of the lexical-semantic network of the Learning Slipnet simulation model
α_P	The noise parameter of the phoneme network of the Learning Slipnet simulation model
τ	The threshold parameter between the lexical-semantic and the phoneme network of the Learning Slipnet simulation model
AAT	Aachen Aphasia Test
ACC	Classification Accuracy
AD	Alzheimer's Disease
BDAE	Boston Diagnostic Aphasia Examination
BNT	Boston Naming Test
GNT	Graded Naming Test
k-NN	k-Nearest Neighbor
MLP	Multi-Layer Perceptron
PALPA	Psycholinguistic Assessment of Language Processing in Aphasia
PNN	Probabilistic Neural Networks
PNT	Philadelphia Naming Test
PPV	Positive Predictive Value
ROC	Receiver Operating Characteristics
SOM	Self-Organizing Map
TNR	True Negative Rate
TPR	True Positive Rate
VaD	Vascular Disease
WAB	Western Aphasia Battery

List of the Original Publications

This thesis is based on the following five publications. In the text they are referred to by their roman numerals.

I. A. Järvelin, M. Juhola, and M. Laine. Neural network modelling of word production in Finnish: coding semantic and non-semantic features. *Neural Computing & Applications*, 15(2):91–104, 2006.

II. A. Järvelin, M. Juhola, and M. Laine. A neural network model for the simulation of word production errors of Finnish nouns. *International Journal of Neural Systems*, 16(4):241–254, 2006.

III. A. Järvelin, M. Juhola, and M. Laine. A neural network model of lexicalization for simulating the anomic naming errors of dementia patients. In M. Fieschi, E. Coiera, and Y.-C. J. Li, editors, *Proceedings of the 11th World Congress of Medical Informatics*, pages 48–51. IOS Press, 2004.

IV. A. Järvelin. Comparison of three neural network classifiers for aphasic and non-aphasic naming data. In L. Azevedo, and A. R. Londral, editors, *Proceedings of the First International Conference on Health Informatics*, pages 186–190. INSTCC Press, 2008.

V. A. Järvelin, and M. Juhola. Comparison of machine learning methods for classifying aphasic and non-aphasic speakers, *Computers in Biology and Medicine*, (submitted)

Reprinted by permission of the publishers.

Chapter 1

Introduction

Word production is a multistaged process where a speaker transforms a semantic representation of a target concept to its phonological representation and finally articulates it. The intricate word production system is also quite sensitive to impairment. In fact, cardinal feature of aphasia, a language disorder following left hemisphere damage, is anomia, a difficulty to find relevant words while speaking. Besides halting or empty spontaneous speech, anomia can be apparent in a confrontation naming task. Here the types of errors produced are of particular interest, since they may inform us about the underlying causes of the patient's aphasia and the functioning of the language production system in general. Commonly encountered error types include semantic errors (mouse → rat), formal errors (mouse → house), neologistic (nonword) errors (mouse → mees) and omission errors (patient says "I don't know", remains silent, etc.).

Machine learning methods can be applied to aphasic naming data in order to better understand the inner dynamics of both normal and disordered word production. In this thesis, two different approaches on applying these methods to aphasic data have been taken. In the first part of studies, the aim was to develop a computational model for simulating the actual naming process. Especially, the developed model simulated the most fundamental part of spoken word production, *lexicalization*, by which a speaker transforms the semantic representation of the word into its abstract phonological representation. Modeling the lexicalization process has theoretical and practical benefits, as the models might provide new insights to the lexicalization process itself and serve as a guide for aphasia treatment. The first part of the present work consists of simulations with the model, where both normal and disturbed lexicalization processes were simulated.

The research problems addressed in the first part of this thesis were as follows. First, suitable encoding techniques for semantic and phonological presentation of words were explored. This was a relevant problem, since to be able to utilize machine learning models and especially neural network models for simulating word production, both semantics and phonology of words must be presented in numerical form. Paper I presents one possible solution for this problem.

The second research problem was to investigate the suitability of the multi-layer perceptron (MLP) neural network architecture to form the basis of a model of word production. In Paper II, the properties of the developed model, Learning Slipnet, were investigated in detail. Especially, the performance patterns of the model's subnetworks were analyzed in order to gain insight to the model's behavior. The most intensive evaluation of the model against patient data was performed in Paper III, where the performance patterns of 22 Finnish-speaking dementia patients and 19 healthy control subjects were simulated with the model.

The latter part of this thesis explores possibilities of applying machine learning classifiers to classify aphasic and non-aphasic speakers into groups based on their aphasia test results. Different classifier types were tested and compared for the task, including various neural network classifiers, decision trees, naïve Bayes classifier, k-means classifier, and nearest neighbor classifier. The rationale of developing classifiers for this task is that classification might give more information on the relationships between different clinical aphasia syndromes, and especially, reveal inconsistencies in the currently used aphasia classification system. On the other hand, these classifiers could be used as a basis on the decision support system utilized by clinicians diagnosing aphasic patients.

The third research problem was thus to find out if certain types of machine learning classifiers would be especially suitable for classifying aphasic and non-aphasic speakers. The problem was investigated by comparing classifiers on three different aphasia data sets. In Paper IV, the classification performance of three neural network classifiers were compared using one aphasia data set. As the results suggested that also other very simple classifiers, such as discriminant analysis classifier, might perform well with the used data set, additional evaluation of classifiers was performed in Paper V. Here eight different machine learning classifiers were compared using three aphasia data sets.

The rest of the introductory part of this thesis is organized as follows: First, the application area is introduced in Chapter 2, including topics such as neuropsychology of spoken word production and aphasia. Chapter 3 gives an overview of machine learning with emphasis especially on classification. In Chapter 4, the roles of the individual papers in this thesis are presented, and Chapter 5 provides discussion and conclusions.

Chapter 2

Aphasia

In this Chapter, the general background for spoken word production – especially for single word production – is first briefly addressed. Then the nature of aphasia is discussed with special interest in word finding difficulties (anomia) which is the most pervasive symptom of aphasia. After this the most important aphasic syndromes are introduced. The aphasia tests applied in diagnosing aphasia are also described, as these tests were used to collect the data that was used in all the research papers of this thesis. Finally, the aphasia rehabilitation methods are briefly addressed.

2.1 Neuropsychology of Spoken Word Production

Word production is a multistaged process where a speaker transforms a semantic representation of a target concept to its phonological representation and finally articulates it. The inner store of words in an adult (the mental lexicon) consists of tens of thousands of words. Nonetheless, a healthy person can select a correct form in less than a second without apparent effort while speaking.

Language-related functions emerge from the structure and functions of the brain. The brain is not homogeneous mass, but different brain areas serve different purposes [20]. Although higher mental functions are not strictly localizable in specific regions of brain, certain brain areas are nevertheless more important for language-related functions than others [38]. The most important brain areas related to language functions are located in the anterior and posterior parts of the left hemisphere [38]. Of particular importance for

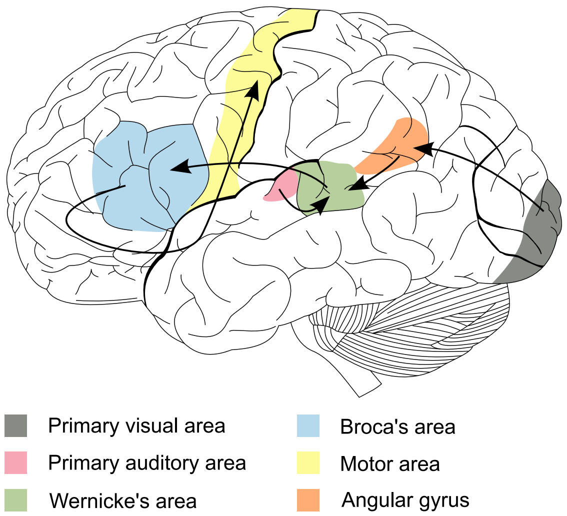

Figure 2.1: Left hemisphere of the human brain with the most important language processing areas highlighted. In this schematic view, the arrows represent the assumed flow of information in the brain when (a) repeating a word (information flow stars from the primary auditory area), and (b) naming a visual object (information flow starts from the primary visual area).

language are the so-called Broca's and Wernicke's areas (see Fig. 2.1[1]). The functions related to production of speech are located in Broca's area, whereas Wernicke's area hosts functions related to phonological skills [20]. These areas are interconnected via subcortical pathways, which enable, for example, effortless repetition of heard words. These "core regions" are connected to other brain areas to enable e.g. links between linguistic and conceptual representations as well as goal-directed linguistic behavior.

Laine and Martin [38] summarize the cognitive processing stages involved in word production as follows. When e.g. naming a picture of a familiar object, less than a second is needed to retrieve

1. sensory qualities of the visual object,
2. its meaning,
3. the corresponding phonological output form,

[1] Fig. 2.1 is based on the figure http://commons.wikimedia.org/w/index.php?title=Image:Brain_Surface_Gyri.SVG&oldid=9338871 published under the Creative Commons Attribution-Share Alike license version 3.0 (see http://creativecommons.org/licenses/by-sa/3.0/). To comply the license terms, Fig. 2.1 is hereby made available under the same license by the author of this thesis.

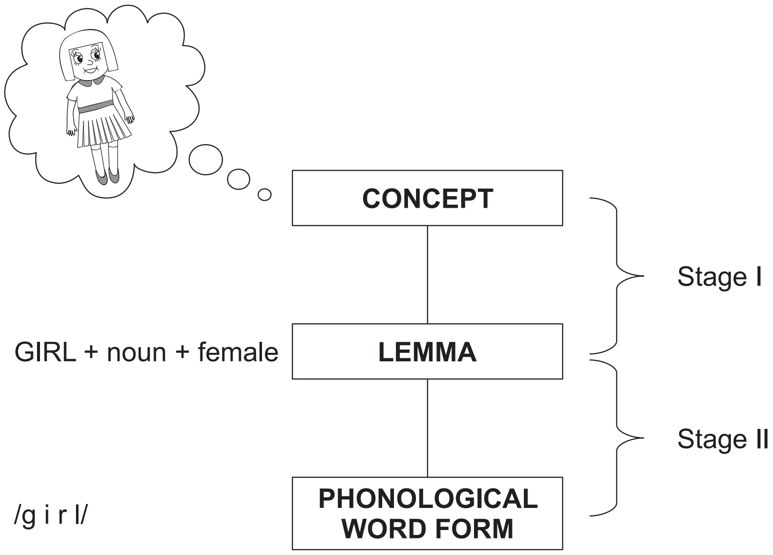

Figure 2.2: The lexicalization process. At the first stage a speaker transforms the semantic representation of the target word into an intermediate representation (i.e. lemma) which is during the second stage transformed into a phonological representation.

4. the syllabic and metric structure of the to-be-produced word, and
5. the phonetic-articulatory program needed for saying the word aloud.

They also note that at each processing stage, many mental representations become activated, even if only a single word will be produced. Furthermore, it seems that semantic and phonological processing are not independent. Although semantic information must be accessed before corresponding phonological information can be activated, there is strong evidence that these two processes overlap and interact with each other. [38]

Stages 2–4 in the description given by Laine and Martin [38] correspond to the two major levels of lexicalization depicted in Fig. 2.2. At first the conceptual representation is transformed into a lexical-semantic representation called lemma which contains syntactic and semantic information about the target word. After this the corresponding phonological representation of the target is retrieved. There are two major theoretical views on the lexicalization process. The advocates of the discrete two-step theory of lexicalization propose that the two processing stages are completely distinct. In their view, at the first stage only one lemma is selected and fed forward to the second stage [42, 43]. Proponents of the interactive activation theory of lexicalization claim the opposite: the two processing stages interact with each other and all activated lemmas may also become more or less phonologically encoded [9]. There exist also theoretical views that are somewhere between the highly discrete and the highly interactive account (e.g. [16]). Currently it seems that the interactive account is more accurate (i.e., is supported by majority of studies) than the highly discrete one [4].

2.2 Aphasic Disorder

2.2.1 The Nature of the Disorder

By definition, aphasic patients have either completely or partially lost the ability to read, write, speak, or understand spoken language [20]. Therefore, problems in language usage that are caused by paralysis, lack of coordination of muscles involved in language production (such as articulatory muscles), or poor vision or hearing are not aphasic *per se*, but may accompany aphasia [17].

Anomia, a difficulty to find highly informative words, is clinically the most common symptom of language dysfunction [38], as the majority of aphasia patients suffer from at least some degree of anomia [55]. Anomia is also the most frustrating and depressing symptom of aphasia, since it has devastating effects on patients' ability to carry on meaningful and effective conversation [55, 60]. Although almost all aphasic patients have limited vocabulary, the ability to produce memorized or automatic sequences, such as numbers, months, alphabets, or nursery rhymes is often preserved [17].

Virtually everyone has experience on occasional slips of tongue or naming difficulties, but the frequency of these difficulties is considerably higher for aphasic patients. In addition to a higher frequency of naming errors, the patients' error type distribution also differs from the distribution of a healthy person, as anomia can result from disorder in semantic processing, with semantic errors dominating the error distribution or phonological processing, with phonological errors dominating the error distribution. However, it should be noted that presence of the semantic errors do not necessarily entail semantical level disorder, as besides semantic errors it would also require a documented comprehension disorder. [38]

Laine and Martin [38] provide a more detailed classification of the most common naming errors encountered with aphasic patients. The phoneme level errors include

- phoneme substitutions (bat → *lat)[2],

[2] Here * refers to a grammatically incorrect word form.

- insertions and deletions (ginger → *gringer, drake → *dake), and
- phoneme movements (candle → *cancle, candle → *dancle).

The word level errors consist of

- semantic substitutions (elbow → knee),
- so-called formal errors (ankle → apple), and
- mixed errors (penguin → pelican)[3].

The fact that word level errors include both semantic and phonological errors suggest that word production is performed in two phases. Furthermore, of the word level errors, the mixed errors have received particular research interest, since they seem to occur more often than one would expect if semantic and phonological errors have totally independent sources [38]. This observation is one of the key evidence of the interactivity between the semantic and phonological processing during the lexicalization. This is one example of the value of the speech errors produced by normals or aphasics in the study of spoken word production.

2.2.2 Major Aphasic Syndromes

Goodglass and Kaplan [17] give a characterization of the major aphasic syndromes. Here a short review of the four major aphasic syndromes, Broca's aphasia, Wernicke's aphasia, anomic aphasia, and conduction aphasia is given based on Goodglass and Kaplan.

Many symptoms of language disorders occur seldom in isolation, but together with other symptoms of language dysfunction. The co-occurrence of the symptoms has given rise to the traditional syndrome approach to aphasia. The existence of more or less specific symptom complexes after localized left hemisphere lesions suggest that certain language functions rely on certain brain areas. However, due to the prominence of mild and severe aphasic patterns (and not the moderately impaired patients) in hospital populations, ca 30 – 80 % of patients are classifiable into the major clinical aphasia syndromes. The figure varies considerable also due to different diagnostic criteria employed. Furthermore, because of the individual differences of the functional organization of the brain, lesions to the same brain area may cause different symptoms, which further complicates the classification of clinical aphasia.

[3] Mixed error is an error that is both semantically and phonologically related to the target word.

The aphasia types can be divided into two main classes based on the fluency of the speech. Non-fluent speech is the result of damage in left anterior regions (including Broca's area) and is characterized by abnormally short utterances, effortful output often coupled with dysarthria (a motor speech disorder characterized by poor articulation). The limited utterances may nonetheless include many words with high information value. This kind of patient, labeled as Broca's aphasics, would typically have rather well preserved auditory comprehension. Degree of anomia in confrontation naming task may vary.

The most common fluent aphasia type, Wernicke's aphasia, usually results from a lesion in Wernicke's and adjacent areas. A typical symptom of Wernicke's aphasia is very weak auditory comprehension, most strikingly occurring even at word level. Symptoms include also fluently articulated but paraphasic speech including phoneme level changes and word level errors. The patients typically suffer from severe naming difficulties.

Anomic aphasics' main problems are word-finding difficulties. Their speech is usually fluent and grammatically correct, but hard to follow due to missing content words (nouns). Anomic aphasia differs from Wernicke's aphasia in that paraphasias may be missing and the auditory comprehension is at the normal level. Although anomic aphasia is frequently associated with angular gyrus lesion, it is the least reliably localizable of the aphasic syndromes.

Conduction aphasia is characterized by difficulties in repetition of (written or) spoken language, although the fluency of the speech and auditory comprehension can be almost at the normal level. In speech production task, patients produce numerous phoneme level changes which they are usually aware of, and hence reject words containing these changes. The more complex/longer the word is, the more likely it becomes phonologically distorted.

Besides the major aphasia syndromes discussed above, there are also other aphasia subtypes. These include transcortical aphasias where repetition is well preserved, global aphasia where all language related functions are severely disturbed, and various pure aphasias, where only one specific language component, such as reading, is disturbed.

2.3 Clinical Diagnosis and Treatment of Aphasia

2.3.1 Aphasia Tests

To be able to systematically analyze and compare patients' linguistic capabilities, standardized aphasia examination procedures are needed. Although the linguistic capabilities of the patients may considerably vary from day to day at the acute phase, they become more predictable after the initial spontaneous recovery [17]. The stability of the symptoms is a prerequisite for reliable testing. According to Goodglass and Kaplan [17] aphasia tests can be used for the following three general aims:

1. diagnosis of the presence and type of aphasic syndrome, leading to inferences concerning lesion localization;

2. measurement of the level of performance over a wide range, for both initial determination and detection of change over time;

3. comprehensive assessment of the assets and liabilities of the patient in all language areas as a guide to therapy.

Many standardized aphasia examination procedures addressing one or more of the three aims exist today, the most prominent ones being the Boston Diagnostic Aphasia Examination (BDAE) [17], the Western Aphasia Battery (WAB) [30], the PALPA (Psycholinguistic Assessment of Language Processing in Aphasia) [29] (in English speaking countries), and the Aachen Aphasia Test (AAT) [24] (in German speaking countries).

Aphasia examinations commonly begin with a free interview of the patient in order to obtain an overall impression of the patient's linguistic abilities [17]. Usually aphasia tests address different parts of the language production system in dedicated subsections, such as object naming, comprehension, or repetition. Several input or output modalities are often used to test the same linguistic domain in order to exactly specify the nature and the reason of the patient's symptoms [17]. For example, the patient's comprehension skills might seem to be impaired when tested with auditory stimuli, but prove to be intact when tested with visual stimuli. In this example it is probable that instead of e.g. a central semantic impairment, the patient's auditory input system is damaged, which might not have been evident if only auditory stimuli had been used to examine the patient.

With regard to naming that is at issue here, it is most commonly assessed by a visual confrontation naming task where a subject is shown pictures of simple objects that they should name [38]. Confrontation naming task is also a sensitive probe for a language disorder, as practically all aphasic patients suffer from anomia [17]. Furthermore, in contrast to many other subtasks of aphasia tests, confrontation naming is rather well controlled situation, where all the main stages of word production have to be activated and accessed [37, 38]. Thus, the confrontation naming task may more clearly reveal the underlying mechanism and the nature of a patient's lexical deficit than, e.g., the analysis of free speech would [6, 9].

There are various confrontation naming tests in use, the Boston Naming Test (BNT) [28] probably being the best known and the most widely utilized [38]. The original English version of the test was first published in 1983 and it has since been adapted into several other languages, including Spanish [15], Korean [31], Swedish [77] and Finnish [36]. BNT consists of 60 line art drawings of objects of various frequency range, which are presented to the patient in an increasing order of difficulty. In Fig. 2.3 four example pictures from the Finnish version of the Boston Naming test are presented. The test is sensitive to relatively mild word-retrieval problems that may appear in variety of neurological conditions, like beginning dementia or developmental language disorders [36]. Other well known naming tests include the Graded Naming Test (GNT) [88], and the Philadelphia Naming Test (PNT) [61].

Although standardized aphasia tests are highly valuable tool at the clinic, Goodglass and Kaplan [17] also note the limitations of such tests. First, the aphasia tests always represent only a small sample a subject's linguistic skills. Secondly, the test scores do not objectively or automatically result in a correct aphasic syndrome classification nor suggest the optimum approach to therapy. Therefore, examiner's personal knowledge and experience is always needed for the interpretation of the test scores and the actions that these results would give rise to.

2.3.2 Aphasia Treatment

The interest in aphasia treatment rose in the first half of the 20th century, and especially after the second world war with the rehabilitation of war veterans [38]. Majority of the aphasia treatment methods developed during the last 100 years have been behaviorally based. According to Nickels [55], the pharmacological treatment of aphasia has only lately started to show some promise, but the treatment seems to be most effective when combined with behavioral language therapy. Therefore, behavioral language therapy will have a central role in aphasia treatment also in the future.

Laine and Martin [38] recognize three approaches to the behavioral language therapy: restoration, reconstruction, and compensation. The advo-

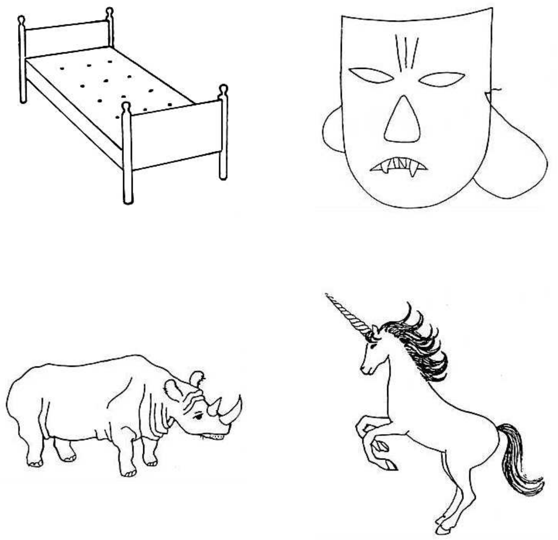

Figure 2.3: Example pictures from the Finnish version of Boston Naming Test in increasing order of difficulty from left to right and top to bottom.

cates of the restoration approach state that one should to rehabilitate the injured parts of the language production system and in that way try to regain the lost language capabilities. The supporters of the reconstructionist view on the other hand, state that the brain could replace the damaged parts of a functional system with new areas adopting the functions of the damaged ones. In this view, the lost language capabilities are regained through reorganization.

In the third, compensatory approach the patient is taught alternative means to bypass the damaged language components by taking advantage of the patient's intact language processes. For example, the patient could be instructed to use the written form of the word to help retrieve the spoken form [55]. Using such a technique, of course, requires that the patient's reading and writing skills are better preserved than the oral skills. Laine and Martin [38] note that the different approaches to behavioral language therapy are not mutually exclusive, and that compensational strategies can be used in tandem with restoration and reconstructionist approaches.

Recovering from a brain damage is a complex process involving physiological, psychological and psychosocial modifications [38]. If the onset of brain damage is sudden such as in a cerebral stroke, most of the spontaneous recovery takes place during the first weeks or months after the onset [35]. After an initial spontaneous recovery (re)learning of the lost language skills plays a major role in the further recuperation of a patient [35]. However, relatively little is known about the relearning process and the physiological diagnosis does not tell what rehabilitation method would be best suited for the patient [35]. Although it is sometimes possible to infer the suitable treatment method for a patient from the functional location of the patient's damage, the results do not necessarily generalize well, and aphasia rehabilitation procedures are not effective in all patients [55]. In anomia treatment, there are case studies indicating that for semantic impairment, semantically driven treatment, and for phonological level disorders phonologically driven treatment is the most effective method. However, contrary effects have also been reported [55]. As the relationship of functional damage and suitable treatment method is unclear, connectionist models have been suggested for simulating the phenomenon [35]. Because both restorationist and reconstructionist view to language therapy postulate plasticity of brain, the connectionist models suit especially well to this simulation [38].

Chapter 3

Machine Learning

This chapter gives an overview of the field known as machine learning. First a short introduction to the topic is given and then different learning strategies that can be used in machine learning are briefly presented. Finally, the processes involved in applying machine methods into real word problems are reviewed.

3.1 Definition

Machine learning refers to the field of how to construct computer programs that automatically improve with experience. It is inherently a multi-disciplinary field including influences from artificial intelligence, computational complexity theory, philosophy, psychology, and statistics. Machine learning methods have been applied to many application areas, such as game playing, natural language processing, and various medical domains. Machine learning methods are especially prominent in data mining, i.e., the search of patterns in large data sets. [53]

Although, as noted by Minsky [51], there are too many notions associated with "learning" to justify defining the term in a precise manner, in context of machine learning the term can be defined in a more restricted manner. Thus, adopting the definition of Mitchell [53] learning in this context can be defined as follows:

> A computer program is said to *learn* from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

In other words, a learning program can use the records of the past as evidence for more general propositions [51].

The most common applications of machine learning methods include *classification* and *prediction*. Classification is the process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the classes of unseen objects with unknown class labels [18]. That is, classification can be seen as predicting categorical labels for the unknown objects. Examples of well known machine learning methods for classification include decision trees [59], artificial neural networks [22], and various clustering algorithms [19], but also many other methods exist. Prediction, on the other hand, refers to the situation where missing or unavailable data values of continuous-valued functions are estimated [18].

Following the above definition of learning, a learning problem of classifying aphasic patients could be specified as follows:

- Task T: recognizing and classifying aphasic patients based on their confrontation naming test results.
- Performance P: percent of patients correctly classified.
- Experience E: a database of confrontation naming test results with given aphasia classifications.

Besides learning, applying machine learning methods to a real world problem includes many additional tasks that need to be concerned. To be able to better introduce these tasks, the following description is focused on classification, as classification related methods are used in the papers constituting this thesis. However, first a short overview of neural networks is given, since various neural network methods have been applied in this thesis. Then the credit assignment problem is introduced, as it provides the basis to a learning process, and after that a very high level review of the learning strategies applied in machine learning is given. Finally, the design process of machine learning classifiers is briefly illustrated. It includes issues like data collection and preprocessing, training the classifier, evaluating the learning output, and selecting suitable classifier for the task.

3.2 Neural Networks

The research of artificial neural networks (neural networks hereafter) has been motivated from the beginning by the fact that brains work differently than a digital computer [22]. Like human brains, neural networks are composed of simple units, (artificial) neurons, that are connected to each others with weighted connections. Each neuron can only evaluate a simple function based

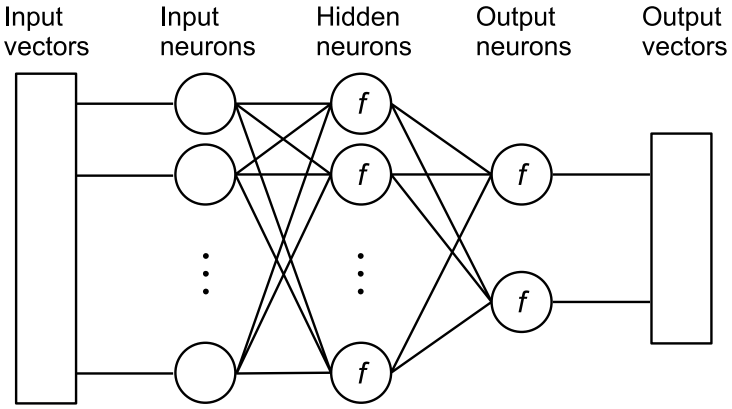

Figure 3.1: An example of MLP neural network with one hidden layer between the input and output layers. The circles represent neurons and the lines the weighted connections between the neurons. Each neuron in the hidden and the output layer calculates its output with function f. The input neurons only transmit the input vector to the neurons of the hidden layer. The information flows in the network from left to right (from the input neurons to the output neurons).

on the inputs it receives and then send the result to other neurons. The complex behavior of the network arises from the interaction of the individual neurons. Usually the neurons in the network are arranged into the layers. The layer connected to the input patterns is called input layer, and the layer, from where the results of the network are read, is called output layer. Often there are one or more hidden layers between the input and output layer as with some neural network types, such as multi-layer perceptrons (MLP), hidden layers increase the computational power of the network [22, 62]. In Fig. 3.1 an example configuration of an MLP network is given.

The first neural network models were presented in 1943 when McClulloch and Pitts [48] published their model for an artificial neuron, which worked as a binary decision unit. Their model was extended in 1960s by Rosenblat [63, 64] with connection weights between the neurons, which resulted in the creation of perceptron neural networks. However, Minsky and Papert [52] analyzed the perceptron model in detail, and showed that without a hidden layer, perceptron was unable to learn non-linear problems, such as exclusive-or-problem. This had a big impact on the interest in the neural network research, because with the perceptron learning rule it was not possible to train networks with hidden layers. The problem was not solved until in 1986 when the back-propagation learning rule for multi-layer perceptrons was popularized by Rumelheart and his colleagues [65], which enabled training networks with hidden layers. This was a major boost for the neural network research, and since the mid 1980s various neural network architectures have been introduced, the best known being MLP networks [22] and self-organizing maps (SOM) [32].

3.3 Learning Strategies

3.3.1 The Credit-Assignment Problem

When learning to play a complex game, such as chess or checkers, one has a definite success criterion: the game is won or lost. However, the result of the game depends on a vast number of internal decisions which are implemented as moves. If the result of the game is successful, how can these individual decisions be credited? The problem can be very difficult, since a game may be lost even if the early moves of the game were optimal [53]. This problem of assigning *credit* or *blame* to the individual decisions made during the game is known as the Credit-Assignment Problem and was formulated by Minsky in [51].

For a machine learning system, the credit-assignment problem is the problem of assigning credit or blame for the overall outcomes to each of the internal decisions made by a learning system which contributed to those outcomes [22]. The learning algorithms are then designed to solve credit-assignment problems arrosen from the specific machine learning model. For example, with MLP neural networks, the structural credit assignment problem is solved by the back-propagation algorithm [22].

3.3.2 Supervised, Unsupervised, and Reinforcement Learning

Learning paradigms can be divided into supervised, unsupervised, and reinforcement learning. The difference between the paradigms is the availability of the external teacher during the learning process. The supervised learning is characterized by the availability of the external teacher having knowledge about the environment in which the machine is operating and how the machine should correct its behavior in order to perform better in the future [22]. The limitation of supervised learning is that without the teacher, the machine cannot learn new knowledge about the parts of the environment that are not covered by the set of examples used during the training of the machine [22]. Examples of supervised machine learning systems include MLP neural networks [22] and decision trees [59].

Unsupervised learning is used for a given input, when the exact result that the learning system should produce is unknown [34, 62]. The practical applications include various data visualization or clustering tasks where the actual class distribution of the data is unknown or the relations between the classes are investigated. Examples of unsupervised machine learning systems are various clustering algorithms, such as k-means algorithms (e.g. [19]), and some neural network types, such as SOM [32].

Reinforcement learning bridges the gap between supervised and unsupervised learning [34]. In reinforcement learning the machine receives only criticism regarding whether or not the responses of the machine are desirable in the environment [51]. Based on the criticism the machine must infer *how* it should correct its behavior [22]. One of the best known reinforcement learning algorithm is Q-learning algorithm [53].

3.4 The Machine Learning Process

3.4.1 Data Representation and Preprocessing

For machine learning purposes, and especially for classification, the data are usually presented as an $n \times p$ data matrix. The rows of the matrix contain n *cases* or *examples* and the columns p *attributes* or *features*, whose values were measured for each case [19]. The cases might be n different aphasia patients, for example, whose naming confrontation performance is recorded as p different error types, such as, number of semantic or phonological errors. The attribute whose class is predicted with classification algorithm is called *class attribute* [19]. To illustrate, Table 3.1 gives an excerpt of a data matrix describing the naming performances of aphasia patients tested with Aachen aphasia test.

The data matrix presented in Table 3.1 contains six cases and eight attributes. The first attribute (diagnosis) is class attribute for which the classification is to be performed. The other attributes are disease, which is the clinical reason for the onset of aphasia, and six attributes P0–P5, which describe patients performance in one subtest of the AAT (spontaneous speech). P0 measures communicative behavior, P1 articulation and prosody, P2 automatized language, and P3 to P5 semantic, phonetic, and syntactic structure of language, respectively [24]. They are measured with scale from 0 to 5, with 0 meaning severely disturbed and 5 normal performance.

At top level attributes can be divided into categorical and quantitative attributes [19]. Quantitative attributes are measured on a numerical scale

Table 3.1: An excerpt of PatLigth aphasia data set describing the results of aphasia patients (the rows of the table) in Aachen aphasia test. The full data set can be browsed in the Internet at http://fuzzy.iau.dtu.dk/aphasia.nsf/PatLight.

Diagnosis	**Disease**	**P0**	**P1**	**P2**	**P3**	**P4**	**P5**
Anomic	ischemic stroke	3	4	5	3	4	4
Broca	ischemic stroke	2	2	3	3	2	2
Conduction	No information	3	5	5	4	2	3
Wernicke	intracranial haemorrhage	1	5	3	2	2	3
No aphasia	ischemic stroke	3	2	5	5	5	5
Undecided	rupture of aneurysm	4	4	5	4	4	4

and can, at least in theory, take any value. They can be divided into two sub categories: interval and ratio scale attributes. Ratio scale attributes have a fixed origin and can be multiplied by a constant without affecting the ratios of the values. With interval attributes the origin is not fixed, but they can still be multiplied by a constant.

Categorical attributes, on the other hand, can take only certain discrete values. Categorical attributes can be further divided into nominal and ordinal attributes. Ordinal attributes posses some natural order, such as the severity of a disease, but nominal attributes simply name the categories and it is not possible to establish any order between the categories [19]. Diagnosis and disease attributes of Table 3.1 are examples of nominal attributes. Attributes P0 to P5, instead, are ordinal attributes, because they can be meaningfully ordered based on the values the attributes can take. In the example data set there are no quantitative attributes present, but a patient's age, had it been recorded, would be an example of a quantitative ratio scale attribute.

The data sets used in machine learning are often incomplete, as they may contain missing values, measurement errors (noise), or human mistakes [18, 19, 78]. For example, in the above data set, the value for disease attribute of conduction aphasic is missing. The data might also come from multiple sources, which have different scales for encoding the attributes. Han and Kamber [18] and Hand et al. [19] introduce many techniques that can be used to preprocess data. These include data cleaning, data integration, data transformation, and data reduction. Using data preprocessing can significantly improve classifier's performance and preprocessing techniques are thus briefly discussed.

Data Cleaning

Data cleaning process includes filling in missing values, and removing outliers and noise [18]. It also includes correcting inconsistencies in data, such as inconsistent use of values for coding date (e.g. 29/07/1978 vs. 1978/07/29) [18]. Many classifiers cannot deal with the missing values in the data, and therefore the problem needs to be addressed before using the classifier. If large amounts of data are available for training the classifier, then it is possible just to ignore cases containing missing values [78]. Cases need to be also ignored if the missing value happens to be the class attribute [18]. As often the amount of available data for training the classifiers is limited, missing values can be filled in manually by an expert or some heuristic can be used instead, if the data set is too large for manual inspection [18, 78]. The heuristic approaches to filling in the missing values include replacing all missing values with a global constant, using the mean or class mean of the attribute as a replacement, and using machine learning techniques to predict the most propable value for the missing values of an attribute [18].

Outliers can cause problems for many machine learning algorithms as outliers can misguide the learning and thus obscure the main point of the classifier [19]. Again, if the number of outliers is very small, they can simply be discarded. On attribute level, outliers can be recognized by using statistical analysis on the attribute that is being investigated. For example, if the attribute is normally distributed, then distance of two times of standard deviation covers 95 % of the values. The remaining 5 % can be treated as outliers and removed [18, 78]. Other statistical methods, such as histograms and boxplots, can be used for outlier detection as well [19].

Outliers can also be processed using binning or clustering [18]. In binning the outliers are smoothed, by sorting the values of the attributes into bins, and then replacing the values with the bin means. Other option is to smooth with bin boundaries, where each value of the bin is replaced with the closest bin boundary value. Binning can also be used to remove noise from data. Clustering can be used in outlier detection by first clustering the data and then calculating the cluster centroids for each cluster. The outliers can be then detected as values that are far from any cluster center [18].

Data Integration and Transformation

Data integration refers to merging of data from multiple data sources [18]. Examples of problems that might occur while merging two data sets include the entity identification problem, data redundancy, and detection and resolution of data value conflicts [18]. Entity identification problem refers to recognizing attributes encoded with different names, but which actually represent the same concept. Meta data, if available, can be used to resolve problem.

An attribute may be redundant if it can be derived from another attribute or set of attributes [18]. Redundancy can also be caused by inconsistent naming of attributes. Correlation analysis can be used to detect some data redundancy. Detection and resolution of data value conflicts is an important issue in data integration, since failing to do so might result in inconsistencies in the data, and thus significantly decrease data quality. An example of data value conflict is an income attribute where the values are measured as Euros in one data set and US Dollars in the other.

The data may also need to be transformed into more suitable forms for the classifier. Techniques that can be applied to data transformation include smoothing, aggregation, generalization, normalization, and attribute construction [18]. Smoothing removes noise from the data and might thus improve the data quality. Techniques like binning and clustering can be used for this purpose [18]. Summarizing or aggregating data over several variables is called data aggregation. An example of data aggregation would be aggregating monthly income data of a person to annual total income. Generalization techniques can be used to transform low level data into higher-level concepts, such as transforming numeric age attribute to higher-level concept, like youth, middle-age, and senior [18]. Normalization can be used to transform attribute values to fall into certain range, like within range $[0, 1]$. This technique can be useful if a machine learning algorithm expects the values of attributes to fall in some specific range. In attribute construction, new attributes are constructed from the existing attributes to improve the understanding of highly dimensional data.

Although data transformation techniques provide a way to improve classifiers' performance, the transformations might also introduce new structures that are artefacts of the used transformation [19]. Domain expert's knowledge should be used to discover these artefacts, and the artefact structures should be rejected. Data transformation may also lose information about the original data, and should thus be used with care [19].

Data Reduction

Data reduction techniques can be used to obtain a reduced representation of the data set, which has much smaller size than the original data set, but still maintains the properties of the original data [18]. According to Han and Kamber [18] data reduction contains the following subtasks: Data aggregation, attribute subset selection (feature selection), dimensionality reduction, numerosity reduction, discretization, and concept hierarchy generation. The data reduction techniques that were relevant for the current study are attribute subset selection and dimensionality reduction, and are thus described in the following.

Attribute subset selection can be used when a data set contains a large number of attributes, some of which are redundant or completely irrelevant for the classifier. Properly selecting the relevant attributes for the classification task can improve classifiers' performance. Also, dropping out redundant attributes reduces the computational costs needed for training and using the classifier. There are many techniques to find good subsets of attributes, some of which are described in [18, 78]. These techniques often include using correlation analysis or tests of statistical significance, such as Student's t-test, to find out which attributes are independent from one another [18, 78] although also other techniques exist.

Even after a suitable subset of attributes has been selected, dimensionality reduction techniques can be used to squeeze down the size of a data set and reduce the computational costs of a classifier, especially by removing redundancy from the data [78]. Dimensionality reduction methods can be either lossy or lossless [18]. With lossy methods some information of the original data set is lost during the transformation, and the original data set cannot be reconstructed from the transformed data set, whereas with lossless methods this is possible.

3.4.2 Training the Classifier

When a suitable classifier for the classification task has been selected, training of the classifier has to be addressed. Methods for selecting a suitable classifier for a given classification task are addressed in Section 3.4.4 and thus here only some general remarks of the training procedure are made.

Training of the classifier includes searching suitable parameter combinations for the classifier and its training algorithm. For some classifiers the task is easier than for others. For example, the k-nearest neighbor classifier [10, 19] requires setting only the number of nearest neighbor value k and selecting a suitable proximity measure that is used during the classification.

On the other hand, with neural networks, such as MLP neural network [22] or SOM [32], many parameter values have to be set before actual training of the classifiers. These include setting first the network parameters, such as the number of hidden neurons of MLP network, or number of neurons and their organization in a SOM network. After the network parameters have been set, various parameters regulating the behavior of the learning algorithm have to be tuned in order to effectively train the networks and ensure their good generalization outside the training set. However, not all neural networks are demanding with this respect, since, e.g., probabilistic neural networks (PNN) [76] require setting only one network parameter before the network can be trained.

Suitable parameter combinations for the selected classifier can be compared using e.g. cross-validation procedure described in Section 3.4.3 and statistical methods described in Section 3.4.4.

3.4.3 Evaluation of the Learning Output

The evaluation of the learning output is the last stage in designing a classifier. However, as Theodoridis and Kountroumbas [78] note, the evaluation stage is not cut off from the previous stages of the classifier design, as the evaluation of the system's performance will determine whether the system complies with the requirements imposed by the specific application and the intended use of the system. Failing to do so may trigger redesign of the system. Theodoridis and Kountroumbas [78] also note that the system's performance indicators can be used as a performance index at the feature selection stage.

The performance of a classifier can be measured using various different performance indicators. Commonly used indicators include classification accuracy (ACC) and error rate, which measure the classifier's performance from different viewpoints [18]. Accuracy measures the percentage of correctly classified samples whereas error rate measures the percentage of false classifications. Other commonly used indicators include true positive rates (TPR), true negative rates (TNR), and positive predictive values (PPV), which are class based performance measures, and receiver operating characteristics (ROC) curves. True positive and true negative rates are also known as sensitivity and specificity [18]. In this study, classification accuracy, true positive rates, and positive predictive values were used in evaluation of the classifiers' performance, and are thus defined here.

The overall performance of a classifier can be evaluated using classification accuracy. It is the proportion of correctly classified samples to all samples and is given by

$$ACC = 100 \cdot \frac{\sum_{c=1}^{C} tp_c}{\sum_{c=1}^{C} p_c} \%, \tag{3.1}$$

where C denotes the number of classes, tp_c the number of true positive classifications for class c, and p_c the size of the class c. True positive rate is a class-based classification accuracy measure. For a given class c the true positive rate TPR_c for the class is calculated with

$$TPR_c = 100 \cdot \frac{tp_c}{p_c}\%. \tag{3.2}$$

Like TPR, also positive predictive value is a class-based performance measure. PPV is a confidence measure for the classifiers' classification decisions for a given class, and is calculated as a proportion of correctly classified samples for class c to all samples classified into class c (correct and false classifications). Thus it can be calculated with

$$PPV_c = 100 \cdot \frac{tp_c}{tp_c + fp_c}\%, \tag{3.3}$$

where fp_c denotes the number of false positive classifications of the class c, i.e., number of samples incorrectly classified into class c.

To be able to evaluate the classifier's ability to generalize outside the training set, the set $\mathcal{D}$ of the labeled training samples can be divided into two disjoint sets called training set and test set (so called *holdout method*). The training set is used to teach the classifier whereas the test set is used to estimate classifier's ability to generalize outside the training set using some performance indicator, such as accuracy [10]. The split of the data into training and test sets should be done so that the training set contains the majority of the patterns, say 90 %, and the test set the rest. Also, the class distribution in both sets should correspond to that of the original data set $\mathcal{D}$.

When the amount of available data is restricted, it is not possible to freely pick many independent training and test sets for evaluating the classifier. In such a case the following methods can be used to estimate classifiers' performance. First, the generalization of training set – test set method called m-fold cross-validation [10, 18] can be used. In m-fold cross validation the training set is randomly divided into m disjoint sets of equal size n/m, where $n = |\mathcal{D}|$, using stratified sampling. The classifier is trained m times with each time holding a different set out as a test set. Sometimes it may be necessary to perform cross-validation several times in order to assure that the partitioning the data set does not influence the results. This kind of cross-validation is called $(k \times m)$-fold cross-validation. It is performed by running m-fold cross-validation k times repartitioning the cross-validation sets after each m-fold cross-validation round. Often in practical applications $m = 10$, i.e., 10-fold cross-validation is used. If there is not enough training data available to perform cross-validation, *leave-one-out* validation can be used instead. It is a special case of cross-validation procedure, where $m = n$, i.e., n-fold cross-validation is performed using the excluded sample as a test case. If cross-validation is used, the performance of a classifier is evaluated by calculating, e.g., the average classification accuracy over the cross-validation folds.

Besides cross-validation described above, also other methods evaluating classifiers performance exist, such as bootstrap [10, 18, 19, 78] and jackknife methods [10, 19, 78]. However, the description of these methods is out of the scope of this thesis.

3.4.4 Classifier Selection

When selecting a classifier for a classification problem, it is reasonable to base the selection on statistically confirmed differences in classifiers. If two or more classifiers are compared using the cross-validation procedure, statistical *hypothesis testing* can be used to test the statistical significance of the differences between the classifiers' classification accuracies. In this case, the two following hypotheses are compared using a statistical test:

H_0: The classifiers' classification accuracies do not differ significantly.

H_1: The classifiers' classification accuracies differ significantly.

H_0 is known as the *null hypothesis* and H_1 as the *alternative hypothesis.* A statistical test is then used to analyze the differences between the classifiers' classification accuracies to determine if H_0 can be rejected and H_1 accepted. Typically, the null hypothesis is rejected when the propability of H_1 exceeds 95 %.

Many statistical tests exist for this purpose, of which a suitable one should be selected and applied with care [71]. For example, suppose that performances of two classifiers are compared, and m-fold cross-validation procedure has been run for both classifiers using the same cross-validation partitioning. Suppose also that the classification accuracies calculated during the cross-validation follow t-distribution (according to Han and Kamber [18], this is often the case). Then the Student's t-test can be applied to evaluate the statistical significance between classifiers' classification accuracies using null hypothesis that there is no difference between the classifiers' accuracies. If a known distribution (e.g. t-distribution) cannot be assumed, then a non-parametric test, like Wilcoxon signed-rank test, should be used for the comparison.

When more than two classifiers are compared, the Student's t -test should not be used to compare the classifiers with each other and then infer the relationships of the classifiers based on the comparisons. Instead, tests designed

Table 3.2: The data matrix for Friedman test. Treatments correspond to different classifiers and blocks to classification results during each cross-validation fold. In this case k classifiers are compared using m-fold cross-validation.

	Treatment (Classifier)			
Block (Fold)	1	2	$\cdots$	k
1	X_{11}	X_{12}	$\ldots$	X_{1k}
2	X_{21}	X_{22}	$\ldots$	X_{2k}
$\cdots$	$\cdots$	$\cdots$	$\ldots$	$\cdots$
m	X_{m1}	X_{m2}	$\ldots$	X_{mk}

especially for this purpose should be used [25]. Otherwise the estimates for the propabilities of the null and alternative hypotheses may be biased. If the cross-validation procedure has been run for all classifiers using the same cross-validation partitioning and if the classification accuracies calculated during the cross-validation follow normal distribution, then two-way analysis of variance can be used to compare the classifiers. However, if the assumption of normality cannot be made, then e.g. the non-parametric Friedman test can be used to compare the classification accuracies. Friedman test can be seen as two-way analysis of variance by ranks (order of observed values), since it depends only on the ranks of the observations in each block [5]. In this study Friedman test was used to compare the statistical significances of the differences between the classification accuracies of the cross-validated classifiers, and it is thus discussed next.

The Friedman test was developed by Milton Friedman in three papers [12, 13, 14] in 1937 - 1940, but the following description of the test is based on Conover [5] as he gives a more recent approach to the test. The data matrix for the Friedman test consists of m mutually independent random variables $(X_{i1}, X_{i2}, \ldots, X_{ik})$, called blocks, $i = 1, 2, \ldots, m$, which in this case correspond to the classifiers' classification accuracies during the ith cross-validation fold (m is the number of folds). Thus random variable X_{ij} is associated with cross-validation fold i and classifier j (treatment in statistical terminology, see Table 3.2). As was noted before, Friedman test can be seen as two-way analysis of variance by ranks. Therefore, let $R(X_{ij})$ be the rank, from 1 to k, assigned to X_{ij} within block i. This means that the values X_{i1}, X_{i2}, $\ldots$, X_{ik} are compared and rank 1 is assigned to the smallest observed value and rank k to the largest observed value. In case of ties average rank is used to substitute the original rank values. For example, if there are two observations with the same value on the second place, then rank 2.5 will be used for the both observations. The rank totals R_j are next calculated for each classifier j with

$$R_j = \sum_{i=1}^{m} R(X_{ij}), \tag{3.4}$$

for $j = 1, \ldots, k$. The Friedman test determines whether the rank totals R_j for each classifier differ significantly from the values which would be expected by chance [75].

To formulate the test, let A_1 be the sum of the squares of the ranks, i.e.,

$$A_1 = \sum_{i=1}^{i=m} \sum_{j=1}^{i=k} \left(R(X_{ij})\right)^2, \tag{3.5}$$

and C_1 a correction factor calculated with

$$C_1 = mk(k+1)^2/4. \tag{3.6}$$

The Friedman test statistics T_1 is calculated with

$$T_1 = \frac{(k-1)(\sum_{j=1}^{k} R_j^2 - mC_1)}{A_1 - C_1}. \tag{3.7}$$

The distribution of T_1 can be approximated with chi-squared distribution with $k-1$ degrees of freedom. However, as noted by Conover [5], the approximation is sometimes poor, and thus test statistic T_2 calculated as a function of T_1 should be used instead. It is calculated with

$$T_2 = \frac{(m-1)T_1}{m(k-1) - T_1}, \tag{3.8}$$

and has the approximate quantiles given by the F distribution with $k_1 = k-1$ and $k_2 = (m-1)(k-1)$ when the null hypothesis (the classifiers' classification accuracies do not differ in statistical sense) is true. The null hypothesis should be rejected at the significance level α if T_2 exceeds the $1-\alpha$ quantile of the F distribution. The approximation is quite good and improves when m gets larger.

If the null hypothesis of Friedman test can be rejected at the chosen α-level, it means that *at least one* of the classifiers differs from *at least one* other classifier [75]. That is, it does not tell the researcher which ones are different, nor does it tell the researcher how many of the classifiers are different from each other. For determining which classifiers actually differ from each other, a multiple comparison method can be used. The classifiers i and j are statistically different if

$$|R_j - R_i| > t_{1-\alpha/2} \left(\frac{(A_1 - C_1)2m}{(m-1)(k-1)} \left(1 - \frac{T_1}{m(k-1)} \right) \right)^{1/2}, \tag{3.9}$$

where $t_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the t-distribution with $(m-1)(k-1)$ degrees of freedom and α has the same value as was used in Friedman test. In other words, if the difference of rank sums of the two compared classifiers exceeds the corresponding critical value given in Eq. (3.9), then the two compared classifiers may be regarded as different.

Although the classifier's performance in a classification task can be seen as the most important criterion when comparing classifiers, also other criteria exist. Depending on the application area these might include the following criteria [18]: the speed of the classifier, its robustness, scalability, and interpretability. The speed of the classifier refers to actual computational costs that training and using the classifier require. These might vary a lot depending on classifier type. Also, the cost of training the classifier and using an actual classifier might vary. This has implications for the types of problems the classifiers are suited for. Nearest neighbor methods [18], for example, are know as "lazy learners", since all actual computation is done during the classification. Therefore, using the classifier with large data sets requires large computational resources, which might render them unsuitable for online usage. On the other hand, MLP classifiers [22] provide an inverse example, since the classification with trained classifier is fast, but the training takes time.

The robustness of a classifier is the ability of a classifier to make correct decision with noisy or incomplete data. MLP classifiers are known to be quite tolerant for noisy data, and they can classify patterns, which they have not been trained for, whereas a k-nearest neighbor classifier is quite sensitive to noise. A classifier is scalable if it is able to perform efficiently even with large amounts of data. The scalability might be an issue for the traditional decision tree algorithms with very large data sets [18].

Interpretability of a classifier refers to the level of understanding and insight that is provided by the classifier. That is, interpretability refers to how easily the decisions made by the classifier can be understood by humans. Interpretability of the classifier might be a very important factor especially in medical expert systems, where it is important to know the reasons why the classifier made a certain decision over another [19]. Although interpretability is a subjective matter [18], some classifiers are easier to interpret than others. For example, acquired knowledge represented in a decision tree classifier is generally in more intuitive form for humans than that of MLP classifiers.

Finally it should be noted, that although it is possible to find a classifier that suits especially well for a particular classification problem, it does not mean that the classifier also performs better with some other, different problem. In fact, if the goal is to maximize the classifier's overall generalization performance, there are no context- or usage-independent reasons to favor one classification method over another [10]. Therefore, the suitability of different machine learning classifiers for classifying aphasic and non-aphasic speakers should be compared and evaluated.

Chapter 4

Roles of the Individual Publications in the Dissertation

This thesis is based on papers addressing two different topics related to each other by the general application area (aphasia) and by methods used to solve the problems in this area (machine learning). The first part of the thesis consists of three papers related to neural network modeling of language production and its disorders. The second part consists of two papers addressing classification of aphasic and non-aphasic speakers based on their results in various aphasia tests. Next, in Sections 4.1 and 4.2 these areas are shortly introduced and an overview of the produced papers is given.

4.1 Modeling of Language Production and its Disorders

Investigation and development of models of language production offer both theoretical and practical benefits [38]. The theoretical benefit of modeling language production is that researchers can create new testable hypotheses about language production based on these models. On the clinical side, models can be used to diagnose language disorders, e.g., deciding to which model's processing level a patient's lesion corresponds. They can also be used to the rehabilitation of aphasic patients, e.g., by examining which processes should be rehabilitated according to the model. Furthermore, the more specific a model of language production is used, the less theoretically justified approaches to treatment there exist [55].

Models of language production are in no way a new innovation. Wernicke and Lichtheim had their first coarse level models of language production in 1874 and 1885, respectively [38]. The models of Wernicke and Lichtheim have played a major role in the creation of the current aphasia profile classification, and the features of the models can still be seen in the current models of word production [38].

From the 1960s to 1980s the behaviorist models of language production were developed. These so called "Box and Arrow" models defined processes needed in language production (boxes) and their relationships (arrows). These functional models have proven to be especially useful in the diagnosis of aphasic patients, because the functional description of cognitive level processes is easier to relate to a patient's symptoms than those of anatomical models. In the 1980s the development of the artificial neural networks enabled the modeling of the processing inside the "boxes" as well as their relationships, which resulted in connectionist modeling of language production. [38]

Since the rise of the connectionist neural network models in the mid 1980s, the neural network modeling of language production and its disorders has gained considerable research interest. Although research has been done before the mid 80s (e.g. [46, 47]), a significant impact on the field was the publication of the highly influential book pair Parallel Distributed Processing vols. 1 and 2 [67, 68] edited by D. E. Rumelhart and J. L. McClelland. The book, among other pioneering work, popularized the back-propagation learning rule for MLP networks [65]. The book also contained a chapter [66] on learning English past tenses, which showed the neural networks suitable for language processing tasks. Soon also MLP-based NETtalk [74] model was published showing that the MLP networks could successfully be used in letter to phoneme mapping problems. It can be heard from the audio tape documenting the learning of the network how the network progresses from the baby babbling via single syllable pronunciation to full text reading.[1]

Also, neural network models designed especially for Finnish have been developed [27, 79, 80, 81, 82, 83, 84, 87]. These neural network models have been applied to nominal inflection [79], transcription of continuous speech [33], diagnostics of speech voicing [33], and to modeling impaired language production system [27, 39, 84, 87]. Neural network models have been successfully applied to the modeling of impaired language production also elsewhere [9, 16, 21, 41, 43, 44, 54, 57, 69, 70, 89]. From language disorders, the modeling of impaired lexical access has been modeled very actively [9, 11, 39, 49]. Usually the models of lexical access focus on single word production, and especially modeling lexicalization.

Two major modeling goals of lexicalization have been modeling the time

[1] The audio tape can be downloaded from the Internet at `http://www.cnl.salk.edu/ParallelNetsPronounce/nettalk.mp3`.

course of the lexicalization process and simulating the effects of brain damage on the lexicalization process. The main focus in the time course studies has been to determine the interaction of semantic and phonological processing during lexicalization, as this has been a major area of debate among the researchers. The other major modeling goal has been the modeling of the naming performances of individual patients (see e.g. [9, 11, 16, 39]). The goal is to investigate if the models are able to simulate the specific symptoms of the patients. Usually this is done by fitting the model to the naming data of the patients. The purpose of the patient data simulations is to (1) evaluate models against empirical data, (2) gain further insight into the functional location of the damage of the patients within a cognitive model of language processing and (3) predict the course of recovery from language impairment.

Traditionally the models of lexicalization have been non-learning as the connection weights of the models have been set externally (e.g. the models used in [8, 9, 11, 16, 39]). From these models, the spreading activation based interactive activation model of Dell et al. [7, 8, 9, 11, 45, 72, 73] is by far most well known and comprehensively tested. However, in order to perform more realistic simulations and to simulate the recovery and rehabilitation process of impaired word production system, learning models of lexicalization are needed. As was mentioned in Chapter 2.3.2, at present, there is a gap between cognitive neuropsychological diagnostics and choice of a treatment method. One reason to this gap may be our lack of understanding the dynamic relearning process during treatment.

There are some models simulating language production and its disorders with capability to learn [23, 50, 56, 58, 85, 89]. Plaut [56], for example, investigated relearning in the connectionist networks after the model had been damaged. However, these kinds of models have not been developed for simulating the lexicalization process. The purpose of the papers constituting the first part of the thesis was to investigate the suitability of the MLP architecture for the basis of such a neural network model. The papers introduce and investigate the properties of the developed Learning Slipnet simulation architecture.

4.1.1 Paper I – Introducing the Learning Slipnet Simulation Model

In this paper the MLP based discrete-two stage model of lexicalization, Learning Slipnet (see Fig. 4.1) is introduced. The developed model consists of MLP neural networks simulating semantic and phonological processing respectively. The effect of the brain damage can be simulated by adding ran-

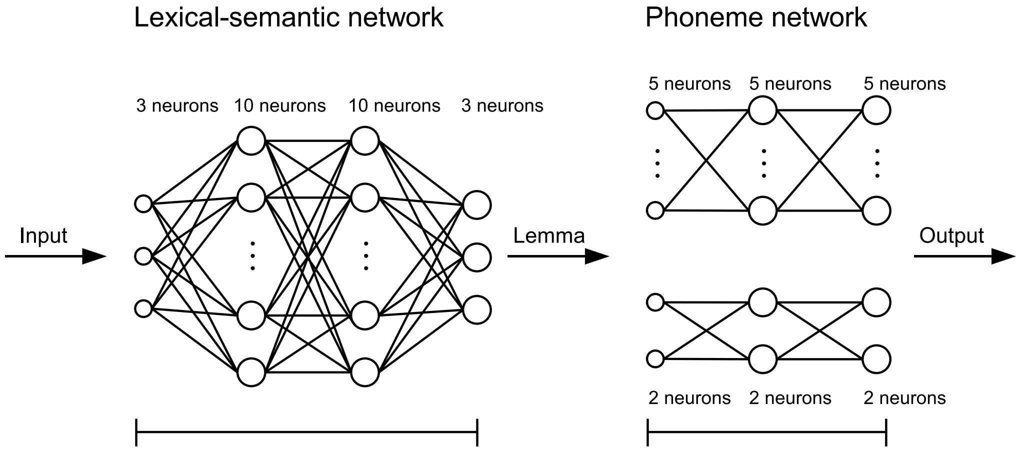

Figure 4.1: Discrete-two stage model of lexicalization, Learning Slipnet, developed and evaluated in Papers I – III. The model consists of lexical-semantic network simulating the first stage of lexicalization, and two phoneme networks, one for vowels and one for consonants, simulating the second stage of the lexicalization.

dom noise to the lexical-semantic or phoneme network of the model, or using threshold between the networks to generate no response (omission) errors. In the following sections the lexical-semantic network's noise parameter is denoted as α_L and the phoneme network's noise parameter with α_P. The threshold between the networks is denoted with τ. Thus, the parameter α_L regulates the amount of semantic errors in the model, the parameter α_P the amount of phonological errors, and threshold parameter τ the amount of omission errors.

In order to automatize the error classification three error classes were used: omissions, semantic errors and other errors. The following rules were applied to the error classification:

1. An output was an omission if the lexical-semantic network generated no response to a given input word.

2. An output was a semantic error if an output word of the lexical-semantic network was different from its input word.

3. An output was phonological error if an output word of the phoneme network was different from its input word, i.e., output of the lexical-semantic network.

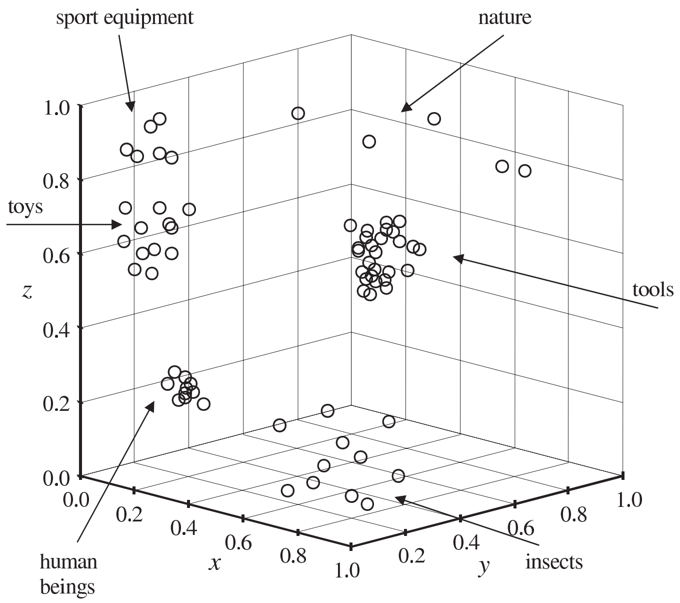

Figure 4.2: Examples of the semantic classes produced by the algorithm developed in Paper I. The concept “living” approximately covers the area of $x \in [0, 1]$, $y \in [0, 1]$, $z \in [0, 0.35]$. Correspondingly, the concept “not living” has the area of $x \in [0, 1]$, $y \in [0, 1]$, $z \in [0.55, 1.0]$.

If the model first produced a semantic error and then phonological error, it was simply classified as phonological error.

Paper I addresses especially the first research problem, the input and output encoding for the model. This problem was interesting and non-trivial, since to be able to train neural network models for simulating word production both the semantics and the phonology of the words must be presented in numerical form. In the paper a new tree based coding technique for semantic features was developed and analyzed. With the algorithm it is possible to generate semantic representations that are compact and easy to modify which renders the coding method suitable for the developed MLP based neural network model of word production. The developed encoding technique is not restricted to be used with Learning Slipnet model only, as it can be utilized also with other neural network or machine learning models. Examples of the semantic classes produced by the algorithm are given in Fig. 4.2.

Some preliminary test runs with the model were also reported using the naming data of four Finnish aphasia patients suffering from word-finding difficulties that Laine et al. [39] used testing their Slipnet model of lexicalization. However, the more detailed evaluation of the model was performed in Papers II and III.

4.1.2 Paper II – Testing Learning Slipnet

The second research problem, the suitability of the MLP neural network architecture for the basis of a model of word production was investigated in Paper II. First, the model's theoretical capabilities to simulate various aphasic naming disorders was investigated by systematically analyzing the effects of the model's parameters (α_L, α_P, and τ) to the error distribution produced by the model. This analysis showed, that at least in theory, the model was capable to account performance patterns of various range of aphasic patients.

After the general analysis of the model's behavior, the model's performance was compared against empirical naming error data from ten aphasic patients. The same patient data was used in this paper as in Paper I, but this time the naming data of all ten patients reported in [39] were simulated with Learning Slipnet. The results presented in Table 4.1 proved the model quite successful in simulating word production errors in this heterogeneous group of aphasic patients. However, the data set consisted of only ten patients, and thus more data would be needed in order to perform rigorous tests with the model.

4.1.3 Paper III – Further Experiments with Learning Slipnet Using Dementia Patients' Naming Data

The testing of the model (the second research problem) was further continued in Paper III, where the naming performances of 22 Finnish-speaking dementia patients and 19 neurologically healthy control subjects were simulated. The dementia data set was originally described by Laine et al. [40] and it contains naming distributions of 12 Alzheimer's disease (AD) patients and 10 vascular disease (VaD) patients. The subjects' naming distributions were based on the half of the items (i.e. 30) in the Finnish version of the Boston Naming Test [36], but Laine et al. [40] allowed 45 seconds for spontaneous naming to produce enough scoreable data. Thus most subjects' error distribution is based on over 30 answers.

The model was able to simulate the naming distributions of the test subjects quite accurately as can be seen from the averaged simulation results presented in Table 4.2 (more detailed results are available in the Paper III). In Fig. 4.3 the simulated subjects are plotted with respect to the simulation

Table 4.1: Confrontation naming test patterns of individual aphasic patients (P) and corresponding averaged simulation results with Learning Slipnet (M) in percent. A1, A2 and A3 are cases with anomic aphasia, B1 and B2 with Broca's aphasia, C1 and C2 with conduction aphasia, and W1, W2 and W3 with Wernicke's aphasia.

Patient	Correct %		Semantic %		Omission %		Phonological %	
	P	M	P	M	P	M	P	M
A1	51.7	53.3	4.2	4.4	44.1	42.4	0.0	0.0
A2	77.1	76.5	8.4	8.5	13.9	14.1	0.6	0.9
A3	39.2	39.2	0.6	0.4	59.6	59.8	0.6	0.6
B1	77.1	77.5	3.0	2.5	19.3	19.4	0.6	0.6
B2	87.3	87.7	3.0	2.7	8.4	8.7	1.2	0.9
C1	65.7	66.4	3.6	3.2	7.2	7.6	23.5	22.8
C2	84.3	85.4	0.6	0.6	4.8	4.0	10.2	9.9
W1	54.8	53.8	4.8	5.1	20.5	20.6	19.9	20.6
W2	42.9	44.2	13.9	13.9	21.1	21.0	22.2	20.9
W3	36.7	36.7	0.6	0.6	37.3	39.0	25.3	23.7

parameters α_L, α_P, and τ. The squares represent AD patients, the circles VaD patients and the diamonds represent the healthy control subjects. The healthy control subjects form a tight cluster, but the clusters of AD and VaD patients are more dispersed. Generally, according to the noise parameter α_L, the VaD patients are closer to the healthy control subjects than AD patients. This is the consequence of the fact that the simulated AD patients tend to make more semantic errors than the simulated VaD patients. On the other hand AD patients were usually closer to the healthy control subjects with respect to the noise parameter α_P. The differences between the AD and VaD patients were not significant with respect to the threshold τ. Based on the simulation results presented in Paper II and III, the model seems to be suitable for simulating naming disorders of patients suffering from various aphasic symptoms.

Table 4.2: Averaged confrontation naming test patterns of Alzheimer's disease patients (AD), vascular disease patients (VaD), and healthy control subjects (Control), and the corresponding averaged performance patterns of the Learning Slipnet model. The performance patterns of the subjects are marked with S, and performance patterns of the model with M. Note that term "other error" is used instead of phonological error in the original Paper III.

Patient	Correct %		Semantic %		Phonological %		Omission %	
	S	M	S	M	S	M	S	M
AD	38.9	39.9	53.7	53.1	8.5	7.7	0.0	0.2
VaD	52.2	52.4	31.3	31.3	5.6	5.3	0.9	1.0
Control	78.8	79.9	17.6	18.6	2.6	2.0	0.0	0.0

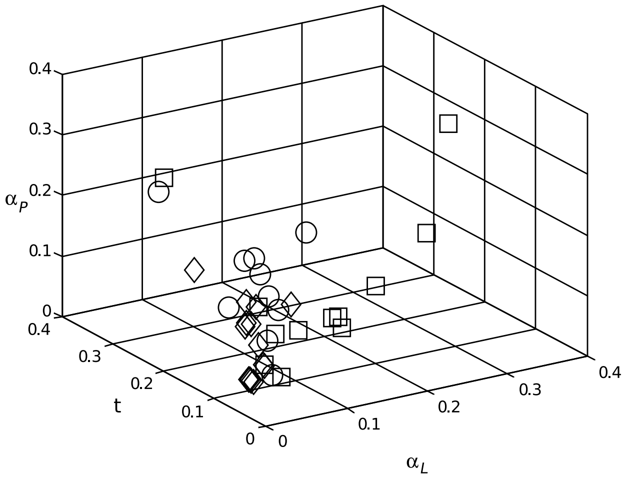

Figure 4.3: The subjects are plotted with respect to the simulation parameters α_L, α_P, and τ. The squares represent Alzheimer's disease (AD) patients, the circles vascular (VaD) disease patients and the diamonds represent the healthy control subjects. The healthy control subjects form a tight cluster, but the clusters of AD and VaD patients are more dispersed.

4.2 Machine Learning on Aphasic Naming Data

The third research problem addressed in the latter part of the thesis was to find out if certain types of machine learning classifiers would be especially suitable to classify aphasic and non-aphasic subjects into groups based on their distribution of scores in various aphasia tests. Although decision support systems in various medical domains have received considerable research interest, aphasic patient classification based on their naming distributions has not been reported much in the literature, the papers of Axer et al. [2, 3, 26] and Tsakonas et al. [86] being the few examples. The goal of the classification research was to evaluate different machine learning classifiers for classifying aphasic naming data using different data sets. These classifiers could then be used for implementing a decision support system which could be utilized by neuropsychologists and speech therapists as an intelligent software tool when investigating aphasic patients.

The research was started by investigating the classification performance of three neural network classifiers separating healthy and aphasic speakers using the data set of Dell et al. [9] (Paper IV). From the tested classifiers a single neuron MLP classifier performed the best, which suggested that the classification problem with the tested data set was a linear problem. However, the problematic feature of the data set was its small size. Therefore artificially generated data sets derived from the data set of Dell et al. [9] were used in the tests.

The good performance of one neuron MLP classifier with the data set of Dell et al. [9] suggested that other simple classifiers should also perform well with the data set. Therefore more classifiers were tested in Paper V. Furthermore, the classifiers were tested with three aphasic data sets, which were:

1. The data set of Dell et al. [9]. The goal was to separate healthy speakers from aphasic patients.

2. Dementia data set used in the Paper III. With this data set the classifiers were trained to separate vascular disease patients from Alzheimer's disease patients.

3. PatLight aphasia data set [1] containing Aachen aphasia test results and diagnoses of 265 aphasic patients. With this data set the goal was on classifying the classical aphasias, anomic, Broca's, conduction, and Wernicke's aphasia. This exclusion of other aphasias left a data set with 146 patients that were used for testing the classifiers.

The Paper V reports the comparison of eight classifier types with the three aforementioned data sets.

4.2.1 Paper IV – Testing Neural Network Classifiers for Aphasic Naming Data Classification

In this paper, the suitability of neural network classifiers for separating healthy individuals from aphasic patients was investigated. Three neural network classifiers, multi-layer perceptrons, probabilistic neural networks, and self-organizing maps, were compared for the classification of aphasic and non-aphasic naming data. The performance of the classifiers were compared using the aphasic naming data reported by Dell et al. [9] as a base data set, which was artificially augmented to suit better for the neural network classifiers. In order to smooth the differences between the augmented data sets, totally ten different data sets were generated from the base set. Each classifier was examined by running a 10 × 10 cross-validation for each data set. The differences between the classifiers were compared by calculating average classification accuracy for each classifier over the ten cross-validated data sets. The differences of the classification accuracies between the classifiers were tested with Friedman test.

The results showed that MLP performed best in separating the aphasic speakers from healthy speakers based on their naming data distributions, although all classifiers had classification accuracy over 90 %. MLP's total classification accuracy was 1 – 2 % higher than the accuracies of other classifier types, and the smaller standard deviation of the classification accuracies proved it also to be a more robust classifier than other tested classifiers. Furthermore, because only one neuron was needed to implement the most successful MLP architecture, it was predicted that other simple classification methods, such as discriminant analysis and Bayes classifiers should also perform well at the classification task. Indeed, evaluation of this prediction was a partial goal of the Paper V.

4.2.2 Paper V – Experimenting with Several Machine Learning Classifiers on Three Aphasic Naming Data Sets

The goal of this study was to investigate the suitability of several machine learning classifiers to separate healthy individuals from aphasic patients and to classify aphasic patients to the groups based on their performance in the aphasia tests. A total of eight classifiers, multi-layer perceptrons (MLP) [22], probabilistic neural networks (PNN) [76], self-organizing maps (SOM) [32], k-nearest neighbor (k-NN) classifier [19], k-means classifier (k-means) [19], decision tree classifier (tree) [59], discriminant analysis [10], and naïve Bayes classifier [10], were tested with three different aphasia data sets.

With the first data set (aphasia data set of Dell et al. [9]), the goal was to separate healthy subjects from the patients. Because the data set is very small it was artificially augmented to suit better for the tested classifiers, as was was done in Paper IV. With the second data set (the dementia data set of Paper III) the goal was to separate the Alzheimer patients from vascular disease patients using the patients' naming distributions. Also, this data set was quite small and needed to be augmented. Finally, with the third data set (PatLight aphasia data set reported by Axer et al. [1]), the goal was to investigate different classifiers' ability to recognize the patients' aphasic syndrome based on their results in Aachen aphasia test.

As in Paper IV, 10×10 cross-validation was used with each data set for validating the classifiers. Total classification accuracy was used as the main performance indicator. For class based evaluation true positive rates and positive predictive values were used. Friedman test was used to test the statistical differences between the classifiers' classification accuracies. Based on the results, no single classifier performed exceptionally well with all data sets. This suggests that with each new aphasia data set, the selection of suitable classifiers for the data set should be based merely on the experiments performed for the data set.

For data set of Dell et al. [9] the decision tree classifier was the best performing classifier with the classification accuracy of 94.4 %. However, as the differences between the classifiers' classification accuracies were generally small, the choice of classification method does not seem to be very crucial for this data set. Based on the TPRs and PPVs all classifiers were biased towards the healthy class.

With the dementia data set, the best performing classifier was the discriminant analysis classifier with the classification accuracy of 69.9 %, but also other classifiers had an accuracy over 60 %. The standard deviations of the classification accuracies were extremely high with all classifiers (ca 30 %) and therefore all tested classifiers were quite unreliable for the task.

For PatLight data set the two best performing classifiers were k-NN and PNN classifiers with the classification accuracies of 90.5 % and 90.0 % respectively. Other classifiers also had classification accuracies of 80 %, except MLP classifier, but k-NN and PNN classifiers were still clearly the two best classifiers. Based on the TPRs, the global aphasic class was the easiest to recognize for the classifiers, although the TPRs varied between classifiers.

The results of the first and the third data sets showed that automatic classification tools might be useful in the aphasia treatment. However, as the results of the second data set suggest, this might not hold with all data sets. Also, the present results showed that the suitability of the individual classifiers should be tested for each new data set, since no single classifier outperformed others with all data sets.

Chapter 5

Conclusions

This thesis focused on applying machine learning methods on aphasic data. The papers constituting the thesis look at the problem from two different angles: in the first part, the naming performances of aphasic and non-aphasic speakers were simulated with specially designed model of lexicalization. Here the goal was to understand the lexicalization process within a computational framework that enables simulating effects of brain damage to the lexicalization process. In the latter part of the thesis, the view is moved from the modeling of the lexicalization process to the investigation of the aphasic data itself. Here classifiers for recognizing various clinical aphasia syndromes as well as separating healthy speakers from aphasic speakers were analyzed.

The first part of the research constituting this thesis started with searching suitable encoding methods for semantic representations of the words. This was a prerequisite for building a model of lexicalization, since the semantics of the words has to somehow be numerically encoded for neural networks before they can be trained. One possible solution to this problem was presented in Paper I, where a new tree based coding technique for semantic features was developed and analyzed. Papers II and III focused on the evaluation of the developed Learning Slipnet model of lexicalization. The results presented in Papers I–III showed that the model was able to simulate the naming performances of both aphasic and non-aphasic speakers, and thus MLP based lexicalization model might be useful also in the clinical aphasia treatment.

However, based on the experience[1] on the lexicalization gained on the Papers I–III, some cautions should be reserved to these results. It seems that modeling lexicalization with learning networks needs to be restricted to simulating learning and relearning processes only. For example, modeling

[1] This refers to personal "silent" expertise gained during the research process, not on the empirical results presented in the Papers I–III.

inner dynamics of the lexicalization process within the interactive two stage framework would require neural networks whose learning algorithms were developed especially for this purpose. For a model implementing the interactive theory, besides of recurrent processing, also selection bonus at the lemma level, which is an essential part of the theory, should be incorporated into the network's learning algorithm. Therefore, if the goal is to model the inner dynamics of the lexicalization process, special hand crafted non-learning models of lexicalization, such as models of Dell et al. [9] or Laine et al. [39], should be used instead, since with these models it is possible to implement the details of the process. However, if the learning or relearning process itself is the target of the modeling, then traditional neural network models, such as MLP or SOM networks, would show some prominence. The examples of such models include the Learning Slipnet model introduced in this thesis, and models of Plaut [56], and Miikkulainen [50]. Especially, SOM networks, on which Miikkulainen's Dislex model is based, should be investigated in the future when modeling learning and relearning processes of aphasic and non-aphasic speakers.

It should also be noted that the ability of a model to simulate *every* possible naming distribution is not a very desirable feature, since then constructing a test that would falsify the model is impossible. Failing to do so severely cripples the model's credibility. Instead, efforts should be put to designing and constructing as simple models as possible, whose performance patterns would be restricted, and follow those of observed with healthy and aphasic speakers. These remarks should be born in mind when designing a new model of spoken word production.

The classification experiments performed in the second part of the thesis showed that machine learning classifiers can be applied to the classification of aphasic and non-aphasic speakers based on their performace in the aphasia tests. The two data sets used in the classification research suffered from their small size, which might bias the results. The experiments with PatLight data set, however, did not suffer from this problem. The classification results obtained with this data set were on par with the results of Axer et al. [3], who had average classification accuracy of 92 % with their best performing MLP classifier. Moreover, it should be noted that Axer et al. [3] used the hold out method in evaluation of their classifiers, whereas the results presented in Paper V were obtained with a cross-validation procedure, which is a more stringent evaluation criterion. If more aphasic data becomes available, then the classification branch of this thesis could be continued, and especially practical applications considered. These would include developing a production ready system for the clinicians to use during aphasia treatment. However, provided the amount of the available data, unfortunately developing such a system is not currently plausible.

The software tools needed in the simulation and classification studies of this thesis, were implemented in the Matlab environment. To be widely and easily accessible for the researchers in the field, of which most are not computer scientists, especially models of lexicalization should be implemented as easy to use stand alone open source programs with graphical user interface. This is one direction that could be taken in the future investigation of the topic.

To sum up, applications of machine learning methods applied to aphasic naming data were investigated. From the results presented in the Chapter 4, it can be concluded that machine learning methods might provide new insight to the spoken word production of aphasic and non-aphasic speakers. However, both application areas would greatly benefit from larger aphasia data sets available, especially describing patients' performance in confrontation naming task, as this would enable even more reliable evaluation of the developed models.

Appendix A

Personal Contributions

Most of the thesis is based on publications containing essential contributions from Martti Juhola and Matti Laine. Therefore, the author of the thesis (referred as AJ hereafter) must separate his individual contributions on each paper. They are as follows:

I. AJ designed the test setup, implemented the model, ran the tests, and analyzed the results. The paper was mostly written by Martti Juhola while Matti Laine provided neuropsychological expertise and AJ wrote some technical parts of the paper.

II. AJ designed the test setup, implemented the model, ran the tests, and analyzed the results. The first version of the paper was written by Martti Juhola, with Matti Laine contributing on the neuropsychological aspects and AJ corresponding on the technical details. Later, during the review process, AJ modified the text heavily.

III. Matti Laine provided the original aphasic naming data used in the article. AJ designed the test setup, preprocessed the naming data, implemented the model, ran the tests, and analyzed the results. The paper was mostly written by AJ with help from Martti Juhola and Matti Laine.

IV. Publication IV is AJ's own work as much as can be while being a member of a scientific research group like Data Analysis Research Group at the University of Tampere. Especially, Martti Juhola gave feedback on the test setup and commented the manuscript.

V. The test setup was designed by AJ and Martti Juhola. AJ implemented the software, performed the experiments, analyzed the results, and wrote the paper.

Bibliography

[1] H. Axer, J. Jantzen, G. Berks, D. Südfeld, and D. G. von Keyserlingk. The aphasia database on the web: Description of a model for problems of classification in medicine. In *ESIT 2000: Proceedings of European Symposium on Intelligent Techniques*, pages 104–110, Aachen, Germany, 2000.

[2] H. Axer, J. Jantzen, and D. G. von Keyserlingk. Aphasia classification using neural networks. In *ESIT 2000: Proceedings of European Symposium on Intelligent Techniques*, pages 111–115, Aachen, Germany, 2000.

[3] H. Axer, J. Jantzen, and D. G. von Keyserlingk. An aphasia database on the Internet: A model for computer-assisted analysis in aphasiology. *Brain and Language*, 75(3):390–398, 2000.

[4] D. Chilant, A. Costa, and A. Caramazza. Models of naming. In A. E. Hillis, editor, *The Handbook of Adult Language Disorders*, pages 123–142. Psychology Press, New York, NY, USA, 2002.

[5] W. J. Conover. *Practical Nonparametric Statistics.* John Wiley & Sons, New York, NY, USA, 3rd edition, 1999.

[6] F. Cuetos, G. Aguado, C. Izura, and A. W. Ellis. Aphasic naming in Spanish: Predictors and errors. *Brain and Language*, 82(3):344–365, 2002.

[7] G. S. Dell, M. F. Schwartz, and N. Martin. Testing the interactive two-step model of lexical access: How we do it and why. *Brain and Language*, 91(1):69–70, 2004.

[8] G. S. Dell, M. F. Schwartz, N. Martin, E. M. Saffran, and D. A. Gagnon. A connectionist model of naming errors in aphasia. In J. A. Reggia, E. Ruppin, and R. S. Berndt, editors, *Neural Modelling of Brain and Cognitive Disorders*, volume 6 of *Progress in Neural Processing*, pages 135–156. World Scientific Publishing, 1996.

[9] G. S. Dell, M. F. Schwartz, N. Martin, E. M. Saffran, and D. A. Gagnon. Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, 104(4):801–838, 1997.

[10] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wiley & Sons, New York, NY, USA, 2nd edition, 2001.

[11] D. Foygel and G. S. Dell. Models of impaired access in speech production. *Journal of Memory and Language*, 43:182–216, 2000.

[12] M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, Dec. 1937.

[13] M. Friedman. A correction: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 34(205):109, Mar. 1939.

[14] M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics*, 11(1):86–92, Mar. 1940.

[15] J. E. Garcia-Albea, M. L. Sanchez-Bernardos, and S. D. Viso-Pabon. Test de Boston para el diagnóstico de la afasia: Adaptación española [Boston Naming Test for aphasia diagnosis: Spanish version]. In H. Goodglass and E. Kaplan, editors, *La evaluación de la afasia y de trastornos relacionados [Assessment of aphasia and related disorders]*, pages 129–198. Editorial Medica Panamericana, Madrid, Spain, 2nd edition, 1986.

[16] M. Goldrick and B. Rapp. A restricted interaction account (RIA) of spoken word production: the best of both worlds. *Aphasiology*, 16(1/2):460–499, 2002.

[17] H. Goodglass and E. Kaplan. *The Assessment of Aphasia and Related Disorders.* Lea & Febiger, Philadelphia, PA, USA, 2nd edition, 1983.

[18] J. Han and M. Kamber. *Data Mining. Concepts and Techniques.* Morgan Kaufmann Publishers, San Fransico, CA, USA, 2nd edition, 2006.

[19] D. J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining.* MIT Press, Cambridge, MA, USA, 1st edition, 2001.

[20] T. Harley. *The Psychology of Language.* Psychology Press, New York, NY, USA, 2nd edition, 2001.

[21] T. A. Harley. Connectionist modeling of the recovery of language functions following brain damage. *Brain and Language*, 52(1):7–24, 1996.

[22] S. Haykin. *Neural Networks. A Comprehensive Foundation.* Prentice-Hall, London, UK, 2nd edition, 1999.

[23] G. E. Hinton and T. Shallice. Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, 98(1):74–95, 1991.

[24] W. Huber, K. Poeck, and D. Weniger. The Aachen aphasia test. In F. C. Rose, editor, *Advances in Neurology. Progress in Aphasiology*, volume 42, pages 291–303. Raven, New York, NY, USA, 1984.

[25] D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval*, pages 329–338, New York, NY, USA, 1993.

[26] J. Jantzen, H. Axer, and D. G. von Keyserlingk. Diagnosis of aphasia using neural and fuzzy techiniques. In *CoIL 2000: Proceedings of Symposium on Computational Intelligence and Learning*, pages 124–132, 2000.

[27] M. Juhola, A. Vauhkonen, and M. Laine. Simulation of aphasic naming errors in Finnish language with neural networks. *Neural Networks*, 8(1):1–9, 1995.

[28] E. Kaplan, H. Goodglass, and S. Weintraub. *The Boston Naming Test.* Lea & Febiger, Philadelphia, PA, USA, 1st edition, 1983.

[29] J. Kay, R. Lesser, and M. Coltheart. *PALPA: Psycholinguistic Asessments of Language Processing in Aphasia.* Psychology Press, East Sussex, UK, 1st edition, 1992.

[30] A. Kertesz. *Western Aphasia Battery Test Booklet.* The Psychological Corporation, New York, NY, USA, 1st edition, 1982.

[31] H. Kim and D. L. Na. Normative data on the Korean version of the Boston naming test. *Journal of Clinical and Experimental Neuropsychology*, 21(1):127–133, 1999.

[32] T. Kohonen. The self-organizing map. *Neurocomputing*, 21(1–3):1–6, 1998.

[33] T. Kohonen. *Self-Organizing Maps.* Springer-Verlag, Berlin, Germany, 3rd edition, 2001.

[34] A. Konar. *Artificial Intelligence and Soft Computing. Behavioral and Cognitive Modeling of the Human Brain.* CRC Press, Boca Raton, FL, USA, 1st edition, 2000.

[35] M. Laine. The learning brain. *Brain and Language*, 71(1):132–134, 2000.

[36] M. Laine, H. Goodglass, J. Niemi, P. Koivuselkä-Sallinen, J. Tuomainen, and R. Marttila. Adaptation of the Boston Diagnostic Aphasia Examination and the Boston Naming Test into Finnish. *Scandinavian Journal on Logopedics & Phoniatrics*, 18:83–92, 1993.

[37] M. Laine, P. Kujala, J. Niemi, and E. Uusipaikka. On the nature of naming difficulties in aphasia. *Cortex*, 28:537–554, 1992.

[38] M. Laine and N. Martin. *Anomia: Theoretical and Clinical Aspects.* Brain Damage, Behaviour and Cognition Series. Psychology Press, New York, NY, USA, 1st edition, 2006.

[39] M. Laine, A. Tikkala, and M. Juhola. Modelling anomia by the discrete two-stage word production architecture. *Journal of Neurolinguistics*, 11(3):275–294, 1998.

[40] M. Laine, E. Vuorinen, and J. O. Rinne. Picture naming deficits in vascular dementia and Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology*, 19:126–140, 1997.

[41] W. J. M. Levelt, A. Roelofs, and A. S. Meyer. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1):1–38, 1999.

[42] W. J. M. Levelt, H. Schriefers, D. Vorberg, A. S. Meyer, T. Pechmann, and J. Havinga. Normal and deviant lexical processing: reply to Dell and O'Seaghdha (1991). *Psychological Review*, 98(4):615–618, 1991.

[43] W. J. M. Levelt, H. Schriefers, D. Vorberg, A. S. Meyer, T. Pechmann, and J. Havinga. The time course of lexical access in speech production: a study of picture naming. *Psychological Review*, 98(1):122–142, 1991.

[44] P. Li, I. Farkas, and B. MacWhinney. Early lexical development in a self-organizing neural network. *Neural Networks*, 17(8–9):1345–1362, 2004.

[45] N. Martin, G. S. Dell, and M. F. Schwartz. Testing the interactive two-step model of lexical access: Part II. Predicting repetition from naming. *Brain and Language*, 91(1):73–74, 2004.

[46] J. L. McClelland and D. E. Rumelhart. An interactive activation model of context effects in letter perception: Part 1. An account of the basic findings. *Psychological Review*, 88(5):375–407, 1981.

[47] J. L. McClelland and D. E. Rumelhart. An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, 89(1):60–94, 1982.

[48] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943.

[49] M. G. McNellis and S. E. Blumstein. Self-organizing dynamics of lexical access in normals and aphasics. *Journal of Cognitive Neuroscience*, 13(2):151–170, 2001.

[50] R. Miikkulainen. Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. *Brain and Language*, 59(2):334–366, 1997.

[51] M. Minsky. Steps toward artificial intelligence. *Proceedings of the Institute of Radio Engineers*, 49(1):8–30, 1961.

[52] M. L. Minsky and S. A. Papert. *Percoptrons: An Introduction to Computational Geometry.* MIT Press, Cambridge, MA, USA, 1st edition, 1969.

[53] T. M. Mitchell. *Machine Learning.* McGraw-Hill, New York, NY, USA, 1st edition, 1997.

[54] P. Monaghan and R. Shillcock. Connectionist modelling of the separable processing of consonants and vowels. *Brain and Language*, 86(1):83–98, 2003.

[55] L. Nickels. Therapy for naming disorders: Revisiting, revising, and reviewing. *Aphasiology*, 16(10/11):935–979, 2002.

[56] D. C. Plaut. Relearning after damage in connectionist networks: towards a theory of rehabilitation. *Brain and Language*, 52:25–82, 1996.

[57] D. C. Plaut. Graded modality-specific specialization in semantics: a computational account of optic aphasia. *Cognitive Neuropsychology*, 19(7):603–639, 2002.

[58] D. C. Plaut and T. Shallice. Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, 10(5):377–500, 1993.

[59] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[60] A. M. Raymer and L. J. G. Rothi. Clinical diagnosis and treatment of naming disorders. In A. E. Hillis, editor, *The Handbook of Adult Language Disorders*, pages 163–182. Psychology Press, New York, NY, USA, 2002.

[61] A. Roach, M. F. Schwartz, N. Martin, R. S. Grewal, and A. Brecher. The Philadelphia Naming Test: Scoring and rationale. *Clinical Aphasiology*, 24:121–133, 1996.

[62] R. Rojas. *Neural Networks. A Systematic Introduction.* Springer-Verlag, 1st edition, 1996.

[63] F. Rosenblat. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[64] F. Rosenblat. *Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanism.* Spartan Books, New York, NY, USA, 1st edition, 1962.

[65] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*, volume 1: Foundations, pages 318–362. MIT Press, Cambridge, MA, USA, 1986.

[66] D. E. Rumelhart and J. L. McClelland. On learning the past tenses of English verbs. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*, volume 2: Psychological and Biological Models, pages 216–271. MIT Press, Cambridge, MA, USA, 1986.

[67] D. E. Rumelhart, J. L. McClelland, and the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.*, volume 1: Foundations. MIT Press, Cambridge, MA, USA, 1986.

[68] D. E. Rumelhart, J. L. McClelland, and the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 2: Psychological and Biological Models. MIT Press, Cambridge, MA, USA, 1986.

[69] W. Ruml and A. Caramazza. An evaluation of a computational model of lexical access: comments on Dell *et al.* (1997). *Psychological Review*, 107:609–634, 2000.

[70] W. Ruml, A. Caramazza, J. R. Shelton, and D. Chialant. Testing assumptions in computational theories of aphasia. *Journal of Memory and Language*, 43(2):217–248, 2000.

[71] S. L. Saltberg. On comparing classifiers: A critique of current research and methods. *Data Mining and Knowledge Discovery*, 1:1–12, 1999.

[72] M. F. Schwartz. Progress in testing the interactive two-step model of lexical access. *Brain and Language*, 91(1):68, 2004.

[73] M. F. Schwartz, G. S. Dell, and N. Martin. Testing the interactive two-step model of lexical access: Part I. Picture naming. *Brain and Language*, 91(1):71–72, 2004.

[74] T. J. Sejnowski and C. R. Rosenberg. Parallel networks that learn to pronounce English text. *Complex Systems*, 1:145–168, 1987.

[75] S. Siegel and N. J. Castellan. *Nonparametric Statistics for Behavioral Sciences*. McGrawn-Hill, New York, NY, USA, 2nd edition, 1988.

[76] D. Specht. Probabilistic neural networks. *Neural Networks*, 3(1):109–118, 1990.

[77] I. M. Tallberg. The Boston naming test in Swedish: Normative data. *Brain and Language*, 94(1):19–31, 2005.

[78] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, San Diego, CA, USA, 3rd edition, 2006.

[79] A. E. Thyme. *A Connectionist Approach to Nominal Inflection: Paradigm Patterning and Analogy in Finnish.* PhD thesis, University of California, San Diego, 1993.

[80] A. Tikkala. *Neural Networks and Lexical Processing.* PhD thesis, University of Kuopio, Kuopio, Finland, 1997.

[81] A. Tikkala. Suggestion for a neural network model for simulating child language acquisition. *Nordic Journal of Linguistics*, 21:47–64, 1998.

[82] A. Tikkala, H.-J. Eikmeyer, J. Niemi, and M. Laine. The production of Finnish nouns: A psycholinguistically motivated connectionist model. *Connection Science*, 9(3):295–314, 1997.

[83] A. Tikkala and M. Juhola. A neural network simulation method of aphasic naming errors: properties and behavior. *Neural Computing & Applications*, 3:191–201, 1995.

[84] A. Tikkala and M. Juhola. A neural network simulation of aphasic naming errors: Network dynamics and control. *Neurocomputing*, 13(1):11–29, 1996.

[85] L. J. Tippett and M. J. Farah. A computational model of naming in Alzheimer's disease: unitary or multiple impairments? *Neuropsychology*, 6(1):3–13, 1994.

[86] A. Tsakonas, G. Dounias, J. Jantzen, H. Axer, B. Bjerregaard, and D. G. von Keyserlingk. Evolving rule-based systems in two medical domains using genetic programming. *Artificial Intelligence in Medicine*, 32(3):195–216, 2004.

[87] A. Vauhkonen and M. Juhola. Convergence of a spreading activation neural network with application of simulating aphasic naming errors in Finnish language. *Neurocomputing*, 30(1-4):323–332, 2000.

[88] E. K. Warrington. The graded naming test: A restandardisation. *Neuropsychological Rehabilitation*, 7(2):143–146, 1997.

[89] J. F. Wright and K. Ahmad. The connectionist simulation of aphasic naming. *Brain and Language*, 59(2):367–389, 1997.