COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS, PHYSICS AND
INFORMATICS

RECURRENT NEURAL NETWORK MODEL OF PHONOLOGICAL DEVELOPMENT USING DISTRIBUTED REPRESENTATIONS

MASTER THESIS

2019
BC. ENDRE HAMERLIK

COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS

RECURRENT NEURAL NETWORK MODEL OF PHONOLOGICAL DEVELOPMENT USING DISTRIBUTED REPRESENTATIONS

MASTER THESIS

Study programme:	Cognitive Science
Field of Study:	2503 Cognitive Science
Training work place:	Department of Applied Informatics
Supervisor:	doc. RNDr. Martin Takáč

Bratislava, 2019
Bc. Endre Hamerlik

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

71307379

THESIS ASSIGNMENT

Name and Surname:	Bc. Endre Hamerlik
Study programme:	Cognitive Science (Single degree study, master II. deg., full time form)
Field of Study:	Cognitive Science
Type of Thesis:	Diploma Thesis
Language of Thesis:	English
Secondary language:	Slovak

Title: Recurrent neural network model of phonological development using distributed representations

Annotation: Takac et al (2017) created a model of early vocabulary and phonology development. The model fits empirical data about neighbourhood density phenomenon in normal and late talkers. The model used localist representations of word meanings and phonemes, which made it easier to analyse the model, but could not account for particular semantic and phonologic errors observed in children.

Aim: The goal of the thesis is to:
a) change representations to distributed,
b) verify that the model still fits the neighbourhood density data
c) explore kinds of semantic and phonological errors the modified model produces.

Literature: Takac, M. and Knott, A. and Stokes, S.: What can Neighbourhood Density effects tell us about word learning? Insights from a connectionist model of vocabulary development. Journal of Child Language 44(2). 346-379. 2017
Dell, G., Juliano, C. & Govindjee, A. (1993). Structure and content in language production: a theory of frame constraints in phonological speech errors. Cognitive Science 17(2), 149–195.

Supervisor:	doc. RNDr. Martin Takáč, PhD.
Department:	FMFI.KAI - Department of Applied Informatics
Head of department:	prof. Ing. Igor Farkaš, Dr.
Assigned:	13.09.2017
Approved:	15.12.2017

prof. Ing. Igor Farkaš, Dr.
Guarantor of Study Programme

.....
Student

.....
Supervisor

Acknowledgment: I would like to thank my supervisor doc. RNDr. Martin Takáč PhD., for his motivation and his constructive feedback during the work on present thesis, and also prof. Kornai András for his helpfulness and guidance in NLP tasks. I would also like to thank my enchanting wife, Csilla for her restless patience and care during the "focusing-period“.

Abstract

The goal of the presented thesis is to build a psycholinguistically plausible computational model of phonological development, inspired by the existing models of Takac, Knott, and Stokes, 2017 and Dell, Juliano, and Govindjee, 1993. Takac's current model of phonological development is considering localistic representations of phonemes and word-meanings (one-hot vectors, more technically). Moreover, according to their study, even the localistic neural activations provide an explanation for several patterns shown by infants during phonological development.

Merely all the neurobiological or even neuro-imaging studies claim, that neural activations are distributed within the brain; Moreover, neither an area, where an activation pattern is exceeding during a given treatment, can be considered as the brain area responsible for phenomenon invoking the activation. Suggesting, that every brain area respond to different treatments to different extent. (Tremblay and Dick, 2016). Our thesis aims to reproduce this phenomenon within the hidden layer of our Artificial Neural Network.

Thus, in the experimental part of our work, we developed a Simple Recurrent Neural Network (SRN) model, trained under circumstances of the source study by Takac, Knott, and Stokes, 2017. However, in our study, phoneme and even the meaning representations are modified in a neurobiologically more plausible way. The model is tested for performance during the training, analogically to children learning their first 300 words, approximately in their first two years. Nevertheless, even the methodology of analyzing infant's data will be maintained, accounting for phonological Neighborhood Density and related phenomena.

Our results suggest, that the phonological Neighborhood Density effect is not necessarily related to localness of representations. In line with our novel findings, we provide an alternative explanation of the present phenomenon.

Furthermore, speech error analysis (in line with Dell, Juliano, and Govindjee, 1993) is presented on the speech errors produced by our model, in order to support our hypotheses based on the Parallel Distributed Processing Paradigm.

Abstrakt

Cieľom tejto práce bolo vyvinúť psycholingvisticky plauzibilný model fonologického vývinu, inšpirovaný prácami Takac, Knott a Stokes, 2017 a Dell, Juliano a Govindjee, 1993. Takáč používal vo svojej práci koncept lokalistických reprezentácii (presnejšie one-hot vektory) tak pre kódovanie foném ako pre kódovanie významu slov. Navyše, ich výsledky a ich analýza nasvedčujú tomu, že práve za pomoci takýchto lokalistických aktivácii neurónov možno psycholingvistické fenomény fonologického vývoja detí vysvetliť.

Na druhej strane, súčasné neurobiologické a neurozobrazovacie štúdie tvrdia, že vzory neurálnych aktivácii v mozgu sa javia byť paralelnými a distribuovanými. Zároveň, aj keď v danej mozgovej časti vyvoláme lokálne zväčšenú aktivitu (aj najväčšiu), nemôžeme túto mozgovú časť považovať za zodpovednú pre fenomén, ktorým som danú aktiváciu vyvolal. Totiž, jednotlivé časti mozgu majú tendenciu reagovať inými aktivačnými črtami na rozdielne vzruchy (Tremblay a Dick, 2016). Náš model sa snaží zachytiť práve tento jav vo svojej skrytej neurálnaj vrstve.

Totiž, v praktickej časti našej práce sme vyvinuli model, jednoduchú rekurentnú sieť, ktorá sa trénovala pri podmienkach východiskovej štúdie Takac, Knott a Stokes, 2017. Avšak, s rozhodujúcim rozdielom: s distribuovanými vektorovými reprezentáciami sme zakódovali fonémy - v súlade s prácou Dell, Juliano a Govindjee, 1993 - aj významy jednotlivých slov (za pomoci GloVe vektorov Pennington, Socher a Manning, 2014). Pri experimente sa zaznamenávali všetky aspekty učenia, všetky produkované sekvencie foném, pre porovnanie s deťmi počas podobného fonologického vývinu, pri učení svojich prvých zhruba 300 slov, v prvých dvoch rokoch svojho života. Zároveň, aj metodológia analýzy a vyhodnocovania dát sú prebraté, sústreďujúc sa na efekt fonologického susedstva (Neighborhood Density) a príbuzných efektov.

Naše výsledky ukazujú, na rozdiel od Takac, Knott a Stokes, 2017, že lokalistické reprezentácie foném a významov nie sú nutnou podmienkou pre produkovanie efektu fonologického susedstva. Zároveň, ponúkame iné vysvetlenie tohto efektu, vzhľadom na naše nové zistenia. Navyše, analýzou rečových chýb si posvietime na vnútornú dynamiku neurónovej siete, pre podporu našej hypotézy opierajúc sa o konekcionistickú argumentáciu.

Contents

List of Figures

List of Abbreviations

ND	**N**eighborhood **D**ensity (phonological)
WF	**W**ord **F**requency
ROI	**R**egion **O**f **I**nterest
PDP	**P**arallel **D**istributed **P**rocessing
AF	**A**rcuate **F**asciculus
ANN	**A**rtificial **N**eural **n**etwork
RNN	**R**ecurrent **N**eural **N**etwork
BPTT	**B**ackpro**p**agation **T**hrough **T**ime
GD	**G**radient **D**escent
SGD	**S**tochastic **G**radient **D**escent
CDI	**C**ommunicative **D**evelopment **I**nventory
IPA	**I**nternational **P**honetic **A**lphabet
LSA	**L**atent **S**emantic **A**nalysis

For my little son, Endi. . .

Introduction

Present work attempts to enclose the biological, philosophical and computational views on phonetic development and phonetic mechanisms within a model of the phonological system and its development. Based on findings of (Sirigu et al., 1998) we consider higher-level speech structures like syntax to be executed in frontal lobes so that we are building a model of the lower-level speech-mechanisms, interfacing the above mentioned higher-level systems as a black-box. The functional architecture of speech mechanism (first proposed by Lichtheim, 1885, shown in Figure 1; Which is trying to catch the core of phonetic-processing; The interaction between the output of Wernicke's and Geschwind's territory, Broca's area, and the sensory-motor cortex.

Later, brain- and neuro-imaging techniques helped us to recognize, neither the speech nor any other brain function is as simple, feedforward, as models of the classical theory. With the rise of neurobiological pieces of evidence depicting other brain areas, and with improvements of Artificial Neural Networks, a new paradigm formed, emphasizing the need and the causalities of the dense interconnection within the neural networks in the brain, called Parallel Distributed Processing. However, the pioneering connectionist works were published in the mid-1980s, distributed representations of concepts in the brain are in question even nowadays. Thus, a current model of phonological development (Takac, Knott, and Stokes, 2017) is considering localistic representations of phonemes and word-meanings. Moreover, it has found, that even the localistic neural activations provide an explanation for several patterns shown

by infants during phonological development. We aim to research this contradiction. Since merely all the neurobiological or neuro-imaging studies claim, that neural activations are distributed within the brain. Also, even if an activation pattern is concentrated within such a brain-area, it cannot be considered as brain area responsible for the given treatment (Tremblay and Dick, 2016). While in turn, another neural focal point is considered to include the area in processing a completely different treatment.

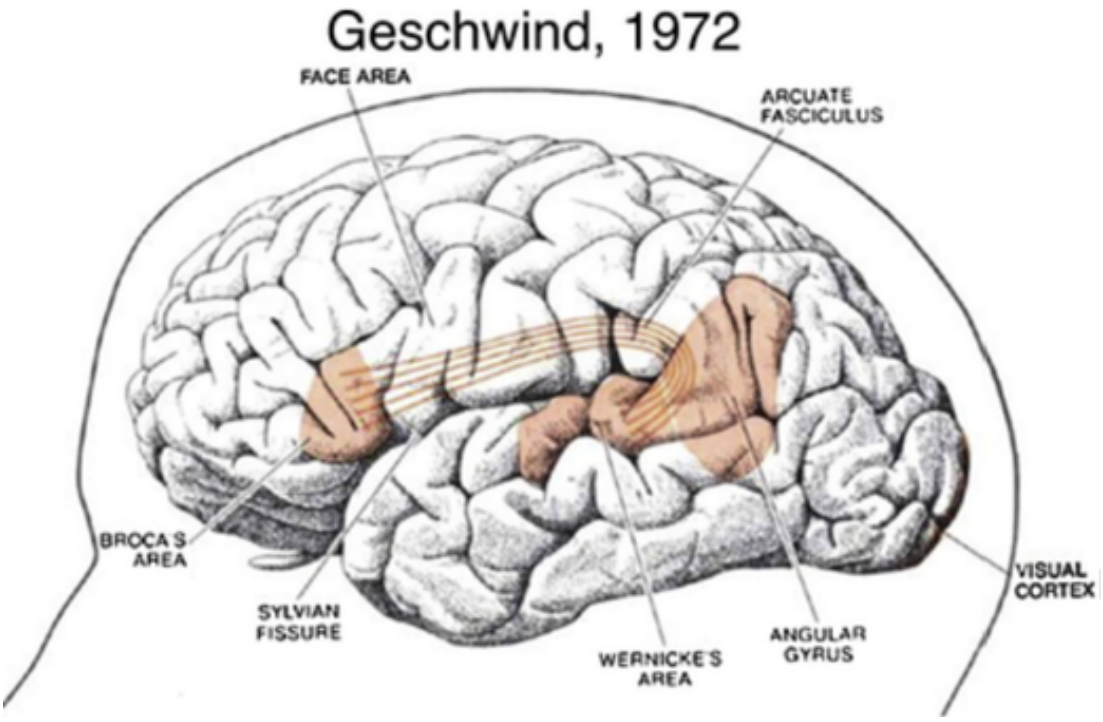

FIGURE 1: Brain areas involved in speech producing by Geschwind, 1970

Our thesis' methodology is scientific modeling, more precisely a validational reimplementation of a source study by (Takac, Knott, and Stokes, 2017) under modified, neurobiologically more plausible conditions. The model in question is a model of phonological development, which is to reproduce the

circumstances, under which infants learn their first ca — 300 words, approximately in their first two years. Nevertheless, even the methodology of analyzing infant's data will be maintained.

However, we are aware of how small, and low-level features of speech production are we modeling, we should emphasize, when such a model gets trained to produce sequences of phonemes, it can even more. Its predictions of new phonemes necessarily express phonotactic constraints of the exposure language. Such as the mappings of meanings, to their phonological forms. Therefore, another pioneering study by (Dell, Juliano, and Govindjee, 1993) will be analyzed, and to a certain extent reproduced; mainly in the field of phonological speech errors and related distributed representations. Present work is to add a level of complexity to existing models, and spice of biological plausibility.

The goal of the present thesis is to compare two perspectives on the phonological part of speech production system through simulation of its development. Unlike the source studies, it will use meaningful parallel distributed vector representations in all levels of representations, articulatory features in case of phonemes, semantic features for the meaning representations. The time-course of the phonetical learning will be followed, and all available features will be likened to the infant's dataset, comparatively analyzed by (Takac, Knott, and Stokes, 2017).

The rest of the thesis is structured as follows. The first chapter reviews relevant previous work: in Chapter 1 we start with investigating the neurobiological background of speech, from the classical „Broca–Wernicke" model to the neuroimaging studies, which were to indicate speech activation patterns in the brain are distributed, running in parallel. Next, a current linguistic approach is presented in Chapter 2, focusing on the speech errors, to be reproduced in our experiments, such as the Neighborhood-density and related phenomena in Section 2.3, where we discuss reproducible aspects of children's phonological

development. The next part summarizes the methods to use in the experiment: Recurrent Neural Networks, parameter optimization techniques, and the two specific applications of such networks for tasks similar to ours. In Chapter 4 the experimental setting is presented. First, the differences are presented between our work and the source studies. Thereafter, our distributed phoneme- and meaning-representations exhibited and discussed in context. Finally, in Chapter 6 we present our results, which are novel in several ways.

Chapter 1

The classical Neuro-biological models of speech production

1.1 Neurobiological perspective

According to the classical "Broca–Wernicke–Lichtheim-Geschwind“ neurobiological perspective, speech-production and -understanding is associated with several brain areas within the cerebral cortex (See Figure 1). In the case of spoken-language-processing, auditory cortex is involved first. Then, in the temporal lobe, the brain matches the phoneme against the vocabulary. Here, in the Wernicke’s area is the meaning assigned to words, and language comprehension is achieved. Then the arcuate fasciculus – AF (a bundle of bidirectional nerve fibers) transmits the activation to Broca’s area in the frontal lobe, where the pre-motor program is being produced for the motor cortex, which controls the muscle movements even for speech production.

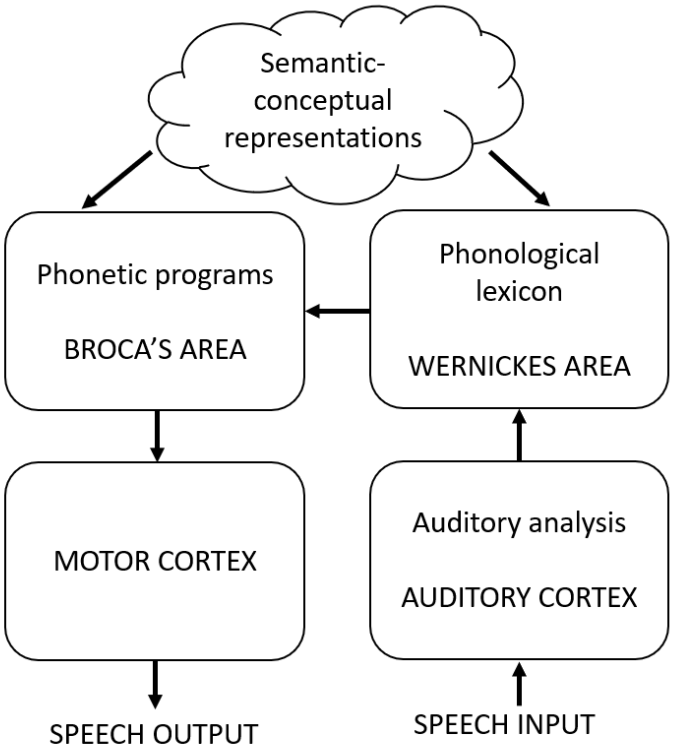

FIGURE 1.1: Wernicke-Lichtenstein's schema of Speech production

Recently, almost every part of the classic theory is in question; Especially the AF being a single fiber pathway, magically ensuring the significant associations between the two, functionally specific areas: Wernicke's and Broca's Graves, 1997. Nevertheless, the most contemporary models are proposing more complex architectures, including areas that had not been considered as parts of speech comprehension or production working as a substantial modular (distributed) system in parallel Tremblay and Dick, 2016. On the one hand, it means, that even our model – considering just few processing nodes (motivated by areas like Broca's, Wernicke's, etc.) - would be so simplifying, to brand it as biologically plausible. On the other hand, it ensures foundations for Parallel Distributed Processing (PDP) paradigm by Rumelhart, McClelland,

and Group, 1987 used in the latter parts of our work.

1.2 The Parallel Distributed Signals and their processing

PDP theory suggests that linguistic units, such as phonemes or semantic concepts are represented by their features rather than a specific neuron, producing Hit-NoHit binary output as it is considered by the opposing localistic approach. Therefore, we are to use feature vectors for the representation of our linguistic units. Thanks to Pulvermüller's event-related MRI experiment Pulvermüller et al., 2006 we know, articulatory features of phonemes – we are to catch in our representations – are processed via motor circuits in the superior temporal lobe. Although the activation map is different when speech is perceived or produced, their involvement is unquestionable. Even the articulatory planning of phonemes like /p/ or /t/ evokes specific activation in the primary and premotor cortex, based on the phoneme's articulatory features. E.g., for /p/ we get an activation pattern in the motor area related to lips; respectively for /t/ we get adequate activation in the motor area of the tongue. For details see Figure 1.2.

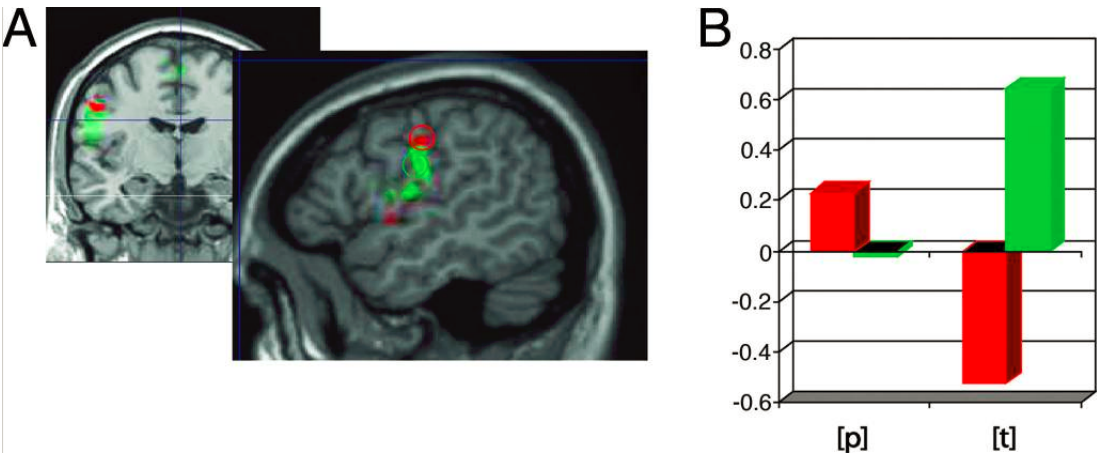

FIGURE 1.2: *„Phoneme-specific activation in precentral cortex during speech perception. It is showing activation during lip (red), and tongue (green) movement along with the Region Of Interests (ROIs) centered in precentral gyrus. (B) Activation (arbitrary units) during listening to syllables including [p] and [t] in precentral brain areas where pronounced activation for lip (red) and tongue (green) movements were found. The significant interaction demonstrates differential activation of these motor areas related to the perception of [p] and [t].“* (Source: Pulvermüller et al., 2006

All in all, we can sum up, that the articulatory motor programs in precentral cortex take a considerable part in phoneme encoding in the brain (if not responsible at all). Therefore, we implement Our Distributed phoneme encodings, based on these distinctive articulatory features in Section 4.3

Chapter 2

Linguistic theory

2.1 Frame constraints in Language

A standard assumption of linguistic theory e.g. by Smith, Kosslyn, and Barsalou, 2007 states, that words and sentences are properly described as a merger of structure and content. A word description contains both a string of phonemes (its structure) and a semantic entity – lexeme, describing the content.

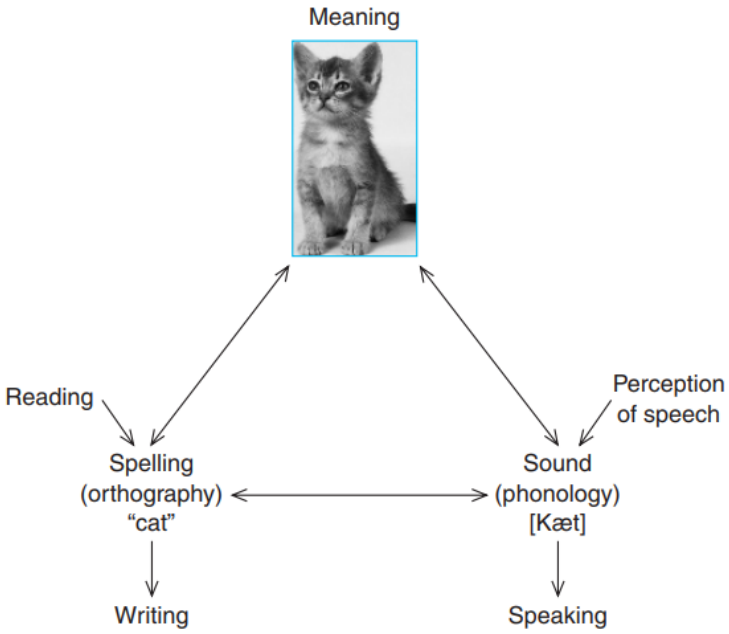

FIGURE 2.1: Triangle model of speech by Smith, Kosslyn, and Barsalou, 2007

The historical view of language theories suggests, that linguistic content and linguistic structure are produced by two distinct mechanisms; Where mental lexicon should ensure the content, and as a slot is inserted to a frame of rules, called frame constraints. These constraints e.g., can set the order of phonemes. Let's consider the word *dog*, which has *Consonant-Vowel-Consonant* (*CVC*) phoneme-order, withal, being the first consonant *onset*, rhyme sub-syllabic unit containing the nucleus, and the final in the remaining part. These phonological-level features make up a phonological frame. For it's tree-representation see Figure 2.2

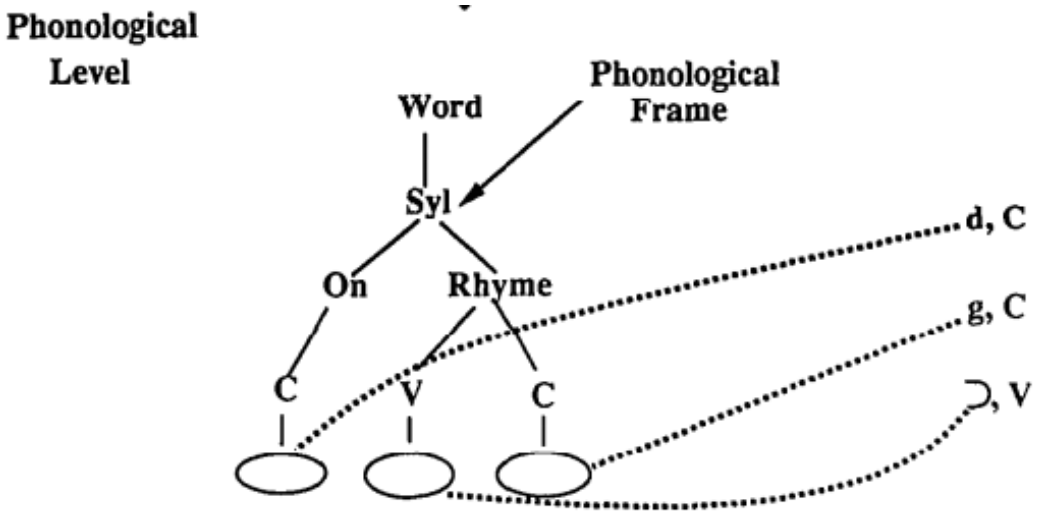

Figure 1. Syntactic and phonological representations in theories of language productior

FIGURE 2.2: Syntactic- and phonological-tree representation in theories of language production (Source: Petruck, 1996)

Phonological frames long have been the basis of language models. E.g., Dell, 1986 proposed a set of hard-wired rules,https://www.overleaf.com/project/5bebe452f04 taking into account prohibition of invalid phoneme-sequences, as a part of the definition for a specific frame. Our model's theoretical background is opposing this kind of simplifying rule-based approach (just like the later model of Dell, Juliano, and Govindjee, 1993, we were inspired by). Here we do not want to

argue or contradict the frame-theories, we present them in order to better understand speech error's nature, and to get used to the terminology of studies exploring and classifying speech errors.

We will consider two classes of frame constraints:

- general constraints, which apply to all phonological errors, both movement and nonmovement errors. E.g. **a reading list** → **a leading list** (phonological movement error, movement of /l/, by Fromkin, 1971)
- movement constraints, which apply only to movement errors. E.g. **Department** → **jepartment** (substitution of textit/j/ for textit/d/, a nonmovement error by Stemberger, 1983a)

2.2 General constraints

2.2.1 Phonotactic Regulatory Effect

Phonotactic Regulatory Effect suggests, that speech errors almost always produce phoneme-sequences that occur in the language; So that the frame reduces all possible phoneme-sequences to legal ones. Many studies refer to this effect as „first law," since it has been observed in about every speech error collection MacKay, 1972. E.g., Stemberger, 1983a found this effect violated in less than 1% of his error corpus. Therefore, no available frame should accept an invalid sequence like „dlorm" in the following example.

2.2.2 Consonant-Vowel Category Effect

This effect appears to be the strongest one among the evidence of speech errors.MacKay, 1970 and Stemberger, 1983b did not report any error-instance proscribing this effect. The rule, itself, says, that the basic phonological categories, consonants, and vowels, should be respected and remained within a

flip or replacement speech error, which means, that vowels can be replaced by vowels and consonants with consonants solely. However, there can be found some counterexamples (such as Mexico being spoken as *„meksakL“*, which can be analyzed as /o/ replaced by syllabic /L/, in Stemberger, 1983b; But they are scarce.

2.2.3 Syllabic Constituent Effect

2.2.4 Initialness Effect

It assumes, initial or onset consonants are more likely to slip than non-initial ones, with two related effects. In behind can be several causes. As can be seen in Figure 2.2, the sub-syllabic structure of a syllable is often divided into onset and the rest (rhyme, coda ...). In such a way, the first phoneme is more often a sub-syllabic unit, much easier to flip Shattuck-Hufnagel, 1987. Pieces of evidence suggest, approximately 80% of consonant-movement errors concern initial consonants. There is also a tendency to of word-initial and even syllable-initial consonants to flip more often than syllable-final ones. According to Dell's review of the topic, the standard explanation for this effect is that word- and syllable-initial consonants are structurally distinct in the phonological frame. Meaning, that they form a phonological constituent on the top of the hierarchical structure of the word, which results from the easier division of the word bat as b-at than ba-t.

These four constraints long have been considered as the foundation of speech production theories of nowadays. However, many findings in speech-related fields (neuroimaging, neurology, computational linguistics) put it in question or refute their biological plausibility; frame constraints remain a shred of quantitatively massive evidence, useful to consider; Especially when evaluating a model of speech production. A model, where no hand-made rules were used,

although, producing (nearly) the same phenomena. We will handle these errors, and constraints, considered to form them, to validate our model's behavior and outputs.

2.3 Phonological Neighborhood density and related empirical phenomena

Phonological Neighborhood Density (ND) is a measure referring to the number of words that can be generated by inserting or deleting one phoneme in a target word or replacing one phoneme with another phoneme in the same position. A word has high ND if there are many words phonologically similar to it. High ND words are more easily learned by infants of all abilities (e.g. Storkel, 2009, Stokes, 2014). From the point of phonological view, Neighborhood Density is a measurement tool of neighborhood size, that refers to the number of words that could be generated by replacing, inserting or deleting one phoneme in some word by another phoneme in the same position.

Recent linguistic studies found that the pND phenomena is correlating to the greatest extent with word learnability (Stokes stokes2010neighborhood, stokes2014impact) for infants of all abilities. By all abilities we mean Late talkers included. A child belonging to the slowest 10-20% of word learners is considered to be Late talker. A Late talker learns by approximately 20% fewer words, than the normal ones until their 2^{nd} birthday. Moreover, Late talkers tend to learn higher-ND words in preference; Rather, than children with ordinary learning course.

Given a specific word from the dictionary and its ND, the age of acquisition is the measure to be followed. It states when having a given word been learned - in what period. For that matter, the trend is decreasing for the ND with an increase of Age of Acquisition. It is supported by empirical evidence of Stokes, 2010; Stokes et al., 2012; Stokes, 2014.

However, the implications of ND-effect are well known, and even supported by empirical evidence; its origin remains in question. On the one hand, high-ND words tend to contain phoneme sequences, which have a higher-than-average frequency in exposure language. Therefore, it long has been considered being purely frequency-phenomena and can be explained by statistics of the language. On the other hand, many studies, including Takac, Knott, and Stokes, 2017, successfully isolated ND-effect from other frequency-related phenomena, so that the focus of research has been moved to cognitive word representations that lead to this effect. Most of the informal explanations of ND-effect suggest, that novel words are easier to learn, if they are phonologically similar to known words. Thus, their phonological representations can be coded as modified existing representations. (Storkel and Lee, 2011). Besides the opposing views on the phonological development, computational linguists agree, that during the learning of either a child or a model the above-described trait can be observed, the so-called Conspiracy effect.

Conspiracy effect stands for, and to a certain extent even explains, how the updates of neural interconnections control the learning course of a neural network. In our particular case, when learning two phonologically similar words, one word's weight update contributes to the potentially needed weight update for its phonologically similar word pair. In such a way, when the network is trained for the second (similar) word, much less effort is needed concerning the updates of connections. We are going to research the direction of causality between Conspiracy effect and phonological Neighborhood Density in the following.

Chapter 3

Computational theory

3.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) are biologically motivated computing systems. These systems are able to learn patterns and tasks based on examples, mainly with supervision. Supervision means labeled examples, i.e., a task with a solution so that the network's parameters are updated due to the difference between the network's output and the teaching signal – the supervision (the right answer).

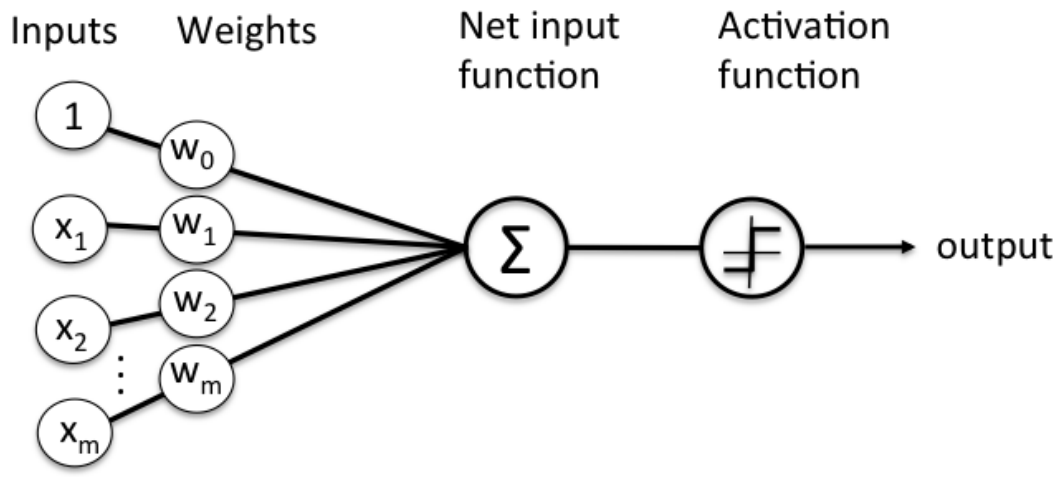

FIGURE 3.1: Schema of a single perceptron

The building blocks of ANNs are perceptrons, single neurons. These units (in Figure 3.1) linearly combine their inputs, in other words, sums up the input signals multiplied by appropriate weights to produce the net signal, the net input function. This signal used to be saturated, normalized to range $[0,1]$-unipolar or $[-1,1]$-bipolar. From a mathematical perspective such network's (in Figure 3.1) output o, is computed as:

$$o = f(net) = f(\overline{w} \cdot \overline{x}) = f\left(\sum_{j=1}^{n+1} w_j x_j\right) = f\left(\sum_{j=1}^{n} w_j x_j - \theta\right) \tag{3.1}$$

Where $\vec{x}$ is the input vector, the pattern, and $\vec{w}$ is the vector of weights. The dot product of w and x to be marked as net as the net's signal. The function f , encompassing that signal is the Activation function of the network. Since one of the main advantages of neural networks is the ability to solve non-linear problems, mostly non-linear neurons are used. Here the activation functions must be non-linear ones. The most commonly used are Logistic function, a.k.a Sigmoid (Unipolar) or Tanh (Bipolar). In case of linear activations Rectified Linear Units are the most common choice. For details see Figure 3.2

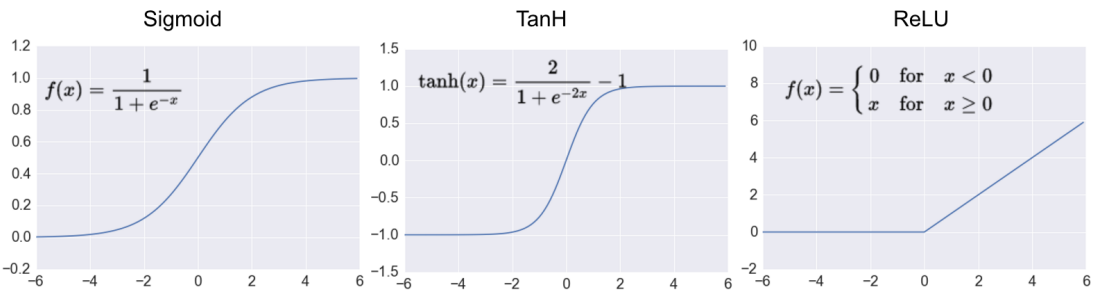

FIGURE 3.2: The three most frequently applied activation functions. In order: Sigmoidal, Hyperbolic tangential, and Rectified Linear Unit

As above mentioned, after the forward pass (obtaining network's output

based on inputs) the weights must be updated to produce more accurate output in the future. In our project, we chose the most widely used algorithm for the update: The Stochastic Gradient Descent (SGD). SGD is an iterative optimization algorithm, optimizing the network's parameters, the weights in order to produce minimal error. Error in time t, $E^{(t)}$ is mostly computed as (squared) sum of the element-wise difference between the network's output, $o_k^{(t)}$ and the target vector $d_k^{(t)}$.

$$E^{(t)} = \frac{1}{2} \sum_{k=1}^{K} \left(d_k^{(t)} - o_k^{(t)} \right)^2 \tag{3.2}$$

Then the weight updates are computed as the partial derivative of the Error with respect to the weights w multiplied by α, the learning rate, controlling the speed of learning.

$$\Delta w_{kj} = -\alpha \frac{\partial E_p}{\partial w_{jj}} = -\alpha \frac{\partial E_p}{\partial \left(ne_k \right)} \frac{\partial \left(net_k \right)}{\partial w_{ij}} = \alpha \delta_{ok} y_j \tag{3.3}$$

Where δ_{ok} is computed as:

$$\delta_{ok} = -\frac{\partial E_p}{\partial \left(net_k \right)} = -\frac{\partial E_p}{\partial o_k} \frac{\partial o_k}{\partial \left(net_k \right)} = \left(d_{pk} - o_{pk} \right) f_k' \tag{3.4}$$

Here d is the target, o is the network's output and f_k' is the derivative of activation function with respect to the net signal. At the end of the day we can assume, that the k-th neuron's j-th weight's change vector's magnitude will depend on the difference between target and output, it's direction will be given by the derivative of activation function multiplied by the j-th input; thus, it gives us the slope of the error surface, the direction to the nearest local or global minimum.

$$\Delta w_{kj} = \alpha \left(d_{pk} - o_{pk} \right) f_k' y_j \tag{3.5}$$

3.1.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) represent a class of Artificial Neural Networks, where besides feedforward connections time-delayed connections are present in the network. Their need stems from phoneme sequences of different length, to be processed, discussed in 4.1. These so called Recurrent Neural Networks are widely used in sequence processing problems.

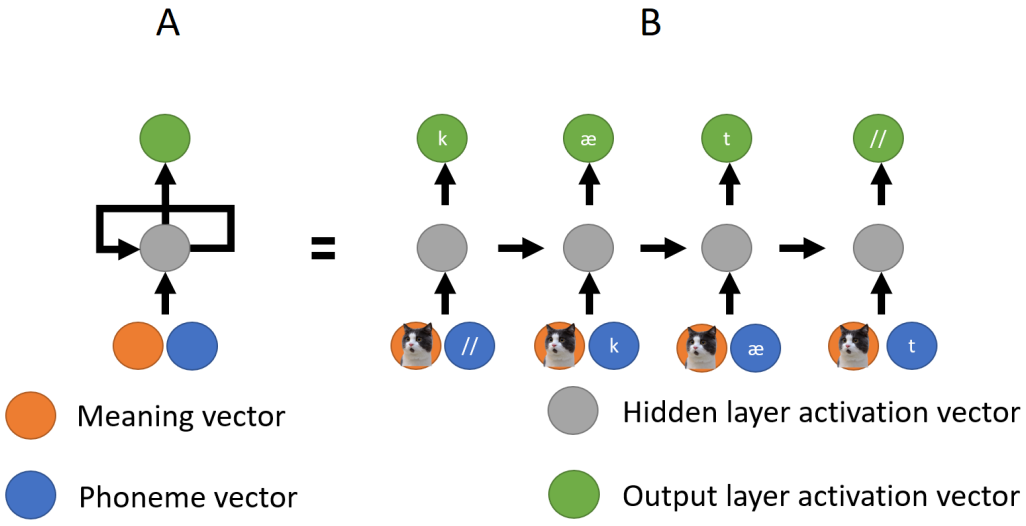

FIGURE 3.3: Schema of a recurrent neural network architecture A:original network, B:unfolded network in 4 timesteps, when processing the word **cat**

RNNs are multi-layer ANNs, in other words: ANNs at least with one hidden layer. The recurrent connection here means a time delayed copy of activations of the whole layer of the network; Mostly, the hidden-, (Elman, 1990) or less often the output-layer (Wilson and Keil, 2001) is delayed. These recurrent connections are fed back to the network as a context layer. The Simple Recurrent Neural Network (In Figure 3.3 A) is a three-layer neural network, where the recurrent connection points from- and to the hidden layer. When unfolded (Figure 3.3 B) the time-resolution becomes apparent.

Consequently, in every time-step t:

- one input vector – x_t and one context vector h_{t-1} presented to the system.
- hidden layer activation is obtained as their dot-products with appropriate weight matrices Wxh and Whh.
- output vector – y_t is obtained after passing hidden layer activation through the weight matrix Why.
- hidden layer activation, h is copied to the context layer.

$$o_{k(t)} = f\left(\sum_{j=1}^{J} Why^{(t)} h_j^{(t)}\right) \tag{3.6}$$

Where:

$$h_j^{(t)} = f\left(\sum_{j=1}^{J} Whh^{(t)} h_j^{(t-1)} + \sum_{i=1}^{I} Wxh^{(t-1)} x_i^{(t-1)}\right) \tag{3.7}$$

Most recurrent settings consider the internal-only feedback loop, copying hidden state to the context, as above. Although, external feedback also used; these are time delaying the output layer's activation to context units – like Dell, Juliano, and Govindjee, 1993 do. Since our network will consider internal-only feedback, the following derivations will reflect exclusively our setup. In our project, the number of time steps will depend on the length of input sequences to be produced. Initial activations in the context layer usually set to zeros. When it comes to Error Backpropagation, a novel approach is needed: Back Propagation Through Time (BPTT) Werbos, 1990. It suggests, that an unfolded network should be treated as a multi-layer network with as many layers, as many timesteps we have - such as in Figure 3.4.

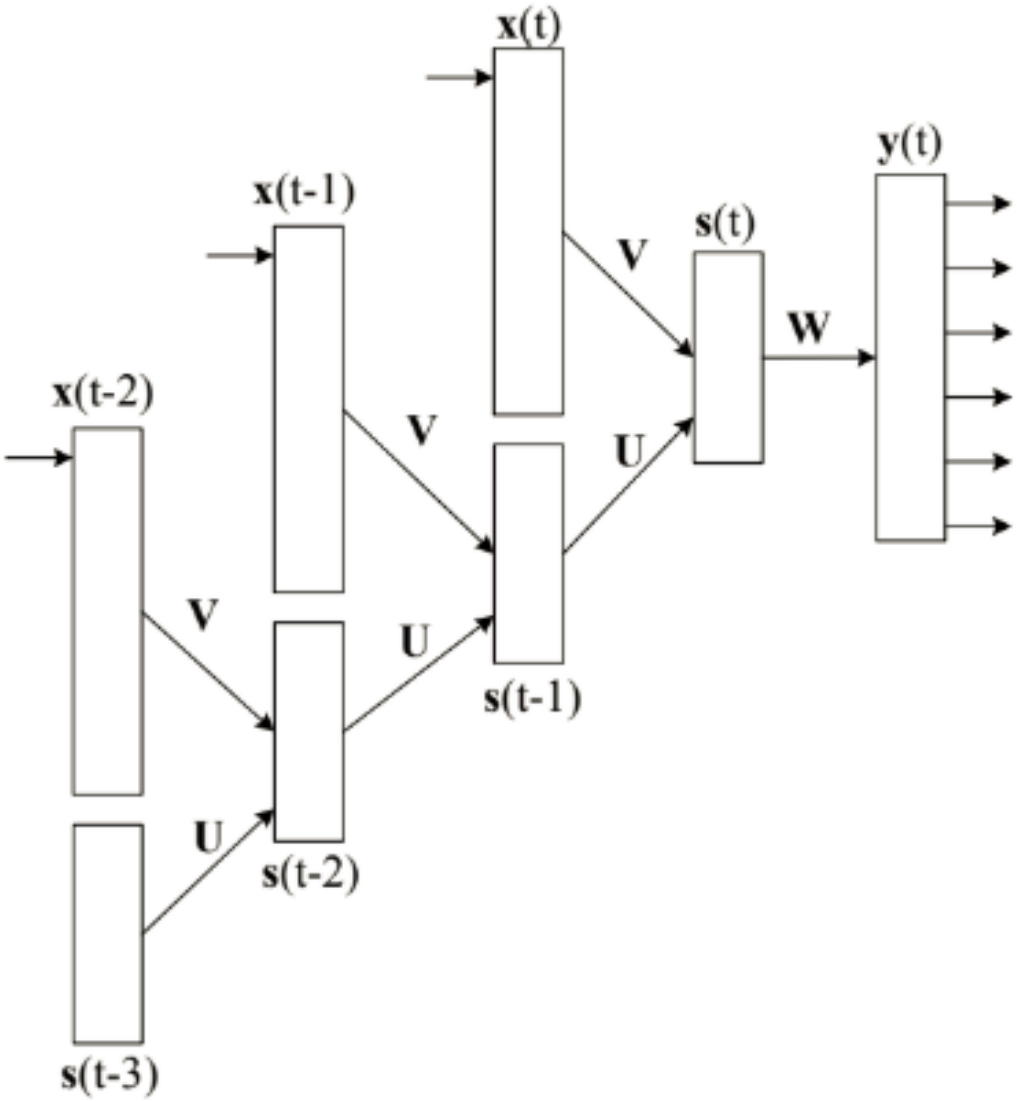

FIGURE 3.4: Recurrent neural network unfolded in time $t = [(t-2);(t)]$

Therefore, the network's parameter updates are computed: For output nodes:

$$\delta_{pk} = -\frac{\partial(C)}{\partial\left(y_{pk}\right)}\frac{\partial\left(y_{pk}\right)}{\partial\left(net_{pk}\right)} = \left(d_{pk} - y_{pk}\right) g'\left(net_{pk}\right) \tag{3.8}$$

For hidden nodes:

$$\delta_{pj} = -\left(\sum_{k}^{o} \frac{\partial(C)}{\partial\left(y_{pk}\right)} \frac{\partial\left(y_{pk}\right)}{\partial\left(net_{pk}\right)} \frac{\partial\left(net_{pk}\right)}{\partial\left(s_{pj}\right)}\right) \frac{\partial\left(s_{pt}\right)}{\partial\left(net_{pj}\right)} = \sum_{k}^{o} \delta_{pk} w_{kj} f'\left(net_{pj}\right) \quad (3.9)$$

The possible activate functions f are in principle the same as in previous section: Logistic, Bipolar or Linear. Cost function can be any function reflecting the difference between output and target, which is differentiable. Sum Squared Error from equation 3.2, as one of the most often used cost function, secures enough simplicity in derivative, and precise difference representation to pick it for our project. So that the weight changes could be easily obtained as:

$$\Delta w_{kj} = \eta \sum_{p}^{n} \delta_{pk} s_{pj} \quad (3.10)$$

For hidden $\rightarrow$ output weight, w_{kj}, the element from kth row and jth column of Why. Then:

$$\Delta v_{ji} = \eta \sum_{p}^{n} \delta_{pj} x_{pi} \quad (3.11)$$

For input $\rightarrow$ hidden weight, v_{kj}, the element from jth row and ith column of Wxh. And finally:

$$\Delta u_{ji} = \eta \sum_{p}^{n} \delta_{pj} s_{ph}(t-1) \quad (3.12)$$

For hidden $\rightarrow$ hidden weight, u_{kj}, the element from jth row and ith column of Whh.

3.1.2 Improvements and adjusting the structure of the network

Momentum

The above derived equations stand for gradient descent (GD). It's standard variant, Stochastic gradient descent (SGD) to be used in our project. It's only

parameter to be adjusted manually is the learning rate, controlling the learning speed. Nowadays, so many ways accessible to improve learning ability of the network by optimizing the GD. Almost every such adjustment's goal is to control the learning rate, to produce a smoother curve on the error surface (see Figure 3.5).

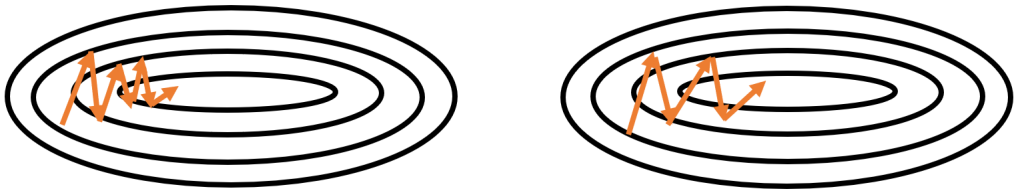

FIGURE 3.5: Comparing SGD without (on the left) and with momentum (on the right) on the contour plot of error surface. Where each point stands for error value, with given parameters. (Source: Montavon, Orr, and Müller, 2012)

Our source studies used momentum, which is a moving average-like algorithm, considering the last updated value v_{t-1} (in time $t-1$) when updating the parameter v_t. Here γ is the momentum factor to be optimized for each task.

$$v_t = \gamma v_{t-1} + \eta \nabla_\theta J(\theta)\theta = \theta - v_t \tag{3.13}$$

Adagrad

Adagrad is an improved momentum-based algorithm, distinguishing between frequent and infrequent activation patterns when updating the parameters. It adapts the learning rate based on these frequencies; larger update for infrequent parameters and smaller ones for frequent parameters.Duchi, Hazan, and Singer, 2011 By the way, Pennington, Socher, and Manning, 2014 also used Adagrad to train GloVe word embeddings (our meaning representation's bank - discussed in 4.3). Adagrad's goal is to modify the learning rate, μ, based on

the previous modified learning rate $\theta_{t,i}$, which is dependent on the previous gradient, $g_{t,i}$. It is calculated as:

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,ii} + \epsilon}} \cdot g_{t,i} \tag{3.14}$$

Where $\theta_{t+1,i}$ denotes updated modified learning rate,μ is the previous learning rate, $G_{t,ii}$ is a diagonal matrix, with sum of squared gradients with respect to the net's output on the diagonal.ϵ is stabilization factor to be fine-tuned for specific tasks; And $g_{t,i}$ is the current gradient.

3.2 Existing Models of phonological development

Processing sequential data (most often language structures like graphemes, phonemes, lexemes ...) is one of the standard tasks for ANNs. Recurrent structures, such as RNNs has shown a great ability of mapping temporal connections between these language structures. Thus, they have been developed and optimized for such tasks Elman, 1990; Elman, 1991. After numerous successful project proving the ANNs' ability to map (even irregular) sequences-to-sequences, computational linguists -mainly connectionists- are to find the relatives of such algorithms, that are implemented in the brain. To enclose these perspectives, choosing a psychological phenomenon, and trying to reproduce its measurable output by a computational model is a standard way. Even our source studies (Takac, Knott, and Stokes, 2017; Dell, Juliano, and Govindjee, 1993) do the same, however in divers depth. So that, such as these source studies, we chose the most straightforward recurrent artificial neural network architecture, the Simple Recurrent Neural Network, the so-called Vanilla Network Elman, 1990 too. Due to its long tradition in word-processing models, it was a safe choice, ensuring processing of phoneme-chains of arbitrary length when combined with the Backpropagation Through Time as training algorithm. Moreover, Takac and Knott, 2015 suggest, it is more than likely, that

even the human brain uses such recurrent connections when it comes to the encoding of sequences such as other sensorimotor routines.

3.2.1 Model of Takac, Knott, and Stokes, 2017

Takac, Knott, and Stokes, 2017 focused on the dynamical aspects of learning; Investigating which words are more likely to be learned earlier under which conditions. The well-defined phenomenon they examined is the Neighborhood Density (ND), thoroughly presented in 2.3. Moreover, they proposed a slightly new explanation of ND-effect and its relation to word meanings. The novelty of their study rooted in generalizing the ND effect based on simulation of realistic phonetical development. Many earlier studies proposed models learning phonological sequences, but with notable limitations, e.g., in the field of word-lengths, or just failed to reproduce cues, showed by the learning children. Nevertheless, their results showed a similar effect than the Late-Talker ND-phenomenon, where a network with limited capacity showed a tendency to learn high-ND words in preference, as Late-Talker children do. However, earlier studies (e.g., Vitevitch and Storkel, 2013) provided mathematical apparatus (an Autoassociative Artificial Neural Network), reproducing the ND-effect, during phoneme-sequence learning; Their control conditions (uniform word-length, uniform word-frequencies) made it harder to compare to the phonological development of real children. Though, their goal was to separate the ND-effect from others (word length, word frequency), correlated with it.

The model, itself was an SRN (Scheme in Figure 3.6), with internal context representation – the network's hidden state is presented as the context with time-delay. The phonemes presented temporally, with a specific word-boundary at the beginning and the end of the sequence. During the production task, the predicted phoneme representation from the output layer is copied to the current input phoneme slot.

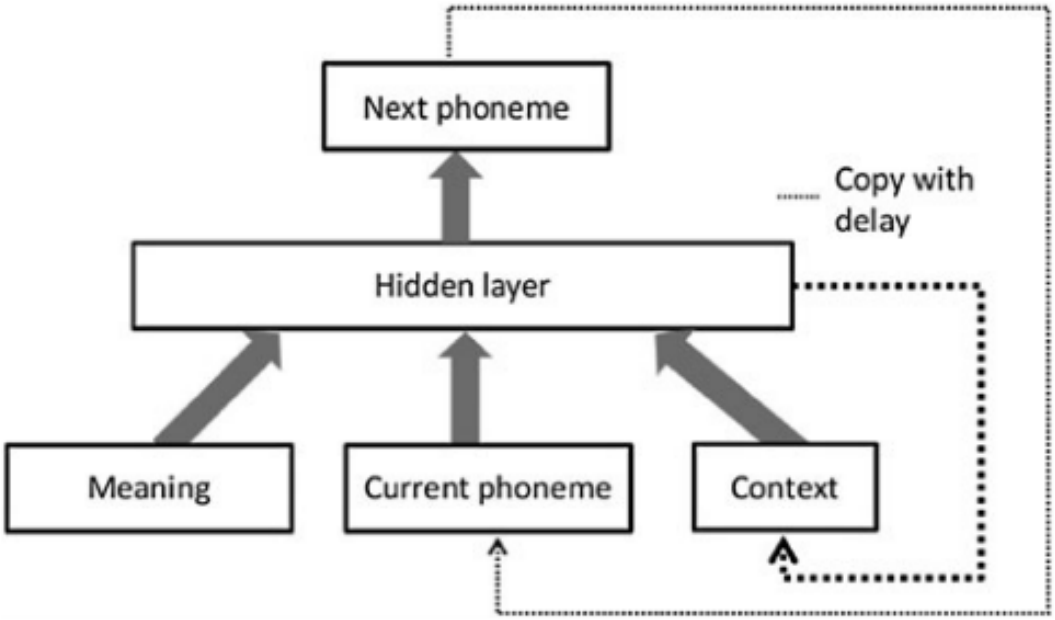

FIGURE 3.6: SRN architecture of the phonological learner model by Takac, Knott, and Stokes, 2017

Its specialty is the composition of the input vector. Hence, it consists of two parts: Meaning- and Phoneme-representation. The network's input-output representations are localistic one-hot vectors. Such vectors must have lengths equal to the number of entities to represent, and naturally, for all distinct entities, another element will be non-zero (mainly 1). So that, these one-hot representations hold information about the identity of the substance and says nothing about its content. The similarity of all localistic representations within a dataset will be equal when comparing them with each other. One localistic vector for the meaning and one for the phonemes are produced and concatenated to the input vector. Since MacArthur-Bates Communicative Development Inventory (CDI) of English words by Fenson et al., 1994 for the „known words" was reduced to 268 words, meaning vector was 268 dimensional, and 44-dimensional vectors for the 44 English phonemes.

Another specialty of their model is to learn even from „not-known words,"

where no meaning input presented to the network (meaning vector is a zero-vector), just the phoneme sequence is passed through the network. For this purpose, they used an extract from the CELEX corpus (Baayen, Piepenbrock, and Van Rijn, 1995 in the amount of 2588 English monosyllables. Training of the network consisted of 100 epochs, presenting the same training set in random order. After the output is obtained, the error is computed and backpropagated through time Werbos, 1990. Between the individual training examples, the context layer was reset, to prevent from the bias of the predecessor. They trained 40 instances of SRNs, with different learning capacities among four classes of networks regarding the hidden-layer sizes: [5, 10, 15 and 20]. So that, each group consisted of 10 artificial participants. Between every two epochs, test session followed, where the network was given a meaning and the word-boundary phoneme. Then, the network's output predicted a phoneme, which was fed back to the input's phoneme slot, until word boundary or more than seven phonemes anticipated. The resulting dataset – the generated phoneme sequences – with all their metadata, was analyzed for to assign to every sequence:

- Age of Acquisition for every word for every instance
- NDaw – Neighborhood Density calculated over all words the network was exposed to
- NDkw – Neighborhood Density calculated over known words only

Their following analysis proved that however, NDaw is due to conspiracy effect, no meaning-specificity found; Unlike in the case of NDkw, which relies on the word-specific biases. Moreover, their regression analysis suggests, that the main effect of ND is due to biasing influence of meanings; proven by effect size, much higher of NDkw. Although, they cannot use this measure when comparing to children data. Thus, it would be tough, if not impossible, to determine the whole vocabulary for a set of individuals. Hence NDkw

should enumerate the neighbors in the dictionary of the child. To sum up, their results show significant similarity to children data, mainly in the field of ND-effect. Thus, they concluded, since meaning plays a crucial role in phoneme sequence generating; even the representation (of the meaning) they chose is indicative. Even their mathematical proof supports that localistic representations of meaning can be the factor which provides a basis for such a realistic behavior of the model.

3.2.2 Model of Dell, Juliano, and Govindjee, 1993

Pioneering work of Dell, Juliano, and Govindjee, 1993 was one of the first models of phonological development with proven speech error convenience. As stated in Chapter 2, speech errors long have been the only empirical basis of speech theories. Frame constraints, which build on these pieces of evidence, were widely accepted so that the explicit frame rules were out of the question. In turn, proponents of PDP paradigm in 1980's published a set of game-changing neural models (Rummelhart's model of past tenses - Rummelhart, 1986; Sejnowski and Rosenberg, 1987), which stated, that no such explicit rules needed to act like a rule-based system. The emerging conclusion in the 1990s became Elman's Elman, 1989 statement, suggesting that not the existence of rules and structure matters, but the way rule-governed effects are produced. Then, all generally accepted theories and even models worked with the concept of distinct linguistic structure and content. That is, what they attempted to change, by their alternative model, respecting the frame constraints without explicitly set rules. The model's architecture was also SRN (Figure 3.7). Lexical state units were either grapheme encodings or random vectors (correlated and uncorrelated respectively), Internal (phonological) state units represented articulatory features. Hidden state units were obtained as mapping of logistic activation function mapping the linear combination of all input units (lexical + phonological). The acquired hidden unit activation then propagated through

one more weight matrix, to produce distributed phoneme representation in the output layer. Training meant updating the weight matrices with standard backpropagation algorithm with momentum factor.

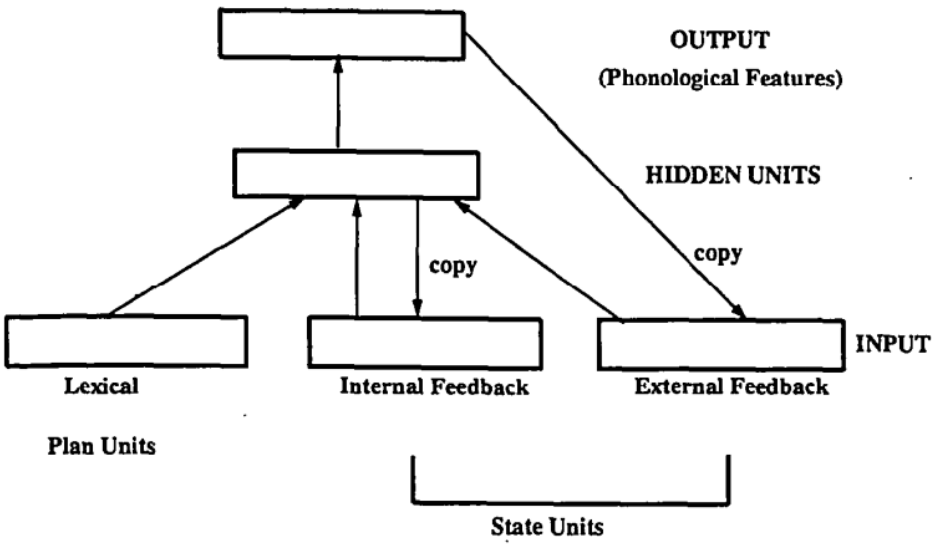

FIGURE 3.7: SRN architecture of the phonological learner model by Dell, Juliano, and Govindjee, 1993

Dell et al.'s model is a model of phonological development; It is learning phoneme sequences based on meanings, and the features of phonemes making it up. In line with PDP paradigm, meanings and phonemes were encoded in distributed vectors or encoded by their features - in other words. His above-cited paper actually contained three studies, let's call them three experiments, where he tested his network model under different circumstances:

- **Factorial manipulation with model** characteristics
  Testing the models' performance with a set of changing parameters. Namely: training vocabulary, input representations (correlated or not) and the nature of state representations (internal only, internal + external, external only)

- **Error generation by degradation**
  In the second part, trained networks, able to produce all words from training set properly, are forced to produce speech errors. Thus, speech errors mostly come from adults, phonologically developed individuals. Therefore, a noise was added to the weights of the network, during the production task.

- **Larger vocabulary**
  Three attempts made in the third part, with three different upgrades to handle enlarged vocabularies (308 and 412 words) including 2-syllable words too. All in all, the performance of their network became less accurate. However, the model reached 99.1% accuracy, all the error-patterns became vaguer, violating frame constraints to a greater extent.

Their results represent substantial evidence of speech errors by an artificial learner, respecting the frame constraints in a significant manner. Dell, Juliano, and Govindjee, 1993 proved the network's ability to act as a rule-based system, with a satisfactory explanation of neural learning. Their frequently spelled phenomenon „well-worn paths" turned out to be similar, as not the same as the conspiracy effect.

3.3 Corpus-based methods for deriving distributed meaning representations

From a connectionist point of view, questions about how semantic units, categories or even entities are represented and stored in our brain are clear. The activations propagated by our simple but massively interconnected networks must serve a set of patterns uniquely (but within a category similarly) defining semantic entities. In other words, concepts in our semantic memory, or lexicon,

are represented by activation patterns, so that a finite set of features can determine entities in the lexicon. Leastwise, in that way, presents the "Semantic Cognition: A Parallel Distributed Processing Approach" book of Rogers and McClelland, 2004. Moreover, they argue, such representations are distributed, and the patterns are served by neural activations governed by the weights between the units, the neurons. This perspective proposes that such weights are acquired by day to day adjustment when processing semantic information.

Nowadays, great support comes from neuroimaging evidence. Even the openings of such studies assume that there is a consensus of memory-contents dependency on the activity of neural ensembles across cortical and subcortical regions Rissman and Wagner, 2012. Hence, the neuro-imaging approach is pivoted from concentrating on peak regional effect, in other words from the localistic approach. The game-changing study by Haxby et al., 2001 hypothesized that however, a given brain region (e.g., ventral temporal cortex, VTC) respond to specific treatments in preference, (e.g. to houses, faces, chairs), the extent to which such region is active (measured by the magnitude of their Blood-oxygen-level dependent response) carries information, about how similar the triggering stimulus is to their preferred ones. Nevertheless, each semantical category is identifiable by its "neural signature," reflecting the mean feature weights for known exemplars from the category. Rissman and Wagner, 2012.For differences in localistic and distributed features see Figure 3.8

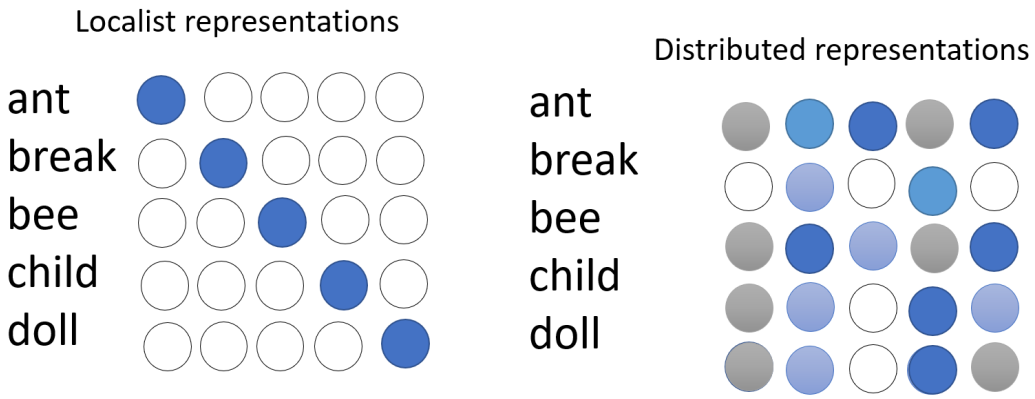

FIGURE 3.8: Comparing features of localistic and distributed vector representations. Each column represent one feature

So, what could a single neuron perceive in above specified massively interconnected distributed network? In line with Rissman and Wagner, 2012, we hypothesize, that every synapse delivers information- or more generally signal - from another neuron or a sub-area in the brain. So, against the localistic approach, where a single neuron is considered to be active under a specific treatment (lexeme), we set our meaning representations to be a collection of its features. Such supposition is present in the field of computational linguistics, and it is called Distributional hypothesis introduced by McDonald and Ramscar, 2001.

The Distributional hypothesis says, that *"meaning of a word can be approximated, or derived from the set of contexts in which it occurs in texts"*. Also, that assumption validated a lack of corpus-based methods for generating valid distributional lexeme representations instead of coding them manually or semi-manually as Rogers and McClelland, 2004 did.

The context in which a given word, or it's lexeme occurs, is defined as n neighboring words in natural text, or the whole corpus. So that, based on neighboring or dependent words we are able to quantify co-occurrences of

	planet	night	full	shadow	shine	crescent
moon	10	19	55	16	35	12
sun	11	5	1	20	53	0
dog	0	1	2	5	0	0

TABLE 3.1: Co-occurrence matrix of related terms from corpus in Figure 3.9

word pairs, and a co-occurrence vector is obtained (Table 3.1). E.g. with a such small corpus (in Figure 3.9) is feasible to find out, which word pairs share more semantic features:

```
he curtains open and the moon shining in on the barely
ars and the cold , close moon " . And neither of the w
rough the night with the moon shining so brightly , it
made in the light of the moon . It all boils down , wr
 surely under a crescent moon , thrilled by ice-white
sun , the seasons of the moon ? Home , alone , Jay pla
m is dazzling snow , the moon has risen full and cold
un and the temple of the moon , driving out of the hug
 in the dark and now the moon rises , full and amber a
bird on the shape of the moon over the trees in front
 But I could n't see the moon or the stars , only the
rning , with a sliver of moon hanging among the stars
 they love the sun , the moon and the stars . None of
the light of an enormous moon . The plash of flowing w
man 's first step on the moon ; various exhibits , aer
 the inevitable piece of moon rock . Housing The Airsh
oud obscured part of the moon . The Allied guns behind
```

FIGURE 3.9: Sample of corpus for occurrences of word “moon"

Calculating the co-occurrences, we get:

Represented in 2-dimensional (co-occurrence with shadow and shine) context:

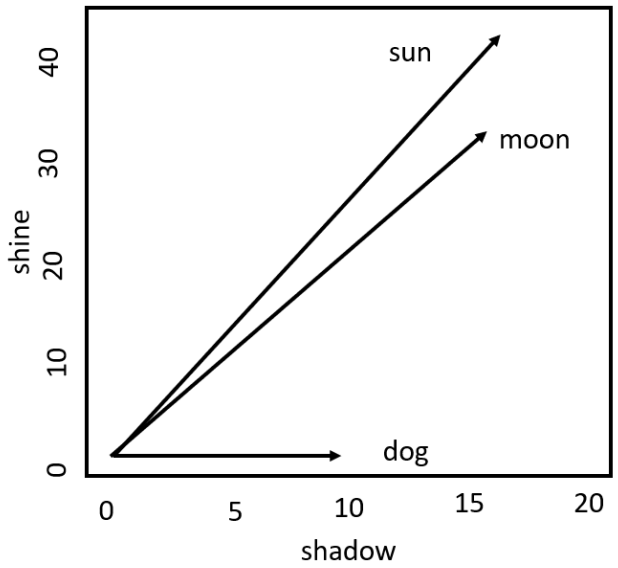

FIGURE 3.10: 2-dimensional (co-occurrence of target word with "shadow" and "shine" words as context) representation of target words: sun, moon, dog.

In Figure 3.10 we can see, that the co-occurrence vectors for semantically related terms moon and sun are similar and dissimilar from a semantically unrelated word: dog.

Above described techniques gave the conceptual framework for the two popular approaches of the 2010s for learning word vectors: global matrix factorization, derived from Latent Semantic Analysis (LSA) Deerwester et al., 1990 (e.g., Levy and Goldberg, 2014b; Levy and Goldberg, 2014a) and local context window methods such as the skip-gram model of Mikolov, Yih, and Zweig, 2013). Skip-gram models train on local co-occurrences in context windows, LSA-methods build upon global co-occurrences. Therefore skip-gram models do better on analogy tasks, LSA-methods on comprehensive statistic

information. Their competition did not take too long, since in 2014 Pennington, Socher and Manning released their GloVe (Global Vectors for Word Representation) model Pennington, Socher, and Manning, 2014, integrating the two models, becoming the etalon, as the state-of-the-art distributed word representation not just for analogy tasks[1]. Proved competency in a wide range of semantical and even syntactic tasks made it appropriate for our purpose, producing systematically different representations of lexemes in line with the PDP paradigm. Therefore, we aim to represent even our meanings in such a way; they would serve their key features, which could be evaluated for all the meanings. For differences in localistic and distributed features see Figure 3.8

GloVe vectors are trained on a dataset that contains:

- 2010 Wikipedia dump with 1 billion tokens
- 2014 Wikipedia dump with 1.6 billion tokens
- Gigaword 5 which has 4.3 billion tokens
- 42 billion tokens of web data, from Common Crawl5

Then, the GloVe vectors' vocabulary has been reduced to the 400,000 most frequent words from the list above, before the matrix of co-occurrences calculated.

[1] https://aclweb.org/aclwiki/Analogy_(State_of_the_art)

Chapter 4

Experiment

4.1 Differences between source studies and our model

As indicated in the previous chapter, the architecture of our model of language acquisition is a Simple Recurrent Network (See Figure 4.2), in many aspects different from models of source studies. We intended to enclose the PDP paradigm with semantic processing. In accordance with (Takac, Knott, and Stokes, 2017) and (Dell, Juliano, and Govindjee, 1993), we assume, the presence of the meaning signal is crucial of learnability; Therefore, appropriate attention must devote to the simulation of such signal. In Section 4.3 we clarified our choice of GloVe vectors, which meant an increase of needed learning capability against all the source studies. On the one hand their dimensionality (300) did not exceed the vector lengths Takac's model significantly, but on the other hand, the density of word vectors caused an increment in computing power needed (in context of non-zero multiplications). Also, even the weight update became more complicated as well since every gradient must have been backpropagated to every non-zero input neuron. Dell chose another two - not surprisingly – computationally also a less complex meaning representation of the 30-dimensional vector. Although, the main difference does not root in in

dimensionality. They assumed a vector of random numbers would be in accordance with the PDP paradigm. We would argue with it; Since the similarity (what they representations lack) between meaning vectors gives the opportunity to produce semantic speech errors, to be evaluated and analyzed in Chapter 5. Nevertheless, all of our arguments hold for biological plausibility from Section 1.2 The other way of creating meaning vectors, the „correlated way" stemmed in an incomprehensible way correlated with the orthography of the word, not the lexemes. This branch to be ignored in our project.

The representations of phonemes differed just from the Takac's model, in the context of their localistic approach. The phoneme-features from Section 4.2, actually, matched the ones in (Dell, Juliano, and Govindjee, 1993). The training vocabulary of present work is taken over from (Takac, Knott, and Stokes, 2017): the MacArthur-Bates Communicative Development Inventory (CDI) of English words by (Fenson et al., 1994) for the "known words" the CELEX corpus (Baayen, Piepenbrock, and Van Rijn, 1995) for the complementary data, the "not-known words." Dell's two 50-word, 3-segment samples meant just another facilitation for their work, against ours. Though, fixed-size consonant-vowel combinations make it somewhat easier to analyze speech errors produced by the model. Moreover, since we are to reproduce children data, with all their learning aspect, we will consider speech errors produced during the vocabulary development, unlike Dell's second experiment. By the way, neither his third experiment will be followed for similar reasons. To be comparable with children data, appropriate treatments must be reproduced including the constant size and consistency of training data. When it comes to network architectures, in training session, the internal-only feedback loop was selected by us. For generating task, (between two epochs), the network's output was copied to the input with a time delay (external recurrence by (Dell, Juliano, and Govindjee, 1993)) In respect of activation functions, unipolar logistic functions were used, in both studies, Only the activation of the hidden layer was set

to bipolar hyperbolic tangent *tanh* in our experiment, as it improved the networks ability to distinguish even negatively among samples from input layer. For that matter, (Takac, Knott, and Stokes, 2017) used linear output neurons and softmax function to match one-hot encodings of phonemes at the output. Our output layer's activation remained sigmoidal, according to the distribution of phoneme vectors. Due to the above particularized computational disparity between source studies and our project (input vector-dimensionality and -density, vocabulary size, word-lengths) about 10x more hidden neurons and 6x longer training time (epochs) were needed to reach human-like performance on phonological development task.

4.2 Phoneme encoding

Most of the phonology textbooks claim – in accord with the International Phonetic Alphabet ((Association, 2005)) that the phonological system of the English language is made up of 44 phonemes, of which 24 are consonants, and 20 are vowels, considering clusters of sounds (e.g., diphthongs) as single phonemes. We used phonemes of the International Phonetic Alphabet (IPA) to support all phonemes of MacArthur-Bates Communicative Development Inventory (CDI) of English words (Fenson et al., 1994), the base dataset of the present project. The complementary data, the "not-known" words came from the CELEX corpus (Baayen, Piepenbrock, and Van Rijn, 1995). Homophones, homographs, and abbreviations excluded in accord with (Takac, Knott, and Stokes, 2017). After Rummelhart and McClelland's (Rumelhart et al., 1986) PDP model of the acquisition of English past tense, many PDP-based vector representations have been proposed to increase generality of their phoneme representation, what turned out to be the weak point of their work. Most of these works came up with word-patterns of fixed-size words. E.g., (Miikkulainen, 1997) used five units on a continuous scale for each phoneme in an arbitrary word

(previous ones considered only monosyllables), classifying the phoneme's features like voicing, sonority, place, and manner of articulation. Their model has also been criticized for presenting the similarity only in phoneme level; So that two words encoded in their way, are dissimilar from many quantitative perspectives. However, representation of phonemes in time could resolve the trouble. Meaning, that one phoneme per timestep to be analyzed could resolve this vectorial dissimilarity issue in our model. In line with (Dell, Juliano, and Govindjee, 1993) and (Miikkulainen, 1997) we created a semi-binary feature vectors for our phonemes; based on above mentioned Distinctive articulatory features on unipolar scale [0,1]. One when one of the following 18 features is true for the maximal extent: Syllabic, consonantal, sonorant, voiced, continuant, nasal, strident, lateral, distributed, affricate, labial, coronal, anterior, high, back, low, round, tense. These specific 18-dimensional vectors are hard to imagine and even check for validity. Therefore, we calculated their Distance matrix, containing the Euclidian distance between each vector with each other, to produce a tree diagram of it. The particular kind of tree diagram in Figure 4.1 is a so-called dendrogram. It is a product of hierarchical clustering of our phoneme vectors so that we can make sure, that phonologically similar phonemes are represented similarly, i.e. belong to the same cluster. Moreover, the number of clusters and their cardinality indicate how unique our representations are, or how many similar phonemes do we have and to what extent are they similar.

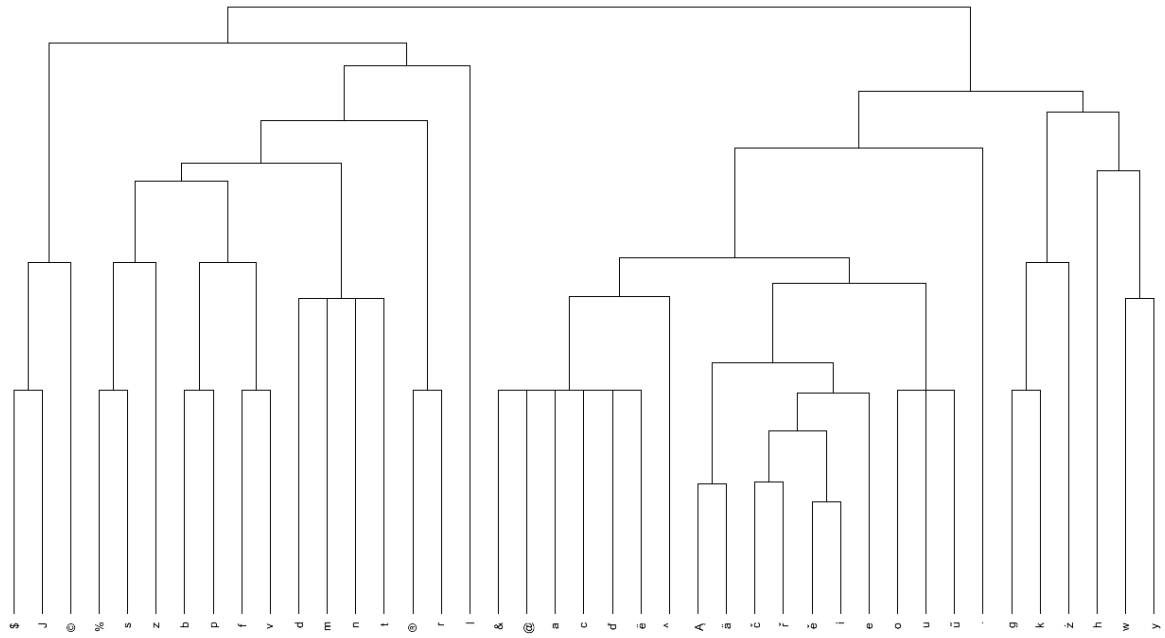

FIGURE 4.1: Result of hierarchical clustering of phonemes by a dendrogram.

4.3 Meaning encoding

As discussed in Section 3.3, GloVe vectors proved competency in a wide range of semantic tasks. Therefore appropriate for our purpose, producing systematically unique representations of lexemes in line with the PDP paradigm. Therefore, our aim is to represent even our meanings in such a way; they would represent their key features, which could be evaluated for all the meanings.

However the these datasets' vocabulary is way more extensive than our dictionary (from child-data), we use the vectors build upon them (upon their co-occurrences) to ensure distinct representations for all the words with appropriate semantic relations between them. These pre-trained GloVe vectors were obtained fromPennington, Socher, and Manning, 2014, and then reduced to the baseline vocabulary - the CDI data inventory by Fenson et al., 1994

4.4 Training regime

Our reference dataset was prepared by (De Cara and Goswami, 2002) from CELEX corpus consisting about 4000 English monosyllables. For our purpose, we reduced it to "more frequent" words, which has more than numerical zero frequency. Moreover, frequencies for CDI words, the "known words" were also assigned from the CELEX corpus in line with (Takac, Knott, and Stokes, 2017). That was the crucial moment for ensuring realistic treatment in the account of probabilities to hear or learn a specific word.

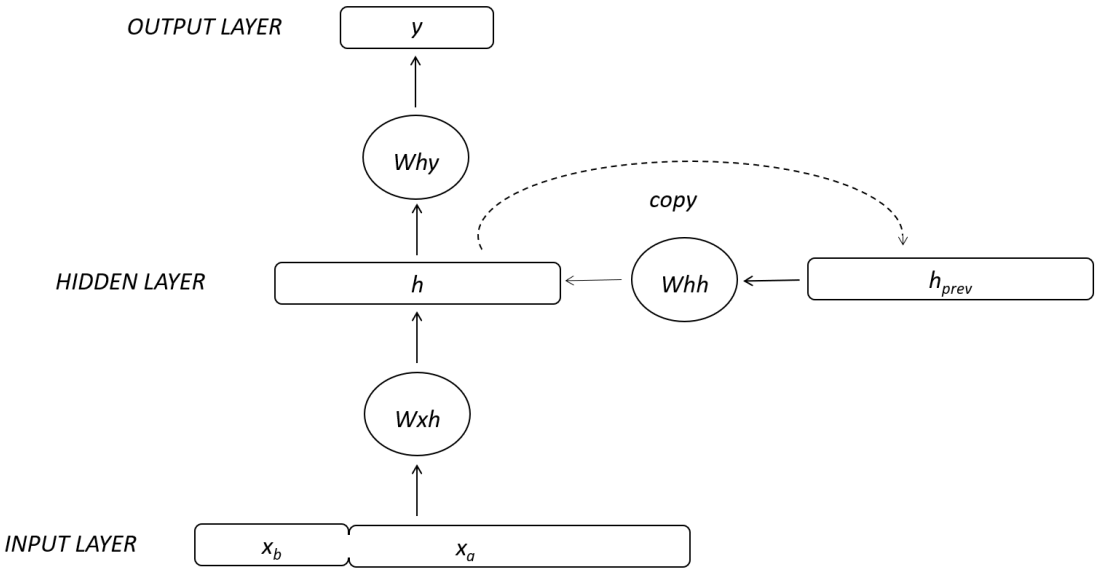

FIGURE 4.2: Schematic architecture of our model, a Simple Recurrent Network.

In every training epoch, the network was exposed to a shuffled batch of training samples reflecting their probability to occur in the CELEX dataset. Meaning, that 268 word-vectors were sampled in every training batch, commonplace words, even multiple times included.

During the training, 18 models were created, three instances for a given number of hidden neurons, 100, 110, 120, 130, 140, 150. Weight matrixes Wxh, Whh, Why initialized as random matrix with uniform distribution over $[-0.05, 0.05]$.

Context vector h_{t-1} was initially set to zeros. In line with time resolution adumbrated in Chapter 3.1.1, one phoneme vector presented in every timestep. For every training sample (phoneme-sequence corresponding to a specific lexeme) an additional input vector created for word boundaries, denoted by /./. So that, the network must predict the next phoneme after the word boundary /./ given, that meaning is x_a. Next, if the first phoneme were predicted to be for instance /k/, the network must predict concerning unchanged meaning x_a and the input phoneme (/k/)-s distributed representation x_b what's the most likely next phoneme. In the case of x_a is the word vector for word *cat*, it should be phoneme /a/ and so on; Until the last phoneme /t/ is predicted. Then the word boundary, /./ to be predicted. In that way, after the forward pass of training sample of length k, for every time step in the range $(0, k)$ error is computed, and recursively accumulated to a global error, E, which is back-propagated through time.

After weight update executed for all lexemes (words) in the batch, test session follows. Within that, a word boundary is given for all meaning vectors, (initial phoneme vector setting) and the network predicts phoneme representations in the output layer y, which is fed back as input meaning x_b until word boundary predicted, or the length of predicted phoneme sequences exceeds number seven. During this session, all the produced phoneme sequences registered and evaluated for further analysis.

Here we want to highlight, that we intended to make every aspect of training as realistic as possible in respect of learning children. The composition of input batches, containing words with, and without meaning vector should reflect the fact, that also children are exposed to phonotactic patterns (words they do not know) without meaning, during their phonological development; influencing their learning process (in accord with (Takac, Knott, and Stokes, 2017). Moreover, the probability distribution of the input batches is also matching the frequencies of words in an average child's environment.

Chapter 5

Results

Eighteen models were created, three instances for every hidden unit amount - for statistical compliance. Hidden layer sizes (100, 110, 120, 130, 140 and 150) were chosen surrounding the model with the best performance (with 120 hidden units, based on preliminary testing). The mean performances of the above-listed parameter settings (in Figure 5.1) show a moderate convergence. After 300 epochs, the mean error got under 10% (7.73%) among all instances in accord with Dell's results.

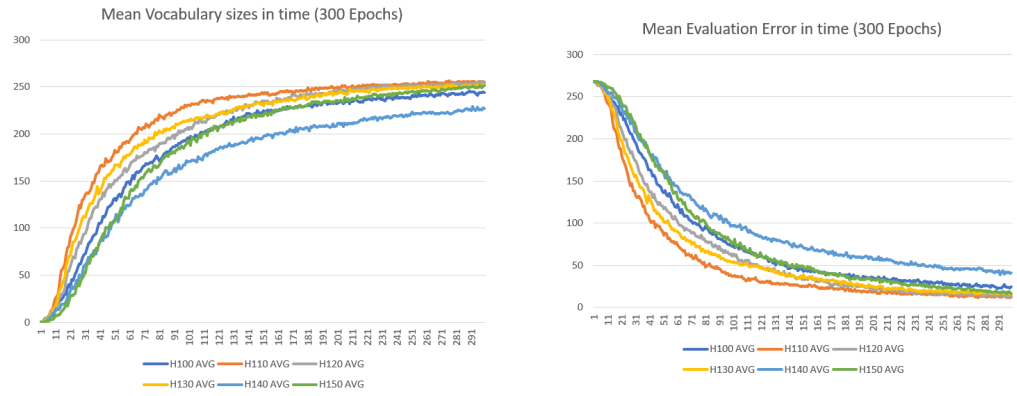

FIGURE 5.1: model performance in Error decrease (left), and vocabulary enlargement (right).

However, we chose network parameters producing the best performance; we did not let the network to catch up to its minimal error. Thus, our aim is

not to find the best approximator of mappings between lexeme and phoneme transitions. That could be done quickly in another way. We needed a massive error-evidence to compare with pieces of evidence of empirical research from many aspects. Therefore, contrary to source studies, testing was launched after every training epoch. Thanks to this setup, 315 159 errors were produced and analyzed, excluding the extremely declinatory predictions (a single word-boundary, single phoneme predicted many times, etc.).

5.1 Phonological errors

Instance	Consonant - Vowel Category Effect	Syllabic Constituent Effect - VC - Rule	Syllabic Constituent Effect - CV Rule	Initialness Effect	Strict Initialness Effect
H100 AVG	100,00%	20,38%	1,46%	76,84%	13,74%
H110 AVG	100,00%	16,36%	1,72%	81,33%	20,89%
H120 AVG	100,00%	21,17%	1,54%	75,94%	16,71%
H130 AVG	100,00%	22,08%	1,62%	76,46%	17,68%
H140 AVG	100,00%	24,42%	1,93%	79,13%	29,04%
H150 AVG	100,00%	24,79%	2,90%	77,45%	22,73%
Overall	100,00%	21,53%	1,86%	77,86%	20,13%
Benchmark (Empirical studies (Dell el al. 1993)	99%	6-20%	2%	62%	40-50%

TABLE 5.1: The Percentual extent of speech errors respecting the Frame constraints (Consonant-Vowel Category, Syllabic Constituency, and Initialness Effect.) 100% means, that all of the speech errors of a given model (in rows) respected the constraint (in columns)

Across 315 159 errors of 18 model trained for 300 Epochs, there was no error violating the Consonant-Vowel Category Effect; In other words, none of the Consonants were slipped with a Vowel and vice versa. There is essentially no difference compared to 99% in (Stemberger, 1983b) collection of natural errors. Every model's behavior reflects the human phenomenon. Syllabic Constituent Effect examines speech errors, where more than one, neighboring phonemes are replaced. The Effect of CV-rule, predicting much more flips in order Vowel-Consonant than Consonant-Vowel is out of the question. The proportion of CV-flips exactly matches 2%, the average among shreds of evidence (Dell, Juliano, and Govindjee, 1993). However, The VC-slips oversteps even the highest measured rate in (Stemberger, 1983b) corpus; we consider it to be acceptable; Especially considering the ratio of CV- and VC-slips. Since the training batches and even the testing setup was on a one-word basis, shift movements (errors of two word-initial Consonants) were not possible. Meaning, that a particular case of Initialness Effect has not been generated at all. It represents a remarkable bias in every aspect of error processing, mainly in Strict Initialness Effect, where nothing but the onset consonant should be slipped.

target	error
drĄ	brĄ
fĄn	vĄn
fĄn	rĄn
fĄn	lĄn
fĄn	sĄn
fĄn	%Ąn

TABLE 5.2: Some errors under the Initialness Effect, flips of onset consonants

(Shattuck-Hufnagel, 1987)'s first phoneme effect was based on, that 66% of their 1520 single-phoneme speech errors from the MIT database involved

word onsets. This is twice as frequent as 33% which is the frequency of word-onset consonants in connected speech (Shattuck-Hufnagel, 1987). That implies that onset phonemes have a special role in word production. Moreover, many frame theorists suggest, they must be processed and represented separately. Although it is hard to measure the legality of a sequence, based on the above-presented results we predict Phonotactic Regularity asymptotically approaching the 100%. Instead of defining what is considered to be a legal sequence, we found just several counter-examples among the studies cited above. These counterexamples mainly focus on appearing a consonant in place of the vowel. Due to the absence of CV Category Effect violation and the significant (.92) correlation between Phonotactic Regularity Effect and CV Category Effect, we assume, the error-patterns are consistent with Phonotactic Regularity Effect. However no exact proportion available, we claim, our model, and it is every instance produces phonotactically legal errors.

5.2 ND Effect

Our ultimate aim was to explore whether the ND Effect of our model befits the patterns shown by children and other models of phonological development. Our first hypothesis held high-ND words (with many phonological neighbors) are more likely to be learned earlier. The high-ND has been operationalized to Average ND in the vocabulary of individuals, the time as the epoch. Thus, we expected the highest-ND words to be learned within the first few epochs, producing high-ND in average in vocabularies. As training proceeds, lower-ND words were learned, smoothly decreasing the mean ND of the vocabulary. Such a smooth course with a fast decrease of average ND we can see in Figure 5.2

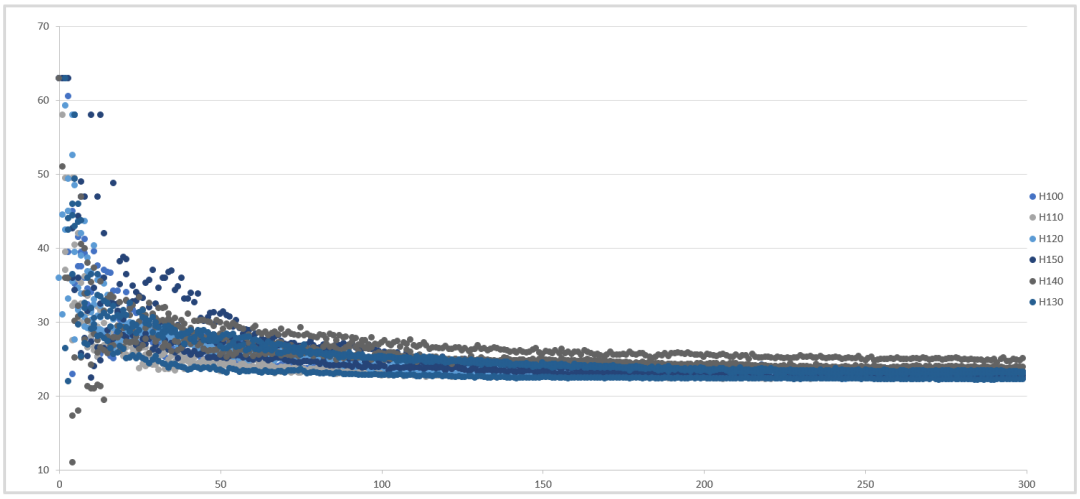

FIGURE 5.2: Average Neighborhood Density in time (epochs) for all the words within vocabulary of the entire population of SRNs in time

The ND-effect in time is significant for the undivided population of artificial learners, such as for individual groups with a given amount of learning capacity (varied sizes of the hidden layer). Notwithstanding, several outlying points found, mainly produced by an instance with 140 hidden neurons (H140). From the training logs (A) we can look up, what was behind (The learning course is presented in Table 5.3). The second instance of our SRNs, with 140 hidden neurons (H140_2), learned a lower-ND word first, the word: FINE / fĄn/ with ND of 36. However, by lower-ND, we do not mean low in general. Since the average ND is 19,7 in not-known (words without added meaning vector during the training) and 22,0 in the dataset of known words (CDI words / with meaning vector) (for detail see Figure 5.3).

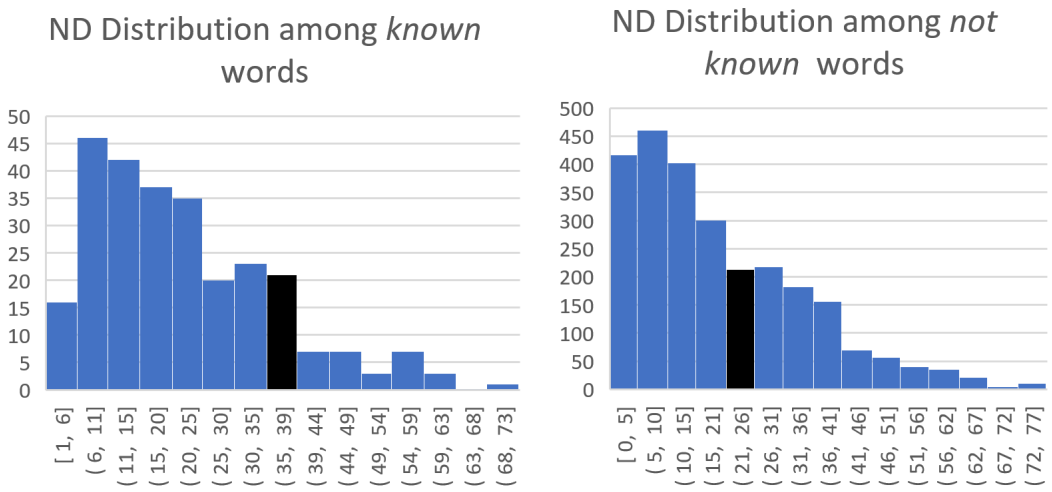

FIGURE 5.3: Histograms of ND Distributions in known, and not-known datasets. The black bin denotes the interval, the word **FINE** belongs to

The ND of **36** is much lower compared to the average ND at the beginning of the training (learning process) whereas the mean ND of the first two epochs among the SRN population is **46**.

Such a relatively low-ND word is quite likely to have unusual structure, resulting an infrequent path of the network becoming „well-worn“ (Dell, Juliano, and Govindjee, 1993; Takac, Knott, and Stokes, 2017). These updated, „well-worn“ path promote potential to another word with the same unusual structure to be the next word learned. In case of H140_2 these were **BROWN** */br&n/* with ND of **10**, **FLY** */flA/* with phonological neighbors in count of **18** and **MOON** */mu:n/* with ND **24** in epoch 4. Their common rare feature is probably the long syllable in the rhyme-constituent represented by diphthongs (**ai, *oi*, au, uu**) and long */u:/*.

Table 5.3 shows, that even if a word has been learned in epoch 3, there is no guarantee, that the word will be correctly pronounced within the production task in next epoch. However the misplaced phonemes are close to each other, our algorithm marked it as a speech error. Although, the incorrect phoneme *ř*

share the feature being learned with preference: Long-syllable and diphthong-like.

epoch 3	fine				
epoch 4	fine	moon	fly	brown	
epoch 5	fine	moon	fly	brown	buy

TABLE 5.3: Known words in epochs (3,4,5). Red color marks the words with not correct pronunciation. In case of fine it was *fřn* instead of *fĄn* (oi-ai)

However, the mean ND within epoch 4 is near to the median of the ND within the source dataset, against, the mean of the all 4th epochs it accumulated approximately a Δ **30 ND** (**17** is the mean of above presented second epoch vs. **46**). In our view, the above-presented violation of ND-effect is in line with the PDP paradigm, especially with the Conspiracy Effect. Moreover, it is strengthening the findings of our source studies. According to (Takac, Knott, and Stokes, 2017), the shared similarities between low-ND words (produced by H140_2 in epoch 4) is represented with similar activation patterns in the network; Therefore, weight update for one of them necessarily overlaps with the weight change potentially needed for the other, phonologically neighboring word. These 'slightly outlying' points we consider being caused by strange weight initialization.

Nevertheless, we must add that the source studies excluded the participants (either artificial and real) with vocabularies with less than 20 words. So, if we would have taken in account the fact, that all of our distant points belong to learners with small (even smaller than 20 words) vocabulary, we could easily exclude them from the dataset, referring to the Heteroscedasticity as our source study did (Takac, Knott, and Stokes, 2017). But not doing so helped us to explore the dynamics of Conspiracy effect.

5.3 Age of Acquisition & Vocabulary Size

Since we took over the training datasets, and many aspects of the training regime, we are able to compare our results with (Takac, Knott, and Stokes, 2017) and even with child language data directly. Our results reflect consistency with the above-cited SRN model in the fields of ND dependency on lexicon size (Figure 5.4)

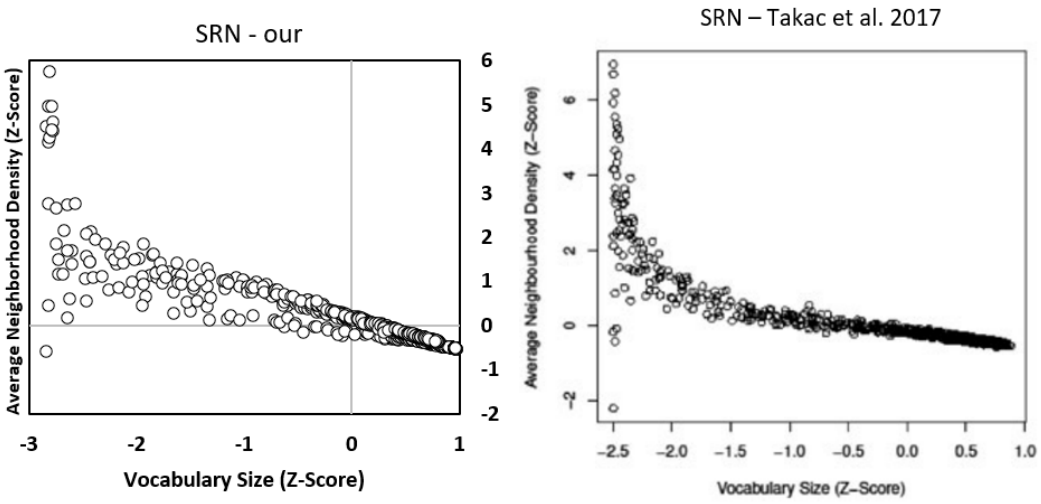

FIGURE 5.4: Scatterplots of ND against Vocabulary Size. Comparing the results of our model (left) to (Takac, Knott, and Stokes, 2017) (right). NDs are calculated within the (Baayen, Piepenbrock, and Van Rijn, 1995, CELEX) dataset

Moreover, conversion of measures to Z-scores (in both Figure 5.4 and 5.5) lets us explore much more analogies even with child data from studies of English and Danish children from (Stokes, 2010; Stokes et al., 2012; Stokes, 2014) in Figure 5.5.

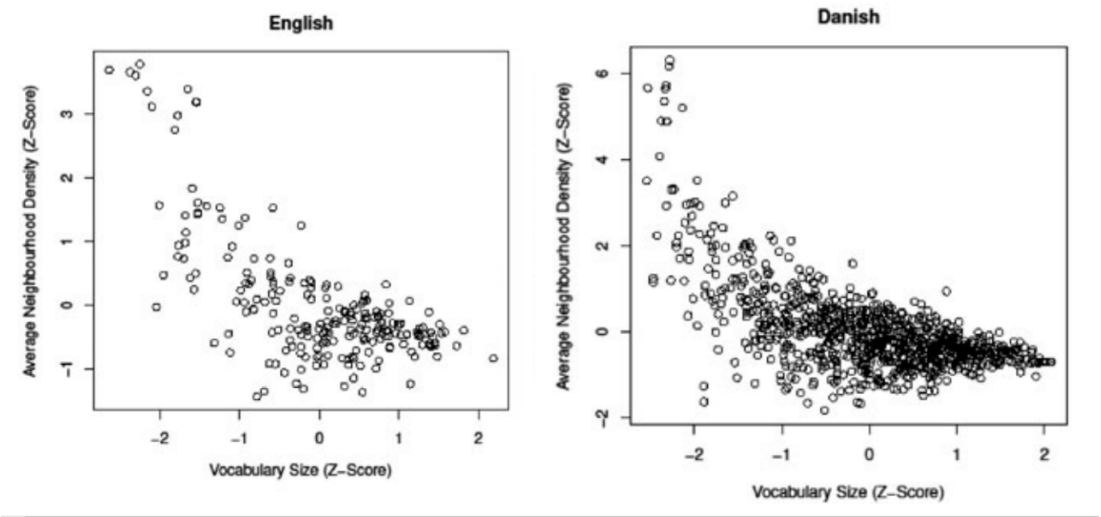

FIGURE 5.5: Scatterplots of average ND vs. Vocabulary Size in English (left) and Danish (right) children. ND-s calculated within the (Baayen, Piepenbrock, and Van Rijn, 1995, CELEX) dataset

Per the above-referred works of Stokes and Takac, we can state, the trend is uniform for children and simulation data, showing a decrease in average ND with growing vocabulary. Nevertheless, another clear trend is shown by both, children and SRN data: the decreasing variance in average ND with increasing Vocabulary Size. The explanation of different variances among children, and SRN data - due to (Takac, Knott, and Stokes, 2017) - might stem in the different memory capacity ranges of neural models and children. While children data came from as many individuals as many samples, the simulation data track just four instances of artificial learners with given memory capacity (number of hidden neurons) in different stages of their development.

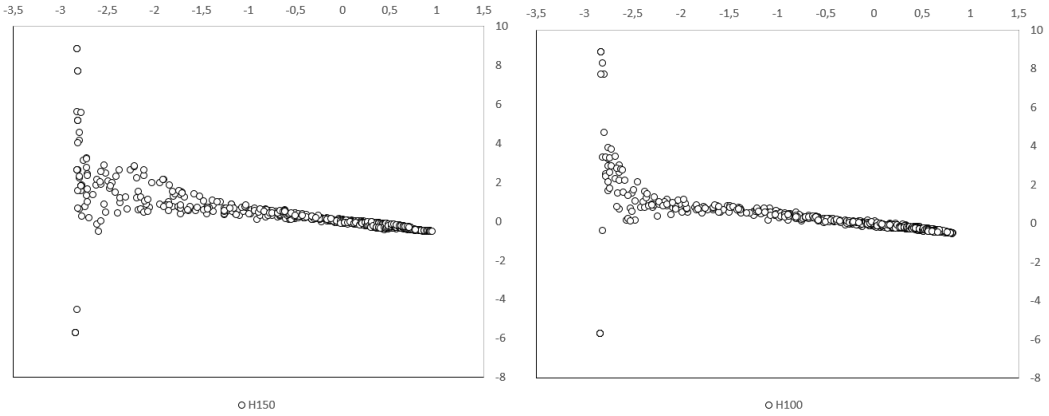

FIGURE 5.6: Comparing results of two artificial SRN participants, with different memory capacity. 100 hidden neurons on the left, 150 on the right hand-side

Both of the scatter plots are produced by a single instance of SRN, with different amounts of hidden neurons. They show reduced variance compared to Figure 5.4, where the whole SRN population is presented. Hence, the network with a higher number of hidden neurons (Figure 5.6, right), results increased the capacity of memory, shows greater variance than the lower-capacity network with 100 hidden units. All in all, Takac's statement seems to be supported by our data. Nevertheless, our experimental setting caused the inability to reproduce late-talker phenomenon. Since even the network with the fewest amount of hidden neurons (H100) was able to learn lower-ND words, to investigate this ability, instances with less hidden neurons must be generated and tested in future work.

Chapter 6

Discussion

The goal of the present thesis was to compare two perspectives on the phonological part of speech production system through simulation of its development. In the first chapter we start with investigating the neuro-biological background of speech, from the classical Broca–Wernicke–Lichtheim- Geschwind" model to the neuroimaging studies, which were to indicate speech activation patterns in the brain are distributed, running in parallel. Next, a current linguistic approach is presented in Chapter 2, focusing on speech errors, to be reproduced in our experiments, such as the Neighborhood-density and related phenomena in Section 2.3. Here we discussed reproducible aspects of children's phonological development. The next part summarizes the methods to use in the experiment: Recurrent Neural Networks, parameter optimization techniques, and the two specific applications of such networks for similar tasks to ours. In Chapter 4 the experimental setting is presented. First, the differences between our work and the source studies. Thereafter our distributed phoneme- and meaning-representations exhibited and discussed in context. In Chapter 5 we present our results, which's novelty we discuss in Chapter 6.

First of all, our model reproduced the ND-effect, in every aspect; however, it should not have happened, according to the source study by Takac, Knott, and Stokes, 2017. They explained the Neighborhood Density phenomena in an even mathematically very convenient way; Suggesting, that the localistic representation of word-meanings might be the cause or at least the support for

such behavior. By all means, their theory should be reviewed, and appropriate interpretation is needed for contradictions between their theory and our results.

In our view, Conspiracy effect and our analysis of the outlying points in Figure 5.2 offer an at least logically passable explanation since in our distributed representations, phonological and even semantic similarity can be found, making possible to explain the semantic ND effect. Furthermore, our distributed representations, are biologically far more plausible, than the source studies' by either Dell, Juliano, and Govindjee, 1993 and Takac, Knott, and Stokes, 2017. As described in Section 1.2, almost every neuro-imaging study of speech production found parallel activation patterns distributed along the brain. If we consider the localistic ones, it will imply, that an infant exactly knows all the English phonemes even before learning the first word.

Our work's originality also stems in speech error analysis of errors produced by artificial children's phonological development. However, a quantitative analysis made only in Chapter 5, the nature of phonological errors produced by the model, could help the speech therapists unfold ponderous phonological patterns, to focus more on, or to avoid during speech therapy of late or even regular talkers.

Beyond the successful parts of the project, words must also fall about the limitations. The first would be the inability to find semantic errors produced by the model. Our searching algorithm (semantic_test.py in Appendix A) searched for mismatch of a known word in the vocabulary. Meaning, that for a positive hit two meaningful phoneme sequences should be flipped; And that did not happen not even once. In the future, a semantic analysis should be improved in several ways. First, we do not know whether semantic errors are explicit or implicit. If a child knows the word cat and pronounces mat instead, what he does not know, will it be considered as a semantic error? What if the child really does not know the meaning of it, just heard several times from

parents that „cat is on the mat"? Anyhow, our analysis should consider even other-than-known words. Another limit represented again, the discretization of the phoneme space in the output layer. Thus, when a network's output vector is computed, it is just a point in 44-dimensional continuous space. So it is essentially impossible to predict exactly the same coordinates, we store in our look-up table of the phoneme encodings. Therefore, the nearest neighbor is chosen from the database. So that we get a discrete phoneme, as a prediction. Although, predicting vector in-between two others, is an ambiguity, discernible in human speech errors. E.g., when we are mixing up two phonemes. Moreover, some of the English phonemes are as close to each other, so that a non-native English speaker would hardly distinguish between them (Wilson and Iacoboni, 2006). So that our decision criterion should have been fuzzier, working with more potential output phonemes, until matching them against the vocabulary, and choose the meaningful one. Furthermore, this proposal is totally in line with the model of Levelt, 1999; Where "self-monitoring" loops make the same comparison and matching against the conceptual representations. Another less convenient way to handle such ambiguous inter-phoneme vector representation in the output layer would be to miss out the phonemes' discretization and let another additional layer of the neural network produce voice sequence instead. Like neural text-to-speech systems do nowadays. Although, if we think about it, the situation would be the same as above; excepting that the priming effect would happen in the brain of the person listening to the ambiguous voice. Again, as people do, e.g., by evidence of Rodd et al., 2013.

The second thing in question is the outlying "not-so-high-ND" words, learned at the beginning of the learning period. As discussed in 5.2, these outlying words were caused by initializing the network's weights randomly. Such that, when the initial weights for non-frequent features are a bit stronger, it is likely to learn a word with non-frequent phonological features, so that with just a

few phonological neighbors. But how about children? How are they learning the first words? What are the limitations of speech organs? According to Jakobson, Fant, and Halle, 1951 we know, that the first words are likely to be semantically related to mother and father of the child. Moreover, phonologically must be as simple as possible - voices with an opened mouth (a) and closed mouth (m, p, b). These phonemes form the optimal consonant-vowel pairs (subject to the energy needed for their production). With sufficient involvement of these constraints to a future model, the learning-course would better fit the children-data.

Nevertheless, we must be aware, that we are modeling a small, low-level subsystem of a robust system producing meaningful sentences, with syntax, semantic relations, and sense along the sentences so that our model must lack an enormous amount of functionality to consider it as a speech production model.

The results of our experiment suggest that changing the representation of phonological and semantic features does not affect the phonological Neighborhood Density phenomenon. Our results - in accord with results of Takac, Knott, and Stokes, 2017 - are representing a decreasing trend of average pND in time during the learning course of artificial learners; Such as children's. It implies that phonological Neighborhood density is not necessarily related to localness of feature representations. Instead, we suggest, that phonological Neighborhood density, and probably even the semantic Neighborhood Density, is due to the Conspiracy effect as we discussed above. To prove our thesis, further work is going to be done, including the changes suggested in this Chapter.

Appendix A

CD contents

The attached CD contains:

- :/Training_logs - including error lists of 4 SRN instances, Calculated linguistic measurements and scripts producing the results
- :/Source_codes - of reimplementation of Takac's model, of our SRN with distributed representations, source datasets and testing scripts.

Bibliography

Association, International Phonetic et al. (2005). "International phonetic alphabet". In: *Revised to*.

Baayen, RH, R Piepenbrock, and H Van Rijn (1995). "The CELEX database". In: *Nijmegen: Center for Lexical Information, Max Planck Institute for Psycholinguistics, CD-ROM*.

De Cara, Bruno and Usha Goswami (2002). "Similarity relations among spoken words: The special status of rimes in English". In: *Behavior Research Methods, Instruments, & Computers* 34.3, pp. 416–423.

Deerwester, Scott et al. (1990). "Indexing by latent semantic analysis". In: *Journal of the American society for information science* 41.6, pp. 391–407.

Dell, Gary S (1986). "A spreading-activation theory of retrieval in sentence production." In: *Psychological review* 93.3, p. 283.

Dell, Gary S, Cornell Juliano, and Anita Govindjee (1993). "Structure and content in language production: A theory of frame constraints in phonological speech errors". In: *Cognitive Science* 17.2, pp. 149–195.

Duchi, John, Elad Hazan, and Yoram Singer (2011). "Adaptive subgradient methods for online learning and stochastic optimization". In: *Journal of Machine Learning Research* 12.Jul, pp. 2121–2159.

Elman, Jeffrey L (1989). "Structured representations and connectionist models". In: *Proceedings of the 11th Annual Conference of the Cognitive Science Society*. Erlbaum Hillsdale, NJ, pp. 17–25.

— (1990). "Finding structure in time". In: *Cognitive science* 14.2, pp. 179–211.

Elman, Jeffrey L (1991). "Distributed representations, simple recurrent networks, and grammatical structure". In: *Machine learning* 7.2-3, pp. 195–225.

Fenson, Larry et al. (1994). "Variability in early communicative development". In: *Monographs of the society for research in child development*, pp. i–185.

Fromkin, Victoria A (1971). "The non-anomalous nature of anomalous utterances". In: *Language*, pp. 27–52.

Geschwind, Norman (1970). "The organization of language and the brain". In: *Science* 170.3961, pp. 940–944.

Graves, Roger E (1997). "The Legacy of the Wernicke-Lichtheim Model". In: *Journal of the History of the Neurosciences* 6.1, pp. 3–20.

Haxby, James V et al. (2001). "Distributed and overlapping representations of faces and objects in ventral temporal cortex". In: *Science* 293.5539, pp. 2425–2430.

Jakobson, Roman, C Gunnar Fant, and Morris Halle (1951). "Preliminaries to speech analysis: The distinctive features and their correlates". In:

Levelt, Willem JM (1999). "Models of word production". In: *Trends in cognitive sciences* 3.6, pp. 223–232.

Levy, Omer and Yoav Goldberg (2014a). "Linguistic regularities in sparse and explicit word representations". In: *Proceedings of the eighteenth conference on computational natural language learning*, pp. 171–180.

— (2014b). "Neural word embedding as implicit matrix factorization". In: *Advances in neural information processing systems*, pp. 2177–2185.

Lichtheim, Ludwig (1885). "On aphasia". In: *Brain* 7, pp. 433–484.

MacKay, Donald G (1970). "Spoonerisms: The structure of errors in the serial order of speech". In: *Neuropsychologia* 8.3, pp. 323–350.

— (1972). "The structure of words and syllables: Evidence from errors in speech". In: *Cognitive Psychology* 3.2, pp. 210–227.

McDonald, Scott and Michael Ramscar (2001). "Testing the Distributioanl Hypothesis: The influence of Context on Judgements of Semantic Similarity".

In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 23. 23.

Miikkulainen, Risto (1997). "Dyslexic and category-specific impairments in a self-organizing feature map model of the lexicon". In: *Brain and language* 59, pp. 334–366.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (2013). "Linguistic regularities in continuous space word representations". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751.

Montavon, Grégoire, Genevieve B. Orr, and Klaus-Robert Müller, eds. (2012). *Neural Networks: Tricks of the Trade - Second Edition*. Vol. 7700. Lecture Notes in Computer Science. Springer. ISBN: 978-3-642-35288-1. DOI: 10.1007/978-3-642-35289-8. URL: https://doi.org/10.1007/978-3-642-35289-8.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Petruck, Miriam RL (1996). "Frame semantics". In: *Handbook of pragmatics*, pp. 1–13.

Pulvermüller, Friedemann et al. (2006). "Motor cortex maps articulatory features of speech sounds". In: *Proceedings of the National Academy of Sciences* 103.20, pp. 7865–7870.

Rissman, Jesse and Anthony D Wagner (2012). "Distributed representations in memory: insights from functional brain imaging". In: *Annual review of psychology* 63, pp. 101–128.

Rodd, Jennifer M et al. (2013). "Long-term priming of the meanings of ambiguous words". In: *Journal of Memory and Language* 68.2, pp. 180–198.

Rogers, Timothy T and James L McClelland (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

Rumelhart, David E, James L McClelland, PDP Research Group, et al. (1987). *Parallel distributed processing*. Vol. 1. MIT press Cambridge, MA.

Rumelhart, David E et al. (1986). "Sequential thought processes in PDP models". In: *Parallel distributed processing: explorations in the microstructures of cognition* 2, pp. 3–57.

Rummelhart, David E (1986). "Learning internal representations by error propagation". In: *Parallel Distributed Processing: I. Foundations*, pp. 318–362.

Sejnowski, Terrence J and Charles R Rosenberg (1987). "Parallel networks that learn to pronounce English text". In: *Complex systems* 1.1, pp. 145–168.

Shattuck-Hufnagel, Stefanie (1987). "The role of word-onset consonants in speech production planning: New evidence from speech error patterns." In:

Sirigu, Angela et al. (1998). "Distinct frontal regions for processing sentence syntax and story grammar". In: *Cortex* 34.5, pp. 771–778.

Smith, Edward E, Stephen Michael Kosslyn, and Lawrence W Barsalou (2007). *Cognitive psychology: Mind and brain*. Vol. 6. Pearson/Prentice Hall Upper Saddle River.

Stemberger, Joseph Paul (1983a). *Speech errors and theoretical phonology: A review*. Indiana University Linguistics Club.

— (1983b). "The nature of [R] and [L] in English/evidence from speech errors". In: *Journal of Phonetics* 11.2, pp. 139–147.

Stokes, Stephanie F (2010). "Neighborhood density and word frequency predict vocabulary size in toddlers". In: *Journal of Speech, Language, and Hearing Research* 53.3, pp. 670–683.

— (2014). "The impact of phonological neighborhood density on typical and atypical emerging lexicons". In: *Journal of Child Language* 41.3, pp. 634–657.

Stokes, Stephanie F et al. (2012). "Statistical learning in emerging lexicons: The case of Danish". In: *Journal of Speech, Language, and Hearing Research* 55.5, pp. 1265–1273.

Storkel, Holly L (2009). "Developmental differences in the effects of phonological, lexical and semantic variables on word learning by infants". In: *Journal of child language* 36.2, pp. 291–321.

Storkel, Holly L and Su-Yeon Lee (2011). "The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children". In: *Language and Cognitive Processes* 26.2, pp. 191–211.

Takac, Martin and Alistair Knott (2015). "A neural network model of episode representations in working memory". In: *Cognitive Computation* 7.5, pp. 509–525.

Takac, Martin, Alistair Knott, and Stephanie Stokes (2017). "What can Neighbourhood Density effects tell us about word learning? Insights from a connectionist model of vocabulary development". In: *Journal of child language* 44.2, pp. 346–379.

Tremblay, Pascale and Anthony Steven Dick (2016). "Broca and Wernicke are dead, or moving past the classic model of language neurobiology". In: *Brain and language* 162, pp. 60–71.

Vitevitch, Michael S and Holly L Storkel (2013). "Examining the acquisition of phonological word forms with computational experiments". In: *Language and speech* 56.4, pp. 493–527.

Werbos, Paul J (1990). "Backpropagation through time: what it does and how to do it". In: *Proceedings of the IEEE* 78.10, pp. 1550–1560.

Wilson, Robert Andrew and Frank C Keil (2001). *The MIT encyclopedia of the cognitive sciences*. MIT press.

Wilson, Stephen M and Marco Iacoboni (2006). "Neural responses to non-native phonemes varying in producibility: Evidence for the sensorimotor nature of speech perception". In: *Neuroimage* 33.1, pp. 316–325.