The Impact of the Older Sibling on First Language Acquisition of the Later Born Child - A Computational Modeling Approach

Sema Karan
STUDENT NUMBER: 2031165

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

Afra Alishahi
Grzegorz Chrupala

Tilburg University
School of Humanities and Digital Sciences
Tilburg, The Netherlands
May 2019

Preface

This thesis is the final product of my Master's program Data Science and Society at Tilburg University. Almost one year ago, when I started to study in this program next to my 4-days part-time job by commuting from Amsterdam in really early mornings, I am often told it is not possible to finish such a master degree on time. Now, I am glad that I came to the end of my master journey by accomplishing all the courses and submitting this thesis. I always knew that determination, ambition and hardworking will reward eventually but it wouldn't be possible without the support and encouragement of the following people.

First and foremost I would like to thank Afra Alishahi for her guidance, insightful comments, and encouragement throughout the process of writing this challenging thesis.

I am sincerely grateful to Lieke Gelderloos for supporting me on the algorithm architecture which was the keystone of this thesis.

I would like to thank my fellow master students Jora Peters, Shelly Van Erp, who also worked on the similar topic, for the stimulating discussions and sharing experiences.

Finally, I would like to express my love and appreciation for all support by my family, friends and colleagues. In particular, Yusuf Soydinc, as loving partner and supportive thoughtful friend was always with me whenever I was stuck.

Sema Karan
Tilburg, June 2019

The Impact of the Older Sibling on First Language Acquisition of the Later Born Child - A Computational Modeling Approach

In this study we investigate whether the presence of an older sibling has an impact on word learning of the later born child, by applying a computational modeling approach. The presence of an older sibling could have an impact on child's word learning. There are different conversational turn-takings in child-parent and sibling-child-parent scenarios. As the linguistic input from the parent may differ in different scenarios, a formalization of dyadic and triadic scenarios is implemented in a word learning model. The model receives annotated Flickr30k Entities images as input, and is trained in two scenarios. In the dyadic condition, the model chooses objects randomly and receives the linguistic input about the object only from the parent, whereas in the triadic condition, the linguistic input comes half of the time from the parent and half of the time from the sibling. The goal of this study is to show how the presence of a sibling impacts child's word learning.
In triadic condition, child's word learning was slower compared to dyadic condition. If the siblings age is older, which is defined by the amount of words learned, it still makes child's word learning slower.

Keywords: *word learning; sibling impact; connectionist model.*

1. Introduction

This section provides an introduction to the study, the problem statement with research questions, and the methodology that has been applied.

1.1 Background

Children are born to complex and interactive environment where they are exposed to social interactions, sounds of speech, noise from different sources and visual inputs of the objects around them during the early childhood. How children learn to speak is still one of the elusive tasks in cognitive science and it has been researched for centuries to enlighten different aspects of language learning such as word production, word comprehension, sentence production, etc. Besides, environmental factors such as cultural background and socioeconomic status of the family are also continuously researched to inform the impact on child's language acquisition.

Language learning is a skill that infants attain by the age of three, even though it's a really complex task (Lust 2006). At the age of seven they are speaking with the full fluency of an adult; by which time they have the ability to process the infinite number of sentences belonging to their language and they are capable of producing an infinite

number of sentences themselves. This would be remarkable even if the procedure for producing (and understanding) sentences could be described by a well defined set of rules: however, human languages are so complicated that we have not yet managed to settle upon a theory that describes all linguistic phenomena satisfactorily (Buttery 2006). The underlying mechanisms for the acquisition of language system has been questioned for long time. Up to now, a number of experimental methodologies have been used to explore how language is learned and used by children. The empirical results of these studies have demonstrated important clues about child language learning. Besides, computational models have simulated the plausible mechanisms of language acquisition to provide insight and these studies are facilitated by the large collections of child-directed and child-produced data which are gathered by researchers. One of the main inputs in both experimental studies and computational models is linguistic input to the child. A question related to the importance of input in language acquisition is whether the quality and the quantity of the linguistic input in word learning improve the child's language learning. Although it is still unknown what the exact quality and quantity of linguistic input should be, Murphy and Slorach (1983) suggest that there is a minimum quantity of linguistic input for language acquisition with a study. They report that a hearing child of deaf parents did not learn spoken English despite having been exposed to television. Linguistic input to the child is an important naturalistic factor that plays a critical role on infant's language acquisition. The quantity of the linguistic input from caregiver has an impact on vocabulary learning and word production of child. Kuhl et al. (1997) indicates that language input provides a rich and detailed source of information that instigates, before word learning, a process of species-specific mapping of information by the brain, a process that alters the infant's perception and perceptual-motor system to conform to a specific language. One study by Narafshan et al. (2014) investigated the effect of the quantity of language input on first language acquisition with six Iranian infants learning Persian as their first language. The infants are followed in two groups during 12 months (24-36 months) where one of the group members received less linguistic input from caregiver (their mothers were mute) compare to the other group (their mothers were normal speaking parents). The infants who received normal input and the infants who received less input, showed some important points which highlighted the important role of input in first language acquisition. Comparing these infants with the infants who received less input, they both followed the same process, but for the children who received less input the process happened with some delay and more slowly.The study showed that a small amount of input may cause language comprehension, but language production will be delayed until the learner receives enough amount of input (Narafshan et al. 2014). Besides the quantity of the linguistic input, the quality of linguistic input also plays an important role on child's word learning. Buttery (2006) indicates that lacking any discerning information, a child is likely to assume that all the utterances she/he hears are grammatical and therefore constitute positive evidence. However, spoken language can contain ungrammatical utterances, perhaps in the form of interruptions, lapses of concentration or slips-of-the-tongue. When a child misclassifies such utterances as positive evidence, an error has occurred. It can be assumed that this ungrammatical linguistic input might be the sibling utterances to the later-born. In child - parent interactions, the learner receives the linguistic input from the parent which is correct input (except the explicit negative feedback), whereas child - sibling - parent interactions the learner might receive correct input from the parent and noisy (or wrong) input from the sibling. Buttery (2006) claims that any simulation or explanation of language acquisition should therefore attempt to

learn from every utterance it encounters and should be robust to errors whether caused by erroneous utterances or general ambiguity.

Household members are the main agents in language learning environment of the infant by providing input about the objects around the infant. Especially older siblings, who do not exist in every child development environment, are additional agents for the later born children alongside a caregiver. Thus far, several studies have shown that the presence of an older sibling creates a linguistic environment for second born children that is qualitatively as well as quantitatively different from that of firstborn children (Oshima-Takane and Robbins 2003). Although most of the studies in child language learning are mainly focused on caregiver-child interaction with(out) intrinsic factors or only child in the family scenario during the first language acquisition, in most of the families, there are also other older sibling(s) in the learning environment of the later born child as a second interaction. According to OECD (Organisation for Economic Co-operation and Development) statistics[1], EU average for the proportion of the families with one child and the proportion of the families with more than one child was almost at the same rate, 15% and 16% respectively in 2016, where 69% of the families were childless. Many children live in households where multiple adults and other children are present for large portions of the day. In such households, young children are likely to hear speech directed to them from older siblings and other household members (Shneidman et al. 2013). Regarding the families with older sibling, the produced words by the sibling can be a noise or a corrective feedback for the later born. It has been found that the older siblings' inability to adjust their speech may be related to the later-borns' slower rate of language acquisition (Tomasello and Mannle 1985). It is also found that preschool-age siblings did not properly adapt their speech when speaking to their infant siblings, although school-age siblings did adjust to some extent (Hoff-Ginsberg and Krueger 1991; Mannle, Barton, and Tomasello 1992).

Besides the quantity and the quality of the input from the early born to later born, the direct input from caregiver may also vary in single-child families compare to the families with more than one child. Downey (2001) indicates that beginning with the assumption of parental resources are finite, resource dilution model explains that as the number of children in the family increases, the proportion of parental resources accrued by any one child decreases and that variations in parental resources or investments have an important influence on children's intellectual development and educational success. Then, do the later-born children rely on input from their older siblings for learning language, and how this additional (potentially noisy) source of input affects the language learning of later-born children? What would be the impact of having an older sibling on the quantity and quality of the words provided by caregiver for the later born in two-children families? Is the presence of an older sibling is an advantage or disadvantage for the later born during the language acquisition? How does the age difference between the siblings impact the language acquisition of the later-born child?

These questions have been researched many times by psychologists, psycho-linguistics researchers with the observational studies to find a relationship between different factors at first language acquisition or with experimental studies to seek an inferential and generalizable conclusion about the impact of different factors to some extent. However, the studies in computational linguistics that address such questions are limited, which is one of the main motivations behind this thesis.

1 http://www.oecd.org/els/family/database.htm

1.2 Problem Statement

Regarding the previous research and substantial evidence that the quality and the quantity of linguistic input are crucial for infants' language development, in this study the impact of an older sibling on the word learning of later born will be investigated since the linguistic input is different in multi-child families compared to single-child families. Specifically, a computational connectionist model of language acquisition will be implemented to study the impact of an older sibling on the language acquisition of the second born child.

In summary, regardless of the negative or positive effect of the birth order, it can be assumed that the caregiver may spend more time to reply to the first-born child and when the second born joins the family environment the attention will be divided between the children. Accordingly, the caregiver will produce less direct input to the second child who will also start to receive linguistic input from the first child in the same learning environment. Besides, the linguistic input form firstborn to the later-born will be noisier compared to the caregiver input. This study will focus on understanding the impact of the linguistic input from the older sibling on the word learning process of the later born child by assuming less linguistic input from caregiver compared to the first born child. Thus, the research question will be;

"RQ: What is the impact of having an older sibling in the language learning environment on the word learning of the later-born child?".

The literature review has led to the following hypotheses to answer the research question regarding the impact of the older sibling presence on the later born child in child-sibling-parent (triadic) scenario;

"H1: The nosier linguistic input from firstborn and the lesser linguistic input(compare to the firstborn) from caregiver will make word learning of later-born slower.".

"H2: By assuming the linguistic input from sibling improves with the age; the older the sibling is, higher the quality of the speech produced by the sibling and this will have a positive impact on the word learning of the later-born child.".

The hypotheses shall be addressed by an experimental study which simulates the child language acquisition in two different scenarios, which are dyadic scenario, the older sibling does not exist and triadic scenario, the older sibling exists.

This topic is worth addressing because, from the societal perspective, the results of such a study may enlighten the families who are raising more than one child, about their children's language development. From the scientific perspective, such an experiment may contribute to our understanding of human language learning.

2. Related Work

In this section, related empirical and computational studies have been researched on this topic will be demonstrated in four different aspects to support the motivation and hypothesis of this project.

2.1 The Development Environment of Second-born Child

The family environment starts to change with the arrival of the second child in various conditions. There are many studies to understand how the environmental conditions, where children grow, have an impact on the development of the children and most of them agree that quality of the house situation, characteristics of the parents and available resources affects the children's development. The resource dilution model posits that parental resources are finite and that as the number of children in the family increases, the resources accrued by any one child necessarily decline (Downey 2001). In the same study, the resource dilution model is featured in three ways which one of them is **the family resources are finite**. Blake (1981) divided these resources in three categories where the first one is settings and cultural objects, the second one is treatments and third one is opportunities. In this thesis project, the dilution of the treatments as a parental resource will be considered for the language acquisition simulation such as attention, intervention and teaching might be decreased for the second-born child. In detail, one of the main building blocks will be the assumption of less attention and intervention as an input from caregiver to second-born child during the language acquisition. Downey (2001) indicates that for some resources, such as parental attention, the dilution model would predict advantages to early-born children because they enjoy less diluted parental resources until subsequent children arrive. As we can presume that siblings can be also a resource for each other since they grow in the same interactive family environment, this argument will be the second building block of the study that there will be attention and intervention as an input from the firstborn to the second-born during the language acquisition of the second-born child. Oshima-Takane and Robbins (2003) claim that linguistic environment of second-born children is qualitatively different from that of firstborn children. With the presence of the second-born child, the conversation environment changes from dyadic context to triadic context where the utterances from the caregiver and the first-born child have also changed. The later-born child begins to overhear conversations between the mother and the older sibling which can influence the linguistic development of the later-born. Even though some studies such as Floor and Akhtar (2006) suggest that when memory demands are not too high, 18-month-old infants can learn words through overhearing, this specific argument is out of the scope of this thesis, since the simulation will cover mainly child directed conversations from the older sibling and the caregiver to the later-born. The other variable in this context to be considered is the age of older sibling which will impact the quality of input to the later-born. Hoff-Ginsberg and Krueger (1991) studied how the older siblings can be a conversation partner to the younger child with an observational study. They observed the children between 1.5-3 years-old in dyadic interaction with their 4- or 5-year-old older siblings, their 7- or 8-year-old older siblings, and their mothers and the significant differences draw two conclusions: 7 to 8 year-olds provide more supportive interactions to their young language learning siblings than do 4 to 5 year-olds, and 7 to 8 year-olds provide less supportive interactions than mothers. To summarise, it's most likely that birth order creates different language learning environments for the second-born child which has not been concluded as detrimental. However, it is known that there are types of conversation and input that children are exposed to in families with more than one child, and having multiple children affects daily routines and interactions for the second-born child.

2.2 Sibling Effect on First Language Acquisition

The research in child development, specifically about the sibling effect on language acquisition are numerous and controversial. Regarding the resource dilution model, it might be presumed that the first-born children have advantage over later-born siblings in terms of parental resources. Some studies found that there are differences in favour of the firstborn children regarding the conversation initiation and language skills (Pine 1995). On the one hand, a study by Jones and Adamson (1987) showed that there is a significant difference between the early-born and the later-born children on a maternal-report measure of vocabulary size in terms of a birth-order effect on children's rate of vocabulary development and it is found in the sample of the study that firstborns did have significantly larger 'Reported Vocabulary Estimates' than later-borns. On the other hand, Oshima-Takane and Derevensky (1990) did not find any differences between first and later-born children on an observational measure of vocabulary size in their studies, but they found a significant advantage for the later-born in the acquisition of personal pronouns. Hoff-Ginsberg (1998) found that the first-born children are more advanced in vocabulary and grammatical development than later-born children, yet later-born children were more advanced in their conversational skills. Pine (1995) also found that the first-born children reached the 50 words milestone earlier than the second-born children, but it is not found that there is a difference at the point of 100 words. Besides, a study with one thousand 18-months-old children by Berglund, Erikson, and Westerlund (2005) found a significant negative effect of birth order on both the word production and the comprehension of words. Another study also indicated that the mothers of the first-born children have been found to make more explicit attempts at eliciting language from their toddlers than the mothers of the later-born children (Jones and Adamson 1987). All these differences between the first-born and the later-born can be assumed to be the result of the different amount and kind of the parental input. Raikes et al. (2006) indicated that the first-born children are read to more often than the later-born children, and as expected these children obtain more linguistic input from their parents and the children can express themselves more explicitly. Besides all these studies, other aspect of the presence of an older sibling might be related to age difference. Bornstein, Leach, and Haynes (2004) conducted a correlation analyses of *vocabulary competence in first- and second-born siblings of the same chronological age* and found that no second-born language variable was related to the age difference between the siblings. Thus, sibling spacing was unrelated to the second-born vocabulary competence.

In this study, we are particularly interested in mother-child (dyadic) versus mother-sibling-child (triads) scenarios when there is a conversational turn-taking between the agents in each scenario. A study about the effect of the linguistic interactions between the language learners and their older siblings found that preschool-age siblings did not properly adjust their speech when speaking to their infant siblings, although school-age siblings did adjust to some extent. It has been suggested that the older siblings inability to adjust their speech may be related to the later-borns' slower rate of language acquisition(Tomasello and Mannle 1985). Therefore, the age of the older sibling might be a hyper-parameter for the computational model in this study and it can be assumed that linguistic input from sibling will be more noisy in early ages and becomes less noisy at later ages of sibling. Another assumption would be that in triadic scenario, later-born will get input from sibling since the input from mother will be shared among the sibling and child.

2.3 Connectionist Modeling Approach

Over the last few decades research in different disciplines such as psycho-linguistics, linguistics, biology, neuroscience, psychology, phonetics and speech technology has resulted in the design of theories and models that account for parts of the chain that links the intentions of the speaker and the comprehension of the listener (ten Bosch et al. 2009). Various computational approaches towards the modeling of language learning are used to distinguish five different categories such as the generative viewpoint, a statistically-based approach, a social/embodiment based approach, a child-based developmental approach, and the cultural evaluation viewpoint (Kaplan, Oudeyer, and Bergen 2008). The first connectionist models used to model language focused on verb morphology. A neural network was built by Rumelhart and McClelland (1986) to predict the past tense of English verbs. The network performed well, trained on a set of mostly irregular verbs. It was even shown to have generalised for patterns within the irregular verbs that were not in the training set (Ingram 1989). As another example, a computational model was developed by ten Bosch et al. (2009) to detect and build word-like representations on the basis of sensory input and showed that a robust word representation can be learned in using less than 100 acoustic tokens (examples) of that word. Another computational modeling study has been performed by (Roy and Pentland 2002) to model speech segmentation, word discovery and visual categorization from spontaneous infant-directed speech paired with video images of single objects and the results demonstrate the possibility of using state-of-the-art techniques from sensory pattern recognition and machine learning to implement cognitive models which can process raw sensor data without the need for human transcription or labeling. Zhao and Li (2005) also used connectionist approach in their study by experimenting a self-organizing neural network model of early word production. The model is used to simulate the early stages of lexical acquisition in children and the simulating results indicate a number of important effects in determining the timing and function of children's word production, such as word frequency and word length effects. The differences between symbolic view and connectionist view over the past research is summarized by Tasnimi (2015) in Table 1.

Traditional symbolic view	Connectionist view
1.Language is rule-governed.	1. Language is only based on construction of associations.
2. Language is viewed modular and language learning is seen as mastering the modules.	2. Language learning is the same as learning other types of knowledge or skills.
3. A distinction is made between competence and performance.	3. No distinction is made between competence and performance.
4. There is a central executive in the mind to control the processing.	4. Control is distributed among the parts of the network.
5. Processing is serial and linear.	5. Processing takes place simultaneously.
6. Higher cognitive functions (such as memory) take place in the mind.	6. Higher cognitive functions (such as memory) are not discussed.
7. Modeling is done through algorithm.	7. Modeling is done through neural networks

Table 1
A comparison between traditional symbolic view and connectionist view(Tasnimi 2015)

2.4 Computational Models of Language Learning.

Connectionist theories are data-rich and process-light: massively parallel systems of artificial neurons use simple learning processes to statistically abstract information from masses of input data (Ellis 1998). In traditional symbolic view language learning is seen as rule governed; therefore, language learners develop complex, internalized rule systems that can be represented symbolically. In contrast, connectionism claims that language learning is not rule governed. Language is only based on construction of associations (Gasser 1990). In computational linguistics, there are some limitations that connectionist models may deal with but traditional symbolic models may not. A connectionist model can account for lack, partial and incorrect storage of knowledge. This can be justified as the strength of the connections between nodes. Items which are not repeated or met frequently have weak interconnections. Therefore, constant practice and reinforcement is necessary to strengthen the interconnections (Waring 1996).

In the field of cognitive science, computational modeling refers to using computational tools and techniques in order to simulate a cognitive process, and explain the observed human behaviour during that process (Alishahi 2010). That being said, a computational model can be implemented to simulate child language acquisition process, specifically word comprehension and word production, by feeding visual object inputs as a representation of the real world and allowing this model to map these objects back to the words, their meanings.

3. Experimental Setup

In this section, the setup for the experiments are described in order to answer the research questions. Data processing, methodology, performance evaluation and required software are explained in each subsection.

3.1 Data

In this project, the Flicker30K dataset (Young et al. 2014) is used as the input of the computational model to represent the visual context of child language development. Young et al. (2014) captured 31,785 images from Flickr and also collect 158,915 captions (five captions per image) via crowdsourcing. The images describes everyday activities and scenes by focusing mainly on animals and people. Each image is described independently by five annotators who are not familiar with images that they are depicting beforehand and different annotators use different levels of specificity, from describing the overall situation to specific actions. Plummer et al. (2015) extended the dataset with 244,035 coreference chains by linking mentions of the same entities across different captions for the same image, and associating them with 275,775 manually annotated bounding boxes as shown in Figure 1. Regarding the purpose of the current research question, which does not require five captions per image, the simplified version of the dataset is used.

Figure 1
Examples of annotated images from Flickr30K data set

The simplified version of the dataset is created by Keijser (2018) for his study where a connectionist modeling approach is applied to investigate the role of curiosity on first language acquisition. To find the descriptions of every object, The Flickr30k Entities sentences file which contains the annotated captions for each image were searched. Every object might be attached to more than one word but for the simplified dataset, the most frequent word was chosen as the single word most likely to describe the object in the image. Such a simplification step has been done after excluding very frequent and irrelevant words such as articles (a, an, the), third-person possessive determiners (his, her, their), the cardinal numbers one through ten, and primary and secondary colors (e.g. orange), including silver and gold. In the case of different objects which are labeled with same word in an image, only the first in the loop was selected. Lastly, images with

less than two objects were removed after reprocessing in order to ensure that there is more than one option that model can still pick one randomly to learn. Table 2 shows the distribution of the images per number of objects included in each image. For example, there are 6481 images in the dataset which include 2 objects.

n of objs	2	3	4	5	6	7	8	9	10	11	12	13
image freq.	6481	512	2860	4988	6280	1295	177	54	19	2	1	1

Table 2
Frequency of images per number of objects included

The simplified dataset contains a total of 86,748 word-object pairs and 4,237 unique words. The least frequent words (e.g. sites and paste) occurred only once and there were 1882 words that occurred only once, while the most frequent word occurred 7,891 times. The five most frequent words were: man (7,891 times), shirt (4,536 times), woman (4,378 times), boy (1,477 times), and girl (1,428 times). The average frequency was 20.47 (SD = 172.33), and the median frequency was 2. After dataset simplification, 24,670 images remained, where 1,000 of them are kept as validation set and another 1,000 of them as test set. When the data is split to test and train, there are 79,749 objects in the training data and 3,943 objects for test data.

Data	Objects	li.baseline	sp.baseline
train data	79,749	0.284	0.091
test data	3,943	0.286	0.089

Table 3
Number of objects and baselines per split

Table 3 shows the total number of objects in the images for train and test splits. There are two parts in the model as it is explained in the next section, the listener and the speaker. The baselines in Table 3 are reference points for the model performance. Listener baseline shows the mean chance of an object being chosen from a visual scene, as each vocabulary in the data may have different frequency. The speaker baseline is the majority baseline of the most frequent word. The baselines tells about the average accuracy obtainable by chance, which are assumed to be the minimal performance expected of the model. If the model performs better than the baselines, it is expected to observe higher accuracy which shows how good the model performs. Since many words occur only once or twice, there are 80 words in the test set that do not occur in the training set, with a token frequency of 80, and 776 words, with a token frequency of 3413 in the test set, that do occur in the training set. These numbers might suppress test accuracy, but should be sufficient to test the model accurately.

3.2 Model

Connectionist modeling is a type of computational algorithm which represents the neural processing in the brain. Connectionist models, also known as artificial neural networks are based upon the architecture of neural networks in the brain. Each network consists of units (which are analogous to neurons) that are connected together by

weighted links (modelling synapses). Each unit is assigned an activation level which determines whether the unit will produce data at its output: much like the activation energy required by a neuron in order to fire. Thus neural networks consist of four parts (units, activations, connections, and connection weights) each of which corresponds to a particular structure or process in biological neural networks (Buttery 2006). The key element of neural networks is the novel structure of the information processing system. It consists of a large number of elements (neurons) which are sometimes at different number of layers and interconnected through weighted links and these neurons work in unison to solve specific problems. Artificial neural networks learn by example, like people, by receiving many input, processing them and transmitting the processed information to other neurons. McLeod, Plunkett, and Rolls (1998) summarised several principles about the connectionist networks that might show individual differences. In addition to initial weight states, they suggested three other sources of variation which are the learning rate of the network, in terms of how quickly weights can be changed in response to learning event, the number of internal units, and the training regime to which the network is exposed. To summarize, computational cognitive modeling is a new and rapidly developing field, but during its short life span, it has been extensively beneficial to cognitive science in general, and the study of natural language acquisition and use in particular (Alishahi 2010).

In this study, the model (represents the child) is an artificial neural network as described above. It is adapted from Lieke Gelderloos' and Daan Keijser's model of cross-situational word learning [2] based on the GroundeR model (Rohrbach et al. 2016). The model has two components as shown in Figure 2.

It consists of a listener and a speaker module. The listener module simulates a child word learning about the objects in the visual modality, by receiving linguistic input from a caregiver who is the conversation partner of the child. The listener learns under supervision. The listener learns to map a given word by a caregiver to its referent in the visual context and then updates its language knowledge accordingly. In the visual scene, there are number of objects. For each visual object in the visual scene, a visual feature vector is extracted by using the VGG-16 object recognition model presented by Simonyan and Zisserman (2014), pre-trained on ImageNet. The last fully connected 4096-dimensional layer is used to extract the high-level visual information about the object. During the training of the listener module, for each object in a given scene, the embedding of the word given by the caregiver is concatenated to the object representation as input to the listener. The input is passed through two hidden layers of 256 units, of which has a ReLu activation function. The listener is trained in supervised manner by using cross-entropy loss on the concatenated output values.

The speaker module learns to produce a word for given an object, which simulates the word production of the child as it maps the object chosen from a scene to the corresponding word. The input to the speaker is a VGG vector which is the representation of the objects with the length 4,096 which are extracted from the image. Then, this input fed to two hidden layers of 256 units. The activation function for the first layer is ReLu. For the second layer, a softmax activation function is applied to get probability distribution of word guesses (between 0 and 1) which is fully connected to the vocabulary-sized output layer as target. The target (label) has been converted to a one-hot coded vector of length 4238, which is vocabulary length plus one for the 'unknown' token, for words

2 The original model is a work of conference paper recently submitted and was shared with me by Ms. Gelderloos. The model is adapted by me to fit the needs of this project.

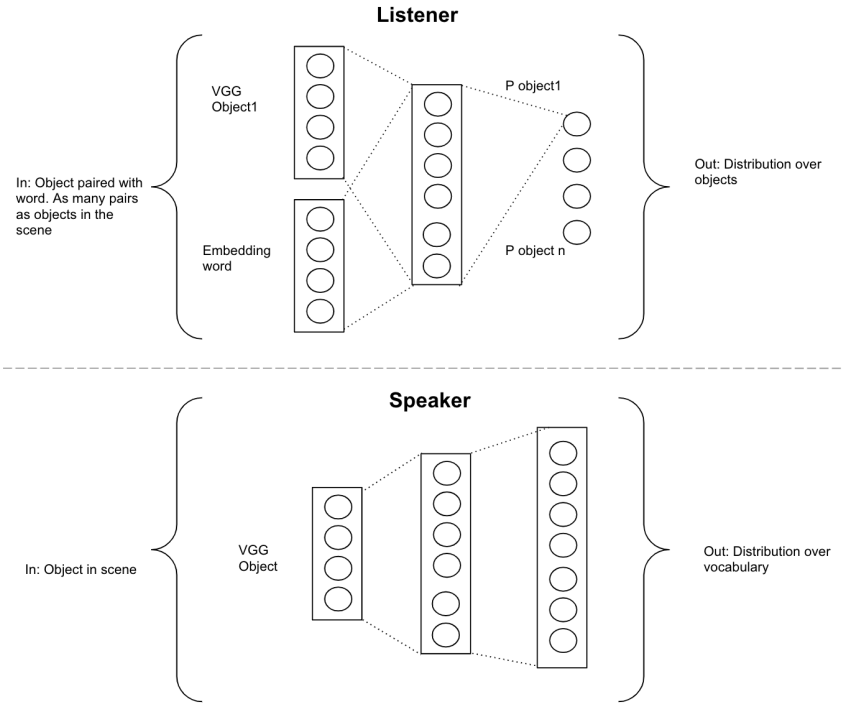

Figure 2
Simplified graphical representation of the model

that the model has not heard. The speaker is fed the sum of the VGG vectors of all objects in the scene, weighted by the output vector of the listener, instead of training on a single object VGG vectors. This step connects the listener and speaker modules of the model, so the word production is done (by the speaker) by using the listener output, which is a supervised learner. The speaker is trained using cross-entropy loss in a self-supervised manner. Since the self-supervision signal consists of the original input word to the listener, the performance of the speaker is dependent on the listener (which is trained under supervision) although the speaker is assumed to be trained in an unsupervised manner.

The model, which represents the child as a learner, observes the scene that consists of from different objects and tries to figure out which object was talked about based on the linguistic input and the representation of the whole visual context. Like a caregiver or a sibling initiates the conversation by choosing an object from the environment surrounded by, the model attempts to produce the word which corresponds to the chosen object. The experimental setup of this study is simulated in two different scenarios to investigate the sibling effect of the child's. Figure 3 shows the symbolic representations of these two scenarios which will be explained in the next sub-sections.

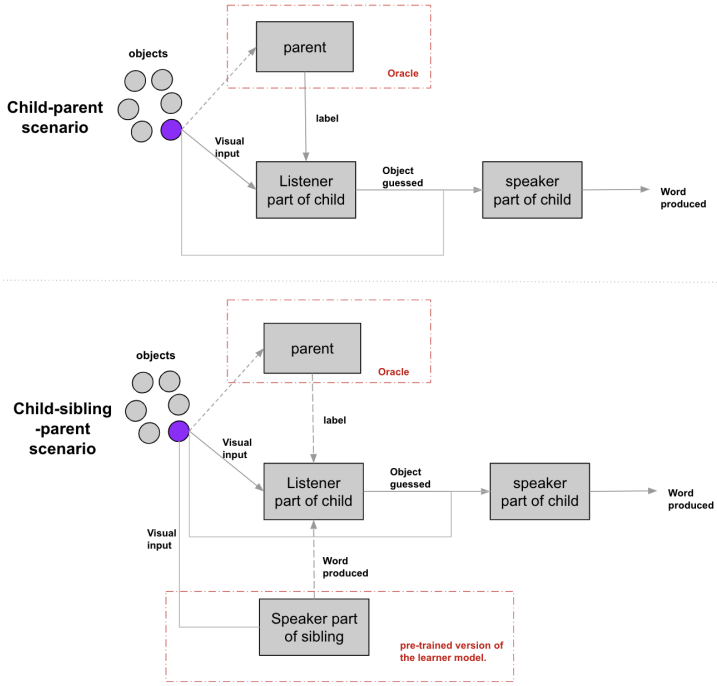

Figure 3
Symbolic representation of conversational turn-taking in dyadic and triadic scenarios

3.2.1 Dyadic Scenario.

In the real world of child-parent scenario (dyadic), the caregiver (oracle) picks an object from the learning environment of the child and gives the corresponding linguistic input of that object to the learner (the child). The learner tries to map the input given to the object chosen and finally produces a word about the object. This flow represents how the child figures out which object is talked about based on the linguistic input and the representation of the whole visual scene. In the simulation of this scenario, the model (the artificial network architecture as explained in the method section) processes the visual input as representation of the randomly chosen object and also receives the corresponding label of that object to produce the word about it through the listener and the speaker modules. The accuracy of the produced word by the model represents the word learning process of the child. The model is trained in dyadic condition two times to investigate the hypotheses separately, and the only difference is the amount of data used during the training. In the first one, the model is trained on the data which consists of one object per image and in the second one the training data consists of two objects per image. The first and second trained models are saved as sibling 1 and sibling 2 for the later use in triadic scenario to represents the older siblings at difference age (based on the amount of data trained). Figure 4 shows the pseudo code of the training algorithm of dyadic condition.

Child - Parent Scenario

```
Load child agent
Initialize speaker gradients to 0
Initialize listener gradients to 0
For each epoch in number of epochs:
   while batch number < number of total batches:
                For each image in batch:
                        Load  objects as visual objects
                        Load labels as language input
                        Add language input also as targets
                        Train listener
                                Get  the language input and visual input
                                Go through hidden layers & activation functions
                                Calculate the loss based on the targets
                                Update the listener weights
                                Return  the listener output
                        Train speaker
                                Get the listener output and visual input
                                Go through hidden layers & activation functions
                                Calculate the loss based on the language input
                                Update the speaker  weights
        Batch +=1
          Calculate total loss of speaker for each batch
          Calculate total loss of listener for each batch
          Calculate accuracy of speaker for each batch
          Calculate accuracy of listener for each batch
Save accuracies and losses for each epoch
{Save the model and parameters for sibling model}
```

Figure 4
Pseudo code of dyadic scenario

3.2.2 Triadic Scenario.

In the real word of child-sibling-parent scenario, the word learning process is as same as dyadic scenario for the child but the child receives the linguistic input of the chosen object half of the time from the caregiver and half of the time from the older sibling which means in triadic condition there is less linguistic input from the caregiver based on the resource dilution model (Blake 1981). The sibling(firstborn) receives linguistic input only from the parent and learns the words with more linguistic input from the parent compared to the later-born child.

The older sibling is the pre-trained model as explained in the dyadic scenario, but only the speaker part of the pre-trained model is used in traidic condition since we are only interested in passing the linguistic output of the sibling - speaker module to the child. Because it is not the goal to train sibling again in triadic condition, instead to use the parameters of the pre-trained model to obtain the noisy feedback from the sibling. Thus, when the linguistic input and visual input come from the caregiver, the model is trained as in dyadic condition, whereas when the linguistic input and visual input come

Child-Sibling-Parent Scenario

```
Load child agent
Load trained sibling - speaker module
Initialize speaker part of child  gradients to 0
Initialize listener part of child gradients to 0
For each epoch in number of epochs:
   while batch number < number of total batches:
                  For each image in batch:
                           If parent turn:
                                    Train the child as same as in dyadic condition
                           Else if sibling turn:
                                    Load  objects as visual objects
                                    Load labels as language input
                                    Add language input also as targets
                                    Get  the targets and visual input
                                    Return to one-hot encoded vector
                                    Train speaker sibling:
                                             Get the one-hot encoded input  and visual input
                                             Go through hidden layers & activation functions
                                             Calculate the loss based on the language input
                                             Update the sibling speaker  weights
                                             Returns word guesses of sibling speaker
                                    Train listener child:
                                             Get word guesses of sibling speaker and visual input
                                             Go through hidden layers & activation functions
                                             Calculate the loss based on the targets
                                             Update the child listener weights
                                             Return  the child listener output
                                    Train speaker child
                                             Get the child  listener output and visual input
                                             Go through hidden layers & activation functions
                                             Calculate the loss based on the language input
                                             Update the child  speaker  weights
                  Batch +=1
         Calculate total loss of speaker for each batch
         Calculate total loss of listener for each batch
         Calculate accuracy of speaker for each batch
         Calculate accuracy of listener for each batch
Save accuracies and losses for each epoch
```

Figure 5
Pseudo code of triadic scenario

from the older sibling, the linguistic input and chosen object is first presented to the pre-trained speaker. To do so, the listener part of pre-trained sibling is not used and instead of the output of the listener (a smooth probability distribution over objects), one-hot vector as weights is fed to the sibling speaker (so that the target object has weight 1, and everything else has 0). Then the outcome of the pre-trained speaker is sent to the child (to the listener and speaker modules respectively).

In the simulation, the linguistic input is either correct word of the chosen object or the produced word by the speaker part of the pre-trained model as representation of the older sibling, which may or may not be correct. Figure 5 shows the pseudo code of the training algorithm of triadic condition.

3.2.3 Model Hyperparameters.

For this experimental setup, there were two main steps to follow to make sure that the right models have been chosen and the conclusions are drawn based on these models. Both dyadic and triadic scenario, it is implemented as below;

1. Choose the best hyper-parameters for child-parent scenario, such as learning rate, number of epoch, batch size. By defining the best model structure with these trials, two things are achieved;
   - The performance of the child in the dyadic condition is considered as a baseline to compare with triadic condition. The results will explain the word learning process in dyadic scenario.
   - The sibling model will be saved to be used in triadic scenario for next steps. The reason to use the same model as sibling; first it will be proved that the best model is chosen for sibling representation and second, word learning process will be kept same for the sibling in triadic scenario and for the child in dyadic scenario since they both represents only child-parent interaction and the way of training the sibling and child will be consistent in terms of hyper-parameters.
2. Train the sibling twice with different amount of input data to represent the sibling with different ages.

Since the results of dyadic scenario create the baseline of the whole experiment, it is attempted to find the best model with the right learning rate, batch size and number of epoch with validation data after training the model on training data. One of the biggest challenges in deep learning is to choose the right hyper-parameters which will not cause overfitting or underfitting. In this experiment, different hyper-parameters are tested regarding the previous studies by Keijser (2018). First, the model is trained up to 20 epoch with different learning rates ranging from 0.01 to 0.00001. It is observed that in each trial, the model was not converging yet, so the number of epoch is set to 50. Then, for each learning rate different batches are tried, 40 batches and 100 batches. The models with a learning rate of 0.0001 performed better for both 40 and 100 batch size compared to the other learning rates. The model with 0.0001 learning rate and 40 batch size showed slightly better results compare to model with 0.0001 learning rate and 100 batch size. In the triadic scenario, the model is trained with a linguistic input which consists of word mixes from parent and sibling (1 input from parent:1 input from sibling).

3.3 Performance Evaluation

There are different performance evaluation metrics for neural network models depending on the type of the outcome predicted. For the model in this study, the accuracy of the model is used to evaluate how correctly the model can produce the words for the objects chosen. If the predictions are totally off, the loss function outputs a higher value and if the predictions are pretty good,then the loss function outputs a lower value. Deep learning neural networks need to be optimized during the training, it is the way how they are learning from input features and in this study Adam optimization algorithm is used. The error of the model should be estimated repeatedly as being part of the

optimization algorithm, so that is why a loss function needs to be chosen. Cross-entropy calculates a score which summarizes the average difference between the actual and predicted probability distributions for all classes in the problem.

The accuracy metric of the model is used to measure the algorithm's performance (accuracy) in an interpretable way. It is usually determined after the model parameters are learned and tuned, so no more learning takes place. Later the test split of the data are passed through the model and the number of correctly predicted outcomes, which are recorded by comparing with real target values, is calculated as model accuracy. It is more intuitive to interpret, basically says how many times the model predict the outcome correctly. In this study, number of epochs for a certain accuracy score is also examined closely with the goal of understanding how slow the sibling makes word learning process of later-born.

3.4 Software

The experiments in this study are done using Python (version 3.6.7) and Pytorch (Paszke et al. 2017) (version 1.0.0), which is Python-based scientific computing package and uses the power of graphics processing units. Pytorch provides two of the most high-level features which are tensor computations with strong GPU acceleration support and building deep neural networks on a tape-based auto-grad systems. The experiments are run on university server (server green) with device type "cuda". The code for running the experiments is documented in Jupyter Notebooks and Python files (.py) which are available on GitHub[3] for reproducible work and future adaptions. The model architecture implemented with Pytorch library in Python environment is as in Figure 2.

4. Results

This section explains the results of the experiments of two scenarios conducted.

4.1 Results of dyadic scenario

As it is explained in hyperparameter section, the results of dyadic scenario create the baseline of the whole experiment. The models with a learning rate of 0.0001 performed better for both 40 and 100 batch size compared to the other learning. The model with 0.0001 learning rate and 40 batch size showed slightly better results compare to model with 0.0001 learning rate and 100 batch size. For both listener and speaker, the accuracy and loss are calculated for each epoch in all possible combination of learning rates and batches as can be seen in appendix 6. To summarize how the best model (lr:0.0001, batch size 40) is chosen;

- Firstly, the accuracy of the listener and the speaker on both training and validation data is higher than the reference points (baselines) as explained in data section, which were around 0.2 and 0.09 respectively. Since the better accuracy is achieved in all possible hyperparameter combinations, it shows the model learns and perform well.

3 https://github.com/karanse/MSc-Thesis-Modeling-Human-Language-Learning

- Train and validation gaps for accuracy and loss graphs the least when the learning rate is 0.0001 both in 40 and 100 batches. And especially validation curves in each hyper-parameter option become more stable when the learning rate is 0.0001.
- The validation accuracy for listener is higher in the option with 40 batch size (0.74) compare to 100 batch size (0.71). The validation accuracy for speaker showed almost same rate in both batch size (0.32)
- The validation loss for listener decreased from 1.3 to 0.6 almost similarly in 40 and 100 batch size. But the gap between training and validation data set were closer in the model with 40 batch size.
- Lastly, the validation loss for speaker shows similar pattern, decreasing from 5.5 to 4.6 with 40 batch size and from 5.8 to 4.2 with 100 batch size. Training and validation lines were again closer with 40 bath size.

As a result, 0.0001 learning rate and 40 batch size have been chosen as hyper-parameters for this experiment. And the baseline for the study (dyadic scenario) is the results of these hyper-parameters as shown in Figure 6. Listener accuracy became stable around 0.74 and the speaker accuracy became stable at around 0.33 for validation dataset. Listener loss decreased at most to 0.63 after 35 epochs and speaker loss decreased at most to 4.2 after 30 epoch.

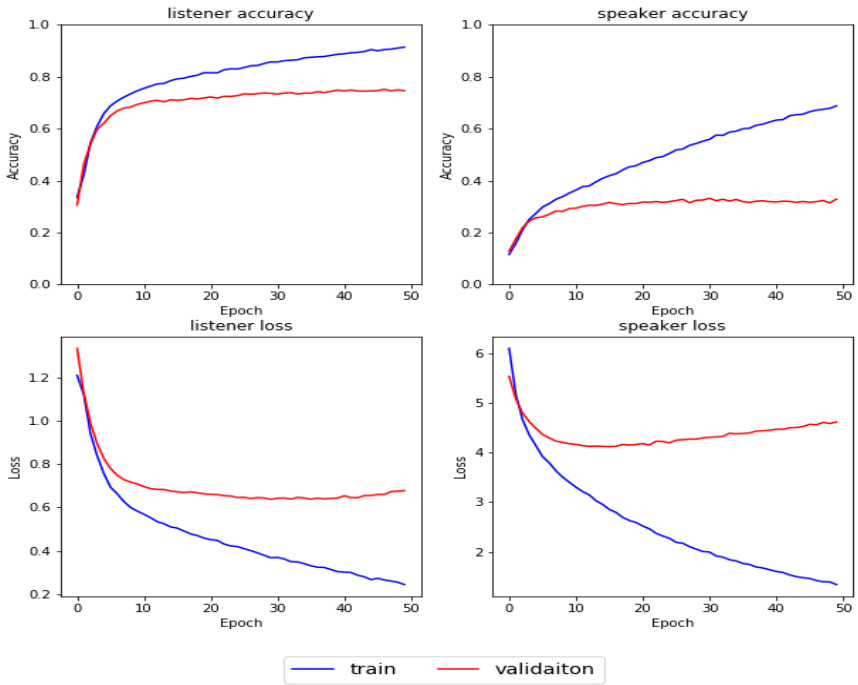

Figure 6
Train and validation results of dyadic condition: Accuracy and loss for child listener and speaker

The model with 0.0001 learning rate and 40 batch size is used going forward to represent the sibling as pre-trained model.

4.2 Results of triadic scenario

The model is trained with 40 batch size and 0.0001 learning rate (chosen as best model) where half of the linguistic inputs come from the parent and the other half come from the sibling. The results on validation data of the triadic scenario is shown in Figure 7. The accuracy for both the listener and the speaker decreased to 0.55 and to 0.23 respectively. The decrease in the accuracy is expected since half of the time the child starts to receive noisy linguistic input from the sibling which is not as accurate as the linguistic input from parent. The loss of the listener and the speaker started to decrease from 1.34 and 5.55 respectively as almost same starting point in the dyadic scenario but the minimum losses reached in triadic scenario were higher than dyadic scenario. In triadic scenario, two different models have been tried to test the second hypothesis as below;

- triadic condition with sibling 1, which is the baseline of the triadic scenario, the sibling was the same model as in dyadic scenario and it has been trained one randomly chosen object per image in the training data. The results of this trial are explained just above, Figure 7
- triadic condition with sibling 2, which is the same model with the same hyper-parameters but trained randomly chosen two objects per image in the training data to simulate the grown sibling and see whether it will affect the word learning of later-born compared to sibling 1. The results of this model can be found in Figure 8.

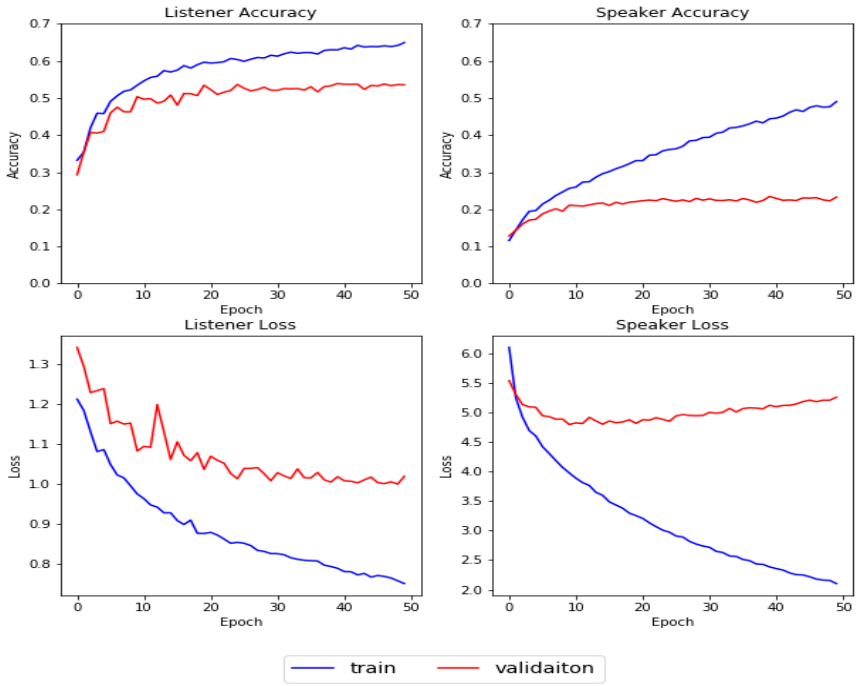

Figure 7
Train and validation results of triadic condition with sibling 1: Accuracy and loss for child listener and speaker

When the triadic scenarios with sibling 1 and sibling 2 compared, it is observed that even the sibling 2 is trained more and supposed the increase the child's word learning accuracy, it shows slightly lower validation accuracy both for listener and speaker modules of the child, which are 0.54 and 0.22 respectively. This might be explained that simulation of a growing sibling and the impact of this growth requires more trial with different hyper-parameters and more training data as well. Increasing the linguistic input 2 times to train a sibling was not enough to observe a clear difference on word learning of later-born child.

So far the dyadic and triadic scenarios are explained separately based on the validation data which allowed us to observe when the model stops learning, at which epoch the model converges by plotting the results of train data (blue lines in all the charts above) and validation data (red lines in all the charts above). Next, the test results of dyadic scenario versus triadic scenario, and triadic scenario with sibling 1 versus triadic scenario with sibling 2 will be compared to see the pattern on unseen data (test set).

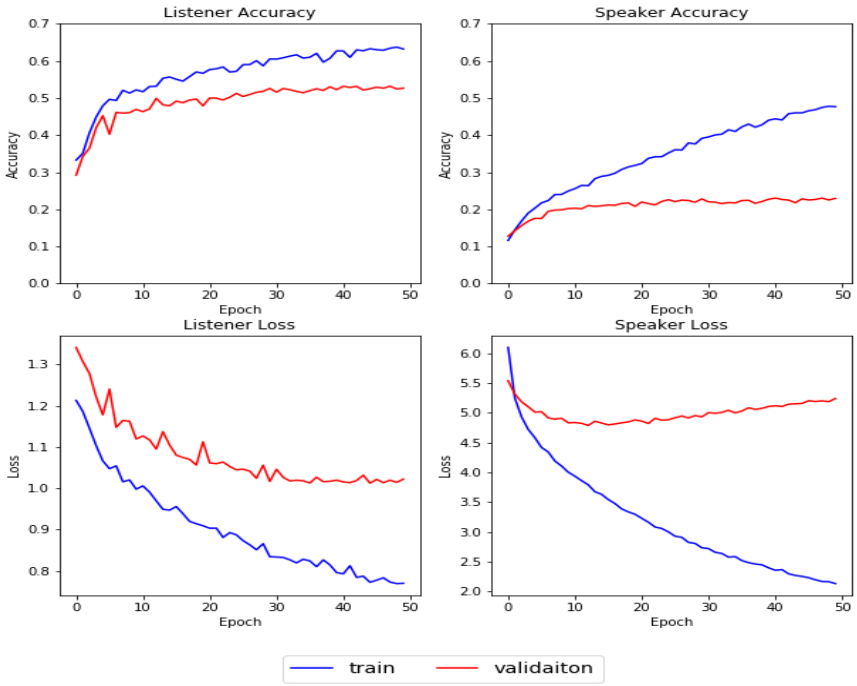

Figure 8
Train and validation results of triadic condition with sibling 2: Accuracy and loss for child listener and speaker

4.3 Comparisons of the scenarios

The results of dyadic and triadic scenario on the test data shows a clear difference in terms of model performance as shown in Figure 9. In both conditions, the word learning performance of the child started to stay stable or slightly change between epoch 20 and 30. The test results are also consistent with validation results that the presence of an older sibling makes word learning process of later-born slower due to the decreased amount of linguistic input from parent and half amount of noisy linguistic input from sibling. The speaker loss for both scenario increase after epoch 30 but it does not decrease speaker accuracy dramatically. The test accuracy of listener decreased from 0.76 (dyadic condition) to 0.55 (triadic condition), whereas the test accuracy of the speaker decreased less compare to the listener accuracy, from 0.32 to 0.23 respectively. The loss changes in both conditions show similar curves, the loss results of dyadic condition is better than triadic condition. When the dyadic and triadic conditions are compared on test data, the results show same increasing trend in terms of model performance, they started to converge at similar iteration and fluctuation of the curves follow the same pattern. This can be explained that in both scenarios, the model is optimized with similar descending pattern. The model does not stuck at local minima in any of the scenario, since we do not see any peak in loss and accuracy graphs, which is also supports the convenience of learning rate.

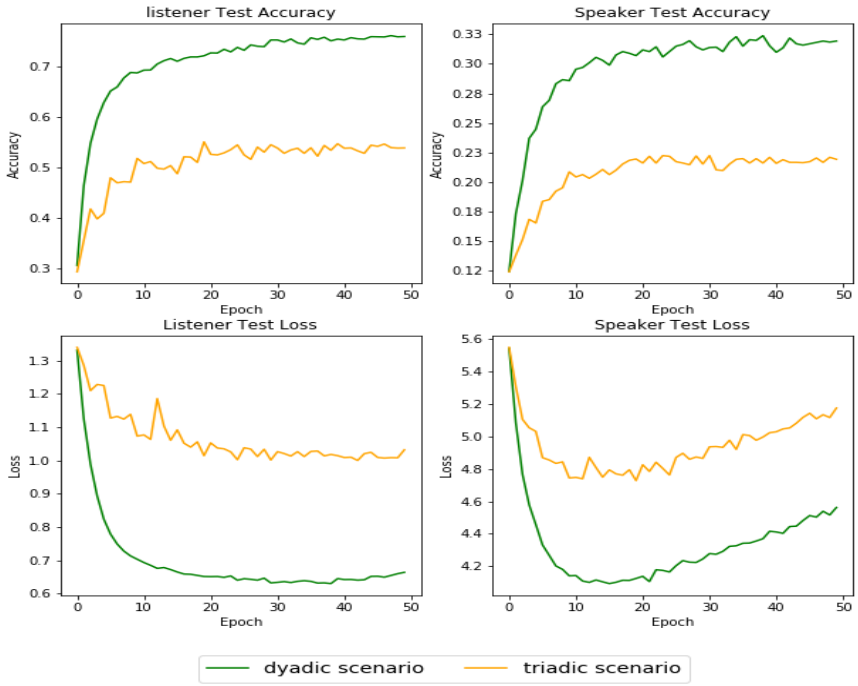

Figure 9
Test results of dyadic vs triadic scenario: Accuracy and loss for child listener and speaker

The tests results of two triadic conditions with sibling 1 and sibling 2 show similar performance and pattern as in Figure 10. The triadic scenario with the older sibling did not showed better results compared to the less trained sibling model. This can be either explained that the model needs to be train more than double size of the baseline condition or the word frequencies could be similar for more words that sibling model can learn better and will not pass unknown words as linguistic input to the child.

All the test results for the trained models in both conditions can be seen in Table 4 in detail.

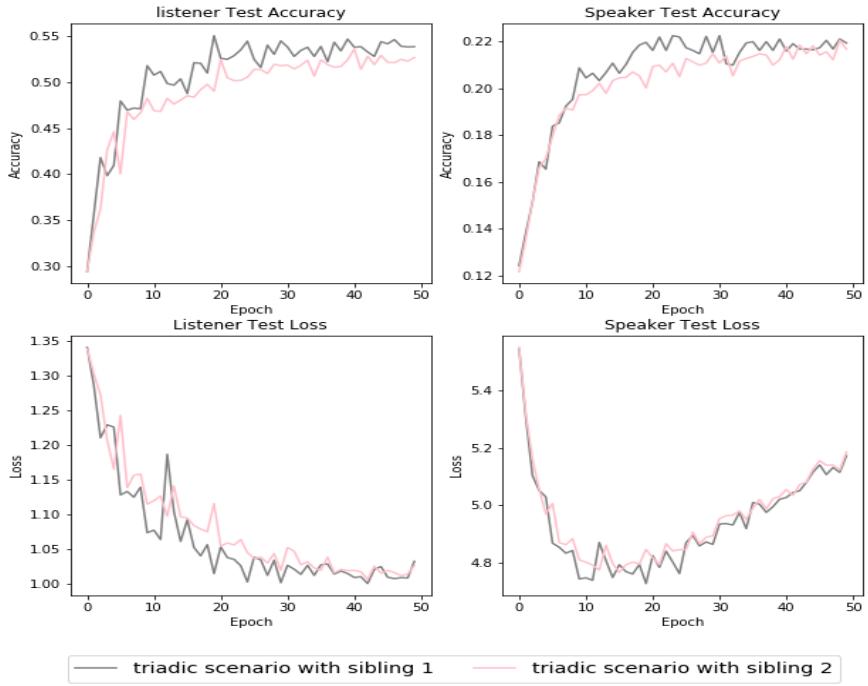

Figure 10
Test results of triadic scenario with sibling 1 and sibling 2 : Accuracy and loss for child listener and speaker
.

Test Results of Child Speaker & Listener **Hyperparameters:** batch size: 40, learning rate: 0.0001					
Scenario Desciption		**Max Listener Accuracy on Test**	**Max Speaker Accuracy on Test**	**Max-Min Listener Loss on Test**	**Max-Min Speaker Loss on Test**
Dyadic Scenario	Child - Parent (also Sibling 1)	**0.76**	**0.32**	**1.33 - 0.63**	**5.53 - 4.09**
	Child - Parent (Sibling 2 trained with more input)	**0.77**	**0.32**	**1.30 - 0.63**	**5.45 - 4.02**
Triadic Scenario	Sibling 1 - Child - Parent	**0.55**	**0.23**	**1.34 - 1.00**	**5.45- 4.72**
	Sibling 2 - Child - Parent	**0.54**	**0.22**	**1.34 - 1.01**	**5.55 - 4.76**

Table 4
The test results for all the scenarios : Accuracy and loss results for the child listener and speaker

5. Discussion

Did the presence of an older sibling impact the performance of the word learning model in triadic scenario? Yes, the accuracy of listener and speaker decreased in triadic condition compare to the dyadic condition. The loss was also lower both for listener and speaker in triadic condition as shown in Figure 9. This results support the dilution model which indicates the advantages to early-born children who enjoy less diluted parental resources until subsequent children arrive (Downey 2001). In dyadic condition, the child is exposed to more child-directed speech compared to the triadic condition and this will yield better performance on word learning. This is explained also as indicated by Raikes et al. (2006) that first-born children are read more often than later-born children, and as expected these children obtain more linguistic input from their parents and the children can express themselves more explicitly. The results of the study is also inline with the empirical studies by Narafshan et al. (2014), who indicated that when there is less linguistic input from the parent, the word learning process becomes slower and by Jones and Adamson (1987), who found that firstborns did have significantly larger 'Reported Vocabulary Estimates' than later-borns as explained in introduction and related work sections.

In the triadic conditions, it is assumed that the child receives the input both from the caregiver and the sibling, which is also indicated by Shneidman et al. (2013) that young children are likely to hear speech directed to them from older siblings and other household members. However, there is a limitation with this assumption. Humans are known to learn preferentially from more expert, it might be the caregiver in this study not the older sibling, but the modeling framework assumes that the child does not make any distinction between the caregiver and the sibling and learns from both indiscriminately. Even Buttery (2006) claims that any simulation or explanation of language acquisition should attempt to learn from every utterance it encounters and should be robust to errors whether caused by erroneous utterances or general ambiguity, it is still unclear how the children differentiate the linguistic input as noisy or not and learns accordingly.

Did the older sibling impact the word learning of the child positive compared to the younger sibling? No, the test accuracy results for the speaker and the listener in different triads did not show a remarkable difference to support the second hypothesis. Tomasello and Mannle (1985) suggested a contrary result about the the different impact of school-age siblings and preschool-age siblings. In this study, it has not been researched deeply how to simulate preschool-age and school-age siblings which may explain why the results in triadic conditions were different than expected. Another reason might be the objects are always chosen randomly, so while training the siblings it is not controlled which words are learnt first and passed to the child. It might be the case that the sibling 2 might be trained with less frequent words and eventually could not give better linguistic input to the child. The results of the triadic conditions are also contrary the study by Hoff-Ginsberg and Krueger (1991) as mentioned in related work section. But the results of second hypotheses is inline with the study by Bornstein, Leach, and Haynes (2004) as they suggested that sibling spacing was unrelated to the second-born vocabulary competence.

One of the limitations is to simulate the age effect of the sibling, since in this study, only the amount of word learned during the training is assumed to represent older sibling. Zajonc and Markus (1975) stated in their study that the older the sibling, the higher their intellectual abilities and thus the less disruptive they are to their younger siblings'development but simulating the intellectual abilities of the sibling in a plausible way is an enormous limitation of this study. Thus, although the present study was well

designed to simulate the effect of siblings with different ages, the age-gap effect remains uncertain.

6. Conclusion and Future Work

Deep learning models are complex architectures with too many hyper-parameters. This complexity also brings flexibility that allows us to build such models for human behaviour simulation. Since we do not have hard and fast rules for configuring such a network for the expected outcome, it is always a lot of trial and error iteration. So, for this study there are still many different hyper-parameters and conditions that should be implemented and the results are carefully tracked. Future work might be a deeper research specifically about the sibling's age impact child's word learning. A fresh approach can be followed to train sibling differently to expect better results from triadic scenario. As explained in the discussion section, finding a plausible mechanism to simulate the siblings with different ages can be one of the future works. Another suggestion for this study can be the conversation initiation in dyadic and triadic conditions. Currently, it is implemented that the object is chosen from a visual scene by the caregiver or the sibling randomly and the turn-taking during the conversation happens respectively but in the real world, the conversation might be initiated by child and the conversation turn-taking can be different, e.g the child may speak three times while the caregiver does once. So, considering such a plausible criterion for future work may yield more realistic results.

Furthermore, the baseline model also can be trained many times to see whether it will produce the similar learning performance. This will ensure a strong foundation for baseline and more satisfactory comparison.

References

Alishahi, Afra. 2010. Computational modeling of human language acquisition. *Synthesis Lectures on Human Language Technologies*, 3(1):1–107.

Berglund, Eva, Marten Erikson, and Monica Westerlund. 2005. Communicative skills in relation to gender, birth order, childcare and socioeconomic status in 18-month-old children. *Scandinavian Journal of Psychology*, 46(6):485–491.

Blake, Judith. 1981. Family size and the quality of children. *Demography*, 18(4):421–442.

Bornstein, Marc H, Diane B Leach, and O Maurice Haynes. 2004. Vocabulary competence in first-and secondborn siblings of the same chronological age. *Journal of Child Language*, 31(4):855–873.

ten Bosch, Louis, Lou Boves, Hugo Van Hamme, and Roger K Moore. 2009. A computational model of language acquisition: the emergence of words. *Fundamenta Informaticae*, 90(3):229–249.

Buttery, Paula J. 2006. Computational models for first language acquisition. Technical report, University of Cambridge, Computer Laboratory.

Downey, Douglas B. 2001. Number of siblings and intellectual development: The resource dilution explanation. *American psychologist*, 56(6-7):497.

Ellis, Nick C. 1998. Emergentism, connectionism and language learning. *Language learning*, 48(4):631–664.

Floor, Penelope and Nameera Akhtar. 2006. Can 18-month-old infants learn words by listening in on conversations? *Infancy*, 9(3):327–339.

Gasser, Michael. 1990. Connectionism and universals of second language acquisition. *Studies in Second language acquisition*, 12(2):179–199.

Hoff-Ginsberg, Erika. 1998. The relation of birth order and socioeconomic status to children's language experience and language development. *Applied Psycholinguistics*, 19(4):603–629.

Hoff-Ginsberg, Erika and Wendy M Krueger. 1991. Older siblings as conversational partners. *Merrill-Palmer Quarterly (1982-)*, pages 465–481.

Ingram, David. 1989. *First language acquisition: Method, description and explanation*. Cambridge university press.

Jones, Celeste Pappas and Lauren B Adamson. 1987. Language use in mother-child and mother-child-sibling interactions. *Child Development*, pages 356–366.

Kaplan, Frederic, Pierre-Yves Oudeyer, and Benjamin Bergen. 2008. Computational models in the debate over language learnability. *Infant and Child Development: An International Journal of Research and Practice*, 17(1):55–80.

Keijser, Daan. 2018. Curious topics: A curiosity-based model of first language word learning. Master's thesis, Tilburg University, the Netherlands.

Kuhl, Patricia K, Jean E Andruski, Inna A Chistovich, Ludmilla A Chistovich, Elena V Kozhevnikova, Viktoria L Ryskina, Elvira I Stolyarova, Ulla Sundberg, and Francisco Lacerda. 1997. Cross-language analysis of phonetic units in language addressed to infants. *Science*, 277(5326):684–686.

Lust, Barbara C. 2006. *Child language: Acquisition and growth*. Cambridge University Press.

Mannle, Sara, Michelle Barton, and Michael Tomasello. 1992. Two-year-olds' conversations with their mothers and preschool-aged siblings. *First language*, 12(34):57–71.

McLeod, Peter, Kim Plunkett, and Edmund T Rolls. 1998. *Introduction to connectionist modelling of cognitive processes*. Oxford University Press.

Murphy, Judith and Neil Slorach. 1983. The language development of pre-preschool hearing children of deaf parents. *International Journal of Language & Communication Disorders*, 18(2):118–127.

Narafshan, Mehry Haddad, Firooz Sadighi, Mohammad Sadegh Bagheri, and Nasrin Shokrpour. 2014. The role of input in first language acquisition. *International Journal of Applied Linguistics and English Literature*, 3(1):86–91.

Oshima-Takane, Y and J Derevensky. 1990. Do later-born children delay in early language development. In *Poster presented at the International Conference on Infant Studies, Montreal, Canada*.

Oshima-Takane, Yuriko and Medina Robbins. 2003. Linguistic environment of secondborn children. *First Language*, 23(1):21–40.

Paszke, Adam, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Pine, Julian M. 1995. Variation in vocabulary development as a function of birth order. *Child Development*, 66(1):272–281.

Plummer, Bryan A, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Raikes, Helen, Barbara Alexander Pan, Gayle Luze, Catherine S Tamis-LeMonda, Jeanne Brooks-Gunn, Jill Constantine, Louisa Banks Tarullo, H Abigail Raikes, and Eileen T Rodriguez. 2006. Mother–child bookreading in low-income families: Correlates and outcomes during the first three years of life. *Child development*, 77(4):924–953.

Rohrbach, Anna, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834, Springer.

Roy, Deb K and Alex P Pentland. 2002. Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1):113–146.

Rumelhart, David E and James L McClelland. 1986. On learning the past tenses of english verbs.

Shneidman, Laura A, Michelle E Arroyo, Susan C Levine, and Susan Goldin-Meadow. 2013. What counts as effective input for word learning? *Journal of Child Language*, 40(3):672–686.

Simonyan, Karen and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tasnimi, Mahshad. 2015. Connectionism: The pros and cons. *International Journal For Research In Educational Studies (ISSN: 2208-2115)*, 1(1):22–38.

Tomasello, Michael and Sara Mannle. 1985. Pragmatics of sibling speech to one-year-olds. *Child Development*, pages 911–917.

Waring, R. 1996. Connectionism and second language vocabulary. *Temple University Japan Occasional papers*.

Young, Peter, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Zajonc, Robert B and Gregory B Markus. 1975. Birth order and intellectual development. *Psychological review*, 82(1):74.

Zhao, Xiaowei and Ping Li. 2005. A self-organizing connectionist model of early word production. In *Proceedings of the twenty-seventh annual conference of the cognitive science society*, pages 2434–2439.

Appendix A: Dyadic Scenario Results

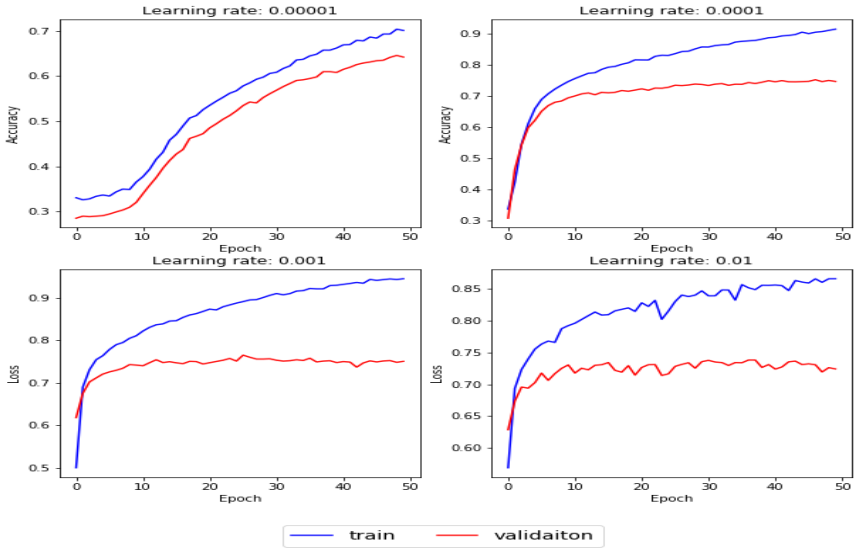

Figure 1
Accuracy results of the listener on train and validation data with different learning rates and 40 batches

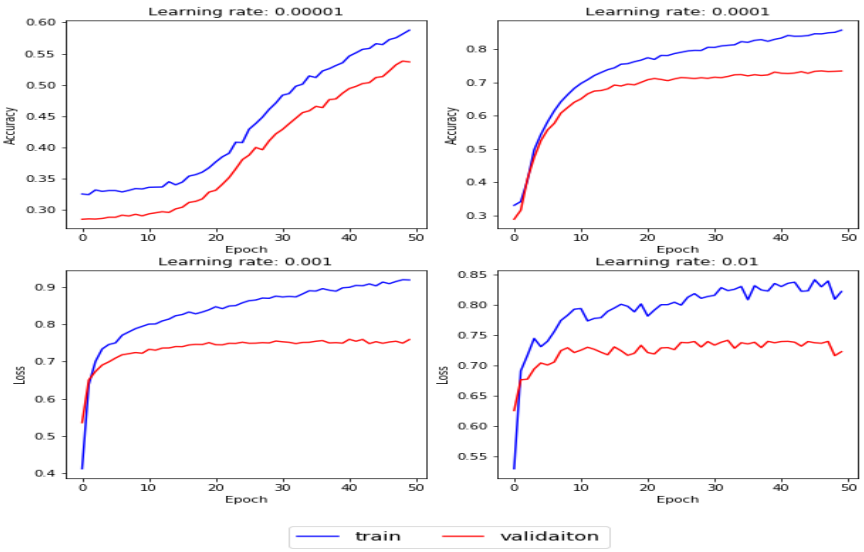

Figure 2
Accuracy results of the listener on train and validation data with different learning rates and 100 batches

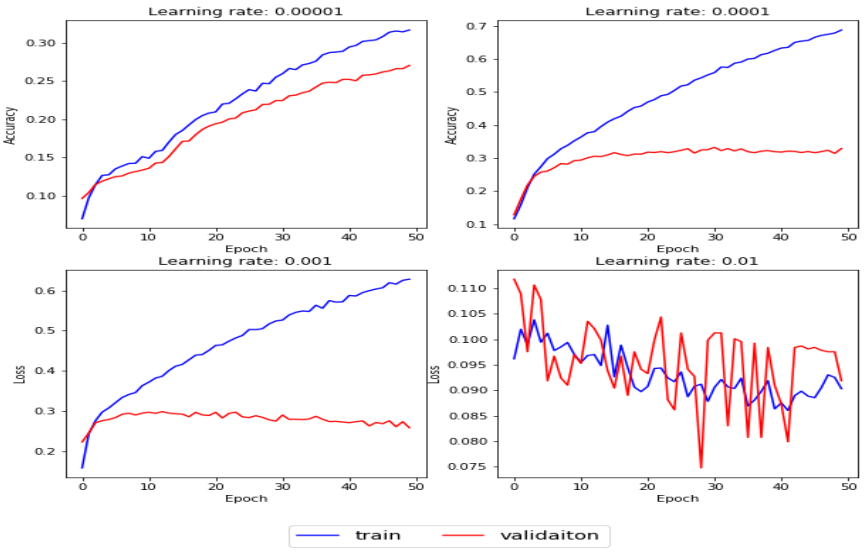

Figure 3
Accuracy results of the speaker on train and validation data with different learning rates and 40 batches

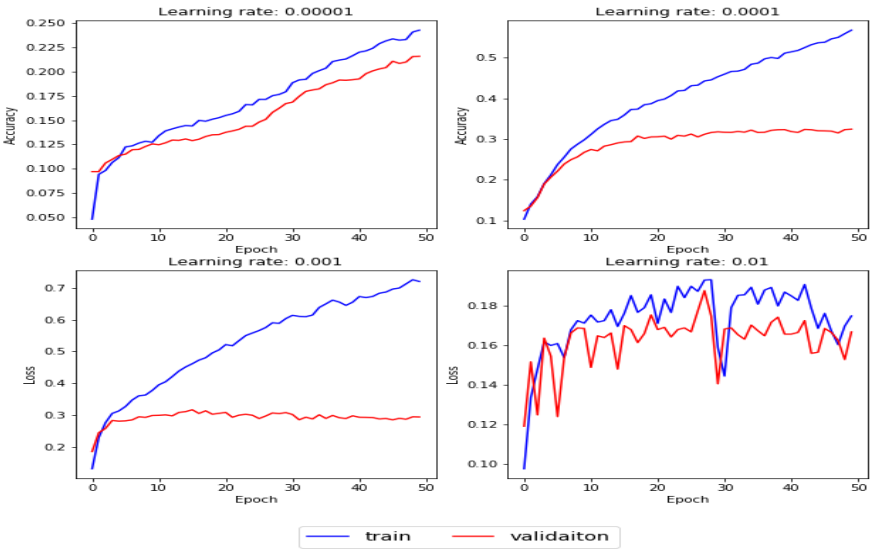

Figure 4
Accuracy results of the speaker on train and validation data with different learning rates and 100 batches

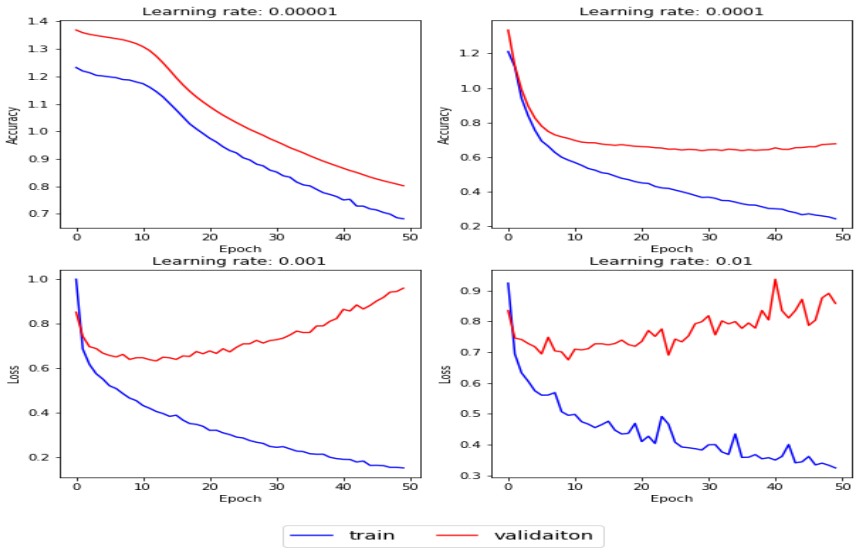

Figure 5
Loss results of the listener on train and validation data with different learning rates and 40 batches

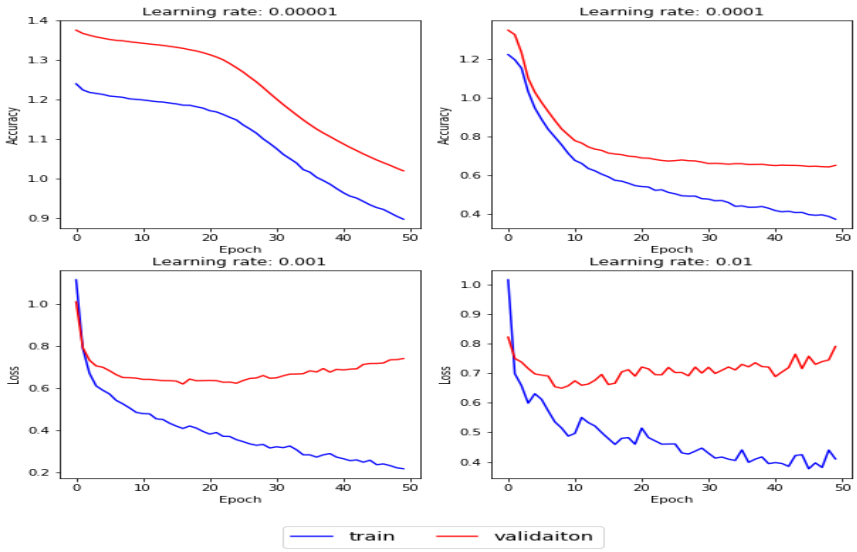

Figure 6
Loss results of the listener on train and validation data with different learning rates and 100 batches

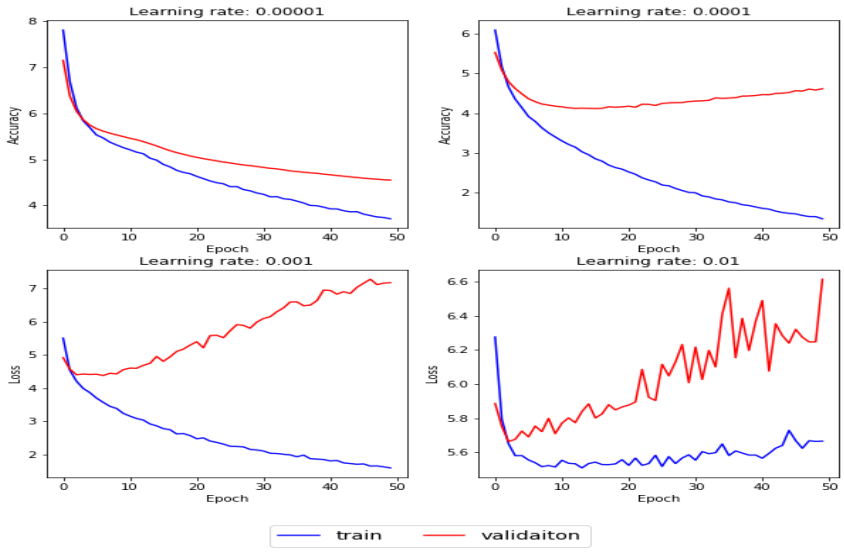

Figure 7
Loss results of the speaker on train and validation data with different learning rates and 40 batches

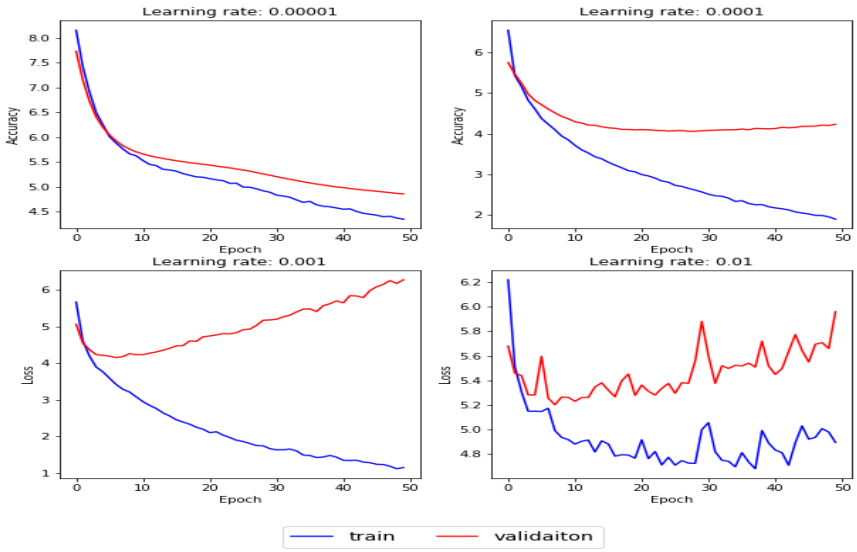

Figure 8
Loss results of the speaker on train and validation data with different learning rates and 100 batches