Lexical emergence from context : exploring unsupervised learning approaches on large multimodal language corpora

William N Havard

Université Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : Sciences du langage Spécialité Informatique et Sciences du Langage

Arrêté ministérial : 25 Mai 2016

Présentée par

William N. HAVARD

Thèse codirigée par **Laurent BESACIER**, Professeur, Université Grenoble Alpes
et codirigée par **Jean-Pierre CHEVROT**, Professeur, Université Grenoble Alpes

préparée au sein du **Laboratoire d'Informatique de Grenoble** et du **Laboratoire de Linguistique et Didactique des Langues Étrangères et Maternelles**
dans **l'École Doctorale Langues, Littératures et Sciences Humaines**

L'émergence du lexique en contexte : apport des méthodes non supervisées sur grands corpus de données multimodales

Lexical emergence from context: exploring unsupervised learning approaches on large multimodal language corpora

Thèse soutenue publiquement le **5 juillet 2021**,
devant le jury composé de :

M. Jean-Luc SCHWARTZ Directeur de Recherche, GIPSA-lab, Université Grenoble Alpes	Président
Mme Odette SCHARENBORG Associate Professor, Technische Universiteit Delft	Rapportrice
M. Laurent PRÉVOT Professeur, LPL, Aix-Marseille Université	Rapporteur
M. Grzegorz CHRUPAŁA Associate Professor, Tilburg University	Examinateur
M. Laurent BESACIER Professeur, LIG, Université Grenoble Alpes	Directeur de thèse
M. Jean-Pierre CHEVROT Professeur, LIDILEM, Université Grenoble Alpes	Directeur de thèse

"La lingua maternal: asi se dize de lo ke se entendya enkaza, ma, en este kavzo, Antonio, la madre no se muere nunka. Siempre se keda fuerte. Puedes azer el mas gran viage, kuando retornas la topas bien en pies. En eya vive tu pasado, en eya te sientes presente a ti mismo. Las palavras son tu verdadero lougar y tu esperanza."

"La langue maternelle : ainsi désigne-t-on ce que l'on entendait à la maison, mais cette mère meurt-elle jamais ? En elle veille notre passé, en elle nous sommes tout à fait présent à nous-mêmes. Et, si les mots sont notre vraie demeure, comment ne seraient-ils aussi une bonne part de notre devenir ?"

"The mother tongue: that's what we called what we spoke at home. Will this mother ever die, Antonio? In her, our past grows old; in her, we are completely present to ourselves. And, if words are our true domain, how could they not also be part of our future?"

Marcel Cohen, *Letra a Antonio Saura*, 1997.
English Translation: Raphael Rubinstein

Résumé

Ces dernières années, les méthodes d'apprentissage profond ont permis de créer des modèles neuronaux capables de traiter plusieurs modalités à la fois. Les modèles neuronaux de traitement de la **P**arole **V**isuellement **C**ontextualisée (PVC) sont des modèles de ce type, capables de traiter conjointement une entrée vocale et une entrée visuelle correspondante. Ils sont couramment utilisés pour résoudre une tâche de recherche d'image à partir d'une requête vocale : c'est-à-dire qu'à partir d'une description orale, ils sont entraînés à retrouver l'image correspondant à la description orale passée en entrée. Ces modèles ont suscité l'intérêt des linguistes et des chercheurs en sciences cognitives car ils sont capables de modéliser des interactions complexes entre deux modalités — la parole et la vision — et peuvent être utilisés pour simuler l'acquisition du langage chez l'enfant, et plus particulièrement l'acquisition lexicale.

Dans cette thèse, nous étudions un modèle récurrent de PVC et analysons les connaissances linguistiques que de tels modèles sont capables d'inférer comme sous-produit de la tâche principale pour laquelle ils sont entraînés. Nous introduisons un nouveau jeu de données qui convient à l'entraînement des modèles de PVC. Contrairement à la plupart des jeux de données qui sont en anglais, ce jeu de données est en japonais, ce qui permet d'étudier l'impact de la langue d'entrée sur les représentations apprises par les modèles neuronaux.

Nous nous concentrons ensuite sur l'analyse des mécanismes d'attention de deux modèles de PVC, l'un entrainé sur le jeu de données en anglais, l'autre sur le jeu de données en japonais, et montrons que les modèles ont développé un comportement général, valable quelle que soit la langue utilisée, en utilisant leur poids d'attention pour se focaliser sur des noms spécifiques dans la chaîne parlée. Nos expériences révèlent que ces modèles sont également capables d'adopter un comportement spécifique à la langue en prenant en compte les particularités de la langue d'entrée afin de mieux résoudre la tâche qui leur est donnée.

Nous étudions ensuite si les modèles de PVC sont capables d'associer des mots isolés à leurs référents visuels. Cela nous permet d'examiner si le modèle a implicitement segmenté l'entrée parlée en sous-unités. Nous étudions ensuite comment les mots isolés sont stockés dans les poids des réseaux en empruntant une méthodologie issue de la linguistique, le paradigme de *gating*, et nous montrons que la partie initiale du mot joue un rôle majeur pour une activation réussie.

Enfin, nous présentons une méthode simple pour introduire des informations sur les frontières des segments dans un modèle neuronal de traitement de la parole. Cela nous permet de tester si la segmentation implicite qui a lieu dans le réseau est aussi efficace qu'une segmentation explicite. Nous étudions plusieurs types de frontières, allant des frontières de phones aux frontières de mots, et nous montrons que ces dernières donnent les meilleurs résultats. Nous observons que donner au réseau plusieurs frontières en même temps est bénéfique en permettant au réseau de prendre en compte la nature hiérarchique de l'entrée linguistique.

Mots clefs : acquisition du langage visuellement contextualisé, modèle de traitement de la parole visuellement contextualisée, acquisition lexicale, acquisition du langage.

Abstract

In the past few years, deep learning methods have allowed researchers to design neural models that are able to process several modalities at once. Neural models of Visually Grounded Speech (VGS) are such kind of models and are able to jointly process a spoken input and a matching visual input. They are commonly used to solve a speech-image retrieval task: given a spoken description, they are trained to retrieve the closest image that matches the description. Such models sparked interest in linguists and cognitive scientists as they are able to model complex interactions between two modalities — speech and vision — and can be used to simulate child language acquisition and, more specifically, lexical acquisition.

In this thesis, we study a recurrent-based model of VGS and analyse the linguistic knowledge such models are able to derive as a by-product of the main task they are trained to solve. We introduce a novel data set that is suitable to train models of visually grounded speech. Contrary to most data sets that are in English, this data set is in Japanese and allows us to study the impact of the input language on the representations learnt by the neural models.

We then focus on the analysis of the attention mechanisms of two VGS models, one trained on the English data set, the other on the Japanese data set, and show the models have developed a language-general behaviour by using their attention weights to focus on specific nouns in the spoken input. Our experiments reveal that such models are able to adopt a language-specific behaviour by taking into account particularities of the input language so as to better solve the task they are given.

We then study if VGS models are able to map isolated words to their visual referents. This allows us to investigate if the model has implicitly segmented the spoken input into sub-units. We further investigate how isolated words are stored in the weights of the network by borrowing a methodology stemming from psycholinguistics, the gating paradigm, and show that word onset plays a major role in successful activation.

Finally, we introduce a simple method to introduce segment boundary information in a neural model of speech processing. This allows us to test if the implicit segmentation that takes place in the network is as effective as an explicit segmentation. We investigate several types of boundaries, ranging from phone to word boundaries, and show the latter yield the best results. We observe that giving the network several boundaries at the same time is beneficial. This allows the network to take into account the hierarchical nature of the linguistic input.

Keywords: grounded language learning, visually grounded speech model, lexical acquisition, language acquisition.

Contents

Introduction

Context and Motivation

In a relatively short period of time, children are able to acquire their native language allowing them to understand it and speak it effortlessly. This is an amazing feat, as they are able to do so without much supervision. They indeed do not need adults around them to explicitly teach them new words or correct them when they speak. Instead, the perceptible surrounding context seems to provide them with all the necessary information they need in order to acquire their native language.

According to Landau & Gleitman (1985, p. 1), "[n]o disagreement arises about the *necessity* for extralinguistic experience. Their [the theorists'] disagreements have to do with the *sufficiency* of experience for learning a language". Indeed, while some language acquisition theories (Chomsky 1969, Pinker 2009) postulate strong innate knowledge where the environment only plays a minor role, other (Skinner 1957, Tomasello 2009) postulate a much larger influence of the environment. However, all these theories agree — to some extent — on the fact that a certain amount of extralinguistic experience is necessary. Indeed, even in the strictest inneist theories, where the child would be born with a knowledge of grammatical and conceptual entities, a mapping still has to operate between the abstract entities of the human mind and their realisation in the physicial world, which is only perceptively accessible.

Landau & Gleitman (1985, p. 7) state that "the child's input consists of sound/situation pairs, but his final output is a set of form/meaning pairs, appropriate to an infinite set of novel but well-circumscribed situations". The question is *how* do children transition from *sound/situation* pairs to *form/meaning* pairs? One step children have to go through is to understand that the sounds they perceive constitute conventional signs used for communication purposes, and not random sounds. This step might not be the first step children go through, but it is a critical step to go from *sound* to *form*. Transitioning from sound/situation to form/meaning pairs also implies a segmentation step. The first segmentation step we could think of is the segmentation of the speech stream into forms. Indeed, the speech stream contains a sequence of forms, which are however connected and not neatly separated from one another. Children thus have to learn how to segment the speech stream appropriately so as to discover the conventional word forms used in their native language. The second segmentation step we could think of consists in analysing the environment so as to identify and extract the stakeholders: who is talking, to whom, about what, etc. While the first step involves identifying patterns in the spoken modality, the other step involves identifying patterns in other modalities (visual, haptic, etc.). We presented these two steps as separate or as if they occurred sequentially but it is in fact not the case, as they do occur simultaneously. Moreover, these two steps seem to be fueled by similar routines in the child's mind (e.g. statistical sensitivity, see Kirkham et al. 2002).

Once forms have been extracted from the speech stream, and referents extracted from the environment, the child has to learn to map both. At this step, the child has transitioned from *sound/situation* to *form/referent* pairs. The associations the child makes might be coarse at first, where a word-form is associated to as specific referent (e.g. dog for the house pet), but the child is ultimately able to abstract the commonalities between different referents of the same word-form and derive a meaning. This involves a certain amount of conceptualisation, and ultimately allows the child to build a set of *form/meaning* pairs. *Lexical acquisition* refers to one of the processes that results from the transition from sound/situation pairs to

a set of form/meaning pairs (other processes include the acquisition of syntax and syntactic frames, the acquisition of phonology, etc.).

The form/meaning pairs the child builds should be stored in such a way they can be easily accessed so as to parse the speech stream and to build its own utterances. Hence, the stored word forms should be specific enough so as to *not* allow for similar sounding word-forms to be recognised, while not being overly specific so as to accomodate for variation and mispronounciations. This step is commonly referred to as *word recognition*.

In this thesis, we propose to study an neural model of visually grounded speech. Visually grounded speech models are models that are trained to solve a speech/image retrieval task. That is, given a spoken description of an image, they should find the closest matching image among a collection of images (or vice-versa, find a spoken description given an image). To do so, such networks have to learn how to appropriately transform the input image and the input spoken description so that it is easy to find one given the other. The task such networks have to solve is thus very close to that of a child acquiring her native language. Indeed, such networks are presented with *sound/situation* pairs (i.e. a spoken description and its matching image), and in order to find the correct matching image, we hypothetise that the network should learn to transform this pair into a *form/meaning* pair. Indeed, in order to find the matching image among a collection of images given a spoken utterance, it should somehow segment the speech stream into sub-units, so that the resulting sub-units refer to objects in the image. The representation of the extracted units should be specific enough so that when prompted with these units, the network only retrieves images featuring instances that the spoken unit refers to. Similarly to humans, on the image side, the network should be able to abstract the referent, so that the target image can be retrieved, even if the object is presented in a non-canonical manner or among a cluttered environment. Consequently, the task neural models of visually grounded speech are trained to solve is a task which is very close to the one of children learning their mother tongue, and more specifically is very close to the task of lexical acquisition. They have to go through the same steps than a child, which are: *segmentation*, *mapping*, and *recognition*.

Contributions

In this thesis, we study a recurrent-based model of visually grounded speech. As such models solve a task similar to that of children, who discover the set form/meaning pairs of their native language, our analyses focus on understanding if neural models do so in the same way.

In this manuscript, we present the following contributions:

(i) We introduce a spoken extension of an image captioning data set that is suitable to train visually grounded speech models. Unlike most data sets of visually grounded speech which are in English, this data set is in Japanese which allows us to study in a contrastive approach the impact of the input language (English or Japanese) on the representations learnt by the model.

(ii) We study the patterns of the attention weights of the attention mechanisms of our models, so as to understand what parts of the speech signal are highlighted, and to what extent they differ from what randomness would predict. We also propose a longitudinal investigation, where we study how these attention weights evolve during the training phase.

(iii) We introduce a methodology stemming from the psycholiguistic litterature, the gating paradigm (Grosjean 1980), that allows us to easily investigate spoken word activation

and competition in our model. To the best of our knowledge, this is the first time such methodology is used to investigate the representations learnt by neural models of speech processing.

(iv) We propose a method to simply introduce prior linguistic information in the form of segment boundaries (phone, syllable, or word boundaries) in a neural model of speech processing. The method we propose allows to integrate several types of boundaries, at different levels of the neural architecture, which allows to take into account the hierarchical nature of the spoken input. This allows us to study if segmenting the speech signal into sub-units allows the models to learn to better map images to their spoken descriptions.

Thesis Outline

This thesis is divided into two parts: **Part I: Background** consists of two background chapters and **Part II: Contributions** consists of four chapters presenting our contributions. The chapters of this manuscript are organised as follows:

Chapter 1: Background on Child Lexical Acquisition: This chapter is structured into three parts. In the first part, we present the strategies infants and children use in order to segment the speech stream into sub-units. In the second part, we review how children are able to map the segmented units they have extracted from the speech stream to their referent, so as to ultimately acquire their meaning. Finally, in the third part, we present several psycholinguistic models of word activation and recognition. We argue that language acquisition, and more specifically lexical acquisition, is only possible because of the multi-modal nature of language, and in particular, we show how the visual modality particularly helps children in this task. This review allows us to understand how lexical acquisition takes place in humans, and consequently, allows us to formulate hypothesis on how the neural model we study carries out its task.

Chapter 2: Background on Speech Processing and Term Discovery: This chapter presents the notion of *unsupervised speech processing* as well as the main approaches to *unsupervised speech segmentation* and *term discovery*. We argue that in order to be most successful, unsupervised approaches to speech segmentation and term discovery should be grounded to another source of information. We then present the basics of machine learning and artificial neural networks, and then review the existing models of visually grounded speech. We end this chapter by examining the difference between *simulation* and *modelling* and by discussing to what extent visually grounded models are indeed grounded.

Chapter 3: Visually Grounded Speech Architectures and Data: In this chapter, we present the data sets we use in this thesis. We also introduce a new data set, the Synthetically Spoken STAIR data set, that we created and used in our experiments. We then present the neural architecture we use and detail the assumptions we make by using such architecture as well as by training it using data sets initially thought for computer vision purposes.

Chapter 4: Attention in a Model of Visually Grounded Speech: In this chapter, we analyse the attention weights of the attention mechanisms of visually grounded speech models trained on three different data sets: two featuring synthetic speech, and one

featuring real human speech. Our analysis mainly focuses on which parts of speech signals are highlighted by the models, and we make connections with child language acquisition. As our models are trained on two languages, English and Japanese, this enables us to observe which behaviours are language-general and which behaviours are language-specific. Finally, we analyse how the attention weights evolve during training.

Chapter 5: Word Activation, Competition, and Recognition: The experiments conducted in this chapter are directly inspired by prior psycholinguistic experiments on word recognition in humans. Using the gating paradigm (Grosjean 1980), we analyse the neural representations of a visually grounded speech model trained on an English data set. More specifically, we investigate if such models are able to recognise isolated words and explore how isolated word activation takes place. In the last part of this chapter, we investigate if word activation and recognition takes place through a process of competition, such as what has been postulated for human word recognition.

Chapter 6: Impact of Prior Linguistic Information: In this final chapter, we investigate if neural models of visually grounded speech have better performances if, instead of being given full utterances, they are given pre-segmented utterances. We investigate if this is the case with different levels of segmentation as well as with a random segmentation. We explore at which layers of the architecture such information should be introduced in order to be most effective. In the last part of this chapter, we explore if using a hierarchical model, where several levels of segmentation are simultaneously given to the model, improve the results, and which combination of segmentation levels proves the most effective.

Conclusion: In this last chapter we summarise the contributions of this thesis, and suggest several future works that could be undertaken.

Part I

Background

Chapter 1

Background on Child Lexical Acquisition

Contents

1.1 Introduction

Language acquisition refers to the process by which a child learns her native language. In order to acquire her native language, one step the child has to perform is to build a mental lexicon (Emmorey & Fromkin 1988). A mental lexicon is a set of phonological word-forms associated to a meaning. It also contains information about the morphology of each form (e.g. plural formation, conjugations, etc.), syntactic information (e.g. part of speech, valency, expected semantic roles), as well as semantic and pragmatic information (e.g. connotation). This mental lexicon is then used by the child to parse the speech stream and understand what is said; or later on, to build her own utterances.

However, building this mental lexicon is far from easy. Indeed, the speech stream does not contain neatly separated words that the child would only have to assign a meaning to. The child first has to segment the speech stream into sub-units and ultimately pair a meaning to each of the resulting sub-units. These two processes may not necessarily occur simultaneously, and for a time being, the mental lexicon can contain word-forms that were not assigned any meaning yet (see Jusczyk & Aslin 1995). The meaning of each form itself is not a ready-to-use object, but has to be inferred and constructed by the child using perceptual information. Finally, the mental lexicon has to be structured in such a way it can be used by the child, notably when parsing the speech stream, in order to activate the meaning of the recognised words so as to interpret what is being said.

These basic processing steps to speech processing and comprehension that children should acquire in order to be able to parse the speech stream correctly — and ultimately acquire their native language — are identified by Di Cristo (2013, pp. 51-52) as *segmentation*, *recognition* and *interpretation*. He also adds another step, which takes place before the segmentation step, which is *decoding* and which is concerned with "extracting the speech signal from its acoustic environment and convert some of the information it contains into linguistic representations".

In this chapter, we will review the strategies that children use in order to parse and segment the spoken input into sub-units (Section 1.2). We will then present several models which account for how humans retrieve words from the mental lexicon (Secion 1.4). Finally, we will show the tools used by the child in order to infer the meaning of a given word, particularly focussing on how vision provides the essential tools to do so (Section 1.3).

1.2 Speech Segmentation

Speech segmentation is generally regarded as a typical example of chicken-and-egg problem. Indeed, in order to correctly parse the speech stream, one should know what a word is, but in order to know what a word is, one should already be able to parse the spoken input. Nonetheless, children are able to solve this riddle, as they all manage to learn their native language at some point. In a relatively short period of time, children are able to identify words in the speech stream without much supervision. In the following sections, we will review some of the cues children use in order to discover words and segment the speech stream.

1.2.1 Speech perception

Language acquisition does not start from the moment the child is born, but rather as soon as the foetus is able to hear, that is, when its body develops a functional proto-auditory system at about 19 weeks of gestational age (Hepper & Shahidullah 1994). Perception of speech evolves as the foetus grows and its auditory systems matures, enabling the brain to

tune to the language environment prior to birth (May et al. 2011). As soon as 35 weeks of gestational age, the foetus is already "capable of discriminating different sounds" (Shahidullah & Hepper 1994). May et al. (2011) showed that neonates' brain responses were different for familiar languages and unfamiliar languages, effectively showing the foetus has learnt a robust enough representation of the language it was exposed to prior to birth so as to recognise it once born. DeCasper & Spence (1986) also demonstrated that newborns are able to recognise samples of stories that were read out loud by their mother during pregnancy. This lets the authors suspect that "foetuses had learned and remembered something about the acoustic cues which specified their particular target passage". These experiments therefore suggest that the exposure to language foetuses have is far from trivial and that they are already accustomed to their future mother tongue.

This pre-natal exposure also develops biases the child will be able to use once born. Hepper & Shahidullah (1994) showed that foetuses are more sensitive to low frequencies (500 Hz and lower) than to higher frequencies. Thus, human voices "form a salient auditory stimulus" as "the fundamental frequency of the human voice is around 225 Hz for females and 128 Hz for males". Newborns indeed attend more to human voices than to other acoustic stimuli (Vouloumanos & Werker 2007).

Hence, as reported by DeCasper & Spence (1986), "newborns do not act like passive and neutral listeners". The newborn child should not be considered as a *tabula rasa*, but is on the contrary already attuned to its mother tongue. These biases will help the child segment the speech signal into sub-units.

1.2.2 Suprasegmental Cues

Suprasegmental cues are cues that are present in the suprasegmental features of speech and that are used by children to discover word boundaries. Suprasegmental features are modulations of the speech signal that may span over more than one segment (i.e. phone or syllable). These suprasegmental features, grouped under the umbrella term or prosody, include rhythm (tempo and pauses), intonation, and stress. As prosodic features are mainly linked to modulation of the fundamental frequency (F0) and that children are particularly sensitive to this frequency, they form salient cues the child can use to find words. Some authors even state that "at the beginning was prosody" (Di Cristo 2013, p. X). Indeed, prosody assumes many functional roles, the main one that may be of use for the infant being its demarcation function, that groups the spoken units together and that children can use to infer word boundaries.

The prosodic bootstrapping hypothesis — initially proposed by Lila R. Gleitman et al. 1982 — states that language acquisition is bootstrapped by the prosodic features the child is able to perceive, both from a lexical point of view than from a syntaxtical point of view (see Jusczyk 2000, p. 38 for a review). An evidence that prosodic features are used by children to segment the speech stream was made by Cristia & Seidl (2011). Their study shows that 6-month-old English-speaking infants' prosodic sensitivity[1] is a good predictor for their productive vocabulary at 24 months. Hence, if children indeed use prosodic cues to segment the speech stream, it appears natural that those that have the greatest sensitivity to prosodical information manage to segment more words than the others. We will see in the next section more precisely which prosodic cues children use in order to bootstrap language acquisition.

[1] As measured, among other measures, by their preference of well-formed prosodic units over ill-formed units.

1.2.2.1 Rhythm

Languages may be divided into three rythmic classes depending on the unit of isochrony used: *stress* in stress-timed languages (e.g. English, Dutch) where the interval between two stressed syllables is equal, *mora*[2] in mora-timed languages (e.g. Japanese, Tamil) where the duration of each mora is equal, and *syllables* in syllable-timed languages (e.g. French, Spanish) where the duration of each syllable is equal. The rythmic segmentation bootstrapping hypothesis (Nazzi & Ramus 2003, Nazzi 2008) postulates that rhythm plays a major role in language acquisition, and that the segmentation units used by children to parse the speech stream depends on the rhythmic class their language belongs to.

A good indication that newborns are indeed sensitive to rhythmic information was made by Nazzi et al. (1998). They showed that children are able to discriminate two languages (English and Japanese) that belong to two different rhythmic classes but failed to distinguish two languages if they belong to the same rhythmic class (such as English and Dutch, or Italian and Spanish). Infants are however able to discriminate two languages belonging to the same rhythmical class if one of them is their native language (Nazzi et al. 2000, Bosch & Sebastián-Gallés 1997). However, when the prosodic information is manipulated — either by shuffling the words of a sentence (Dehaene-Lambertz & Houston 1998), or by manipulating the pitch (Chong et al. 2018) — children fail to discriminate two languages, even if one of them is their native language. These studies thus effectively show that infants are sensitive to the rythmic properties of their native language, which they can use to discover word boundaries.

Cutler & Norris (1988) showed that English native adults tend to use rythmic information such as stressed syllables as evidence for a probable word boundary.[3] Jusczyk, Cutler & Redanz (1993) investigated if English-speaking infants had a preference for trochaic words — i.e. words that have a Strong-Weak (SW) stress pattern such as "GARden" — and showed it was the case, hinting that they could use this information to segment the speech signal, such as what was observed for adults. Yet, this preference was only found for 9-month-old infants but not for 6-month-old infants suggesting such preference might only be a result of the linguistic environment, and hence takes time to develop. In a later study, Jusczyk, Houston & Newsome (1999) tested if 7.5-month-old children use trochaic patterns to segment speech and "treat strong syllables as markers of word onsets". To do so, in a familiarisation phase, they embedded words with a Weak-Strong (WS) pattern in a sentence (such as "[...] guiTAR is [...]" or "[...] deVIce to [...]") and later tested if children showed a preference for SW or WS words: "guiTAR" vs. "TARis" and "deVIce" vs. "VIceto". Their results indeed show that 7.5-month-old children prefer "TARis" and "VIceto". This demonstrates, first, that 7.5-month-old infants are sensitive to SW patterns contrary to 6-month-old infants, and second, that they segment the input so that the resulting words have a SW pattern. Hence, 7.5-month-old English-speaking infants use stressed syllables as evidence for word boundaries, such as what was observed for adults.

Similarly to English, Dutch is a language which has more SW-patterned words than WS words. Houston et al. (2000) showed that 9-month-old Dutch children are able to

[2] A mora is an intermediary unit between the phone and the syllable that determines the syllable's weight. They might be a perfect overlap between the number of moræ and the number of syllables as in 建物 ("building") → たてもの /ta.te.mo.no/ with 4 moræ and 4 syllables; but it is most often not the case such as in 漢字 ("kanji") → かんじ with 2 syllables (/kan.ʑi/) and 3 moræ (/ka-n-ʑi/) or 東京 ("Tokyo") → とうきょう with 2 syllables (/toː.kjoː/) and 4 moræ (/to-o-kjo-o/). The final ん of a syllable and the long vowel う count as single moræ.

[3] They do so by showing that their test subjects take longer to notice the presence of the word "mint" in MIN-TAYVE (Strong-Strong (SS) pattern) than in MIN-tesh (Strong-Weak (SW) pattern), as in the former, detecting "mint" supposes to reassemble word parts that the listener considered to be two different words.

recognise — and hence segment — words with a SW pattern from fluent speech after having been familiarised to individual words, such as what was observed for English-speaking infants. The same study also shows that English-speaking infants are equally successful in segmenting SW Dutch words, suggesting they apply the SW segmenting strategy to segment other languages than their native language.

Jusczyk, Houston & Newsome (1999) revealed that, at 10 months old, children are able to segment both SW and WS words correctly — words such as "guiTAR" and "deVIce" — showing that even though children initally rely on SW patterns to segment the speech stream, they shift to other cues later on, enabling them to segment WS-patterned words. Thiessen & Saffran (2007) further showed that English-speaking infants can be trained to segment WS disyllabic words after an habituation phase, revealing the SW pattern only emerges as a result of the linguistic environment, which uses more SW words than WS words.

Nazzi et al. (2006) explored word segmentation in 8-month-old, 12-month-old, and 16-month-old French-speaking infants to see if infants with different languages develop the same abilities at the same age. To do so, infants were familiarised with individual units (e.g. *toucan*) and in the test phase the infants had the choice to listen to two different passages: one that contained the target word embedded in fluent speech (e.g. [...]*toucan*[...]) or another one that contained another disyllabic (e.g. *putois*, polecat) word also embedded in fluent speech ([...]*putois*[...]). If infants chose to listen more often to the passage containing the target word, it can be concluded they recognised, and hence segmented, the target word embedded in the passage. No evidence of segmentation was found in 8-month-old infants as they did not favour one passage over another, while they seem to segment bisyllabic words at 16 months, as they listened preferably to the *toucan* passage than the other passage. In between, at 12 months old, infants are able to segment the final syllable of word (*can* from *toucan*) only if they were familiarised with *can* beforehand, but not the whole word. They are able to recognise the initial syllable of a bisyllabic word (*tou* from *toucan*) if and only if it is an exact match to the word heard in the familiarisation phase (*tou* spliced from *toucan*), but are not able to recognise *toucan* as a whole. This let Nazzi et al. (2006) conclude that French-speaking infants display a systematic pattern of segmentation *later* than English- or Dutch-speaking infants. The authors also conclude French-speaking infants segment the speech stream using syllables as the basic processing unit. Nonetheless, they also seem to be sensitive to stress as they should otherwise have been able to segment *tou* without requiring an exact match. Such syllabic segmentation was also observed in children whose native language was Castilian and/or Catalan (Bosch et al. 2013) which also are syllable-timed languages.

As shown by Nazzi (2008) and Nazzi et al. (2006), it seems the basic computation unit used by children to segment the speech stream depends on the rhythmic class of their native language: English- or Dutch-speaking infants segment units based on their stress as their language is a stressed-time language, while French-speaking infants' segmentation is based on syllables and not on stress patterns as their language is a syllable-timed language. Hence, while English- or Dutch-speaking infants extract SW patterns, French-speaking infants extract syllable-sized units.

For Japanese, it is less clear what the basic segmentation unit used by children is. Indeed, as Japanese is a mora-time language, the rhythmic segmentation hypothesis would imply that Japanese-speaking infants start by segmenting mora-like units. Even if it is in fact what is observed for adults (Otake et al. 1993), recent works (Inagaki et al. 2000) suggest that pre-literate children segmentation is part mora-based, part syllable-based; but

as literacy increases, segmentation gradually becomes solely mora-based.[4]

1.2.2.2 Phonological Phrases & Utterance Boundaries

Prosody groups words together into units that are prosodically marked which are called *phonological phrases* (which correspond more-or-less to syntactic-based chunks Abney 1992). Soderstrom et al. (2005) showed that children prefer to listen to clauses that form a single prosodic unit (e.g. "Leafy vegetables taste so good") rather than to the same clause (in terms of the sequence of words) that straddles over two sentences (e.g. "They must buy **leafy vegetables. Taste so good** helps their families.") and which consequently does not constitute a phonological unit anymore. Hence, children are able to detect phonological phrase boundaries and might use this information to infer word boundaries.

Christophe et al. (2003) showed that word recognition in 13-month-old infants is indeed constrained by phonological phrase boundaries. Children effectively recognise a familiarised target word (e.g. paper) if it is part of phonological phrase (e.g. "[The college] [with the biggest **paper** forms] [is best]"), but not if it straddles over two phonologial phrases (e.g. "[The butler] [with the highest **pay**] [**per**forms the most]") thus indicating that 10-month-old infants know how to detect phonological phrase boundaries, and that a word may not straddle two phonological phrases. Prosodical phrase boundaries may be signalled by a pause as it is (theoretically) the only legal place where one can pause to catch one's breath (Di Cristo 2013, p. 15). However, most of the time, such boundaries are signalled by other acoustic cues, such as pre-boundary vowel lengthening (Vaissière 1983) which could be a universal cue to detect prosodical phrase boundaries, or other allophonic cues which will be presented in the following section. Therefore, by paying attention to phonological phrases — whose "primary purpose is to segment statements and discourse into groups of meanings" Di Cristo (2013, p. 65) — children are able to posit word boundaries.

Another prosodic cue infants use to segment the speech stream is to pay specific attention to utterances-edges,[5] a strategy called "Edge Hypothesis". In their study, Johnson et al. (2014) showed that children as young as 6 months old that hear a novel word placed either at the beginning and/or at the end of the familiarisation utterance recognised the target word when heard in isolation, but not if this word appeared sentence medially. This shows that infants consider utterance boundaries as evidence for word boundaries and are able to effectively use this information afterwards.

1.2.2.3 Conclusion on Suprasegmental Cues

All in all, by paying attention to prosodic features, the child is able to posit boundaries in the speech signal. According to Di Cristo (2013, p. 63) children's sensitivity to rythmic cues "[is] likely to favour the establishment of routines, which will help identify word boundaries". Similarly, he argues that even though the main function of prosody is not to give direct cues to word boundaries, it gives a "metric framework" that the listener can use in order to posit word boundaries, or at least, posit boundaries for large linguistic units such as chunks, which necessarily correspond to word boundaries.

1.2.3 Segmental Cues

Contrary to the previous section where we presented *supra*segmental cues, this section is concerned with *segmental* cues, that is, cues directly linked to the segments (i.e. phone[me]s)

[4] As a direct consequence of using the *kana* writing system which is mora-based, see footnote 2. For more details on this debate, see Kubozono (2015, §V.17.5.3)

[5] Which are generally prosodically marked by a falling intonation.

of a language. As Mattys et al. (2005), we group under the term of segmental cues both phonotatic cues and acoustico-phonetical cues.

1.2.3.1 Phonotactics

Phonotactics refers to the study of the sound sequences of a language. Some sound sequences are said to be *legal* (i.e. allowed) while some are *illegal* (i.e. not allowed) in a given language. What constitutes (or not) a legal sequence depends on the language, and sequences that are legal in one language might be illegal in another language. For example, English allows very complex sound sequences within one syllable (such as CCCVCCCC[6] as in "strangles" /stɹæŋglz/[7]) while such a sequence is not allowed in French. Moreover, some sequences might only be allowed at certain positions, but not at others (for example, /ŋ/ is only allowed word-finally in French). All these constraints provide the child with potential cues to infer word boundaries.

First, let us consider phonotactics at word level only. Jusczyk, Friederici, Wessels, Svenkerud & Jusczyk (1993) showed that 9-month-old children preferred listening to word lists whose words matched the phonotactic regularities of their native language than to a word list that violated these regularities, hence showing that infants are sensitive to the phonotactics of their language and intuitively know what sequences are legal. Jusczyk et al. (1994) further showed that within the same language, 9-month-old children prefer a word list that contains words with very frequent sound patterns over a list consisting of words with infrequent sound patterns.[8] Thus, by knowing which sound patterns are legal or not, the child can immediately know if the candidate sound sequence she has isolated may or may not constitute a word. If it does contain an illegal sequence, then it should not be regarded as a single sequence, but should be further segmented or discarded. This is in fact what was observed by MacKenzie et al. (2012): 12-month-old infants are able to associate English CVCV (e.g. *panu*), Japanese CVCV (e.g. *hashi*) sequences to novel objects. They are also able to do so with English CCVC non word (e.g. *plok*) but not Czech CCVC (e.g. *ptak*) sequences as the /pt/ sequence is illegal as a word onset in English, while /pl/ is not.

The phonotactic regularities the child uncovers are not restricted to adjacent units, but may also include non-adjacent units. Also, the regularities the child is able to discover is not only limited to legal/illegal patterns but also includes frequent/infrequent phonotactic patterns. For example, Nazzi et al. (2009) put forward a Labial-Coronal (LC) bias in speech perception in French infants: they prefer to listen to pseudo-words in which a labial consonant comes before a coronal consonant – LC words such as /**bu**d**e**/ – than the reverse pattern, that is, CL words such as /**de**b**y**/; as LC words are more frequent in French than CL words. This bias toward LC words has been shown to also constrain word learning: Gonzalez-Gomez et al. (2013) indeed showed that 14-month-old children are able to learn a mapping between a novel object and a word with a LC pattern (e.g. "bat"), but are unable to do so when the label of the word has a CL pattern (e.g. "tab"). It is only later on, at 16 months that they are able to learn both. Hence, not only children detect which phonotactic patterns are the most frequent, but the frequent patterns condition which words are learnt first.

Once children have enough information about the canonical shape of the individual words of their language, they can apply this information to segment the speech stream. Friederici & Wessels (1993) showed that 9-month-old Dutch-speaking infant could detect

[6] C = consonant sound, V = vowel

[7] Though /l/ tends to be realised syllabically [stɹæŋgl̩z] or [stɹæŋgəlz].

[8] Note that both lists featured words comprising very frequent sounds. Hence, the difference can only be explained by the frequency of the sound sequences and not the frequency of the individual sounds.

legal and illegal sound sequences in connected speech (i.e. they show a preference towards the passage that has legal phonotactic patterns) provided certain conditions are met.[9] Mattys et al. (1999) showed that English-speaking infants know what sequences are legal (or more probable) between words and within words. Infants presented with CVC·CVC[10] disyllabic words accented with a SW pattern prefer sequences where the internal C·C is more probable to appear within words than between words.

A real evidence that infants do use phonotactic patterns to segment speech into sub-units was made by Mattys & Jusczyk (2001*b*). In their experiment they show that 9-month-old children are able to recognise a CVC word spoken in isolation if it was previously heard within a sentence (i.e. [...]C·CVC·C[...]) so that the phonotactic patterns provide good indications as to where the edges of the words are (i.e. both initial and final C·C have consonants that have a low probability to co-occur within the same word). When the word is embedded in a sentence where the phonotactic patterns do not make this obvious (i.e. initial and final C·C have a high probability to occur within a word), infants fail to show preference for the target word. Children are also able to detect the target word even if it is only the onset or the offset of the word that displays obvious phonotactic patterns signalling a boundary. Hence, word boundaries need not be phonotactically signaled from both ends in order to be correctly segmented.

Consequently, the child is able to detect the phonotactic regularities of its languages and uses this information to segment the speech stream into words and evaluate if the segmented pattern constitutes a good word candidate or not.

1.2.3.2 Vowel Harmony

Vowel harmony is found in many languages (e.g. Turkic languages, Finno-Ugric languages, etc.) and places a hard constraint on the form of the words. It may be considered as a special type of non-adjacent phonotactic constraint that only affects vowels within one word. Vowel harmony consists in an assimilation by which all the vowels of a word share a common feature (either rounded/unrounded, front/back, etc.).[11] Therefore, if the child succeeds in discovering that its mother tongue has vowel harmony, it could use such cue to segment the spoken input. Ketrez (2013) indeed showed that when vowel harmony is broken, it most likely signals a word boundary, and suggest that children might use this property so infer word boundaries.

Mintz et al. (2018) showed in an artificial language learning (ALL) experiment that children indeed segment words based on vowel harmony. 6- to 7-month-old English-speaking infants where familiarised with the sequence *pidigitokobogetepedubuku* composed of four different words: *pidigi*, *tokobo*, *getepe*, and *dubuku*[12] which was repeated 45 times on a loop. Their results show that infants indeed segmented sequences corresponding to words (e.g. *toboko*) and not to part-words (e.g. *bukupi*, or *gitoko*). In a further experiment, they observe that infants can also segment words with a more complex harmony pattern. After hearing a sequence *ditepubobidetupo* consisting of four words with either front (*dite* or *bide*) or back harmony (*pubo* or *tupo*), children in the test phrase listen longer to words (e.g. *dite*) than to part-words (e.g. *detu*) showing the children extracted disyllabic pseudo-words based

[9] Most notably in their study, the spoken sample should be pronounced in a Child Directed fashion in order to observe this phenomenon.

[10] The symbol ·marks a syllable boundary.

[11] For example, in Turkish "Türkler" (Turkish men) where "ü" /y/ and "e" /ɛ/ are front vowels and "Fransızlar" (Frenchmen) where "a" /ɑ/ and "ı" /ɯ/ are back vowels.

[12] Note that this is a very restrictive (and maybe facilitative) case of vowel harmony as all the vowels in a given word are the same. Such vowel harmony, as far as we know, has never been reported in any of the world's languages.

on their vowel harmony pattern. This result is coherent with other research: Hohenberger et al. (2016) indeed showed that Turkish children as young as 6 months old prefer to listen to harmonic words over non-harmonic words. However, it seems that in their study children were only sensitive to the front/back harmony and not to the rounded/unrounded harmony.

All in all, these studies show that from a fairly early age, infants are sensitive to subtle differences such as vowel harmony and are able to use this information in order to segment the speech stream. More generally, it shows that infants are able to detect non-adjescent regularities "where X and Y are separated by intervening, unpredictable elements – such that listeners might be exposed to XAY, XBY, XCY – participants are able to learn that X predicts a following Y" (Thiessen & Erickson 2015); vowel harmony being only a special case of non-adjescent pattern. Newport & Aslin (2004), however, showed that children are able to keep track of non-adjacent regularities at phoneme level, but not a syllable level.

1.2.3.3 Allophones & Coarticulation

Allophones are sounds that correspond to different acoustic realisations of a given phoneme. Context usually governs the choice of one variant over another (i.e. in some cases the variant is systematically realised). For example, English /p/ may be realised [p] (after /s/ as in "s**p**ark"), [p^h] (word-initially as in "**p**ark"), or [p˺] (word-finally as in "po**p** music"). Hence, if the child has learnt the rules that govern the appearance of one variant over another, it can use this information to segment the speech stream (e.g. boundary before [p^h], boundary after [p˺], etc.).

A first indication that infants are sensitive to allophonic variations was made by Christophe et al. (1994). They showed that 3- or 4-day-old children are able to detect a difference between a CVCV disyllabic sequence spliced from a word (e.g. /mɑ̃ta/ in *sédi**menta**tion*) and the same CVCV sequence straddling over a boundary (e.g. /mɑ̃#ta/ in *déguise**ment ta**lentueux*). Therefore, they are able to differentiate two different realisations of the same phonological sequence. For English, Hohne & Jusczyk (1994) showed that while 2-month-old English infants are able to distinguish two sequences based on allophonic cues (e.g. "nitrate" [naɪt̚^{h}˳ɹeɪt˺] and "night rate" [naɪt˺ɹeɪt˺]), it is only at 10.5 months they are able to effectively use this information to segment the speech stream (Jusczyk, Hohne & Bauman 1999).

Mattys & Jusczyk (2001*a*) showed how sensitivity to allophonic cues helps children in segmenting words and are not simply responding to acoustic patterns. First, 8.5-month-old children were familiarised with a set of isolated words (e.g. *dice*). In the test phase, infants were presented with passages that either contain the target word (e.g. *Two **dice** can be rolled without difficulty.*) and other passages that do not contain the target word. Results show that children listen significantly longer to the passage containing the target word, hence showing they did recognise the word they had only seen in isolation in connected speech. In a similar experiment, children are once again familiarised with isolated words (e.g. *dice* [daɪs]). However, this time in the test phrase, infants are presented with passages that either contain a sequence in which the target word is split over two words (e.g. *d#ice* [d˺ʔaɪs] as in *The rink had been sprinkled with sprea**d ice***) and two passages that do not contain the target sequence. If infants listen longer to the passage that contains the target sequence split over two words, it means they would have incorrectly recognised *dice* in *weir**d ice***. If not, it would mean they have identified an allophonic variation that signals a word boundary. Results confirm it is not the case, hence showing that they noticed the word boundary in the similar sounding sequence (*weir**d#ice***). These set of experiments thus confirm that allophonic cues are used by children in order to segment the speech stream into words.

Coarticulation refers to the fact that a segment is never exactly pronounced the same way given the preceding and following sounds. Indeed, sounds tend to be blended together. Nonetheless, the "level" of blending depends on several factors. Johnson & Jusczyk (2001) indeed note that there is "less overlap of adjacent sound segments [...] in word-final and word-initial consonant articulations belonging to different prosodic domains". Consequently, if two phonemes display a low level of coarticulation, it might be they belong to two different words. Thus, such information might help the child segment its input. Curtin et al. (2001) showed in an ALL task that 7.5-month-old, are able to recognise isolated word-like patterns if these isolated patterns had coherent coarticulation effects between syllables, but could not if the syllables comprising the patterns where spliced in such a way that the coarticulation patterns where incoherent. This thus shows that children encode coarticulation information when remembering a new word and re-use this information later on to recognise words in the speech stream.

1.2.3.4 Transitional Probabilities

Children may also use statistics in order to decide if there is a word boundary between two adjacent syllables or not. If two syllables frequently co-occur, it seems reasonable to treat them as belonging to a single unit than to two different units.

In an ALL task, Saffran et al. (1996) tested if 8-month-old infants were sensitive to statistical (ir)regularities, and more specifically transitional probabilities (TP)[13] in order to segment the speech stream into words. To do so, they assembled a set of four artificial words (*pabiku*, *tibudo*, *golatu*, and *daropi*) so that the whole sequence would be two minutes long. The sequence was synthesised so that only statistical cues might be used (i.e. monotone voice, no stress, no pause, etc.). Because the words were repeated in a random order, the TP between two syllables of the same word is always 1 (e.g. *bi* is always followed by *ku*) while the TP of two syllables belonging to different words is 0.25 (e.g. *ku* might be either followed by *ti*[budo], *go*[latu], *da*[ropi] or *pa*[biku]). Hence, if children are indeed sensitive to the transitional probabilities of the syllables, they should consider the aforementioned words as valid units, but not *tudaro*, *pigola*, etc. It is in fact what the authors observe, letting them conclude that children do use transitional probabilities between syllables as a cue for word segmentation.

Even though the previous experiment does show that children are able to use TP to segment speech in a relatively short time, the input does not reflect the actual difficulty of real life language. Pelucchi et al. (2009*b*) hence tested if 8-month-old English-speaking infants were able to segment speech based on TP when presented with Italian words (e.g. *fuga*, getaway) embedded in grammatically and idiomatically correct Italian sentences. The sentences were constructed in such a way that the TP between *fu* and *ga* was always 1 (i.e. there were no other occurrences of the syllable *fu* except in the word *fuga*). Another real Italian word was used in the sentences (*bici*, bike), however this time the syllables *bi* was also used to construct other words (*bira*, *bigia*, *bivio*, etc.); hence, the TP between *bi* and *ci* is lower than 1.[14] At test time, if children indeed used TP to segment speech into sub-units, they should consider *fuga* as a real unit while *bici* should be discarded. It is in fact what is experimentally observed, therefore suggesting that using TP is not only used by children in ALL task, but also when confronted to real speech.

Not only are children able to keep track of forward transition probabilities (FTP), but

[13] Probability that the next syllable is Y given the fact that the current syllable is X is given by $TP = P(Y|X) = \frac{freq.XY}{freq.X}$.

[14] Such as in this sentence: "La cavia **Bi**da è in fuga da casa per aver giocato con le **bi**lie blu". (The guinea pig Bida is on the run from home for playing with the blue balls)

research suggest they can also keep track of backward transition probabilities (BTP).[15] To study this, Pelucchi et al. (2009*a*) embedded four target words into sentences in Italian (such as Pelucchi et al. 2009*b*): *fuga*, *melo*, *bici*, and *casa*. In all cases, FTP(Y|X) = P(*ga*|*fu*) = P(*lo*|*me*) = P(*ci*|*bi*) = P(*sa*|*ca*) = 1 (i.e. no other words than the target words start by *fu*, *me*, *bi*, or *ca*). However, BTP probability was different for half of the words: in one case BTP(X|Y) = P(*fu*|*ga*) = P(*me*|*lo*) = 1 (i.e. no other words than *fuga* or *melo* have the final syllable *ga* or *lo*), while in the other case BTP(X|Y) = P(*bi*|*ci*) = P(*ca*|*sa*) < 1 (i.e. other words than *bici* and *casa* end with *ci* or *sa*). If infants also pay attention to BTP, they should consider *fuga* and *melo* as better word candidates than *bici* and *casa*, as the former have both FTP and BTP = 1, while the latter have a BTP < 1. It is what is observed experimentally, letting the author conclude that infants do keep track of both FTP and BTP, and that such knowledge is used for multiple tasks, one of them being word segmentation.

Estes et al. (2007) showed that children may segment a continuous speech stream into words using TP without necessarily immediately assigning a meaning to the segmented units. First, 17-month-old children were exposed to an artificial language of four words (e.g. *timay#dobu#gapi#moku*) for which the only cues for word segmentation were TP. Then, the infants had to learn a mapping between a novel object and a novel word that was either one of the words used to build the initial sound sequence (e.g. *timay*), or a novel word that did not occur in the sequence (e.g. *nomay*), or a part word that was made up of two word parts (e.g. *pimo*). The only children that managed learning a novel word-object mapping were those for which the novel word was one used to construct the sequence. Hence, the authors conclude that not only children used TP to segment the speech stream, but that prior exposure to word facilitates subsequent word-referent mapping. This result is coherent with previous results: Jusczyk & Aslin (1995) indeed showed the child's proto-lexicon might also only contain word-forms that are not associated to any meaning at first.

1.2.4 Lexical Cues

Children may use lexical cues in order to parse the spoken input and segment it. Indeed, as soon as the child has isolated a certain number of word-forms, she can use them in order to isolate the words that precedes and/or follows the known word. White et al. (2010) coined the term "segmentation by lexical subtraction" and defines it as "the use lexical knowledge to impose a segmentation structure on the speech input".[16]

Jusczyk & Aslin (1995) showed that 6-month-old children are able to detect known words in the speech stream if they already heard the same word in isolation. Further study by Jusczyk & Hohne (1997) observe that infants are able to remember words for relatively long time spans (as long as two weeks in their study). Thus, isolated words may be used by children to segment the speech stream. Brent & Siskind (2001) estimate that 9% of infant directed speech are isolated words. Bortfeld et al. (2005) later showed that 6-month-old infants already used familiar words ("mommy", "daddy", or the child's own name) to segment the following word in the speech stream. Indeed, as soon as the familiar word is recognised, its end necessarily constitutes the beginning of the next word. Bortfeld et al. (2005) state (without experimentally testing) that the familiar words might not be

[15] Probability that the previous syllable is X given the fact that the current syllable is Y is given by $BTP = P(X|Y) = \frac{freq.XY}{freq.Y}$.

[16] This term, originally introduced in the aforementioned publication, did not refer to child language acquisition (the original study explored the segmentation strategies used by Hungarian that are L2 speakers of English). However, we do believe that this term also applies to child language acquisition as it adequately describes the method used by children to segment the spoken input.

restricted to those already mentioned, but might also include other familiar words such as "diaper" or "bottle" that are part of the child's everyday life. Bergelson & Swingley (2012) later showed that other common words known by children included words refering to body parts ("feet", "hands", "mouth", or "eyes") or refering to food ("yogurt", "banana", or "milk"). Hence, all these words might serve as support for the child to segment the speech stream into sub-units.

Other types of words such as functional words (i.e. closed-class words such as determiners, conjunctions, prepositions, etc.) are also used by the child to segment the speech stream. For example, Shi et al. (2006) showed that the determiner "the" facilitates the recognitions of a test pseudo-word while non-sense words ("kuh" and "ler") or a least frequent determiner ("her") do not. A similar effect was found for 8-month-old French-speaking infants by Shi & Lepage (2008) where infants better recognised a novel target word in isolation if it was heard during the familiarisation phase either with a highly frequent determiner before ("des") or less frequent determiner ("mes"), but not if it was presented with a non-sense similar-sounding word ("kes"). They however note that this effect is not observed with all determiners as they do not observe any difference when the target word is presented with "vos" or with "kos", thus showing that in order to be used, the determiners should be frequent. Höhle et al. (2004) found a similar behaviour for 14-month-old German infants. Finally, Haryu & Kajikawa (2016) ran a similar analysis on Japanese and the use of particles by children. Contrary to Indo-European languages, particles are postpositions and are used to indicate the function of the preceding word (subject, object, etc.). They found that 15-month-old Japanese toddlers use particles, and more specifically the "ga" particle which is the subject particle, in order to segment speech. Thus, the pattern of using functional words as cues for word segmentation seems to be valid cross-linguistically.

Finally, Johnson et al. (2014) report in their study that interjections (such as "oh", "no", "wow") account for 80% of the words heard in isolation in child directed speech. Such words, that do not serve any grammatical function but act merely as phatic and conative interjections (Ameka 1999), might help the child segment the speech stream.

Therefore, children use the lexical knowledge they have to parse the speech stream and acquire novel word. However, as noted by Curtin & Hufnagle (2009), "familiar words are not enough for infants to excel at segmentation". Indeed, in order to fully segment the speech stream, the child should develop other routines such as those already discussed thereafter.

1.2.5 Other cues

Another cue that children could use in order to infer the lexicon of their language derives from the nature of the spoken input they hear. Indeed, Child Directed Speech (CDS)[17] has certain properties (at various levels: acoustics, syntax, etc.) which are not found in Adult Directed Speech (ADS). Fernald et al. (1989) reports for example that both mothers and fathers have a higher fundamental frequency (F0) with exaggerated pitch contours and long pauses between utterances. CDS is also reported to have a slower rate (syllable/seconds). Nevertheless, Church et al. (2005) reports that this is to be explained by the elongated pre-boundary final vowel lengthening that is greater in CDS than in ADS. If this final syllable is excluded in the speech rate calculation, CDS and ADS speech rates are similar. Johnson et al. (2014) also reports a higher proportion of isolated words in infant-directed speech (IDS, up to 90-days-old infants) than in CDS (i.e. from 90 day old up to 2.5 years old) and the proportion decreases again in ADS. Their study also reveals that nouns are more likely to appear at utterance-edges in IDS than in CDS or ADS.

[17] Also known as *motherese*, *parentese*, *baby talk*, etc.

It should be noted that IDS/CDS is not universal, as in some cultures children are not (or barely) addressed to until they are able to speak (Cristia et al. 2019) or are addressed to like adults. Thus, this type of speech is not mandatory for children to learn their language, but seems to help the child when it is used.

Nelson et al. (1989) indeed showed that 8-month-old infants prefer listening to speech excerpts in CDS that preserved clausal boundaries over excerpt in CDS that do not preserve clausal boundaries. They showed no such preference for ADS, whether clausal boundaries where respected are not. Infants thus seem to be able to better detect what a legal prosodical unit is when CDS was used but not when ADS was used. Therefore, CDS seem to provide useful cues to the child in order to segment the spoken input, at least at clausal level.

Thiessen et al. (2005) further tested in a ALL task if CDS was easier to segment for 7.5-month-old infants than ADS. To do so, they had infants listen to a sequence of words (*dibo*, *kuda*, *lagoti*, *nifopa*) whose only cue to word boundaries was the probability of co-occurence of each syllable.[18] The sentences were read both as CDS and ADS by the same speaker. In the test phase, children payed more attention to well-formed word (e.g. *lagoti*) than to part-words (i.e. *gotini* or *dalago*) when having been familiarised with CDS, but show no difference when familiarised with ADS. Hence, the authors conclude that CDS indeed improves word segmentation for young children.

Ma et al. (2011) explored if CDS facilitated word learning compared to ADS. In their experiment, 21-month-old children were presented with novel objects (displayed on a screen) and were told the name of the novel object in a familiarisation passage (e.g. *modi* in the passage "Look here! It's a *modi*! See the *modi*. That's the *modi* [...]") either in CDS or ADS. During the test phase, Ma et al. (2011) checked if the children looked longer to the object that was presented as a *modi* than to the distractor object when hearing the sentence "*Modi*. Where's the *modi*". Children that were familiarised with a CDS passage showed such a pattern, but not those that were familiarised with the ADS passage.

Hence, this body of work suggests that, even though CDS is not found in every culture, it enhances the ability of children to extract words from the speech stream. Di Cristo (2013, pp. 172-173) indeed states "the repetitive use of lexical stress in the speech addressed to the child helps to fix the child's attention and at the same time helps the child to identify word boundaries". Not only IDS/CDS help children identify word boundaries, they also help them map the extracted word-form to their referents. For Di Cristo (2013, p. 175), it may be because "the effect of the accentual function is to emphasise the element that is the subject of this prosodic distinction [... making] it possible to establish a link between a phenomenon of physical or perceptual salience and the promotion of a mental or cognitive salience."

1.2.6 Multiple Cues: from Mutual Exclusion to Combination

The previous sections presented the numerous cues children may use to segment the speech stream. Yet, not all cues are available nor used at the same time. Some cues are only available from a certain age, such as phonotactic cues as they require some amount of exposure to the target language, while some are available since the child was born, such as the ability to detect allophones. Similarly, even if multiple cues may be used at the same time, research shows that children might favour some cues over others at a certain point of their development.

Johnson & Jusczyk (2001) tested in an ALL task if 8-month-old infants relied more on coarticulation or statistical cues to decide whether there was a boundary or not. The same

[18] The syllables *mo* and *fa* was added respectively at the beginning and the end of the sequence so that the sequence beginning or end would not be used as a cue

sequence of words as Saffran et al. (1996) was used (*...pabiku#tibudo#golatu#daropi...*, see §1.2.3.4) unless this time, instead of splicing a part-word (e.g. ku#tibu) from the original sequence, it was resynthetised. Contrary to the previous experiments of Saffran et al. (1996), the new part-word had a coherent coarticulation pattern at word boundary — or what should theoretically be considered a word boundary from statistics alone (i.e. between *ku* and *ti*). Hence, according to the coarticulation pattern, there should not be a word boundary while according to statistical cues there should be. If children favour the part-word this time, it means that they discarded the TP information and only considered the acoustic cues. They results show that infants consider *ku#tibu* as a valid word and hence favoured coarticulation information over statistical information.

Mattys et al. (1999) similarly showed that when phonotactic cues and prosodic cues provide conflicting information about the word boundary, 9-month-old infants only rely on the information provided by the prosodic cues. They favour a disyllabic CVC·CVC SW sequence with a frequent between-word C·C internal cluster over a disyllabic WS sequence with frequent within-word C·C internal cluster. Hence, more weight is given to prosodic cues than to phonotactic cues.

Jusczyk, Houston & Newsome (1999) revealed that while 7.5-month-old children are unable to segment WS words, 10-month-old children are able to do so. They cannot therefore do it using SW pattern as a cue, as the word that is extracted does not have this accentuation pattern. Consequently, at some point, infants need to discard the information provided by the primary (i.e. first used) cues and integrate other cues to segment the speech stream. Such pressure could arise from what is known as the Possible Word Constraint (PWC, see Norris et al. 1997) which states that segmentation should leave a low number of un-analysed chunks (ideally none) and that the chunks resulting from the segmentation process should be existing words. Johnson (2003) showed that segmentation in 12-month-old infants is indeed conditioned by such PWC. Hence, in the "guiTAR is" experiment, the segmentation that conforms the best to the PWC is "guiTAR#is" despite the WS pattern and not "gui#TARis" as the latter leaves a non attested word "gui".

Finally, segmentation cues may also be used in conjunction rather than in isolation. There are infinitely many ways of segmenting a speech signal into sub-units, but the multiple cues children have access to constrain the possible segmentations. Indeed, "constraints should make a 'combinatorial explosion' less likely." (Thiessen & Erickson 2009, p. 44).

In the previously described experiments of Pelucchi et al. (2009*b*,*a*), where children found disyllabic words embedded in fluent Italian sentences, the Italian speaker was asked to read the sentences with a "lively voice, pretending to be in front of a baby". Hence, it might also be the combined effect of TP and CDS that helped children segment the spoken input, and not TP alone.

The use of TP to segment the speech stream also seems effective when combined with another cue. Johnson & Tyler (2010) showed that 5.5- and 8-month-old children do manage to extract words from the speech stream using TP if the words are uniform in length. However, when the words are not uniform in length (three-syllable and two-syllable long), infants do not seem to be able to segment the speech stream correctly. Therefore, this suggests that in order to be used effectively, TP should be combined with other cues, as real life language has words with a different number of syllables.

A similar result as Johnson & Tyler (2010) was observed by Mersad & Nazzi (2012). In their experiments, they tested if 8-month-old French-speaking infants could segment an artificial language consisting of words of non uniform length (e.g. respectively bisyllabic words *pabi*, *tibu*, and *mãma* and trisyllabic words *golatu*, *daropi*) where TP were the only cues to properly segment the speech stream. As in Johnson & Tyler (2010), French infants were not able to segment the speech stream into words. Yet, when the pseudo word *mãma*

was replaced with *mamã* ("mum"), the infants succeeded in segmenting the speech stream, even if it consists of non-uniform words. Consequently, the authors suggest that children were able to segment the speech stream using these two cues in conjunction. Indeed, if TP alone had been enough, the infants should have managed to segment the speech stream in the *mãma* condition, which is not the case. The authors thus suggest that infants first perform a segmentation by lexical subtraction (see Section 1.2.4) — therefore discovering also the beginning and end of the other words — and then used TP to segment the rest: with the help of these two cues, the child managed to segment the speech stream, when if only one cue or the other had been present, the child would not have been able to do so.

1.2.7 Conclusion on Speech Segmentation

In this section, we presented the multiple cues children use in order to segment the speech stream. We showed that children are sensitive to suprasegmental cues (e.g. prosody), segmental cues (e.g. phonotactics), and subsegmental cues (e.g. coarticulation). Once segments have been isolated, they can then use statistical cues (e.g. TP) to further segment the speech stream. Additionally, we showed that children also make use of isolated words in order to segment the speech stream (segmentation by lexical subtraction) and we highlighted the benefits of using CDS over ADS.

It thus appears that children use a combination of top-down (e.g. isolated words) and bottom-up cues (e.g. phonotactics, TP, etc.) in order to segment the speech stream. Both approaches are necessary — as, for example, discovering the regular phonotactic patterns within words requires the knowledge of a set of individual words — and complementary — as in return, the inferred phonotactic patterns enable the learner to segment even more words.

We will conclude this section by introducing the Hierarchical Framework by Mattys et al. (2005) which shows how the various cues we presented are organised in a *mature* segmentation system (Figure 1.1). As can be seen, the cues that are the most important are lexical cues. That is, lexicality places such a hard constraint on word segmentation (with the Possible Word Constraint), that it is the cue that is used primarily by adults to segment the speech stream. If, for some reason, the lexical level is not directly accessible, when for example the speech signal is noisy, then adults fall back on non-lexical cues to parse the speech stream (segmental cues, word stress).

Contrary to adults, speech segmentation in children is a *developing system*, and seems to be bootstrapped by the cues that adults tend to use the less favourably (i.e. prosodic cues). However, we believe that this diagram should also include an arrow at the top, as we have seen that children also hear a significant proportion of words in isolation (up to 9% such as *mommy*, *bottle*, etc.) and even more if interjections are counted.[19] Johnson et al. (2014) reports that 29% of the words in CDS are interjections and 80% of the words pronounced in isolation are interjections. These words are extremely frequent and can be used by the child to bootstrap speech segmentation. Hence, speech segmentation in a developping system uses, as in a mature system, both top-down and bottom-up cues, the only difference being the relative importance given to each approach. Given the little vocabulary children have, bottom-up cues are used more favorably than top-down cues.

This review allowed us to observe the different strategies used by children to segment the spoken input into sub-units. We will explore in this thesis if a neural model of visually grounded speech also segments its input into subunits as children do.

[19] They are usually discarded in the statistics of IDS/CDS speech corpora.

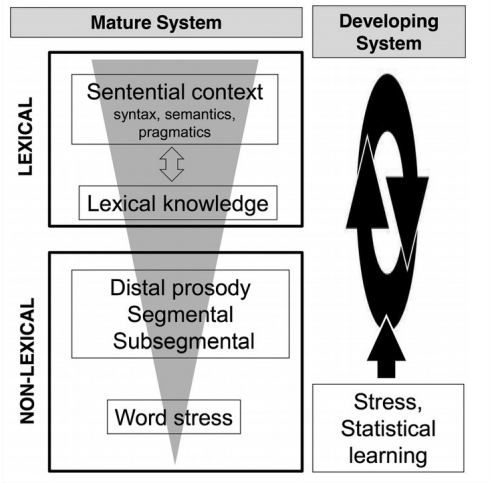


Figure 1.1: Figure taken from Mattys & Bortfeld (2015) presenting the cues used in a mature speech segmentation (i.e. adult) system and in a developing system (i.e. child) with the black arrows illustrating the back and forth interactions between top-level cues (semantics, lexical) and low-level cues (prosody). The grey triangle indicates the importance of each set of cues (with the wider end signalling the greater importance of such cues).

1.3 Word Mapping

In the previous sections, we presented the cues children use to segment the speech signal into sub-units. However, we somehow considered language *in vacuo* as if words existed *per se*, disconnected from any context. This is of course not the case:

> The important point in all this is that language only has existence with respect to the physical and mental order of things. Language is not any sort of independent system, to be tapped into as required. Aspect of language have significance only as they relate to aspects of the world. (Dixon 2012, p. 434)

This point, though obvious, should be mentioned as the NLP community often tends to consider language as having an existence of its own, disconnected from the real world. We will further discuss this point in the following chapter (see Section 2.1.3).

In this section, we highlight how children assign a meaning to the extracted word-forms they have isolated in the speech signal, and show why context is necessary to acquire a language. Recent research show, that the more often a word is heard in different contexts, the better it is learnt by the child (Roy et al. 2015). Hence, contextual information plays a decisive role in lexical acquisition. We will mainly focus on the visual modality, as it is the modality that is used by our model, and show how it influences language acquisition.

1.3.1 Prerequisites

Once children have isolated a certain number of word-forms, they should assign them a meaning. Meaning is not a ready-to-be-used object that the child would only have to map to a given word-form; but rather should be constructed by the child on the basis of its experience with the physical world.

Hence, assigning each word-form a meaning is only the last step. The first step consists in finding the referent of each word-form. Finding the right referent can only be done in context, as a given word-form relates to a given object in the physical world, or a given

percept. Potentially, the referent may change from one occurence of a word-form to another. For example, the word-form "dog" might be uttered when refering to different dogs, each one being a referent of the word-form "dog" when it was uttered. Once the child has infered the referents of a given word-form, it should find the commonalities shared by the referents in order to abstract a potential meaning.

However, finding which is the referent of a word-form is far from an easy task. Indeed, there are an infinite number of possibilities the child has to choose from (see Quine's famous *gavagai* example, Quine 1964). The problem is even trickier if one thinks of abstract words such as *pride*, or *love* for which the referent is not perceptible and refers to inner sensations. A purely associanistic approach, where word-form/referent co-occurence alone would solve the problem, seems unrealistic. Indeed, if word learning were only done using cross-situational statistics, then children would necessary have to hear a word several times in order to learn it, which is not what is observed in practice. Indeed, children are able to do fast-mapping (Carey & Bartlett 1978), that is, learn to quickly map a word-form to its referent and derive a meaning with a very few number of examples (in some cases only one example). This led researchers such as Bloom to argue that "statistical covariation between word and percept is neither necessary nor sufficient for word learning." (Bloom 2002, p. 59). Therefore, word/referent mapping and ultimately the acquisition of meaning seems to require something more.

1.3.2 Theory of Mind and Shared Attention

For Tomasello (2009, p.3), acquiring a language involves two skills: pattern-finding, and intention reading. Pattern-finding involves all the statistical computations the child makes in order to segment the speech stream, as well as recurrent sequence detection and extraction in the speech stream (which we presented in the past sections). Intention reading is, according to Tomasello, the necessary ingredient for children to acquire their language, and particularly when it comes to inferring what is the referent of a given word-form.

Bloom (2002) similarly argues that "children use their *naive psychology* or *theory of mind* to figure out what people are referring to when they use words". Theory of mind could be defined as the capacity humans have to infer the mental states of others and relate them with their own. This allows them to interpret and understand the emotions, intentions, beliefs, and behaviours people around them have (Astington & Dack 2008). Specifically for language, having a theory of mind is necessary to appropriately contextualise what is said. Hence, Bloom argues that children are able to do fast-mapping only because they have a theory of mind.

Having a theory of mind enables children to enter in shared attention frames (also called joint attention frames). A shared attention frame may be defined as "a triadic episode of interaction involving a caregiver, an infant/toddler and an object" (Rudd & Johnson 2011). To be more precise, there is shared attention when both the child and the caregiver are attending to the same object, while being mutually aware of what the other is attending to (see Figure 1.2).

According to (Bloom 2002, p. 46), if having a theory of mind is a necessary prerequisite for building a lexicon, there should be a correlation between the onset of word learning and the first evidence of the child having a theory of mind. Morales et al. (1998, cited by Bloom 2002) observe that in six-month-old infants, the ability to follow the caregiver's gaze — and hence initiate a shared attention state — is a good predictor of their receptive vocabulary at 12 months. This result is coherent with previous findings: Tomasello & Todd (1983) for example observed that infants that have long periods of shared attention with their caregiver at 6 months have overall a larger perceptive vocabulary than those for which

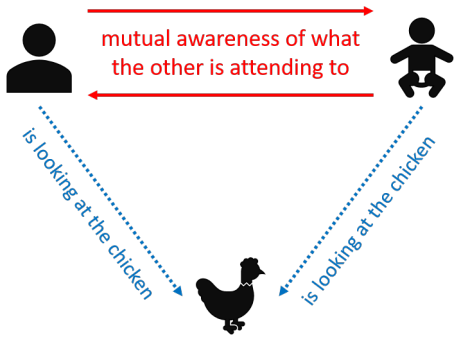


Figure 1.2: Illustration of shared attention between a caregiver (top left) and a toddler (top right), attending to the same objet, a chicken (bottom).

shared attention moments are infrequent. Moore et al. (1999) showed that shared attention was a cue that outweighted saliency when 24-month-old children learn a new word. That is, children are better able to learn a novel word if this word was part of a shared attention frame, even if it was not visually salient. On the contrary, children did not learn words when the object was not part of a shared attention frame. Hence, having a theory of mind, and being able to enter in shared attention frames seems crucial for lexical acquisition.

1.3.3 Assumption and Biases

Children's ability to assign a meaning to a word-form seems to obey a certain number of assumptions and biases. Markman (1990) lists several biases children seem to have when acquiring novel words: the *whole object assumption*, the *taxonomic assumption*, and the *mutual exclusivity assumptions*.

The *whole object assumption* states that children treat word labels as refering to an object as whole (e.g. a tree) and not a sub-part of it (e.g. trunk, leaf, branch). Hollich et al. (2007) recently gathered evidence in favour of this hypothesis. When 12 and 19-month-old children are presented with novel objects with detachable parts, they associate the novel word to the whole object rather to one or the other detachable part. The *taxonimic assumption* states that the usage of novel words is extended to similar objects, that is, object with similar characteritics (e.g. dog label for a German Shepherd and a Bulldog), and not to objects thematically related (e.g. dog and leash) (see Markman & Hutchinson 1984). Finally, *the mutual exclusivity assumption* states there is a bijection between a word-form and its referents: that is, for a given label there is only one associated (conceptual) referent (e.g. /dɔg/ for dogs), and a single referent can only be referred to by a single label. According to Markman (1990), this assumption balances the previously presented assumptions. Indeed, children necessarily have to acquire meaning for object parts and other properties of objects (e.g. color, texture, etc.). Such constraint would thus "lead them to analyze the object for some other property to label", and more generally, analyse in greater depth the world around them in the search for new referents.

Other constraints were postulated, such as the *principle of constrast* and the *principle of conventionality* (Clark 1987). The principle of constrast says that "every two [word] forms constrast in meaning". That is, the child would consider that no two word-forms are exact synonyms. Such principle would also encourage the child to explore its environment so as to find a possible referent. The principle of conventionality states that "for certain meanings, there is a conventional form that speakers expect to be used in the language community". That is, children should expect stability in the linguistic system they are learning, and

hence that word-forms are regularly associated to a given meaning. This is also what is expected with the mutual exclusitivity assumption.

1.3.4 Mismappings

Even though lexical acquisition seems to be guided by a certain number of biases, children often make mistakes. When they fail to correctly map a word-form to its referent, they usually *overextend* its usage (e.g. dog for every furry animal with four legs) or *underextend* it (e.g. dog for German Shepherds only). Even though over- and underextensions are most appearant in production as a suppletive mechanism to refer to something the child does not know the word for yet (Clark 1978), they also occur in perception (Behrend 1988). Rescorla (1980) provides a typology of the main causes of overextensions: *categorical overinclusions* where a label is used for a referent close to the real referent (e.g. *baby* for children), *analogical overextensions* where a label is used for a referent which bears a similarity with the true referent (e.g. *ball* for marbles or apples), and *predicate statements* which are holophrastic (e.g. *dog* when pointing at a basket to show the dog is not here). Rescorla (1980) estimates that one third of the child's first words are overextended, and about three quarters of the extentions being either categorical overinclusions or analogical overextensions, most of them being done on the basis of "perceptual similarity" (as opposed to functional similarity).

Regarding overextensions, most of them seem to be explainable by a *shape-bias* (Landau et al. 1988); that is, most words are overextended based on the global shape of their referent (e.g. dog for any four-legged animal). Even though most authors state that children overextend words based on "perceptual" features (i.e. shape, size, texture), in most experiments, the visual modality is usually the only modality available to the subjects.[20] Nonetheless, this seems reasonable to some extent, as "[s]hapes of certain things (those too large, distant, or gossamer to explore by hand, such as mountains, birds, and fog) are visible only in practice" (Landau & Gleitman 1985, p. 13).

Hence *word* acquisition and *world* perception seem to evolve jointly: first, word-forms are rudimentarily associated to referents, and second, the aforementioned assumptions and principles encourage the learner to further analyse its environement, so as to discover new referents that can be associated to new word-forms. Association between a word-form and its referent is done on the account of perceptual characteristics, for which vision seems to play a major role.

1.3.5 Visual Modality

In order to map the segmented word-forms to their referent, vision plays a central role. Indeed, "it is largely vision that directs the infant's attention to persons, objects and events; vision is important for the co-ordination of attention between parent and child; and it often constrains (especially in Anglo-American culture) the content of language addressed to the child" (Andersen et al. 1993). Vision hence provides the necessary interpretative context, which is necessary to get "language learning off the ground floor" (Landau & Gleitman 1985) and seems to be the modality used the most by children in order to map word-forms to their referent (this will be discussed more extensively in the following section 1.3.5.2).

[20] From Landau et al. (1988): "subjects were *shown* a single standard (i.e. the Dax) and told its name ("This is a Dax"). The standard remained in *view* throughout the entire procedure. Subjects were then *shown* the seven test objects one by one" Emphasis added.

1.3.5.1 Word Mapping and Vision

It seems that infants' ability to systematically associate a visual referent to a word-form only appears at around 12 months of age. Thiessen (2010) tested adults and 8-month-old infants on their ability to segment words from fluent speech, and associate the target word with a visual referent. During the familiarisation phase, an image depicting the target word appeared on a screen at the same time the word was uttered, and disappeared from the screen once the word's offset was reached. Therefore, in this experiment there is a regular association between the boundaries of the target word and its visual appearance, which should facilitate both segmentation and mapping. The author observes that while adults do benefit from the synchronicty between both modalities, 8-month-old children do not. This result is in line with previous results, Werker et al. (1998) indeed showed that 8- to 12-month-old infants are not able to learn word-object pairing. They however found that 13- or 14-month-old infants are able to do so.

Friedrich (2008) made similiar observations by studying the event-related brain potential of children between 12 and 19 months. The subjects were primed with a visual stimulus which was followed by either its matching audio word-form (e.g. "cat" for the picture of a cat) or a mismatching one (e.g. "dog" for the picture of a cat) and observed if there is a N400 priming effect. The N400 potential is linked to semantic processing in the brain. The expectation is that if the two modalities match (e.g. "cat" with cat) then an N400 effect should be observed, and no effect should be observed if the modalities do not match. While the N400 effect was observed for infants from 14 months on, it was not for 12-month-old infants, showing they did not seem to map the word-form to its visual referent.

Smith & Yu (2008) tested if 12- and 14-month-old children are able to quickly acquire the meaning of a novel word and effectively resolve the ambiguities that might arise if several referents are possible by simply using cross-situational statistics. To test this, they present two novel objects to infants, each paired with a novel noun (e.g. a picture of a ball and a baseball bat, paired with their respective name). Hence, the child is unable at this stage to tell which object is paired with which noun. Nonetheless, by seeing other occurences of one of the two objects paired with another noun (e.g. picture of a ball and a dog, paired with their respective name), the child is able to learn the correct word-object mapping. While this is not suprising for 14-month-old infants, Smith & Yu (2008) also found it was the case for 12-month-old infants, which seems contradictory with the result of Friedrich (2008) and Werker et al. (1998). However, as this ability seems to develop in a very short time frame (between the 12th and 13th month of age), it is reasonable to observe slight differences from an experiment to another, especially if the experimental settings are different.

Vouloumanos & Werker (2009) tested if 18-month-old infants are able to map a word-form to its referent when a given word-form is not always uttered at the time its referent is simultaneously perceptible. Their study reveal that children are able to map the word-form to its referent, even if the word-form is used in contexts where the referent is not perceptible. It shows that children are able to maintain several hypotheses of the meaning associated to a given word-form, and seem to select the most frequent one.

The visual modality and the intention reading skills of infants are combined so as to make sense of what is said. Goldin-Meadow (2009) argues that "referring to an object in gesture could facilitate learning the word for that object in toddlers at the early stages of language learning" as the referent of a given word-form would be unambiguously signaled to the child. Similarly, being able to interpret other gestures, and more specifically gaze direction helps children learning novel words, by noticing what the caregiver is looking at (Law et al. 2012). Computational simulations (Smith & Yu 2008, Frank et al. 2009) find that word learning by cross-situational statistics is significativelly enhanced by speakers'

intention information.

1.3.5.2 Language Acquisition and Blindness

Even though vision seems to be the most prominent sense children would use to learn a language, lexical acquisition is still possible when vision is lacking: indeed, blind people are as proficient in their native language as sighted native speakers.[21] However, being congenitally blind might slightly slow the learning process and affect the quality of what is being learnt — mainly phonetics and semantics — at least in the early stages. We will thus review in this section what is known of language acquisition in blind children. This will enable us to better understand what role vision plays in sighted children when they acquire their native language.

Most authors observe a tendency in blind children to talk about past events instead of the "here and now". This tendency to favour past events over what is occurring in the "here and now" is to be explained by an absence of obvious shared attention between the child and the caregiver. Such absence of shared attention is mainly explained by an inability of the caregiver to interpret what the child does and an inability to understand what the child is interested in (Andersen et al. 1993). By focusing on past events, the child tries to recreate a situation where she knows the caregiver and herself have shared a common experience they can talk about.[22]

Research has also shown that blind children also suffer from a lack of decentration,[23] that is, they fail to consider anything else other than themselves and their immediate surroundings. Thus, the main topics that are initiated by blind children refer to action or states involving the child herself (see Andersen et al. 1984 and Peltzer-Karpf 1994, p. 41) and not to the outside world. It is easily understandable why blind children do so, as it is impossible for them to discuss or get interested in persons or objetcs that are not directly noticeable (either by touch, hearing, or smell) contrary to sighted children who can get interested in objects that are not in their immediate surroundings. Consequently, by being deprived of sight, blind children can only initiate interactions that involve persons or objects that are in their immediate surrounding. Unsurprisingly, blind children mainly rely on haptic input (i.e. touch) to compensate for the lack of visual input (see Dunlea 1989, p. 160; or Peltzer-Karpf 1994, p. 58), but it can only compensate for so much.

Consequently, this lack of decentration generates interactions with the caregivers that are different from the interactions the same caregiver would have with sighted children. As caregivers usually fail to interact properly with blind children — at least at a very early stage — the nature of the conversations are affected. For example, Andersen et al. (1993) reports that caregivers repeatedly "bombard" the child with labels instead of providing a description of their environment as they would usually do. She also reports that caregivers tend to be more directive so as to encourage children to explore their environment. She finally notes that caregivers tend to "restrict their verbal input".

Naturally, both the child's comprehension and production is affected by the nature of the interactions. Andersen et al. (1984) reports in her study that deictic terms were only

[21] Blind people also make the same conceptual differences as sighted people for verbs such as "look" and "see", the only difference in the way of looking or seeing being the modality. See Landau & Gleitman (1985).

[22] "Talking about shared past events allows the blind children to maximize the probability that they and their addressees have a common focus of attention, just as the use of visual cues aids sighted children in talking about the here and now." (Dunlea 1989, p. 157).

[23] "The gradual progression of a child away from egocentrism toward a reality shared with others. [...] It can also be extended to the ability to consider many aspects of a situation, problem, or object, as reflected, for example, in the child's grasp of the concept of conservation." in `https://dictionary.apa.org/decentration` consulted on 16/11/2020.

used by sighted children and not by blind children. On the other side, Peltzer-Karpf (1994) note that blind children tend to use twice as much interrogative pronouns as sighted children which seems natural as it is the only way for them to explore the world.[24]

Blindness also seems to affect the generalisation capacities of the child. Indeed, the learning paradigm in sighted children and blind children is different. As noted by Dunlea (1989, p.10) "the act of touching and feeling is a search for information; it implies a conscious effort to obtain sensory stimulation" while sighted children do not need to do a conscious effort in order to interpret and understand their environment. This has also a more insidious effect that blind children are only able to fully grasp (or even notice) some concepts once they are able to physically explore their environment, as noted by Peltzer-Karpf (1994, p. 40): "While sighted children formulate concepts such as ego, others, agent, action and object through observation even before the crawling age, the blind child lacks decisive possibilities here".[25] Thus, the "computational load" blind children face once they are able to freely explore their environment is higher than for sighted children.[26]

Because blind children are only able, to some extent, to conceptualise a large part of their environment later than sighted children, Dunlea (1989) postulates that they fail to grasp the symbolic nature of language and fail to apply appropriately the words or chunks of words they learn to new situations, whereas sighted children typically display a pattern of overgeneralisation. Indeed, Andersen et al. (1984) notes that blind children tend to parrot what they hear without further analysis (i.e. segmentation). Mills (1983, p. 145) also has the same observations where he notes that the blind "child's language may remain tied to familiar social routines [...] and a reliance on phrases acquired wholesale is often couple with a marked propensity for imitation and echolalia." Concerning the lack of generalisation, Andersen et al. (1984) observe "word-referent isomorphism, where children treat words as if they are proper names". Vision thus seems to act as a facilitator for abstracting and conceptualising the world, as it is easier to perceive what makes a set of objects similar when all the objects can be perceived simultaneously, such as noted by (Dunlea 1989, p. 87): "Vision is unique in that it provides instant simultaneous access to information which is otherwise segmented by space and time. Without vision, most information is necessarily perceived sequentially through haptic exploration". When overgeneralisations do actually occur, Dunlea (1989) reports that they are mainly based on tactile or auditory information, but not on olfactory or taste information.

Finally, regarding phonetics and phonology, Dunlea (1989, p. 15) reports that blind children tend to confuse only acoustically similar sounds while sighted children also confuse sounds because of their visual similarity (i.e. mouth movements). Peltzer-Karpf (1994, p. 24) notes that even though blind children succeed in acquiring the phonological system of their native languages, it generally takes longer for them to do so than for sighted children.

Hence, it appears that the lack of visual input slightly hinders language learning at an early stage. As we mentioned, this is to be explained by the fact that vision helps children conceptualise the world their live in. They can do so very early on as they do not need

[24] "Dieses Ergebnis ist hinsichtlich der Art der Sinnesminderung nicht überraschend, weil anzunehmen ist, dass das Fragen ein wichtige Strategie der Blinden darstellt, die Welt zu erforschen." ("This result is not surprising in view of the sensory impairment, because it is understandable that questioning represents an important strategy for the blind [child] to explore the world.").

[25] Original citation in German: "Während sehende Kinder noch vor dem Krabbelalter durch Beobachtung Konzepte wie Ich, Andere, Agent, Handlung und Objekt formulieren, fehlen dem blinden Kind hier entscheidenden Möglichkeiten"

[26] "Für das blinde Kind stellt der Wortschatzerwerb eine enorme intellektuelle Leistung dar, da es aufgrund des fehlenden visuellen Inputs Wörter, die sich auf Farbeindrücke, räumliche Distanzen oder Grössordnungen beziehen, nur schwer begreifen kann." ("For the blind child, the acquisition of vocabulary represents an enormous intellectual achievement, as the lack of visual input makes it difficult for him or her to grasp words that refer to colour impressions, spatial distances or dimensions.") Peltzer-Karpf (1994)

to physically explore their environment, but simply have to watch and see. Contrary to sighted children, blind children have to be able to walk (or crawl) in order to discover their environment and make sense of it. Even though blind children and sighted children have similar lexemes, the meanings they associate to them are different (Dunlea 1989, p. 59). Only when they have conceptualise their world can lexical acquisition take place successfully. Vision also helps children make sense of what is said to them, and help the child segment the spoken input into smaller units.

1.3.6 Conclusion on Word Mapping

As we put forward in this section, language acquisition is a multimodal phenomenon, that relies both on sensory perciptible cues, such as vision, and social cues, by using, for example, shared attention frames in order to map an acoustic stimulus to its referent. In this thesis, the model we use will only have access to perceptible cues — in the form of images — but naturally will not be able to use any of the social cues generally used by children. Similarly, our model does not have any Theory of Mind. Thus, we will investigate if using a simple associanistic approach suffices to learn a reliable mapping between an acoustic stimulus and a visual context.

1.4 Word Recognition

Once the child has isolated word-forms, they must be stored in the mental lexicon in such a way they can be retrieved afterwards, so as to create her own utterances or to interpret heard utterances. However, what appears to be simple, that is, storing word-forms in the mental lexicon, is far from trivial:

> The scope of phonological representation is a fine balance. On one hand, memories for words have to be sufficiently broad to incorporate the rampant variability inherent in natural discourse and to normalize for accents, different voices, emotions, and other factors. On the other hand, memories have to be sufficiently specific to not incorrectly equate minimal pairs or tolerate mispronunciations. (Singh et al. 2012)

Spoken word recognition can be defined as the process of accessing lexical items in the mental lexicon from phonological patterns in the speech signal (Magnuson et al. 2013). That is, it implies mapping what is being heard with lexical items stored in the mental lexicon.

We will review in this section several models of spoken word recognition that have been proposed to explain how humans, children included, manage to do so. This will allow us to compare — later on in this thesis — word activation and recognition in humans and in a neural model of visually grounded speech, and observe if the patterns that are learnt by a neural model are similar to that of humans.

1.4.1 Cohort Model

One of the very first models trying to account for how humans recognise and extract words from the speech stream is the Cohort model by Marslen-Wilson & Welsh (1978). According to this model, spoken word recognition proceeds in three steps that occur simultaneously which are: *access*, *selection*, and *integration*.

Access consists in activating a set of word-forms in the mental lexicon which corresponds to the acoustic input. Each word-form is paired to an activation unit, allowing for the

activation of multiple word-forms at once. These units are only activated if the perceived acoustic input exactly matches the internalised phonological form. If there is a mismatch between the perceived acoustic input and the phonological form, the unit is deactivated and the word coded by this unit is not considered a valid candidate anymore. Hence, the initial phone of the speech stream activates a set of units whose word they code for all start with the perceived phone. This set of activated word-forms is called a COHORT. The word-forms of the cohort remain activated as long as the perceived acoustic input matches the internalised phonological form of each word of the cohort. The words comprising the initial cohort that do not match the perceived input are pruned. The process of interactive pruning of the cohort is called *selection*. The *integration* process is part of the selection process, but is however more concerned with the nature and function of the words of the cohort rather than with form. Indeed, not only should the phonological form of the words in the cohort match what is perceived, but they should also be be semantically coherent with the context and syntactically fit within the sentence. Thus, words that do not fit either syntactically or semantically are removed from the word cohort. Hence, this model integrates bottom-up activation signals — from the speech signal to the phonological form of a word — as well as top-down inhibition signal — essentially by controling for the syntactic and semantic validity of the activated words.

This initial version of the COHORT model has several problems. First, it requires the perceived acoustic stimuli to exactly match the internalised phonological forms. However, it is well known that this is not how humans process speech. Indeed, humans are able to recognise a word even when it is mispronounced and might not even notice any mispronunciation (Cole 1973); when for example one sound is replaced by a similar sound ([m]/[n], [s]/[ʃ], etc.). As in this model, mismatching words are removed from the cohort, it would thus be impossible to recognise mispronounced words. Second, it requires the onset of the speech stream to necessarily be the onset of the word, which is unrealistic as one might drop in a conversation or overhear a conversation and still identify what word was being uttered when dropping in. Finally, this model implies that words should necessarily be recognised once their offset has passed, or there are otherwise deemed unrecognised and unrecognisable and removed from the cohort. However, Grosjean (1985) showed that in some cases word recognition could only occur *after* word offset and not *at* word offset. This is the case for words that may constitute the beginning of another word. For example, in the sentence "I saw the **bun** in the store" (Grosjean 1985) the recognition of "bun" can only occur when the /ð/ of "the" is perceived. Indeed, the word "bunny" ("[...] bun in [...]") could be a possible word. Similarly, the /n/ could be the beginning of another coherent word (such as in "[...] bunny nibbling [...]"). However, once the /ð/ is perceived, the only option left is to analyse the sequence as "bun in the". Hence, word recognition might only occur well after the word's offset has passed, situation for which the COHORT model does not account for.

The REVISED COHORT model (Marslen-Wilson 1987*a*) relaxes some of the constraints of the first model. This model does allow for a word to be activated even if the perceived form does not match the internalised form. Nonetheless, only slight mispronunciations are tolerated and are able to activate a word (i.e. such as swapping /p/ for /b/ that only differ by one articulatory feature: voicing). Also, this version of the cohort relaxes the effect of context (*integration*) as "inappropriate words can [...] be readily perceived and identified, so long as they are unambiguously specified in the signal".[27] Therefore, in this version, top-down inhibition is removed so as to allow for non semantically or syntactically expected words to be recognised. However, to compensate for this top-down inhibition, this model incorporates the effect of word frequency, such that more frequent words may be activated

[27] The authors give the sentence "John slept the guitar" as example, where the listener is able to activate the word "guitar" and hence recognise it even if it is grammatically incorrect to use it in this context.

more readily. Yet, this version of the cohort does still not account for the fact that a large majority of words can only be recognised passed their offset and not at word offset.

1.4.2 TRACE

The TRACE model builds upon the COHORT model by conserving its "major positive features" (McClelland & Elman 1986*a*), which are simultaneous word activation and interactive selection and pruning, while trying to overcome the problems already discussed there above. The TRACE model is structured in three layers, where each layer represents a particular linguistic unit: *feature*, *phoneme*, and *word*.

The *feature layer* constitutes one of the major differences with the COHORT model as COHORT supposes the input is perceived as a sequence of discrete units. Here, the acoustic input is perceived sequentially in terms of features: power, vocalic, consonant, voiced, etc.[28] Each time slice (25ms) of the speech signal is represented by these features that can take a value ranging from 1 to 8: 8 representing the highest activation possible, and 1 the lowest. This feature layer is connected to the *phone layer*. The phones that match to a certain extent the activated features are themselves activated. For example, /b/ would be activated by the features *voiced* and *burst* (among other features). This feature layer allows for several phonemes – that only differ by one or two features – to be activated at once by the same acoustic input. For example, a [b] sound would also activate /p/. Indeed /p/ and /b/ only differ by the value of the voicing feature, all the others being equal. However, the phoneme that does not exactly correspond to the acoustic input will be less activated than the one that exactly matches the input. Hence, this model, by simply adding a feature layer allows for mispronunciations to activate the mispronounced phoneme, which was not the case in the COHORT model, making the model much more flexible. Finally, the *phone layer* is connected to the *word layer*. The activated phonemes activate the words in which they appear. Contrary to the COHORT model, which gives the word onset a very high importance, in this model words can be activated from any point on. For example, the phoneme string /pɔɹt/ is able to simultaneously activate words such "port", "air*port*", "im*port*ant" or "*port*able".

Units inside the same layer (phoneme, or word) are linked through inhibitory connexions so that the units that are the most activated inhibit to a certain extent the activation of the others if they do not fit the input as well as the most activated units. The proportion of inhibition is based on the proportion of overlap: "the strength of the inhibition between two word units depends on the number of time slices in which they overlap". A word is deemed recognised if its activation value is above a given threshold. Contrary to the COHORT model, this model is able to handle string of connected words. In this model, the speech stream is implicitly segmented when the words are recognised.

This model brings a major improvement over the COHORT or REVISED COHORT: words can be recognised after their offset, and words may be activated from any point on, thus relaxing the constraint on exact matching word onset the COHORT supposes. However, we believe the biggest improvement the TRACE model brings is the active competition between words, as words really compete between one another. Indeed, in the COHORT model, words are considered as competitors only because they belong to the same cohort. Yet, there is no active competition between words, that is, there is no process by which a very activated word inhibits the activation of another. In the TRACE model however, there is an active competition between words, where highly activated words inhibit the activation of other words, until the one that matches the input the most eventually "wins".

[28] For more details on the features see (McClelland & Elman 1986*a*, p. 15).

1.4.3 Shortlist

The computational implementation of the TRACE model, though functional, made it unrealistic. Indeed, theoretically, this model supposes that for each time slice, each and every word of the vocabulary is considered a potential match, even if it hardly matched the spoken input. Hence, the computational load — both in a computational implementation and for a human brain — makes this model unrealistic when using a large vocabulary. The SHORTLIST model (Norris 1994) is a two-stage model that circumvents this limitation. The first step resembles the COHORT model, as it consists in building a shortlist (i.e. akin to a cohort) of words that match the current input. The second stage is a competition phase akin to that found in the TRACE model.

At each time step, words that begin with the current phoneme as well as words that start with the previously perceived phonemes are activated and added to the shortlist: taking the word /kæt/ as an example, the initial /k/ activates a list of words that start with [k], the following [æ] activates words that start either with [æ] or [kæ], etc. (Norris 1994, p. 204). The words figuring in this initial list are removed if their activation score — which represents how well the phonological representation they stand for matches the perceived acoustic input — is below a certain threshold. The second phase is the competition phase, where the words that most closely match the input compete between one another for recognition. As in the TRACE model, the proportion of competition is moduled by the overlap — here, in terms of phonemes — between the competing words.

Consequently, this model is computationally lighter than the TRACE model — as only a subset of the lexicon is considered a valid candidate — while preserving most of its original features. The SHORTLIST model however is purely bottom-up (i.e. there are excitatory connexions between the phonemic level and lexical level), but does not implement the top-down excitatory (from the lexical level to the phonemic level)[29] that is found in the TRACE model. They deem this constraint "redundant" with lexical inhibition. Nonetheless, this decision does not conform with reality, as lexical processing influence phoneme perception (see Warren (1970) for the phoneme restoration effect or Ganong (1980) for the so-called "Ganong effect"). Furthermore, contrary to the TRACE model that operates principally on handcrafted pseudo-acoustic features, SHORTLIST operates on strings of phonemes. They also (see §10 in Norris 1994) introduce an input representation that is a bit more realistic than discrete phonemes: mid-class transcription (i.e. coarse-grained transcription: (un)voiced fricative/stop/etc.) to account for "uncertainty, or ambiguity in the input". Nevertheless, this solution is still less realistic than the feature input adopted by the TRACE model.

1.4.4 Distributed Cohort Model

The Distributed Cohort Model (DCM) sets apart from the previously presented models (Gaskell & Marslen-Wilson 1997).[30] While the preceding models explicitaly incorporated several levels of computation (feature layer, phone layer, lexical layer), the DCM does not. Indeed, the DCM uses a simple recurrent neural network (Elman 1990, see Section 2.2.3) where information is distributed over several processing units. Hence, instead of using discrete representations, this model uses continuous representations. Even though artificial neural models are structured in layers, these layers cannot be interpreted as representing solely a single type of linguistic unit, where for example the first layer would represent features, the second phonemes, etc. Instead, each layer might represent several linguistic

[29] Called "feedback" in the original TRACE paper.

[30] It also sets apart from other word recognition models not presented here, see Weber & Scharenborg (2012) and Magnuson et al. (2012) for an extensive review.

units at once; the network learning on its own which is the best representation that each layer should output in order to solve its task.

The network is inputted with a sequence of phonetic features and is trained to predict a semantic vector and a phonological vector. While the target phonological vector was constructed based on the phonemes of the target word, the target semantic vector is simply a vector of ones and zeros randomly assigned to each word of the training set. Given its nature, the model is purely bottom-up, and consequently no top-down contraints were implemented. However, given its recurrent nature, the prediction at a given timestep depending on the previous ones, this constraint is somehow implicit in the network.

In this model, the concept of activation is different from the other models. Indeed, while in the other models a scalar value could be attributed to each word of the lexicon — scalar value representing the strength of the activation — it is not possible here to have such value. Instead, activation is represented by the closeness of the predicted semantic vector to that of the words of the lexicon. Simultaneous activation of several words is still possible, but is however implicit. Indeed, the predicted semantic vector may be considered as "a 'blend' of the relevant representations". Activating a word more than another one can be thought of "modifying this blend" and "can be viewed in terms of movement through semantic space". A word is deemed recognised when the semantic vector corresponds exactly to the semantic vectors of one of the words of the lexicon.

An interesting property of this model is that it is easy to test if word frequency influences recognition. Indeed, to do so, it is only necessary to let the network see more instances of one word than another at training time. The authors find that their network tends to favour more frequent words by predicting a semantic vector which is closest to the most frequent word, particularly when the number of competitor words is high, that is, when many words start with the same sequence of phonemes. This behaviour is coherent with previous research: Goldinger et al. (1989) indeed showed that more frequent words are more easily recognised than less frequent words.

Hence, in such model, the notion of feature, phone, and lexical unit is implicit as all the layers are able to reprensent this information simultaneously. Similarly, simultaneous activation of lexical units is pervasive, by the prediction of a semantic vector which "blends" the representation of possible words. However, in their experiment, they only test the recognition of individual words and not full sentences.

1.4.5 Conclusion on Word Recognition

All word recognition models assume that word recognition proceeds in several steps: an activation, a competition, and the final recognition. Nonetheless, the way these three steps are carried out varies from one model to another. While the Cohort model requires the acoustic input to exactly match the internalised representations, other models, such as Trace and Shortlist do not, and allow for a word to be activated even if the internalised representation does not exactly match the perceived acoustic realisations. The competition step also varies. In the Cohort, there is no active competition between words as the degree of activation of a given word does not influence that of others. This is not the case in the other models where there is a true competition for attention between words.

The DCM sets apart from the other models and shows that word recognition may be done using continuous representations only, and does not need, as for the other models, for the intermediate representations to be discrete (i.e. phones). This model is relatively close to recent models of speech recognition which rely on RNNs to decode speech.

However, all these models agree on the fact that word recognition necessarily implies the simultaneous activation of a set of word candidates. This set of words is iteratively pruned

if the acoustic input is too distant from the internalised representation so that ultimately only one word remains.

1.5 Conclusion

In this chapter, we reviewed the strategies used by children to segment the spoken input into sub-units. We showed that children use a wide range of strategies to find words. They first start by using supra-segmental cues (i.e. prosody) as they are very sensitive to variation of the fundamental frequency from the moment they are born. We also showed that children hear a high proportion of words in isolation. Using these two strategies, they are able to infer the canonical shape of the word-forms of their native language, so as to later apply more advanced segmentation strategies. Indeed, they are able to compute forward and backward transition probabilities between units in order to posit word boundaries, or use phonotactic rules in order to extract segments.

We then presented several word recognition models that try to account for how humans are able to activate and retrieve lexical units from the mental lexicon. All word recognition models agree on the fact that a successful recognition strategy consists in simultaneously activating several lexical units at once. Yet, the process by which these lexical units are activated varies from one model to another. While some models require the acoustic input to exactly match the internal representations, some are less strict, and allow for mispronunciations.

Finally, we showed that language acquisition is a multimodal phenomemon that involves all of the child's perceptual abilities. We showed that intention reading and shared attention were crucial components for the child to acquire its native language. Lastly, we showed that vision provided the child with invaluable information to acquire her mother tongue. Children that are deprived from this perceptual input, even though they are eventually able to acquire their native language, are hindered, as it is harder to conceptualise the world when such perceptual input is lacking.

Chapter 2

Background on Speech Processing and Term Discovery

Contents

2.1 Introduction

2.1.1 Unsupervised Speech Processing

The work we conduct in this thesis fits into the scheme of unsupervised speech processing. Speech Processing has always required a large amount of human supervision, for example in Automatic Speech Recognition (ASR), in the form of the priors incorporated into the models themselves — when for example using Hidden Markov Models (HMM) and Gaussian Mixture Models (GMMs) which explicitly model phonemes (through GMMs) or syllables

(through the use of triphone-based HMMs) — and more importantly in the form of manual transcriptions. However, gathering and having human experts transcribe and annotate data is a time-consuming process which requires expert-knowledge and is costly. Also, not all languages have a written form, or some may have several non standardised written forms, which make current approaches not usable on such languages. Hence, while traditional approaches are sustainable for developed languages,[1] they are not for less-developed languages.

Another motivation, which also constitutes the motivation of the work carried out in this thesis, is to devise speech processing models that are closer to human speech processing. Indeed, humans do not require much supervision to learn how to process speech, especially when acquiring their native language. Notably, they do not require textual labels, but instead use weak supervision signals, such as visual cues. These signals constitute weak supervision signals as they constrain the learning process by contextually grounding it.

The work carried out in this thesis belongs to the sensory-based approaches of speech processing models, as defined by Glass (2012). Sensory-based models of speech processing are models that only require speech paired with sensory data in order to operate. That is, such models do not require any annotated data nor human expertise to operate and "closely match that of human spoken language acquisition" according to Glass (2012).

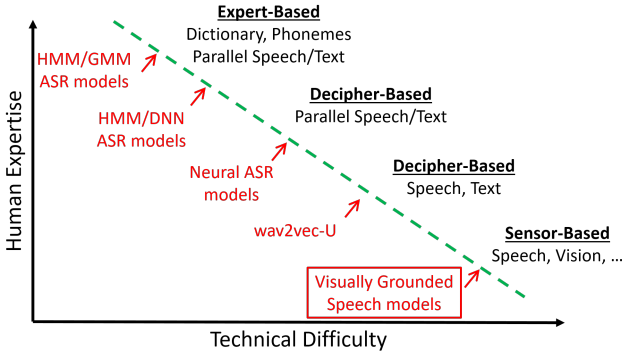


Figure 2.1: Unsupervised Speech Processing Hierarchy. Figure reproduced from Glass (2012) except for the annotations in red which we added. The boxed text shows where the work conducted in this thesis fits into this hierarchy.

In the following sections of this introduction, we will present models that also belong to the class of sensory-based approaches to speech processing. We will focus more specifically on models targeted on unsupervised term discovery and speech segmentation. We will then argue that grounding speech to another modality is necessary if one wants to be able to associate a meaning to the extracted word-forms, and hence be able to devise computational models of lexical acquisition.

2.1.2 Spoken Term Discovery and Speech Segmentation

Most of the work done on unsupervised speech processing aims at segmenting an audio signal into sub-units — generally phones or words — or without fully segmenting the audio, aims at finding large lexical units, such as words or multiple word expressions.

[1] "developed" should be understood in the sense given by Ferguson (Fishman et al. 1968, p. 31-32), see https://www.ethnologue.com/language-development.

Lee et al. (2015) distinguishes between two terms: *spoken term discovery*, and *unsupervised lexicon discovery*.[2] The first term, *spoken term discovery* simply refers to a keyword spotting task. That is, given a spoken utterance, the goal is to find similar acoustic segments across a collection of utterances. According to Lee et al. (2015), this task typically involves using "unsupervised pattern discovery methods" such as Dynamic Time Warping (DTW, Sakoe & Chiba 1978). This task only aims at finding keywords, or a least chunks of words, but does not seek to obtain a *full segmentation*, where the boundary of each word in the speech stream would be marked. In such keyword spotting task, the goal is only to find one or more keywords, which results in the discovery of two boundaries — one at the beginning of the word, one at the end — leaving the rest of the speech stream unanalysed. *Unsupervised lexicon discovery* refers to the process which results in a complete segmentation where ideally all word boundaries are found. According to the authors, such task usually implies a two stage modelling, where a model simultaneously learns a sub-word (i.e. syllabic, morphemic, etc.) distribution and a word distribution. Such two stage modelling generates top-down and bottom-up constraints that result in a better segmentation overall, as it explicitly takes into account the hierarchical linguistic structure of the input.

Spoken Term Discovery. Park & Glass (2005) introduced segmental dynamic time warping (S-DTW) for unsupervised term discovery. This method is based on dynamic time warping which is a pattern similarity measure between two sequences (Rabiner & Juang 1993, p. 226). While DTW operates at a global scale, giving the similarity measure of two sequences globally, S-DTW operates at a local scale, giving a similarity measure for two sub-sequences taken from two larger sequences. The idea behind S-DTW is simple: even though two spoken utterances are different, they might contain similar words. Hence, even though the similarity measure might be low at a global scale, it might be high at a local scale. Such methods allow to detect and eventually extract word-like units from raw acoustic input without the need to result to a transcription at any point (Park & Glass 2008). Jansen & Van Durme (2011) further improved on vanilla S-DTW by introducing algorithms with lower complexities, making S-DTW more scalable on large data sets. Recently, Räsänen & Blandón (2020) proposed a two-staged probabilistic variation of S-DTW making S-DTW usable on larger data sets.

Finally, we should mention the work of Räsänen et al. (2015) which introduced a method for unsupervised term discovery without the need for S-DTW, by first slicing the input speech signal into syllables — by considering local minima in the amplitude envelop as boundaries — and then grouping the syllables together by finding frequent n-grams.

Unsupervised lexicon discovery. The goal of unsupervised lexicon discovery is to obtain a full segmentation, and this usually involves Bayesian approaches which model several levels of linguistic units at once. These methods are, to a large extent, inspired by prior work on dealing with the segmentation of strings of discrete units (Goldwater 2006, Johnson et al. 2007). Taniguchi et al. (2016) introduced the Bayesian Double Articulation Analyser that aims at learning in an unsupervised fashion a language model and an acoustic model using a non-parametric Bayesian approach. Lee et al. (2015) combined unsupervised acoustic unit discovery with Bayesian segmentation in order to unsupervisedly segment a speech stream into lexical units. Kamper (2017) propose a model that jointly segments and clusters, where first, boundaries are sampled, the resulting segments are then embedded and clustered by a Bayesian GMM, which in turn gives information on the likelihood of the initial segmentation, which is used to sample better boundaries.

Recently, neural architectures were proposed to solve this task; most of them using auto-encoders. For example, Bhati et al. (2020) proposed a new auto-encoder architecture that

[2] Actually, a distinction is made with a third term, *word segmentation*, which is similar to unsupervised lexicon discovery, but carried out on string of symbols (e.g. graphemes or phonological transcriptions).

allows to segment the speech stream into phone-like units which are then clustered. Their idea relies on unsupervisedly learning similar embeddings for audio segments that belong to the same phone unit. Similarly to Räsänen et al. (2015), words are recomposed using a n-gram strategy. Chen et al. (2019) propose an auto-encoder in which discrete boundary decisions are taken at each time step. The intuition being that if the models take good boundary decisions (i.e. place boundaries near true boundaries), the reconstruction process should be facilitated. A similar approach was also proposed by Elsner & Shain (2017).

2.1.3 Grounding Language

A surprising fact of the previously mentioned approaches to unsupervised term discovery and lexicon discovery, is that word spotting and segmentation are only done on the basis of form alone, without any other contextual cue being used. To this extent, these works are very reminiscent of the work carried out in linguistics that also abstract meaning from the segmentation task (see Section 1.2). This is even truer since the creation of the *Zero Recource Speech Challenge*.[3] The original goal of this challenge is "unsupervised discovery of linguistic units from raw speech in an unknown language" without using any linguistics resource — hence the name — such as phonetic or orthographic transcriptions. The motivation of such challenge is twofold: first, a technological one, which is to develop methods to handle languages which do not have a written form (or language that might have one, but which do not have a standardised orthography); and second, a psycholinguistic motivation, which is to reproduce the cognitive processes at work in the human child when she learns her mother tongue, as she does so without any supervision.[4] Even though the website mentions the term *zero resource* should be understood in the sense of "zero *linguistic* resource" and not "zero information besides audio (visual, limited human feedback)", the only data which is given to the participants is audio data. Hence, the entries to the contest rarely use any other external resource than the one provided, and consequently rely solely on form.

When examining the literature, we notice that the words *term* and *lexicon* in *unsupervised term discovery* and *lexicon discovery* are used as synonyms for *word-form*, when from a linguistics perspective, a *term* or a *lexicon* is much more than a word-form (or a list of word-forms respectively), but also includes information about meaning (among other information). Language is not an independent system, and what gives it its substance is that it is tied to the physical world. We used the example of blind children to show that, when deprived from one type of sensory input, namely vision, language acquisition was substantially affected. Notably, blind children show a propensity to repeat phrases and sentences tied to a particular context (see 1.3.5.2) without further analysing them. To draw an analogy, term discovery and segmentation models are actually trained as sensory-deprived toddlers, that are only trained on form. Word segmentation might lead to erroneous results if done without contextual information: multi-word expression might be considered as a sequence of several words instead of unique words (e.g. French for *potato* "pomme de terre"). The decision as to whether these words should be considered as one unit or as several can only be done on the basis of contextual information, that is, their linguistic referent. Similarly, the phonological system of a language can only be inferred in light of the surrounding context and whether the two word-forms that use these similar sounds have the same referent or not. Therefore, it seems an adequate speech segmentation strategy should include a

[3] https://zerospeech.com/

[4] The pscholinguistic motivation tends to disappear in the recent editions (2020, 2021), but is explicit in the first edition (2015): "[provide] adaptable algorithms that [...] aid infant language acquisition research by providing scalable quantitative models that can be compared to psycholinguistic data" Versteegh et al. (2015).

meaning component. In their work, Bender & Koller (2020) argue that:

> if form is augmented with grounding data of some kind, then meaning can conceivably be learned to the extent that the communicative intent is represented in that data

Consequently, grounding linguistic data to external knowledge seems to be a necessary step to gain access to meaning. Yet, what does *grounding* more precisely mean? If grounding is simply adding external knowledge, ASR can be thought of textually grounded model, as text, which is external knowledge, is added to the model. Roy (2005) defines grounding as follows:

> The relationship between words and the physical world, and consequently our ability to use words to refer to entities in the world, provides the foundations for linguistic communication. Current approaches to the design of language processing systems are missing this critical connection, which is achieved through a process I refer to as *grounding*. [...] *Language grounding* refers to processes specialized for relating words and speech acts to a language user's environment. (Roy 2005)

Grounding thus implies adding to linguistic data external *non-linguistic* data which somehow reflects the physical world. The external data should also reflect to some extent the communicative intent in the linguistic data as mentioned by Bender & Koller (2020). This latter constraint is also visible in Roy (2005)'s definition, as "speech acts" necessarily result from a communicative intent. Grounding language to another source of knowledge acknowledges the simple fact that language cannot be considered *in vacuo* – that is, as a whole complete in itself and disconnected from any context – but is actually produced *in* and tied *to* a particular context. Even though considering language *in vacuo* does not preclude from learning anything, accurate segmentation requires access to meaning. It should be noted that grounding language to another source of information does not necessarily imply a change in modality. For example, it is possible to ground the word-form /kaʊ/ ("cow") to the sounds [muː] ("moo") in which case it is the nature – linguistic v. non-linguistic – of the acoustic stimulus that changes, but not the modality itself.

Hence, *grounded* models refer to a class of computational models that process some linguistic form — either text or speech — in conjunction with another source of information from the physical world. The fact that two modalities are processed in conjunction is necessary, but is not sufficient: both modalities should occur simultaneously in the physical world so that both are tied. Thus, grounded models do not process form alone, but are able — or at least given the ability — to link this form to referents in the physical world, or representation thereof.

2.2 Background

In the following section, we will present computational models that are able to ground language, and more specifically speech, by using visual data. We will show that such models constitute viable test beds to study lexical acquisition. But first, we will present artificial neural networks, which are used to build visually grounded computational models.

2.2.1 Machine Learning

Machine Learning (ML) is a sub-field of Computer Science concerned with the creation and "study of computer algorithms that improve automatically through experience" (Mitchell

1997, p. xv). Instead of designing a specific algorithm to solve a specific task, ML seeks to design algorithms that *learn* how to solve a specific task:

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. (Mitchell 1997, p. 3)

Experience E consists of a set of examples — called a data set — which the program will learn from. Learning occurs through a process of trial and error where the program tries to solve the task T using the available data E and changes its parameters so as to increase its performance measure P.

Among the several learning schemes that exist, two important ones can be distinguished: *supervised learning*, and *unsupervised learning*. We use the latter in this thesis. In the supervised case, the data set consists of a set of annotated (i.e. labeled) examples and the task of the algorithm is to learn the relationship between the examples and their labels. An example of such a task would be to learn to segment a string of characters given a training set where each character is labelled so as to indicate if it constitutes a boundary or not. In the unsupervised case however, the data set would consist of unannotated examples and the latent structure of the data should be inferred by the algorithm. For example, in this case the task would be to infer word boundaries from the string of characters alone.

Hence, contrary to the unsupervised case, where the model is only given raw data, in the supervised case, the data is augmented with supplementary labels which will be the learning target of the algorithm (Goodfellow et al. 2016, p. 105). More formally, Goodfellow et al. (2016, p. 105) define unsupervised and supervised learning as follows: given a set of examples $\mathcal{X}$, unsupervised learning consists in learning the probability distribution $p(x)$ given several random vectors $x \in \mathcal{X}$, while supervised learning consists in learning to predict a set of values y given several random vectors x and their associated label y where $(x, y) \in \mathcal{X}$, which can be interpreted as learning $p(y|x)$.

Goodfellow et al. (2016, p. 105) argues that the lines between supervised and unsupervised learning is blurry. It is the case for example of *self-supevised learning*. In this case, the targets that the model should predict are drawn from the input data itself. That is, the input data acts as its own label. Given a set of examples $\mathcal{X}$, it could be more formally defined as learning $p(x)$ where $x \in \mathcal{X}$ by learning to model $p(y|x_{\setminus y})$ where $y \in x$. This is the case of most recent language models (BERT), audio embeddings models (wav2vec), or unsupervised learning of visual representations.

The data set used for learning is commonly split in three uneven sets: the training set which is the largest ($\approx 80\%$), the development (or validation) set ($\approx 10\%$), and the test set ($\approx 10\%$). The training set, as its name suggests, is used to train the model. The development set is used to test the model so as to select the best model checkpoint, while the test set is only used once the best model is selected to report the final score.

2.2.2 Artificial Neural Networks

Artificial Neural Networks (ANN) are a type of ML algorithms which are based on the work of the the American psychologist Rosenblatt (1958). ANNs belong to the connectionist approach of Artificial Intelligence (AI):

> These models assume that information processing takes place through the interactions of a large number of simple processing elements called units, each sending excitatory and inhibitory signals to other units. (Rumelhart, McClelland & PDP Research Group 1986, p. 10)

The central idea of connexionism is that a complex behaviour can be approximated by a complex computation, that can in turn be broken down into smaller and simpler computations realised by individual processing units. The inter-connexion of these units (and hence computation) gives rise to the global behaviour. Connectionism was also referred to as "parallel distributed processing" (PDP), where knowledge is distilled and distributed over each processing units that constitute the network.

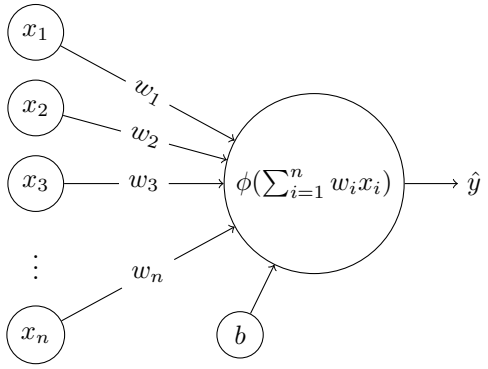


Figure 2.2: Example of an artificial neuron

These processing units, originally called *perceptrons*, are now commonly called *neurons* and may be represented as in Figure 2.2. A neuron receives n inputs $X = x_0, x_1, ..., x_n$ from which an output $\hat{y}$ is computed. Given the input features, the neuron is trained to predict an output $\hat{y}$ that should be as close as possible to the desired output y. In order to do so, the input vector is weighted by a set of weights $W = w_0, w_1, ..., w_n$ which are learnable parameters, and the output is further transformed by a non-linear activation function ϕ. A bias neuron may also be added.[5] The final computation of a single neuron can thus be summarised as follows:

$$\hat{y} = \phi(W \cdot x + b) = \phi(\sum_{i=1}^{n} w_i x_i + b) \tag{2.1}$$

Real world problems are too complex to be solved by a single neuron. Thus, an arbitrary number of neurons can be used in conjunction so as to learn complex transformation functions. These neurons are grouped into *layers* and form what is commonly referred to as a *Multilayer Perceptron* (MLP). The first layer is called the *input layer*, the final layer *the output layer*, and the one or more layers in-between are called the *hidden layers*. The topology of the network is the result of how the different layers are connected. There are several ways to connect consecutive layers of neurons, the simplest being by connecting all the neurons of a layer to all the neurons of the next layer (*fully connected layers*). However, this is not the only option: it is possible to introduce cycles (*recurrent neural networks*) where the output of a layer serves as input for the same layer; or a layer can be connected to the next layer but also the second-next or n-next layer (*skip connections*); or only partially connect one layer to the next (*sparse connections*).

2.2.3 Recurrent Neural Networks

Theoretically, MLP can handle any type of data provided that the input layer is large enough to fit it. For example, it is possible to train a simple image recognition model that

[5] Similarly to the b term of a linear regression $y = ax + b$, where b controls the y-intercept of the function.

uses a MLP provided the size of the input image is fixed and known in advance (e.g. only 32×32 pixels images). In such cases, the data is *flattened* to a one-dimensional vector (e.g. 1×1024 for a 32×23 grey-scale image) which is then fed to the MLP. Such solution might be satisfactory for simple problems, but it is unusable when handling sequential data, as usually the size of the sequence to be processed is unknown.

Recurrent neural networks (RNN) are neural networks that are able to process sequential data (Elman 1990). RNN can be considered as feed-forward networks augmented with feedback connections.[6] RNN can be formalised as follows:

$$h_t = \phi(Wx_t + Uh_{t-1} + b) \tag{2.2}$$

where h_t is the hidden state at timestep t, x_t is the current input, h_{t-1} is the previous hidden state, b is a bias term, ϕ is a non-linear activation function (usually sigmoid or hyperbolic tangent), and where W, U and b are learnable parameters. Note that the only difference with a feed-forward layer is simply the additional term Uh_{t-1}: the output at a time step t does not only depend on the current input but also on the output at the previous time step $t-1$. Thus, such computation allows to model the temporal dependency that exists between consecutive vectors. The output of a RNN consists of a sequence T of vectors. The final vector of the sequence (at timestep T) can be viewed as the compact representation of the whole input sequence, as the final vector depends on the computation of all the previous vectors of the sequence. When the first element of the sequence x_1 is processed, h_{t-1} does not exist. In such case, this previous hidden state, noted h_0, is set to be a vector of 0, though in some cases the initial state might also be a learnable parameter of the network.

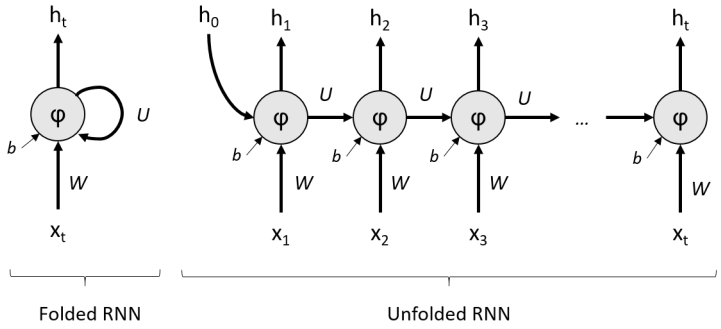


Figure 2.3: Illustration of a RNN folded (right) and unfolded over time (left). The diagram is annotated according to Equation 2.2. Note the initial state h_0 which is added for the first timestep.

Note that processing a sequence with an RNN is different from running a simple feed-forward network over each vector of the sequence. Indeed, in such case, the output at a given time step would be totally independent of the previous output. Such computation would not be able to model the sequential nature of the data, and would be a simple projection of the input vectors. Hence, the feed-back connection of RNNs really constitute the backbone of this neural model.

Even if RNNs are able to model sequential data and keep track of local dependencies relatively well, they are unable to capture long term dependencies in the input sequence. Indeed, as the sequence size T grows longer, the vector h_t contains less and less information

[6] as defined by Hochreiter & Schmidhuber (1997).

about the beginning of the sequence. The information about the past fades away as the previous hidden vector h_{t-1} is combined at each timestep with new information. Gated units were introduced to solve this problem.

2.2.4 Gated Recurrent Units

LSTM (Hochreiter & Schmidhuber 1997) are a type of gated units and are able to learn both long term and short term dependencies.[7] LSTM are equipped with a *memory cell* that is used to store information about the past timesteps. Two gates (the *input gate* and the *forget gate*) control which information is added to the memory cell and which information should be removed. Such gating mechanism thus enable the LSTM cell to keep track of both long term and short term dependancies by adding to the memory cell information that needs to be preserved over several timesteps. However, LSTM are hard to train as they have many parameters, notably because of the parameters used for the memory cell and additional gates.

Cho et al. (2014) introduced Gated Recurrent Units (GRUs) that are also able to keep track of long term and short term dependencies such as LSTMs, but with fewer parameters which make them easier to train. A GRU is formally defined as follows:

$$\begin{aligned} z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\ r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\ \hat{h}_t &= \tanh(W_h x_t + U_h(r_t * h_{t-1}) + b_h) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \hat{h}_t \end{aligned} \tag{2.3}$$

where $W_z, U_z, W_r, U_r, W_h, U_h$ are learnable matrices, b_z, b_r, b_h are bias terms, and σ represents the sigmoid activation function. A GRU has two gates: an *update gate* z, and a *reset gate* r that both independently compute a scalar value at each time step t (z_t and $r_t \in [0, 1]$). Given the current input x_t, the value of z_t represents a ratio between how much information should be carried on from the previous time step $[(1 - z_t) * h_{t-1}]$ and how much new information should be integrated $[z_t * \hat{h}_t]$. This gate allows to control for long term dependencies. r_t represents how much the new state should depend on the previous time step $[r_t * h_{t-1}]$ thus controlling for local dependencies.

Even though LSTM and GRU are able to keep track of long- and short-term dependencies, encoding the meaning of a whole sequence in a single vector is still challenging and the recurrent unit might forget the beginning of the sequence, especially when the sequence is particularly long. One solution is to encode the sequence from both ends, where one GRU processes the sequence from left to right and another GRU processes the input sequence from right to left. The final vector is a concatenation of both vectors. Such processing is know as *bidirectional* processing, as opposed to *unidirectional* processing (which is usually done from left to right).

2.2.5 Attention Mechanism

Even when using bidirectional recurrent cells, encoding the meaning of a whole sequence in a single vector remains a difficult task. Attention Mechanisms (Bahdanau et al. 2015) were introduced to solve this problem. This solution was originally thought for machine

[7] LSTM were introduced to solve the *vanishing/exploding gradient problem* which we will not discuss here. However, the fact that vanilla RNN cells are unable to keep track of long term dependencies is a direct consequence of the vanishing gradient problem.

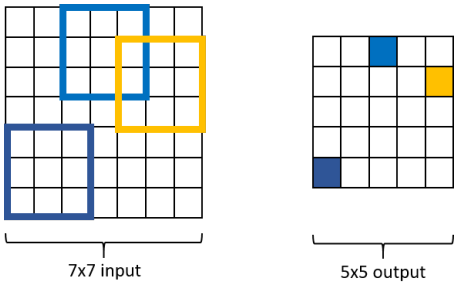


Figure 2.4: Illustration of a 2D-convolution on a 7×7 input producing a 5×5 output when using 1 kernel of size 3×3 sliding from left to right from top to bottom with a stride of 1, and zero padding. The convolutional kernel is depicted with borders of different colours at different places; the result of the convolution is shown by a shaded box with the same colour. For illustration purposes, we only depicted three positions where the kernel could be, when naturally the convolutional kernel scans the whole input. Usually more than one kernel is used.

translation purposes, but found a wide echo and is now used for many other tasks. The intuition of such attention mechanisms is simple: instead of keeping the last vector of the sequence — which should encode the meaning of the whole sentence — all the vectors that were computed at each time step are kept and a weight is assigned to each of them so as to give more importance to some vectors than others.

The weights are computed by the network itself and learnt at training time. Hence, the network learns which vector(s) in the input sequence should be given more importance in the final representation without any supervision. An attention mechanism can be formalised as follows:

$$\begin{aligned} c &= \sum_{t=1}^{T} \alpha_t h_t \\ \alpha_t &= \frac{\exp(\text{score}(h_t))}{\sum_{t'=1}^{T} \exp(\text{score}(h_{t'}))} \end{aligned} \tag{2.4}$$

where c, the context vector, is the weighted sum of the hidden vectors, and α_t is the attention weight for time step t, where $\alpha_t \in [0, 1]$ and $\sum_{t=1}^{T} \alpha_t = 1$. The formalisation we present is very general: the scoring function *score* (usually a MLP) computes a scalar value for each h_t. However, the precise way this scalar value is computed depends on the particular implementation of the attention mechanism and may not only be a function of h_t.

With such attention mechanism, vectors for which the attention weight α is high will be highly represented in the final vector c, while vectors which were assigned a low attention weight won't.

2.2.6 Convolutional Neural Network

Convolutional cells (Fukushima 1980, LeCun et al. 1999) were originally introduced for image processing. As already mentioned, even though MLP may be used to process images, they become unusable in practice when the size of the image grows, as the size of the input

of the MLP has to fit the whole image (e.g. 32M for a 32Mpixel image). This makes MLP unusable in practice on such large images. Also, as the input data is flattened, information about which pixel is adjacent to which in the original becomes harder to pick up.

Convolutional cells enable to solve this problem by repetitively applying a *convolutional kernel* (also called a *filter*) to the input image. The image is hence processed piece-wise, one portion at a time. While such kernel processing was known before neural networks,[8] the convolutional neural network allows for the kernels to be learnt: the kernel becomes part of the trainable parameters of the network. It hence allows the network to learn the kernels, as usually more than one kernel are used, so as to best solve the task the network is trained for. Convolutions are also easy to implement, as they just consist in matrix multiplication: the sub-matrix that represents the portion of the image being processed and one for the convolutional kernel.

Three types of convolutions exist, depending on the axes on which the kernel slides: 1D convolution (where the kernel slides along one axis), 2D convolution (slides along 2 axes), and 3D convolutions (along 3 axes). For our purpose here, we will only consider 1D convolutions as they constitute the most used type of convolutions for speech processing[9] when working with MFCC vectors; such as we do in this thesis.

Convolutions are parametrised by:

- k the number of convolutional kernels used to scan the input (in Figure 2.5 on the *Kernel* Axis, $k = 4$). If only one kernel is used, the output of the convolution will be a 1D vector of length S_{out} if more than one kernel is used, the output will be of dimension $S_{out} \times k$.

- *size* which is the height of the kernel which represents how many input vectors are processed at once (in Figure 2.5, $size = 2$). In 1D convolutions, the length of the kernel usually is the same as the length of the input vectors (in Figure 2.5, the lenght of the kernel is equal to the dimension of the MFCC vectors: *13*); though depending on the implementation, the convolutional kernel can be split over several groups that only see one part of the vectors.[10]

- *stride* represents the shift along the Time axis (in Figure 2.5, $stride = 1$), meaning the kernel will slide along the Time axis by moving by *stride* vectors at each step.

- *padding* represents how many blank vectors are added before and after the actual vectors to be processed. This allows to control for the output size of the sequence S_{out} given the original input size S_{in}.

The length of the output sequence S_{out} can be computed as follows:

$$S_{out} \approx \lfloor \frac{S_{in} - \text{size} + \text{padding start} + \text{padding end}}{\text{stride}} \rfloor + 1 \tag{2.5}$$

where *padding start* refers to the number of padding vectors added before the input sequence, and *padding end* to the vector added after the input sequence; the precise number of vectors added at each end depending on the padding mode used.

1D-convolutions are widely used for speech processing as each convolutional kernel aggregates several vectors along the time dimension into a single value. This is particularly useful for speech, as each phone of the speech stream straddles over several MFCC vectors

[8] Where each type of kernel (edge detector, blur, etc.) was hand crafted. See example https://en.wikipedia.org/wiki/Kernel_(image_processing).

[9] Though 2D convolution may also be used when working with spectrograms.

[10] see https://pytorch.org/docs/stable/generated/torch.nn.Conv1d.html#torch.nn.Conv1d.

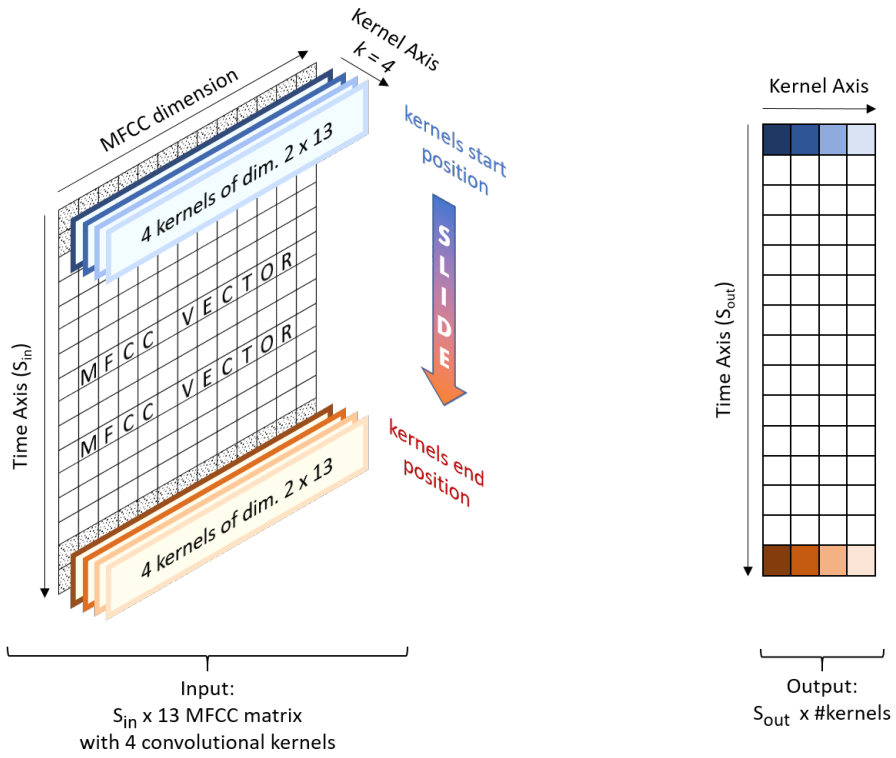


Figure 2.5: Illustration of a 1D-convolution on a $S_{in} \times 13$ input MFCC matrix producing a $S_{out} \times 4$ output (see Equation 2.5 on how to compute S_{out}) when using 4 convolutional kernels of size 2×13 sliding along the *Time* axis with a stride of 1, and zero padding. The position of the convolutional kernels are depicted with borders of different colours — blue and red — to materialise the evolution of their position over time; and the result of the convolution operation is shown by a shaded box with the same colour in the output matrix. For illustration purposes, we only depicted two positions where the kernels could be (at *kernels start position* and *kernels end position*), when naturally the convolutional kernels scan the whole input. We also only depicted 4 kernels — shown on the *Kernel* axis, with different shades of the same colour — when usually more are used.

— which usually represent 10ms of the original input signal, that is, less than the length of a phone. The use of different kernels also allows the network to select different features in the original input, and encode different correlations between the input frames. Also, the convolved input is usually shorter than the original input given appropriate choice of the kernel size, stride, and padding (the longer the stride and the kernel, the shorter the output sequence). This allows for the network to learn to properly downsample the input, and hence reduce the computational load, particularly when using recurrent layers afterwards.

2.2.7 Loss Function and Backpropagation

The perceptron, and neural networks in general, have a biological motivation (i.e. the nervous system) and authors regularly use biological terms (e.g. "stimuli", "excitatory/inhibitory inpulse", see Rosenblatt 1958) to describe how such models work and compare them to biological systems. Nevertheless, artificial neurons widely differ from their biological counterparts in the way the computation is realised, in their organisation, and in the way actual learning takes place: *gradient descent*. Using gradient descent as a learning

scheme imposes certain constraints when designing the network: every operation should be differentiable.[11]

The goal of the network is given an input x to predict an output $\hat{y}$ which is as close as possible to the desired (true) output y. The success of the operation is measured using a loss function (see Equation 2.6). During training, the network is encouraged to minimise the difference between the predicted output and the true output, and the value computed by the loss function will serve as the basis to update all the weights of the network so as to reach its goal.

$$\mathcal{L}(y, \hat{y}) = |y - \hat{y}| \tag{2.6}$$

Because the output $\hat{y}$ directly depends on the output of the network, which in turn depends on the parameters of the network θ, it is possible to change them so as to tune the network so that the predicted output is closer to the desired output. This operation is done through *gradient descent* using the backopropagation algorithm (Rumelhart, Hinton & Williams 1986). This operation consists in computing the derivative of the loss function with respect to the parameters of the network. Intuitively, this measures how responsible a given weight is for the discrepancy between the predicted output and the expected output:

$$\theta_{t+1} = \theta_t - \eta \frac{d\mathcal{L}(y, \hat{y})}{d\theta} \tag{2.7}$$

where the weights at θ_{t+1} are updated by a factor η which is called the learning rate.[12] The learning rate controls the strengths of the updates of the network's weights. If it is set too high, the network might miss the optimal solution, while a too small learning rate will make learning slower and additionally may have the network stuck in a local optimum. Note that the equation 2.7 corresponds to *stochastic gradient descent*, which corresponds updates the weights of the network based on the loss value of one training example (see Goldberg 2017, p. 31). Even though this approach is effective, it might take time to reach the global minimum, hence, usually batched gradient descent is used, where the loss is averaged over a *batch* of examples, that is, over several examples randomly sampled from the training set.

Hence, for each backward pass (i.e. gradient descent) all the weights of the network are updated so as to try to minimise the training cost. Even though neural networks are biologically inspired, the way learning takes place is not cognitively plausible (forward pass and backward pass use the same path, when biological synapses are unidirectional; artificial neurons are organised in well defined layers, and not in a biological brain, etc.).

2.3 Visually Grounded Speech

In this section, we then present Visually Grounded Speech (VGS) models and the analysis methods that have been developed in order to understand what such models have learnt. Finally, we will examine to what extent VGS models are able to model lexical acquisition.

[11] This is not entirely true: for example, Reinforcement Learning does not require the loss function to be differentiable. However, in most cases, all the operations are differentiable.

[12] We here take the simplifying assumption that gradient descent is done for only one instance x and not a batch of items.

2.3.1 Models

2.3.1.1 The CELL Model

One of the very first models of Visually Grounded Speech (VGS) is the CELL (Cross-channel Early Lexical Learning) model developed by Roy & Pentland (2002) and Roy (2003). This model was explicitly developed so as to understand how the interaction of visual and auditory stimuli enabled lexical acquisition. The goal of the CELL model is to learn audio-visual mappings (called audio-visual prototype) between various objects and their spoken word-forms so as to constitute a proto-lexicon such as a child would. This model strives to implement components that reproduced known cognitive abilities such as short term memory (STM) and long term memory (LTM) so as to simulate the memory abilities of a child.

In order to learn audio-visual prototypes, the CELL model is inputted with a visual context (image) and its paired utterance. The data used to train the model was gathered specially for this experiment and consists of child directed speech uttered by a caregiver while a child was playing with various objects. The CELL model searches for recurrent audio patterns, which are then stored in the STM along with the object considered at the time the discovered pattern was uttered. The LTM then scans the STM so as to find repeated occurrences of a given audio-visual pair — called AV event — and if several are found, they are copied into the LTM. The AV events stored in the LTM are then further processed to remove spurious pairs. Indeed, some words may occur frequently (e.g. determiners) while not being linked to a specific visual context. Hence, spurious AV events are removed based on a mutual information criterion: the lower it is, the less the two modalities (i.e. speech and vision) are tied, and hence do not form a valid AV pair. If mutual information is above a certain threshold, the AV event becomes a lexical item which is stored permanently in the LTM.

The results show that the CELL model is able to learn semantically valid AV items (e.g. shoe, key, dog, doggie) and reports 85% semantic accuracy of the discovered pairs. Some of the AV items also contain non standard "words" — which are not counted as semantically accurate pairs — such as onomatopoeic sounds (e.g. barking, engine). This is an interesting result, as onomatopoeias are also acquired by young children and used both in perception and production (see Laing 2019). In order to compare the lexicon that would be acquired when no visual stimuli is used, the authors also ran experiments using a "blind" model, that is, a model that does not use visual context. The results are very different with 0% semantic accuracy. Most of the AV pairs in this setting only comprise onomatopoeias and recurrent patterns not linked to the object in consideration (e.g. "what you gonna do", "really good", etc.). Hence, the key finding of the CELL model is that visually grounding speech increases the accuracy of word learning and speech segmentation compared to a model that would operate on speech only; and this model reproduced some behaviours observed in children (such as learning onomatopoeias).

This model however made several simplifying assumptions, notably the fact that speech is perceived categorically in terms of phonemes[13] and images had to be pre-processed so as to first detach the background from the foreground, and then isolate the object from the image. Also, the visual input was deliberetaly simplified so that it only contained one object only, hence facilitating the task of the model.

[13]Children do in fact perceive speech categorically. However, categorical perception for children should be understood in terms of *phones* (e.g. [b] is perceived differently from [p], or [p] from [p˺] or [p^{h}]) but *not* in terms of *phonemes*. It is only later that they learn that the acoustic difference between [b] and [p] is meaningful — and are thus two phonemes /b/ and /p/ — while the acoustic difference between [p] and [p˺] and [p^{h}] is meaningless in English.

In a model similar to Roy & Pentland (2002), Yu et al. (2005) test if audio-visual mapping was enhanced by gaze information. To test this, they had a picture book in a foreign language read by a native speaker and recorded. The speaker also had a gaze tracker so as to pinpoint where in the picture they were looking while reading a particular sentence. Their model is inputted with a visual context (image from the picture book) and an acoustic stimuli (sentences from the story, in the form a string of phonemes). The model then tries to find repeated patterns in the acoustic stimuli and learns to pair them with the visual context using an EM (expectation maximisation) algorithm. They train their model on two different conditions: the intention-cued condition, and the audio-visual condition. In the first case, the model is inputted with a string of phonemes and the particular object that was watched by the speaker; in the second case, the model is still inputted with a string of phonemes, but instead of a particular object watched by the speaker, the model is given all instances of the objects in the image. Therefore, in the latter, finding a good AV pair is more challenging, as the model has to learn which object is most likely to have occurred with a given recurring audio sequence.

Their results shows that audio-visual mapping was easier when gaze information was available for the model than when it was not. Consequently, their experiment shows that co-occurrence statistics do not seem to be enough, even though they enable computational models to learn a few reliable word-object mappings such as in Roy & Pentland (2002). Word-object mapping can be learned more reliably when additional information (such as attention of the speaker) is available. Their results are consistent with research in child language acquisition which shows that shared-attention is a critical parameter for children to acquire their language. Recall that for blind children, it is the lack of shared attention (or its impoverishment due to the lack of a visual context) that leads them to acquire their native language slower than their sighted peers.

2.3.1.2 CNN-based Neural Models

Neural Networks enabled researchers to model even more complex interactions between the visual and the spoken modalities. Gabriel et al. (2014) introduced, to the best of our knowledge, the first neuronal (CNN-based) VGS model, where the model is trained to map images to isolated spoken words. Their model has two branches, an audio branch and a visual branch which vectorises the input image and word. The model is trained to minimise the cosine squared distance between matching vectors, while making this distance greater for mismatching vectors. Their model is then evaluated on a speech$\rightleftarrows$image retrieval task: retrieve the matching image given an input word, or vice-versa. Their results indicate that their network was indeed able to capture cross-modal links and effectively learn to map a spoken word to its visual context. Their model paved the way for models handling more complex acoustic stimuli, such as full captions instead of isolated words.

Harwath & Glass (2015) building upon the work of Gabriel et al. (2014) proposed the first model that could handle full spoken captions instead of isolated words. Their model can be considered as an upgraded version of the CELL model as it makes also a few simplifying assumptions. First, the images are processed by a CNN object detector that outputs 20 bounding boxes (19 bounding boxes around each detected object in an image and an additional bounding box for the the whole image); and second, the captions are pre-segmented at word level. The goal of the network is to align each word in the audio caption to each bounding box in the image. Finally, a global similarity score is computed for each image/utterance pair, and the network is encouraged to make the similarity score

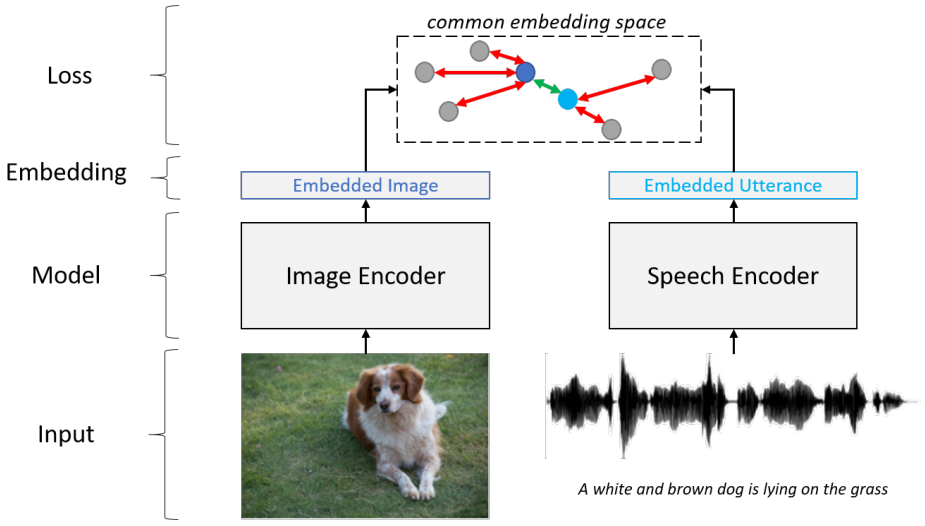


Figure 2.6: Traditional architecture of neuronal VGS models. Figure inspired by Chrupała et al. (2017*a*) and Khorrami & Räsänen (2021). The difference between CNN-based and RNN-based model depends on neural cells used for the speech encoder. Note that the embedded images and utterances need not be vectors but may also be matrices. The dashed box represents the shared embedding space in which the similarity measures between matching (green arrow) and mismatching (red arrows) image and utterance pairs is realised. Note that the shape of the embedding space depends on the nature of the encoders. The embedding space used by RNN-based models is typically a d-dimensional hypersphere, where d denotes the dimension of the embedding vectors (see Section 3.3.2). Transcription of the audio signal is only given for illustration purposes and is not inputted to the model. Black arrows show the direction of the information flow through the network.

between a matching image/utterance pair higher than with mismatching pairs.[14] Similarly to Gabriel et al. (2014) their model is evaluated on a speech⇄image retrieval task. Their results show similar trends to that of Gabriel et al. (2014): it is easier to find a caption given an image than vice-versa. Nevertheless, in both cases the results are much better than what randomness would predict. Harwath & Glass (2015) observed their model reliably learnt to map isolated words to their correct visual referent via the bounding boxes.

Harwath et al. (2016) improved on their previous work so that their model could use full captions instead of pre-segmented captions. Also, instead of running an object detector on the image, they simply use the penultimate activation of a pre-trained VGG network, yielding a 4096-dimensional vector representing the input image. The audio input is processed by three layers of 1D-convolutions with multiple filters at each layer. Contrary to their previous work, they use the dot-product of the image vector and the speech vector as similarity measure. This work allowed them observe that some parts of the captions have a higher degree of similarity with the image than others, suggesting their model was able to detect highly relevant acoustic sequences in the speech signal.

Harwath & Glass (2017) built upon their previous work by trying to find image regions regularly associated with spectrogram regions, so as to build audio-visual pairs. While they used the same model as their previous work, their loss function is different and is the same as in their first work Harwath & Glass (2015), that is a contrastive loss function. In order to infer an audio-visual collection, they first apply a grid over the image so as to divide the original image into sub-sections. They then run a VGG network over all possible

[14] Following the objective function used by Karpathy & Li (2017) for image/caption pairs, which will also be used in this thesis (see section 3.3.2).

groupings of consecutive image blocks so as to have several embeddings for one image. They proceed similarly for the captions, slicing them into subsections and embedding each subsection. They then compute a similarity measure between the representation of each image region embedding and each caption region embedding to find which have the highest degree of similarity. This effectively allows them to extract fine-grained image-audio pairs in an unsupervised fashion. This is even more impressive when one recalls that the network was trained on unsegmented images and unsegmented captions, hence showing the network somehow segmented the spoken input into word-like units.

Finally, in their latest work, Harwath, Recasens, Surís, Chuang, Torralba & Glass (2018) completely change their model so as to have a fully convolutional model: instead of using VGG vectors, they use the full VGG network up until the last convolutional layer (i.e. removing the fully connected layers), and use 1D convolutions for the input caption. Also, contrary to their previous work where the output of their audio branch is a 1D vector, they here keep the whole feature map, yielding a 2D matrix. In order to join the output of both branches, they use a dot product as in their previous work. Keeping the feature maps of both branches allows them to detect saliency regions between the convolved image and the convolved spoken input. Hence, they are able to highlight specific regions in the spectrogram that correspond to specific regions in the image, allowing them to build fine-grained image/audio pairs in an unsupervised fashion, contrary to their previous work which required computing similarity for each possible subpart of the image and spectrogram manually.

In their work, Harwath and colleagues thus went from finding coarse audio/visual pairs from pre-segmented images and pre-segmented captions to finding fine-grained audio/visual pairs without any supervision from raw speech paired to raw images. Overall, their work shows it is possible to extract word-like units from raw speech using images as a form of weak supervision. Most importantly, while their network was trained to minimise the distance between an image and its matching caption at a global scale, the network derived similarities at a local scale. Hence, lexical acquisition seems to be a by-product of the main task, and appears "naturally" in the model.

Kamper, Anastassiou & Livescu (2019) also proposed a visually grounded CNN-based model to learn speech segment embeddings in a query-by-example search task: given an audio excerpt, the network should find *semantically*, and not only *phonetically*, related audio excerpts among a collection of utterances. Contrary to the model of Harwath et al. (2016) that uses raw images, here images are represented as a "bag-of-visual (semantic) tags": a vector, where each dimension represents a given object in the picture and is assigned 1 if the object that is represented at a given dimension is in the image, 0 otherwise. Speech utterances are sliced into sub-parts ranging from a minimum length up to a defined maximum length. The goal of the network is to embed each speech excerpt and predict the bag-of-visual semantic tags — which boils down to a classification task. Here, the images only serve a grounding purpose and are only used when training the network. The intuition is that the network will be able to learn how to generate similar embeddings for speech segments which have the same referent in the image. For example, the words "flower" and "rose", while being phonetically different, are semantically related, and might refer to the same object in the paired image. Hence, the embeddings of these two words should be similar. Therefore, at testing time, the network should be able to retrieve audio excerpts corresponding to the word "flower" even when prompted with the word "rose". This is in fact what is observed at testing time, the network is able to retrieve semantically related words which are not phonetically close. Using images to ground the meaning of the audio excerpts proves successful.

Räsänen & Khorrami (2019) recently introduced a CNN model specifically designed to

study child language acquisition. Their model is trained to encode the spoken input, in the form of a spectrogram, into a bottleneck representation that is used simultaneously in two sub-tasks: the first task is to predict a visual context (in the form of a 60-dimensional vector representing objects) as traditional VGS models do; and the second task, a decoding task, which consists in predicting the original spectrogram. The task thus assumes that the ideal representation should both be able to correctly predict the visual context, and should incorporate enough details so as to be able to reconstruct the spoken input. Their model is able to learn to correctly map a spoken input to the visual referent while making some plausible mistakes such as confusing similar sounding words (e.g. cat/hat) or confusing semantically related words (e.g. nose/mouth). Most importantly, they show that removing the prediction branch of their model yields worse results, suggesting that joint learning of perception and production is essential.

2.3.1.3 RNN-based Neural Models

Contrary to CNNs, RNNs are designed to process sequences such as speech. Even though CNNs are also able to process sequences, RNNs are more plausible from a cognitive perspective. Indeed, as already mentioned (see Section 2.2.3), the output at a given timestep depends on the predictions of the past timesteps, which is not the case for CNNs. While RNNs are able to model long-term dependencies, CNNs are unable to so, as their predictions do not rely on the predictions made at the previous timesteps. Consequently, the way speech is processed in RNN-based models is closer to how humans process speech, that is, sequentially and incrementally.

The first RNN-based VGS model we know of is that of Chrupała et al. (2017*a*). This model builds upon their previous work on textual models (Chrupała et al. 2015) and their analysis (Kádár et al. 2015). The architecture used is very similar to that of the previously mentioned works and has two branches: an audio branch and a visual branch; and uses the same type of loss as Harwath & Glass (2015). Similarly to Harwath et al. (2016), they use VGG vectors as image features. Nevertheless, contrary to Harwath et al. (2016), the speech encoder consists of stacked recurrent highway networks (RHN, Zilly et al. 2017) instead of convolutional layers.[15] The final embedding of the audio branch is computed by an attention mechanism (see Section 3.4) which learns how to weight the vectors of each timestep, so as to give more weight to specific parts in the speech signal. They evaluate their model on a speech→image retrieval task, that is, given an input spoken caption, the network should retrieve the matching image. They report better results than Harwath & Glass (2015), showing RNN-based model are also able to adequately map both modalities.

More recently, Merkx et al. (2019) built upon the architecture of Chrupała et al. (2017*a*), adding an attention mechanism at each layer (implemented as in Chrupała et al. 2017*a*), and using bidirectional recurrent cells instead of unidirectional cells. Their model reached even higher performance levels than Chrupała et al. (2017*a*) and Harwath & Glass (2015) on the same data set. Several optimisations were also made, such as the use of cyclic learning rate (Smith 2017) which consists in having a learning rate which increases and decreases several times between an upper and lower bound, instead of a having a strictly decreasing learning rate as what is usually done. This allows the model to get out of local minima it could be stuck in and converge to a overall better solution.

However, these better results come at the cost of cognitive plausibility. Indeed, as this model uses bidirectional cell, it processes the spoken input from both ends at the same

[15] RHN are recurrent cells, similar to LSTMs or GRUs, which are however able to perform more transformation steps for each recurrent transition. This allows to learn better representations while reducing the size of the stack of recurrent cells, which in turn makes learning more efficient.

time, which is of course not possible for humans. Furthermore, contrary to the previously mentioned approaches (be they CNN of RNN-based), their model uses bottleneck features Fér et al. (2017) extracted from a multilingual neural network trained to predict phonemes. Hence, by using such features, their model has some knowledge about the basic units of speech which might help the model converge more rapidly as the input is more informative than plain MFCCs.

Recently, Krishnamohan et al. (2020) explored if VGS models are able to do few-shot learning of novel noun/object pairs. The goal of few shot learning is to give to a neural network as few examples as possible — 1 for one-shot learning, n for n-shot learning — and see if the model is able to learn from these few examples and generalise. To do so, they used novel (fake) objects from the NOUN (Novel Object and Unsual Name, Horst & Hout 2015) data base where fake objects (3D generated images with various views of the same object) are given fake names (e.g. kakimense, tanzerposk, etc.). They define n-shot learning in their experiment as presenting "n augmented variants [i.e. views] of the stimulus [i.e. novel image/noun pair] before the evaluation". Their model also has two branches, a visual branch which extracts bottleneck features, and an audio branch which is inputted with ASR bottleneck features which are then processed by an LSTM. As in the aforementioned models, their model is trained to minimise the distance between the encoded image and the encoded isolated word referring to the object in the image. Their network is first pre-trained on a set of existing noun/object pairs using a computer vision data set, and they then perform transfer learning on the novel objects. Their result shows that the VGS model they train is able to do 1-shot learning, that is, to learn the association between a novel object and associated noun when presented with one instance, and generalise to novel instances of the same object/noun pair. They compare their result with human data and show their model is on par with human results. We could however argue that they consider an idealised case of few-shot learning, where the label is presented in isolation. This is rarely the case, as usually the novel word is embedded in a full sentence, and humans, and more specifically chidren, are still able to do few-shot learning in such case. It remains to be investigated if it is also the case for neural networks.

We should mention a last model before ending this section which, even though it uses text instead of speech, remains interesting in its approach: that of Hill et al. (2020) and Hermann et al. (2017) which also aims at studying lexical acquisition. Contrary to the previous approaches, their model is trained using a 3D virtual environment, which the model, in the form of a virtual agent, can actively explore. The goal of the model is, given an instruction (e.g. "Find and bump into a pencil"), to learn to distinguish the target object (*pencil*) in the virtual environment from a distractor object (e.g. *fridge*). If the model bumps into the correct object, then it is said to have learnt the word for the target object. The model is not only trained to distinguish object based solely on their shape, but also on their colour (e.g. "Find and bump into the blue object"), pattern, or position (e.g. "Find and bump into the object furthest to the left as you look"). The model is free to explore (i.e. "walk") in this virtual world so as to fulfil the instruction it is given. Hence, this task requires the model to understand the instruction it is given and to actively explore its environment so as to understand what the target object looks like. Contrary to the previous approaches were the model is given static data (i.e. the model cannot change the data it is given so as to gain more insight), here the model is free to get a better understanding of the object by seeing different views of the same object by moving in the virtual environment. Thus, even though the input is less realistic and does not reflect the complexity of the real world, the model here is given the possibility of exploring its environment, such as what a child would do, which is impossible with the previous approaches.

The visual branch of their model is a CNN which processes the images of this virtual

world (of shape $84 \times 84 \times 3$[RGB colours]). The language branch of their model is a simple feed forward layer which embeds the instruction, presented as a single world. The output of both branches are then merged and further fed to a LSTM which, given the previously taken action (e.g. move right) and the current input, predicts what to do next. Their results are interesting as they correlate well with what is observed for children. Their model also display a slow learning curve at the beginning and then has a vocabulary spurt such as what is observed for children. The authors however observe a color bias when humans usually display a shape bias (i.e. children tend to overgeneralise based on the shape of the object and not so much on their colour). The authors explain this by the fact that colours and shapes are perfectly balanced in their training data, while in real life shape terms are more used than colour terms. Hence, shape bias seems only to depend on the input the child receives and not by its perceptible environment.

Hence, RNN-based models are able to learn speech/image mapping as well as CNN-based models. They also seem to be quite flexible models as they are able to quickly learn the mapping between novel object/noun pairs. The fact that these models use RNN cells instead of CNNs also make them more cognitively plausible, and hence constitute ideal test-beds to simulate lexical acquisition.

2.3.1.4 Representation Analysis

Chrupała et al. (2017*a*) tried to understand the representations learnt by their model. First, they used a classifier to test if the embedding of a full spoken caption contains information about the individual words of the caption, and at which layer this information is the most reliably encoded. They found that not all layers are equally informative about the presence of a word. Particularly, they found the lower layers of their architecture to be the least informative, while the second-to-last was the most reliable. This suggests that word-like units are progressively constructed as the information flows through the network, and that a certain amount of computation is necessary so that information appears. A similar observation was made more recently by Merkx et al. (2019).

Chrupała et al. (2017*a*) also explored to what extent the learnt embeddings encoded semantics. Their study reveal that the lower layers of their architecture encodes forms while higher layers encode semantics. Surprisingly, their results show that the last layer of their architecture as well as the final embedding encode the utterance semantics less reliably than the previous layers. The authors explain this is the case because the final embedding should correspond as closely as possible to the embedding of the paired image. Hence, the network might tune the final vector according to the visual modality and discard the information which is not useful in the speech signal.

Alishahi et al. (2017) explore to what extent such VGS model encoded phonology. Their study shows that the network's encodings approximately group the English phones by sound class (plosives, fricatives, affricates, etc.). However, the grouping is not perfect and seems to be done on the basis of acoustic factors (formants) rather than on deeper linguistic factors: we could have expected vowel/consonant dichotomy, a clear voiced/unvoiced dichotomy which is not the case here. This study also investigates the encoding of synonyms (e.g. *store*/*shop*) and whether the network is able to distinguish them or not, and at which layer precisely. Their results reveal that synonym discrimination is very low for the representations extracted from all recurrent layers except for the last and the final embedding. It thus shows that form is encoded in the lower recurrent layers while meaning is encoded in the top recurrent layer and the final embedding, making synonym detection much harder as no (or little) information on form is present.

Harwath and colleagues also performed analysis on the representations learnt by their

- an up close picture of an elephants [*sic*] trunk and eye.
- a baby elephant holding it's [*sic*] trunk out on a dirt ground.
- an elephant raising its truck [*sic*] with a tree in background.
- a close up of an elephant's face showing eyelashes and opening at end of trunk.
- an elephant pointing its trunk upward as it looks down.

Figure 2.7: Example image taken from MSCOCO (left) along with five descriptive captions (right).[16]

model. Harwath et al. (2016) explored if their network reliably encoded several occurrences of the same word, which are pronounced by different speakers. Their results reveal that the network is able to cluster several occurrences of a given word together. This shows the network was able to remove inter- and intra-speaker variability. Their analysis also suggests that their network performs an implicit segmentation of the audio input into sub-units, which was confirmed in a further study (Harwath & Glass 2017). Drexler & Glass (2017) showed that some neurons were specifically activated by certain sequences of phonemes. Interestingly, and similarly to the observation of Chrupała et al. (2017*a*) and Alishahi et al. (2017), the lower layers of the network are more concerned with form than with sense. They observe this by clustering the activation extracted at different levels of the architecture and show that the lower activations tend to cluster according to speaker identity while activation from the upper layers cluster according to meaning. Harwath & Glass (2019) showed that the second layer of their architecture was particularly sensitive to phone boundaries. Similarly to Alishahi et al. (2017), their study shows that the representation learnt by the network encodes coarse phonetic categories (fricatives, plosives, etc.).

2.3.1.5 Data Sets

Because the number of freely available data sets that feature both images and spoken descriptions is limited, most of the aforementioned models (Harwath & Glass 2015, Chrupała et al. 2017*a*, Merkx et al. 2019) are trained on extensions of data sets initially created for image vision purposes. The data sets these models use were conceived to train image captioning models, that is, models that generate textual descriptions of images passed as input. The main data sets used for this task are Flickr8k (Rashtchian et al. 2010, Hodosh et al. 2013) and COCO (Lin et al. 2014) which feature images paired with 5 descriptive captions written by humans (see Figure 2.7). These data sets naturally feature grounded language, as each descriptive caption is paired to an image, which constitutes, to some extent, knowledge of the physical world. As the captions were written by annotators upon having seen the image, the image naturally reflects the communicative intent expressed in the captions. An extensive presentation of the audio extension of these data sets is done in Section 3.2 as we use these data sets in this thesis.

Recently, new data sets that closely capture what children see were introduced. Slone et al. (2018) introduced a methodology to collect audio-visual data from children interacting with objects using eye-tracking devices and head-mounted cameras. Such data can then be used to train visually grounded speech or text models and study lexical acquisition. Tsutsui et al. (2020) for example use such data to study word-object mapping. In their experiment,

[16]Image Credit: Josh More, Flickr, BY-NC-ND 2.0, MSCOCO ID 4477.

they used videos captured with a head-mounted camera that shows the perspective of the child. They also used an eye tracker to pinpoint where the child was looking at each time frame and simulate visual acuity by blurring the image except at the focal point. Hence, the data is much more realistic than any other computer vision data set. The model used in Tsutsui et al. (2020) is a supervised model which is trained to classify an image as featuring (or not) a given object (e.g. bed, turtle, etc.). Their results show that word-object association is more successful when using data with the visual perspective of the child than with the visual perspective of the caregiver. Following a similar methodology, the SEEDLingS project Bergelson & Aslin (2017)[17] gathered data of child/caregiver interaction using a head mounted camera and microphones. This data was used to train the model of Räsänen & Khorrami (2019) presented above.

The results obtained by Tsutsui et al. (2020) and Räsänen & Khorrami (2019) highlight the need for more realistic data sets if one wants to use computational simulations to understand child language acquisition as advocated by Dupoux (2018). Indeed, data sets such as Flickr8k or COCO only reflect the perspective of adults and not the perspective a child could have of the same situation. Hence, models trained only on these data sets might have sub-optimal results. However, realistic data sets are often not publicly available, which explains why so few models are trained using realistic data.

2.3.2 Neural Models and Language Acquisition

2.3.2.1 Simulation or Modelling?

Most of the aforementioned models are inspired by child language acquisition, or at least make an explicit link between the ultimate goal of their model and child language acquisition. It seems reasonable to ask to what extent these models do *model* child language acquisition or do *simulate* child language acquisition, or in fact do something else. This distinction is important as it might influence the weight of the conclusions that will be drawn.

We may try to answer this question by referring to Marr's (Marr 1977) levels of information processing (see more specifically Marr 1983, p. 24). Marr distinguishes three levels of information processing: the *computational* level, the *algorithmic* level, and the *implementation* level. The computational level is concerned with the problem to be solved. It defines *what* is the object of the computation (i.e. the input and the output) and *why* those are the objects of the computation. This level also defines abstractly the constraints that should be satisfied by the computational process. The second level, the algorithmic level, is concerned more precisely with *how* the problem should be solved. Each step of the computational process is precisely defined. Note that this level is different from the computational level, as the computational level defines the task rather abstractly and what constraints it should obey to, while the algorithmic level precisely defined how the computational process takes place. Indeed, "there may be many algorithms that implement the same computation" Marr (1977). The final level, which is for us of little concern, is the implementation level, which also defines how the computations are done, this time at the physical level: are the computations supported by biological objects (i.e. real neurons) or artificial objects (i.e. silicon chips).

The difference between simulation and modelling is concerned with the first two levels only, that is the computational level and the algorithmic level. Rieder (2003, p. 818) makes a difference between the two terms: "simulation conveying the action of imitating reality and [the] model representing the vehicle". Simulation is thus "the act of presenting

[17] https://bergelsonlab.com/seedlings/

the appearance, or interacting with the behavior, of a system without the reality" while modeling would be "the generation of a facsimile or representation of the real system", the latter being "physical, mathematical, procedural [i.e. algorithmic], or some combination".[18] Hence, referring to Marr's hierarchy, the model is defined at the algorithmic level while the simulation is defined at the computational level. There could well be several possible models (i.e. several algorithms) possible that would generate the same simulation.

We believe that if one wants to model child language acquisition, the model should be as close as possible to what humans do algorithmically, and for example incorporate the same priors as humans, which is not the case with any of the aforementioned models. Hence, even though they do *simulate* child language acquisition to some extent, and more specifically lexical acquisition, they do not *model* child language acquisition. Nonetheless, not implementing the same "vehicle" as humans does not preclude these implementations to display similar patterns in the final simulation, hence making their implementation and study worthwhile so as to test hypotheses.

2.3.2.2 Perfect Simulation and Groundedness

According to Dupoux (2018) (and others), a successful approach to study child language acquisition using computational approaches should satisfy the following constraints:

> construct[ing] *scalable* computational systems that can, when fed with *realistic* input data, *mimic* language acquisition as it is observed in infants (Dupoux 2018)

If one wants to mimic language acquisition as it is observed in infants — which would be what we define as a *perfect simulation* — the aforementioned models are not entirely suitable. Indeed, the models we presented are inputted with an audio caption which they have to learn to map to a paired image. Consequently, these models only incorporate the perceptive abilities of a child but not its productive skills. Hence, this leads us to question to what extent these models really are grounded.

Roy (2005) further defines grounding as "an interactive process of predictive control and causal feedback." and describes the ideal grounding model (which he calls a *computational semiotics framework*) in a figure which we reproduced in Figure 2.8. The feedback loops — language production, and physical action — enable the child to act on the physical world, either by moving and gaining another view of what she is looking at, or by speaking to the persons around her, which prompts a response back from them. Therefore, the ideal grounded model should be both able of perception and production so as to really mimic child language acquisition.

None of the models we presented do implement all the feed-back loops necessary to have a fully-fledge grounded model. They all only implement a sub-set of it, and hence are not able to entirely simulate child language acquisition. Indeed, the models of (Harwath et al. 2016, Chrupała et al. 2017*a*, Merkx et al. 2019) only have a perceptive ability. While the model of Räsänen & Khorrami (2019)[19] is trained to reconstruct the input utterance — and somehow incorporating productive skills — it does not allow for the model to actively explore its environment, whereas the situation is opposite for the model of Hill et al. (2020). However, despite not incorporating all the aspects of child language acquisition, they do incorporate a sub-set of it, which is *lexical acquisition*. This process indeed implies building

[18] Rieder (2003) however notes that "another viewpoint is that the term modeling includes both the construction of models and the manipulations of these models (the simulations)".

[19] and other models such as Wang et al. (2020) and a model we have been working on during an internship in Japan.

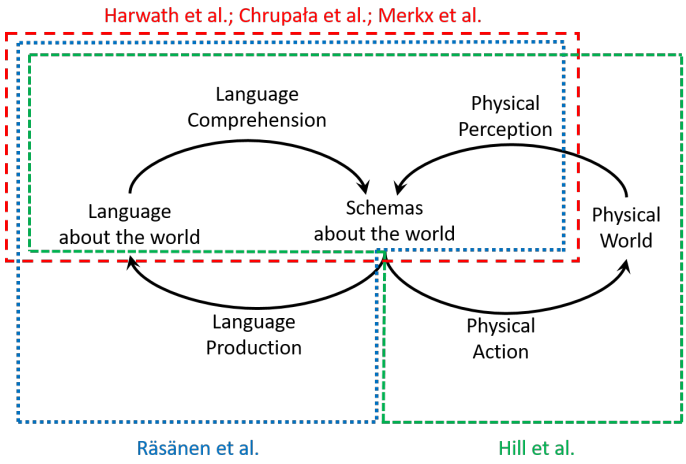


Figure 2.8: Figure reproduced from Roy (2005). We annotated the figure according to the models we presented, so as to reflect to what extent these models are fully-fledged grounded models.

schemas about the world (i.e. neural representations) using a lingusitic input (i.e. audio captions) and a non linguistic input (i.e. a visual context in the form of still images).

In this thesis, we will focus on studying lexical acquisition in RNN-based models, and more specifically in a model similar to that of Chrupała et al. (2017*a*). Note that the conclusions we will be able to draw and the links we will make with lexical acquisition as observed in children will necessarily be limited. We further describe this issue in Section 3.3.4.

2.4 Conclusion

Previous research shows that VGS models, be they CNN or RNN-based, are able to successfully learn how to map a spoken stimulus to a visual stimulus. These works also reveal that, despite having only been trained to minimise a distance between an image and its spoken description so as to map both modalities accurately, these models have developed deeper linguistics abilities. This is an interesting property, as these linguistic abilities only appear as a by-product of the main task. This seems somehow similar to how children learn. They indeed do not have an explicit linguistic task laid out before them, but their linguistic abilities rather emerge as by-product of the interactions with the world around them.

While these abilities were widely explored for CNN-based models, they were not for RNN-based model. We thus aim in this thesis to better understand what linguistic abilities VGS RNN-based models are able to develop. More specifically, we aim at answering the following questions:

- Do RNN-based models highlight specific part(s) of the spoken input that are particularly relevant to predict the target image, such as what was shown for CNN-based models? (Harwath & Glass 2017)
- If so, what specific parts of the input are highlighted (which specific word-forms, or part-of-speech)?
- If they do highlight specific parts in the input, is this ability cross-linguistically valid,

even if the model is trained with a language typologically different from English, say Japanese?

- To what extent such linguistic abilities develop over time? Are VGS models able to quickly acquire some linguistic abilities (i.e. with a relatively small amount of data) or not?
- Previous research (Chrupała et al. 2017*a*; Merkx et al. 2019) shows that the internal representations of such network encode the presence of the individual words of the spoken captions. This raises the question whether this behaviour is valid for all the words of a given caption, or only specific words?
- If the presence of individual words are encoded in the internal representations of the network, it means the network has learnt what constitutes a word and has stored this information in its weights. This raises the question of how the network recognises (i.e. activates) the representation of a given word based on an acoustic input.
- VGS models are trained with full unsegmented captions and obtain fair results. However, their textual counterparts (i.e. networks trained on written captions) do better. This raises several questions, the main one being: if VGS networks were presented with segmented captions, would they do better?

Part II

Contributions

Chapter 3

Visually Grounded Speech Architectures and Data

Contents

3.1 Introduction

Children learn their native language in context. That is, raw speech alone is not sufficient for them to acquire a language: language needs to be grounded in order for them to make sense of what is being said. Grounding is necessary to map linguistics signs — signifier/signified pairs — to their referents. Linguistic signs are not readily available objects that are to be found in the real world but they also have to be constructed from it instead. Through repeated exposure to referents in conjunction with linguistic signs, the child is able to build form/meaning pairs. Grounding may thus occur through various sensory-motor experiences: vision, touch, smell, taste, hearing, thermoception, social interactions, etc.

Neural architectures of Visually Grounded Speech (VGS), be they CNN-based or RNN-based, have recently become popular as they enable to model complex interactions between two modalities, namely speech and vision. Such architectures can be used to model child language acquisition and more specifically, lexical acquisition. Indeed, these architectures are trained to solve a speech-image retrieval task. This task involves identifying lexical units that might be relevant in the spoken input, detecting which objects are present in the image, and finally pairing the detected objects to the detected spoken lexical units. Their task is very close to that of a child learning her mother tongue, who is surrounded by a visually perceptible context and who tries to match parts of the acoustic input to surrounding visible scenes.

In this chapter, we present the VGS architecture that is studied throughout this thesis. We also present the data sets that are used to train such architecture and introduce a new

data set in Japanese consisting of image paired to audio descriptive captions. Finally, we mention the assumptions we make on language learning by using the particular architecture we use in conjunction with the data sets we present.

3.2 Data

Deep learning architectures require large amounts of data in order to be trained effectively and eventually converge. In our case, the architecture we train requires images that are paired to spoken descriptions. While image recognition and captioning data sets are numerous, only a few number of data sets featuring images paired with audio descriptions are freely available. We here present the three data sets that are used in the experiments of this thesis, two of them featuring synthetic speech (COCO and STAIR) and one featuring real human speech (Flickr8k).

3.2.1 COCO and STAIR

Two of the data sets we use for our experiments are based on the Microsoft Common Objects in Context (MSCOCO) data set (Lin et al. 2014).[1] This data set, initially introduced for computer vision purposes, consists of a set of images paired to five descriptive textual captions. The images show scenes of everyday life which feature at least one instance of the 80 target objects the data set was conceived around. The images comprising the data set were gathered on Flickr,[2] thus insuring a diversity of contexts and views in which the 80 object instances are presented. The captions were written by native English speakers recruited on Amazon Mechanical Turk.

In order to have a data set consisting of images and spoken descriptions, Chrupała et al. (2017*a*) introduced the Synthetically Spoken COCO data set. It consists in a synthesised version of the original MSCOCO captions using Google's Text-to-Speech (TTS) system (Chrupała et al. 2017*b*). The audio captions were synthesised using one synthetic female voice with an American accent. This resulted in 616,435 spoken captions for 123,287 images. The original training, validation and test splits of Vendrov et al. (2016) were kept and consist of 566,435, 25,000, and 25,000 captions respectively.[3] From now on, this data set will be referred to as COCO.

One of today's pitfall of NLP and SP research is that most of the work is carried out on mainstream languages[4] that are, for most of them, Indo-European languages. This has several impacts, the main one in our case being that it could bias our analysis. Indeed, Indo-European languages are very close typologically. Consequently, it might be that the patterns that are uncovered when analysing neural network might only stem from the fact that the languages are typologically related and the analysis might miss the bigger picture. Hence, in this thesis we introduce a speech/image data set in Japanese.

Our Japanese data set is based on the STAIR data set by Yoshikawa et al. (2017). Using the same methodology as Lin et al. (2014), Yoshikawa et al. (2017) collected five captions in Japanese for each image of the original MSCOCO data set. We chose this data set as it has several advantages over other text/image data sets in other languages. First, it uses the same set of images as the English data set (as opposed to the English-German Multi30k data set (Elliott et al. 2016) which uses a different set of images). Second, the STAIR data

[1] https://cocodataset.org/#home

[2] https://www.flickr.com/

[3] Representing 551h, 25.27h, and 25.24h of speech respectively; with an average duration of 3.6s±0.8/caption

[4] In terms of political power of their native countries, not in terms of absolute number of speakers.

set has captions for the whole original MSCOCO data set (as opposed to the Chinese version (Li et al. 2019) which only has captions for a small subset of the original data set). Finally, the captions were written by native Japanese speakers. They are original captions and are not translations of the English captions (as opposed to the Japanese captions introduced by Havard et al. (2017) which used computer generated translations). Following the same methodology as Chrupała et al. (2017*a*), we used Google's TTS system to synthesise speech for each of the Japanese captions (Havard et al. 2019*b*). In order to enable a fair comparison between the models trained on the English version of the data set and those trained on the Japanese version, we kept the exact same training, validation, and testing splits.[5]

From a typological point of view, Japanese is widely different from English, be it phonetically, phonologically (mora-based language), morphologically (agglutinative) or syntactically (OV language, explicit function marking with particles). This will thus enable us to test whether the behaviour of our architecture varies according to the language used for training, and if so, what kind of language-specific strategies the resulting models develop.

3.2.2 FLICKR8k

The audio quality of both the COCO and STAIR data sets is very high as both feature synthetic speech. However, as noted by Chrupała et al. (2017*a*), the generated speech is much clearer and simpler than real human speech: there is no inter-speaker variation, as only one synthetic voice is used, and there is very few intra-speaker variation: two occurrences of a word in the same context will be pronounced exactly the same. Also, real human speech is characterised by false starts, repetitions, corrections, and hesitations which are not present in synthetic speech. We therefore felt the need to validate our conclusion on real human speech using the Flickr8k data set.

As for MSCOCO, the Flickr8k data set (Rashtchian et al. 2010, Hodosh et al. 2013) was originally conceived for computer vision purposes. It contains 8,000 images, each paired to five written descriptions written by humans also recruited on Amazon Mechanical Turk. Harwath & Glass (2015) had each of the captions read by native English speakers, resulting in 40,000 audio captions. As this data set features 183 different speakers, it is much more challenging than the previous ones. The audio quality is also uneven from a caption to another, some being very neatly recorded while other have a lot of background noise.

As Chrupała et al. (2017*a*), who also used this data set in their original experiments, we kept the original splits provided by Karpathy & Li (2017) resulting in a training, validation, and testing set consisting of 30,000, 5,000 and 5,000 audio captions respectively.[6] Needless to say, we expect the results obtained using this data set to be lower than when using either COCO or STAIR, as modelling real speech is notably harder than synthetic speech and also because this data set is much smaller than the other two.

3.2.3 Data Format

Contrary to the VGS model of Harwath & Glass (2015) which used raw images and spectrograms, the architecture we used requires the data to be pre-processed before being inputted to the network.

For COCO and STAIR, which both use the same set of images, we use image vectors extracted from a VGG network (Simonyan & Zisserman 2015) trained on ImageNet. The

[5] Representing 729h, 32.53h, and 32.14h of speech respectively; with an average duration of 4.6s±1.2/caption

[6] Representing 34.39h, 5.76h, and 5.73h of speech respectively; with an average duration of 4.1s±2.0/caption

features we used for our experiments are those provided by Vendrov et al. (2016) which are averaged over 10 crops.[7] For Flickr8k, we used the image vectors provided by Karpathy & Li (2017),[8] also extracted from a VGG network.

The vectors we use do not correspond to the activation of the last layer of the VGG network,[9] but we use the activations of the last fully connected layer of the VGG network. Thus, the vectors computed at this layer are informative of the content of the image without being restricted to a fixed number of objects to predict.

Both for COCO and Flickr8k we use the acoustic features provided by Chrupała et al. (2017*b*) which were extracted using Python Speech Features.[10] The audio features for COCO consist of 12 Mel-Frequency Cepstral Coefficients (MFCC, Picone 1993) and energy. For Flickr8k, audio features consist of 12 MFCCs with deltas, delta-deltas and energy. We extracted MFCC features for the synthetic STAIR data set using the same parameters as for COCO.

3.2.4 Metadata

For each data set, we force-aligned the speech with its transcription at word and phone level using the *Maus* forced aligner (Kisler et al. 2017).[11] We also tagged each caption: we used TreeTagger (Schmid 1997) for English part-of-speech (POS) tagging using the default model and KyTea (Neubig et al. 2011) for Japanese POS-tagging using the default model. Because both taggers use different tag sets, that the tag sets are unnecessarily fine-grained (especially for English) and for comparison purposes, we converted each tag to its Universal POS equivalent (Petrov et al. 2012) which uses a coarser tag set.

While mapping PennTreeBank tags to their Universal POS equivalent was rather straightforward, it was not the case for KyTea's POS tag set. KyTea's analysis is closer to that of a morphological tagger than to that of a coarse-grained POS tagger as exemplified by the example thereafter:

Japanese	バナナ	が	たくさん	積ま		れ	て		いる	
KyTea Tags	banana/N	ga/PRT	takusan/ADV	tsu/V	ma/TAIL	re/AUXV	te/PRT		i/V	ru/TAIL
Gloss	banana	SUBJ	many	piled-up		PASS	connective particle		be/exist	
							PROG			
Translation	Lots of bananas are piled up									

Table 3.1: Example sentence in Japanese taken from the test set.

For example, the verb "積ま" (base form 積まる) is split in two parts by KyTea, respectively "積" ("tsu", VERB)" and "ま" ("ma", TAIL). For our analyses, we considered both parts as forming a single token which we labelled as a VERB. Such overanalysis also occurs for adjectives. In such cases, we considered the sequence ADJ+TAIL as forming only one token of type ADJ.

Another example of KyTea's over-segmentation and analysis is shown in Table 3.2 where the word "work vehicle" is split in two parts. In our analysis, we considered such suffixes as having the same POS as the preceding word.

[7] http://www.cs.toronto.edu/~vendrov/order/coco.zip

[8] https://cs.stanford.edu/people/karpathy/deepimagesent/flickr8k.zip

[9] which is a softmax over all the dimensions of the vector, which represents the probability that the object represented at dimension n is present in the image.

[10] https://github.com/jameslyons/python_speech_features with default configuration. Default configuration: window size: 25ms, window step: 10ms, pre-emphasis: 0.97

[11] Available at https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic (visited on February 28, 2020)

Japanese	作業車	
KyTea Tags	sagyō/N	sha/SUF
Gloss	work	vehicle
Translation	work vehicle	

Table 3.2: Example word oversegmented by KyTea

3.3 Architecture

The model we use for our experiments is based on that of Chrupała et al. (2017*a*).[12] The modifications we made to the architecture are detailed in the following sections. This implementation uses Python's Theano deep learning library (James Bergstra et al. 2010). As most VGS architectures (Harwath & Glass 2015, Kamper, Shakhnarovich & Livescu 2019, Merkx et al. 2019), the architecture we use has two main components: an image encoder, and a speech encoder (see Figure 3.1). Such models are trained to solve a speech/image retrieval task: given a input spoken description, they retrieve the image that matches the description the closest. To do so, such models project the image and its matching description in a common vector space so that matching pairs lie close in the representation space while mismatching pairs lie far apart. Therefore, the speech encoder and the image encoder have to learn how to trasnform these two modalities appropriately so the resulting image vectors and matching speech vector lie near in the final embedding space.

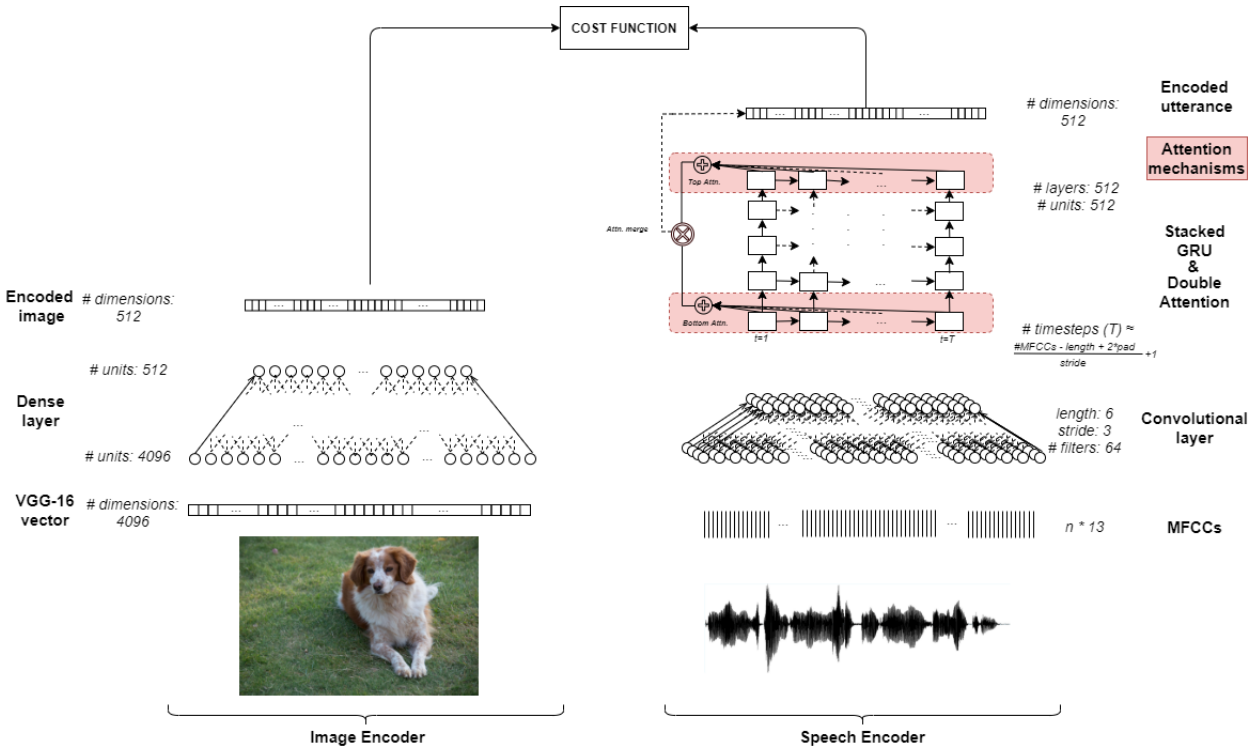


Figure 3.1: Architecture of the visually grounded speech models used in the following experiments. The red boxes show the position of the two attention mechanisms used in the following experiments.

[12] https://github.com/gchrupala/visually-grounded-speech

3.3.1 Encoders

Contrary to Harwath & Glass (2015) and Harwath & Glass (2017) who use a full VGG network as image encoder, the image encoder of our architecture simply consists of a linear layer. Indeed, as previously mentioned (see 3.2.3), our architecture uses pre-computed VGG vectors instead of raw images. The image encoder thus learns how to reduce the original 4096 dimensional input vector to a lower dimension. The resulting vector is then normalised to the unit ℓ^2-norm.[13] Note that it would still be possible to use full image encoder – such as a VGG or ImageNet encoder. However, it would substantially lengthen the training time and would be computationally heavier. As we wish to focus on the representation learnt for speech, we deemed such modification unnecessary and kept the original architecture. Using vector extracted from a trained neural network (such as done by Chrupała et al. 2015 and hence this thesis, or Merkx et al. (2019)) or using a the full pretrained network (such as Harwath & Glass 2017) is a common practise known as transfer learning. Nonetheless, such *transfer learning* is not necessary, and VGS models are still able to converge – sometimes even better – when the image encoder is trained from scratch such as shown by Mortazavi (2020).

The speech encoder consists of a 1D-convolutional layer that subsamples the input vectors followed by five layers of Gated Recurrent Units (GRU) (Cho et al. 2014) with residual connections (He et al. 2016). Contrary to the original architecture of Chrupała et al. (2017*a*) which has only one attention mechanism, we use two attention mechanisms (see section 3.3.3 for details): one after the first recurrent layer, and a second after the fifth recurrent layer. The final vector produced by the speech encoder is an element-wise product of the vectors produced by both attention mechanisms. This vector is then normalised to the unit ℓ^2-norm.

Contrary to the original implementation that uses Recurrent Highway Units (RHN) (Zilly et al. 2017), we decided to use GRUs as these cells are widely used by the research community, and also because their behaviour is better understood. Furthermore, it is to be noted we used unidirectional GRUs and not bidirectional GRUs. Indeed, unidirectional GRUs process the input sequentially from left to right, thus respecting the temporal dimension of speech, whereas bidirectional GRUs process input both from left to right *and* from right to left at the same time. We thus decided to use unidirectional GRUs as those are more cognitively plausible as humans process speech from left to right and not both ends at the same time.[14]

3.3.2 Contrastive Loss Function

The network is trained to minimise the following triplet loss function (as used by Chrupała et al. (2017*a*) and orignally proposed by Weinberger & Saul (2009)):

$$\mathcal{L}(u,i,\alpha) = \sum_{u,i}\left(\sum_{u'}\max[0,\alpha + d(\vec{u},\vec{i}) - d(\vec{u'},\vec{i})] + \sum_{i'}\max[0,\alpha + d(\vec{u},\vec{i}) - d(\vec{u},\vec{i'})]\right) \tag{3.1}$$

where $\vec{u}$ is an encoded utterance, $\vec{i}$ an encoded image, $\vec{u'}$ and $\vec{i'}$ are mismatching utterances (respectively images) with respect to image i (respectively utterance u). This loss function

[13] that is $x = \frac{x}{\|x\|_2}$ where ℓ^2-norm $= \|x\|_2 \triangleq \sqrt{\sum_{i=1}^{d} x_i^2}$ so that all vector lie on a d-dimensional hypersphere.

[14] Humans can also process the spoken input in reverse (in case one misunderstood what was said for example). However, this happens once the spoken utterance has been processed once, thus the reverse processing only takes place after, whereas for bidirectional RNNs, the input is processed from both ends, independently and at the same time.

encourages the network to minimise by a margin α the distance $d(\vec{u}, \vec{i})$ between the encoded image $\vec{i}$ and the encoded utterance $\vec{u}$ belonging to matching image/utterance pairs while making the distance greater for mismatching image/utterance pairs. The loss is computed at batch level, that is, all the images inside a batch (except image i) are considered as contrastive mismatching examples for utterance u, while all the other utterances (except image u) inside the same batch serve as contrastive mismatching example for image i.[15] In our case, the distance d used is the cosine distance,[16] defined as follows:

$$\begin{aligned} d(\vec{u}, \vec{i}) &\triangleq 1 - \text{cosine similarity} \\ &\triangleq 1 - \frac{\vec{u} \cdot \vec{i}}{\|\vec{u}\| \left\|\vec{i}\right\|} \end{aligned} \tag{3.2}$$

The domain of cosine similarity is $[-1, 1]$, where $cos(0°) = 1$; $cos(90°) = 0$; and $cos(180°) = -1$. Hence, if two vectors are collinear (with the same sign), their cosine similarity will be 1, and consequently the cosine distance will be 0. Consequently, orthogonal vectors will have a cosine distance of 1, and opposite vectors (collinear with different signs) a cosine distance of 2. The loss function encourages the network to produce collinear vectors for matching speech/image pairs.

Even though the batches are created by randomly selecting examples in the training set, it could be that, in the same batch, two images are adequately described by the same audio caption. Such pairs would act as mismatching examples even though in reality they are not truly mismatching pairs. In practice, it does not impede learning.

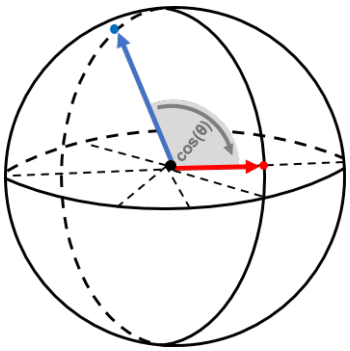


Figure 3.2: Illustration of Cosine Similarity on a d-dimensional hypersphere between two vectors.

This loss function is called *contrastive* as it does not only enable the model to learn what makes matching pairs similar (e.g. that the acoustic unit [dɔg] corresponds to the furry animal in the picture), but also enable the model to learn what makes mismatching pairs dissimilar (e.g. that the acoustic unit [dɔg] does not corresponds to green background in the picture; and the furry animal in the picture does not correspond to the acoustic unit [gɛˤl]). Such loss function enables us to train our architecture fully unsupervisedly.

[15] Note that it is not the only possibility. Merkx et al. (2019), for example, uses importance sampling in order to only select the hardest examples inside a batch and obtains better results.

[16] It should be noted that despite being referred to as a *distance* in the literature, it is not a true distance. However, as in our case both vector A and B are ℓ^2-normed, it is proportional to the Euclidean distance.

3.3.3 Attention Mechanism

One of the key components of this architecture is its attention mechanism (as presented in Section 3.4). We remind to the reader that the attention mechanism computes the context vector c as follows:

$$c = \sum_{t=1}^{T} \alpha_t h_t \tag{3.3}$$

where T is the length of the sequence, h_t is the hidden stated produced by a recurrent cell at time step t, and α_t is a learnable parameter. Recall that the attention mechanism learns how to assign a weight α_t to each input vector. The higher the weight, the more importance is given to the vector in the final representation. In our case, high attention weights over specific vectors means the network has paid more attention to specific portions of the input signal.

We decided to use two attention mechanisms at two different levels of the network in order to see if the network highlighted the same type of information at different layers or not. The final vector is the dot product of the vectors computed by both attention mechanisms:

$$c = \left(\sum_{t=1}^{T} \alpha_t^{GRU1} h_t^{GRU1} \right) \cdot \left(\sum_{t=1}^{T} \alpha_t^{GRU5} h_t^{GRU5} \right) \tag{3.4}$$

where T it the total sequence length, h_t^{GRU1} and h_t^{GRU5} respectively represent the hidden vector computed at time step t by the 1^{st} and 5^{th} layer, and α_t^{GRU1} and α_t^{GRU5} is the attention weight computed respectively by the attention mechanism following the 1st and the 5th recurrent layer. The final vector is normalised to the unit ℓ^2-norm.[17]

The idea of using more than one attention mechanism stemmed from the article by Mattys et al. (2005) which shows that human attend to different cues when processing speech in order to isolate words from the speech stream.[18] The first attention mechanism we use is placed after the first recurrent layer, hence very close to the acoustic information. The network could thus use acoustic-related information (such as pitch) to highlight relevant units. The second attention mechanism being higher in the architecture, we expect the information to be more abstract, and thus we except the attention mechanism to highlight units based on semantic cues. We hence aimed at checking if the neural network highlighted different type of units by analysing the weights of the two attention mechanisms.

3.3.4 Assumptions in the Model[19]

In this thesis, we will draw analogies between the learning processes and strategies of a neural network and those of a child learning her mother tongue. However, it is worth mentioning that our comparisons will be limited as there are many differences between how learning occurs in a machine and in a human. By design, the model and data we use also makes certain (limiting) assumptions which we will review here. We may classify the assumptions that are made in this thesis in three categories: assumptions on the data, assumptions on the task, and assumptions on the computations.

[17] We also tried to do a simple sum to merge the two vectors computed by the two attention mechanisms, but we found dot product to be more efficient.

[18] At the time we introduced this modification, using more than one attention mechanism was not a common practice and was popularised by Vaswani et al. (2017). Using multiple levels of attention in a model of VGS was then adopted by Merkx et al. (2019) who uses an attention mechanism after each recurrent layer.

[19] Title borrowed from Roy & Pentland (2002).

We first examine the limitations that ensue from the data set we use. Our data set is biased by the nature of the data, as first, it only contains images that, by nature, are still while the real world is not; and second, because these images were selected according to a predefined set of objects they should depict, they do not reflect the distribution of the objects children encounter in their daily life. Therefore, the salient parts of the images are static objects and not (or rarely) actions the persons depicted in the images could undertakeare undertaking. Thus, the descriptions are equally biased and mainly describe the objects present in the image and not actions. Additionally, in our experiments every picture is paired to a spoken description. This is a very strong assumption which occurs quite rarely in real life. Indeed, even though children and adults tend to pay attention to the same things in their environment (see Section 1.3.2) and thus tend to speak about what is being attended to, it is often the case that adults refer to absent persons, objects, or situations when they speak. In short, it is not because we speak in the context of a particular situation that the discourse that is held necessarily refers to this situation, and if it refers to the particular situation, it might not be a description of the situation.

Second, in our experiments we make assumptions on how learning occurs. This assumption ensues from the task our neural network is trained to solve. The neural network we use is trained to solve a very specific and well-defined task which is a speech-image retrieval task: given a spoken description of an image, the network should retrieve the image that matches the input description the closest. This task is artificial and does not correspond to what humans do when they learn their language. Humans do not restlessly try to predict a visual context from what they hear. Nonetheless, the core skill that solving such task implies, that is, learning a mapping between a visual stimulus and an acoustic stimulus, is a skill that children develop when they learn their language. However, this core skill is just one of the many skills children need to develop to learn their language, but not the only one. Even though visual stimuli play an important role in language learning, vision is not the only way to ground speech: smell, touch, taste, social interactions, for example, are equally (or even more) important than vision to learn a language (see Section 1.3.5.2). Yet, for the particular neural network we study, vision will be the only modality used to ground speech.

Finally, in this thesis we make assumptions on the memory capacities and the nature of the computation done by humans. One assumption we make by using the model we use is that image processing is decoupled from audio processing. Indeed, in our model, both processes are done in parallel, but independently while in human cognition both visual and audio processing is done jointly (see McGurk & MacDonald (1976) for an example). Thus, in our model there is no true interaction between the visual stimuli and the audio stimuli. The interaction between both modalities only occurs through the loss function. Moreover, as our network uses pre-trained VGG vectors to represent the content of the images it implies that one should first learn how to see before mapping what is visually perceived to what is heard. It would thus imply that children have from the beginning on a clear and robust categorical perception of their visual surroundings, which we know is not the case: babies have a blurry vision and shorter sight angle, which implies they are not able to perceive distant objects as clearly as adults.

The way the spoken input is processed by the network makes also false assumptions on what we know of human speech processing. MFCC vectors are hand-engineered features that aim at reproducing how speech is perceived by humans. However, the speech encoder we use does not implement several innate priors which we know children have, such as categorical (discrete) perception of speech. Our speech encoder must thus be considered as a blank slate that needs to learn how to encode speech, which is not the case when children are born.

The loss function we use (presented in Section 3.3.2) also makes several (false) assumptions about the memory capacities of a child or humans. This function computes a loss value at batch level. This implies a child would be able to compare a given situation to several other situations, and would have, by doing so, stored in mind the exact sentence and image representation of theses situations. This is of course impossible at a large scale.

Hence, the conclusions we draw should be understood in light of the presented assumptions and limitations. If vision – in form of still images – were to be the only modality used to make sense of the surrounding speech, and that the surrounding speech always referred to the visual context, what regularities should we expect children to have picked up?

3.4 Chapter Summary

In this chapter, we presented the VGS speech model we used and detailed the modifications we made to the original architecture of Chrupała et al. (2017*a*), and detailed each part of our architecture. We also showed the assumptions using such a model implied and showed the limitations of using the model and data we use in our conclusions.

We presented the data sets used to train such architectures (Flickr8k and COCO). The true contribution we made in this chapter was introducing a new audio/image data set based on the original text STAIR data set (Yoshikawa et al. 2017) that we created following the same methodology as Chrupała et al. (2017*a*). This data set enables us to compare the performance of our network on two comparable corpora, the only changing factor being the language: either English (COCO) or Japanese (STAIR). This data set, entitled Synthetically Spoken STAIR (Havard et al. 2019*b*), is openly available to the research community: `https://zenodo.org/record/1495070#.Xyv1dCgzZhE`.

Finally, we also made available the POS tags and forced alignments of both the Sythetically Spoken COCO and Synthetically Spoken STAIR data sets: `https://github.com/William-N-Havard/VGS-dataset-metadata` so that the research community is able to reproduce our results.

Chapter 4

Attention in a Model of Visually Grounded Speech

The work presented in this chapter is based on the article we published at ICASSP2019 Havard et al. (2019*a*). This work is inspired by prior work by Chrupała et al. (2017*a*), notably Figure 3 of the aforementioned article.

Contents

4.1 Introduction

In the previous chapter, we presented the model of VGS we used and mentionned that this model is trained to solve a speech-image retrieval task. The main hypothesis we have in this thesis is that, in order to learn a reliable mapping between and image and its spoken description, the model should implicitly learn to segment the spoken input into sub-units. As the images used mainly figure objects and not actions, we hypothesise that the model should implicitly learn to segment noun-like units.

In this chapter, we study the attention mechanisms of our models and analyse which parts of the input signal are highlighted. Indeed, as previously mentioned, *attention* is a tool the network is given so as to favour some part of the input signal over other. We

hypothesise that our network uses this tool as a form of segmentation module in order to highlight particularly relevant words in the spoken input.

We focus on analysing where attention is located, that is, we study which parts of the input speech signal are more favourably highlighted than others. We first start by showing that the learnt attention weights are not random and do highlight specific parts of the speech signal. We then study how the distribution of the attention weights evolves over time. Indeed, as the attention mechanism is a trainable component of our network, the distribution of the attention weights evolves over time. Finally, we compare what is known of language acquisition in children to the behaviour we observe in our models.

We believe the main originality of this chapter resides in the fact that the work we present is the first to study the behaviour of attention in two typologically different languages: English and Japanese. Such methodology enables us to isolate the language general behaviours from the language specific behaviours of our models.

4.2 Is Attention Explanation?

Before diving into the experiments of this chapter, we have to address an issue that was recently raised regarding the reliability of studying the attention weights of a given model. Jain & Wallace (2019) claim in their paper "Attention is not Explanation" that attention weights do not provide a meaningful explanation of a neural network's predictions. Notably, they show it is possible to find alternative distributions of the attention weights while keeping the predictions intact. They introduce two ways of doing so: the first one consists in randomly shuffling the attention weights and the other one consists in constructing adversarial attention weights whose distribution differs as widely as possible from the original distribution while leaving the final prediction unchanged. They come to the conclusion that because attention weights can be modified without changing the ouput of the neural network they study, attention weights do not constitute a reliable source of information to understand the predictions of a neural network.

Wiegreffe & Pinter (2019) in their response paper "Attention is not not Explanation"[1] however mitigate the initial statement of Jain & Wallace (2019). They first state that the goal of studying attention is not to explain all of a model's behaviour by only looking at its attention weights: "attention scores are used as providing *an* explanation; not *the* explanation". We totally subscribe to this view. Notably, they observe that the adverserially computed attention weights usually perform worse than the initial attention weights showing "the relationship between tokens and prediction [...] cannot be easily 'hacked' adversarially". The experiments of both papers suggest that manipulating the attention weights yields uneven results depending on the data set, suggesting that in some cases it is legitimate to consider attention weights as a form of explanation.

Even though the experimental settings of Jain & Wallace (2019) widely differ from ours — their task is sentiment analysis while ours is speech-image retrieval, they use discrete units as input whereas we use acoustic vectors that are continuous by nature, their output consists in a binary classification while ours consists in predicting a vector in a continuous space, etc. — we believe the issue they raised is important. In order to ascertain that the attention weights in our models do constitute a form of explanation, we will perform a few sanity checks. We will randomly shuffle the attention weights of both attention mechanisms. If the scores obtained with shuffled weights are worse than those obtained with the original weights, we will be able to conclude the original attention weights are useful for the models' prediction and can thus constitute a good way of understanding our models' behaviour.

[1] Emphasis added.

4.3 Studying Attention

In this chapter, we aim at understanding the behaviour of the attention mechanisms of our architecture. More specifically, we study which parts of the spoken input the model pays attention to by analysing the attention weights. Our analyses focus on the following points:

- Which POS are highlighted;
- Which words are highlighted;
- Which parts of a given word are more specifically highlighted;
- How the distribution of the attention weights evolves over time.

After having trained models on either the English data set or the Japanese data set, we encode each caption of the test set and extract the attention weights α (see 3.3.3) for both attention mechanisms. Recall that the higher the weight is for a given vector, the more importance this vector will have in the final representation. We then use a peak detection algorithm[2] to detect the local maxima (that from now on will be referred to as *peaks*) in the attention weights. We considered as peaks local maxima that were at least 60% as high as the highest detected peak. For the rest of this thesis, we will say that attention *highlights* a specific part of a unit (POS, word, etc.) if there is a local maxima among the attention weights assigned to the vectors that comprise this unit. A visualisation of how attention weights are distributed on a specific caption is shown in Figure 4.1 where high attention weights that are considered to be peaks are marked with a large orange marker.

In order to understand if the highlighted units are different from what chance would predict, we could compare the proportions of peaks highlighting a given unit to the token frequency of the same unit and see if both are close or not. However, doing so would introduce a bias. Indeed, spoken units vary in length (nouns are typically longer than determiners or prepositions), and tokens do not account for this length difference. In order to have a baseline reference, we instead use random peaks as a baseline. For each caption, we sample $10 \times n$ random peaks, where n is the number of true detected peaks in the caption. A random peak is thus a randomly selected time step in the vector sequence, where each time step is equally likely to be selected as the other. We then compute word and POS distribution under such random peaks. Doing so enables us to account for the size difference of different words and enables us to estimate peak distribution if peaks were to be randomly distributed.

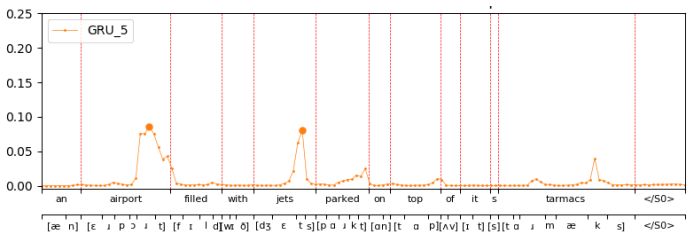
For brievety reasons, from now on, the attention weights computed by the attention mechanism following the first layer and fifth GRU layers will be referred to as "GRU1" and "GRU5" respectively. Therefore, the sentence "GRU1 highlights more prepositions than determiners" should be understood as "the attention mechanism computing the attention weights for the hidden states of GRU1 highlights more prepositions than determiners".

4.4 Experiments on Synthetic Speech: COCO & STAIR

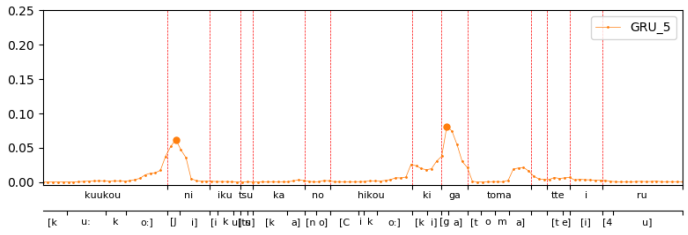
4.4.1 Experimental Settings

We used the same experimental settings for both COCO and STAIR in order to enable a fair comparison and used the same experimental parameters as Chrupała et al. (2017*a*): a 1D convolution layer with 64 filters, a window size of 6 and a stride of 3, followed by 5 recurrent layers consisting of 512 units with residual connections. Attention dimension was

[2]Peak detection is done by taking the first-order difference of the input sequence. Python module available at https://bitbucket.org/lucashnegri/peakutils/src/master/.



(a)



(b)

(c)

Figure 4.1: Example of the distribution of GRU5's attention weights over (4.1a) an English caption and (4.1b) Japanese caption along with (4.1c) reference picture.

set to 512, Adam optimiser with a initial learning rate of 0.0002, margin size α of 0.2. As in the original implementation, we did not use any dropout as we found out that it hindered performance. The model was trained for 15 epochs.

4.4.2 Results

In order to insure that our results were not due to chance, we trained five models, each with a different seed. The results presented in Table 4.1 are an average ($\pm$ standard deviation) of the results obtained across the five models. Models are evaluated in terms of recall@k (R@k) on a speech-image retrieval task. For each spoken query, the images are ranked from the closest matching to the least matching image. R@k evaluates if the target image (i.e. the image truly paired with the query caption) is ranked among the first k images. We also report median rank $\widetilde{r}$ which informs us at which rank the true images are ranked on average.

Model	R@1	R@5	R@10	$\widetilde{r}$
English (COCO)	5.52 ± 0.31	18.26 ± 0.82	28.64 ± 1.15	27.4 ± 1.51
Japanese (STAIR)	5.3 ± 0.16	17.88 ± 0.41	27.92 ± 0.36	29.2 ± 0.44
English (Chrupała et al. 2017*a*)	11.1	31.0	44.4	13

Table 4.1: Recall at 1, 5, and 10 (in %) as well as median rank $\widetilde{r}$ ($\pm$standard deviation) on a speech-image retrieval task on the test set of our data sets (5k images) averaged over five runs with different seeds. Models were selected according to the highest R@1 on the validation set. We report the results obtained by Chrupała et al. (2017*a*) with the original RHN implementation. Chance R@k are 0.0002/0.001/0.002, chance median rank $\widetilde{r}$ is 2500.5.

Overall, our results are worse than the results obtained by (Chrupała et al. 2017*a*). This is to be explained by the fact we swap the RHN cells for GRU cells. Even though our results

are lower, they still remain much better than chance scores, showing the models we trained did effectively learn to map an image with its spoken description.

4.4.3 Random Attention

The first question we need to answer before analysing the patterns of our attention mechanisms is the following: is attention really useful? To do so, we either shuffled the attention weights of both attention mechanisms at the same time, or only shuffled alternatively one of the two so as to estimate the contribution of each in the final prediction. We did so on the best trained model (selected on the validation set) of each of the five runs for both COCO and STAIR. We present the results averaged over five runs ($\pm$ standard deviation) of randomly shuffling the attention weights in Table 4.2.

When the attention weights of both attention mechanisms are shuffled, we observe that R@1 is barely above 0% showing the network is barely able to find the correct image given a spoken query. It clearly shows that the learned attention weights do highlight very specific parts in the spoken input, and that the highlighted parts are essential for the network to correctly encode the spoken input.

We observe a different pattern when we only shuffle the weights of only one of the two attention mechanisms. When we shuffle the attention weights for GRU1 but leave those of GRU5 intact, we notice that the network obtains better results than when we do the opposite. In both cases, the results are worse than when we leave the attention weights of both attention mechanisms intact (see Table 4.1), but better than when both are shuffled. This shows that the attention weights of GRU5 are more important than those of GRU1, as shuffling the weights of the former has much more (negative) impact than shuffling the weights of the latter.

	Rand. Attn GRU1 Rand. Attn GRU5	**Rand. Attn GRU1 True Attn GRU5**	**True Attn. GRU1 Rand Attn GRU5**
COCO	0.32 ± 0.13	2.66 ± 1.17	1.02 ± 0.18
STAIR	0.20 ± 0.07	1.94 ± 0.65	1.00 ± 0.16

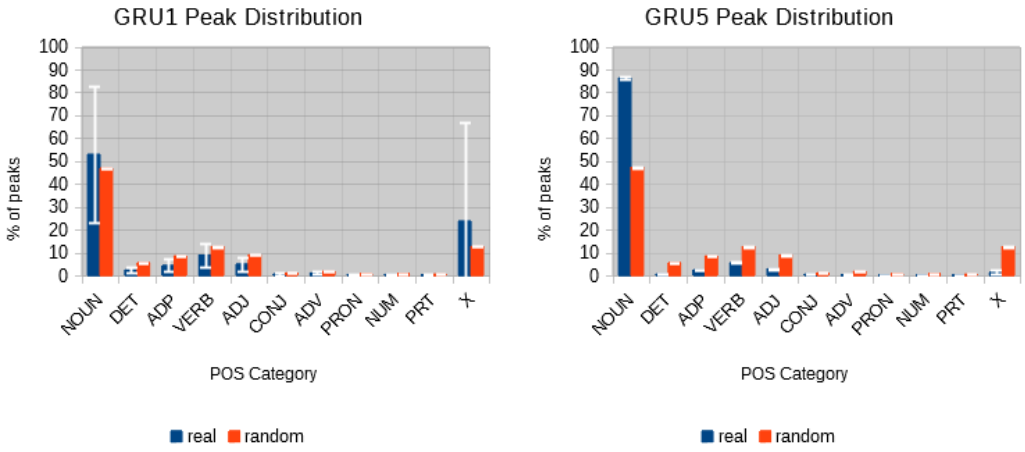
Table 4.2: Recall at 1 (averaged over 5 runs $\pm$ standard deviation) of trained models where attention weights are randomly shuffled.

Thus, this experiments allows us to conclude that both attention mechanisms highlight units that are useful for the networks' predictions and hence, that analysing the attention weights is a legitimate endeavour.

4.4.4 Highlighted POS

Following the methodology introduced in Section 4.3 we analyse which POS both attention mechanisms highlight.

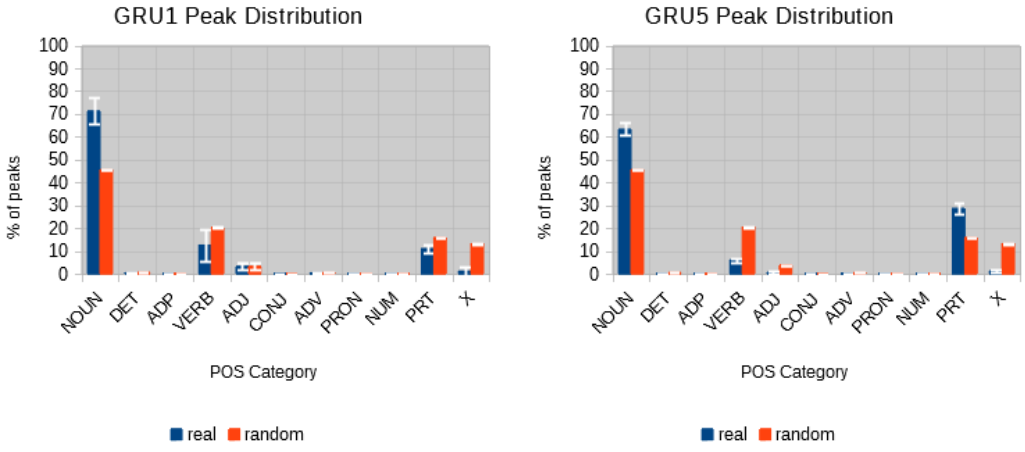
For COCO (Figure 4.2), we notice a large asymmetry in the POS that are highlighted by GRU1 (Figure 4.2a) and GRU5 (Figure 4.2b). We observe that GRU5 highly focuses on nouns ($85.89\% \pm 0.38$ of the peaks) and barely focuses on other POS. We also notice that the proportion of highlighted nouns is very different from what randomness would predict ($47.15\% \pm 0.09$), showing the network has learnt to detect and focus specifically on nouns. The behaviour of GRU1 appears different and seems much closer to a random behaviour: $52.37\% \pm 29.46$ peaks are located above nouns when randomness would predict $46.92\% \pm 0.15$. However, this result is due to an outlier run (hence the large standard deviation) where attention was concentrated at the end of the captions. Once this outlier



(a) (b)

Figure 4.2: Bar plots showing the proportion of attention peaks above each POS for (4.2a) GRU1 and (4.2b) GRU5 on the COCO data set. Real peaks are shown in blue and random peaks are shown in red. The results are averaged over five runs with different seeds. Error bars represent ± standard deviation.

removed, we also observe that a large part of the attention peaks are located above nouns (65.46% ± 3.77). Other POS are less highlighted than what randomness would predict (particularly for GRU5) showing the network has learnt to detect the most relevant type of words (which correspond to nouns) and ignore the others.



(a) (b)

Figure 4.3: Bar plots showing the proportion of attention peaks above each POS for (4.3a) GRU1 and (4.3b) GRU5 on the STAIR data set. Real peaks are shown in blue and random peaks are shown in red. The results are averaged over five runs with different seeds. Error bars represent ± standard deviation.

For STAIR (Figure 4.3), we also notice that there is a large difference between the random peaks and the true attention peaks, showing that for Japanese also, the attention mechanisms highlight different units than what chance would predict. Both attention mechanisms have learnt to focus on nouns: 71.33% ± 5.94 for GRU1 (Figure 4.3a) and

63.30% ± 2.76 for GRU5 (Figure 4.3b) when randomness would predict 45%. We notice a slight asymmetry in the highlighted POS between both attention mechanisms: while GRU5 focuses on particles 28.63% ± 2.15 (where randomness would predict 15.88% ± 0.07), GRU1 highlights such POS less than what randomness would predict and only focuses on nouns instead. In Japanese, particules are clitics that may follow a noun, a verb, an adjective, or even a sentence and that cover a wide range of functions: some have clear grammatical functions and act as case markers such as "は" /wa/ (topic marker), "が" /ga/ (subject marker), or "を" /o/ (object marker) while others are only interjections such as "ね" (tag question, used to express surprise).

The study of which POS are highlighted by the attention mechanisms of the English and Japanese models reveals that the models have adopted a specific behaviour according to the language of the captions. In English, nouns are the most highlighted POS as they refer to objects present in the image. It should be noted that verbs are not completely ignored as there are some cases in which they are the only informative words of the captions. (e.g. "a woman skiing down a track beside some trees" for which the action of "skiing" is the main information). In Japanese however, the models took advantage of how the language works by not only highlighting nouns, but by adopting a language-specific behaviour when highlighting particles.

4.4.5 Highlighted Words

In this section, we take a closer look at which words are specifically highlighted by the best model of each language. To do so, we compute the proportion of peaks that are above a given word-form. This gives us a finer understanding of which specific words are the most highlighted by each attention mechanism for each language. The 40 most highlighted words for both COCO and STAIR are shown in Appendix B.

For COCO (Tables B.1 and B.2), we notice a slight difference in the words highlighted by the two attention mechanisms. For GRU5, all of the top 10 highlighted words are nouns referring to concrete objects in the pictures (train, tennis, toilet, baseball, etc.). For GRU1, while there are a few nouns (table, train, baseball), the network also highlighted prepositions such as "in" and "of". We believe the network did so because the words preceding preposition are nouns that refer to the main object of the image (e.g. a *stop sign* in a town). Surprisingly, GRU1's attention mechanism also highlighted the determiner "a" even though such word (or the preceding word) does not convey much information on the content of the image.

When we take a closer look at which words are highlighted by the best STAIR model (Tables B.3 and B.4), we notice that out of the top 10 highlighted words, respectively 3 and 4 words for GRU1 and GRU5 are particles. For GRU1 the most highlighted particles are "ga" (subject marker), "no" (genitive marker) and "o" (object marker). For GRU5, the most highlighted particles are "ga", "no", "o" and "ni" (locative marker). Even though GRU1 highlights less particles overall, the 3 most highlighted words remain particles. For GRU5, we notice that the network has learnt to highlight what seems to be the most useful particle overall which is the "ga" particle which is used to signal that the preceding word is the subject of the sentence. Focusing on this particle is especially useful as there is a strong probability that the preceding word refers to the main object of the image. For GRU5, 13.18% of the attention peaks are located above this "ga" particle, when randomness would predict 3.45% of peaks on this word. This shows that the network has learnt to deliberately focus on this particle. Otherwise, as for the models trained on the COCO data set, the other words among the top 10 words refer to nouns which are concrete objects widely represented in the corpus: skateboards, dogs, cats, etc.

The strategy of the network to highlight particles is very interesting as, by design, because of the uni-directional nature of the recurrent units used; and because particles are suffixed words, the vectors that constitute a particle contain a lot of information concerning the previous word. Consequently, highlighting particles is the optimal strategy.

4.4.6 Peak Position

In this section, we analyse specifically where the attention peaks are located above words. To do so, we divided each word beneath a peak into four equal parts and we count the percentage of peaks located above a given part. This allows us to analyse if attention waits until the end of a word to peak (and thus attention peaks could be considered as word boundary markers) or if attention tends to peak before the end of a word. Results are shown in Table 4.3.

For the models trained on English, we observe that the attention peaks are not located precisely at the end of a word, but are rather located between the middle and the end of the highlighted words. This seems to indicate that the network does not focus on the representation of the entire word, but rather on the vectors representing the first half or two thirds of a word. We also observe that GRU1 and GRU5 have a different pattern: while GRU5's peaks clearly highlight word endings, attention peaks for GRU1 seem to highlight in a higher proportion the middle of words.

For Japanese, we observe a similar behaviour as both GRU1 and GRU5's peaks are globally located above word endings. However, the situation is less as clear-cut as for English as we notice that also a large part of the peaks are located above other word parts and peak distribution tends to be more uniform. We explain this by the fact that attention peaks over Japanese captions are located above particles. We suspect that particles are able to trigger the attention mechanism as usually attention peaks are located at the very beginning of the particle or at the boundary with the preceding word (see both peak of Figure 4.1b located at the beginning of the "ni" and "ga" particles). Thus, the distribution of the attention peaks above a given part of word tends to be more uniform.

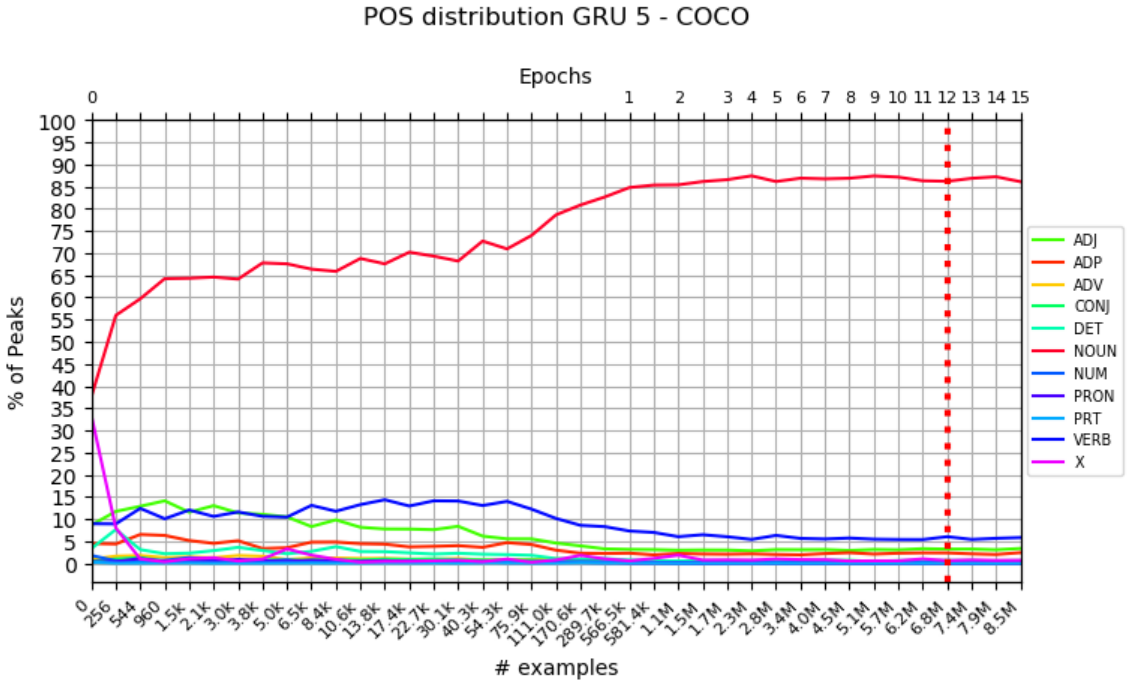
Data Set	Attn. Pos.	Beginning	Middle-Beginning	Middle-End	End
COCO	**GRU1**	17.47 ± 10.95	23.31 ± 11.65	**44.46** ± 28.58	14.75 ± 7.25
	GRU5	3.90 ± 0.48	14.11 ± 2.07	**43.32** ± 1.73	38.66 ± 2.50
STAIR	**GRU1**	17.98 ± 4.29	25.03 ± 0.77	**35.71** ± 2.18	21.28 ± 4.08
	GRU5	20.69 ± 1.29	14.56 ± 1.02	22.07 ± 2.57	**42.68** ± 0.87

Table 4.3: Dsitribution of the attention peaks above words for the COCO and STAIR data sets.

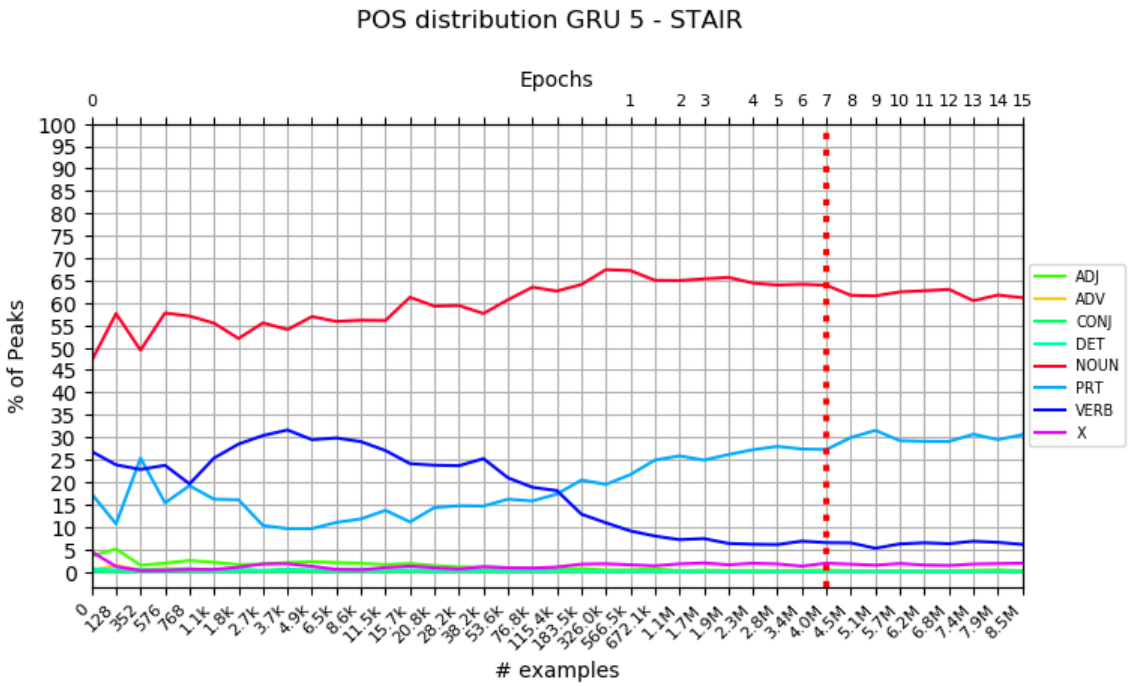
The fact that attention mechanisms peak before the end of a word seems to show that the models do not need to have access to the full version of the word in order to recognise it, but rather peak as soon as the minimum information necessary to recognise it is present. This matter is explored further in Chapter 5.

4.4.7 Longitudinal Study

In the previous sections, we showed that the models trained on the COCO and the STAIR data sets used attention to highlight specific words in the captions. We showed that such behaviour was learnt as the highlighted units differ widly from what would be expected if attention peaks were randomly positionned. This begs however the question of when and how quickly this behaviour is acquired. To explore this question, we regularly saved the



(a)



(b)

Figure 4.4: Evolution across epochs of the proportion of peaks above a given POS for (4.4a) COCO and (4.4b) STAIR. The percentage of peaks above a given POS is shown on the y-axis. The number of training examples seen at any given point of the graph is shown on the bottom x-axis. The point marking the last saving step of an epoch is shown on the top x-axis. Best saving time step (selected on the validation set) is shown by a bold dotted red vectical line. Results obtained with randomly initialised weights correspond to the first tick of both x-axes labelled "0".

models during training so as to have a longitudinal vision of how learning takes place. To do so, we saved the models' weights every time the computed loss decreased by 4% of its initial value. We also save the model at each epoch.

We will here focus on the learning pattern of the best model (out of the five trained models),[3] and more specifically we will analyse the learning pattern of GRU5 as we have shown that, for both COCO and STAIR, this is the attention mechanism that has the most interesting pattern.

We first notice that for both COCO (4.4a) and STAIR (4.4b) the savings steps belonging to the first epoch represent a large part of the graph, showing the loss dropped quite fast and that the most important learning phase is actually concentrated in the first epoch. This impression is confirmed when we observe the evolution of the mean rank $\widetilde{r}$ (reported in Appendix D) which drops from 2504 to 42 at the end of the first epoch for COCO and from 2529 to 44 for STAIR, while the minimum $\widetilde{r}$ of 27 is obtained 7 epochs later for COCO and the minimum $\widetilde{r}$ of 29 for STAIR is obtained 6 epochs later.

For COCO, we notice a clear gap between the proportion of peaks highlighting nouns when the model uses randomly initialised weights (first tick on the x-axis) and the next saving steps where peaks above nouns have gone from 37.4% up to 55.9% after only 8 batches (256 examples). After 30 batches (960 examples), the proportion of nouns under attention peaks already reaches 64.2%. This shows that it is possible for the network to focus on very specific parts of the spoken input with very a few number examples. At the end of the first epoch, the proportion of peaks above nouns does not evolve much and remains steady, hovering at 85%. We may also notice that, while there are few peaks above other POS during the first epoch, they are then barely highlighted past the first epoch. Interestingly, it seems the network uses verbs at the beginning (as the proportion slightly increases in the first part of the first epoch) and stop highlighting such units later on (-10pp when we compare the proportion at the end of the second epoch with the middle of the first epoch).

For STAIR, the behaviour of attention displays a very different pattern compared to COCO and seems to be much more exploratory. Indeed, we notice that in the first steps of the first epoch, there is a competition between POS, whereby attention highlights simultaneously nouns, verbs, and particles. In the first steps (up to 4.9k examples), the proportion of peaks highlighting particles decreases while the proportion of verbs increases. Then, this behaviour stops, and the number of peaks highlighting particles surpasses that of verbs. We notice such behaviour on 4 out of the 5 models trained on Japanese. Therefore, it seems that in order to properly converge, the models first need to highlight verbs, and then switch to highlighting particles. Also, contrary to COCO where attention does not evolve much after the first epoch, we observe the contrary for STAIR. Nevertheless, we also notice that after the first epoch, the proportion of peaks above nouns decreases as much as the proportion of peaks above particles increases. It shows the model is still adapting its decision strategy after the first epoch and favours particles over nouns.

4.5 Experiments on Natural Speech: Flickr8k

In the previous sections, we presented the results we obtained on two data sets that consist of synthetically spoken captions in English and in Japanese. We justified the use of synthetically spoken captions as no bilingual data set is currently openly available.[4] This

[3] All models have a similar behaviour.

[4] Bilingual speech-image data sets actually do exist and are based on the data set introduced by Harwath et al. (2016): namely, a Hindi data set Harwath, Chuang & Glass (2018) and more recently a Japanese data

enabled us to put forward that neural models adapt their learning strategy according to the language used. However, using synthetic speech introduces a bias as it is clearer and simpler than human speech. We thus wanted to test the network's behaviour with real human speech to see if we obtained consistent results.

4.5.1 Experimental Settings and Results

We trained the model on the Flickr8k data set which is much smaller than the COCO and the STAIR data sets (see 3.2.2). This data set consists of real human speech, which features multiple speakers. The quality of the recordings is very uneven (some have a very high background noise), making this data set really challenging. We thus expect the results to be very lower than those previously presented. However, as previously shown, only a few examples are necessary for the network to highlight nouns. Consequently, if the network has a similar behaviour on natural speech, we should also be able to observe this fact.

The network's architecture used for this experiment is the same as in the previous experiments. However, we used 1024 recurrent units instead of 512 as the acoustic vectors are larger (13 MFCC coefficients, delta, and delta-deltas). This also leaves enough space for the network to take into account and encode intra- and inter-speaker variation.

Model	**R@1**	**R@5**	**R@10**	$\widetilde{r}$
Our Models	2.08 ± 0.24	7.96 ± 0.57	12.94 ± 1.20	113 ± 9.08
Chrupała et al. (2017*a*)	5.5	16.3	25.3	48

Table 4.4: Recall at 1, 5, and 10 (in %) as well as median rank $\widetilde{r}$ (±standard deviation) on a speech-image retrieval task on the test set of our data sets (1k images) averaged over five runs with different seeds. Models were selected according to the highest R@1 on the validation set. We report the results obtained by Chrupała et al. (2017*a*) with the original RHN implementation. Chance R@k are 0.001/0.005/0.01, chance median rank $\widetilde{r}$ is 500.

As previously observed for the COCO data set, our results are lower than those reported by Chrupała et al. (2017*a*). The results we obtain are also worse than those obtained on the COCO and STAIR data sets. This demonstrates how difficult it is for the network to model the variation found in real speech. Nevertheless, even though the results are not particularly good, they are still far better than chance, showing the network was still able to make sense of the data and has learnt a reliable speech-to-image mapping.

4.5.2 Random Attention

As with the previous experiments, we shuffled the attention weights in order to understand if they were meaningful or not. The results are presented in Table 4.5.

Here also, we notice that randomly shuffling the attention weights of both attention mechanisms yields a R@1 which is much lower than the original R@1 (−1.54pp). Once again, this shows that attention plays a vital role for the models. However, contrary to the previous experiments, we notice that while randomly shuffling the attention weights of the first attention mechanism really hurts the performance of the model, shuffling the attention weights of the fifth attention mechanism results in a very small loss of the R@1. This behaviour was observed across all 5 runs. We observed the opposite pattern for COCO and STAIR.

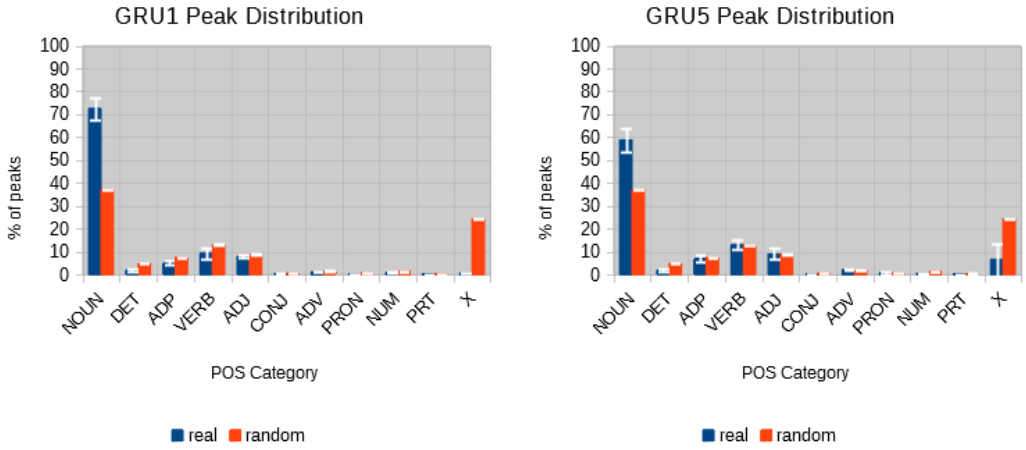
set Ohishi et al. (2020). However, these data sets are not publicly available, and it was thus impossible to use them.

This tends to show that the majority of the models' predictions rely on the context vector computed by the first attention mechanism rather than that of the fifth.

	Rand. Attn GRU1 Rand. Attn GRU5	**Rand. Attn GRU1 True Attn GRU5**	**True Attn. GRU1 Rand Attn GRU5**
Flickr8k	0.54 ± 0.82	0.44 ± 0.22	1.76 ± 0.22

Table 4.5: R@1 of models where attention weights are randomly shuffled.

4.5.3 Highlighted POS and Highlighted Words



(a) (b)

Figure 4.5: Bar plots showing the proportion of attention peaks above each POS for (4.3a) GRU1 and (4.3b) GRU5 on the Flickr8k data set. Real peaks are shown in blue and random peaks are shown in red. The results are averaged over five runs with different seeds. Error bars represent ± standard deviation.

As for COCO, we notice that the models have learnt to focus primarily on nouns. Both attention mechanisms highlight more nouns than what randomness would predict (which would be 37%) also demonstrating the network deliberately focuses on nouns rather than any other POS. However, in this case and contrary to COCO and STAIR, GRU1 highlighted more nouns overall ($72.58\% \pm 4.94$) than GRU5 ($58.73\% \pm 5.30$). The difference in the R@1 we observed when shuffling the attention weights of GRU1 and GRU5 tends to show that highlighting nouns is not sufficient, otherwise we should have observed much of a gap in R@1. What thus seems most important is to highlight the key nouns in the captions.

When looking more closely at which specific words are highlighted by each attention mechanism of the best model (shown in Appendix B), we notice that the top 10 words highlighted by GRU1 are only nouns referring to objects present in the images (dog, man, girl, boy, dogs, people, woman, child, ball, water). These words are much more highlighted than what randomness woud predict: for example, 13.49% of GRU1's peaks highlight the word "dog" when only 1.53% random peaks highlight the same word. This shows the model has learnt to detect and highlight specific acoustic patterns corresponding to what most likely is the main object of the image. GRU5 top 10 words also contain nouns, but also

contain silences, prepositions and determiners (</s>, water, <sil>, beach, snow, grass, shirt, street, a, in). Hence, this attention mechanism seems less specialised than the first one. 5.29% of GRU5's peak highlight the final silence of the caption, showing this attention mechanism favours a representation that accounts for the whole sentence rather than specific parts of the input signal.

4.5.4 Peak Position and Longitudinal Study

In this experiment also, we analysed above which word parts the attention peaks are specifically located. We observe that peaks also tend to favour word endings. As for COCO and STAIR, peaks are not located at the very end of the highlighted words, but seem to be more concentrated in the middle as the highest proportions of peaks are to be found in the middle-beginning and middle-end parts of words. Once again, these results tend to show that the model does not need to have access to the full word in order to recognise it and properly highlight it. We however notice for GRU5 that the peaks seem to be more evenly distributed over word parts, which confirms our previous intuition that GRU5 is an attention mechanism that highlights less specific units when dealing with natural speech.

Data Set	Attn. Pos.	Beginning	Middle-Beginning	Middle-End	End
Flickr8k	GRU1	12.09 ± 2.28	31.18 ± 2.75	$\mathbf{42.35} \pm 2.00$	14.38 ± 2.74
	GRU5	18.72 ± 7.95	29.38 ± 5.03	$\mathbf{31.05} \pm 6.58$	23.84 ± 2.98

Table 4.6: Repartition of the attention peaks above words for the Flickr8k data sets.

Figure 4.6 shows how the proportion of highlighted POS evolves over time. We will focus on GRU1 as it is the most interpretable attention mechanism of the model. We notice also the same phenomenon as what was observed for GRU5 for COCO and STAIR: we observe a large gap between the proportion of peaks highlighting nouns when the model uses randomly initialised weights (first tick on the x-axis) and the next saving steps, going from 40% up to 60% two saving steps later (amounting to 2,912 seen examples). The proportion of peaks highlighting other POS varies slightly in the first few steps and quickly decreases and stabilises to its final value. This shows that the model identified nouns as being useful to find the matching image. The proportion of peaks above nouns increases steadily during the first epoch. Yet, contrary to COCO and STAIR where the proportion did not evolve much after the first epoch, we notice here that the proportion of peaks above nouns still evolves after the first epoch. Indeed, at the end of the first epoch, the proportion of peaks above nouns is 72%, while it is 77% at the last epoch. This confirms that the model needs more time in order to precisely identify all the important words in the caption. In our opinion, this is to be explained by the fact that Flickr8k consists of real human speech. Thus, the model needs more time in order to account for intra- and inter-speaker variation.

4.6 Relationship with Language Acquisition

One question we are drawn to ask at this point is the following: are there any commonalities between what we know of language acquisition in children and the learning patterns we put forward in the VGS models we studied?

The fact that our models preferably put forward nouns over all other POS seems coherent to what is observed in the language acquisition literature. This phenomenon is commonly referred to as the "noun bias". Indeed, it has been found that children's lexicon – both in perception and production – contains a higher proportion of nouns than verbs or any other

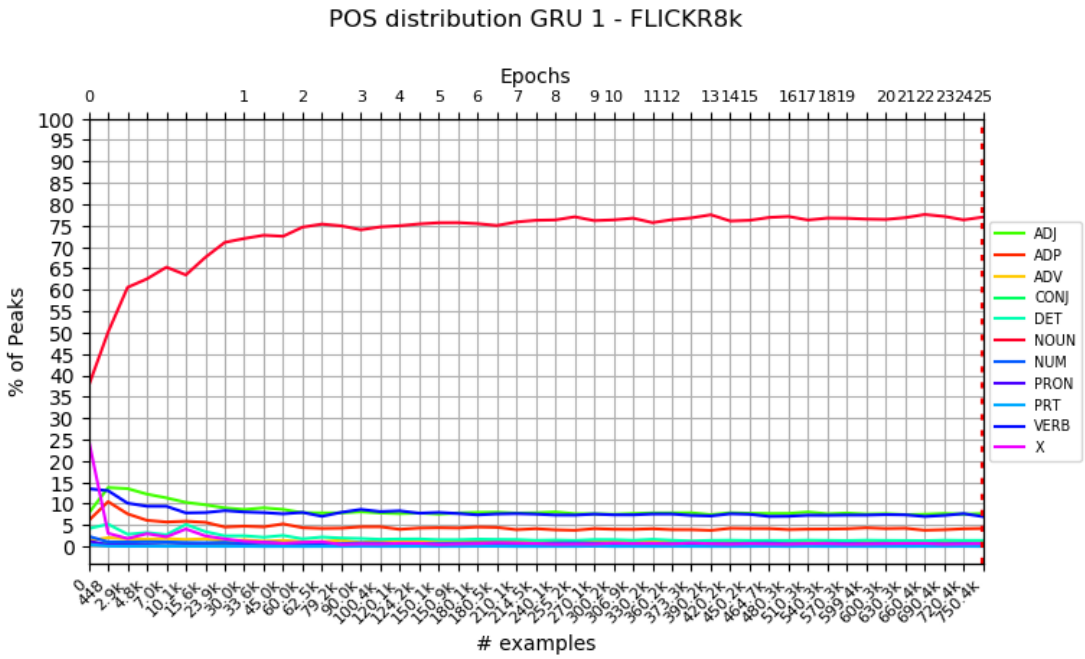


Figure 4.6: Evolution across epochs of the proportion of peaks above a given POS on the Flickr8k data set.

POS. Gentner (1982) was the first to postulate that nouns are easier to learn than verbs. She explains this by the fact that "words that refer to concepts are easy to learn because the child has already formed object concepts, and need only match words and concepts". Our neural network is in such case as it uses pre-trained VGG vectors that encode which objects are present or not in the image. Also, she explains that "noun bias" is also to be explained by the fact that nouns represent objects that are "more salient, or more stable" as opposed to verbs, whose main function is to describe action and therefore whose referent is more evanescent. McDonough et al. (2011) use the notion of "imageability" to explain this behaviour, by which nouns or verbs for which it is easier to recall a mental image are learnt before other words. Even though a few studies have concluded to the absence of a "noun bias", recent studies (Bornstein et al. 2004) have shown that Gentner (1982) conclusions hold true for a wide range of languages. Nevertheless, we should not lose sight of the fact that the task our models solve is not the same as that of a child discovering the world and that our data is biased. First, our data set consists of static images and not video clips. Second, our data sets consist of images where there are more pictures depicting static scenes than dynamic scenes, thus prompting the annotators to use more nouns than verbs. Given these biases, it seems only natural that the network highlighted more nouns than verbs. We suppose that VGS models that operated on video clips, such as that recently proposed by Rouditchenko et al. (2020) should highlight a higher proportion of verbs.

We showed that our Japanese model develops a language-specific behaviour when it mainly focuses on "ga" particles. Haryu & Kajikawa (2016) observed that Japanese children (from 15 months on) also make use of "ga" to segment speech. Our models thus adopted the same strategy as Japanese children to segment the adjacent noun. Haryu & Kajikawa (2016) state that "it is clear that noun particles are not the earliest cue infants use for word segmentation". We do observe this type of pattern for our models where particles are barely

used in the first learning steps, but become more and more highlighted, while nouns are less prevalent than at the beginning.

Finally, is there any evidence that children pay more attention to word endings, such as what we observed with attention peaks which tend to highlight last half of the words? Such a proposal was made by Slobin (Ferguson & Slobin 1973, p. 191) in his list of Operating Principles, where he states that children "Pay attention to the ends of words". He proposed this operating principle as he noticed that post-verbal and post-nominal grammatical markers (and more generally every grammatical inflexion that is suffixed) are acquired by children before other prefixed grammatical realisations. Clark (Slobin 1985, Chapter 7, p. 761) notes that this principle "pertains to children who are trying to understand the language being spoken around them" and does not apply to production. This principle seems to also be used by our models, as they highlight word endings more than any other word parts. However, this behaviour is more clearly illustrated by our Japanese model where 42.6% of GRU5' peak are located at the end of the word. This is to be explained by the fact that our models favour particles which are postposed.

4.7 Chapter Summary

In this chapter, we trained VGS models on two languages (English and Japanese) and we analysed the behaviour of their attention mechanisms. Our analysis revealed that attention adopted a language general behaviour by which it learns to detect and highlight specific nouns in the spoken input. Our experiments thus confirm the intuition of Chrupała et al. (2017*a*) stating that "the speech model's attention mechanism enables it to cherry pick key fragments of [...] utterances". But we also showed that attention could also adopt a specific behaviour based on the language the network is processing, such as what we observed for Japanese where attention mainly heavily relies on particles. Our analyses also revealed that attention quickly learns to focus on nouns and that such behaviour does not require much training data. With less than a thousand examples, attention already focuses on nouns more than what randomness would predict.

Also, we observed that contrary to what we expected, attention does not peak at word boundaries but rather peaks between the middle and the end of words. This shows that attention does not need to have access to the full word. This however begs the question of whether the implicit segmentation carried out by the network segments word into units that correspond to written chunks, or rather segments words into shorter units. We explore this question in the following chapter.

The fact we do not observe much difference between the two attention mechanisms (be it between the highlighted POS, peak position, etc.) might only be a consequence of the way both attented vectors are merged. Indeed, we used dot-product to merge both vectors which equivalates an un-normalised cosine similarity. Indeed, recall that cosine similarity is $\frac{\vec{u}\cdot\vec{i}}{\|\vec{u}\|\left\|\vec{i}\right\|}$, which normalises the dot-product by the norm of both vectors $\|\vec{u}\|\left\|\vec{i}\right\|$. This surely explains why the difference in attention weights and peaks is minimal.

Finally, we tried to show the relationships that existed between what is known of language acquisition in children and the behaviour of attention in a neural network and have shown that some behaviours seem to be similar, the main one being that nouns are the most favoured POS categories.

Chapter 5

Word Activation, Competition, and Recognition

The work presented in this chapter is based on the article we published at CoNLL2019 Havard et al. (2019*c*).

Contents

5.1 Introduction

In the previous chapter, we showed that VGS models were able to highlight specific words in the speech stream using their attention mechanisms. This implies that the model is able to recognise the words which are being highlighted. This aptitude raises a few questions which we will try to answer in the present chapter.

First, we explore if the model is able to map an individual word to its visual referent. That is, instead of presenting full captions to the network we present individual words. If the network is able to retrieve images that feature the object the spoken word refers to, it means the network was able to segment that word from the speech stream and thus proceeded to an implict segmentation.

Second, we explore if all the words that correspond to the main objects the COCO data set was conceived around are equally well recognised by the network or not. We explore a few factors that could influence speech/image mapping in the network.

Third, we examine how words are recognised by the network. The fact that specific words can be highlighted implies that the model was able to store a certain representation of the most important words of the data set and is able to activate these representations when processing the spoken caption. To explore this question, we introduce a methodology originating from psycholinguistics – the gating paradigm (Grosjean 1980) – and use it to analyse the representation learned by our model. We put forward in the previous chapter that the attention mechanisms need not see full words in order to highlight them, but rather peaked before the end of the words. This seems to imply that the network is able to recognise a specific word from a partial input. This is what we will study, namely how much of a word the network needs to see in order to be able to recognise it and map it to its visual referent.

Finally, we examine if there is a form of lexical competition when the network activates words that start with the same sequence of phonemes, such as it was found in human speech processing.

5.2 Word Recognition

The fact that VGS models are able to recognise individual words was already explored by several studies. For instance, Chrupała et al. (2017*a*) and more recenlty Merkx et al. (2019) showed that the utterance embeddings computed by RNN-based VGS models contain information about the presence of individual words. To do so, they simply encode a full caption using the speech encoder of their network, encode an individual word, and train a probing classifier to predict whether the individual word was present or not in the full caption.

However, these studies did not show for what type of words this behaviour holds true and if the model had learnt to map these individual words to their visual referents. Also, none of these studies explored the factors that influence word recognition and why some words seem to be fairly well recognised while other are not recognised at all. It should be noted that Harwath & Glass (2017) and Harwath, Recasens, Surís, Chuang, Torralba & Glass (2018) already observed that CNN-based models can reliably map word-like units to their corresponding visual reference. Therefore, we expect RNN-based VGS model to display a similar behaviour.

5.2.1 Isolated Word Mapping

To explore if the model is able to map isolated words to their visual referent, we selected a set of 80 words corresponding to the name of 80 object categories in the MSCOCO data set.[1] We expect our model to be very efficient with these specific words, as these refer to the main objects featured in the COCO data set.

We generated speech signals for these 80 isolated words using Google's TTS system and then extracted MFCC features for each of the generated words. In this experiment, we evaluate the ability of the model to rank the images so that at least one image in the first top 10 images contains an object instance corresponding to the target word (Precision@10, abbreviated P@10).[2] Contrary to Chrupała et al. (2017*a*) who uses Recall@k, we use

[1] List available at `https://github.com/amikelive/coco-labels/blob/master/coco-labels-2014_2017.txt`

[2] Evaluation is performed on the test set containing 5000 images.

Precision@k as there are several images that may be correctly associated to a single target word: e.g. if the network is prompted with the word "elephant", every image that features an elephant should be considered a valid association. An important point to remember is that at training time, the network was only given full captions and not isolated words. Hence, if the network is able to retrieve images featuring instances of the target word, it shows that implicit segmentation was carried out at training time.

Results are shown in Figure 5.1. 40 words out of the 80 target words have a P@10$\geq$ 0.8. This shows that the network is able to reliably map isolated words to their visual referent despite never having seen them in isolation. Conversely, we also notice that 15 words are not mapped to their visual referents.

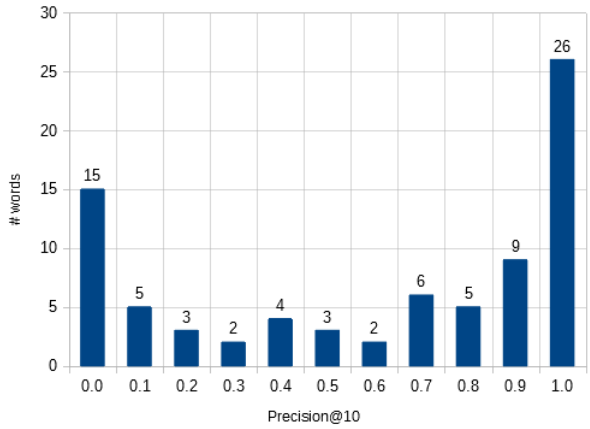


Figure 5.1: Precision@10 for the 80 isolated words corresponding to MSCOCO categories.

A list of the 80 words used for this experiment along with P@10 for each of them is available in Appendix D. Among the words that are the best recognised are animals (elephant, zebra, sheep, giraffe) and objects (truck, bus, boat, airplane) that are very frequent in our data set. Among the words that are the least well recognised are objects (fork, knife, vase, toaster) or animals (mouse) that are quite infrequently mentionned in the captions as they might be too small and not described by the annotators. However, among the least well recognised words are also common objects (such as skis or frisbee) that are very frequent in the captions. Manual exploration of the attention weights shows the model also peaks on such words (see Appendix B where "skis" and "frisbee" are respectively the 28^{th} and 33^{rd} most highlighted words). We explain this absence of recognition here by a competition effect whereby the network favours the representation of other words. For example, "skis", "snow", "hill" and "snowboard" frequently cooccur. The network might thus favour the other words in order to learn a reliable image mapping instead of our target word.

5.2.2 Factors Influencing Word Mapping

We explore here in more details the factors that could come at play in the recognition of isolated words. We explore 2 types of factors: image-related factors and speech-related factors. For the latter, we consider word frequency in the training set and length of the speech signal. Concerning image related factors we consider object instance frequency in the images of the training set, average number of neighbouring object instances, and average area of each object. We decided to consider these variables as failures to map a spoken

word to its visual referent could only be due to the fact that the pre-trained VGG network, from which the image features are extracted, fails to detect and thus encode the presence of a given object.

To model the relationship between all these variables, we fit a multiple linear regression model with R where we try to predict Precision@10 using the previously mentioned factors. Results are shown in Table 5.1. We notice that the only effect that plays a significant role in word recognition is the frequency of the word in the training set. The positive effect shows that the more frequent the word is in the training set, the better it is recognised. Word length has a mild positive effect which tends to show that longer words are better recognised than shorter words. Object frequency and number of neighbouring object have no effect. Object size seems to play a mild effect also. This seems to be coherent with the fact that more frequent words are better recognised, as we expect objects that occupy a large place in picture to be more often described than smaller objects. Consequently, such results seem to confirm our previously exposed hypothesis stating that words denoting small objects are often not described.

Factors		**Estimate**	**p-value**
Images	# Neighbouring Objects	-0.054178	0.123
	Object Size	+0.010932	0.048
	Object Frequency	-0.006001	0.421
Speech	Word Frequency	+0.165963	0.007
	Word Length	+0.384533	0.090

Table 5.1: Factors influencing word recognition performance in our model.

Our results thus show that individual words are indeed reliably mapped to their visual referent by the network. The main factor of success in this task being the frequency of the target words in the caption as well as the size of the objects in the image. Hence, words that are very frequent and refer to large objects are better recognised than others.

5.3 Word Activation

In this section we describe how individual words are activated by the network. To do so, we perform an ablation experiment (similar to that of Grosjean (1980) which was conducted on humans) where the neural model is inputted with only a truncated version of the 80 target words. Such a method is also called *gating* in the psycholinguistic literature.

5.3.1 Gating Paradigm

The gating paradigm was introduced by Grosjean (1980) and involves the following procedure:

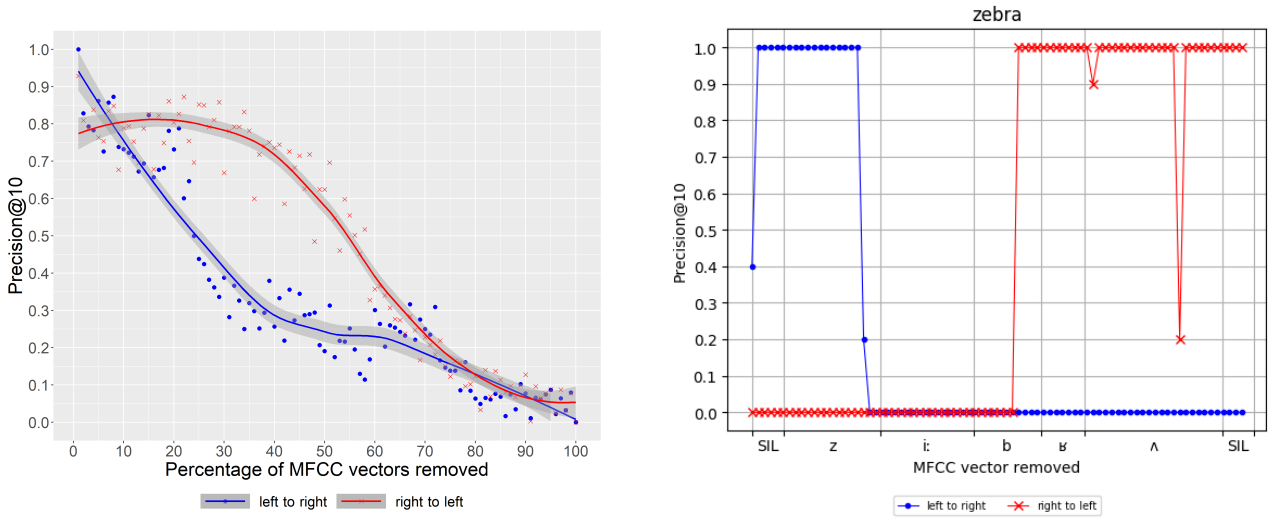
> The gating paradigm involves the repeated presentation of a spoken stimulus (in this case, a word) such that its duration from onset is increased with each successive presentation. This is done until the entire word has been presented. After each presentation (or gate), subjects are asked to write down the word being presented and to rate their confidence in each guess. (Cotton & Grosjean 1984)

In our case, it means the neural model is fed with truncated versions of a given target word, each truncated version comprising a larger part of the target word. Contrary to

Grosjean (1980) who only proceeds to a truncation that preserves the beginning of a word, we also proceed to a truncation that only preverses the end of a word. That is, in our case, truncation is either done left to right (the model only has access to the end of the word) or right to left (the model only has access to the beginning of the word). Also, contrary to the original setting where experimenters are asked to write the target word after each successive presentation, we evaluate the models' ability to rank the images so that the top k images contain instances of the target object.[3] The COHORT model, in its initial version (Marslen-Wilson 1987*b*), stipulates that word onset plays a crucial role in word recognition. The aim of this experiment is to test whether word onset plays a role in word recognition for the network or not. If it is the case, we expect the network to fail to recover images of the target word if the word is truncated left to right, but not — or less — when the word is truncated right to left, hence motivating the truncation from both ends.

Truncation is operated on the MFCC vectors computed for each individual word, meaning that MFCC vectors are iteratively removed either from the beginning of the word or the end of the word, but not from both sides at the same time. Each truncated version of the word is then fed to the speech encoder which outputs an embedding vector. As in our previous experiment, the model's performance is evaluated in terms of P@10.

5.3.2 Effect of Gating



(a) (b)

Figure 5.2: 5.2a Evolution of Precision@10 averaged over 80 test words as a function of the percentage of MFCC vectors removed for each word. 5.2b Evolution of Precision@10 for each ablation step of the word "zebra", with time-aligned phonemic transcription /ziːbʁʌ/ at the bottom. "SIL" signals silences. For both 5.2a and 5.2b, blue lines show P@10 when ablation is carried out left to right, meaning that at any given part on the blue curve, the model has only had access to the rightmost part of the word. (e.g. /iːbʁʌ/ without initial /z/). Red line displays scores when ablation was carried out right to left, meaning that at any given part on the red curve, model has only had access to the leftmost part of the word. (e.g. /ziː/ without the final /bʁʌ/).

[3] In fact, it is impossible to reproduce the original experimental settings with our model as it does not predict discrete labels. It would be however possible to use the original settings with a network trained to predict word labels, such as an Automatic Speech Recognition model.

Figure 5.2b shows the evolution of P@10 for one of the target words ("zebra"). When MFCC vectors corresponding to the first phoneme are removed (/z/), precision plummets from 1 to 0. However, when MFCC vectors belonging to the end of the word are removed, precision plateaus at 1 until /iː/ is reached and then plunges to 0. This shows the model successfully retrieved pictures of zebras when only prompted with /ziːb/ but not when prompted with /iːbʁʌ/ even though the latter comprises a longer part of the target word. It also seems that little acoustic differences may yield very different representation as depicted by the red trough, were the only difference with the previous two points being the addition or deletion of an MFCC vector.

Figure 5.2a shows evolution of P@10 averaged over the 80 test words. As can be seen from the graph, precision evolves differently according to which part of the word was truncated. When the target words are truncated left to right, precision drops quicker than when truncated right to left. These results show that the model is robust to truncation when it is carried out right to left but not when it is carried out left to right: when the initial phonemes of the words are removed, the network fails to retrieve the target image, but when only presented with the initial phonemes, the network is globally able to retrieve images of the target word. These results suggest that the model does not rely on a vague acoustic pattern to activate the semantic representation of a given concept, but needs to have access to the first phonemes in order to yield an appropriate representation.

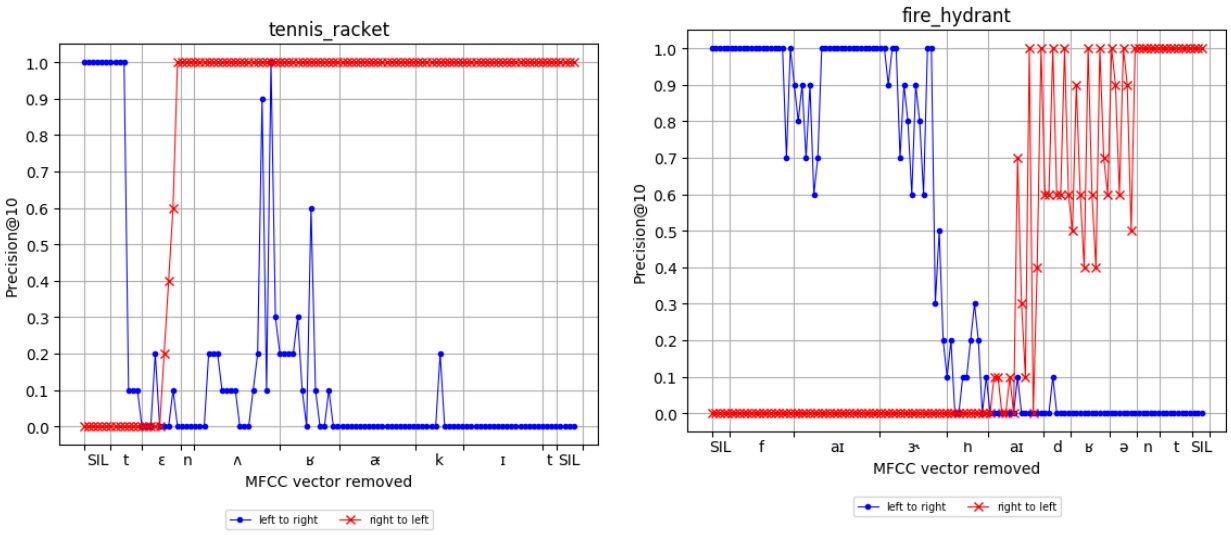
5.3.3 Activated Pseudo-Words

Such ablation experiments also enable us to infer on what units the network relies on to make its predictions. Figure 5.3 for example allows us to see what are the pseudo-words that were internalised by the network for the word "tennis racket". When truncation is done left to right (blue curve), we notice that at the beginning precision is high (1.0), then reaches 0 when only /ɛnɪsrækɪt/ is left, but suddenly increases up to 1 when the only part left is /rækɪt/. This suggests that the network mapped both "tennis racket" as a whole and "racket" as referring to the same object.

Figure 5.3b shows that removing the first part of the word "fire hydrant" has nearly no effect on P@10. Indeed, even if a large part of the word "fire" is removed, P@10 remains high. Yet, once the /ɝ/ is reached, P@10 abruptly decreases. We also notice it is possible to remove a large part of the word "hydrant" up to the /d/ without hurting much the performances of the network. This tends to show that only a sub-part of the word is necessary for the network to produce the appropriate representation and not the word "fire hydrant" as a whole. It thus seems that the minimal units that are necessary to activate the word fire hydrant are /ɝhaɪ/. Therefore, we need to take caution when stating that the network has isolated words, as the words internalised by the network might not always match the human gold reference.

Finally, Figure 5.2b shows that when only prompted with /ziːb/ the network manages to find pictures of zebras. This suggests that only the first part of the word is necessary to activate the representation of "zebra". This might be due to the fact that the vocabulary the network has to deal with is very limited and thus the network is confident from that point on that the rest of the word will refer to a zebra. This is similar to what happens in French with the pseudo-word "coquelic" which activates the word "coquelicot" (poppy) as this is the only word in French which starts with "coquelic".

This experiment hence shows that the pseudo-words internalised by the network do not necessarily correspond to a graphic word. It also shows that only sub-parts of a given word are necessary for the network to activate the target representation.



(a) (b)

Figure 5.3: Evolution of P@10 for each ablation step of the word 5.3a "tennis racket" and 5.3b "fire hydrant" with time aligned phonemic transcription at the bottom

5.3.4 Gradual or Abrupt Activation?

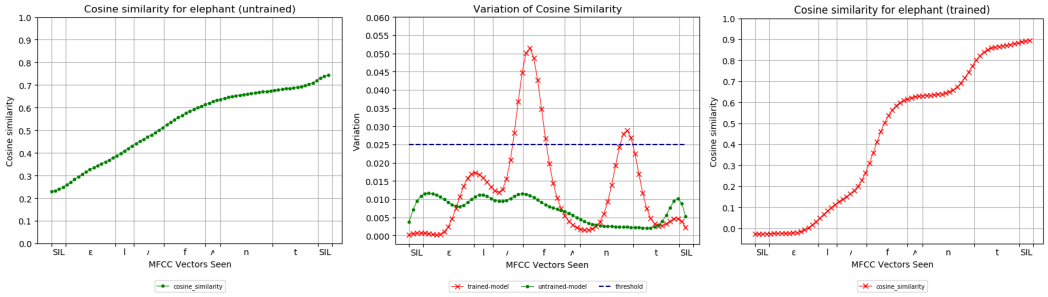
As previously mentioned, little acoustic differences yield wide differences in the final representation. Thus, in this section we analyse how representation is being constructed over time and explore if some MFCC vectors play a more important role than others in the activation of the final representation.

We progressively let the network see more and more of the MFCC vectors composing the word, iteratively feeding it with MFCC vectors starting from the beginning of the word until the network has had access to the full word. We then compute the cosine similarity between the embedding computed for each of the truncated version of the word and the embedding corresponding to the full word. The closer the cosine similarity is to 1, the more similar the two representations are. Thus, if each MFCC vector equally contributes to the final representation of the word, we expect cosine similarity to evolve linearly. However, if some MFCC vectors have a determining factor in the final representation, we expect cosine similarity to evolve in steps rather than linearly. To detect steps that could occur in the evolution of cosine similarity, we approximate its derivative by computing first order difference. High steps should thus translate into peaks (e.g. Figure 5.4b). We compute the evolution of cosine similarity for the 80 target words encoded with the best trained model (e.g. Figure 5.4c) and also consider a baseline evolution by encoding the 80 target words with an untrained model (e.g. Figure 5.4a).[4] To avoid micro-steps of yielding peaks and hence creating noise, we smooth cosine evolution curves with a gaussian filter. We consider peaks higher than 0.025 as translating a high step in the evolution of cosine similarity.

On average, they are 1.25 peaks per word for the trained model against 0.1 peak per word that are above our 0.025 threshold for our baseline condition (untrained model), showing that cosine evolution is linear in the latter, but not in the former. Consequently, in our baseline condition (untrained model), each MFCC vector equally contributes to the final representation, whereas in our trained model some MFCC vectors are more decisive for

[4] Thus consisting only of randomly initialised weights.

the final representation than others. Indeed, some MFCC vectors trigger a high step in the cosine evolution suggesting that the embedding suddenly gets closer to its final value. Figure 5.4c shows the evolution of the cosine similarity for the word "elephant". As it can be seen, cosine similarity does not tend linearly towards 1, but rather evolves in steps. Adding the MFCC vectors corresponding to the transition from /ɪ/ to /f/ triggers a large difference in the embedding as the cosine similarity suddenly jumps to a higher value, showing it is getting closer to its final representation. After the vectors of /f/ are added, the evolution of cosine similarity is modest, suggesting the final part of the words plays a less important role than the initial part in the final representation of the word.



(a) (b) (c)

Figure 5.4: Figure 5.4a shows evolution of the cosine similarity between the embeddings produced for each truncated version of the target word and the embedding for the full word using a model with randomly initialised weights. Figure 5.4c shows the same measure with the embeddings produced by a trained model. Figure 5.4b shows peaks indicating the inflection points of curve 5.4a (green) and 5.4c (red). For our experiments, we only considered inflection point to be significant if the resulting peak was higher than 0.025 (blue line).

We conclude from this experiment that the network needs not see the full word in order to activate the representation that enables images featuring this very word to be retrieved. More specifically, we put forward that the network needs to have access to the first part of the target word in order to activate the correct representation and that some word parts are more crucial than other in order to compute the appropriate representation.

5.4 Word Competition

Some psycholinguistic models (see Section 1.4) assume that the first phoneme of a word activates all the words starting by the same phoneme. The word that the speaker wants to pronounce and gradually utters is called the "target" word. The words that are activated but which do not correspond to the target word are called "competitors". As the listener perceives more and more of the target word, some competitors are deactivated. This means that they are not considered as the potential word any more, as they do not match what is being perceived.[5]

For example, let us consider that a speaker has started to utter the following sentence "I have a ..." and the following lexicon that only consists of three words that start with a /b/: /beɪbi/ (baby), /beɪsmənt/ (basement), and /beɪsbɔːl bæt/ (baseball bat). The first sound of last word of the sentence all start with /b/. Thus, all words would be activated and at this point the sentence could as well be "I have a baby", "I have a basement" or "I

[5]Note that phonetic factors are not the only factors taken into account to deactivate a set competitors, morpho-syntactic factors also come at play.

have a baseball bat". Once /beɪs/ is reached, "baby" would not be considered a competitor any more, and once /beɪsb/ is reached the only word activated would be "baseball bat" as it is the only word whose beginning corresponds to the perceived sounds. The speaker would know from that point on that the sentence is "I have a baseball bat".

5.4.1 Methodology

We tested if the network displays such lexical competition patterns. To do so, we selected a set of 57 word pairs that could potentially compete between one another. We selected the word pairs according to the following criteria:

- Words should at least appear 500 times or more in the captions of the training set, so that the network would have been able to learn a mapping between these words and their referent;
- Words forming a pair should at least start with the same phoneme;[6]
- Words should not be synonyms and clearly refer to a different visual objects (thus excluding pairs such as "motorcycle" and "motorbike").

Out of these 53 word pairs, we assessed if each word of each word pair was reliably mapped to a visual referent. We considered it was the case if at least 5 of the first 50 ranked images contained an instance of this word. Finally, out of the initial 57 word pairs, only 12 remained.[7]

For each word pair, we selected one of the word which we consider as the target word, progressively let the network see more and more of the MFCC vectors composing this word. At each time step, the network produces an embedding, which we use to rank the images from the closest matching image to the least matching image.[8] Then, for the 50 closest matching images, we check if at least one of the captions contains either the target word or the competitor. We have to use the captions to check if the concept is present in the images, as among the word pairs, most contain words that do not belong to the 80 object categories of MSCOCO. We thus have to rely on the caption to check whether the object is present in the image or not.

As the competitor and the target word start with the same phonemes, we expect the network to produce an embedding that activates both the competitor and the target word at the beginning and then, when the acoustic signal does not match the competitor any more, we expect the network to be able to activate only the target word. For each word pair, each word is alternatively used as the target word.

5.4.2 Results

We present the competition plots of three specific word pairs as they are the most representative of the patterns observed with the other word pairs.

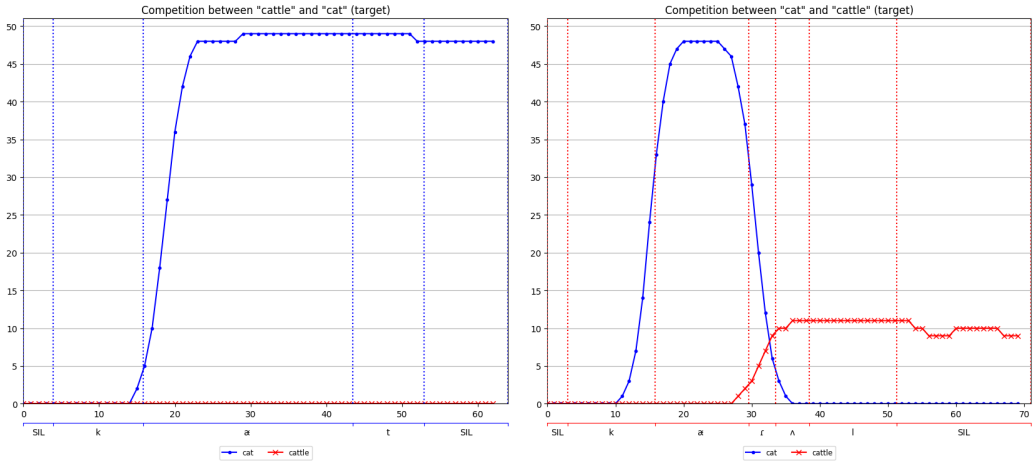
5.4.2.1 Cat/Cattle: mild competition

Figure 5.5 shows an example of competition plots between two words: "cat" and "cattle". In Figure 5.5a, the network is prompted with iteratively longer spans of the word "cat".

[6] Phonemic transcription found in CMU Pronouncing Dictionary was used

[7] wii/window, cat/cattle, cat/cow, cat/catcher, cattle/catcher, floor/flower, fridge/frisbee, kid/kitchen, player/plate, tree/train, meat/meter, and train/truck.

[8] That is, we compute the cosine distance between the embedding produced at time step t and all the images (5000) of our collection.



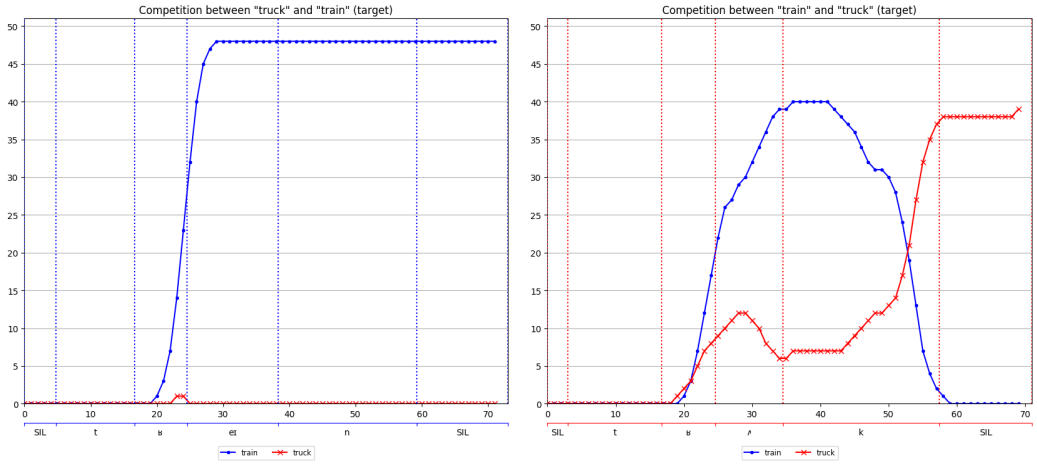
(a) (b)

Figure 5.5: Illustration of lexical competition between "cat" and "cattle". 5.5a shows competition plots when the word "cat" is used as target. 5.5b shows competition plots when the word "cattle" is used as target. Numbers in 1^{st} x-axis correspond to the number of MFCC frames of the target word; 2^{nd} x-axis corresponds to time-aligned phonemic transcription of the target word; y-axis shows number of images for which at least one caption (out of 5) contains the target or competitor word. Vertical colour bars are projection of phoneme boundaries of the target word. In order to have smoother curves, we used a gaussian blur with a standard deviation of 2.

We notice that this word only activates the representation of "cat" as out the 50 images, 49 contain a caption with the word "cat" and none of them with the word "cattle". Figure 5.5b shows what happens when we use the word "cattle" as target, where the network is prompted with iteratively longer spans of the word "cattle". We notice that at the beginning, the representation which is activated is that of "cat" as once again, more than 45 of the top 50 images have a caption that contains the word "cat". However, at the end of [æ] and the beginning of [ɾ], we notice that the number of images with a cat strongly decreases and the number of images with cattle increases. Once [ʌ] is reached, no more picture of cats are retrieved, showing this representation was totally deactivated. This example is interesting as it seems the network was able to notice fine acoustic variation such as [t]/[ɾ] in order to recognise the target word.

5.4.2.2 Train/Truck: strong competition

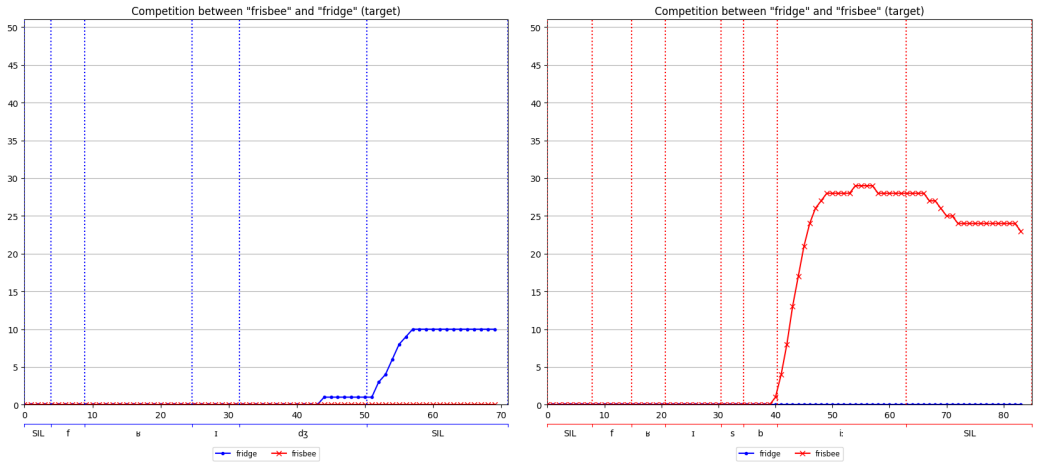
Figure 5.6 shows the competition plots between "train" and "truck". Contrary to the previous example, we notice a clear competition between "train" and "truck" when "truck" is used as target. We observe that both representations are activated at the same time (after [r]). However, we notice that the magnitude of the activation is very different as more pictures of trains are retrieved than pictures of trucks. We also notice that even though the network activated the representation of "train" when prompted with "truck", there are less pictures of "train" than when the network is prompted with the word "train". We also observe that at first the activation is very similar (both blue curves start to increase at the beginning of [r]), but once [ʌ] is reached, the curve is less steep in Figure 5.6b than in Figure 5.6a, showing the network notice it is not the canonical pronunciation of the word "train". It is only at the very end of the word that the representation of "train" deactivates and that the network retrieves pictures of trucks.



(a) (b)

Figure 5.6: Illustration of lexical competition between “train” and “truck”. 5.6a shows competition plots when the word “train” is used as target. 5.6b shows competition plots when the word “truck” is used as target.

5.4.2.3 Frigde/Frisbee: no competition



(a) (b)

Figure 5.7: Illustration of lexical competition between “fridge” and “frisbee”. 5.7a shows competition plots when the word “fridge” is used as target. 5.7b shows competition plots when the word “frisbee” is used as target.

Figure 5.7 shows the competition (or lack thereof) between “fridge” and “frisbee”. We observe that even though both words start with the same sequence of phonemes, there is no competition between them. “frigde” only activates pictures of fridges and none of “frisbee” and vice-versa. In both cases, the network needs to have access to large part of the word in order to retrieve the target word.

5.4.3 Is there any Competition?

(REVISED) COHORT Marslen-Wilson & Welsh (1978), Marslen-Wilson (1987*b*) and TRACE McClelland & Elman (1986*b*) both state that competing words are all activated at the same time, that is when the first phoneme is perceived. Even though some examples conform to this statement – such as that presented in Figure 5.6b – it is not the case of all word pairs. In some cases, there is only little competition as the representation of both words is activated sequentially (as in Figure 5.5) and in some other cases, there is literally no competition between the two words (such as 5.7) despite both words starting in a similar way.[9] Therefore, the behaviour of the network appears to be very unclear. It seems however that the network activates the word that is the most common. For example, there are many more pictures of trains than pictures where trucks constitute the main object of the image. Thus, the network seems to activate preferentially the representation of the object that is the most frequent in the images and captions.

5.5 Replication on Natural Speech

We wanted to replicate our initial experiments on natural speech using the FLICKR8K data set. Nevertheless, we were not able to do so as this was done before we had the time to do it ourselves. Indeed, Scholten et al. (2020) followed the methodology we introduced in our article Havard et al. (2019*c*) and applied it to analyse word activation and recognition in the VGS model of Merkx et al. (2019). The results they obtain confirm the trends we presented in section 5.2: they also observe that not all words are equally well recognised when presented in isolation, and report a median P@10 of 0.4 (when we report a median P@10 of 0.8). Such difference is to be expected as natural speech is harder to model than synthetic speech. Contrary to us, where we observed that longer words are better recognised than shorter words, they observe the reverse pattern. However, they also found that frequency had a positive effect such as we did, that is, words that occur more frequently in the training set are those that are recognised better in isolation.

In their study, they also find that word activation operates through a process of competition. Indeed, they show that word recognition is harder for words that have a high number of similar-sounding words – that is, words which start with the same sequence of phonemes.

Their study also puts forward that word recognition is possible even with a partial input, such as we did (see the "cat" example in Figure 5.5a, where /k/ only is enough to activate the representation of the full word). Finally, their study also show that word recognition also occurs in steps (which they call "leaps") where P@10 abruptly jumps from 0 to 1.

5.6 Chapter Summary

In this chapter, we showed that a VGS model is able to map individual words to their visual referents despite having been trained on full captions. van Zon (1997, p. 8) notes that in the COHORT and TRACE models "segmentation is the result of recognition". We believe that indiviual word recognition shows that the model implicitly segmented its inputs into subunits, and propose the converse formulation "recognition is proof of segmentation". An important point we made in this chapter is however that the resulting segmentation might not always correspond to the gold standard.

[9] However, it could be that those two words compete with other words, but not specifically these two.

In this chapter, we introduced several methodologies to analyse how the representation of individual words is built over time. Notably, we adapted the gating paradigm of Grosjean (1980) so as to analyse how words are activated by the network. We observed that word representation is not built linearly by the network and that recognition may occur with a partial input, thus corroborating that "recognition often occurs before the acoustic offset of the word" (van Zon 1997, p. 8). We showed that in order to be able to activate the correct representation of a given word, the network needs to have access to the first phonemes of this word, as when they are removed, the network is unable to activate the correct representation. Thus, when word recognition is observed with a partial input, it is only when the partial input encompasses the first part of the word, but not when it only encompasses the final part of the word.

Finally, we tried to see if word activation was done through a process of word competition, such as what was postulated in human word recognition. We observed mixed results, as we did observe competition in some cases, but this behaviour was far from systematic. Also, word competition models assume that competitor words are totally deactivated when the acoustic input does not match the internalised representation, but we found it was not the case for most of the word pairs we tested.

The work we presented in this chapter has some limitations. The first one being that we only used synthetic speech, however, this issue was addressed by Scholten et al. (2020) and confirmed the results we present. One of the issue we faced was that we were limited to a given set of words. Indeed, we were limited by the data set we used and the most frequent words it contains. We chose words according to the 80 object types the data set was conceived around. Nonetheless, all of those words are not equally well recognised by the network, and therefore the activation pattern of some words was rather unclear.

Also, the analysis of the network's representation is uneasy as we predict a vector in a continuous space, and the only way of interpreting the network's behaviour is by looking at which images are retrieved. Studying if there is any form of competition in an ASR model would be much easier as the last layer consists in a softmax over the vocabulary.

Chapter 6

Impact of Prior Linguistic Information

The work presented in this chapter is based on the article we published at CoNLL2020 Havard et al. (2020).

Contents

6.1 Introduction

In the previous chapters, we showed that RNN-based VGS speech models use their attention mechanism to highlight words that are relevant to retrieve the correct image, and that

such models implicitly segment the spoken input into sub-units. In this chapter, instead of understanding what types of units were implicitly segmented by the network, we approach the problem the other way round and ask ourselves the following question: what segmentation maximises the performance (recall@k) of a VGS model if speech were to be segmented? In order to answer this question, we explore *how* boundary information can be integrated, *which* type of boundary is the most efficient (either phone, syllable, or word), and finally *where* – that is, at which layer – such boundary should be introduced in the network. Finally, we also explore hierarchical models for which we provide multiple segmentation levels at the same time to understand the effect of explicitly modelling the structure nature of speech.

Another motivation for this experiment stems from the linguistic litterature, which shows that litteracy (i.e. the ability to read) affects language acquisition, and more specifically lexical acquisition. Havron et al. (2018) indeed showed in an artificial language learning task that preliterate children have difficulties associating a novel word to its referent while litterate children were easily able to do so. Several factors may explain this, the main one being that prelitterate infant struggle to segment the speech stream, while litterate infants are aided in this task by having seen visual cues (i.e. blanks) separating words.

Contrary to the previous chapters where we used the Theano-based model of Chrupała et al. (2017*a*), we switch to a PyTorch-based model.[1] Also, in this chapter, we only used one attention mechanism instead of two as in the previous chapters.

6.2 Boundary Information

6.2.1 Boundary Types

As previously stated, we wish to give linguistic information to our network, and more specifically with segment boundary information. In this chapter, we define a *segment* as either being a phone, a syllable, or a word. Segment boundaries were derived from the forced alignment metadata (see § 3.2.4) so as to know which MFCC vector constitutes a boundary or not. As the force aligner does not provide alignment at the syllable level, we wrote a custom script to recreate syllables from the phonemic transcription.

We consider two different types of syllables in this work: indeed, when we speak, words are not uttered one after the other in a disconnected fashion, but are rather blended together through a process called "resyllabification". In English, this phenomenon is visible when a word ending with a consonant is followed by a word starting with a vowel. In this case, the final consonant of the first word tends to be detached from it and attached to the next word, thus crossing the word boundary. This phenomenon is illustrated in Example (1) where phonemes in red indicate a resyllabification phenomenon.

(1) This is an example.

	Transcription[2]	/ðɪs#ɪz#ən#ɪgzæmpəl/
a.	*No resyllabification*	/ðɪs.ɪz.ən.ɪg.zæm.pəl/
b.	*With resyllabification*	/ðɪ.sɪ.zə.nɪg.zæm.pəl/

The two types of syllables that we consider in this work are the following: "syllable-word" that refers to syllables that result from a segmentation that does not take into account resyllabification (1-a), and "syllable-connected" that refers to syllables that result from a segmentation that takes into account resyllabification (1-b). It should be noted that in

[1] https://github.com/gchrupala/vgs

[2] We use "#" to signal word boundaries and "." to signal syllable boundaries.

the syllable-connected condition, most word boundaries are lost.[3] In the syllable-word condition however all word boundaries are preserved and the segmentation inside a word may occasionally result in a morphemic segmentation (as for example in "runway" /ɹʌn.weɪ/ or "air.plane" /ɛɹ.plein/). Yet, this is not always the case, especially for longer words that are of non-germanic origin (such as "elephant" /ɛ.lɛ.fant/ or "computer" /kəm.pju.tɚ/).

Thus, for each caption we have a sequence X of length T of d-dimensional acoustic vectors $X = \left[x_1^d, x_2^d, ..., x_T^d\right]$ and a corresponding sequence of scalars B of length T representing boundaries $B = [b_1, b_2, ..., b_T]$, $b_t \in \{0, 1\}$, where $b_t \triangleq 1$ if x_t is a segment boundary, 0 otherwise.

6.2.2 Integrating Boundary Information

In order to integrate boundary information in our network, we take advantage of its design, and more specifically of the recurrent cells and how such cells compute their output. Recurrent cells, as already mentioned in the first part of this thesis (see 2.2.3), are particularly suited to handle sequences – such as speech – where each acoustic vector cannot be independently considered, but rather depends on the preceding acoustic vectors. Recurrent neural networks can thus be formalised as follows:

$$h_t = f\left(h_{t-1}, x_t; \theta\right) \tag{6.1}$$

where the hidden state at timestep t, noted h_t is a function f of the previous timestep h_{t-1} and the current input vector x_t, and where θ is a set of learnable parameter of the function f. Hence, the final vector computed at timestep T depends on all the previous vectors, thus effectively modelling the sequential nature of the input. A special case arises for the first timestep $t = 1$ as the previous hidden state h_{t-1} does not exist. In such case, the initial state h_{t-1} – noted h_0 and called the initial state – is set to be vector of zeros.

Our approach to integrate boundary information into the recurrent layers of our network can be formalised as follows:

$$h_t = \begin{cases} f\left(h_0, x_t; \theta\right), & \text{if } b_{t-1} = 1 \\ f\left(h_{t-1}, x_t; \theta\right), & \text{otherwise} \end{cases} \tag{6.2}$$

In our approach, h_t is only dependant on the previous timestep h_{t-1} if the previous timestep was not an acoustic vector corresponding to segment boundary ($b_{t-1} \neq 0$). If the previous timestep corresponds to a segment boundary ($b_{t-1} = 1$), we reset the hidden state so that it is equal to h_0. Hence, vectors in the same segment are temporally dependent, but vectors belonging to two different segments are not. The GRUs that use this computing scheme will from now on be referred to as $\text{GRU}_{\text{PACK.}}$, as vectors belonging to the same segment are "packed" together.

As for our previous experiments, we use GRUs, but our methodology could be applied to any other type of recurrent cells such as LSTM of vanilla RNN.

6.2.3 All and Keep Conditions

From this initial $\text{GRU}_{\text{PACK.}}$ setting, we derived to different conditions: ALL and KEEP. In the ALL condition (see Figure 6.1b), all the vectors belonging to a segment are forwarded to the next layer (which can either be a recurrent layer, or an attention mechanism depending

[3] Word boundaries are not lost in the following cases: V#V and C#C when CC is not an allowed complex onset. C and V respectively refer to "consonant" and "vowel".

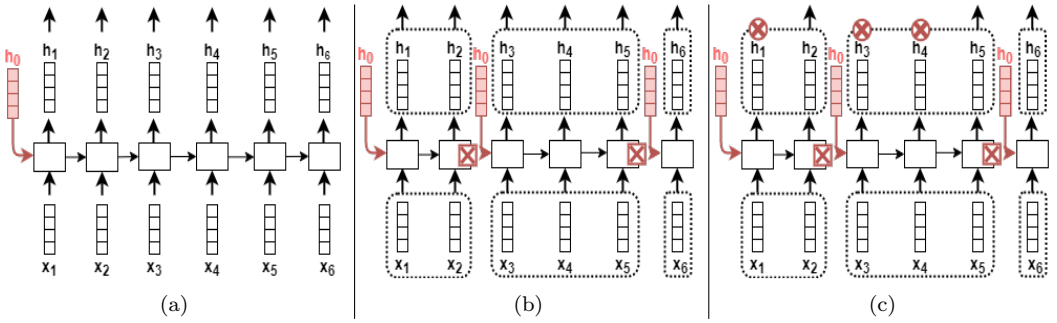


Figure 6.1: Graphical representation of the different GRUs used in our experiments: Figure 6.1a shows a Vanilla GRU. Figure 6.1b shows $\text{GRU}_{\text{PACK.}}$ in the ALL condition where all the vectors produced at each time step are passed on to the next layer. 6.1c shows $\text{GRU}_{\text{PACK.}}$ in the KEEP condition where only the last vector of a segment is passed on to the next layer, thus resulting in an output sequence shorter than the input sequence. The red crosses inscribed in a square (⊠) signal that the output vector computed at a given timestep is not passed on to the next timestep and that the initial state h_0 is passed on instead. The red crosses inscribed in a circle (⊗) signal that the output vector computed at a given timestep is not passed on to the next layer. Dotted lines group vectors belonging to a same segment (either phone, syllable-connected, syllable-word, or word). Note that h_0 is only passed on to the next state at the end of a segment, thus effectively materialising a boundary by manually resetting the history. Also note that the $x_1, x_2, ..., x_t$ figured in this representation could either be the original input sequence (in our case, acoustic vectors) or could also be the output of the previous recurrent layer

on the position of the $\text{GRU}_{\text{PACK.}}$ layer). In the KEEP condition, only the last vector of each segment is forwarded to the next layer (see Figure 6.1c). The length of the output and input sequence stays the same in the ALL condition. However, it should be noted that in the KEEP condition, the length of the output sequence is shorter than the input sequence.[4]

The difference between ALL and KEEP is motivated by the fact that we believe that keeping the last vector of a segment could constrain the network to learn more consistent representations for different occurrences of the same segment. Indeed, in the ALL condition, even though a vector at timestep t is different from its neighbours, we expect consecutive vectors to share a substantial amount of information as the acoustic vectors are extracted within a very short timeframe with an overlapping window (see 3.2.3, footnote 10). Thus, boundary information might be watered down by the fact that the preceding and following vector contain redundant information, and the network might not use this information effectively. However, in the KEEP condition, as only the last vector of a given segment is forwarded to the next layers, the network is forced to "cram" as much information as possible in a single vector so as to pack the information effectively. Also, as only the final vector is kept, it should be as informative as possible as the subsequent layers will have less information to rely on, making it difficult to reconstruct the missing data.

Our $\text{GRU}_{\text{PACK.}}$–KEEP segmentation approach was inspired by prior work by Kreutzer & Sokolov (2018) on dynamic segmentation for MT. However, we also found that a nearly identical approach was proposed by Chen et al. (2019) in an Audio-Word2Vec experiment. Instead of being given gold segment boundaries, a classifier outputs a probability that a given frame constitutes a segment boundary, and if so, the RNN's history is reset and

[4]Potentially, the length of the sequences can be different for different items inside a batch as the captions have a different number of segments (be they phones, syllables or words). For this reason, and as the subsequent layers expect a 3D rectangular matrix (of shape batch size × sequence × embedding dimension) we add padding vectors on the sequence dimension until all the elements of the batch have the same sequence length.

only the last vector of a segment is forwarded to the next layers. Working with pre-segmented speech is also analogous to prior work done by Harwath & Glass (2015) where the spectrogram was broken into words. Nonetheless, our work is different from theirs as we also explore other types of segments (not only words, but also phones and syllables) and also because we introduce this information at different layers of the architecture.

6.2.4 Experimental Settings

6.2.4.1 $\text{GRU}_{\text{PACK.}}$ Position

In order to understand where boundary information should be introduced (that is, at which level of the architecture), we train as many models as the number of recurrent layers, where each time one layer of GRUs is replaced with one $\text{GRU}_{\text{PACK.}}$ layer. For example, "$\text{GRU}_{\text{PACK.}}$–3" refers to a model where the third layer of GRUs is a $\text{GRU}_{\text{PACK.}}$ layer and other layers (1^{st},2^{nd},4^{th}, and 5^{th} layer) are vanilla GRU layers. This setting will allow to explore *where* introducing boundary information is the most efficient.

6.2.4.2 Random Boundaries

In order to understand if introducing boundary information helps the network in its task, we compare the performance of the models using boundary information with a baseline model which does not use any (thus, all the recurrent layers of the baseline architecture are Vanilla GRU layers). This model will from now on be referred to as BASELINE. We also introduce another condition, where, instead of training models with real segment boundaries (which from now on will be referred to as TRUE), we train models with random boundaries (which from now on will be referred to as RANDOM). Indeed, it could be that randomly slicing speech into sub-units leads to better results, even though the resulting units do not constitute any linguistically meaningful units. Therefore, training models with random boundaries will enable us to verify this claim. Random boundaries were generated by simply shuffling the position of the real boundaries (vector B introduced in §6.2.1), resulting in as many randomly positioned boundaries as they are real boundaries. Note that we do still expect the models to have reasonable results even when using random boundaries, as the acoustic vectors are kept untouched. However, we expect that placing random boundaries will hinder the network's learning process and thus yield results significantly lower than when using true boundaries. We expect the results to be significantly lower in the RANDOM-KEEP condition as this condition is equivalent to randomly subsampling the input, and thus removing a lot of information.

6.2.4.3 Evaluation

Models are evaluated in term of Recall@k (R@k). Given a spoken query, R@k evaluates the models' ability to rank the target paired image in the top k images. In order to evaluate if the results observed in our different experimental conditions (TRUE-ALL, TRUE-KEEP, RANDOM-ALL, RANDOM-KEEP) are different from one another and from the BASELINE condition, we used a two-sided proportion Z-Test. This test is used to assess if there is a statistical difference between two independant proportions. As for each spoken query there is only one target image, R@k becomes a binary value which equals 1 if the target image is ranked in the top k images and 0 otherwise. In our case, the proportion that we test is the number of successes over the number of trials (which corresponds to the number of different caption/image pairs in the test set).

6.2.5 Results

Overall, our experimental settings led to the training of 81 different models per data set.[5] BASELINE results are shown in Table 6.1. Results for the TRUE/RANDOM conditions obtained

Data set	R@1	R@5	R@10
COCO	9.0	27.0	39.5
Flickr8k	4.3	13.4	21.4

Table 6.1: Mean recalls at 1, 5, and 10 (in %) on a speech-image retrieval task COCO and Flickr8k in the BASELINE condition on the test set (based on the results of the best epoch on the validation set). Chance scores are 0.0002/0.001/0.002 for COCO and 0.001/0.005/0.01 for Flickr8k. Results are different than those presented in Chapter 4 as we use a different code that uses a different Deep Learning framework.

on the COCO and Flickr8k data sets are shown in Table 6.2 and 6.3 respectively. We obtain lower results on Flickr8k than on COCO which shows how difficult the task is on natural speech. Note that the results obtained on synthetic speech are also very low compared to their textual counterpart.[6]

6.2.5.1 True and Random Boundaries

The first question our experiments aim at answering is whether introducing boundary information helps the network in solving its task or not. Overall, models trained on FLICKR8K with TRUE boundary information have a significantly better R@1 than their baseline counterparts and models trained with RANDOM boundaries (which are either on a par with the baseline, or worse). This indicates that the models did effectively use the provided boundary information.

For COCO, we observe that some models trained with random boundaries have significantly better scores than the baseline (particularly when using phone boundaries in the KEEP condition). We explain this by the fact that randomly subsampling the input signal might act as a form of regularisation for the network. This effect disappears when using larger units, such as words, for which the results are not significantly different from the baseline. This regularisation effect might only be due to the fact that COCO uses synthetic speech with only one voice and hence, has very low intra-speaker variation. Thus, even though we randomly subsample the input, as there is very little intra-speaker variation, the network is much more likely to figure out from which units the subsampled vector came from.

6.2.5.2 All and Keep

The results shown above should also be analysed in regards to the ALL and KEEP conditions. Indeed, there is an interaction effect between using TRUE and RANDOM boundaries either in the ALL or KEEP condition.

This effect is particularly noticeable for FLICKR8K. Indeed, in the RANDOM-ALL condition, no result is statistically better than the baseline, while in the RANDOM-KEEP condition, the results are statistically worse than the baseline. This clearly shows that keeping the last vector only of segment rather than all of the vectors has a real effect, and that this effect is particularly noticeable when using RANDOM boundaries. Hence, using random boundaries

[5] (Seg. type $\in$ {phone,syl.-connected,syl.-word,word}$\times$ $\text{GRU}_{\text{PACK.}}${1,2,3,4,5}$\times$ {TRUE,RANDOM}$\times$ {ALL,KEEP}) + BASELINE

[6] Merkx & Frank (2019) reports R@1 = 27.5 on a GRU-based model using characters as input.

	COCO | ALL condition							
GRU	Phones		Syl.-Co.		Syl.-Word		Word	
Pack.	T	R	T	R	T	R	T	R
5	**09.7** $^{+}$	09.4	09.1	**09.5**	**09.5**	09.4	09.3	**09.5**
4	**09.3**	09.2	**09.4**	09.1	**09.6**	09.2	09.0	**09.4**
3	**09.5**	09.2	**09.2**	09.1	**09.4**	09.1	**09.4**	09.2
2	**09.4**	08.9	**09.7** $^{+}$	09.1	**09.6**	09.2	***09.7*** $^{+}$	*08.9*
1	**09.8** $^{+}$	09.4	**09.6**	09.4	***10.0*** $^{+}$	*09.1*	**09.5**	09.1

	COCO | KEEP condition							
GRU	Phones		Syl.-Co.		Syl.-Word		Word	
Pack.	T	R	T	R	T	R	T	R
5	9.4	**9.6**	**9.1**	**9.1**	**9.6**	9.1	***9.4***	*8.7*
4	10.0 $^{+}$	**10.5** $^{+}$	**10.2** $^{+}$	9.6	**10.4** $^{+}$	9.9 $^{+}$	***10.6*** $^{+}$	*9.5*
3	**10.5** $^{+}$	10.1 $^{+}$	**10.4** $^{+}$	9.8 $^{+}$	**10.5** $^{+}$	10.1 $^{+}$	***11.0*** $^{+}$	*9.7*
2	***10.7*** $^{+}$	*9.8* $^{+}$	***10.5*** $^{+}$	*9.4*	***10.9*** $^{+}$	*9.3*	***11.3*** $^{+}$	*8.8*
1	***10.1*** $^{+}$	*7.9* $^{-}$	***9.7***	*7.1* $^{-}$	***10.2*** $^{+}$	*7.0* $^{-}$	***10.3*** $^{+}$	*7.0* $^{-}$

Table 6.2: Maximum R@1 (in %) for each model trained on the COCO data set. "T" stands for TRUE (boundaries) and "R" stands for RANDOM (boundaries). "Syl-Co." and "Syl-Word" stand for "Syllable-Connected" and "Syllable-Word" respectively. Each line shows the results for when a specific recurrent layer is a $\text{GRU}_{\text{PACK.}}$ layer. The 1^{st} layer is the lowest layer (right after the 1D convolutions and acoustic vectors) and the 5^{th} the highest (right after the four preceding recurrent layers and before the attention mechanism). The highest R@1 in the table is shown in **red**. The best results between each TRUE and RANDOM pair (columnwise) are shown in **bold**. $\circ^{+}$ and $\circ^{-}$ indicate that the results are statistically better (respectively worse) than the baseline. Results in *italics* show statistical significance (two-sided Z-Test, p-value $< 1e^{-2}$, see §6.2.4.3) between each TRUE and RANDOM pair (columnwise).

which do not delimit meaningful linguistic units really hurts the performance of the network. Furthermore, we notice that the results between ALL either in the RANDOM or TRUE condition are not statistically different from one another while they are in the KEEP condition. This tells us that boundary information is not used effectively in the ALL condition, even when TRUE boundaries are given. The only exception to this statement is when using word boundaries at the first layer in the ALL condition. It seems that boundaries are only effectively used here. This confirms our intuition that the KEEP condition effectively constrains the network to learn better representations, and that in the ALL condition boundary information is watered-down by the neighbouring vectors, thus leading to a suboptimal use of such information.

For COCO as well, we notice an asymmetry between the ALL and KEEP conditions. In the ALL condition, the results are rarely significant while there are significantly better than the baseline in the KEEP condition. Here also, we observe an interaction between the TRUE-RANDOM and KEEP-ALL condition. In the ALL condition, the results between TRUE and RANDOM are rarely significantly different from one another, while they are in the KEEP condition.

	Flickr8k | ALL condition							
GRU	Phones		Syl.-Co.		Syl.-Word		Word	
Pack.	T	R	T	R	T	R	T	R
5	**4.0**	3.9	**4.1**	**4.1**	**4.3**	3.9	3.4	**4.2**
4	4.0	**4.4**	3.9	**4.1**	**4.3**	3.8	**4.5**	**4.5**
3	**4.5**	4.4	**4.3**	4.2	**4.4**	4.2	**4.5**	3.8
2	**4.5**	3.8	***4.8***	*3.6*	**4.4**	4.2	**4.7**	4.1
1	**4.3**	3.4	**4.0**	**4.0**	**4.4**	4.3	***5.3*** $^{+}$	*4.1*

	Flickr8k | KEEP condition							
GRU	Phones		Syl.-Co.		Syl.-Word		Word	
Pack.	T	R	T	R	T	R	T	R
5	3.6	**3.7**	***3.6***	*2.5* $^{-}$	**3.3**	3.0	**3.2**	**3.2**
4	**3.8**	**3.8**	**4.4**	3.5	***3.9***	*2.6* $^{-}$	***5.2*** $^{+}$	*2.5* $^{-}$
3	***4.9*** $^{+}$	*3.8*	***4.5***	*3.1*	***5.3*** $^{+}$	*3.1*	***4.9*** $^{+}$	*3.3*
2	**4.8** $^{+}$	3.9	***5.1*** $^{+}$	*3.6*	***4.8***	*3.4*	***5.4*** $^{+}$	*3.4*
1	***4.8***	*2.4* $^{-}$	***3.4***	*1.9* $^{-}$	***4.4***	*2.0* $^{-}$	***3.9***	*1.9* $^{-}$

Table 6.3: Maximum R@1 (in %) for each model trained on the Flickr8k data set. The same naming conventions of Table 6.2 are used for this table.

6.2.5.3 Phone, Syllable, or Word

In our experiments, we used four different types of segments corresponding two different types of linguistic units: phones, syllables-connected, syllables-word, and words. These different types of segments vary in *length* (words and syllables are longer than phones), *quantity* (there are more phones and syllables than words), and *intrinsic linguistic information*: phones only show which are the basic acoustic units of the language, while word segments represent meaningful units, and syllables-word and syllables-connected are a higher form of acoustic units that may contain morphemic information. Given the task the network is trained for (speech-image retrieval), we do not expect these different units to perform equally well. Indeed, as this task implies mapping an image vector describing which objects are present in a picture and a spoken description of an image, we expect word-like segments (or segments that preserve word boundaries and that bear a substantial amount of semantic information) to perform better.

This is in fact what we observe in practice. Word units obtain statistically better results than the baseline for both FLICKR8K and COCO ($R@1 = 5.4$, +1.1pp and $R@1 = 11.3$, +2.3pp respectively). Syllables-word also bring significant improvement ($R@1 = 5.3$ for FLICKR8K and $R@1 = 10.9$ for COCO), however, slightly less than when using word units. It should be noted that syllables-connected segments obtain also statistically significant improvement over the baseline ($\text{GRU}_{\text{PACK.}}$–2 for both data sets) despite not preserving all word boundaries. However, these results are slightly worse than the syllables-word and word segments suggesting that preserving word boundaries is a property that helps the network. Using syllable-connected segments yields results that are on par with phone units. It appears that the size of a segment is a very important parameter. Indeed, phone segments (naturally) preserve word boundaries but of course naturally lack the internal cohesion of a morpheme or a word as nothing links two adjacent phonemes together, while syllables-connected do not preserve word boundaries, but present a higher internal cohesion. Hence, it seems that segments that preserve both meaning and word boundaries (such as

words) or from which meaning can be more easily recomposed (syllable-word) may facilitate the network's task. The fact that syllable-word segments perform as well as word segments might only be an artefact of using English where a high proportion of word is monosyllabic.[7] Working on a language where the syllable-to-morpheme ratio is higher would be a future line of work that would enable to test this hypothesis.

6.2.5.4 $\text{GRU}_{\text{PACK.}}$ Layer Position

We introduced boundary information at different levels of our architecture in order to better understand at which layer it is the most useful to add such information. We will focus in this section on the results obtained in the KEEP condition, as the ALL condition brings little improvement over the BASELINE condition.

Our results clearly show that introducing boundary information at different layers has a substantial impact on the results: using such information at the first or the fifth layer is useless, as we notice it either yields similar results as the baseline ($\text{GRU}_{\text{PACK.}}$–1) or worsens the results regardless of the type of boundary used ($\text{GRU}_{\text{PACK.}}$–5). When using syllables-word segments the best results are obtained when introducing the information at $\text{GRU}_{\text{PACK.}}$–3 ($\text{GRU}_{\text{PACK.}}$–2 for COCO), and $\text{GRU}_{\text{PACK.}}$–2 for word segments. This results are in line with that of Chrupała et al. (2017*a*) who found that the representations learnt by the fifth layer is the less informative in predicting word presence, while lower layers encode this information better. This confirms that the middle layers of our architecture are better suited to deal with lexical units whereas the fifth layer encodes information that disregards that type of information.

All in all, for FLICKR8K word-like segments seem to be the most robust representation to be used as they yield *significantly* better results at three different layers ($\text{GRU}_{\text{PACK.}}$–2,3,4). For COCO, we observe mixed results where introducing boundary information at the first four layers improve our baseline results. However, the results tend to be higher when boundary information is introduced at the second layer, confirming this layer is the one that benifit the most from boundary information overall.

6.2.6 Segmentation as a means for compression

Recall that in the KEEP condition, only the last vector comprising a segment is kept while the other vectors are discarded. This can be interpreted as a form of "guided" subsampling, as usually subsampling does not take into consideration linguistic factors. To understand how much information is kept between the input and the output of a $\text{GRU}_{\text{PACK.}}$ layer in the KEEP condition, we compute an average compression rate (in %) for each of the segment types for Flickr8k. The results are the following: phones = 90.57%, syllables-connected = 93.41%, syllables-word = 94.36%, and words = 94.90%.[8]

When we re-analyse our results in light of this information, it appears we can remove a large part of the original input (up to 94.90% if using word segments) while conserving or increasing the original R@1. It is not simply the effect of subsampling that helps, but subsampling with *meaningful* linguistic units. The effect of informed subsampling is striking when we compare R@1 for RANDOM-KEEP, which are always below the BASELINE, while TRUE-KEEP are on a par with the BASELINE or better. This effect is particularly visible for $\text{GRU}_{\text{PACK.}}$–1 in the RANDOM condition where the more vectors are discarded (syllable-like and word segments), the worse the results are. This can only be explained by the fact that

[7] Jespersen (1929) estimates that at least 8,000 commonly frequent words are monosyllabic in English.

[8] Note that the compression rate for syllables-word and words is very close, suggesting there is a significant overlap between syllables-word units and word units.

randomly subsambling removes important information that the network is unable to recover in the four subsequent layers. A counter-intuitive finding of our experiments is that it is better to subsample early on (in the first layers) and thus remove most of the information early on than later on. Subsampling with word segments in $\text{GRU}_{\text{PACK.}}$–2 (and thus only keeping 5.1% of the original amount of information for the subsequent layers) yields better results than subsampling with the same resolution at $\text{GRU}_{\text{PACK.}}$–5.

6.3 Hierarchical Information

6.3.1 Integrating Hierarchical Information

In the aforementioned experiments, we supply the network with only one type of boundary (either phone, syllable, or word) but not multiple at the same time, as if several units could not coexist at the same time. Yet, this does not correspond to how speech is really structured. Indeed, multiple spoken units exist at the same time, and they are structured hierarchically: words can be broken down into syllables, which can be in turn be broken down into phones.[9] In order to model such hierarchical nature of speech, we can stack as many $\text{GRU}_{\text{PACK.}}$ as desired, where one layer handles one type of segment (e.g. phone) and the following $\text{GRU}_{\text{PACK.}}$ layer handles another type of segment, that is hierarchically above the preceding (e.g. syllable, or word).[10] A graphical representation of a hierarchical $\text{GRU}_{\text{PACK.}}$ architecture is shown in Figure 6.2.

We expect such hierarchical architecture to perform even better than a single-layered $\text{GRU}_{\text{PACK.}}$. Indeed, a two-layered architecture should constrain the network to learn even better representations than a single-layered architecture, especially for longer segments such as words. Words might be long units and capturing all their details in a single pass might be challenging. By breaking them first into sub-units, such as phones or syllables, we expect the network to learn more consistent representation and help the network better distinguish between words differing by only one phoneme.

Harwath et al. (2020) explored such hierarchical architecture using a CNN-based model that incorporated multiple vector quantisation layers and found that it improved the network ability to retrieve the target image. Our work thus attempts to verify if it is also the case for an RNN-based model, when provided with gold boundaries.

6.3.2 Experimental Settings

We explore the effect of using a hierarchical architecture on the Flickr8k data set only, as it features real human speech and might thus be more challenging, but also more realistic. Contrary to our previous experiments, we will only consider hierarchical architectures that use $\text{GRU}_{\text{PACK.}}$–keep layers, as we have shown that $\text{GRU}_{\text{PACK.}}$–all bring little improvement over vanilla GRUs.

We test various architectures that handle different type of boundaries simultaneously with two out of the five recurrent layers being $\text{GRU}_{\text{PACK.}}$ layers (6.3.3) or with three $\text{GRU}_{\text{PACK.}}$ layers (6.3.4). In both cases, we test all possible positions as well as boundary type to understand what the best combination is.

[9] Note that for Japanese, we could add another intermediate step between syllables and phones which would be moræ

[10] Note that it could also be possible to use larger units, such as chunks.

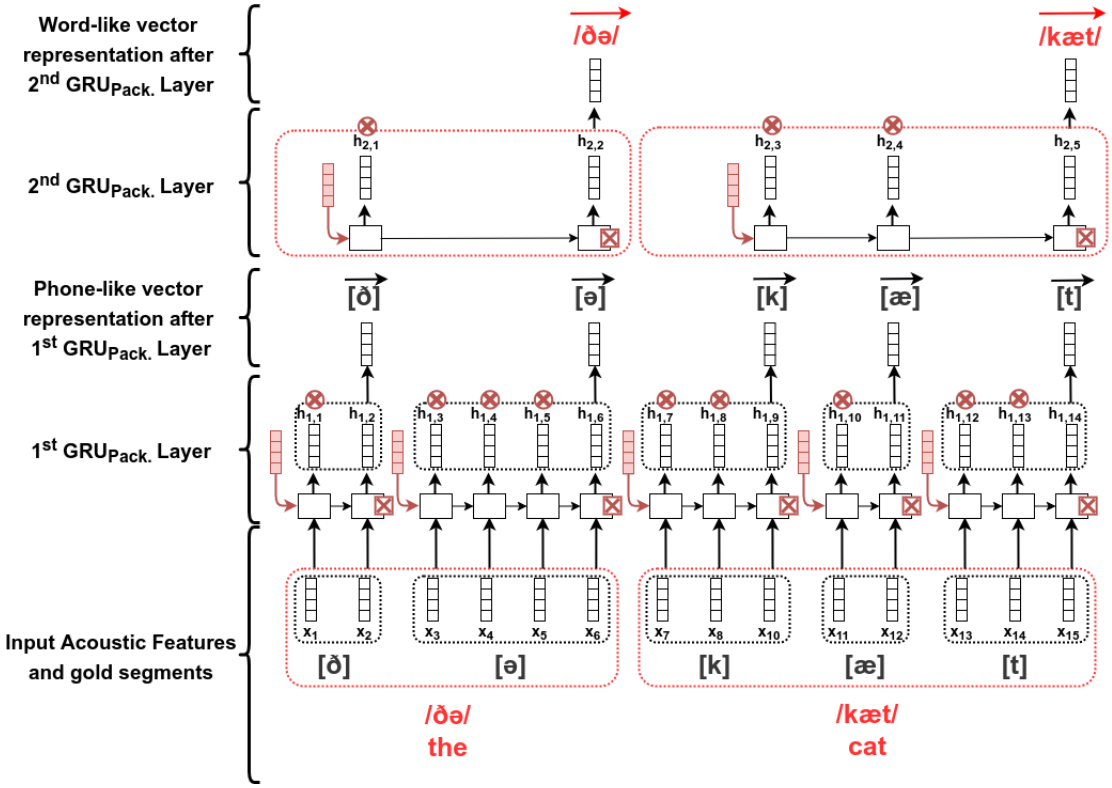


Figure 6.2: Graphical representation of a hierarchical architecture that uses two $\text{GRU}_{\text{PACK.}}$ layers. The 1^{st} $\text{GRU}_{\text{PACK.}}$ layer is supplied with phone boundary information. This layer outputs vectors representing phones that are recomposed from the input vectors. The next layer, $\text{GRU}_{\text{PACK.}}$–2 takes these phone vectors as input and recombines them into words and outputs vector representing whole words. This architecture effectively models the hierarchical nature of speech, where words consist of a sequence of phones. Note that both $\text{GRU}_{\text{PACK.}}$ are in the KEEP condition where only the last vector of a given unit is kept. As in Figure 6.1, note that $x_1, x_2, ..., x_t$ need not be acoustic vectors, but could also be the output of the previous vector.

6.3.3 Two $\text{GRU}_{\text{PACK.}}$ Layers

In this section, we experiment an architecture with two $\text{GRU}_{\text{PACK.}}$ each handling different boundary types: phones and words (6.3.3.1), phones and syllables (6.3.3.1), and syllables and words (6.3.3.3). We vary the position of the $\text{GRU}_{\text{PACK.}}$ layers used in our architecture and test all the combinations of possible positions.

6.3.3.1 Phones and Words

When using phones and words together, the results (Table 6.4) are higher than the baseline architecture and than the single-layered $\text{GRU}_{\text{PACK.}}$ architecture. Indeed, we obtain a maximum R@1 of 8.2% when using $\text{GRU}_{\text{PACK.}}$ at the layer 2 and 3 which is +3.9pp over the baseline and +2.8pp over a single-layered architecture. We also notice that using a hiearchical architecture also allows us to train shallower networks while improving the results over our baseline architecture. For example, a two-layered architecture (where both layers are $\text{GRU}_{\text{PACK.}}$) has a R@1 of 6.4 which is +2.1pp over our 5-layered baseline architecture that does not use any boundary information.

Here also, we observe that placing a $\text{GRU}_{\text{PACK.}}$ layer at the last layer yields worse results than when the last layer is a vanilla GRU layer. Indeed, we observe that for all the architectures, when the last layer is a $\text{GRU}_{\text{PACK.}}$ layer, the results are substantially lower when placed one layer before. Once again, this confirms that the last layer of the network – which is just before the attention mechanism – handles semantic information and is not concerned with form any more. These results also indicate that the $\text{GRU}_{\text{PACK.}}$ layers are more effective when placed in the middle of the architecture, than when used at the beginning: both for a five-layered architecture or a four-layered architecture, the best results are obtained when the $\text{GRU}_{\text{PACK.}}$ are placed at layers 2 and 3, suggesting that speech requires a certain form of pre-processing in the lower layers in order to use boundary information more effectively.

Architecture	**5 layers**					**4 layers**				**3 layers**			**2 layers**	
$1^{st}\text{GRU}_{\text{PACK.}}$ \ $2^{nd}\text{GRU}_{\text{PACK.}}$	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**1**	**2**
1		7.7	7.7	7.3	3.9		7.6	7.9	5.7		**8.1**	5.3		**6.4**
2			**8.2**	7.6	5.8			**8.1**	6.3			7.3		
3				7.1	6.5				6.7					
4					6.1									
5														
Baseline	4.3					4.4				3.4			3.5	

Table 6.4: R@1 obtained on the test set of the Flickr8k data set with a hierarchical architecture consisting of two $\text{GRU}_{\text{PACK.}}$ layers using phone and word segments (models were selected based on the maximum R@1 on the validation set). Best score overall is shown in red. Best score (layer-wise) is shown in **bold**. Greyed out cells signal impossible configurations. We also indicate R@1 obtained on a baseline architecture that does not use any $\text{GRU}_{\text{PACK.}}$ layers.

In our previous experiments, we observed that even random boundaries could yield statistically better results than the baseline and showed it could only be explained by a regularisation effect. It could be that stacking two $\text{GRU}_{\text{PACK.}}$ layers yields better results not because of the hierarchical nature of the architecture, but simply because of a double regularisation effect. To investigate if such effect arises or not, and in order to confirm our previous observations, we select our best performing architecture (5-layered architecture with $\text{GRU}_{\text{PACK.}}$ layers at layer 2 and 3, see Table 6.4) and retrain it using random boundaries.[11] We thus have the following training and testing settings:

- TRUE phones + TRUE words : the first $\text{GRU}_{\text{PACK.}}$ layer receives true phone boundaries and the second $\text{GRU}_{\text{PACK.}}$ layer receives true word boundaries. This is the ideal condition, and we expect models trained with such boundaries to have the best results (this setting corresponds to the R@1 = 8.2 in the previous table).

- TRUE phones + RANDOM words : the first $\text{GRU}_{\text{PACK.}}$ layer receives true phone boundaries while the second $\text{GRU}_{\text{PACK.}}$ layer receives random word boundaries, randomly sampled from the phone boundaries.

- RANDOM phones + TRUE words : the first $\text{GRU}_{\text{PACK.}}$ layer receives random phone boundaries while the second $\text{GRU}_{\text{PACK.}}$ layer receives true word boundaries. Because the second layer requires true word boundaries, we also need to keep some boundaries intact in the phone boundaries provided to the first $\text{GRU}_{\text{PACK.}}$ layer. Thus, the random boundaries provided to the first layer are not entirely random, and

[11] When using random phone or random word boundaries, we make sure to keep the number of random boundaries the same as the number of true boundaries.

they are as many true phone boundaries as the number of words. The true phone boundaries we keep correspond to the last phoneme of a word.

- RANDOM phones + RANDOM words : the first $\text{GRU}_{\text{PACK.}}$ receives random phone boundaries and the second $\text{GRU}_{\text{PACK.}}$ receives random word boundaries, randomly sampled from the phone boundaries. This is the worst condition, and we expect models trained with such boundaries to have the worst results.

The results we obtained using random phones and/or random word boundaries are shown in Table 6.5. We notice that all the results are lower when using random phones and/or random words than when using true phones and true words together. This thus shows that the better results obtained with a hierarchical architecture are not to be explained by a double regularisation effect, as if it had been the case, we should have observed similar results with random boundaries. Hence, we conclude that using a hierarchical structure is indeed beneficial because it accounts for the hierarchical nature of the spoken units and helps the network build better (more consistent) representations.

Configuration / $\text{GRU}_{\text{PACK.}}$ Pos.	TRUE Phones TRUE Words	TRUE Phones RANDOM Words	RANDOM Phones TRUE Words	RANDOM Phones RANDOM Words
2 and 3	**8.2**	4.3	3.6	1.0

Table 6.5: R@1 obtained on the test set of the Flickr8k data set with a hierarchical architecture consisting of two $\text{GRU}_{\text{PACK.}}$ layers using random phones and/or random word boundaries (models were selected based on the maximum R@1 on the validation set). Best score overall is shown in red. Best score (layer-wise) is shown in **bold**.

6.3.3.2 Phones and Syllables

The results we obtain when using phones and syllables is shown in Table 6.6. Here also, we notice that the results are better than the baseline results (without any $\text{GRU}_{\text{PACK.}}$) and also better than when using only one $\text{GRU}_{\text{PACK.}}$: $R@1 = 7.9$, +2.5pp. Contrary to the previous experiment, the 4-layered architecture converges better than the 5-layered architecture. Nonetheless, the results are lower than when using jointly phones and word (−0.3pp) indicating that using jointly phones and syllables is not the ideal combination. Once again, we do observe the same tendencies as we previously put forward: weaker results

Architecture	**5 layers**					**4 layers**				**3 layers**			**2 layers**	
$1^{st}\text{GRU}_{\text{PACK.}}$ / $2^{nd}\text{GRU}_{\text{PACK.}}$	1	2	3	4	5	1	2	3	4	1	2	3	1	2
GRU PACK														
1		6.6	6.0	6.3	4.3		6.9	6.5	4.6		**6.6**	4.9		**4.6**
2			**7.5**	6.8	5.7			**7.9**	4.7			6.1		
3				6.5	4.8				4.7					
4					4.6									
5														
Baseline	4.3					4.4				3.4			3.5	

Table 6.6: R@1 obtained on the test set of the Flickr8k data set with a hierarchical architecture consisting of two $\text{GRU}_{\text{PACK.}}$ layers using phones and syllable-word (models were selected based on the maximum R@1 on the validation set). The same naming conventions of Table 6.4 are used for this table

when the last layer is a $\text{GRU}_{\text{PACK.}}$ layer; and the best results are obtained when $\text{GRU}_{\text{PACK.}}$ layers are placed in the middle layers. We however observe that the two-layered architecture

performs worse than in the previous experimental setting. Indeed, in the previous setting, the result obtained with such an architecture was +2.1pp over our baseline result, while the gain here is negligible: only +0.3pp over the baseline. This shows that even though using word boundaries or syllable boundaries results in a higher compression rate (see 6.2.6), it does not compensate for the inherently lower semantic contribution of the provided segments.

6.3.3.3 Syllables and Words

The results we obtain when using syllables and words is shown in Table 6.7. Once again, we notice that R@1 is higher than a single-layered $\text{GRU}_{\text{PACK.}}$ architecture ($R@1 = 7.6, +2.2$pp) but worse than a two-layered architecture handling phones and words: -0.6pp. The best result is also lower than the best result when using phones and syllables -0.3pp. However, we observe that the two-layered architecture performs better than the baseline (+1.0pp) and better than the two-layered architecture that uses jointly phones and syllables (+0.7pp).

These results suggest two things: first, in order to use a very shallow architecture, segments that bear a lot of semantic information should be preserved (as we observe better results with such architectures when word segments are preserved). Second, keeping low-level segments such as phones is also necessary when training deeper models. The fact that we do not observe improvement when using jointly syllables and words with deeper architecture might show that the information brought by both levels overlap and might be redundant.

Architecture	**5 layers**					**4 layers**				**3 layers**			**2 layers**	
$1^{st}\text{GRU}_{\text{PACK.}}$ \ $2^{nd}\text{GRU}_{\text{PACK.}}$	1	2	3	4	5	1	2	3	4	1	2	3	1	2
GRU PACK														
1		5.7	5.5	5.7	4.5		5.7	6.8	5.2		6.0	5.2		**5.3**
2			**7.3**	7.1	6.1			**7.6**	6.0			**6.3**		
3				6.8	5.7				6.0					
4					5.5									
5														
Baseline	4.3					4.4				3.4			3.5	

Table 6.7: R@1 obtained on the test set of the Flickr8k data set with a hierarchical architecture consisting of two $\text{GRU}_{\text{PACK.}}$ layers using syllable-word and word segments (models were selected based on the maximum R@1 on the validation set). The same naming conventions of Table 6.4 are used for this table

6.3.3.4 Section Conclusion

Our experiments show that the best result using two $\text{GRU}_{\text{PACK.}}$ layers is obtained when using jointly phone and word boundaries, when the $\text{GRU}_{\text{PACK.}}$ layers are placed in the middle of the recurrent stack. We also observe that overall our results are maximal when the layers immediately follow one another. Finally, this set of experiments allowed us to show that the network converges best when low-level segments (phones) and high-level segments (words) are used jointly. We explain this by the fact that this allows the model to learn robust representations for the phone units, while having high level units that bear a lot of semantic information. Consequently, using intermediate segments such as syllables is not useful as they are not short enough to learn a consistent representation while being too short in regards to the amount of semantic information they bear.

6.3.4 Three GRU$_{\text{PACK.}}$ Layers: Phones, Syllables, and Words

Finally, we integrated three segment levels in a single model. As in our previous experiments, we experiment with a different number of layers (from 3 to 5), each time with 3 GRU$_{\text{PACK.}}$ layers at each possible position. The results of this experiment are presented in Table 6.8.

We observe that the best result obtained with this architecture ($R@1 = 9.6$) is far better than the baseline (+5.3pp), better than the best result of a single-layered architecture (+4.2pp) but also better than the best result of a double-layered architecture (+1.4pp over the phone-word architecture). Our best result is obtained by a five-layered architecture with GRU$_{\text{PACK.}}$ in position 1, 3 and 4. However, we notice that the four-layered architecture obtains more consistent results across all layers, the maximum result being only −0.3pp away from best five-layered architecture. We also notice that the 3 layered architecture obtains a very high R@1 of 8.0 which is about two times over the baseline results. However, the result obtain with such settings is worse than what was obtained when using a three-layered architecture with two GRU$_{\text{PACK.}}$ layer (−0.1pp). We believe this is to be explained by the fact that in such setting, there necessarily is a GRU$_{\text{PACK.}}$ at the last layer, and we observe such architecture degraded the results overall.

GRU$_{\text{PACK.}}$ \ Architecture	**5 layers**	**4 layers**	**3 layers**
1 + 2 + 3	8.5	**9.3**	8.0
1 + 2 + 4	8.1	8.6	
1 + 2 + 5	7.8		
1 + 3 + 4	**9.6**	8.4	
1 + 3 + 5	7.9		
1 + 4 + 5	7.8		
2 + 3 + 4	8.8	8.3	
2 + 3 + 5	8.5		
2 + 4 + 5	8.3		
3 + 4 + 5	7.8		

Table 6.8: R@1 obtained on the test set of the Flickr8k data set with a hierarchical architecture consisting of three GRU$_{\text{PACK.}}$ layers using phone, syllable-word and word segments (models were selected based on the maximum R@1 on the validation set). The same naming conventions of Table 6.4 are used for this table

Even though we notice that R@1 improved when adding another boundary level in this last experiment, we also observe that the leap in the results is not as big as what we observed before. When using one GRU$_{\text{PACK.}}$, we observe an improvement of +1.1pp in R@1 compared to our baseline architecture. When using two GRU$_{\text{PACK.}}$, we observe an improvement of +2.8pp compared to using one GRU$_{\text{PACK.}}$. And finally, when using a third GRU$_{\text{PACK.}}$, we observed an improvement of +1.4pp compared to using two GRU$_{\text{PACK.}}$. Thus, even though introducing more structure into the network is beneficial, we also observe that some levels are more critical than others and have a bigger effect on the final result.

6.4 Chapter Summary

In this chapter we studied the impact of prior speech segmentation in a VGS model. We presented a simple method to introduce boundary information at any recurrent layer of our architecture. We do so by simply resetting the RNN's history every time there is a segment boundary. From this initial condition, we derived two conditions, the ALL and

KEEP conditions, where we either keep all the vectors of a segment or only the last vector. We showed that the latter is more efficient, as, first, it helps the network to learn consistent representations, and second, it reduces the computational charge on the upper layers (as less vectors are forwarded to the next layers).

The goal of this chapter was to see if segmenting speech in sub-units was beneficial – that is, enabled the network to learn a better speech-to-image mapping – and if so, which units maximise the performance. It is indeed the case that segmenting speech into sub-units helps. This results is coherent with prior linguistic observation as well as previous results with textual models. As to which segment obtains the best performance we observe mixed results. Indeed, word segmentation yields better results than phone segmentation, but we do also observe that syllable-like segmentation also gives results that are in the same ballpark as word segmentation. Nevertheless, word segmentation seems to be a *more robust* representation compared to syllable as such word segments consistently yield better results at various levels of our architecture. Our experiments thus allowed us to observe that, such as for humans, the use of large units, such as words, is indeed the most efficient solution to learn a reliable speech-to-image mapping.

Additionally, we observed different results depending on the level at which segmentation information is introduced. We observed negative effects if boundary information is introduced too late (last layer of our architecture). This tends to show that the last layer is not concerned with form any more and thus that boundary information should be introduced before. This result is coherent with previous observation by Chrupała et al. (2017*a*) and Merkx et al. (2019) who showed that the last layer of such network does not encode word presence (or absence) effectively. Surprisingly, if boundary information is introduced too early (first layer), we do not observe any improvement over our baseline results. Thus, in order to be used effectively, boundary information should be provided to the middle layers. This shows that speech requires a certain amount of pre-processing before it is segmented.

Nonetheless, even though if introducing boundary information is useful, it only mildly improves the performance of the network. It is only when different levels are combined that the performance of the network reaches its peak. Our $\text{GRU}_{\text{PACK.}}$ setting allowed us to simply introduce such hierarchy in a neural network by simply stacking $\text{GRU}_{\text{PACK.}}$ layers and providing different boundary information to each of them. Our experiments reveal that having a structure that uses low-level segments (i.e. phones) jointly with high level segments (i.e. words) is better than using segments that are more or less of the same size (i.e. jointly using syllables and words). Adding boundary information also allowed us to reduce the number of layers while increasing the performance of the network. Therefore, explicitly taking into account the hierarchical nature of speech proves useful.

Conclusion

7.1 Summary of Findings

In this thesis, we studied an RNN-based model of Visually Grounded Speech. Our goal was to analyse the representations learnt by our model in order to better understand what type of linguistic knowledge neural models are able to acquire in an unsupervised fashion. We compared these representations to what is known of human speech processing. More specifically, we focused on lexical acquisition and found commonalities between the processes at work in the models we studied and the processes reported in the child language acquisition literature.

More specifically, the main contribution of this thesis is threefold:

- **Synthetically Spoken STAIR data set.** We introduced the "Synthetically Spoken STAIR data set" which is based on the STAIR data set (Yoshikawa et al. 2017). This data set constitutes the Japanese equivalent of the "Synthetically spoken COCO data set" (Chrupała et al. 2017*a*) for English. We introduced this data set as we believe working on typologically distinct languages is of paramount importance if one wants to truly understand the modelling capacities of deep neural networks. Despite recent work on this subject (Linzen et al. 2018, 2019, Alishahi et al. 2020, Belinkov & Glass 2019), most of the analyses focus on analysing deep neural networks trained on data sets in English. As both the STAIR data set and the COCO data set use the same set of images, this allowed us to compare the learning process of neural models trained on the exact same data, the only changing parameter being the language spoken.

- **Analysis of attention.** We showed that RNN-based models are able to detect recurring specific patterns in the acoustic input. More specifically, we showed such models did so by tuning the weights of their attention mechanism so as to give more importance to parts of the acoustic input that are particularly discriminative to predict a visual context. The models we trained focus specifically on concrete words such as nouns, as those refer to objects that are particularly salient in the images. We observed this behaviour for two typologically distinct languages, English and Japanese, hence showing this behaviour does not depend on a particular language. We concluded that the models we studied displayed a noun bias, such as what is also found in humans during the process of lexical acquisition, but we however highlighted that such preference might only be a consequence of the data set used in the experiments, which consists of still images.

  We also showed that despite adopting a language-general processing behaviour, the model could also develop a language-specific behaviour in order to better solve its task. Specifically, we put forward in this thesis that the Japanese models have learned to detect and highlight particles (such as the "ga" particule) showing a language-specific behaviour. By doing so, the model adopted the same behaviour as Japanese toddlers in order to segment the speech stream, who also use particles to detect nouns in the speech stream.

  The networks' ability to highlight nouns — and particles for Japanese — is a behaviour that the networks learned to adopt with a very few examples — less than 500

image/caption pairs — showing the network quickly learns which are the most important parts of the captions. This is an interesting finding, as it is generally admitted that large neural networks require a large amount of data to be trained effectively. We indeed find that the models we train require large amounts of data to solve their primary task effectively (see Appendix D), but they are however able to focus on the important parts of the speech inputs with very few examples. This suggests such networks could be used on low-resourced languages by linguists so as to automatically detect nouns in the spoken input.

- **Analysis of individual word knowledge.** We showed that the network was able to map individual nouns to their correct visual referent. This suggest that the network implicitly segmented its input into sub-units, and later mapped them to a visual context. However, we observed that the network was not able to map all isolated nouns to their visual referents equally well. Indeed, while the network was able to learn very reliable mappings between a word-form and its visual referent for about half of the most important words of the data set, the other half were not mapped to their visual referent, suggesting the model's lexicon is restricted to a set of words. We observed that this phenomemon was mainly due to the frequency of the word in the caption: the more frequent a word-form is, the better the model maps it to its visual referent.

  We then studied how the network activates an isolated word and compared it to model of word activation and recognition in humans. We showed that the network necessarily needed to have access to the first phoneme of a word in order to activate the representation of the target word. This result is similar to what is postulated in the COHORT model of speech recognition, where word onsets are of particular importance to activate and recognise a word. Using an algorithmic equivalent of the gating paradigm Grosjean (1980, 1985), a methodology stemming from psycholinguistics, we were able to observe that word activation does not occur linearly, but rather evolves in steps. This enabled us to conclude the model was able to recognise a word from a partial input, that is before its offset. We investigated if word recognition was carried out through a process of simultaneous activation of a cohort of words, that would then compete for recognition. We found some evidence of simultaneous activations for some words and competition between them, however, this process seems far from systematic, as we have shown that some words are activaed without competing with similar sounding words.

- **Introduction of prior linguistic information.** Finally, we investigated if introducing prior linguistic information in the form of boundary information was beneficial. We indeed found the network benefited from such information, particularly when the network was given word boundaries. The network benefited from phone and syllable boundary information, however, not as much as word boundaries. More importantly, we found that taking into account the hierarchical nature of speech, by simultatneously giving to the network phone, syllable and word boundaries, yielded even better results.

  We observed that introducing this information at different layers impacted significantly the results, and that this information was better handled by the intermediate layers. This experiment also allowed us to observe that keeping only the last vector of a segment (KEEP condition) which is equivalent to subsampling the speech signal in a linguistically meaningful way was more effective than keeping all the vectors of a segment. Hence, compressing the information in such a way that it is linguistically meaningful enables the network to acquire better representations.

In his book, Bloom (2002, p. 60) argues that "children learn the meanings of words through theory of mind. If this is right, then a direct connectionist implementation of word learning, in which sounds are associated with percepts, is unfeasible. (And this does preclude all connectionist theories of word learning that [he is] aware of.)". We believe the experiments we carried in this thesis, as well as previous work by Harwath et al. (2016), Harwath & Glass (2017), Chrupała et al. (2017*a*), Merkx et al. (2019) (among others) show that purely connexionist models are able to directly associate sounds with percepts, here, in the form of vector representations of visual stimuli. Hence, connexionist models are able to learn words to some extent. Of course, we do not argue that child lexical acquisition is only done *via* a purely associative mechanism, but it might be that a purely associative learning mechanism bootstraps lexical acquisition for children. An argument against this fact could be that connexionist approaches require large amounts of data to be trained effectively. However, our experiments show that our models learned to focus on specific nouns with a very few number of examples, suggesting associationist bootstrapping constitutes a viable mechanism to acquire a lexicon.

Given the architecture we use in our experiments, the lexical acquisition process we simulate is necessarily limited, as our architecture is not fully grounded (see Figure 2.8 in Chapter 2). Also, an ideal model would not only include visual stimuli, but also olfactory, tactile, etc. and would be a model that would be possible to interact with. In this thesis, we consider word-form/object mapping as a proxy for word learning. Yet, word learning entails much more than simply mapping a word-form to its visual referent, but should also include morphological information, syntactic information, and pragmatic information. It is nonetheless possible the models we trained learned other information, but this remains to be tested.

The experiments we conducted in this thesis allow us to conclude that VGS models implicitly segment their input into sub-units and associate these sub-units to their visual referent. This process seems to only emerge as a by-product of their main task which is to minimise a distance between an acoustic and a matching visual stimulus. Our conclusion is in line with the very recent work of Khorrami & Räsänen (2021) who conclude that "both sub-lexical and lexical representations can gradually emerge from the interaction of rich multimodal experiences available to the learner" and that these representations emerge as "a byproduct of multimodal sensing and interaction with the environment". However, even though an implicit segmentation seems to take place in such networks, this segmentation seems to be less effective than explicitly segmenting the input, as our experiments suggest. This result is in line prior researches in language acquisition: Havron et al. (2018) show that literacy — and consequently the knowledge of word boundaries — facilitates the acquisition of novel nouns.

7.2 Future Works

Given the work we conducted in this thesis, several future works could be carried out:

- **Word Activation in CNN-based models.** We studied how RNN-based models store lexical units and activate the representation of individual words by using the gating paradigm. This methodology could also be applied to analyse the representations learnt by CNN-based models in order to understand how the representation of a given word is activated in such models. It would be interesting to know if CNN-based and RNN-based models converge to the same type of representations, and for example, if word activation in CNN-based models proceeds in steps as what we observed for RNN-based models.

Also, CNN-based models such as that of Harwath & Glass (2017) are able to highlight specific parts in the visual input.[12] This would allow us to better experiment the competition phase by constructing images that contain instances of objects that could compete with a chosen target word, also figured in the same image. By observing the attention maps over the image, one could precisely measure competition between words. Moreover, using such a CNN-model would enable us to directly reproduce linguistic experiments that measure word recognition and competition using eye-tracking devices such as that of Huettig & McQueen (2007). Such an experiment was recently attempted by Duta & Plunkett (2020), however, the model they use does not work on raw images. Hence, a CNN-based model would allow us to exactly reproduce psycholinguistic experiments.

- **Image-To-Speech.** During an intership we did at the Nara Institute of Science and Technology (NAIST) in 2019, we tried to implement an Image-To-Speech model similar to those recently proposed by (Hsu et al. 2020, Wang et al. 2020) which directly produce a spoken description using an image as input, without requiring an intermediary textual representation. The model we implemented — which we did not describe in this thesis nor in any article — was only able to produce isolated sequences and not full captions. We believe our model was not able to learn how to produce full captions as we did not use any discrete units such as in the model proposed by Hsu et al. (2020), and we used an image encoder that processed raw images, instead of using an object detector first, as in the work of Wang et al. (2020).

  We would like to further work on such model to understand why our model did not succeed in predicting full captions. We would also like to study how such networks gradually learn to produce their first sentences, and compare this evolution to that of children. Indeed, when children learn how to speak, do not start by uttering full sentences, but rather start by producing isolated words. Later on, they produce two-word-long sentences and from that point on start to produce full sentences. It would be interesting to investigate if an Image-To-Speech model goes through the same steps as children, and if not, investigate why it is the case. Similarly, such experiment would provide insight on the linguistic development children go through by studying the representations learnt by such a model.

- **Discrete segmentation.** The experiments conducted in Chapter 6 reveal that giving an explicit segmentation improves the models' ability to corretly map a spoken caption to its visual context. Even though our experiments show that when a explicit segmentation is not given, the models implicitily segment the spoken input into sub-units, this implicit segmentation appears to be less effective.

  Hence, it would be desirable that the network learns to *explicitily* segment the spoken input into sub-units. Several options, which we started to explore, could be possible. Kreutzer & Sokolov (2018) and Chen et al. (2019) introduced a mechanism that allows recurrent neural networks to learn how to segment an input stream into sub-units. These methods essentially consist in predicting a binary value (1 or 0) if the current vector constitutes a segment boundary or not. Hence, given the $\text{GRU}_{\text{PACK.}}$ architecture we develop, it would simply consist in adding a binary classifier for each time frame. Yet, our initial experiments reveal that the definition of the loss function is critical so that the network does not learn to either consider each input vector as a

[12] They implicitly do so as the final matrix is a dot-product between the feature map of the image and the feature map of the spoken caption. Hence, some parts of the final matrix reflect close proximity between the two modalities.

segment of its own, or on the contrary, discard all the vectors except the last one and consider the full caption as one segment.

Recent work by Shain & Elsner (2020) suggest such approach is possible. However, their work reveal that their model is unable to learn large linguistic units such as words but only phoneme-like units. We believe this is the case as the spoken input is not grounded. Hence, by using a corpus similar to those used in this thesis, we believe such a model would be able to learn to segment both phoneme-like and word-like units.

In conclusion, neural models of visually grounded speech models offer invaluable opportunities to study and test hypotheses about child language acquisition, thanks to their ability to model complex interactions across several modalitites. New data sets, such as the SEEDLingS data set (Bergelson & Aslin 2017) or the data set recently collected by Tsutsui et al. (2020), which are large scale recordings collected in ecological environments (Dupoux 2018), will allow researchers to simulate language acquisition with more realistic data than ever before.

Appendix A

Personal Bibliography

- Havard, W. N., Besacier, L. & Chevrot, J.-P. (2020), 'Catplayinginthesnow: Impact of Prior Segmentation on a Model of Visually Grounded Speech', in 'Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL)', Association for Computational Linguistics, pp. 291–301.
  URL: https://www.aclweb.org/anthology/2020.conll-1.22

- *Zanon Boito M., *Havard W. N., Garnerin M., Le Ferrand E & Besacier L. (2020), 'MaSS: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the Bible'. In Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC2020), Marseille, France. European Language Resources Association (ELRA), pp. 6486–6493
  URL: https://www.aclweb.org/anthology/2020.lrec-1.799

- Havard, W. N., Chevrot, J.-P. & Besacier, L. (2019), Models of Visually Grounded Speech Signal Pay Attention to Nouns: A bilingual Experiment on English and Japanese, in 'ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 8618–8622.
  URL: https://doi.org/10.1109/icassp.2019.8683069

- Havard, W. N., Chevrot, J.-P. & Besacier, L. (2019), Word Recognition, Competition, and Activation in a Model of Visually Grounded Speech, in 'Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)', Association for Computational Linguistics, Hong Kong, China, pp. 339–348.
  URL: https://www.aclweb.org/anthology/K19-1032

- He, X., Tran, Q., Havard, W. N., Besacier, L., Zukerman, I. & Haffari, G. (2018) 'Exploring textual and speech information in dialogue act classification with speaker domain adaptation'. In 'Proceedings of the Australasian Language Technology Association Workshop 2018', Dunedin, New Zealand, pp. 61–65.
  URL: https://www.aclweb.org/anthology/U18-1007/

- Havard, W. N., Besacier, L. & Rosec, O. (2017), SPEECH-COCO: 600k Visually Grounded Spoken Captions Aligned to MSCOCO data Set, in 'Proceedings of the GLU 2017 International Workshop on Grounding Language Understanding', pp. 42–46.
  URL: http://dx.doi.org/10.21437/GLU.2017-9

- Havard, W. N., Chevrot, J.-P., & Besacier L. (2018), 'Emergence of attention in a neural model of visually grounded speech'. Learning Language in Humans and in Machines 2018 conference, ENS Paris, Poster (Peer Reviewed Abstract).
  URL: https://hal.archives-ouvertes.fr/hal-01970514/document

Appendix B

Attention: Highlighted Words

Word Rank	TRUE Peaks		RANDOM Peaks	
	Word	% peak	Word	% Peak
1	table	3.19	</s>	9.23
2	a	2.22	a	3.48
3	train	1.89	<s>	2.48
4	baseball	1.64	on	2.04
5	room	1.61	of	1.30
6	bathroom	1.52	with	1.27
7	of	1.23	in	1.21
8	in	1.11	man	1.12
9	sitting	1.03	standing	1.07
10	giraffe	0.97	and	1.02
11	bus	0.95	sitting	1.00
12	skateboard	0.94	is	0.88
13	on	0.93	<sil>	0.86
14	kitchen	0.90	the	0.84
15	with	0.84	table	0.83
16	ball	0.84	people	0.77
17	cat	0.77	woman	0.70
18	people	0.76	next	0.65
19	</s>	0.76	holding	0.58
20	and	0.74	two	0.58
21	horse	0.74	street	0.57
22	board	0.73	large	0.57
23	woman	0.72	person	0.56
24	parked	0.69	white	0.55
25	truck	0.69	to	0.54
26	umbrella	0.68	field	0.47
27	snow	0.68	plate	0.44
28	standing	0.67	cat	0.37
29	car	0.67	building	0.37
30	grass	0.65	top	0.37
31	teddy	0.60	group	0.36
32	dog	0.59	tennis	0.36
33	boat	0.58	water	0.35
34	bowl	0.57	down	0.35
35	surfboard	0.55	walking	0.35
36	park	0.51	riding	0.35
37	traffic	0.51	dog	0.34
38	motorcycle	0.50	pizza	0.34
39	tower	0.50	small	0.34
40	hydrant	0.49	black	0.33

Table B.1: Top 40 highlighted words on the COCO data set by GRU1.

Word Rank	TRUE Peaks		RANDOM Peaks	
	Word	**% peak**	**Word**	**% Peak**
1	train	2.04	</s>	9.15
2	tennis	1.73	a	3.47
3	toilet	1.53	<s>	2.47
4	baseball	1.50	on	1.96
5	skateboard	1.46	with	1.26
6	dog	1.45	in	1.22
7	cat	1.44	of	1.18
8	giraffe	1.39	man	1.08
9	pizza	1.35	and	1.04
10	kitchen	1.35	standing	1.02
11	street	1.34	sitting	0.93
12	sign	1.26	<sil>	0.87
13	bench	1.15	is	0.87
14	clock	1.12	the	0.86
15	bed	1.10	people	0.79
16	snow	1.10	table	0.76
17	cake	0.99	street	0.66
18	motorcycle	0.98	woman	0.63
19	room	0.96	next	0.61
20	beach	0.92	holding	0.59
21	bus	0.91	person	0.57
22	bathroom	0.86	two	0.56
23	bear	0.85	large	0.52
24	sink	0.84	white	0.52
25	horse	0.84	to	0.51
26	laptop	0.83	bathroom	0.47
27	phone	0.79	field	0.44
28	skis	0.73	train	0.44
29	food	0.73	skateboard	0.43
30	elephant	0.73	baseball	0.42
31	plate	0.72	riding	0.41
32	giraffes	0.71	down	0.41
33	frisbee	0.70	cat	0.37
34	boat	0.70	water	0.35
35	building	0.67	top	0.34
36	flying	0.65	group	0.34
37	surfboard	0.64	small	0.34
38	sheep	0.59	walking	0.34
39	is	0.59	building	0.33
40	sandwich	0.57	young	0.33

Table B.2: Top 40 highlighted words on the COCO data set by GRU5.

Word	TRUE Peaks			RANDOM Peaks		
Rank	Word	Translation	% peak	Word	Translation	% Peak
1	ga	SUBJ	3.88	</s>	*END OF SENTENCE*	9.96
2	no	GEN	3.17	i+ru	to be	5.73
3	o	OBJ	2.78	no	GEN	3.99
4	neko	cat	2.18	ga	SUBJ	3.50
5	sukeetoboodo	skateboard	2.00	ni	LOC, ALL	2.58
6	hikou	part of "hikouki" ("aeroplane") "hikou" ("aviation")	1.62	<s>	*START OF SENTENCE*	2.50
7	sukii	ski	1.27	o	OBJ	2.02
8	pasokon	Personnal Computer	1.26	dansei	man	1.68
9	piza	pizza	1.21	te	particle of reason, state	1.61
10	saafuboodo	surfboard	1.17	de	LOC	0.97
11	basu	bus	1.17	a+ru	to be	0.95
12	baiku	bike	1.16	josei	woman	0.85
13	uma	horse	1.16	shiro+i	white	0.79
14	banana	banana	0.98	to	enumeration particle	0.74
15	zou	elephant	0.96	<sil>	*<sil>ENCE*	0.64
16	shingou	traffic light	0.81	takusan	lots of	0.62
17	to	enumeration part.	0.81	sukeetoboodo	skateboard	0.54
18	saafin	surf	0.80	kuro+i	black	0.46
19	shimauma	zebra	0.77	shi	part of "shiteiru" ("is/are doing"), inflected form "suru" (to do)	0.44
20	mo+tte	holding	0.76	no+tte	riding	0.44
21	de	LOC	0.73	aka+i	red	0.44
22	kuruma	car	0.73	hito	man	0.41
23	kasa	umbrella	0.72	ta+tte	even if	0.41
24	kuma	bear	0.71	toma+tte	stopping	0.39
25	kiiro+i	white	0.70	naka	inside	0.37
26	benchi	bench	0.67	mo+tte	with	0.36
27	denwa	train	0.63	suwa+tte	sitting	0.36
28	kitchin	kitchen	0.63	ue	above	0.33
29	orenji	orange	0.61	ookina	big	0.33
30	doonatsu	donut	0.59	ta	PST	0.33
31	douro	road	0.59	hashi+tte	running	0.32
32	sunooboodo	snowboard	0.59	re	part of "reru" (passive) voice	0.31
33	beddo	bed	0.59	kodomo	children	0.31
34	suwa+tte	sitting	0.57	furisubii	frisbee	0.30
35	yuki	snowboard	0.53	teeburu	table	0.30
36	sara	plate	0.52	tokei	watch	0.29
37	takusan	lots of	0.50	kirin	giraffe	0.29
38	ha	TOP	0.50	saafuboodo	surboard	0.28
39	ni	LOC, ALL	0.49	ao+i	blue	0.28
40	mae	in front	0.48	neko	cat	0.28

Table B.3: Top 40 highlighted words on the STAIR data set by GRU1.

Word Rank	TRUE Peaks			RANDOM Peaks		
	Word	**Translation**	**% peak**	**Word**	**Translation**	**% Peak**
1	ga	SUBJ	13.18	</s>	*END OF SENTENCE*	9.97
2	no	GEN	5.35	i+ru	to be	5.65
3	o	OBJ	3.51	no	GEN	4.00
4	ni	LOC, ALL	3.34	ga	SUBJ	3.45
5	</s>	END OF SENTENCE	1.67	ni	LOC, ALL	2.55
6	kirin	giraffe	1.61	<s>	*START OF SENTENCE*	2.47
7	inu	dog	1.39	o	OBJ	2.05
8	uma	horse	1.27	dansei	man	1.79
9	baiku	bike	1.18	te	particle of reason, state	1.64
10	sukeetoboodo	skateboard	1.11	a+ru	to be	0.95
11	keeki	cake	1.09	de	de	0.95
12	shimauma	zebra	0.98	josei	woman	0.91
13	dansei	man	0.97	to	enumeration particle	0.75
14	pasokon	PC	0.96	shiro+i	white	0.74
15	nuigurumi	stuffed toy	0.96	<sil>	*SILENCE*	0.66
16	sha	end of jitensha ("bus") mistakenly tagged as particle	0.93	takusan	lots of	0.65
17	tokei	watch	0.88	no+tte	riding	0.46
18	hyoushiki	sign, mark	0.88	aka+i	red	0.45
19	furisubii	frisbee	0.85	kuro+i	black	0.45
20	piza	pizza	0.85	shi	four	0.44
21	kuruma	car	0.85	hito	man	0.42
22	shingou	sign	0.75	teeburu	table	0.40
23	benchi	bench	0.74	pasokon	PC	0.38
24	kitchin	kitchen	0.73	naka	inside	0.36
25	de	LOC	0.71	ta+tte	even if	0.36
26	burokkorii	broccoli	0.71	suwa+tte	sitting	0.35
27	woman	woman	0.67	ue	above	0.35
28	tenisu	tennis	0.62	mo+tte	with	0.35
29	toire	toilet	0.62	re	causative particle used for verbs	0.34
30	hitsuji	sheep	0.62	ta	PST	0.34
31	beddo	bed	0.61	toma+tte	stopping	0.33
32	to	enumeration particle	0.61	neko	cat	0.32
33	ryouri	cookery	0.60	sukeetoboodo	skateboard	0.31
34	raketto	racket	0.58	ookina	big	0.31
35	doonatsu	donut	0.58	kuruma	car	0.31
36	re	nominalising suffix	0.55	nuigurumi	stuffed toy	0.30
37	kanban	signboard	0.55	kodomo	children	0.30
38	densha	train	0.54	ki+ta	wear.PST	0.29
39	terebi	television	0.52	sukii	ski	0.29
40	senmen	part of "senmendai" ("washbasin"), senmen∼washing	0.51	o+ka	to put	0.28

Table B.4: Top 40 highlighted words on the STAIR data set by GRU5.

Word Rank	TRUE Peaks		RANDOM Peaks	
	Word	% Peak	Word	% Peak
1	dog	13.49	<s>	15.64
2	man	7.34	</s>	5.35
3	girl	5.42	<sil>	3.51
4	boy	5.38	a	3.02
5	dogs	4.02	in	1.80
6	people	3.37	man	1.65
7	woman	3.23	dog	1.53
8	child	2.12	on	1.21
9	ball	1.82	the	1.05
10	water	1.49	is	0.89
11	girls	1.38	two	0.84
12	children	1.26	with	0.80
13	boys	1.09	water	0.79
14	basketball	0.94	and	0.78
15	men	0.91	people	0.77
16	a	0.90	black	0.76
17	on	0.90	boy	0.71
18	while	0.87	woman	0.70
19	walking	0.80	white	0.66
20	wall	0.80	girl	0.64
21	person	0.79	wearing	0.62
22	black	0.76	playing	0.56
23	white	0.75	dogs	0.53
24	snowboarder	0.73	standing	0.51
25	rock	0.72	of	0.48
26	football	0.71	young	0.46
27	bike	0.69	child	0.46
28	little	0.64	person	0.42
29	snow	0.61	ball	0.41
30	wearing	0.55	grass	0.40
31	women	0.55	red	0.40
32	skateboarder	0.51	brown	0.40
33	<sil>	0.47	beach	0.39
34	with	0.46	shirt	0.38
35	skateboard	0.46	blue	0.38
36	large	0.41	little	0.36
37	walks	0.41	jumping	0.35
38	rides	0.41	running	0.34
39	walk	0.36	large	0.34
40	in	0.36	are	0.33

Table B.5: Top 40 highlighted words on the Flickr8k data set by GRU1.

Word Rank	TRUE Peaks		RANDOM Peaks	
	Word	% Peak	Word	% Peak
1	</s>	5.29	<s>	15.51
2	water	4.01	</s>	5.25
3	<sil>	2.80	<sil>	3.41
4	beach	1.91	a	2.95
5	snow	1.77	in	1.73
6	grass	1.74	dog	1.65
7	shirt	1.46	man	1.59
8	street	1.35	on	1.15
9	a	1.13	the	1.09
10	in	0.96	is	0.95
11	standing	0.89	two	0.89
12	is	0.69	woman	0.79
13	player	0.67	with	0.75
14	to	0.63	boy	0.75
15	field	0.62	black	0.74
16	dirt	0.61	girl	0.74
17	air	0.59	and	0.72
18	pool	0.57	people	0.69
19	swing	0.57	white	0.67
20	snowboarder	0.56	wearing	0.65
21	snowy	0.54	dogs	0.62
22	sitting	0.53	water	0.57
23	sunglasses	0.53	playing	0.57
24	soccer	0.48	standing	0.53
25	through	0.47	young	0.53
26	ocean	0.47	child	0.46
27	sand	0.46	person	0.46
28	the	0.44	of	0.45
29	surfer	0.44	brown	0.44
30	on	0.42	shirt	0.43
31	outside	0.42	red	0.42
32	her	0.42	running	0.42
33	his	0.42	small	0.40
34	with	0.41	grass	0.40
35	grassy	0.41	little	0.38
36	bicycle	0.41	are	0.38
37	camera	0.40	blue	0.37
38	skateboard	0.40	children	0.37
39	small	0.40	ball	0.37
40	together	0.39	jumping	0.35

Table B.6: Top 40 highlighted words on the FLICKR8K data set by GRU5.

Appendix C

Attention: Best Models' Scores

Model	R@1	R@5	R@10	Ranks
Epoch 15 - Saving Step 39	0.052	0.179	0.279	28.000
Epoch 15 - Saving Step 38	0.052	0.179	0.279	28.000
Epoch 14 - Saving Step 37	0.055	0.180	0.286	27.000
Epoch 13 - Saving Step 36	0.055	0.185	0.291	26.000
Epoch 12 - Saving Step 35	**0.058**	**0.188**	**0.292**	**26.000**
Epoch 11 - Saving Step 34	0.058	0.184	0.286	27.000
Epoch 10 - Saving Step 33	0.055	0.185	0.290	27.000
Epoch 9 - Saving Step 32	0.054	0.177	0.280	28.000
Epoch 8 - Saving Step 31	0.058	0.185	0.290	27.000
Epoch 7 - Saving Step 30	0.056	0.183	0.288	28.000
Epoch 6 - Saving Step 29	0.053	0.177	0.276	28.000
Epoch 5 - Saving Step 28	0.052	0.174	0.271	29.000
Epoch 4 - Saving Step 27	0.053	0.170	0.269	30.000
Epoch 3 - Saving Step 26	0.048	0.159	0.254	32.000
Epoch 3 - Saving Step 25	0.047	0.158	0.249	33.000
Epoch 2 - Saving Step 24	0.042	0.144	0.237	36.000
Epoch 2 - Saving Step 23	0.033	0.127	0.206	42.000
Epoch 1 - Saving Step 22	0.035	0.127	0.204	42.000
Epoch 1 - Saving Step 21	0.028	0.104	0.175	52.000
Epoch 1 - Saving Step 20	0.023	0.087	0.148	63.000
Epoch 1 - Saving Step 19	0.018	0.072	0.126	74.000
Epoch 1 - Saving Step 18	0.015	0.061	0.107	91.000
Epoch 1 - Saving Step 17	0.013	0.055	0.097	107.000
Epoch 1 - Saving Step 16	0.010	0.045	0.082	131.000
Epoch 1 - Saving Step 15	0.007	0.034	0.062	180.000
Epoch 1 - Saving Step 14	0.007	0.031	0.056	209.000
Epoch 1 - Saving Step 13	0.005	0.028	0.050	253.000
Epoch 1 - Saving Step 12	0.005	0.022	0.041	314.000
Epoch 1 - Saving Step 11	0.004	0.021	0.038	370.000
Epoch 1 - Saving Step 10	0.003	0.015	0.028	462.000
Epoch 1 - Saving Step 9	0.003	0.015	0.027	523.000
Epoch 1 - Saving Step 8	0.003	0.012	0.023	623.000
Epoch 1 - Saving Step 7	0.002	0.009	0.017	743.000
Epoch 1 - Saving Step 6	0.002	0.007	0.013	974.000
Epoch 1 - Saving Step 5	0.001	0.005	0.010	1108.000
Epoch 1 - Saving Step 4	0.001	0.004	0.008	1240.500
Epoch 1 - Saving Step 3	0.001	0.004	0.007	1515.000
Epoch 1 - Saving Step 2	0.001	0.003	0.006	1619.000
Epoch 1 - Saving Step 1	0.000	0.002	0.005	1905.000
Epoch 0 - Saving Step 0	0.000	0.001	0.003	2504.000

Table C.1: Scores obtained by our best model (selected on the validation set) on the COCO data set. Best epoch is shown in **bold**.

Model	R@1	R@5	R@10	Ranks
Epoch 15 - Saving Step 39	0.049	0.167	0.265	32.000
Epoch 15 - Saving Step 38	0.049	0.167	0.265	32.000
Epoch 14 - Saving Step 37	0.051	0.172	0.271	30.000
Epoch 13 - Saving Step 36	0.050	0.169	0.266	31.000
Epoch 12 - Saving Step 35	0.053	0.175	0.276	29.000
Epoch 11 - Saving Step 34	0.052	0.177	0.278	29.000
Epoch 10 - Saving Step 33	0.055	0.178	0.279	29.000
Epoch 9 - Saving Step 32	0.054	0.179	0.280	29.000
Epoch 8 - Saving Step 31	0.053	0.174	0.274	30.000
Epoch 7 - Saving Step 30	**0.055**	**0.182**	**0.282**	**29.000**
Epoch 6 - Saving Step 29	0.051	0.177	0.274	29.500
Epoch 5 - Saving Step 28	0.051	0.175	0.275	31.000
Epoch 4 - Saving Step 27	0.051	0.166	0.264	32.000
Epoch 4 - Saving Step 26	0.049	0.160	0.251	33.000
Epoch 3 - Saving Step 25	0.047	0.156	0.250	34.000
Epoch 2 - Saving Step 24	0.041	0.141	0.226	38.000
Epoch 2 - Saving Step 23	0.034	0.126	0.203	44.000
Epoch 1 - Saving Step 22	0.035	0.122	0.199	44.000
Epoch 1 - Saving Step 21	0.027	0.099	0.170	53.000
Epoch 1 - Saving Step 20	0.022	0.087	0.150	63.000
Epoch 1 - Saving Step 19	0.019	0.073	0.129	77.000
Epoch 1 - Saving Step 18	0.015	0.063	0.110	91.000
Epoch 1 - Saving Step 17	0.012	0.051	0.091	113.000
Epoch 1 - Saving Step 16	0.011	0.042	0.078	149.000
Epoch 1 - Saving Step 15	0.010	0.039	0.070	186.000
Epoch 1 - Saving Step 14	0.007	0.031	0.058	248.500
Epoch 1 - Saving Step 13	0.005	0.023	0.044	326.000
Epoch 1 - Saving Step 12	0.004	0.018	0.036	418.000
Epoch 1 - Saving Step 11	0.003	0.015	0.027	516.000
Epoch 1 - Saving Step 10	0.002	0.011	0.021	635.000
Epoch 1 - Saving Step 9	0.002	0.009	0.018	688.500
Epoch 1 - Saving Step 8	0.002	0.009	0.016	793.000
Epoch 1 - Saving Step 7	0.001	0.005	0.011	917.000
Epoch 1 - Saving Step 6	0.001	0.003	0.006	1243.000
Epoch 1 - Saving Step 5	0.001	0.004	0.007	1256.000
Epoch 1 - Saving Step 4	0.001	0.004	0.006	1491.000
Epoch 1 - Saving Step 3	0.001	0.003	0.006	1512.000
Epoch 1 - Saving Step 2	0.000	0.002	0.005	1872.000
Epoch 1 - Saving Step 1	0.001	0.002	0.005	2068.000
Epoch 0 - Saving Step 0	0.000	0.001	0.002	2529.000

Table C.2: Scores obtained by our best model (selected on the validation set) on the STAIR data set. Best epoch is shown in **bold**.

Model	R@1	R@5	R@10	Ranks
Epoch 25 - Saving Step 47	0.024	0.086	0.142	107.000
Epoch 25 - Saving Step 46	**0.024**	**0.086**	**0.142**	**107.000**
Epoch 24 - Saving Step 45	0.024	0.081	0.137	109.500
Epoch 23 - Saving Step 44	0.021	0.081	0.139	112.000
Epoch 22 - Saving Step 43	0.022	0.079	0.135	103.000
Epoch 21 - Saving Step 42	0.024	0.082	0.140	102.000
Epoch 20 - Saving Step 41	0.021	0.084	0.141	109.000
Epoch 20 - Saving Step 40	0.020	0.082	0.144	106.000
Epoch 19 - Saving Step 39	0.024	0.083	0.141	106.000
Epoch 18 - Saving Step 38	0.020	0.078	0.134	104.000
Epoch 17 - Saving Step 37	0.021	0.080	0.133	102.000
Epoch 16 - Saving Step 36	0.021	0.078	0.132	101.000
Epoch 16 - Saving Step 35	0.019	0.080	0.137	102.000
Epoch 15 - Saving Step 34	0.020	0.076	0.135	99.000
Epoch 14 - Saving Step 33	0.018	0.075	0.132	100.000
Epoch 13 - Saving Step 32	0.018	0.077	0.132	102.000
Epoch 13 - Saving Step 31	0.017	0.075	0.127	104.000
Epoch 12 - Saving Step 30	0.019	0.075	0.131	101.000
Epoch 11 - Saving Step 29	0.018	0.075	0.126	103.000
Epoch 11 - Saving Step 28	0.019	0.074	0.126	104.000
Epoch 10 - Saving Step 27	0.017	0.068	0.120	104.000
Epoch 9 - Saving Step 26	0.013	0.063	0.115	104.000
Epoch 9 - Saving Step 25	0.017	0.069	0.117	107.000
Epoch 8 - Saving Step 24	0.016	0.069	0.117	108.000
Epoch 8 - Saving Step 23	0.013	0.059	0.103	110.000
Epoch 7 - Saving Step 22	0.013	0.062	0.109	111.500
Epoch 7 - Saving Step 21	0.013	0.056	0.101	117.000
Epoch 6 - Saving Step 20	0.013	0.057	0.102	118.500
Epoch 6 - Saving Step 19	0.011	0.052	0.097	121.000
Epoch 5 - Saving Step 18	0.013	0.052	0.093	123.000
Epoch 5 - Saving Step 17	0.012	0.052	0.092	131.000
Epoch 4 - Saving Step 16	0.013	0.049	0.085	134.000
Epoch 4 - Saving Step 15	0.010	0.042	0.074	151.000
Epoch 3 - Saving Step 14	0.008	0.036	0.069	158.000
Epoch 3 - Saving Step 13	0.006	0.038	0.064	162.000
Epoch 3 - Saving Step 12	0.009	0.033	0.057	168.000
Epoch 2 - Saving Step 11	0.008	0.031	0.055	171.000
Epoch 2 - Saving Step 10	0.007	0.021	0.044	189.000
Epoch 2 - Saving Step 9	0.005	0.023	0.043	206.000
Epoch 1 - Saving Step 8	0.006	0.023	0.040	212.000
Epoch 1 - Saving Step 7	0.004	0.017	0.034	222.000
Epoch 1 - Saving Step 6	0.001	0.014	0.029	252.000
Epoch 1 - Saving Step 5	0.002	0.012	0.024	279.000
Epoch 1 - Saving Step 4	0.003	0.013	0.023	314.000
Epoch 1 - Saving Step 3	0.003	0.010	0.023	320.500
Epoch 1 - Saving Step 2	0.002	0.012	0.022	382.000
Epoch 1 - Saving Step 1	0.002	0.007	0.012	484.000
Epoch 0 - Saving Step 0	0.001	0.008	0.012	498.000

Table C.3: Scores obtained by our best model (selected on the validation set) on the Flickr8k data set. Best epoch is shown in **bold**.

Appendix D

Isolated Word Recognition

Concept		P@10	Concept		P@1
1	zebra	1.0	**41**	wine_glass	0.7
2	truck	1.0	**42**	tie	0.7
3	train	1.0	**43**	person	0.7
4	tennis_racket	1.0	**44**	orange	0.7
5	teddy_bear	1.0	**45**	cell_phone	0.7
6	surfboard	1.0	**46**	bear	0.7
7	sink	1.0	**47**	remote	0.6
8	sheep	1.0	**48**	donut	0.6
9	pizza	1.0	**49**	horse	0.5
10	kite	1.0	**50**	bottle	0.5
11	giraffe	1.0	**51**	apple	0.5
12	fire_hydrant	1.0	**52**	umbrella	0.4
13	elephant	1.0	**53**	potted_plant	0.4
14	dog	1.0	**54**	cup	0.4
15	dining_table	1.0	**55**	chair	0.4
16	clock	1.0	**56**	handbag	0.3
17	cat	1.0	**57**	backpack	0.3
18	cake	1.0	**58**	spoon	0.2
19	bus	1.0	**59**	oven	0.2
20	boat	1.0	**60**	book	0.2
21	bird	1.0	**61**	vase	0.1
22	bicycle	1.0	**62**	traffic_light	0.1
23	bed	1.0	**63**	knife	0.1
24	baseball_glove	1.0	**64**	fork	0.1
25	banana	1.0	**65**	carrot	0.1
26	airplane	1.0	**66**	tv	0.0
27	toilet	0.9	**67**	toothbrush	0.0
28	skateboard	0.9	**68**	toaster	0.0
29	refrigerator	0.9	**69**	suitcase	0.0
30	parking_meter	0.9	**70**	snowboard	0.0
31	laptop	0.9	**71**	skis	0.0
32	keyboard	0.9	**72**	scissors	0.0
33	car	0.9	**73**	mouse	0.0
34	bowl	0.9	**74**	microwave	0.0
35	baseball_bat	0.9	**75**	hot_dog	0.0
36	stop_sign	0.8	**76**	hair_drier	0.0
37	sports_ball	0.8	**77**	frisbee	0.0
38	sandwich	0.8	**78**	couch	0.0
39	motorcycle	0.8	**79**	broccoli	0.0
40	cow	0.8	**80**	bench	0.0

Table D.1: List of the 80 target words used for isolated word recognition with P@10.

List of Figures

List of Tables

Bibliography

Abney, S. P. (1992), Parsing By Chunks, *in* R. C. Berwick, S. P. Abney & C. Tenny, eds, 'Principle-Based Parsing: Computation and Psycholinguistics', Springer Netherlands, Dordrecht, pp. 257–278.
URL: *https://doi.org/10.1007/978-94-011-3474-3_10* (Cited in page 12.)

Alishahi, A., Barking, M. & Chrupała, G. (2017), Encoding of phonology in a recurrent neural model of grounded speech, *in* 'Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)', Association for Computational Linguistics, Vancouver, Canada, pp. 368–378.
URL: *https://www.aclweb.org/anthology/K17-1037* (Cited in pages 54, 55 et 185.)

Alishahi, A., Belinkov, Y., Chrupała, G., Hupkes, D., Pinter, Y. & Sajjad, H., eds (2020), *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Association for Computational Linguistics, Online.
URL: *https://www.aclweb.org/anthology/2020.blackboxnlp-1.0* (Cited in page 119.)

Ameka, F. K. (1999), 'Interjections', *Concise encyclopedia of grammatical categories* pp. 213–216. (Cited in page 18.)

Andersen, E. S., Dunlea, A. & Kekelis, L. (1993), 'The impact of input: language acquisition in the visually impaired', *First Language* **13**(37), 23–49.
URL: *http://journals.sagepub.com/doi/10.1177/014272379301303703* (Cited in pages 25, 27 et 174.)

Andersen, E. S., Dunlea, A. & Kekelis, L. S. (1984), 'Blind children's language: resolving some differences', *Journal of Child Language* **11**(3), 645–664. (Cited in pages 27, 28 et 174.)

Astington, J. & Dack, L. (2008), Theory of Mind, *in* M. M. Haith & J. B. Benson, eds, 'Encyclopedia of Infant and Early Childhood Development', Academic Press, San Diego, pp. 343–356.
URL: *https://www.sciencedirect.com/science/article/pii/B9780123708779001638* (Cited in page 23.)

Bahdanau, D., Cho, K. & Bengio, Y. (2015), Neural Machine Translation by Jointly Learning to Align and Translate, *in* 'ICLR 2015', San Diego, California, USA, pp. 3104–3112. (Cited in pages 43 et 181.)

Behrend, D. A. (1988), 'Overextensions in early language comprehension: evidence from a signal detection approach', *Journal of Child Language* **15**(1), 63–75.
URL: *https://doi.org/10.1017/s0305000900012058* (Cited in page 25.)

Belinkov, Y. & Glass, J. (2019), 'Analysis Methods in Neural Language Processing: A Survey', *Transactions of the Association for Computational Linguistics (TACL)* **7**, 49–72.
URL: *https://doi.org/10.1162/tacl%5Fa%5F00254* (Cited in page 119.)

Bender, E. M. & Koller, A. (2020), Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data, *in* 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)', Association for Computational Linguistics, Online, pp. 5185–5198.
URL: *https://www.aclweb.org/anthology/2020.acl-main.463* (Cited in pages 39, 178 et 179.)

Bergelson, E. & Aslin, R. N. (2017), 'Nature and origins of the lexicon in 6-mo-olds', *Proceedings of the National Academy of Sciences* **114**(49), 12916–12921.
URL: *https://doi.org/10.1073/pnas.1712966114* (Cited in pages 56, 123 et 210.)

Bergelson, E. & Swingley, D. (2012), 'At 6-9 months, human infants know the meanings of many common nouns', *Proceedings of the National Academy of Sciences* **109**(9), 3253–3258.
URL: *https://doi.org/10.1073/pnas.1113380109* (Cited in pages 18 et 171.)

Bhati, S., Villalba, J., Zelasko, P. & Dehak, N. (2020), Self-Expressing Autoencoders for Unsupervised Spoken Term Discovery, *in* H. Meng, B. Xu & T. F. Zheng, eds, 'Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020', ISCA, pp. 4876–4880.
URL: *https://doi.org/10.21437/Interspeech.2020-3000* (Cited in page 37.)

Bloom, P. (2002), *How Children Learn the Meanings of Words*, A Bradford Book, MIT Press. (Cited in pages 23, 121, 172, 173 et 209.)

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., Pecheux, M.-G., Ruel, J., Venuti, P. & Vyt, A. (2004), 'Cross-Linguistic Analysis of Vocabulary in Young Children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English', *Child Development* **75**(4), 1115–1139.
URL: *https://doi.org/10.1111/j.1467-8624.2004.00729.x* (Cited in page 86.)

Bortfeld, H., Morgan, J. L., Golinkoff, R. M. & Rathbun, K. (2005), 'Mommy and Me: familiar names help launch babies into speech-stream segmentation', *Psychological Science* **16**(4), 298–304.
URL: *https://doi.org/10.1111/j.0956-7976.2005.01531.x* (Cited in pages 17 et 171.)

Bosch, L., Figueras, M., Teixidó, M. & Ramon-Casas, M. (2013), 'Rapid gains in segmenting fluent speech when words match the rhythmic unit: evidence from infants acquiring syllable-timed languages', *Frontiers in Psychology* **4**.
URL: *https://doi.org/10.3389/fpsyg.2013.00106* (Cited in pages 11 et 169.)

Bosch, L. & Sebastián-Gallés, N. (1997), 'Native-language recognition abilities in 4-month-old infants from monolingual and bilingual environments', *Cognition* **65**(1), 33–69.
URL: *https://doi.org/10.1016/s0010-0277(97)00040-1* (Cited in page 10.)

Brent, M. R. & Siskind, J. M. (2001), 'The role of exposure to isolated words in early vocabulary development', *Cognition* **81**(2), B33–b44.
URL: *https://doi.org/10.1016/s0010-0277(01)00122-6* (Cited in page 17.)

Carey, S. & Bartlett, E. (1978), Acquiring a Single New Word, *in* 'Proceedings of the Stanford Child Language Conference 15', pp. 17–29. (Cited in page 23.)

Chen, Y., Huang, S., Lee, H., Wang, Y. & Shen, C. (2019), 'Audio Word2vec: Sequence-to-Sequence Autoencoding for Unsupervised Learning of Audio Segmentation and Representation', *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **27**(9), 1481–1493. (Cited in pages 38, 106, 122 et 210.)

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. & Bengio, Y. (2014), Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, *in* 'Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, pp. 1724–1734.
URL: *https://aclweb.org/anthology/D14-1179* (Cited in pages 43, 68, 181 et 189.)

Chomsky, N. (1969), *Aspects of the Theory of Syntax*, The MIT Press, MIT Press. (Cited in page 1.)

Chong, A. J., Vicenik, C. & Sundara, M. (2018), 'Intonation Plays a Role in Language Discrimination by Infants', *Infancy* **23**(6), 795–819.
URL: *https://doi.org/10.1111/infa.12257* (Cited in page 10.)

Christophe, A., Dupoux, E., Bertoncini, J. & Mehler, J. (1994), 'Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition', *The Journal of the Acoustical Society of America* **95**(3), 1570–1580.
URL: *https://doi.org/10.1121/1.408544* (Cited in pages 15 et 170.)

Christophe, A., Gout, A., Peperkamp, S. & Morgan, J. (2003), 'Discovering words in the continuous speech stream: the role of prosody', *Journal of Phonetics* **31**(3-4), 585–598.
URL: *https://doi.org/10.1016/s0095-4470(03)00040-8* (Cited in pages 12 et 170.)

Chrupała, G., Gelderloos, L. & Alishahi, A. (2017*a*), Representations of language in a model of visually grounded speech signal, *in* 'Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)', Association for Computational Linguistics, pp. 613–622.
URL: *https://aclweb.org/anthology/P17-1057* (Cited in pages 50, 52, 54, 55, 57, 58, 59, 64, 65, 67, 68, 72, 73, 75, 76, 83, 87, 90, 104, 111, 118, 119, 121, 184, 185, 187, 188, 189, 193, 195, 197, 198, 208 et 209.)

Chrupała, G., Gelderloos, L. & Alishahi, A. (2017*b*), 'Synthetically Spoken Coco', [Data set]
http://doi.org/10.5281/zenodo.400926.
URL: *https://zenodo.org/record/400926* (Cited in pages 64, 66 et 188.)

Chrupała, G., Kádár, Á. & Alishahi, A. (2015), Learning language through pictures, *in* 'Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP) (Volume 2: Short Papers)', Association for Computational Linguistics, Beijing, China, pp. 112–118.
URL: *https://www.aclweb.org/anthology/P15-2019* (Cited in pages 52 et 68.)

Church, R., Bernhardt, B., Shi, R. & Pichora-Fuller, K. (2005), 'Infant-directed speech: Final syllable lengthening and rate of speech', *The Journal of the Acoustical Society of America* **117**(4), 2429–2430.
URL: *https://doi.org/10.1121/1.4786663* (Cited in page 18.)

Clark, E. V. (1978), 'Strategies for Communicating', *Child Development* **49**(4), 953.
URL: *https://doi.org/10.2307/1128734* (Cited in page 25.)

Clark, E. V. (1987), The principle of contrast: A constraint on language acquisition, *in* B. MacWhinney, ed., 'Mechanisms of Language Acquisition', Lawrence Erlbaum Assoc., Hillsdale, NJ, pp. 1–33. (Cited in pages 24 et 173.)

Cole, R. A. (1973), 'Listening for mispronunciations: A measure of what we hear during speech', *Perception & Psychophysics* **13**(1), 153–156.
URL: *https://doi.org/10.3758/bf03207252* (Cited in page 30.)

Cotton, S. & Grosjean, F. (1984), 'The gating paradigm: A comparison of successive and individual presentation formats', *Perception & Psychophysics* **35**(1), 41–48.
URL: *https://doi.org/10.3758/BF03205923* (Cited in pages 92 et 199.)

Cristia, A., Dupoux, E., Gurven, M. & Stieglitz, J. (2019), 'Child-Directed Speech Is Infrequent in a Forager-Farmer Population: A time Allocation Study', *Child Development* **90**(3), 759–773.
URL: *https://srcd.onlinelibrary.wiley.com/doi/abs/10.1111/cdev.12974* (Cited in pages 19 et 172.)

Cristia, A. & Seidl, A. (2011), Sensitivity to prosody at 6 months predicts vocabulary at 24 months, *in* N. Danis, K. Mesh & H. Sung, eds, 'BUCLD 35: Proceedings of the 35th annual Boston University Conference on Language Development', Cascadilla Press, Somerville, Mass, pp. 145–156. (Cited in page 9.)

Curtin, S. & Hufnagle, D. (2009), Speech perception, *in* E. L. Bavin, ed., 'The Cambridge Handbook of Child Language', Cambridge Handbooks in Language and Linguistics, Cambridge University Press, p. 107–124. (Cited in page 18.)

Curtin, S., Mintz, T. & Byrd, D. (2001), Coarticulatory Cues Enhance Infants"Recognition of Syllable Sequences in Speech, *in* 'Proceedings of the 25th Annual Boston University Conference on Language Development', Vol. 1, pp. 191–201. (Cited in page 16.)

Cutler, A. & Norris, D. (1988), 'The role of strong syllables in segmentation for lexical access.', *Journal of Experimental Psychology: Human Perception and Performance* **14**(1), 113–121.
URL: *https://doi.org/10.1037/0096-1523.14.1.113* (Cited in pages 10 et 169.)

DeCasper, A. J. & Spence, M. J. (1986), 'Prenatal maternal speech influences newborns' perception of speech sounds', *Infant Behavior and Development* **9**(2), 133–150.
URL: *http://linkinghub.elsevier.com/retrieve/pii/0163638386900251* (Cited in page 9.)

Dehaene-Lambertz, G. & Houston, D. (1998), 'Language Discrimination Response Latencies In Two-Month-Old Infants', *Language and Speech* **41**, 21–43. (Cited in page 10.)

Di Cristo, A. (2013), *La prosodie de la parole*, Voix, parole, langage, De Boeck Supérieur. (Cited in pages 8, 9, 12, 19 et 169.)

Dixon, R. (2012), *Basic Linguistic Theory Volume 3: Further Grammatical Topics*, Basic Linguistic Theory, OUP Oxford. (Cited in pages 22 et 172.)

Drexler, J. & Glass, J. (2017), Analysis of Audio-Visual Features for Unsupervised Speech Recognition, *in* 'Proc. GLU 2017 International Workshop on Grounding Language Understanding', pp. 57–61.
URL: *http://dx.doi.org/10.21437/GLU.2017-12* (Cited in pages 55 et 185.)

Dunlea, A. (1989), *Vision and the emergence of meaning: blind and sighted children's early language*, Cambridge University Press, Cambridge [England] ; New York. (Cited in pages 27, 28, 29 et 174.)

Dupoux, E. (2018), 'Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner', *Cognition* **173**, 43–59.
URL: *http://www.sciencedirect.com/science/article/pii/S0010027717303013* (Cited in pages 56, 57, 123, 185, 186 et 210.)

Duta, M. & Plunkett, K. (2020), A neural Network Model of Lexical Competition during Infant Spoken Word Recognition, *in* S. Denison, M. Mack, Y. Xu & B. C. Armstrong, eds, 'Proceedings of the 42th Annual Meeting of the Cognitive Science Society - Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020, virtual, July 29 - August 1, 2020', cognitivesciencesociety.org.
URL: *https://cognitivesciencesociety.org/cogsci20/papers/0794/index.html* (Cited in page 122.)

Elliott, D., Frank, S., Sima'an, K. & Specia, L. (2016), Multi30K: Multilingual English-German Image Descriptions, *in* 'Proceedings of the 5th Workshop on Vision and Language', Association for Computational Linguistics, Berlin, Germany, pp. 70–74.
URL: *https://www.aclweb.org/anthology/W16-3210* (Cited in page 64.)

Elman, J. L. (1990), 'Finding Structure in Time', *Cognitive Science* **14**(2), 179–211.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1402_1* (Cited in pages 32, 42, 176 et 180.)

Elsner, M. & Shain, C. (2017), Speech segmentation with a neural encoder model of working memory, *in* 'Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Copenhagen, Denmark, pp. 1070–1080.
URL: *https://www.aclweb.org/anthology/D17-1112* (Cited in page 38.)

Emmorey, K. D. & Fromkin, V. A. (1988), The mental lexicon, *in* 'Linguistics: The Cambridge Survey', Cambridge University Press, pp. 124–149.
URL: *https://doi.org/10.1017/cbo9780511621062.006* (Cited in pages 8 et 169.)

Estes, K. G., Evans, J. L., Alibali, M. W. & Saffran, J. R. (2007), 'Can Infants Map Meaning to Newly Segmented Words?', *Psychological Science* **18**(3), 254–260.
URL: *https://doi.org/10.1111/j.1467-9280.2007.01885.x* (Cited in page 17.)

Fér, R., Matějka, P., Grézl, F., Plchot, O., Veselý, K. & Černocký, J. H. (2017), 'Multilingually trained bottleneck features in spoken language recognition', *Computer Speech & Language* **46**, 252–267.
URL: *http://www.sciencedirect.com/science/article/pii/S0885230816302947* (Cited in page 53.)

Ferguson, C. & Slobin, D. (1973), *Studies of child language development*, New York : Holt, Rinehart and Winston. (Cited in page 87.)

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B. & Fukui, I. (1989), 'A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants', *Journal of Child Language* **16**(3), 477–501.
URL: *https://doi.org/10.1017/s0305000900010679* (Cited in page 18.)

Fishman, S., Fishman, J., Ferguson, C. & Dasgupta, J. (1968), *Language Problems of Developing Nations*, Wiley. (Cited in page 36.)

Frank, M. C., Goodman, N. D. & Tenenbaum, J. B. (2009), 'Using Speakers' Referential Intentions to Model Early Cross-Situational Word Learning', *Psychological Science* **20**(5), 578–585.
URL: *https://doi.org/10.1111/j.1467-9280.2009.02335.x* (Cited in page 26.)

Friederici, A. D. & Wessels, J. M. I. (1993), 'Phonotactic knowledge of word boundaries and its use in infant speech perception', *Perception & Psychophysics* **54**(3), 287–295.
URL: *https://doi.org/10.3758/bf03205263* (Cited in page 13.)

Friedrich, M. (2008), 6. Neurophysiological correlates of picture-word priming in one-year-olds, *in* 'Early Language Development', John Benjamins Publishing Company, pp. 137–160.
URL: *https://doi.org/10.1075/tilar.5.08fri* (Cited in page 26.)

Fukushima, K. (1980), 'Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position', *Biological Cybernetics* **36**(4), 193–202.
URL: *https://doi.org/10.1007/bf00344251* (Cited in page 44.)

Gabriel, S., Maarten, V. & Dupoux, E. (2014), Learning Words from Images and Speech, *in* 'NIPS Workshop on Learning Semantics'. (Cited in pages 49, 50 et 183.)

Ganong, W. F. (1980), 'Phonetic categorization in auditory word perception.', *Journal of Experimental Psychology: Human Perception and Performance* **6**(1), 110–125.
URL: *https://doi.org/10.1037/0096-1523.6.1.110* (Cited in page 32.)

Gaskell, M. G. & Marslen-Wilson, W. D. (1997), 'Integrating Form and Meaning: A distributed Model of Speech Perception', *Language and Cognitive Processes* **12**(5-6), 613–656.
URL: *https://doi.org/10.1080/016909697386646* (Cited in pages 32 et 176.)

Gentner, D. (1982), 'Why nouns are learned before verbs: Linguistic relativity versus natural partitioning', *Language* **2**, 301–334. (Cited in pages 86 et 196.)

Glass, J. (2012), Towards unsupervised speech processing, *in* '2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)', IEEE, pp. 1–4.
URL: *https://doi.org/10.1109/isspa.2012.6310546* (Cited in pages 36, 141 et 178.)

Goldberg, Y. (2017), *Neural Network Methods for Natural Language Processing*, Vol. 10, Morgan & Claypool Publishers LLC.
URL: *https://doi.org/10.2200/s00762ed1v01y201703hlt037* (Cited in page 47.)

Goldin-Meadow, S. (2009), From gesture to word, *in* E. L. Bavin, ed., 'The Cambridge Handbook of Child Language', Cambridge Handbooks in Language and Linguistics, Cambridge University Press, p. 145–160. (Cited in page 26.)

Goldinger, S. D., Luce, P. A. & Pisoni, D. B. (1989), 'Priming lexical neighbors of spoken words: Effects of competition and inhibition', *Journal of Memory and Language* **28**(5), 501–518.
URL: *https://doi.org/10.1016/0749-596x(89)90009-0* (Cited in page 33.)

Goldwater, S. (2006), Nonparametric Bayesian Models of Lexical Acquisition, PhD thesis, Brown University. (Cited in pages 37 et 178.)

Gonzalez-Gomez, N., Poltrock, S. & Nazzi, T. (2013), 'A "Bat" Is Easier to Learn than a "Tab": Effects of Relative Phonotactic Frequency on Infant Word Learning', *PLoS ONE* **8**(3), e59601.
URL: *https://doi.org/10.1371/journal.pone.0059601* (Cited in page 13.)

Goodfellow, I., Bengio, Y. & Courville, A. (2016), *Deep Learning*, Adaptive Computation and Machine Learning series, MIT Press. (Cited in page 40.)

Grosjean, F. (1980), 'Spoken word recognition processes and the gating paradigm', *Perception & Psychophysics* **28**(4), 267–283.
URL: *https://doi.org/10.3758/BF03204386* (Cited in pages 2, 4, 90, 92, 93, 101, 120, 168, 199, 201 et 208.)

Grosjean, F. (1985), 'The recognition of words after their acoustic offset: Evidence and implications', *Perception & Psychophysics* **38**(4), 299–310.
URL: *https://doi.org/10.3758/bf03207159* (Cited in pages 30, 120 et 208.)

Harwath, D., Chuang, G. & Glass, J. (2018), Vision as an Interlingua: Learning Multilingual Semantic Embeddings of Untranscribed Speech, *in* '2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 4969–4973. (Cited in page 82.)

Harwath, D. & Glass, J. (2015), Deep multimodal semantic embeddings for speech and images, *in* '2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)', pp. 237–244. (Cited in pages 49, 50, 52, 55, 65, 67, 68, 107, 183, 184 et 188.)

Harwath, D. & Glass, J. R. (2017), Learning Word-Like Units from Joint Audio-Visual Analysis, *in* R. Barzilay & M. Kan, eds, 'Proceedings of the 55th Annual Meeting of the Association for Coputational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers', Association for Computational Linguistics, pp. 506–517.
URL: *https://doi.org/10.18653/v1/P17-1047* (Cited in pages 50, 55, 58, 68, 90, 121, 122, 187, 209 et 210.)

Harwath, D. & Glass, J. R. (2019), Towards Visually Grounded Sub-word Speech Unit Discovery, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019', Ieee, pp. 3017–3021.
URL: *https://doi.org/10.1109/ICASSP.2019.8682666* (Cited in page 55.)

Harwath, D., Hsu, W.-N. & Glass, J. (2020), Learning Hierarchical Discrete Linguistic Units from Visually-Grounded Speech, *in* '8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020', OpenReview.net.
URL: *https://openreview.net/forum?id=B1elCp4KwH* (Cited in page 112.)

Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A. & Glass, J. (2018), Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input, *in* V. Ferrari, M. Hebert, C. Sminchisescu & Y. Weiss, eds, 'Computer Vision – ECCV 2018', Springer International Publishing, Cham, pp. 659–677. (Cited in pages 51, 90 et 183.)

Harwath, D., Torralba, A. & Glass, J. R. (2016), Unsupervised Learning of Spoken Language with Visual Context, *in* 'Proceedings of the 30th International Conference on Neural Information Processing Systems', Nips'16, Curran Associates Inc., Red Hook, NY, USA, p. 1866–1874. (Cited in pages 50, 51, 52, 55, 57, 82, 121, 183, 184, 185 et 209.)

Haryu, E. & Kajikawa, S. (2016), 'Use of bound morphemes (noun particles) in word segmentation by Japanese-learning infants', *Journal of Memory and Language* **88**(C), 18–27. (Cited in pages 18, 86, 171 et 196.)

Havard, W., Besacier, L. & Chevrot, J.-P. (2020), Catplayinginthesnow: Impact of Prior Segmentation on a Model of Visually Grounded Speech, *in* 'Proceedings of the 24th Conference on Computational Natural Language Learning', Association for Computational Linguistics, Online, pp. 291–301.
URL: *https://www.aclweb.org/anthology/2020.conll-1.22* (Cited in page 103.)

Havard, W., Besacier, L. & Rosec, O. (2017), SPEECH-cOCO: 600k Visually Grounded Spoken Captions Aligned to MSCOCO data Set, *in* 'Proc. GLU 2017 International Workshop on Grounding Language Understanding', pp. 42–46.
URL: *http://dx.doi.org/10.21437/GLU.2017-9* (Cited in page 65.)

Havard, W. N., Chevrot, J.-P. & Besacier, L. (2019*a*), Models of Visually Grounded Speech Signal Pay Attention to Nouns: A bilingual Experiment on English and Japanese, *in* 'ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 8618–8622.
URL: *https://doi.org/10.1109/icassp.2019.8683069* (Cited in page 73.)

Havard, W. N., Chevrot, J.-P. & Besacier, L. (2019*b*), 'Synthetically Spoken STAIR', [Data set]
https://zenodo.org/record/1495070.
URL: *https://zenodo.org/record/1495070* (Cited in pages 65, 72 et 188.)

Havard, W. N., Chevrot, J.-P. & Besacier, L. (2019*c*), Word Recognition, Competition, and Activation in a Model of Visually Grounded Speech, *in* 'Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)', Association for Computational Linguistics, Hong Kong, China, pp. 339–348.
URL: *https://www.aclweb.org/anthology/K19-1032* (Cited in pages 89 et 100.)

Havron, N., Raviv, L. & Arnon, I. (2018), 'Literate and preliterate children show different learning patterns in an artificial language learning task', *Journal of Cultural Cognitive Science* **2**(1-2), 21–33.
URL: *https://doi.org/10.1007/s41809-018-0015-9* (Cited in pages 104 et 121.)

He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep Residual Learning for Image Recognition, *in* '2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016', IEEE Computer Society, pp. 770–778.
URL: *https://doi.org/10.1109/CVPR.2016.90* (Cited in page 68.)

Hepper, P. G. & Shahidullah, B. S. (1994), 'Development of fetal hearing', *Archives of Disease in Childhood Fetal and Neonatal edition* **71**(2), F81–f87.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1061088/* (Cited in pages 8 et 9.)

Hermann, K. M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., Szepesvari, D., Czarnecki, W. M., Jaderberg, M., Teplyashin, D., Wainwright, M., Apps, C., Hassabis,

D. & Blunsom, P. (2017), 'Grounded Language Learning in a Simulated 3D world'. (Cited in page 53.)

Hill, F., Clark, S., Hermann, K. M. & Blunsom, P. (2020), Understanding Early Word Learning in Situated Connectionist Agents, *in* 'Proceedings of the 42th Annual Meeting of the Cognitive Science Society - Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020, virtual, July 29 - August 1, 2020', cognitivesciencesociety.org.
URL: *https://cognitivesciencesociety.org/cogsci20/papers/0155/index.html* (Cited in pages 53 et 57.)

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural computation* **9**(8), 1735–1780. (Cited in pages 42 et 43.)

Hodosh, M., Young, P. & Hockenmaier, J. (2013), 'Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics', *Journal of Artificial Intelligence Research* **47**, 853–899.
URL: *https://doi.org/10.1613/jair.3994* (Cited in pages 55, 65, 185 et 188.)

Hohenberger, A., Altan, A., Kaya, U., Özgün Köksal Tuncer & Avcu, E. (2016), Sensitivity of Turkish infants to vowel harmony, *in* 'The Acquisition of Turkish in Childhood', John Benjamins Publishing Company, pp. 29–56.
URL: *https://doi.org/10.1075/tilar.20.02hon* (Cited in page 15.)

Höhle, B., Weissenborn, J., Kiefer, D., Schulz, A. & Schmitz, M. (2004), 'Functional Elements in Infants' Speech Processing: The Role of Determiners in the Syntactic Categorization of Lexical Elements', *Infancy* **5**(3), 341–353.
URL: *https://doi.org/10.1207/s15327078in0503_5* (Cited in page 18.)

Hohne, E. A. & Jusczyk, P. W. (1994), 'Two-month-old infants' sensitivity to allophonic differences', *Perception & Psychophysics* **56**(6), 613–623.
URL: *https://doi.org/10.3758/bf03208355* (Cited in pages 15 et 171.)

Hollich, G., Golinkoff, R. M. & Hirsh-Pasek, K. (2007), 'Young children associate novel words with complex objects rather than salient parts.', *Developmental Psychology* **43**(5), 1051–1061.
URL: *https://doi.org/10.1037/0012-1649.43.5.1051* (Cited in page 24.)

Horst, J. S. & Hout, M. C. (2015), 'The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research', *Behavior Research Methods* **48**(4), 1393–1409.
URL: *https://doi.org/10.3758/s13428-015-0647-3* (Cited in page 53.)

Houston, D. M., Jusczyk, P. W., Kuijpers, C., Coolen, R. & Cutler, A. (2000), 'Cross-language word segmentation by 9-month-olds', *Psychonomic Bulletin & Review* **7**(3), 504–509.
URL: *https://doi.org/10.3758/bf03214363* (Cited in pages 10 et 169.)

Hsu, W.-N., Harwath, D., Song, C. & Glass, J. (2020), 'Text-Free Image-to-Speech Synthesis Using Learned Segmental Units'. (Cited in pages 122 et 210.)

Huettig, F. & McQueen, J. M. (2007), 'The tug of war between phonological, semantic and shape information in language-mediated visual search', *Journal of Memory and Language* **57**(4), 460–482.
URL: *https://doi.org/10.1016/j.jml.2007.02.001* (Cited in pages 122 et 210.)

Inagaki, K., Hatano, G. & Otake, T. (2000), 'The Effect of Kana Literacy Acquisition on the Speech Segmentation Unit Used by Japanese Young Children', *Journal of Experimental Child Psychology* **75**(1), 70–91.
URL: *https://doi.org/10.1006/jecp.1999.2523* (Cited in page 11.)

Jain, S. & Wallace, B. C. (2019), Attention is not Explanation, *in* 'Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long and Short Papers)', Association for Computational Linguistics, Minneapolis, Minnesota, pp. 3543–3556.
URL: *https://www.aclweb.org/anthology/N19-1357* (Cited in pages 74 et 192.)

James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde Farley & Yoshua Bengio (2010), Theano: A CPU and GPU Math Compiler in Python, *in* Stéfan van der Walt & Jarrod Millman, eds, 'Proceedings of the 9th Python in Science Conference', pp. 18–24. (Cited in pages 67 et 189.)

Jansen, A. & Van Durme, B. (2011), Efficient spoken term discovery using randomized algorithms, *in* '2011 IEEE Workshop on Automatic Speech Recognition Understanding', pp. 401–406. (Cited in page 37.)

Jespersen, O. (1929), Monosyllabism in English : Biennial Lecture on english philology, *in* H. Milford, ed., 'Proceedings of the British Academy', Vol. Xiv, London : British academy. (Cited in page 111.)

Johnson, E. (2003), 'Lexical viability constraints on speech segmentation by infants', *Cognitive Psychology* **46**(1), 65–97.
URL: *https://doi.org/10.1016/s0010-0285(02)00507-8* (Cited in page 20.)

Johnson, E. K. & Jusczyk, P. W. (2001), 'Word Segmentation by 8-Month-Olds: When Speech Cues Count More Than Statistics', *Journal of Memory and Language* **44**(4), 548–567.
URL: *https://doi.org/10.1006/jmla.2000.2755* (Cited in pages 16 et 19.)

Johnson, E. K., Seidl, A. & Tyler, M. D. (2014), 'The Edge Factor in Early Word Segmentation: Utterance-Level Prosody Enables Word Form Extraction by 6-Month-Olds', *PLoS ONE* **9**(1), e83546.
URL: *https://doi.org/10.1371/journal.pone.0083546* (Cited in pages 12, 18, 21, 170 et 171.)

Johnson, E. K. & Tyler, M. D. (2010), 'Testing the limits of statistical learning for word segmentation', *Developmental Science* **13**(2), 339–345.
URL: *https://doi.org/10.1111/j.1467-7687.2009.00886.x* (Cited in pages 20 et 172.)

Johnson, M., Griffiths, T. & Goldwater, S. (2007), Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models, *in* B. Scholkopf, J. Platt & T. Hofmann, eds, 'Advances in Neural Information Processing Systems 19 - Proceedings of the 2006 Conference', MIT Press, pp. 641–648. (Cited in page 37.)

Jusczyk, P. (2000), *The Discovery of Spoken Language*, A Bradford Book, MIT Press. (Cited in page 9.)

Jusczyk, P. & Aslin, R. (1995), 'Infants' Detection of the Sound Patterns of Words in Fluent Speech', *Cognitive Psychology* **29**(1), 1–23.
URL: *https://doi.org/10.1006/cogp.1995.1010* (Cited in pages 8 et 17.)

Jusczyk, P., Friederici, A., Wessels, J., Svenkerud, V. & Jusczyk, A. (1993), 'Infants' Sensitivity to the Sound Patterns of Native Language Words', *Journal of Memory and Language* **32**(3), 402–420.
URL: *https://doi.org/10.1006/jmla.1993.1022* (Cited in pages 13 et 170.)

Jusczyk, P. W., Cutler, A. & Redanz, N. J. (1993), 'Infants' Preference for the Predominant Stress Patterns of English Words', *Child Development* **64**(3), 675.
URL: *https://doi.org/10.2307/1131210* (Cited in pages 10 et 169.)

Jusczyk, P. W. & Hohne, E. A. (1997), 'Infants' Memory for Spoken Words', *Science* **277**(5334), 1984–1986.
URL: *https://doi.org/10.1126/science.277.5334.1984* (Cited in page 17.)

Jusczyk, P. W., Hohne, E. A. & Bauman, A. (1999), 'Infants' sensitivity to allophonic cues for word segmentation', *Perception & Psychophysics* **61**(8), 1465–1476.
URL: *https://doi.org/10.3758/bf03213111* (Cited in page 15.)

Jusczyk, P. W., Houston, D. M. & Newsome, M. (1999), 'The Beginnings of Word Segmentation in English-Learning Infants', *Cognitive Psychology* **39**(3-4), 159–207.
URL: *https://doi.org/10.1006/cogp.1999.0716* (Cited in pages 10, 11 et 20.)

Jusczyk, P. W., Luce, P. A. & Charles-Luce, J. (1994), 'Infants' Sensitivity to Phonotactic Patterns in the Native Language', *Journal of Memory and Language* **33**(5), 630–645.
URL: *https://doi.org/10.1006/jmla.1994.1030* (Cited in pages 13 et 170.)

Kádár, Á., Chrupała, G. & Alishahi, A. (2015), Lingusitic Analysis of Multi-Modal Recurrent Neural Networks, *in* 'Proceedings of the Fourth Workshop on Vision and Language', Association for Computational Linguistics, Lisbon, Portugal, pp. 8–9.
URL: *https://www.aclweb.org/anthology/W15-2804* (Cited in page 52.)

Kamper, H. (2017), Unsupervised neural and Bayesian models for zero-resource speech processing, PhD thesis, University of Edinburgh.
URL: *http://hdl.handle.net/1842/25432* (Cited in pages 37 et 178.)

Kamper, H., Anastassiou, A. & Livescu, K. (2019), Semantic Query-by-example Speech Search Using Visual Grounding, *in* 'ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 7120–7124. (Cited in page 51.)

Kamper, H., Shakhnarovich, G. & Livescu, K. (2019), 'Semantic Speech Retrieval With a Visually Grounded Model of Untranscribed Speech', *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **27**(1), 89–98. (Cited in page 67.)

Karpathy, A. & Li, F. F. (2017), 'Deep Visual-Semantic Alignments for Generating Image Descriptions', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(4), 664–676.
URL: *https://doi.org/10.1109/TPAMI.2016.2598339* (Cited in pages 50, 65, 66 et 188.)

Ketrez, F. N. (2013), 'Harmonic cues for speech segmentation: a cross-linguistic corpus study on child-directed speech', *Journal of Child Language* **41**(2), 439–461.
URL: *https://doi.org/10.1017/s0305000912000724* (Cited in pages 14 et 170.)

Khorrami, K. & Räsänen, O. (2021), 'Can phones, syllables, and words emerge as side-products of cross-situational audiovisual learning? - A computational investigation'.
URL: *https://doi.org/10.31234/osf.io/37zna* (Cited in pages 50, 121 et 209.)

Kirkham, N. Z., Slemmer, J. A. & Johnson, S. P. (2002), 'Visual statistical learning in infancy: evidence for a domain general learning mechanism', *Cognition* **83**(2), B35–b42.
URL: *https://doi.org/10.1016/s0010-0277(02)00004-5* (Cited in page 1.)

Kisler, T., Reichel, U. & Schiel, F. (2017), 'Multilingual processing of speech via web services', *Computer Speech & Language* **45**, 326–347. (Cited in pages 66 et 189.)

Kreutzer, J. & Sokolov, A. (2018), 'Learning to Segment Inputs for NMT Favors Character-Level Processing', *Proceedings of the International Workshop on Spoken Language Translation October 29-30, 2018 Bruges, Belgium* **1**, 166–172.
URL: *https://workshop2018.iwslt.org/downloads/Proceedings%5FIWSLT%5F2018.pdf* (Cited in pages 106, 122 et 210.)

Krishnamohan, V., Soman, A., Gupta, A. & Ganapathy, S. (2020), Audiovisual Correspondence Learning in Humans and Machines, *in* 'Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020', pp. 4462–4466.
URL: *http://dx.doi.org/10.21437/Interspeech.2020-2674* (Cited in pages 53 et 185.)

Kubozono, H., ed. (2015), *Handbook of Japanese Phonetics and Phonology*, De Gruyter.
URL: *https://doi.org/10.1515/9781614511984* (Cited in page 12.)

Laing, C. (2019), 'A role for onomatopoeia in early language: evidence from phonological development', *Language and Cognition* **11**(02), 173–187.
URL: *https://doi.org/10.1017/langcog.2018.23* (Cited in page 48.)

Landau, B. & Gleitman, L. (1985), *Language and Experience: Evidence from the Blind Child*, Cognitive Science Series, Harvard University Press. (Cited in pages 1, 25, 27, 167 et 174.)

Landau, B., Smith, L. B. & Jones, S. S. (1988), 'The importance of shape in early lexical learning', *Cognitive Development* **3**(3), 299–321.
URL: *https://doi.org/10.1016/0885-2014(88)90014-7* (Cited in page 25.)

Law, B., Houston-Price, C. & Loucas, T. (2012), Using Gaze Direction to Learn Words at 18 Months: Relationships with Later Vocabulary, *in* 'Language Studies Working Papers', pp. 3–14. (Cited in page 26.)

LeCun, Y., Haffner, P., Bottou, L. & Bengio, Y. (1999), *Object Recognition with Gradient-Based Learning*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 319–345.
URL: *https://doi.org/10.1007/3-540-46805-6_19* (Cited in page 44.)

Lee, C.-y., O'Donnell, T. J. & Glass, J. (2015), 'Unsupervised Lexicon Discovery from Acoustic Input', *Transactions of the Association for Computational Linguistics (TACL)* **3**, 389–403.
URL: *https://www.aclweb.org/anthology/Q15-1028* (Cited in pages 36, 37 et 178.)

Li, X., Xu, C., Wang, X., Lan, W., Jia, Z., Yang, G. & Xu, J. (2019), 'COCO-cN for Cross-Lingual Image Tagging, Captioning, and Retrieval', *IEEE Transactions on Multimedia* **21**(9), 2347–2360. (Cited in page 65.)

Lila R. Gleitman, E., Wanner, E., Gleitman, L. & Press, C. U. (1982), *Language Acquisition: The State of the Art*, Cambridge University Press. (Cited in page 9.)

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014), Microsoft COCO: Common Objects in Context, *in* D. Fleet, T. Pajdla, B. Schiele & T. Tuytelaars, eds, 'Computer Vision – ECCV 2014', Springer International Publishing, Cham, pp. 740–755. (Cited in pages 55, 64, 185 et 188.)

Linzen, T., Chrupała, G. & Alishahi, A., eds (2018), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Association for Computational Linguistics, Brussels, Belgium.
URL: *https://www.aclweb.org/anthology/W18-5400* (Cited in page 119.)

Linzen, T., Chrupała, G., Belinkov, Y. & Hupkes, D., eds (2019), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Association for Computational Linguistics, Florence, Italy.
URL: *https://www.aclweb.org/anthology/W19-4800* (Cited in page 119.)

Ma, W., Golinkoff, R. M., Houston, D. M. & Hirsh-Pasek, K. (2011), 'Word Learning in Infant- and Adult-Directed Speech', *Language Learning and Development* **7**(3), 185–201.
URL: *https://doi.org/10.1080/15475441.2011.579839* (Cited in pages 19 et 171.)

MacKenzie, H., Curtin, S. & Graham, S. A. (2012), '12-Month-Olds' Phonotactic Knowledge Guides Their Word-Object Mappings', *Child Development* **83**(4), 1129–1136.
URL: *https://doi.org/10.1111/j.1467-8624.2012.01764.x* (Cited in pages 13 et 170.)

Magnuson, J. S., Mirman, D. & Harris, H. D. (2012), Computational Models of Spoken Word Recognition, *in* 'The Cambridge Handbook of Psycholinguistics', Cambridge University Press, pp. 76–103.
URL: *https://doi.org/10.1017/cbo9781139029377.006* (Cited in page 32.)

Magnuson, J. S., Mirman, D. & Myers, E. (2013), *Spoken Word Recognition*, Oxford University Press.
URL: *https://doi.org/10.1093/oxfordhb/9780195376746.013.0027* (Cited in pages 29 et 174.)

Markman, E. M. (1990), 'Constraints Children Place on Word Meanings', *Cognitive Science* **14**(1), 57–77.
URL: *https://doi.org/10.1207/s15516709cog1401_4* (Cited in pages 24 et 173.)

Markman, E. M. & Hutchinson, J. E. (1984), 'Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations', *Cognitive Psychology* **16**(1), 1–27.
URL: *https://doi.org/10.1016/0010-0285(84)90002-1* (Cited in page 24.)

Marr, D. (1977), 'Artificial intelligence—A personal view', *Artificial Intelligence* **9**(1), 37–48.
URL: *https://doi.org/10.1016/0004-3702(77)90013-3* (Cited in page 56.)

Marr, D. (1983), *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, Henry Holt and Company. (Cited in page 56.)

Marslen-Wilson, W. D. (1987*a*), 'Functional parallelism in spoken word-recognition', *Cognition* **25**(1-2), 71–102.
URL: *https://doi.org/10.1016/0010-0277(87)90005-9* (Cited in page 30.)

Marslen-Wilson, W. D. (1987*b*), 'Functional parallelism in spoken word-recognition', *Cognition* **25**(1), 71–102. Special Issue Spoken Word Recognition. (Cited in pages 93, 100 et 201.)

Marslen-Wilson, W. D. & Welsh, A. (1978), 'Processing interactions and lexical access during word recognition in continuous speech', *Cognitive Psychology* **10**(1), 29–63.
URL: *http://www.sciencedirect.com/science/article/pii/001002857890018X* (Cited in pages 29, 100, 175 et 201.)

Mattys, S. & Bortfeld, H. (2015), Speech segmentation, *in* G. Gaskell & J. Mirkovic , eds, 'Speech Perception and Spoken Word Recognition', Taylor and Francis. (Cited in pages 22 et 141.)

Mattys, S. L. & Jusczyk, P. W. (2001*a*), 'Do infants segment words or recurring contiguous patterns?', *Journal of Experimental Psychology: Human Perception and Performance* **27**(3), 644–655.
URL: *https://doi.org/10.1037/0096-1523.27.3.644* (Cited in pages 15 et 171.)

Mattys, S. L. & Jusczyk, P. W. (2001*b*), 'Phonotactic cues for segmentation of fluent speech by infants', *Cognition* **78**(2), 91–121.
URL: *https://doi.org/10.1016/s0010-0277(00)00109-8* (Cited in pages 14 et 170.)

Mattys, S. L., Jusczyk, P. W., Luce, P. A. & Morgan, J. L. (1999), 'Phonotactic and Prosodic Effects on Word Segmentation in Infants', *Cognitive Psychology* **38**(4), 465–494.
URL: *https://doi.org/10.1006/cogp.1999.0721* (Cited in pages 14 et 20.)

Mattys, S. L., White, L. & Melhorn, J. F. (2005), 'Integration of Multiple Speech Segmentation Cues: A hierarchical Framework.', *Journal of Experimental Psychology: General* **134**(4), 477–500.
URL: *https://doi.org/10.1037/0096-3445.134.4.477* (Cited in pages 13, 21, 70 et 170.)

May, L., Byers-Heinlein, K., Gervain, J. & Werker, J. F. (2011), 'Language and the Newborn Brain: Does Prenatal Language Experience Shape the Neonate Neural Response to Speech?', *Frontiers in Psychology* **2**.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3177294/* (Cited in page 9.)

McClelland, J. L. & Elman, J. L. (1986*a*), 'The TRACE model of speech perception', *Cognitive Psychology* **18**(1), 1–86.
URL: *https://doi.org/10.1016/0010-0285(86)90015-0* (Cited in pages 31 et 175.)

McClelland, J. L. & Elman, J. L. (1986*b*), 'The TRACE model of speech perception', *Cognitive Psychology* **18**(1), 1–86.
URL: *http://www.sciencedirect.com/science/article/pii/0010028586900150* (Cited in pages 100 et 201.)

McDonough, C., Song, L., Hirsh-Pasek, K., Golinkoff, R. M. & Lannon, R. (2011), 'An image is worth a thousand words: why nouns tend to dominate verbs in early word learning', *Developmental Science* **14**(2), 181–189.
URL: *https://doi.org/10.1111/j.1467-7687.2010.00968.x* (Cited in page 86.)

McGurk, H. & MacDonald, J. (1976), 'Hearing lips and seeing voices', *Nature* **264**(5588), 746–748.
URL: *https://doi.org/10.1038/264746a0* (Cited in page 71.)

Merkx, D. & Frank, S. L. (2019), 'Learning semantic sentence representations from visually grounded language without lexical knowledge', *Natural Language Engineering* **25**(4), 451–466. (Cited in page 108.)

Merkx, D., Frank, S. L. & Ernestus, M. (2019), Language Learning Using Speech to Image Retrieval, *in* 'Interspeech 2019, 20st Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019', pp. 1841–1845.
URL: *http://dx.doi.org/10.21437/Interspeech.2019-3067* (Cited in pages 52, 54, 55, 57, 59, 67, 68, 69, 70, 90, 100, 118, 121, 184, 185, 187, 198 et 209.)

Mersad, K. & Nazzi, T. (2012), 'When Mommy Comes to the Rescue of Statistics: Infants Combine Top-Down and Bottom-Up Cues to Segment Speech', *Language Learning and Development* **8**(3), 303–315.
URL: *https://doi.org/10.1080/15475441.2011.609106* (Cited in pages 20 et 172.)

Mills, A. E., ed. (1983), *Language acquisition in the blind child: normal and deficient*, Croom Helm ; College-Hill Press, London : San Diego. (Cited in page 28.)

Mintz, T. H., Walker, R. L., Welday, A. & Kidd, C. (2018), 'Infants' sensitivity to vowel harmony and its role in segmenting speech', *Cognition* **171**, 95–107.
URL: *https://doi.org/10.1016/j.cognition.2017.10.020* (Cited in pages 14 et 170.)

Mitchell, T. M. (1997), *Machine Learning*, 1 edn, McGraw-Hill, Inc., Usa. (Cited in pages 39, 40 et 179.)

Moore, C., Angelopoulos, M. & Bennett, P. (1999), 'Word learning in the context of referential and salience cues.', *Developmental Psychology* **35**(1), 60–68.
URL: *https://doi.org/10.1037/0012-1649.35.1.60* (Cited in pages 24 et 173.)

Morales, M., Mundy, P. & Rojas, J. (1998), 'Following the direction of gaze and language development in 6-month-olds', *Infant Behavior and Development* **21**(2), 373–377.
URL: *https://doi.org/10.1016/s0163-6383(98)90014-5* (Cited in page 23.)

Mortazavi, M. (2020), Speech-Image Semantic Alignment Does Not Depend on Any Prior Classification Tasks, *in* 'Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020', pp. 3515–3519.
URL: *http://dx.doi.org/10.21437/Interspeech.2020-3024* (Cited in page 68.)

Nazzi, T. (2008), 'Segmentation précoce de la parole continue en mots : évaluation inter-linguistique de l'hypothèse d'initialisation rythmique'.
URL: *https://www.persee.fr/doc/psy_0003-5033_2008_num_108_2_30973* (Cited in pages 10, 11 et 169.)

Nazzi, T., Bertoncini, J. & Bijeljac-Babic, R. (2009), 'A perceptual equivalent of the labial-coronal effect in the first year of life', *The Journal of the Acoustical Society of America* **126**(3), 1440–1446.
URL: *https://doi.org/10.1121/1.3158931* (Cited in page 13.)

Nazzi, T., Bertoncini, J. & Mehler, J. (1998), 'Language discrimination by newborns: Toward an understanding of the role of rhythm.', *Journal of Experimental Psychology: Human Perception and Performance* **24**(3), 756–766.
URL: *https://doi.org/10.1037/0096-1523.24.3.756* (Cited in pages 10 et 169.)

Nazzi, T., Iakimova, G., Bertoncini, J., Fredoni, S. & Alcantara, C. (2006), 'Early segmentation of fluent speech by infants acquiring French: Emerging evidence for crosslinguistic differences', *Journal of Memory and Language* **54**(3), 283–299.
URL: *https://doi.org/10.1016/j.jml.2005.10.004* (Cited in pages 11, 169 et 170.)

Nazzi, T., Jusczyk, P. W. & Johnson, E. K. (2000), 'Language Discrimination by English-Learning 5-Month-Olds: Effects of Rhythm and Familiarity', *Journal of Memory and Language* **43**(1), 1–19.
URL: *https://doi.org/10.1006/jmla.2000.2698* (Cited in page 10.)

Nazzi, T. & Ramus, F. (2003), 'Perception and acquisition of linguistic rhythm by infants', *Speech Communication* **41**(1), 233–243.
URL: *https://doi.org/10.1016/s0167-6393(02)00106-1* (Cited in page 10.)

Nelson, D. G. K., Hirsh-Pasek, K., Jusczyk, P. W. & Cassidy, K. W. (1989), 'How the prosodic cues in motherese might assist language learning', *Journal of Child Language* **16**(1), 55–68.
URL: *https://doi.org/10.1017/s030500090001343x* (Cited in page 19.)

Neubig, G., Nakata, Y. & Mori, S. (2011), Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, *in* 'Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)', Association for Computational Linguistics, pp. 529–533.
URL: *https://www.aclweb.org/anthology/P11-2093* (Cited in pages 66 et 189.)

Newport, E. L. & Aslin, R. N. (2004), 'Learning at a distance I. Statistical learning of non-adjacent dependencies', *Cognitive Psychology* **48**(2), 127–162.
URL: *https://doi.org/10.1016/s0010-0285(03)00128-2* (Cited in page 15.)

Norris, D. (1994), 'Shortlist: a connectionist model of continuous speech recognition', *Cognition* **52**(3), 189–234.
URL: *http://www.sciencedirect.com/science/article/pii/0010027794900434* (Cited in page 32.)

Norris, D., Mcqueen, J. M., Cutler, A. & Butterfield, S. (1997), 'The Possible-Word Constraint in the Segmentation of Continuous Speech', *Cognitive Psychology* **34**(3), 191–243.
URL: *https://doi.org/10.1006/cogp.1997.0671* (Cited in page 20.)

Ohishi, Y., Kimura, A., Kawanishi, T., Kashino, K., Harwath, D. & Glass, J. (2020), Trilingual Semantic Embeddings of Visually Grounded Speech with Self-Attention Mechanisms, *in* 'ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 4352–4356. (Cited in page 83.)

Otake, T., Hatano, G., Cutler, A. & Mehler, J. (1993), 'Mora or Syllable? Speech Segmentation in Japanese', *Journal of Memory and Language* **32**(2), 258–278.
URL: *https://doi.org/10.1006/jmla.1993.1014* (Cited in page 11.)

Park, A. & Glass, J. R. (2005), Towards unsupervised pattern discovery in speech, *in* 'IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.', pp. 53–58. (Cited in pages 37 et 178.)

Park, A. S. & Glass, J. R. (2008), 'Unsupervised Pattern Discovery in Speech', *IEEE Transactions on Audio, Speech, and Language Processing* **16**(1), 186–197. (Cited in pages 37 et 178.)

Peltzer-Karpf, A. (1994), *Spracherwerb bei hörenden, sehenden, hörgeschädigten, gehörlosen und blinden Kindern*, number 403 *in* 'Tübinger Beiträge zur Linguistik', G. Narr, Tübingen. (Cited in pages 27, 28 et 174.)

Pelucchi, B., Hay, J. F. & Saffran, J. R. (2009*a*), 'Learning in reverse: Eight-month-old infants track backward transitional probabilities', *Cognition* **113**(2), 244–247.
URL: *https://doi.org/10.1016/j.cognition.2009.07.011* (Cited in pages 17, 20 et 171.)

Pelucchi, B., Hay, J. F. & Saffran, J. R. (2009*b*), 'Statistical Learning in a Natural Language by 8-Month-Old Infants', *Child Development* **80**(3), 674–685.
URL: *https://doi.org/10.1111/j.1467-8624.2009.01290.x* (Cited in pages 16, 17, 20 et 171.)

Petrov, S., Das, D. & McDonald, R. (2012), A universal Part-of-Speech Tagset, *in* N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis, eds, 'Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)', European Language Resources Association (ELRA), Istanbul, Turkey. (Cited in pages 66 et 189.)

Picone, J. W. (1993), 'Signal modeling techniques in speech recognition', *Proceedings of the IEEE* **81**(9), 1215–1247. (Cited in page 66.)

Pinker, S. (2009), *Language Learnability and Language Development, With New Commentary by the Author*, Harvard University Press.
URL: *https://doi.org/10.2307/j.ctvjsf414* (Cited in page 1.)

Quine, W. V. O. (1964), *Word and object / Willard van Orman Quine,...*, The M.I.T. press paperback series, The Massachusetts Institute of Technology Press, Cambridge, Mass. (Cited in pages 23 et 172.)

Rabiner, L. & Juang, B.-H. (1993), *Fundamentals of Speech Recognition*, Prentice-Hall, Inc., Usa. (Cited in page 37.)

Räsänen, O. & Blandón, M. A. C. (2020), Unsupervised Discovery of Recurring Speech Patterns Using Probabilistic Adaptive Metrics, *in* 'Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020', pp. 4871–4875.
URL: *http://dx.doi.org/10.21437/Interspeech.2020-1738* (Cited in page 37.)

Räsänen, O., Doyle, G. & Frank, M. C. (2015), Unsupervised word discovery from speech using automatic segmentation into syllable-like units, *in* 'Interspeech 2015, 16st Annual Conference of the International Speech Communication Association, Dresden, Germany, 6-15 September 2015', pp. 3204–3208. (Cited in pages 37 et 38.)

Räsänen, O. & Khorrami, K. (2019), A computational Model of Early Language Acquisition from Audiovisual Experiences of Young Infants, *in* 'Interspeech 2019, 20st Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019', pp. 3594–3598.
URL: *https://doi.org/10.21437/interspeech.2019-1523* (Cited in pages 51, 56, 57 et 185.)

Rashtchian, C., Young, P., Hodosh, M. & Hockenmaier, J. (2010), Collecting Image Annotations Using Amazon's Mechanical Turk, *in* 'Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk', Association for Computational Linguistics, Los Angeles, pp. 139–147.
URL: *https://www.aclweb.org/anthology/W10-0721* (Cited in pages 55, 65, 185 et 188.)

Rescorla, L. A. (1980), 'Overextension in early language development', *Journal of Child Language* **7**(2), 321–335.
URL: *https://doi.org/10.1017/s0305000900002658* (Cited in pages 25 et 173.)

Rieder, W. G. (2003), Simulation and Modeling, *in* 'Encyclopedia of Physical Science and Technology', Elsevier, pp. 815–835.
URL: *https://doi.org/10.1016/b0-12-227410-5/00692-x* (Cited in pages 56, 57 et 186.)

Rosenblatt, F. F. (1958), 'The perceptron: a probabilistic model for information storage and organization in the brain.', *Psychological review* **65 6**, 386–408. (Cited in pages 40 et 46.)

Rouditchenko, A., Boggust, A., Harwath, D., Joshi, D., Thomas, S., Audhkhasi, K., Feris, R., Kingsbury, B., Picheny, M., Torralba, A. & Glass, J. (2020), 'AVLnet: Learning Audio-Visual Language Representations from Instructional Videos'. (Cited in page 86.)

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M. & Roy, D. (2015), 'Predicting the birth of a spoken word', *Proceedings of the National Academy of Sciences* **112**(41), 12663–12668.
URL: *https://doi.org/10.1073/pnas.1419773112* (Cited in page 22.)

Roy, D. (2003), 'Grounded spoken language acquisition: experiments in word learning', *IEEE Transactions on Multimedia* **5**(2), 197–209. (Cited in pages 48 et 182.)

Roy, D. (2005), 'Semiotic schemas: A framework for grounding language in action and perception', *Artificial Intelligence* **167**(1-2), 170–205.
URL: *https://doi.org/10.1016/j.artint.2005.04.007* (Cited in pages 39, 57, 58, 141, 179 et 186.)

Roy, D. K. & Pentland, A. P. (2002), 'Learning words from sights and sounds: a computational model', *Cognitive Science* **26**(1), 113–146. (Cited in pages 48, 49, 70 et 182.)

Rudd, L. C. & Johnson, L. E. (2011), Joint Focus of Attention, *in* 'Encyclopedia of Child Behavior and Development', Springer US, pp. 849–849.
URL: *https://doi.org/10.1007/978-0-387-79061-9_1559* (Cited in pages 23 et 173.)

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986), 'Learning representations by back-propagating errors', *Nature* **323**(6088), 533–536.
URL: *https://doi.org/10.1038/323533a0* (Cited in pages 47 et 182.)

Rumelhart, D. E., McClelland, J. L. & PDP Research Group, C., eds (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, MIT Press, Cambridge, MA, USA. (Cited in page 40.)

Saffran, J. R., Aslin, R. N. & Newport, E. L. (1996), 'Statistical Learning by 8-Month-Old Infants', *Science* **274**(5294), 1926–1928.
URL: *https://doi.org/10.1126/science.274.5294.1926* (Cited in pages 16, 20 et 171.)

Sakoe, H. & Chiba, S. (1978), 'Dynamic programming algorithm optimization for spoken word recognition', *IEEE Transactions on Acoustics, Speech, and Signal Processing* **26**(1), 43–49. (Cited in page 37.)

Schmid, H. (1997), Probabilistic Part-of-Speech Tagging Using Decision Trees, *in* D. Jones & H. Somers, eds, 'New Methods in Language Processing', Studies in Computational Linguistics, UCL Press, London, GB, pp. 154–164. (Cited in pages 66 et 189.)

Scholten, S., Merkx, D. & Scharenborg, O. (2020), 'Learning to Recognise Words using Visually Grounded Speech'. (Cited in pages 100 et 101.)

Shahidullah, S. & Hepper, P. G. (1994), 'Frequency discrimination by the fetus', **36**(1), 13–26. (Cited in page 9.)

Shain, C. & Elsner, M. (2020), Acquiring language from speech by learning to remember and predict, *in* 'Proceedings of the 24th Conference on Computational Natural Language Learning', Association for Computational Linguistics, Online, pp. 195–214.
URL: *https://www.aclweb.org/anthology/2020.conll-1.15* (Cited in pages 123 et 210.)

Shi, R., Cutler, A., Werker, J. & Cruickshank, M. (2006), 'Frequency and form as determinants of functor sensitivity in English-acquiring infants', *The Journal of the Acoustical Society of America* **119**(6), El61–el67.
URL: *https://doi.org/10.1121/1.2198947* (Cited in pages 18 et 171.)

Shi, R. & Lepage, M. (2008), 'The effect of functional morphemes on word segmentation in preverbal infants', *Developmental Science* **11**(3), 407–413.
URL: *https://doi.org/10.1111/j.1467-7687.2008.00685.x* (Cited in pages 18 et 171.)

Simonyan, K. & Zisserman, A. (2015), Very Deep Convolutional Networks for Large-Scale Image Recognition, *in* 'Proceedings of ICLR 2015', pp. 1–14. (Cited in pages 65 et 189.)

Singh, L., Reznick, J. S. & Xuehua, L. (2012), 'Infant word segmentation and childhood vocabulary development: a longitudinal analysis', *Developmental Science* **15**(4), 482–495.
URL: *https://doi.org/10.1111/j.1467-7687.2012.01141.x* (Cited in page 29.)

Skinner, B. (1957), *Verbal Behavior*, Century psychology series, Appleton-Century-Crofts. (Cited in page 1.)

Slobin, D. I. (1985), Introduction: Why Study Acquisition Crosslinguistically?, *in* D. I. Slobin, ed., 'The Crosslinguistic Study of Language Acquisition', Psychology Press.
URL: *https://doi.org/10.4324/9781315802541* (Cited in page 87.)

Slone, L. K., Abney, D. H., Borjon, J. I., hsin Chen, C., Franchak, J. M., Pearcy, D., Suarez-Rivera, C., Xu, T. L., Zhang, Y., Smith, L. B. & Yu, C. (2018), 'Gaze in Action: Head-mounted Eye Tracking of Children's Dynamic Visual Attention During Naturalistic Behavior', *Journal of Visualized Experiments* **1**(141).
URL: *https://doi.org/10.3791/58496* (Cited in page 55.)

Smith, L. N. (2017), Cyclical Learning Rates for Training Neural Networks, *in* '2017 IEEE Winter Conference on Applications of Computer Vision (WACV)', pp. 464–472. (Cited in pages 52 et 184.)

Smith, L. & Yu, C. (2008), 'Infants rapidly learn word-referent mappings via cross-situational statistics', *Cognition* **106**(3), 1558–1568.
URL: *https://doi.org/10.1016/j.cognition.2007.06.010* (Cited in pages 26 et 174.)

Soderstrom, M., Nelson, D. G. K. & Jusczyk, P. W. (2005), 'Six-month-olds recognize clauses embedded in different passages of fluent speech', *Infant Behavior and Development* **28**(1), 87–94.
URL: *https://doi.org/10.1016/j.infbeh.2004.07.001* (Cited in page 12.)

Taniguchi, T., Nagasaka, S. & Nakashima, R. (2016), 'Nonparametric Bayesian Double Articulation Analyzer for Direct Language Acquisition From Continuous Speech Signals', *IEEE Transactions on Cognitive and Developmental Systems* **8**(3), 171–185. (Cited in page 37.)

Thiessen, E. D. (2010), 'Effects of Visual Information on Adults' and Infants' Auditory Statistical Learning', *Cognitive Science* **34**(6), 1093–1106.
URL: *https://doi.org/10.1111/j.1551-6709.2010.01118.x* (Cited in page 26.)

Thiessen, E. D., Hill, E. A. & Saffran, J. R. (2005), 'Infant-Directed Speech Facilitates Word Segmentation', *Infancy* **7**(1), 53–71.
URL: *https://doi.org/10.1207/s15327078in0701_5* (Cited in pages 19 et 171.)

Thiessen, E. D. & Saffran, J. R. (2007), 'Learning to Learn: Infants' Acquisition of Stress-Based Strategies for Word Segmentation', *Language Learning and Development* **3**(1), 73–100.
URL: *https://doi.org/10.1080/15475440709337001* (Cited in page 11.)

Thiessen, E. & Erickson, L. (2009), Statistical learning, *in* E. L. Bavin & L. R. Naigles, eds, 'The Cambridge Handbook of Child Language', Cambridge University Press, pp. 37–60.
URL: *https://doi.org/10.1017/cbo9781316095829.003* (Cited in pages 20 et 172.)

Thiessen, E. & Erickson, L. (2015), Statistical learning, *in* E. L. Bavin & L. R. Naigles, eds, 'The Cambridge Handbook of Child Language', Cambridge University Press, pp. 37–60.
URL: *https://doi.org/10.1017/cbo9781316095829.003* (Cited in page 15.)

Tomasello, M. (2009), *Constructing a Language*, Harvard University Press. (Cited in pages 1, 23 et 173.)

Tomasello, M. & Todd, J. (1983), 'Joint attention and lexical acquisition style', *First Language* **4**(12), 197–211.
URL: *https://doi.org/10.1177/014272378300401202* (Cited in pages 23 et 173.)

Tsutsui, S., Chandrasekaran, A., Reza, M. A., Crandall, D. J. & Yu, C. (2020), A computational Model of Early Word Learning from the Infant's Point of View, *in* 'Proceedings of the 42th Annual Meeting of the Cognitive Science Society - Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020, virtual, July 29 - August 1, 2020', cognitivesciencesociety.org.
URL: *https://cognitivesciencesociety.org/cogsci20/papers/0180/index.html* (Cited in pages 55, 56, 123, 185 et 210.)

Vaissière, J. (1983), Language-Independent Prosodic Features, *in* A. Cutler & D. R. Ladd, eds, 'Prosody: Models and Measurements', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 53–66.
URL: *https://doi.org/10.1007/978-3-642-69103-4_5* (Cited in page 12.)

van Zon, M. (1997), Speech processing in Dutch: A cross-linguistic approach, PhD thesis, Tilburg University. (Cited in pages 100, 101 et 201.)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. & Polosukhin, I. (2017), Attention is All you Need, *in* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett, eds, 'Advances in Neural Information Processing Systems 30', Curran Associates, Inc., pp. 5998–6008.
URL: *http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf* (Cited in page 70.)

Vendrov, I., Kiros, R., Fidler, S. & Urtasun, R. (2016), Order-Embeddings of Images and Language, *in* Y. Bengio & Y. LeCun, eds, '4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings'.
URL: *http://arxiv.org/abs/1511.06361* (Cited in pages 64 et 66.)

Versteegh, M., Thiollière, R., Schatz, T., Cao Kam, X.-N., Anguera, X., Jansen, A. & Dupoux, E. (2015), The Zero Resource Speech Challenge 2015, *in* 'Proc. Interspeech 2015', pp. 3169–3173. (Cited in page 38.)

Vouloumanos, A. & Werker, J. F. (2007), 'Listening to language at birth: evidence for a bias for speech in neonates', *Developmental Science* **10**(2), 159–164.
URL: *https://doi.org/10.1111/j.1467-7687.2007.00549.x* (Cited in page 9.)

Vouloumanos, A. & Werker, J. F. (2009), 'Infants' learning of novel words in a stochastic environment.', *Developmental Psychology* **45**(6), 1611–1617.
URL: *https://doi.org/10.1037/a0016134* (Cited in page 26.)

Wang, X., Feng, S., Zhu, J., Hasegawa-Johnson, M. & Scharenborg, O. (2020), 'Show and Speak: Directly Synthesize Spoken Description of Images'. (Cited in pages 57, 122 et 210.)

Warren, R. M. (1970), 'Perceptual Restoration of Missing Speech Sounds', *Science* **167**(3917), 392–393.
URL: *https://doi.org/10.1126/science.167.3917.392* (Cited in page 32.)

Weber, A. & Scharenborg, O. (2012), 'Models of spoken-word recognition', *Wiley Interdisciplinary Reviews: Cognitive Science* **3**(3), 387–401.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/wcs.1178* (Cited in page 32.)

Weinberger, K. Q. & Saul, L. K. (2009), 'Distance Metric Learning for Large Margin Nearest Neighbor Classification', *J. Mach. Learn. Res.* **10**, 207–244. (Cited in pages 68 et 189.)

Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M. & Stager, C. L. (1998), 'Acquisition of word–object associations by 14-month-old infants.', *Developmental Psychology* **34**(6), 1289–1309.
URL: *https://doi.org/10.1037/0012-1649.34.6.1289* (Cited in pages 26 et 174.)

White, L., Melhorn, J. F. & Mattys, S. L. (2010), 'Segmentation by lexical subtraction in Hungarian speakers of second-language English', *The Quarterly Journal of Experimental Psychology* **63**(3), 544–554.
URL: *https://www.tandfonline.com/doi/abs/10.1080/17470210903006971*, (Cited in pages 17 et 171.)

Wiegreffe, S. & Pinter, Y. (2019), Attention is not not Explanation, *in* 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)', Association for Computational Linguistics, Hong Kong, China, pp. 11–20.
URL: *https://www.aclweb.org/anthology/D19-1002* (Cited in pages 74 et 192.)

Yoshikawa, Y., Shigeto, Y. & Takeuchi, A. (2017), STAIR captions: Constructing a Large-Scale Japanese Image Caption Dataset, *in* 'Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 2: Short Papers)', Association for Computational Linguistics, pp. 417–421.
URL: *https://www.aclweb.org/anthology/P17-2066* (Cited in pages 64, 72, 119, 188 et 208.)

Yu, C., Ballard, D. H. & Aslin, R. N. (2005), 'The Role of Embodied Intention in Early Lexical Acquisition', *Cognitive Science* **29**(6), 961–1005.
URL: *https://doi.org/10.1207/s15516709cog0000_40* (Cited in page 49.)

Zilly, J. G., Srivastava, R. K., Koutník, J. & Schmidhuber, J. (2017), Recurrent Highway Networks, *in* D. Precup & Y. W. Teh, eds, 'Proceedings of the 34th International Conference on Machine Learning', Vol. 70 of *Proceedings of Machine Learning Research*, Pmlr, International Convention Centre, Sydney, Australia, pp. 4189–4198.
URL: *http://proceedings.mlr.press/v70/zilly17a.html* (Cited in pages 52, 68, 184 et 189.)

Résumé Étendu en Français

1 Introduction

Dans un laps de temps relativement court, les enfants sont capables d'acquérir leur langue maternelle, ce qui leur permet de la comprendre et de la parler sans effort. Il s'agit d'un exploit étonnant, car ils sont capables de le faire sans beaucoup de supervision. Au contraire, le contexte perceptible environnant semble leur fournir toutes les informations nécessaires à l'acquisition de leur langue maternelle.

Landau & Gleitman (1985, p. 7) affirment que "l'*input* de l'enfant consiste en des paires son/situation, mais que son *output* final est un ensemble de paires forme/sens, appropriées à un ensemble infini de situations nouvelles mais bien circonscrites". La question est la suivante : comment les enfants passent-ils de paires son/situation à des paires forme/sens ?

L'une des étapes que les enfants doivent franchir est de comprendre que les sons qu'ils perçoivent constituent des signes conventionnels utilisés à des fins de communication, et non des sons aléatoires. Cette étape n'est peut-être pas la première que les enfants franchissent, mais elle est essentielle pour passer de *son* à *forme*. La transition des paires son/situation aux paires forme/sens implique également une étape de segmentation. La première étape de segmentation à laquelle nous pouvons penser est la segmentation du flux de parole en formes. En effet, le flux de parole contient des séquences de formes sont cependant connectées les unes aux autres. Les enfants doivent donc apprendre à segmenter le flux de parole de manière appropriée afin de découvrir les formes conventionnelles des mots utilisés dans leur langue maternelle. La deuxième étape de segmentation à laquelle on peut penser consiste à analyser l'environnement afin d'identifier et d'extraire les acteurs : qui parle, à qui, à propos de quoi, etc. Alors que la première étape consiste à identifier des schémas récurrents dans la modalité parlée, l'autre étape consiste à identifier des schémas récurrents dans d'autres modalités (visuelles, haptiques, etc.). Une fois que les formes ont été extraites du flux de la parole et que les référents ont été extraits de l'environnement, l'enfant doit apprendre à les mettre en correspondance. À ce stade, l'enfant est passé de paires *son/situation* à des paires *forme/référent*. L'enfant devient capable d'abstraire les points communs entre différents référents de la même forme et d'en déduire une signification et cela lui permet finalement à l'enfant de construire un ensemble de paires *forme/sens*. Les paires forme/sens que l'enfant construit doivent être stockées afin d'être facilement accessibles pour analyser le flux de parole et pour construire ses propres énoncés. Par conséquent, la forme des mots qui est stockée dans le lexique mental doit être suffisamment spécifique pour ne *pas* permettre la reconnaissance de formes de mots dont la prononciation serait similaire, sans pour autant être trop spécifique afin de tenir compte des variations et d'éventuelles erreurs de prononciation. Cette étape est communément appelée *reconnaissance des mots parlés*.

Dans cette thèse, nous nous proposons d'étudier un modèle neuronal de la parole visuellement contextualisée (*visually grounded speech models*). Les modèles de parole visuellement contextualisée sont des modèles qui sont entraînés pour résoudre une tâche de recherche d'image à partir d'une requête orale. C'est-à-dire qu'étant donné la description orale d'une image, ils doivent trouver l'image la plus proche de la description fournie parmi une collection d'images (ou vice-versa).Pour ce faire, ces réseaux doivent apprendre à transformer de manière appropriée l'image d'entrée et la description orale d'entrée afin qu'il soit facile de trouver l'une en fonction de l'autre. La tâche que ces réseaux doivent résoudre est donc très proche de celle d'un enfant qui acquiert sa langue maternelle. En effet, de tels

réseaux ont comme *input* des paires *son/situation* (c'est-à-dire une description parlée et son image correspondante), et afin de trouver l'image correspondante correcte, nous faisons l'hypothèse que le réseau doit apprendre à transformer cette paire *son/situation* en une paire *forme/sens*. En effet, pour trouver l'image correspondante parmi une collection d'images à partir d'un énoncé, le réseau doit d'une manière ou d'une autre segmenter le flux de parole en sous-unités, de sorte que les sous-unités résultantes se réfèrent à des objets dans l'image. La représentation des unités extraites doit être suffisamment spécifique pour que, lorsqu'une de ces unités est donnée en entrée au réseau, il ne récupère que les images présentant des instances auxquelles l'unité parlée fait référence. De même, comme pour les humains, du côté de l'image, le réseau doit être capable d'abstraire le référent, de sorte que l'image cible puisse être récupérée, même si l'objet principal de l'image est présenté de manière non canonique ou dans un environnement encombré. Par conséquent, la tâche pour laquelle les modèles neuronaux de parole visuellement contextualisée sont entraînés est très proche de celle des enfants qui apprennent leur langue maternelle, et plus précisément de la tâche d'acquisition lexicale. Ils doivent passer par les mêmes étapes qu'un enfant, que sont : la *segmentation*, le *appariement forme/référent*, et la *reconnaissance*.

Contributions

Dans cette thèse, nous étudions un modèle récurent de parole visuellement contextualisée (PVC). Comme de tels modèles résolvent une tâche similaire à celle des enfants, qui découvrent l'ensemble des paires forme/sens de leur langue maternelle, nos analyses visent à comprendre si les modèles neuronaux le font de la même manière. Dans ce manuscrit, nous présentons les contributions suivantes :

(i) Nous présentons une extension vocale d'un ensemble de données de description textuelle d'images (*image captioning*) qui permet d'entraîner de modèles de parole visuellement contextualisée. Contrairement à la plupart des jeux de données image/parole qui sont en anglais, ce jeu de données est en japonais, ce qui nous permet d'étudier, dans une approche contrastive, l'impact de la langue d'entrée (anglais ou japonais) sur les représentations apprises par le modèle.

(ii) Nous étudions les poids d'attention des mécanismes d'attention de nos modèles, afin de comprendre quelles parties du signal de parole sont mises en avant, et dans quelle mesure elles diffèrent de ce que le hasard prédirait. Nous proposons également une investigation longitudinale, où nous étudions comment les poids d'attention évoluent pendant la phase d'entraînement.

(iii) Nous introduisons une méthodologie issue de la littérature psycholinguistique, le paradigme du *gating* (Grosjean 1980), qui nous permet d'étudier facilement l'activation et la compétition des mots parlés dans notre modèle. A notre connaissance, c'est la première fois qu'une telle méthodologie est utilisée pour étudier les représentations apprises par les modèles neuronaux de traitement de la parole.

(iv) Nous proposons une méthode permettant d'introduire simplement des informations linguistiques préalables sous la forme de frontières de segments (frontières de phonèmes, de syllabes ou de mots) dans un modèle neuronal de traitement de la parole. La méthode que nous proposons permet d'intégrer plusieurs types de frontières, à différents niveaux de l'architecture neuronale, ce qui permet de prendre en compte la nature hiérarchique de l'entrée parlée. Cela nous permet d'étudier si la segmentation du signal vocal en sous-unités permet aux modèles d'apprendre une meilleure correspondance entre les images à leurs descriptions orales.

2 Informations générales sur l'acquisition du langage chez l'enfant

L'acquisition du langage désigne le processus par lequel un enfant apprend sa langue maternelle. Afin d'acquérir sa langue maternelle, une étape que l'enfant doit franchir est la construction d'un lexique mental (Emmorey & Fromkin 1988). Cependant, la construction de ce lexique mental est loin d'être facile. En effet, le flux de la parole ne contient pas de mots bien séparés auxquels l'enfant n'aurait qu'à attribuer un sens. L'enfant doit d'abord segmenter le flux de la parole en sous-unités et finalement associer un sens à chacune des sous-unités résultantes.

Ainsi, les étapes fondamentales du traitement et de la compréhension de la parole que les enfants doivent acquérir pour pouvoir correctement analyser le flux de la parole et construire un lexique mental — et finalement acquérir leur langue maternelle — sont identifiées par Di Cristo (2013, pp. 51-52) comme la *segmentation*, la *reconnaissance* et l'*interprétation* ; étapes que nous détaillons ci-après.

2.1 Segmentation

Les indices utilisés par les enfants pour segmenter le signal de parole en sous-unités peuvent être classés ainsi : les indices suprasegmentaux, les indices segmentaux, et les indices lexicaux. On peut également y ajouter d'autres types d'indices qui seront détaillés à la fin de cette présente section, qui ne rentrent nettement dans aucune des catégories pré-citées.

2.1.1 Indices suprasegmentaux

Les indices suprasegmentaux désignent les indices qui sont présents dans les caractéristiques suprasegmentales de la parole. Les caractéristiques suprasegmentales sont des modulations du signal vocal qui peuvent s'étendre sur plus d'un segment (c'est-à-dire un phone ou une syllabe). Ces caractéristiques suprasegmentales, regroupées sous le terme générique de prosodie, comprennent le rythme (tempo et pauses), l'intonation et l'accentuation. La prosodie assure de nombreux rôles fonctionnels, le principal qui peut être utile au nourrisson étant sa fonction de démarcation, qui regroupe les unités parlées et que les enfants peuvent utiliser pour inférer les frontières des mots.

Les enfants utilisent notamment des informations rythmiques pour segmenter le flux sonore en sous unités. Nazzi et al. (1998) ont montré que les enfants sont sensibles aux informations rythmiques portées par le signal de parole puisqu'ils sont en effet capables de distinguer deux langues appartenant à deux classes rythmiques différentes, mais sont incapables de distinguer deux langues différentes si celles-ci appartiennent à la même classe rythmique.

Les informations rythmiques étant (principalement) portée par les syllabes accentuées en anglais, Jusczyk, Cutler & Redanz (1993) ont exploré si les enfants segmentaient le flux sonore selon l'accentuation des syllabes ; la présence d'une frontière de mot étant fortement liée à la présence d'une syllabe accentuée (Cutler & Norris 1988). Jusczyk, Cutler & Redanz (1993) concluent que les enfants de 7,5 mois segmentent le signal sonore de sorte que les unités résultantes aient le motif accentué/non-accentué confirmant ainsi l'hypothèse des auteurs. Ce résultat a également été trouvé pour des enfants néerlandophones (Houston et al. 2000). Pour les enfants français, les travaux de Nazzi et al. (2006) et Nazzi (2008) (et de Bosch et al. 2013 pour le castillan et catalan) montrent que l'unité de segmentation adoptée par les enfants est la syllabe ; puisque ce sont des langues où l'unité de battement utilisée pour marquer le rythme est la syllabe et non l'accentuation. Ainsi, les sous unités dégagées

ne sont pas extraites selon l'accentuation des mots, mais uniquement selon la conformation à une unité syllabique correcte. Cependant, Nazzi et al. (2006) notent que les enfants francophones segmentent le flux sonore de manière systématique plus tardivement que les enfants anglophones ou néerlandophones (16 mois chez les enfants francophones contre 10 pour les enfants anglophones sur une tâche de segmentation de mots dissyllabiques).

Outre les indices suprasegmentaux déjà mentionnés, les enfants sont également sensibles à des indices suprasegmentaux affectant des unités plus larges, telles que celles des phrases phonologiques. Christophe et al. (2003) ont montré que la reconnaissance des mots chez les enfants de 13 mois est effectivement contrainte par les frontières des phrases phonologiques. Les enfants reconnaissent effectivement un mot cible familier (par exemple, *paper*) s'il fait partie d'une phrase phonologique ("*[The college][with the biggest* ***paper*** *forms][is best]*") mais pas s'il est à cheval sur deux phrases phonologiques ("*[The butler][with the highest* ***pay****][****per****forms the most]*"). Ainsi, en prêtant attention aux indices signalant les frontières de phrases prosodiques, les enfants peuvent déduire les frontières de mots, puisqu'une frontière de phrase prosodique correspond nécessairement à une frontière de mot. Johnson et al. (2014) montrent d'ailleurs que les enfants prêtent une attention particulière aux fins de phrases phonologiques et sont capables de reconnaitre un mot si celui-ci est présent au début ou à la fin d'une phrase phonologique.

2.1.2 Indices segmentaux

Les indices segmentaux sont des indices directement liés aux segments (c'est-à-dire phone[me]s) d'une langue. Comme Mattys et al. (2005), nous regroupons sous le terme d'indices segmentaux les indices phonotactiques et les indices acoustico-phonétiques.

Jusczyk, Friederici, Wessels, Svenkerud & Jusczyk (1993) ont montré que les enfants de 9 mois étaient sensibles aux indices phonotactiques, c'est-à-dire aux informations qui indiquent si une séquence de phonèmes constitue une séquence légale ou non dans son enchaînement. Ainsi les enfants de 9 mois préfèrent les pseudo-mots dont les séquences de phonèmes sont légales et non des pseudo-mots constitués de séquences illégales. Jusczyk et al. (1994) montrent que les enfants préfèrent les séquences légales les plus fréquentes face à celles qui sont légales, mais moins fréquentes. Mattys & Jusczyk (2001*b*) ont montré que les enfants se servent effectivement d'indices phonotactiques afin de segmenter la parole en sous unités et segmentent une séquence [...]C·CVC·C[...] en [...]C#CVC#C[...] si C·C constituent des séquences phonotactiques plus susceptibles d'apparaître en frontière de mot qu'à l'intérieur d'un mot. MacKenzie et al. (2012) ont montré que la connaissance d'informations phonotactiques contraint l'acquisition de mots nouveaux. Ainsi, chez des enfants anglophones de 12 mois, des (pseudo-)mots nouveaux ne peuvent être associés à des objets nouveaux que si ceux-ci comportent des séquences phonotactiques légales. L'harmonie vocalique, qui peut être considérée comme un cas spécial de contrainte phonotactique ; les segments concernés n'étant simplement pas adjacents ; est également un indice qui peut être utilisé par les enfants afin d'extraire des mots du signal sonore. Ketrez (2013) a en effet montré qu'en turc, lorsque l'harmonie vocalique est rompue, cela signale très probablement une limite de mot, et suggère que les enfants pourraient utiliser cette propriété pour inférer les limites des mots. Mintz et al. (2018) montrent ainsi dans une tâche d'apprentissage d'une langue artificielle que des enfants anglophones segmentent effectivement les mots en fonction de l'harmonie des voyelles et que cet indice constitue donc un indice viable pour extraire des unités lexicales.

Les enfants sont également sensibles à de fines variations acoustiques et se servent d'indice de coarticulation ainsi que d'indices d'ordre allophonique afin de déterminer la présence d'une frontière de mot. Christophe et al. (1994) ont par exemple montré que des

nourrissons de 3 jours font la différence entre un séquence dissyllabique CVCV extraite d'un mot (par ex. /mãta/ dans "*sédi**menta**tion*") et la même séquence CVCV à cheval sur une frontière (par exemple /mã#ta/ dans "*déguise**ment ta**lentueux*"). Ils sont donc capables de différencier deux réalisations différentes d'une même séquence phonologique à partir d'indices acoustiques fins (expliquant ainsi la sensibilité aux phrases phonologiques). Des résultats similaires ont été observés en anglais par Hohne & Jusczyk (1994) (différence "nitrate" [naɪt̪ʰ̥ɹeɪt˺] et "night rate" [naɪt˺ɹeɪt˺]). Plus récemment, Mattys & Jusczyk (2001*a*) ont montré que les enfants utilisent cette sensibilité aux variations allophoniques pour déterminer si une séquence constitue un mot ou non, et ainsi segmenter le flux sonore en sous-unités.

Finalement, les enfants utilisent également des indices statistiques afin segmenter le flux sonore en sous-unités. Saffran et al. (1996) ont notamment montré que les enfants utilisent les probabilités de transition entre syllabes afin de poser des frontières de mots, considérant les séquences des syllabes qui ont une forte probabilité de transition entre elles comme des unités, tandis que les séquences de syllabes ayant une faible probabilité de transition entre elles ne sont pas considérées comme formant une unité. Pelucchi et al. (2009*b*) ont confirmé ces résultats sur de l'italien. Les travaux de Pelucchi et al. (2009*a*) montrent que les enfants se servent aussi bien des probabilités de transitions progressives (*forward transition probability*) que de probabilités de transition régressives (*backward transition probability*)

2.1.3 Indices lexicaux

Dès que l'enfant a isolé un certain nombre de formes de mots, il peut les utiliser afin d'isoler les mots qui précèdent et/ou suivent le mot connu. White et al. (2010) ont inventé le terme "segmentation par soustraction lexicale" et le définit comme "l'utilisation de la connaissance lexicale pour imposer une structure de segmentation à l'entrée de la parole". Bortfeld et al. (2005) ont ainsi montré que les nourrissons de 6 mois utilisaient déjà des mots familiers ("mommy", "daddy", ou le propre nom de l'enfant) pour segmenter le mot suivant dans le flux de parole. Plus récemment, Bergelson & Swingley (2012) ont montré que les très jeunes enfants connaissent de nombreux mots courants (yaourt, banane, etc.), mots qui peuvent ainsi être utilisés pour aider la segmentation. Shi et al. (2006) et Shi & Lepage (2008) ont montré respectivement que chez des enfants anglophones et francophones, les déterminants sont utilisés par les enfants afin de segmenter le signal sonore en sous-unités. Un phénomène similaire a été mis en avant en japonais par Haryu & Kajikawa (2016), en montrant notamment que les enfants japonais se servent de la particule "ga" afin de déterminer des frontières de mots. Finalement, Johnson et al. (2014) rapportent que 80% des mots entendus en isolation par les enfants sont des interjections (oh, ah, etc.) et que ceux-ci pourraient également servir de support à la segmentation.

2.1.4 Autres indices

Les enfants peuvent également se servir d'autres indices afin de déterminer les frontières de mots. Notamment, le langage bébé (également appelé mamanais) permettrait aux enfants de mieux segmenter le flux sonore. En effet, le langage bébé est caractérisé par de nombreuses variations par rapport au langage adulte, notamment, une exagération de la variation de la fréquence fondamentale ou bien encore des pauses plus longues. Thiessen et al. (2005) concluent dans leur étude que le langage bébé permet aux enfants de 7,5 mois de mieux segmenter le flux sonore en mots, et Ma et al. (2011) ont montré que le langage bébé facilite l'acquisition (en termes de rétention) de mots nouveaux. Cependant, même si le langage bébé peut aider les enfants il n'est en rien nécessaire. Il existe en effet des

populations dans lesquels la parole est peu adressée aux enfants tant qu'ils ne sont pas capables de parler (voir Cristia et al. 2019), et qui arrivent malgré tout à apprendre leur langue maternelle.

Les indices que nous avons présentés l'ont été en isolation, comme si les enfants faisaient usage d'un seul indice de segmentation à la fois. Or, ce n'est pas le cas et les enfants utilisent de nombreux indices de manière concomitante cela permettant de "rendre une 'explosion combinatoire' moins probable" (Thiessen & Erickson 2009, p. 44). Johnson & Tyler (2010) montrent ainsi que l'utilisation de probabilités de transition ne semble être pleinement efficace que lorsqu'elles sont associées à d'autres indices. Cela est confirmé par les travaux de Mersad & Nazzi (2012) qui ont montrés dans leur expérience que les enfants utilisaient correctement les probabilités de transition entre syllabes seulement après être parvenus à extraire de la séquence présentée un mot connu (dans leur expérience, le mot *maman*). Si ce mot connu ne figurait pas dans la séquence, alors les enfants ne parvenaient pas à utiliser les probabilités de transitions.

2.1.5 Conclusion

Ainsi, les enfants utilisent de nombreux indices pour segmenter le flux de parole. Ceux-ci sont de nature variée, allant d'indice supra-segmentaux à des indices infra-segmentaux en passant par des indices segmentaux. Il semble donc que les enfants utilisent une combinaison d'indices descendants (par exemple, des mots isolés) et ascendants (par exemple, la phonotactique, les transitions de probabilité, etc.) afin de segmenter le flux de parole. Ces deux approches sont nécessaires — car, par exemple, la découverte des schémas phonotactiques réguliers dans les mots nécessite la connaissance d'un ensemble de mots individuels — et complémentaires — car en retour, les schémas phonotactiques déduits permettent à l'apprenant de segmenter encore plus de mots.

2.2 Appariement

Dans les sections précédentes nous avons considéré le langage *in vacuo* comme si les mots existaient pour et par eux-mêmes, déconnectés de tout contexte. Ce n'est bien sûr pas le cas :

> Le point important[...] est que le langage n'a d'existence que par rapport à l'ordre physique et mental des choses. Le langage n'est pas une sorte de système indépendant dans lequel on peut puiser à volonté. Les aspects de la langue n'ont de signification que dans la mesure où ils se rapportent à des aspects du monde. (Dixon 2012, p. 434)

Ainsi, les informations contextuelles jouent un rôle décisif dans l'acquisition lexicale : on peut notamment citer les informations contextuelles liées à des percepts (vision, toucher, odorat) qui vont venir guider l'enfant dans l'acquisition de sa langue maternelle.

2.2.1 Théorie de l'esprit et attention conjointe

Des chercheurs tels que Bloom affirme que "la covariation statistique entre le mot et le percept n'est ni nécessaire ni suffisante pour l'apprentissage des mots." (Bloom 2002, p. 59) et que la mise en correspondance des mots et des référents et, en fin de compte, l'acquisition du sens semblent exiger quelque chose de plus. En effet, il existe un nombre infini de possibilités parmi lesquelles l'enfant doit choisir (voir le célèbre exemple de Quine, *gavagai*, Quine 1964).

Pour Tomasello (2009), la lecture d'intention (*intention reading*) est l'ingrédient nécessaire à l'acquisition de la langue par les enfants, notamment lorsqu'il s'agit d'inférer le référent d'une forme verbale donnée. De la même manière, Bloom (2002) soutient que "les enfants utilisent leur *psychologie naïve* ou *théorie de l'esprit* pour comprendre à quoi les gens font référence lorsqu'ils utilisent des mots" et qu'il est nécessaire de disposer d'une théorie de l'esprit pour contextualiser de manière appropriée ce qui est dit. Le fait de posséder une théorie de l'esprit permet aux enfants d'entrer dans des moments d'attention conjointe. Un moment d'attention conjointe peut être défini comme "un épisode triadique d'interaction impliquant un interlocuteur, un nourrisson et un objet" (Rudd & Johnson 2011). La nécessité de partager des moments d'attention conjointe semble être confirmée par des travaux tels que ceux de Tomasello & Todd (1983) qui montrent que les enfants qui partagent de longues périodes d'attention partagée avec leurs interlocuteurs ont un vocabulaire perceptif globalement plus étendu que ceux pour lesquels les moments d'attention conjointe sont peu fréquents. Les travaux de Moore et al. (1999) semblent aller également dans ce sens en montrant que les enfants sont plus aptes à apprendre un nouveau mot si celui-ci a été vu dans un moment d'attention conjointe.

2.2.2 Assomption, biais et erreurs

Même si une théorie de l'esprit et les moments d'attention conjointe semblent décisifs pour acquérir une langue, il semble que l'acquisition du lexique soit guidée par de nombreux biais et assomptions que les enfants feraient sur le monde. On peut notamment citer les travaux de Markman (1990) qui postule l'existence de plusieurs biais que sont l'*hypothèse de l'objet entier*, l'*hypothèse taxonomique*, et l'*hypothèse d'exclusivité mutuelle*. Ainsi le premier guiderait les enfants afin que les étiquettes des mots se réfèrent aux objets dans leur globalité et non à des portions de ceux-ci. Le second stipule que les étiquettes de mot s'étendent aux objets similaires, c'est-à-dire aux objets ayant des caractéristiques similaires (par exemple, chien pour un berger allemand ou un bouledogue). Le dernier affirme qu'il existe une bijection entre une forme verbale et ses référents : pour une étiquette donnée, il n'existe qu'un seul référent (conceptuel) associé, et un référent ne peut être désigné que par une seule étiquette canonique. (Clark 1987) mentionne deux principes complémentaires : le *principe de contraste* et le *principe de conventionnalité*. Le principe de contraste dit que "chaque paire de mots entre en contraste par leur sens". Autrement dit, l'enfant considère que deux formes verbales ne sont pas des synonymes exacts. Ce principe encouragerait l'enfant à explorer son environnement afin de trouver un référent possible afin de ne pas assigner deux étiquettes différentes à un même mot. Le principe de conventionnalité stipule que "pour certaines significations, il existe une forme conventionnelle que les locuteurs s'attendent à voir utilisée dans la communauté linguistique".

Il arrive parfois que les enfants fassent des erreurs d'appariement et attribuent la mauvaise étiquette à un référent donné. Rescorla (1980) fournit une typologie des principales causes de surextensions : *sur-inclusions catégorielles* où une étiquette est utilisée pour un référent proche du référent réel (*bébé* pour les enfants), *sur-extensions analogiques* où une étiquette est utilisée pour un référent qui présente une similarité avec le référent réel (*tex*-titballe pour les billes ou les pommes), et *énoncés prédicats* qui sont holophrastiques (par exemple *chien* en montrant un panier désigner un chien absent). Rescorla (1980) estime qu'un tiers des premiers mots de l'enfant sont des surextensions, et qu'environ trois quarts des extensions sont soit des surinclusions catégorielles, soit des surextensions analogiques, la plupart d'entre elles étant faites sur la base de la "similarité perceptuelle" (par opposition à la similarité fonctionnelle). Cependant, bien souvent, la similarité perceptuelle évoquée dans les études se limite à une similarité visuelle.

2.2.3 Modalité visuelle

La vision joue un rôle central dans l'acquisition du langage. En effet, "c'est en grande partie la vision qui dirige l'attention du nourrisson sur les personnes, les objets et les événements ; la vision est importante pour la coordination de l'attention entre le parent et l'enfant ; et elle contraint souvent (surtout dans la culture anglo-américaine) le contenu du langage adressé à l'enfant" (Andersen et al. 1993).

Il semble que la capacité des nourrissons à associer systématiquement un référent visuel à mot n'apparaisse que vers l'âge de 12 mois. Werker et al. (1998) montrent par exemple qu'à 8 et 12 mois les enfants en sont incapables, cependant ils le sont à 13 ou 14 mois. Smith & Yu (2008) quant à eux trouvent que des enfants de 12 mois en sont capables. Ce dernier résultat semble donc être en contradiction avec Werker et al. (1998). Cependant, comme cette capacité semble se développer dans un laps de temps très court (entre le 12e et 13e mois), il est raisonnable d'observer de légères différences d'une expérience à l'autre, surtout si les paramètres expérimentaux sont différents.

Pour mesurer l'important de l'input visuel sur l'acquisition du langage, nous pouvons nous intéresser à l'acquisition du langage chez les enfants aveugles de naissance. La cécité, même si elle n'empêche une maitrise parfaite de la langue maternelle à l'âge adulte (Landau & Gleitman 1985), peut entraîner des retards dans la prime enfance. Notamment, on constate chez les enfants aveugles une tendance à parler d'événements passés plutôt que de "l'ici et maintenant" . Ainsi, en se concentrant sur des événements passés, l'enfant essaie de recréer une situation dans laquelle il sait que la personne qui s'occupe de lui et lui-même ont partagé une expérience commune dont ils peuvent parler et essaye ainsi de créer une situation d'attention partagée, à défaut de pouvoir en créer ou percevoir une dans "l'ici et le maintenant" (Andersen et al. 1993). Pour les mêmes raisons, les enfants aveugles souffrent également d'un manque de décentration et ont une forte tendance à initier des conversations centrées sur eux-mêmes ou leur environnement immédiat (Andersen et al. 1984, Peltzer-Karpf 1994). On constate également chez les enfants aveugles une capacité de généralisation moindre. En effet, le fait d'être privé de la vue les prive de l'accès simultané à de nombreuses caractéristiques (forme, couleur, texture, etc.) qui constituent de précieuses informations pour généraliser. De plus, pour l'enfant aveugle, explorer le monde requiert un effort conscient d'exploration de son environnement, ce qu'il ne peut faire que lorsqu'il est en mesure de se déplacer, alors que pour l'enfant voyant, percevoir le monde (au moins visuellement) ne nécessite pas d'efforts conscients (Dunlea 1989, p.10). De même que pour la généralisation, on constate une capacité de conceptualisation moindre chez l'enfant aveugle. Selon Dunlea (1989), ils ne parviennent pas à saisir la nature symbolique du langage et ne parviennent pas à appliquer de manière appropriée les mots ou les morceaux de mots qu'ils apprennent à de nouvelles situations et d'après Andersen et al. (1984) ont tendance à répéter mot pour mot ce qu'ils entendent sans analyse supplémentaire (c'est-à-dire sans segmentation) et tendent à répéter des routines pré-faites.

Ainsi, la vision semble fournir des informations essentielles au jeune enfant afin de lui permettre d'acquérir sa langue maternelle. Plus particulièrement, la vision semble aider l'enfant à conceptualiser le monde et à établir des généralisations.

2.3 Reconnaissance

La reconnaissance des mots parlés peut être définie comme le processus d'accès aux éléments lexicaux du lexique mental à partir des motifs phonologiques perçus dans le signal de parole (Magnuson et al. 2013). En d'autres termes, cela implique de mettre en correspondance ce qui est entendu avec les éléments lexicaux stockés dans le lexique mental. Nous passons en

revue dans cette section plusieurs modèles de reconnaissance des mots.

2.3.1 Modèle de la Cohorte

Le modèle de la COHORT (Marslen-Wilson & Welsh 1978) est l'un des tout premier modèle tentant de rendre compte de la manière dont les humains reconnaissent et extraient les mots du flux de la parole. Selon ce modèle, la reconnaissance des mots parlés se déroule en trois étapes simultanées : *accès*, *sélection*, et *intégration*.

L'*accès* consiste à activer un ensemble de mot-formes dans le lexique mental qui correspondent à l'entrée acoustique. Chaque mot-forme est associée à une unité d'activation, ce qui permet l'activation de plusieurs formes de mots à la fois. Ces unités ne sont activées que si l'entrée acoustique perçue correspond exactement à la forme phonologique internalisée. Tous les mots qui commencent par la séquence sonore perçue sont activés, et forment ainsi une cohorte. Lorsqu'il y a une incompatibilité entre la forme internalisée d'un mot et l'*input* perçu, alors ce mot est retiré de la cohorte. Ce processus d'élagage des mots de la cohorte constitue l'étape de *sélection*. Finalement l'*intégration* consiste à vérifier que les propriétés sémantiques et syntaxiques des mots activés concordent avec le reste de la phrase, sinon, le ou les mots incompatibles sont retirés de la cohorte. Ce processus continue jusqu'à ce qu'il ne reste plus qu'un mot dans la cohorte.

On peut citer plusieurs problèmes avec ce modèle de cohorte, notamment le fait qu'il suppose que les stimuli acoustiques perçus correspondent exactement aux formes phonologiques internalisées. Or ceci n'est pas réaliste puisque les humains peuvent reconnaitre un mot même mal prononcé. De même il est possible de reconnaitre un mot bien prononcé mais sémantiquement incorrect dans le contexte phrastique, bousculant ainsi l'hypothèse de l'étape d'intégration.

2.3.2 Modèle TRACE

Le modèle TRACE s'appuie sur le modèle COHORT en conservant ses "caractéristiques positives majeures" (McClelland & Elman 1986*a*), que sont l'activation simultanée, la sélection, et l'élagage interactif, tout en essayant de surmonter les problèmes qui ont été abordés ci-dessus.

L'*input* acoustique, contrairement au modèle de la COHORT, n'est plus perçu en termes de phonèmes, mais en termes de stimuli acoustiques. Le modèle TRACE est organisé en couches, la première traitant les stimuli acoustiques, la deuxième les phonèmes et la troisième les mots. La couche acoustique active certaines unités représentant des phonèmes dans la couche des phonèmes, qui eux-mêmes activent des unités représentant des mots dans la couche lexicale. L'activation dans ce modèle dépend de l'adéquation entre les percepts acoustiques avec les représentations internalisés, mais ne nécessite pas une adéquation parfaite ; autorisant ainsi les erreurs de prononciation et/ou de perception. Les phonèmes activent les mots dans lesquels ils apparaissent, peu importe leur position. Les unités à l'intérieur d'une même couche (phonème, ou mot) sont reliées par des connexions inhibitrices de sorte que les unités les plus activées inhibent dans une certaine mesure l'activation des autres si elles ne correspondent pas à l'entrée acoustique aussi bien que les unités les autres unités plus activées.

Contrairement au modèle COHORT, ce modèle est capable de gérer des chaînes de mots connectés. Dans ce modèle, le flux vocal est implicitement segmenté lorsque les mots sont reconnus. Ce modèle apporte une amélioration majeure par rapport au modèle COHORT : les mots peuvent être reconnus après leur fin, et les mots peuvent être activés à partir de n'importe quel point, relâchant ainsi la contrainte de correspondance exacte de l'apparition

des mots que suppose le modèle de la COHORT. La plus grande amélioration apportée par le modèle TRACE est la compétition active entre les mots, car les mots sont réellement en compétition les uns avec les autres grâce aux connexions inhibitrices entre les unités d'une même couche.

2.3.3 Shortlist

Le modèle TRACE dans son implémentation est computationnellement irréaliste. En effet, il suppose qu'un phonème donné active l'*ensemble* des mots dans lequel il apparait. Cela est possible pour un vocabulaire limité, mais pas pour un vocabulaire d'une taille plus élevée.

Le modèle Shortlist peut être vu comme un modèle hybride entre le modèle de la COHORT et TRACE. Certes, un phonème active, comme dans TRACE, l'ensemble des mots dans lequel il apparait, mais les mots pour lesquels l'activation est trop faible ne sont plus considérés comme des cibles potentielles. Ainsi, le nombre de mots activés de manière simultanée est limité à un petit nombre (*a shortlist*) et la compétition inter-mots limitée à ceux de la *shortlist*. Par conséquent, ce modèle est plus léger en termes de calculs que le modèle TRACE— car seul un sous-ensemble du lexique est considéré comme un candidat valide — tout en préservant la plupart de ses caractéristiques originales.

2.3.4 Modèle de la Cohorte distribuée

Le modèle de la cohorte distribuée (MCD, Gaskell & Marslen-Wilson 1997) constitue un changement de paradigme. Alors que les modèles précédents incorporaient explicitement plusieurs niveaux de calcul (couche de caractéristiques, couche phonétique, couche lexicale), le MCD ne le fait pas. En effet, le MCD utilise un simple réseau neuronal récurrent (Elman 1990).

Le réseau reçoit en entrée une séquence de mesures phonétiques et est entraîné à prédire un vecteur sémantique et un vecteur phonologique. Dans ce modèle, le concept d'activation est différent des autres modèles. En effet, alors que dans les autres modèles, une valeur est attribuée à chaque mot du lexique — cette valeur représentant la force de l'activation — il n'est pas possible ici d'avoir une telle valeur. Au lieu de cela, l'activation est représentée par la proximité du vecteur sémantique prédit avec celui des mots du lexique. L'activation simultanée de plusieurs mots est toujours possible, mais elle est cependant implicite. En effet, le vecteur sémantique prédit peut être considéré comme un "mélange" des représentations pertinentes. Activer un mot plus qu'un autre peut être considéré comme "modifier ce mélange" et "peut être interprété en termes de mouvement dans l'espace sémantique".

Ainsi, dans ce modèle, la notion de caractéristique acoustique, de phonème et d'unité lexicale est implicite puisque toutes les couches sont capables de représenter cette information simultanément. Cependant, dans leur expérience, les auteurs ne testent que la reconnaissance de mots individuels et non de phrases complètes. Ainsi, même si ce modèle obtient des résultats intéressants (en montrant par exemple que les mots les plus fréquents sont ceux qui sont activés prioritairement), il ne réalise aucune segmentation de l'entrée parlée.

2.3.5 Conclusion sur la reconnaissance de mots parlés

Tous les modèles de reconnaissance des mots supposent que la reconnaissance des mots se déroule en plusieurs étapes : une activation, une compétition et la reconnaissance finale. Néanmoins, la manière dont ces trois étapes sont réalisées varie d'un modèle à l'autre. Cependant, tous ces modèles s'accordent sur le fait que la reconnaissance de mots implique nécessairement l'activation simultanée d'un ensemble de mots candidats. Cet ensemble de

mots est itérativement élagué si l'entrée acoustique est trop éloignée de la représentation internalisée, de sorte qu'il ne reste finalement qu'un seul mot.

2.4 Conclusion

Dans ce chapitre, nous avons examiné les stratégies utilisées par les enfants pour segmenter l'entrée parlée en sous-unités. Nous avons montré que les enfants utilisent un large éventail de stratégies pour trouver des mots en se servant d'indices suprasegmentaux, segmentaux et infrasegmentaux. Nous avons ensuite présenté plusieurs modèles de reconnaissance de mots qui tentent de rendre compte de la manière dont les humains sont capables d'activer et de récupérer les unités lexicales du lexique mental. Tous les modèles de reconnaissance de mots s'accordent sur le fait qu'une stratégie de reconnaissance réussie consiste à activer simultanément plusieurs unités lexicales à la fois. Enfin, nous avons montré que l'acquisition du langage est un phénomène multimodal qui implique toutes les capacités perceptives de l'enfant, et notamment que la vision fournissait à l'enfant des informations précieuses pour acquérir sa langue maternelle.

3 Informations générales sur le traitement automatique de la parole et la découverte d'unités lexicales

3.1 Traitement non supervisé de la parole & notion de *grounding*

Le travail que nous menons dans cette thèse s'inscrit dans le cadre du traitement non supervisé de la parole, et la principale motivation de celle-ci est de concevoir des modèles de traitement automatique de la parole plus proches du traitement humain de la parole, qui ne requiert pas ou peu de supervision. Par exemple, un enfant qui apprend sa langue maternelle n'a pas besoin d'étiquettes textuelles, mais utilise plutôt des signaux de supervision faibles, tels que des indices visuels. Ces signaux constituent des signaux de supervision faibles, car ils limitent le processus d'apprentissage en l'ancrant dans le contexte.

Le travail effectué dans cette thèse appartient donc aux "approches sensorielles" des modèles de traitement de la parole, telles que définies par Glass (2012). Les modèles "sensoriels" de traitement de la parole sont des modèles qui n'ont besoin que de parole couplée à des données sensorielles pour fonctionner. En d'autres termes, ces modèles ne nécessitent pas de données annotées ni d'expertise humaine pour fonctionner et "se rapprochent de l'acquisition du langage oral humain" selon Glass (2012).

L'une des premières approches de traitement non supervisé de la parole est celle introduite par Park & Glass (2005). La méthodologie qu'ils proposent permet de découvrir des unités lexicales à partir d'un signal de parole, sans avoir besoin que celui-ci soit préalablement transcrit (Park & Glass 2008). La méthodologie proposée de Segmental-DTW est basée sur la mesure du DTW (*Dynamic Time Warping*) qui permet de mesurer la similarité de deux séquences. Les autres modèles de traitement non supervisé de la parole, qui ont pour but d'extraire des unités lexicales ou de segmenter l'entrée parlée se base principalement sur des modèles bayésiens. Lee et al. (2015) proposent par exemple de combiner la découverte non supervisée d'unités acoustiques avec des modèles de segmentation bayésien tels que celui de (Goldwater 2006). Plus récemment, on peut citer le travail de Kamper (2017) qui propose un modèle bayésien permettant de conjointement inférer des frontières dans le signal de parole tout en clusterisant les unités qui résultent de la segmentation menée.

Un fait surprenant des travaux de découverte non supervisée de termes et de lexique évoquée ci-dessus est que le repérage et la segmentation des mots sont effectués sur la base de la forme uniquement, sans qu'aucun autre indice contextuel ne soit utilisé. Le langage n'est pas un système indépendant, et ce qui lui donne sa substance est qu'il est lié au monde physique. Nous avons utilisé l'exemple des enfants aveugles pour montrer que, lorsqu'ils sont privés d'un type d'entrée sensorielle, à savoir la vision, l'acquisition du langage est considérablement affectée. Pour établir une analogie, les modèles de découverte et de segmentation des termes sont en fait entraînés comme des nourrissons privés de sens, qui ne se fondent que sur la forme. Il semble ainsi qu'une stratégie de segmentation de la parole adéquate doive inclure une composante *sémantique*. Dans leurs travaux, Bender & Koller (2020) soutiennent que :

> si la forme est complétée par des données contextuelles de quelque nature que ce soit, il est concevable que le sens puisse être appris dans la mesure où l'intention communicative est représentée dans ces données.

Par conséquent, l'ancrage des données linguistiques à des connaissances externes semble être une étape nécessaire pour accéder au sens. Pourtant, que signifie plus précisément le terme *grounding* ? Si l'ancrage contextuel consiste simplement à ajouter des connaissances

externes, les modèles de reconnaissance vocale peuvent être considérés comme des modèles textuellement ancrés, puisque le texte, qui est une connaissance externe, est ajouté au modèle. Roy (2005) définit l'ancrage comme suit :

> La relation entre les mots et le monde physique, et par conséquent notre capacité à utiliser des mots pour faire référence à des entités dans le monde, fournit les fondements de la communication linguistique. Les approches actuelles de la conception des systèmes de traitement du langage passent à côté de cette connexion critique, qui est obtenue par un processus que j'appelle *grounding*. [...] Le *grounding* linguistique fait référence aux processus spécialisés dans la mise en relation des mots et des actes de parole avec l'environnement de l'utilisateur de la langue. (Roy 2005)

Le *grounding* implique donc d'ajouter aux données linguistiques des données externes *non-linguistiques* qui reflètent d'une certaine manière le monde physique. Les données externes doivent également refléter dans une certaine mesure l'intention communicative des données linguistiques, comme le mentionne Bender & Koller (2020). Cette dernière contrainte est également visible dans la définition de Roy (2005), puisque les "actes de parole" résultent nécessairement d'une intention communicative. Par conséquent, les modèles *grounded* désignent une classe de modèles informatiques qui traitent une certaine forme linguistique — du texte, ou bien de la parole — en conjonction avec une autre source d'information provenant du monde physique. Le fait que deux modalités soient traitées en conjonction est nécessaire, mais pas suffisant : les deux modalités doivent se produire simultanément dans le monde physique pour que les deux soient liées. Ainsi, les modèles *grounded* ne traitent pas uniquement la forme, mais sont capables — ou du moins ont la capacité — de lier cette forme à des référents dans le monde physique, ou à une représentation de celui-ci.

3.2 Apprentissage automatique

L'apprentissage machine (ML) est un sous-domaine de l'informatique qui s'intéresse à la création et à "l'étude d'algorithmes informatiques qui s'améliorent automatiquement grâce à l'expérience" (Mitchell 1997, p. xv). Au lieu de concevoir un algorithme spécifique pour résoudre une tâche spécifique, le ML cherche à concevoir des algorithmes qui *apprennent* comment résoudre une tâche spécifique :

> On dit d'un programme informatique qu'il apprend de l'expérience E sur une certaine classe de tâches T et selon une mesure de performance P, si sa performance à la tâche T, telle que mesurée par P, s'améliore avec l'expérience E. (Mitchell 1997, p. 3)

L'expérience E consiste en un ensemble d'exemples — appelé ensemble de données — à partir duquel le programme va apprendre. L'apprentissage se fait par un processus d'essais et d'erreurs au cours duquel le programme tente de résoudre la tâche T à l'aide des données disponibles E et modifie ses paramètres de manière à augmenter sa mesure de performance P.

L'ensemble de données utilisé pour l'apprentissage est généralement divisé en trois ensembles inégaux : l'ensemble d'apprentissage qui est le plus important ($\approx 80\%$), l'ensemble de développement (ou de validation) ($\approx 10\%$), et l'ensemble de test ($\approx 10\%$). L'ensemble d'entraînement, comme son nom l'indique, est utilisé pour entraîner le modèle. L'ensemble de développement est utilisé pour tester le modèle afin de sélectionner le meilleur point de

contrôle du modèle, tandis que l'ensemble de test n'est utilisé qu'une fois le meilleur modèle sélectionné pour rapporter le score final.

3.3 Réseaux neuronaux artificiels

Les réseaux de neurones artificiels (ANN) sont un type d'algorithmes d'apprentissage qui appartiennent à l'approche connexionniste de l'intelligence artificielle (IA). L'idée centrale du connexionnisme est qu'un comportement complexe peut être approximé par un calcul complexe, qui peut à son tour être décomposé en calculs plus petits et plus simples réalisés par des unités de traitement individuelles. L'interconnexion de ces unités (et donc du calcul) donne naissance au comportement global.

Ces unités de traitement, appelées à l'origine *perceptrons*, sont maintenant communément appelées *neurones*. Un neurone reçoit n entrées $X = x_0, x_1, ..., x_n$ à partir desquelles une sortie $\hat{y}$ est calculée. Étant donné les caractéristiques d'entrée, le neurone est entraîné à prédire une sortie $\hat{y}$ qui doit être aussi proche que possible de la sortie souhaitée y. Pour ce faire, le vecteur d'entrée est pondéré par un ensemble de poids $W = w_0, w_1, ..., w_n$ qui sont des paramètres entraînables, et la sortie est ensuite transformée par une fonction d'activation non linéaire ϕ. Un neurone de biais peut également être ajouté. Le calcul final d'un seul neurone peut donc être résumé comme suit :

$$\hat{y} = \phi(W \cdot x + b) = \phi(\sum_{i=1}^{n} w_i x_i + b) \tag{D.1}$$

Les problèmes du monde réel sont trop complexes pour être résolus par un seul neurone. Ainsi, un nombre arbitraire de neurones peut être utilisé en conjonction afin d'apprendre des fonctions de transformation complexes. Ces neurones sont regroupés en *couches* et forment ce que l'on appelle communément un *perceptron multicouche* (*MultiLayer Perceptron*, MLP). La première couche est appelée la *couche d'entrée*, la dernière couche *couche de sortie*, et la ou les couches intermédiaires sont appelées les *couches cachées*.

3.3.1 Réseaux de neurones récurrents et unités récurrente à portes

Les réseaux neuronaux récurrents (RNN) sont des réseaux neuronaux capables de traiter des données séquentielles (Elman 1990). Les RNN peuvent être considérés comme des réseaux *feed-forward* complétés par des connexions de rétroaction. Les RNN peuvent être formalisés comme suit :

$$h_t = \phi(W x_t + U h_{t-1} + b) \tag{D.2}$$

où h_t est l'état caché au pas de temps t, x_t est l'entrée actuelle, h_{t-1} est l'état caché précédent, b est un terme de biais, ϕ est une fonction d'activation non linéaire (généralement sigmoïde ou tangente hyperbolique), et où W, U et b sont des paramètres entraînables. Notez que la seule différence avec une couche *feed-forward* est simplement le terme supplémentaire Uh_{t-1} : la sortie à un pas de temps t ne dépend pas seulement de l'entrée courante mais aussi de la sortie au pas de temps précédent $t-1$. Ainsi, un tel calcul permet de modéliser la dépendance temporelle qui existe entre des vecteurs consécutifs. La sortie d'un RNN consiste en une séquence T de vecteurs. Le vecteur final de la séquence (au pas de temps T) peut être considéré comme la représentation compacte de toute la séquence d'entrée, car le vecteur final dépend du calcul de tous les vecteurs précédents de la séquence. Lorsque le premier élément de la séquence x_1 est traité, h_{t-1} n'existe pas. Dans ce cas, cet état caché précédent, noté h_0, est défini comme un vecteur de 0, bien que dans certains cas, l'état initial puisse également être un paramètre entraînable du réseau.

Même si les RNN sont capables de modéliser des données séquentielles et de suivre les dépendances locales relativement bien, ils sont incapables de capturer les dépendances à long terme dans la séquence d'entrée. En effet, à mesure que la taille de la séquence T s'allonge, le vecteur h_t contient de moins en moins d'informations sur le début de la séquence. L'information sur le passé s'estompe à mesure que le vecteur caché précédent h_{t-1} est combiné à chaque pas de temps avec de nouvelles informations. Les unités à porte ont été introduites pour résoudre ce problème.

Cho et al. (2014) a introduit les Gated Recurrent Units (GRU) qui sont des cellules récurrentes à ports (gated) qui sont également capables de garder la trace des dépendances à long terme et à court terme. Un GRU est formellement défini comme suit :

$$\begin{aligned} z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\ r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\ \hat{h}_t &= \tanh(W_h x_t + U_h(r_t * h_{t-1}) + b_h) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \hat{h}_t \end{aligned} \tag{D.3}$$

où $W_z, U_z, W_r, U_r, W_h, U_h$ sont des matrices entraînables, b_z, b_r, b_h sont des termes de biais, et σ représente la fonction d'activation sigmoïde. Un GRU possède deux portes : une *porte de mise à jour* z, et une *porte de réinitialisation* r qui calculent toutes deux indépendamment une valeur scalaire à chaque pas de temps t (z_t et $r_t \in [0, 1]$). Compte tenu de l'entrée courante x_t, la valeur de z_t représente un rapport entre la quantité d'informations à reporter depuis le pas de temps précédent $[(1 - z_t) * h_{t-1}]$ et la quantité de nouvelles informations à intégrer $[z_t * \hat{h}_t]$. Cette porte permet de contrôler les dépendances à long terme. r_t représente le degré de dépendance du nouvel état par rapport au pas de temps précédent $[r_t * h_{t-1}]$, ce qui permet de contrôler les dépendances locales.

3.4 Mécanisme d'attention

Même en utilisant des cellules récurrentes bidirectionnelles, l'encodage de la signification d'une séquence entière dans un seul vecteur reste une tâche difficile. Les mécanismes d'attention (Bahdanau et al. 2015) ont été introduits pour résoudre ce problème. L'intuition de tels mécanismes d'attention est simple : au lieu de conserver le dernier vecteur de la séquence — qui devrait encoder le sens de la phrase entière — tous les vecteurs qui ont été calculés à chaque pas de temps sont conservés et un poids est attribué à chacun d'entre eux de façon à donner plus d'importance à certains vecteurs qu'à d'autres.

Les poids sont calculés par le réseau lui-même et appris au moment de l'entraînement. Ainsi, le réseau apprend quel(s) vecteur(s) de la séquence d'entrée doit(vent) se voir accorder plus d'importance dans la représentation finale, et ce de manière non supervisée. Un mécanisme d'attention peut être formalisé comme suit :

$$c = \sum_{t=1}^{T} \alpha_t h_t \quad \text{avec} \quad \alpha_t = \frac{\exp(\text{score}(h_t))}{\sum_{t'=1}^{T} \exp(\text{score}(h_{t'}))} \tag{D.4}$$

où c, le vecteur de contexte, est la somme pondérée des vecteurs cachés, et α_t est le poids d'attention pour le pas de temps t, où $\alpha_t \in [0, 1]$ et $\sum_{t=1}^{T} \alpha_t = 1$. La formalisation que nous présentons est très générale : la fonction de notation *score* (généralement un MLP) calcule une valeur scalaire pour chaque h_t. Cependant, la manière précise dont cette valeur scalaire est calculée dépend de l'implémentation particulière du mécanisme d'attention et

peut ne pas être uniquement une fonction de h_t. Avec un tel mécanisme d'attention, les vecteurs pour lesquels le poids de l'attention α est élevé seront fortement représentés dans le vecteur final c, tandis que les vecteurs auxquels on a attribué une faible valeur scalaire α seront peu ou pas représentés dans le vecteur final c.

3.4.1 Fonction de coût et rétropropagation

Le but du réseau est de prédire, à partir d'une entrée x, une sortie $\hat{y}$ qui soit aussi proche que possible de la sortie désirée y. Le succès de l'opération est mesuré à l'aide d'une fonction de coût. Pendant l'apprentissage, le réseau est encouragé à minimiser la différence entre la sortie prédite et la sortie réelle, et la valeur calculée par la fonction de coût servira de base pour mettre à jour tous les poids du réseau afin d'atteindre son objectif.

$$\mathcal{L}(y, \hat{y}) = |y - \hat{y}| \tag{D.5}$$

Comme la sortie $\hat{y}$ dépend directement de la sortie du réseau, qui dépend elle-même des paramètres du réseau θ, il est possible de les modifier afin que la sortie prédite soit plus proche de la sortie souhaitée. Cette opération se fait par une opération appelée *descente de gradient* en utilisant l'algorithme de rétro-propagation du gradient d'erreur (*backpropagation*, Rumelhart, Hinton & Williams 1986). Cette opération consiste à calculer la dérivée de la fonction de coût par rapport aux paramètres du réseau. Intuitivement, cela mesure la responsabilité d'un poids donné dans l'écart entre la sortie prédite et la sortie attendue :

$$\theta_{t+1} = \theta_t - \eta \frac{d\mathcal{L}(y, \hat{y})}{d\theta} \tag{D.6}$$

où les poids à θ_{t+1} sont mis à jour par un facteur η qui est appelé le taux d'apprentissage. Le taux d'apprentissage contrôle la force des mises à jour des poids du réseau. S'il est trop élevé, le réseau risque de rater la solution optimale, tandis qu'un taux d'apprentissage trop faible ralentira l'apprentissage et risque en outre de bloquer le réseau dans un optimum local. On utilise généralement la descente de gradient par lots, où la perte est moyennée sur un lot (*batch*) d'exemples, c'est-à-dire sur plusieurs exemples échantillonnés de manière aléatoire dans l'ensemble d'apprentissage.

3.5 Modèle de parole visuellement contextualisée

3.5.1 Modèle CELL

Le premier modèle de parole visuellement contextualisée (PVC, en anglais *Visually Grounded Speech Models* ou *VGS models*) est le modèle CELL (Cross-channel Early Lexical Learning) développé par Roy & Pentland (2002) et Roy (2003). Ce modèle a été explicitement développé afin de comprendre comment l'interaction des stimuli visuels et auditifs permet l'acquisition lexicale. Le but du modèle CELL est d'apprendre des appariements audio-visuels (appelés prototypes audio-visuels) entre divers objets et leurs formes verbales afin de constituer un proto-lexique comme le ferait un enfant. Afin d'apprendre des prototypes audiovisuels, le modèle CELL reçoit en entrée un contexte visuel (image) et son énoncé apparié. Le modèle CELL recherche des séquences audio récurrentes qui sont stockées avec l'objet qui était manipulé lorsque la séquence a été extraite. Les paires objet/son très fréquentes sont conservées puis ensuite filtrées selon un critère d'information mutuelle afin d'enlever les paires de mauvaise qualité (par exemple les paires objets/déterminant, objet/pronom, qui sont certes fréquentes mais non discriminantes).

Les résultats montrent que le modèle CELL est capable d'apprendre des paires audio-visuelle (AV) sémantiquement valides (par exemple, chaussure, clef, chien, toutou) et indique une précision sémantique de 85% des paires découvertes. Certains des éléments AV contiennent également des "mots" non standard — qui ne sont pas comptés comme des paires sémantiquement exactes — tels que des sons onomatopéiques (par exemple aboiement, bruit de moteur), bien que de telles paires soit exactes d'un point de vue de l'acquisition du langage, le vocabulaire des enfants étant aussi constitué d'onomatopées.

Ce modèle a cependant fait plusieurs hypothèses simplificatrices, notamment le fait que la parole est perçue de manière catégorielle en termes de phonèmes et les images devaient être prétraitées de manière à détacher d'abord l'arrière-plan du premier plan, puis à isoler l'objet de l'image. En outre, l'entrée visuelle était délibérément simplifiée de manière à ne contenir qu'un seul objet, ce qui facilitait la tâche du modèle. Cependant, le modèle CELL constitue l'un des tout premier modèle computationnel d'acquisition du langage en contexte, et est le précurseur des modèles neuronaux actuels.

3.5.2 Modèles neuronaux

Les réseaux neuronaux ont permis aux chercheurs de modéliser des interactions encore plus complexes entre les modalités visuelles et orales. Gabriel et al. (2014) a introduit, à notre connaissance, le premier modèle PVC neuronal (basé sur un CNN), dans lequel le modèle est entraîné à faire correspondre des images à des mots parlés isolés. Leur modèle comporte deux branches, une branche acoustique et une branche visuelle qui vectorise respectivement le mot d'entrée et l'image. Le modèle est entraîné à minimiser la distance cosinus entre les vecteurs correspondants, tout en augmentant cette distance pour les vecteurs non correspondants (approche contrastive). Leur modèle est ensuite évalué sur une tâche de recherche image/parole : retrouver l'image correspondante à partir d'un mot en entrée, ou vice-versa. Leurs résultats indiquent que leur réseau est effectivement capable de capturer les liens intermodaux et d'apprendre efficacement à associer un mot parlé à son contexte visuel. Leur modèle a ouvert la voie à des modèles traitant des stimuli acoustiques plus complexes, tels que des légendes complètes au lieu de mots isolés et a posé les bases de l'ensemble des modèles neuronaux utilisés aujourd'hui, par l'utilisation d'une fonction de coût contrastive notamment.

Harwath & Glass (2015), s'appuyant sur les travaux de Gabriel et al. (2014), ont proposé le premier modèle capable de gérer des légendes complètes au lieu de mots isolés. Leur modèle peut être considéré comme une version améliorée du modèle CELL car il fait également quelques hypothèses simplificatrices (les légendes sont segmentées en mots et les images pré-traitées afin d'en extraire les objets). L'objectif du réseau est d'aligner chaque mot de la légende audio sur chaque boîte de délimitation de l'image. Harwath et al. (2016) ont amélioré leurs travaux précédents de sorte que leur modèle puisse utiliser des légendes complètes au lieu de légendes pré-segmentées. De plus, au lieu d'exécuter un détecteur d'objets sur l'image, ils utilisent simplement l'avant-dernière activation d'un réseau VGG pré-entraîné. Finalement, Harwath, Recasens, Surís, Chuang, Torralba & Glass (2018) changent complètement leur modèle afin d'avoir un modèle entièrement convolutif : au lieu d'utiliser des vecteurs VGG, ils utilisent le réseau VGG complet jusqu'à la dernière couche convolutive (c'est-à-dire en supprimant les couches entièrement connectées), et utilisent des convolutions 1D pour la légende d'entrée. Afin de joindre les résultats des deux branches, ils utilisent un produit scalaire entre la matrice représentant l'image et celle représentant le signal de parole. Le fait de conserver les *feature maps* des convolutions des deux branches leur permet ainsi de détecter les zones de saillance entre l'image convoluée et l'entrée parlée convoluée avec un niveau de précision très fin. Leur modèle est en mesure de mettre en

évidence des régions spécifiques du spectrogramme qui correspondent à des régions spécifiques de l'image, ce qui leur permet de construire des paires image/audio de manière non supervisée.

Dans leur travail, Harwath et collègues sont donc passés de la recherche de paires audio-visuelles à partir d'images pré-segmentées et de légendes pré-segmentées à la recherche de paires audio-visuelles à un fin niveau de granularité sans aucune supervision à partir de parole brute appariée à des images brutes. Globalement, leur travail montre qu'il est possible d'extraire des unités de type mot à partir de la parole brute en utilisant des images comme une forme de supervision faible. Plus important encore, alors que leur réseau a été entraîné à minimiser la distance entre une image et sa légende correspondante à une échelle globale, le réseau a déduit des similarités à une échelle locale. Par conséquent, l'acquisition lexicale semble être un sous-produit de la tâche principale, et apparaît "naturellement" dans le modèle.

Contrairement aux CNN, les RNN sont conçus pour traiter des séquences telles que la parole. Même si les CNN sont également capables de traiter des séquences, les RNN sont plus plausibles d'un point de vue cognitif. En effet, comme nous l'avons déjà mentionné, la sortie à un temps donné dépend des prédictions des pas de temps précédents, ce qui n'est pas le cas pour les CNNs.

Le premier modèle PVC à base de RNN que nous connaissons est celui de Chrupała et al. (2017*a*). L'architecture utilisée est très similaire à celle des travaux précédemment cités et comporte deux branches : une branche audio et une branche visuelle ; et utilise le même type de fonction de coût que Harwath & Glass (2015). De même que Harwath et al. (2016), ils utilisent des vecteurs VGG pour représenter l'image. Néanmoins, contrairement à Harwath et al. (2016), l'encodeur acoustique est constitué de Recurrent Highway Networks (RHN, Zilly et al. 2017) empilés au lieu de couches convolutives. L'intégration finale de la branche audio est calculée par un mécanisme d'attention (voir section 3.4) qui apprend à pondérer les vecteurs d'entrée, de manière à donner plus de poids à des parties spécifiques du signal vocal. Ils évaluent leur modèle sur une tâche de recherche d'images vocales, c'est-à-dire qu'à partir d'une légende parlée, le réseau doit retrouver l'image correspondante. Ils obtiennent de meilleurs résultats que Harwath & Glass (2015), montrant que les modèles basés sur les RNN sont également capables de mettre en correspondance les deux modalités de manière adéquate.

Plus récemment, Merkx et al. (2019) se sont basés sur l'architecture de Chrupała et al. (2017*a*), en ajoutant un mécanisme d'attention à chaque couche (implémenté comme dans Chrupała et al. 2017*a*), et en utilisant des cellules récurrentes bidirectionnelles au lieu de cellules unidirectionnelles. Leur modèle a atteint des niveaux de performance encore plus élevés que Chrupała et al. (2017*a*) et Harwath & Glass (2015) sur le même jeu de données. Plusieurs optimisations ont également été faites, comme l'utilisation du taux d'apprentissage cyclique (*cyclic learning rate*, Smith 2017) qui consiste à avoir un taux d'apprentissage qui augmente et diminue plusieurs fois entre une borne supérieure et une borne inférieure, au lieu d'avoir un taux d'apprentissage strictement décroissant comme ce qui est fait habituellement. Cela permet au modèle de sortir des minima locaux dans lesquels il pourrait être bloqué et de converger vers une solution globalement meilleure. Cependant, ces meilleurs résultats se font au détriment de la plausibilité cognitive. En effet, comme ce modèle utilise des cellules bidirectionnelles, il traite l'entrée parlée des deux côtés en même temps, ce qui n'est bien sûr pas possible pour les humains.

Ainsi, les modèles basés sur les RNN sont donc capables d'apprendre la correspondance parole/image de manière aussi fiable que les modèles basés sur les CNN. Ils semblent également être des modèles assez flexibles, car ils sont capables d'apprendre rapidement la

correspondance entre de nouvelles paires objet/nom comme le montre les travaux de Krishnamohan et al. (2020). Le fait que ces modèles utilisent des cellules RNN au lieu de CNN les rend également plus plausibles sur le plan cognitif, et constituent donc des bancs d'essai idéaux pour simuler l'acquisition lexicale.

3.5.3 Analyse des représentations

Chrupała et al. (2017*a*) ont essayé de comprendre les représentations apprises par leur modèle. Ils ont constaté que toutes les couches ne sont pas également informatives sur la présence d'un mot. En particulier, ils ont constaté que les couches inférieures de leur architecture étaient les moins informatives, tandis que l'avant-dernière était la plus fiable. Cela suggère que les unités ressemblant à des mots sont progressivement construites au fur et à mesure que l'information circule dans le réseau, et qu'une certaine quantité de calcul est nécessaire pour que l'information apparaisse. Une observation similaire a été faite plus récemment par Merkx et al. (2019). Leur étude révèle également que les couches inférieures de leur architecture encodent la forme (c'est à dire, des informations phonétique) tandis que les couches supérieures encodent la sémantique. Alishahi et al. (2017) explorent dans quelle mesure un tel modèle PVC encode la phonologie. Leur étude montre que les encodages du réseau regroupent approximativement les phonèmes de l'anglais par classe de sons (plosives, fricatives, affricées, etc.). Cependant, le regroupement n'est pas parfait et semble se faire sur la base de facteurs acoustiques (formants) plutôt que sur des facteurs linguistiques plus profonds.

Harwath et collègues ont également effectué une analyse des représentations apprises par leur modèle. Les travaux de Harwath et al. (2016) suggèrent que leur réseau effectue une segmentation implicite de l'entrée audio en sous-unités. Drexler & Glass (2017) a montré que certains neurones étaient spécifiquement activés par certaines séquences de phonèmes. De façon intéressante, et de façon similaire à l'observation de Chrupała et al. (2017*a*) et Alishahi et al. (2017), leur étude révèle que les couches inférieures du réseau encode des représentations liées plus à la forme qu'à la sémantique.

3.5.4 Jeux de données

Étant donné que le nombre d'ensembles de données librement disponibles qui comportent à la fois des images et des descriptions parlées est limité, la plupart des modèles mentionnés ci-dessus sont entraînés sur des extensions d'ensembles de données initialement conçus pour entraîner des modèles qui génèrent des descriptions textuelles d'image. Les principaux ensembles de données utilisés pour cette tâche sont Flickr8k (Rashtchian et al. 2010, Hodosh et al. 2013) et MSCOCO (Lin et al. 2014) qui présentent des images associées à 5 légendes descriptives écrites par des humains. Ces ensembles de données présentent naturellement un langage contextualisé, car chaque légende descriptive est associée à une image, ce qui constitue, dans une certaine mesure, une connaissance du monde physique. Comme les légendes ont été écrites par des annotateurs après avoir vu l'image, l'image reflète naturellement l'intention communicative exprimée dans les légendes.

Cependant, ces données sont artificielles, dans le sens où elles ne reflètent pas véritablement ce à quoi un enfant est confronté. De récents résultats, utilisant des jeux de données réalistes (voir Tsutsui et al. 2020 et Räsänen & Khorrami 2019) montrent que leur modèle respectif peine à apprendre. Cela souligne la nécessité de disposer de jeux de données plus réalistes si l'on veut utiliser des simulations informatiques pour comprendre l'acquisition du langage chez l'enfant, comme le préconise Dupoux (2018).

3.5.5 Simulation ou modélisation de l'acquisition du langage ?

La plupart des modèles susmentionnés s'inspirent de l'acquisition du langage chez l'enfant, ou du moins établissent un lien explicite entre le but ultime de leur modèle et l'acquisition du langage chez l'enfant. Il semble raisonnable de se demander dans quelle mesure ces modèles *modélisent* l'acquisition du langage chez les enfants ou *simulent* l'acquisition du langage chez les enfants.

La différence entre simulation et modélisation ne concerne que les deux premiers niveaux de la hiérarchie de Marr, à savoir le niveau computationnel , et le niveau algorithmique ; le premier s'intéressant au problème à résoudre et à l'objet du calcul, tandis que le second s'intéresse à la manière dont le calcul est mené et définit chaque étape du calcul précisément. Rieder (2003, p. 818) fait une différence entre simulation et modélisation. Pour lui, la "simulation traduit l'action d'imiter la réalité et [le] modèle représentant le véhicule". La simulation est donc "l'action de présenter l'apparence, ou d'interagir avec le comportement, d'un système sans la réalité", tandis que la modélisation serait "la génération d'un fac-similé ou d'une représentation du système réel", ce dernier pouvant être "physique, mathématique, procédural [c'est-à-dire algorithmique], ou une combinaison de ces éléments".

Nous pensons que si l'on souhaite modéliser l'acquisition du langage chez l'enfant, le modèle doit être aussi proche que possible de ce que font les humains sur le plan algorithmique, et par exemple incorporer les mêmes *priors* que les humains, ce qui n'est le cas d'aucun des modèles PVC susmentionnés. Par conséquent, même s'ils *simulent* dans une certaine mesure l'acquisition du langage chez l'enfant, et plus particulièrement l'acquisition lexicale, ils ne *modélisent* pas l'acquisition du langage chez l'enfant. Néanmoins, le fait de ne pas implémenter le même "véhicule" que les humains n'empêche pas ces implémentations d'afficher des résultats similaires dans la simulation finale, ce qui rend leur implémentation et leur étude intéressantes pour tester des hypothèses.

3.5.6 Simulation parfaite et *grounding*

Selon Dupoux (2018) (et d'autres), une approche réussie pour étudier l'acquisition du langage chez l'enfant à l'aide d'approches computationnelles devrait satisfaire aux contraintes suivantes :

> construire des systèmes informatiques évolutifs capables, lorsqu'ils sont alimentés par des données d'entrée réalistes, d'imiter l'acquisition du langage telle qu'elle est observée chez les enfants.Il faut (Dupoux 2018)

Si l'on veut imiter l'acquisition du langage telle qu'elle est observée chez les nourrissons — ce qui serait ce que nous définissons comme une *simulation parfaite* — les modèles susmentionnés ne sont pas entièrement adaptés. En effet, ces modèles n'intègrent que les capacités perceptives et non les capacités productives d'un enfant. Ainsi, pour Roy (2005) le modèle idéal devrait être à la fois capable de perception et de production afin d'imiter réellement l'acquisition du langage chez l'enfant et ayant des boucles de rétroactions. Aucun des modèles que nous avons présentés ne met en œuvre toutes les boucles de rétroaction nécessaires pour obtenir un modèle totalement *grounded* au sens de Roy (2005). Cependant, bien qu'ils n'intègrent pas tous les aspects de l'acquisition du langage chez l'enfant, ils en intègrent un sous-ensemble, à savoir l'*acquisition lexicale*. Ce processus implique en effet de construire des schémas sur le monde (c'est-à-dire des représentations neuronales) en utilisant une entrée linguistique (c'est-à-dire des sous-titres audio) et une entrée non linguistique (c'est-à-dire un contexte visuel sous forme d'images fixes). Ainsi, ces modèles se prêtent à la simulation de l'acquisition lexicale chez l'enfant.

3.6 Conclusion

Des recherches antérieures montrent que les modèles PVC, qu'ils soient basés sur des CNN ou des RNN, sont capables d'apprendre avec succès comment apparier un stimulus acoustique à un stimulus visuel. Ces travaux révèlent également qu'en dépit du fait qu'ils n'ont été entraînés qu'à minimiser la distance entre une image et sa description orale afin d'apparier les deux modalités avec précision, ces modèles ont développé des capacités linguistiques plus profondes. Il s'agit d'une propriété intéressante, car ces capacités linguistiques apparaissent comme un sous-produit de la tâche principale. Cela ressemble en quelque sorte à la façon dont les enfants acquièrent leur langue maternelle.

Alors que ces capacités ont été largement explorées pour les modèles basés sur les CNN, elles ne l'ont pas été pour les modèles basés sur les RNN. Nous cherchons donc dans cette thèse à mieux comprendre quelles capacités linguistiques les modèles PVC à base de RNN sont capables de développer. Plus précisément, nous cherchons à répondre aux questions suivantes :

- Les modèles basés sur les RNN mettent-ils en évidence des parties spécifiques de l'entrée parlée qui sont particulièrement pertinentes pour prédire l'image cible, comme cela a été montré pour les modèles basés sur les CNN ? (Harwath & Glass 2017)
- Si oui, quelles parties spécifiques de l'entrée sont mises en évidence ?
- S'ils mettent effectivement en évidence des parties spécifiques de l'entrée, cette capacité est-elle valable d'une langue à l'autre, même si le modèle est entraîné avec une langue typologiquement différente de l'anglais, par exemple le japonais ?
- Dans quelle mesure ces capacités linguistiques se développent-elles avec le temps ? Les modèles PVC sont-ils capables d'acquérir rapidement certaines capacités linguistiques (c'est-à-dire avec une quantité de données relativement faible) ou non ?
- Des recherches antérieures (Chrupała et al. 2017*a* ; Merkx et al. 2019) montrent que les représentations internes d'un tel réseau encodent la présence des mots individuels des légendes parlées. Cela soulève la question de savoir si ce comportement est valable pour tous les mots d'une légende donnée, ou seulement pour des mots spécifiques ?
- Si la présence de mots individuels est encodée dans les représentations internes du réseau, cela signifie que le réseau a appris ce qui constitue un mot et a stocké cette information dans ses poids. Cela soulève la question de savoir comment le réseau reconnaît (c'est-à-dire active) la représentation d'un mot donné sur la base d'une entrée acoustique.
- Les modèles PVC sont entraînés avec des légendes complètes non segmentées et obtiennent des résultats corrects. Cependant, leurs homologues textuels (c'est-à-dire les réseaux entraînés sur des légendes écrites) obtiennent de meilleurs résultats. Cela soulève plusieurs questions, la principale étant : si les réseaux PVC étaient présentés avec des légendes segmentées, obtiendraient-ils de meilleurs résultats ?

4 Modèles de parole visuellement contextualisée et jeux de données

Les enfants acquièrent leur langue maternelle en contexte. Ainsi, l'accès à des stimuli uniquement oraux n'est pas suffisant pour qu'ils acquièrent leur langue maternelle. En effet, il est nécessaire que ces stimuli soient contextualisés (*grounded*) afin que les enfants puissent faire sens de ce qui est dit autour d'eux. Le contexte leur permet ainsi d'apparier des signes linguistiques (paire signifié/signifiant) à leurs référents. Les modèles neuronaux de parole visuellement contextualisée permettent de modéliser les interactions complexes qui ont lieu entre la modalité visuelle et la modalité auditive. Ainsi, ces modèles peuvent être utilisés afin de simuler l'acquisition du langage chez l'enfant. Ce chapitre présente donc les modèles qui seront utilisés pour les expériences décrites dans cette thèse ainsi que les données nécessaires à l'entraînement de ceux-ci.

4.1 Données

Les jeux de données que nous utilisons dans cette thèse sont des jeux de données initialement conçus pour une tâche de description textuelle d'images (*image captioning*). Ces corpus contiennent un ensemble d'images, chacune décrite par 5 descriptions textuelles écrites par des humains. Ces jeux de données sont le corpus MSCOCO (Lin et al. 2014), STAIR Yoshikawa et al. (2017) et Flickr8k (Rashtchian et al. 2010, Hodosh et al. 2013). Nous utilisons des extensions des jeux de données originaux, qui en plus d'avoir une description textuelle de chacune des images, incluent une version parlée de ces mêmes descriptions. L'extension audio de MSCOCO et de STAIR inclut une description audio générée par un système de synthèse vocale, tandis que l'extension audio de Flickr8k n'utilise pas de voix de synthèse mais des voix naturelles.

COCO L'extension audio de MSCOCO (Synthetically Spoken COCO Chrupała et al. 2017*b* que nous appellerons désormais COCO) que nous utilisons a été créée par Chrupała et al. (2017*a*). Celle-ci a été générée en utilisant le système de synthèse vocale de Google et ne comprend qu'une seule voix de synthèse. Ainsi, le jeu de données COCO comprend 616 435 descriptions audio pour un total de 123 287 images. Nous avons conservé le partitionnement original de données, à savoir 566 435 paires audio/image pour l'entraînement, 25 000 pour la validation et 25 000 pour le test.

STAIR L'extension audio de STAIR (Synthetically Spoken STAIR Havard et al. 2019*b* que nous appellerons désormais STAIR) est une extension que nous avons nous-mêmes créée pendant notre thèse. Le jeu de données original STAIR utilise les mêmes images que MSCOCO, mais contient, au lieu de descriptions textuelles en anglais, des descriptions textuelles en japonais. Il est important de préciser que les descriptions en japonais sont des descriptions originales et non des traductions des descriptions anglaises. Nous avons suivi la même méthodologie que Chrupała et al. (2017*a*) et avons généré des signaux acoustiques pour chacune des descriptions japonaises en utilisant le système de synthèse vocale de Google. Nous avons conservé le même partitionnement de données que Chrupała et al. (2017*a*) et Karpathy & Li (2017).

FLICKR8k Enfin, le dernier jeu de données est Flickr8k (Harwath & Glass 2015) qui contient de la parole naturelle : les descriptions ont été enregistrées par plus de 180 locuteurs. Ce corpus est le plus petit et ne contient que 8 000 images pour 40 000 descriptions. Les descriptions orales sont des verbatims des descriptions écrites, qui ont simplement été lues par des locuteurs natifs de l'anglais sur Amazon Mechanical Turk. 30 000 paires images/audio sont utilisées pour l'entraînement, 5 000 pour la validation et enfin 5000 pour le test.

Métadonnées Nous avons procédé à un alignement forcé des signaux de parole avec le texte pour chacun des trois corpus afin d'avoir des informations de frontières précises pour chacun des mots des descriptions. Cet alignement a été fait au moyen de l'outil en ligne Maus Forced Aligner (Kisler et al. 2017). De plus, nous avons également procédé à une annotation en partie du discours. Pour les corpus en anglais, cette annotation a été faite en utilisant TreeTagger (Schmid 1997), et pour le japonais, celle-ci a été faite en utilisant KyTea. Puisque l'annotation faite par KyTea (Neubig et al. 2011) était plus proche d'une annotation morphologique que morphosyntaxique, nous avons pris le parti d'éditer les annotations en fusionnant notamment certaines étiquettes entre elles (morphèmes "TAIL" et "SUF" notamment). Afin d'avoir des étiquettes comparables entre anglais et français, nous avons porté les étiquettes de chacun des deux étiqueteurs vers leur équivalent Universal POS (Petrov et al. 2012).

4.2 Architecture

L'architecture que nous utilisons est basée sur celle de Chrupała et al. (2017*a*) et est construite au moyen de la librairie Python Theano (James Bergstra et al. 2010). Comme toutes les architectures neuronales de parole visuellement contextualisée, cette architecture à deux parties principales, un encodeur d'image et un encodeur de parole. Cette architecture est entraînée pour réaliser une tâche de recherche d'image à partir d'une description audio. Ainsi, lorsqu'une description orale d'une image est donnée en entrée, le réseau retrouve l'image qui correspond à la description. Pour ce faire, le réseau projette l'image et le signal de parole dans un espace de représentation commun, de sorte que l'image et sa description correspondante soient proches dans l'espace de représentation alors qu'une image et une description ne formant pas une paire (la description correspond à la description d'une autre image) soit éloignées dans l'espace de représentation.

Encodeurs L'encodeur d'image consiste en une simple couche linéaire, chargée de réduire la dimension du vecteur d'image d'entrée. Ainsi, le vecteur VGG (Simonyan & Zisserman 2015) représentant l'image, extrait à partir d'un réseau de vision par ordinateur, se voit réduit pour avoir la même dimension que le vecteur représentant le signal acoustique. Concernant l'encodeur de parole, celui-ci prend en entrée des vecteurs MFCC qui sont ensuite passés à une couche convolutionnelle 1D. Cette couche convolutionnelle est suivie par 5 couches récurrentes composées de GRU (Gated Recurrent Units, Cho et al. 2014). Contrairement à l'implémentation de base qui utilise des cellules RHN (Zilly et al. 2017), nous avons fait le choix d'utiliser des GRU car ceux-ci sont mieux documentés dans la littérature. Nous utilisons deux mécanismes d'attention, l'un produisant un vecteur à partir des vecteurs de la première couche récurrente, et un second produisant un vecteur à partir des vecteurs de la cinquième couche récurrente. Le vecteur final de l'encodeur de parole correspond au produit terme à terme de chacun des deux vecteurs d'attention. Pour finir, le vecteur de parole et le vecteur d'image sont normalisés à la norme L2.

Fonction de coût La fonction de coût que nous utilisons est la fonction de coût contrastive (*contrastive loss*, également appelée *hinge loss* ou *triplet loss* Weinberger & Saul 2009):

$$\mathcal{L}(u,i,\alpha) = \sum_{u,i}\left(\sum_{u'}\max[0,\alpha + d(\vec{u},\vec{i}) - d(\vec{u'},\vec{i})] + \sum_{i'}\max[0,\alpha + d(\vec{u},\vec{i}) - d(\vec{u},\vec{i'})]\right) \tag{D.7}$$

où $\vec{u}$ est un vecteur représentant l'énoncé, $\vec{i}$ un vecteur représentant l'image, $\vec{u'}$ et $\vec{i'}$ sont des énoncés (respectivement des images) ne formant pas une paire avec l'image i (respectivement l'énoncé u). Cette fonction de coût encourage le réseau à minimiser la distance $d(\vec{u},\vec{i})$ entre l'image $\vec{i}$ et l'énoncé $\vec{u}$ appartenant à des paires image/énoncé concordantes pour

que celle-ci soit inférieure à une marge α, tout en augmentant la distance pour les paires image/énoncé non concordantes afin qu'elles soient éloignées d'une distance au moins égale à α dans l'espace de représentation. Le coût est calculé au niveau du *batch*, c'est-à-dire que toutes les images à l'intérieur d'un *batch* (à l'exception de l'image i) sont considérées comme des exemples d'inadéquation contrastive pour l'énoncé u, tandis que tous les autres énoncés (à l'exception de l'image u) à l'intérieur du même *batch* servent d'exemple d'inadéquation contrastive pour l'image i. Dans notre cas, la distance d utilisée est la distance cosinus.

Attention L'un des éléments clef de notre modèle est son mécanisme d'attention. On rappelle au lecteur que l'attention se calcule ainsi :

$$c = \sum_{t=1}^{T} \alpha_t h_t \tag{D.8}$$

où T est la longueur de la séquence, h_t est l'état caché produit par une cellule récurrente au moment t, et α_t est un paramètre qui est appris. Rappelons que le mécanisme d'attention apprend à attribuer un poids α_t à chaque vecteur d'entrée. Plus le poids est élevé, plus le vecteur a de l'importance dans la représentation finale. Dans notre cas, des poids d'attention élevés sur des vecteurs spécifiques signifient que le réseau a accordé plus d'attention à des parties spécifiques du signal d'entrée. Comme mentionné précédemment, nous utilisons deux modèles d'attentions, l'un après la première couche récurrente, l'autre après la cinquième couche récurrente, le vecteur final étant calculé ainsi :

$$c = \left(\sum_{t=1}^{T} \alpha_t^{GRU1} h_t^{GRU1} \right) \cdot \left(\sum_{t=1}^{T} \alpha_t^{GRU5} h_t^{GRU5} \right) \tag{D.9}$$

où T est la longueur totale de la séquence, h_t^{GRU1} et h_t^{GRU5} représentent respectivement le vecteur caché calculé au moment t par la 1^e et la 5^e couche récurrente, et α_t^{GRU1} et α_t^{GRU5} sont les poids d'attention calculés respectivement par le mécanisme d'attention suivant la 1^e et la 5^e couche récurrente. Le vecteur final est normalisé à la norme unitaire ℓ^2.

Présupposés et Biais Nous pouvons classer les hypothèses et biais que nous faisons implicitement en trois catégories : hypothèses sur la *nature des données*, hypothèses sur la *tâche*, et enfin hypothèses la *nature des opérations computationnelles.*

De par la nature même des images d'entrée, le contexte visuel fourni au modèle sera nécessairement limité. En effet, nous ne percevons pas en tant qu'humain des images fixes, mais des séquences d'images. De même, les images des corpus ont été soigneusement sélectionnées afin de mettre en avant des objets particuliers, et ainsi ne reflètent que peu le contexte visuel qu'un enfant pourrait percevoir, et notamment ne sont pas appropriées pour dépeindre des actions. Aussi, les descriptions liées aux images traduisent cela, et ne décrivent que peu les éventuelles actions qui ont lieu dans les images. De plus, en choisissant de telles données, nous faisons implicitement l'hypothèse que l'enfant entendrait systématiquement une description de son environnement. Cela est évidemment loin d'être le cas. Il arrive en effet très souvent que l'on évoque des personnes ou des objets qui ne sont pas physiquement présents.

Nous faisons également des hypothèses la tâche. En effet, le réseau est entraîné à résoudre une tâche très spécifique : retrouver une image à partir d'une description orale. Cela est une tache peu réaliste et ne correspond pas à ce que font des humains qui apprennent leur langue maternelle. Cependant, cette tâche met en jeu une capacité similaire à celle des humains, à savoir apprendre à faire un appariement entre un percept visuel et un percept acoustique. De plus, les humains ne se limitent pas à une seule modalité, mais utilisent de nombreuses autres modalités (odorat, toucher, proprioception, etc.), rendant ainsi la tâche que fait le réseau très spécifique et peu réaliste.

Enfin, nous faisons de nombreuses hypothèses sur la nature des opérations faites. Notre réseau traite les images et les signaux de paroles de manière indépendante, et ainsi la perception visuelle n'influe pas sur la perception auditive. Ce n'est pas le cas chez l'humain qui traite conjointement les deux modalités et où la perception d'une modalité peut influer sur la manière dont est perçue l'autre. Ainsi, il n'y a au sein de notre réseau pas de véritable interaction entre les deux modalités.

Ainsi, de par la nature des données que nous utilisons, de la tâche pour laquelle est entraîné le réseau et de la nature des opérations faites, les comparaisons que nous ferons avec l'acquisition du langage chez l'humain ne pourront être que limitées. Aussi, les conclusions que nous tirerons doivent être comprises à la lumière des hypothèses et des limites présentées : si la vision — sous la forme d'images fixes — devait être la seule modalité utilisée pour donner un sens à la parole environnante, et si la parole environnante se référait toujours au contexte visuel, quelles régularités devrions-nous attendre des enfants ?

5 Attention dans un modèle de parole visuellement contextualisée

L'hypothèse principale de cette thèse est que, afin d'apprendre une correspondance fiable entre une image et sa description orale, le modèle devrait implicitement apprendre à segmenter l'entrée orale en sous-unités. Comme les images utilisées représentent principalement des objets et non des actions, nous supposons que le modèle devrait apprendre implicitement à segmenter des unités nominales. Ainsi, nous étudions les mécanismes d'attention de nos modèles et analysons quelles parties du signal d'entrée sont mises en évidence, pour voir si effectivement le modèle met en avant des unités lexicales nominales.

5.1 Attention et méthodologie

Jain & Wallace (2019) affirment dans leur article "Attention is not Explanation" que les poids d'attention ne fournissent pas une explication significative des prédictions d'un réseau neuronal. Ils montrent notamment qu'il est possible de trouver des distributions alternatives des poids d'attention tout en conservant les prédictions intactes et arrivent à la conclusion que les poids d'attention ne constituent pas une source d'information fiable pour comprendre les prédictions d'un réseau neuronal. Wiegreffe & Pinter (2019) dans leur document de réponse "Attention is not not explication" atténuent toutefois la déclaration initiale de Jain & Wallace (2019) : pour eux les "scores d'attention sont utilisés comme fournissant *une* explication ; pas *l*'explication". Ils affirment qu'il est légitime de considérer les poids d'attention comme une forme d'explication étant donné qu'il est peu facile de trouver une distribution alternative qui conserve la prédiction du réseau intact ou qui ne crée pas de chute de performance trop importante.

Afin de nous assurer que l'attention de notre modèle explique en partie les prédictions, et pour mesurer objectivement son importance, nous allons mélanger de façon aléatoire les poids d'attention des deux mécanismes d'attention. Si les scores obtenus avec les poids mélangés sont moins bons que ceux obtenus avec les poids originaux, nous pourrons conclure que les poids d'attention originaux sont utiles pour la prédiction des modèles.

5.2 Mesure de l'attention

Dans ce chapitre, nous étudions les parties de l'entrée parlée auxquelles le modèle prête attention en analysant les poids d'attention. Nos analyses se concentrent sur les points suivants :

- Quels parties du discours (POS) sont mises en évidence ;
- Quels mots sont mis en évidence ;
- Quelles parties d'un mot donné sont plus spécifiquement mises en évidence ;
- Comment la distribution des poids d'attention évolue-t-elle dans le temps.

Après avoir entraîné les modèles sur l'ensemble de données anglais ou japonais, nous encodons chaque légende de l'ensemble de test et extrayons les poids d'attention α pour les deux mécanismes d'attention. Nous utilisons ensuite un algorithme de détection de pics pour détecter les maxima locaux dans les poids d'attention. Nous dirons que l'attention *se focalise sur* ou *met en évidence* une unité spécifique (POS, mot, etc.) s'il existe un maximum local parmi les poids d'attention attribués aux vecteurs qui composent cette unité. Pour des raisons de brièveté, à partir de maintenant, les poids d'attention calculés par le mécanisme d'attention suivant la première couche et la cinquième couche de GRU seront appelés respectivement "GRU1" et "GRU5".

Afin de comprendre si les unités mises en évidence sont différentes de ce que le hasard pourrait prédire, nous utilisons des pics aléatoires comme base de référence. Pour chaque légende, nous échantillonnons des pics aléatoires et calculons la distribution des mots et des POS sous ces pics aléatoires et regardons dans quelle mesure celle-ci est différente du hasard.

5.3 Résultat sur les corpus synthétiques

5.3.1 Résultats de référence

Afin de nous assurer que nos résultats ne sont pas dus au hasard, nous avons entraîné cinq modèles, chacun avec une graine différente. Les résultats présentés dans le tableau 4.1 sont une moyenne (± écart-type) des résultats obtenus pour les cinq modèles. Les modèles sont évalués en termes de rappel@k (R@k) sur une tâche de recherche d'images à partir d'une requête vocale. Pour chaque requête vocale, les images sont classées de l'image la plus proche à l'image la moins proche. Le R@k évalue si l'image cible (c'est-à-dire l'image réellement appariée avec la légende de la requête) est classée parmi les k premières images. Nous indiquons également le rang médian $\widetilde{r}$ qui nous informe sur le rang moyen des images réelles. Nos résultats (Table 4.1) révèlent que le réseau est parvenu à apprendre un appariement correct dans les deux langues. Cependant, nos résultats en anglais sont moins bons que ceux de Chrupała et al. (2017*a*) (R@1 de 5.5 contre 11.1) ce qui s'explique par le changement des cellules RHN par des GRU. Cependant, nos résultats sont bien meilleurs que le hasard (R@1 de 0.02).

5.3.2 Attention aléatoire

Pour vérifier la pertinence des poids d'attention nous avons soit mélangé les poids des deux mécanismes d'attention en même temps, soit mélangé alternativement l'un des deux afin d'estimer la contribution de chacun dans la prédiction finale. Nous l'avons fait sur le meilleur modèle entraîné (sélectionné sur l'ensemble de validation) de chacun des cinq entraînements pour COCO et STAIR (voir Table 4.2).

Lorsque les poids d'attention des deux mécanismes d'attention sont mélangés, nous observons que le R@1 est à peine supérieur à 0%, ce qui montre que le réseau est à peine capable de trouver l'image correcte à partir d'une requête orale. Lorsque nous mélangeons les poids d'attention de GRU1 mais laissons ceux de GRU5 intacts, nous remarquons que le réseau obtient de meilleurs résultats que lorsque nous faisons l'inverse. Dans les deux cas, les résultats sont moins bons que lorsque nous laissons les poids d'attention des deux mécanismes d'attention intacts (voir Tableau 4.1), mais meilleurs que lorsque les deux sont mélangés. Cela montre que les poids d'attention de GRU5 sont plus importants que ceux de GRU1, car le remaniement des poids du premier a un impact négatif beaucoup plus important que le remaniement des poids du second.

Ainsi, ces expériences nous permettent de conclure que les deux mécanismes d'attention mettent en évidence des unités utiles pour les prédictions des réseaux et que, par conséquent, l'analyse des poids d'attention est une entreprise légitime.

5.3.3 Parties du discours (POS) et mots

Pour COCO (Figure 4.2), nous observons que GRU5 se concentre fortement sur les noms ($85.89\% \pm 0.38$ des pics) et à peine sur les autres POS, et ce d'une façon très différente de ce que le hasard prédirait ($47.15\% \pm 0.09$). Le comportement de GRU1 est différent et semble beaucoup plus proche d'un comportement aléatoire : $52.37\% \pm 29.46$ des pics sont situés

au-dessus des noms alors que l'aléatoire prédirait $46.92\% \pm 0.15$. Cependant, ce résultat est dû à un entraînement fautif (d'où l'écart-type important) où l'attention était concentrée à la fin des légendes.

Pour STAIR (Figure 4.3), les deux mécanismes d'attention ont appris à se concentrer sur les noms : $71.33\% \pm 5.94$ pour GRU1 (Figure 4.3a) et $63.30\% \pm 2.76$ pour GRU5 (Figure 4.3b) alors que le hasard prédirait 45%. Nous remarquons que GRU5 se concentre principalement sur les particules $28.63\% \pm 2.15$ alors que l'aléatoire prédirait $15.88\% \pm 0.07$ (en japonais, les particules sont des clitiques qui indiquent notamment la fonction du mot précédent dans la phrase. Par exemple, "は" /wa/ marque le topique, et "が" /ga/ marque le sujet), tandis que GRU1 se concentre principalement sur les noms.

L'étude des POS mis en évidence par les mécanismes d'attention des modèles anglais et japonais révèle que les modèles ont adopté un comportement spécifique en fonction de la langue des légendes. En anglais, les noms sont les POS les plus mis en avant car ils font référence aux objets présents dans l'image. En japonais, cependant, les modèles ont tiré parti du fonctionnement de la langue en ne mettant pas seulement en évidence les noms, mais en adoptant un comportement spécifique à la langue lors de la mise en évidence des particules.

Pour COCO (Tableaux B.1 et B.2), les mots les plus mis en évidence sont des noms faisant référence à des objets concrets dans les images (train, tennis, toilettes, baseball, etc.). Pour STAIR (Tableaux B.3 et B.4), on remarque que sur les 10 premiers mots mis en évidence, respectivement 3 et 4 des mots pour GRU1 et GRU5 sont des particules. ("ga" (marqueur sujet), "no" (marqueur génitif) et "o" (marqueur objet) et aussi "ni" (marqueur locatif) pour GRU5) les autres mots étant des noms. La stratégie du réseau de mettre en évidence les particules est très intéressante car, en raison de la nature unidirectionnelle des unités récurrentes utilisées ; et parce que les particules sont des mots suffixés, les vecteurs qui constituent une particule contiennent beaucoup d'informations concernant le mot précédent. Par conséquent, la mise en évidence des particules est la stratégie optimale.

5.3.4 Position des pics

Nous analysons spécifiquement où les pics d'attention sont situés au-dessus des mots. Pour ce faire, nous avons divisé chaque mot situé sous un pic en quatre parties égales et nous comptons le pourcentage de pics situés au-dessus d'une partie donnée. Les résultats sont présentés dans le tableau 4.3.

Pour les modèles entraînés sur l'anglais, les pics d'attention ne sont pas situés précisément à la fin d'un mot, mais sont plutôt situés entre le milieu et la fin des mots mis en évidence. Cela semble indiquer que le réseau ne se concentre pas sur la représentation du mot entier, mais plutôt sur les vecteurs représentant la première moitié ou les deux tiers d'un mot. Pour le japonais, nous observons un comportement similaire puisque les pics de GRU1 et GRU5 sont globalement situés au-dessus des fins de mots. Cependant, la distribution des pics a tendance à être plus uniforme. Nous expliquons cela par le fait que les pics d'attention sur les légendes japonaises sont situés au tout début de la particule ou à la frontière avec le mot précédent (voir les deux pics de la figure 4.1b situés au début des particules "ni" et "ga"). Ainsi, la distribution des pics d'attention au-dessus d'une partie donnée du mot tend à être plus uniforme.

5.3.5 Étude longitudinale

Nous avons régulièrement sauvegardé les modèles pendant l'entraînement afin d'avoir une vision longitudinale de la façon dont l'apprentissage se déroule et avons donc sauvegardé

les poids des modèles chaque fois que le coût diminuait de 4% de sa valeur initiale. Nous avons sauvegardé également le modèle à chaque époque.

Nous remarquons d'abord que pour COCO (4.4a) et STAIR (4.4b), le nombre d'étapes de sauvegarde appartenant à la première époque représente une grande partie du graphique, ce qui montre que le coût a chuté assez rapidement et que la phase d'apprentissage la plus importante est en fait concentrée dans la première époque. Pour COCO, nous remarquons une évolution claire dans la proportion de pics mettant en évidence des noms où les pics au-dessus des noms sont passés de 37,4% à 55,9% après seulement 8 *batches* (256 exemples). Cela montre qu'il est possible pour le réseau de se concentrer sur des parties très spécifiques de l'entrée parlée avec très peu d'exemples. À la fin de la première époque, la proportion de pics au-dessus des noms n'évolue pas beaucoup et reste stable, oscillant autour de 85%. Pour STAIR, nous remarquons que dans les premières étapes de la première époque, il y a une compétition entre les POS, l'attention mettant en évidence simultanément les noms, les verbes et les particules. Puis à la fin de la première epoch, le nombre de pics mettant en évidence les particules dépasse celui des verbes, pour ne plus mettre en avant que des noms ou des particules à la fin de l'apprentissage. Contrairement à COCO, où l'attention n'évolue plus après la première epoch, on constate pour STAIR que l'attention et les unités sur lesquelles elle se focalise évolue tout au long de l'entraînement.

5.4 Résultat sur le corpus de parole naturelle FLICKR8k

5.4.1 Résultats de référence et attention aléatoire

Comme observé précédemment pour l'ensemble de données COCO, nos résultats sont inférieurs à ceux rapportés par Chrupała et al. (2017*a*) qui rapportent un R@1 de 5,5 alors que le nôtre n'est que de 2,08. Les résultats que nous obtenons sont également moins bons que ceux obtenus sur les ensembles de données COCO et STAIR. Néanmoins, même si les résultats ne sont pas particulièrement bons, ils sont toujours bien meilleurs que le hasard, ce qui montre que le réseau a été capable de donner un sens aux données et a appris une correspondance parole-image fiable.

Comme pour les expériences précédentes, nous avons mélangé les poids d'attention afin de comprendre s'ils étaient significatifs ou non. Les résultats sont présentés dans le tableau 4.5. Ici aussi, nous remarquons qu'en mélangeant aléatoirement les poids d'attention des deux mécanismes d'attention, nous obtenons un R@1 beaucoup plus faible que le R@1 original (-1.54pp). Une fois encore, cela montre que l'attention joue un rôle essentiel pour les modèles. Cependant, contrairement aux modèles COCO et STAIR, la majorité des prédictions des modèles reposent sur le vecteur contextuel calculé par le premier mécanisme d'attention plutôt que sur celui du cinquième.

5.4.2 Parties du discours (POS) et mots

Comme pour COCO, nous remarquons que les modèles ont appris à se concentrer principalement sur les noms. Les deux mécanismes d'attention mettent en évidence plus de noms que ce que le hasard aurait prédit (qui serait 37%), ce qui démontre également que le réseau se concentre délibérément sur les noms plutôt que sur tout autre POS. Cependant, dans ce cas et contrairement à COCO et STAIR, GRU1 a mis en évidence plus de noms dans l'ensemble ($72,58\% \pm 4,94$) que GRU5 ($58,73\% \pm 5,30$).

En examinant de plus près les mots spécifiquement mis en évidence par chaque mécanisme d'attention du meilleur modèle (présentés dans l'annexe B), nous remarquons que les 10 premiers mots mis en évidence par GRU1 sont uniquement des noms faisant référence à des objets présents dans les images (chien, homme, fille, garçon, chiens, personnes, femme,

enfant, balle, eau). Les 10 premiers mots de GRU5 contiennent également des noms, mais aussi des silences, des prépositions et des déterminants (</s>, eau, <sil>, plage, neige, herbe, chemise, rue, a, in). GRU5 semble donc moins spécialisé que GRU1.

5.4.3 Position des pics et étude longitudinale

Dans cette expérience également, nous avons analysé au-dessus de quelles parties de mots les pics d'attention sont spécifiquement situés (Tableau 4.6). Nous observons que les pics ont également tendance à favoriser les fins de mots. Comme pour COCO et STAIR, les pics ne sont pas situés à la toute fin des mots mis en évidence, mais semblent être plus concentrés au milieu, puisque les plus fortes proportions de pics se trouvent au milieu du début et de la fin des mots. Une fois de plus, ces résultats tendent à montrer que le modèle n'a pas besoin d'avoir accès au mot complet pour le reconnaître et le mettre en évidence correctement.

La figure 4.6 montre comment la proportion de POS mis en évidence évolue dans le temps. Nous nous concentrerons sur GRU1 car c'est le mécanisme d'attention le plus interprétable du modèle. Comme pour COCO et STAIR, on constate que le modèle se focalise vite sur les noms et ce avec peu d'exemples (400 exemples suffisent). On constate que la proportion de pics au-dessus des noms augmente ensuite régulièrement au cours de la première époque. Cependant, contrairement à COCO et STAIR où la proportion n'a pas beaucoup évolué après la première époque, nous remarquons ici que la proportion de pics au-dessus des noms évolue encore après la première époque. Cela confirme que le modèle a besoin de plus de temps afin d'identifier précisément tous les mots importants de la légende. Cela s'explique par le fait que Flickr8k est composé de parole naturelle. Ainsi, le modèle a besoin de plus de temps pour prendre en compte les variations intra- et inter-locuteurs.

5.5 Acquisition du langage

Le fait que nos modèles mettent préférentiellement en avant les noms semble cohérent avec ce qui est observé dans la littérature sur l'acquisition du langage. Ce phénomène est communément appelé le "*noun bias*". En effet, il a été constaté que le lexique des enfants — tant en perception qu'en production — contient une plus grande proportion de noms que de verbes ou de tout autre POS. Gentner (1982) a été la première à postuler que les noms sont plus faciles à apprendre que les verbes. Elle explique cela par le fait que "les mots qui se réfèrent à des concepts sont faciles à apprendre parce que l'enfant a déjà formé des concepts d'objet, et n'a besoin que de faire correspondre les mots et les concepts". Notre réseau neuronal est dans ce cas puisqu'il utilise des vecteurs VGG pré-entraînés qui codent les objets présents dans l'image.

Nous avons montré que notre modèle japonais développe un comportement spécifique à la langue lorsqu'il se concentre principalement sur les particules "ga". Haryu & Kajikawa (2016) ont observé que les enfants japonais (à partir de 15 mois) font également usage de la particule "ga" pour segmenter le discours. Nos modèles ont donc adopté la même stratégie que les enfants japonais pour segmenter le nom adjacent. Haryu & Kajikawa (2016) affirment qu'"il est clair que les particules liées au nom ne sont pas les premiers indices que les enfants utilisent pour la segmentation des mots". Nous observons effectivement ce type de schéma pour nos modèles où les particules sont à peine utilisées dans les premières étapes d'apprentissage, mais deviennent de plus en plus mises en avant, alors que les noms sont moins présents qu'au début.

5.6 Conclusion

Dans ce chapitre, nous avons entraîné des modèles PVC sur deux langues (anglais et japonais) et nous avons analysé le comportement de leurs mécanismes d'attention. Notre analyse a révélé que l'attention a adopté un comportement général par lequel elle apprend à détecter et à mettre en évidence des noms spécifiques dans l'entrée parlée. Nos expériences confirment ainsi l'intuition de Chrupała et al. (2017*a*) selon laquelle "le mécanisme d'attention du modèle de parole lui permet de sélectionner des fragments clefs des [...] énoncés". Nous avons également montré que l'attention pouvait aussi adopter un comportement spécifique à la langue que le réseau traite en mettant par exemple en avant les particules du japonais. Finalement, nous avons montré que la focalisation de l'attention sur les noms est un comportement appris, et qui apparaît avec relativement peu d'exemples.

6 Activation, compétition et reconnaissance lexicale

Dans le chapitre précédent, nous avons montré que les modèles PVC étaient capables de mettre en évidence des mots spécifiques dans le flux de la parole en utilisant leurs mécanismes d'attention. Ceci implique que le modèle est capable de reconnaître les mots qui sont mis en évidence. Cette aptitude soulève quelques questions auxquelles nous allons tenter de répondre dans le présent chapitre.

6.1 Reconnaissance de mots

Le fait que les modèles PVC soient capables de reconnaître des mots individuels a déjà été exploré par plusieurs études. Par exemple, Chrupała et al. (2017*a*) et plus récemment Merkx et al. (2019) ont montré que les embeddings d'énoncés calculés par les modèles PVC basés sur les RNN contiennent des informations sur la présence des mots de l'énoncé d'entrée. Cependant, ces études n'ont pas montré pour quel type de mots ce comportement est vrai et si le modèle a appris à associer ces mots individuels à leurs référents visuels. De plus, aucune de ces études n'a exploré les facteurs qui influencent la reconnaissance des mots et pourquoi certains mots semblent être assez bien reconnus alors que d'autres ne le sont pas du tout. C'est ce que nous cherchons à faire ici.

6.1.1 Appariement de mots isolés

Afin de déterminer si le modèle est capable d'associer des mots isolés à leurs référents visuels, nous avons sélectionné un ensemble de 80 mots qui correspondent aux noms de 80 catégories d'objets dans l'ensemble de données MSCOCO. Nous nous attendons à ce que notre modèle soit très efficace avec ces mots spécifiques, car ils font référence aux principaux objets présentés dans l'ensemble de données MSCOCO. Nous évaluons la capacité du modèle à classer les images de manière à ce qu'au moins une image parmi les 10 premières contienne l'instance de l'objet correspondant au mot cible présenté (Précision@10, abrégé P@10). Il est important de rappeler qu'au moment de l'entraînement, le réseau n'a reçu que des légendes complètes et non des mots isolés. Par conséquent, si le réseau est capable de retrouver des images contenant des occurrences du mot cible, cela montre qu'une segmentation implicite a été effectuée au moment de l'entraînement.

Les résultats sont présentés dans la figure 5.1. 40 mots sur les 80 mots cibles ont une P@10$\geq$ 0.8. Cela montre que le réseau est capable d'associer de manière fiable des mots isolés à leurs référents visuels, même s'il ne les a jamais vus isolément. À l'inverse, on remarque également que 15 mots ne sont pas mis en correspondance avec leurs référents visuels. Parmi les mots les mieux reconnus, on trouve des animaux (éléphant, zèbre, mouton, girafe) et des objets (camion, bus, bateau, avion) qui sont très fréquents dans notre jeu de données. Parmi les mots les moins bien reconnus, on trouve des objets (fourchette, couteau, vase, grille-pain) ou des animaux (souris) qui sont assez rarement mentionnés dans les légendes car ils peuvent être trop petits et non décrits par les annotateurs.

6.1.2 Facteurs influençant l'appariement

Nous explorons 2 types de facteurs : les facteurs liés à l'image et les facteurs liés à la parole. Pour ces derniers, nous considérons la fréquence des mots dans l'ensemble d'entraînement et la longueur du signal vocal. En ce qui concerne les facteurs liés à l'image, nous considérons la fréquence des instances d'objets dans les images de l'ensemble d'entraînement, le nombre moyen d'instances d'objets voisins et la surface moyenne de chaque objet. Pour modéliser la relation entre toutes ces variables, nous avons ajusté un modèle de régression

linéaire multiple avec R où nous essayons de prédire la Précision@10 en utilisant les facteurs mentionnés précédemment.

Les résultats sont présentés dans le Tableau 5.1. Nous remarquons que le seul effet qui joue un rôle significatif dans la reconnaissance des mots est la fréquence du mot dans l'ensemble d'entraînement : plus le mot est fréquent dans l'ensemble d'entraînement, mieux il est reconnu. La longueur du mot a un léger effet positif qui tend à montrer que les mots longs sont mieux reconnus que les mots courts. La fréquence des objets et le nombre d'objets voisins n'ont aucun effet. La taille des objets semble également avoir un léger effet.

Nos résultats montrent donc que les mots individuels sont effectivement mis en correspondance de manière fiable avec leur référent visuel par le réseau. Le principal facteur de réussite dans cette tâche est la fréquence des mots cibles dans la légende ainsi que la taille des objets dans l'image. Ainsi, les mots qui sont très fréquents et qui font référence à des objets de grande taille sont mieux reconnus que les autres.

6.2 Activation lexicale

6.2.1 Paradigme du *gating*

Le paradigme de gating a été introduit par Grosjean (1980) et implique la procédure suivante :

> Le paradigme du *gating* implique la présentation répétée d'un stimulus oral (dans ce cas, un mot) de telle sorte que sa durée depuis l'apparition est augmentée à chaque présentation successive. On procède ainsi jusqu'à ce que le mot entier ait été présenté. Après chaque présentation (ou *gate*), on demande aux sujets de noter le mot présenté et d'évaluer leur confiance dans chaque supposition. (Cotton & Grosjean 1984)

Dans notre cas, cela signifie que le modèle neuronal est alimenté par des versions tronquées d'un mot cible donné, chaque version tronquée comprenant une plus grande partie du mot cible. Dans notre cas, la troncation se fait soit de gauche à droite (le modèle n'a accès qu'à la fin du mot), soit de droite à gauche (le modèle n'a accès qu'au début du mot). Nous évaluons la capacité des modèles à classer les images de façon à ce que les k premières images contiennent des instances de l'objet cible (nous utilisons la Précision@10, P@10). Le but de cette expérience est de tester si le début du mot joue un rôle dans la reconnaissance des mots pour le réseau ou non. Si c'est le cas, nous nous attendons à ce que le réseau ne parvienne pas à récupérer les images du mot cible si le mot est tronqué de gauche à droite, mais pas — ou moins — lorsque le mot est tronqué de droite à gauche, ce qui motive la troncature des deux côtés.

6.2.2 Effets du *gating*

La figure 5.2a montre l'évolution moyenne de la P@10 sur les 80 mots de test. Comme on peut le voir sur le graphique, la précision évolue différemment selon la partie du mot qui a été tronquée. Lorsque les mots cibles sont tronqués de gauche à droite, la précision chute plus rapidement que lorsqu'ils sont tronqués de droite à gauche. Ces résultats montrent que le modèle est robuste à la troncation lorsqu'elle est effectuée de droite à gauche mais pas lorsqu'elle est effectuée de gauche à droite : lorsque les phonèmes initiaux des mots sont supprimés, le réseau ne parvient pas à retrouver l'image cible, mais lorsqu'on ne lui présente que les phonèmes initiaux, le réseau est globalement capable de retrouver des images correspondant au mot cible.

6.2.3 Activation abrupte ou graduelle ?

Nos expériences révèlent que de petites différences acoustiques produisent de grandes différences dans la représentation finale. Il semble que certains vecteurs MFCC jouent un rôle plus important que d'autres dans l'activation de la représentation finale.

Nous laissons progressivement le réseau voir de plus en plus de vecteurs MFCC composant un mot donné, en lui donnant itérativement des segments de vecteurs MFCC de plus en plus longs en commençant par le début du mot jusqu'à ce que le réseau ait eu accès au mot complet. Nous calculons ensuite la similarité cosinus entre l'embedding calculé pour chacune des versions tronquées du mot et l'embedding correspondant au mot complet. Plus la similarité cosinus est proche de 1, plus les deux représentations sont similaires. Si chaque vecteur MFCC contribue de manière égale à la représentation finale du mot, la similarité cosinus évoluera linéairement sinon la similarité cosinus évoluera par "sauts" plutôt que linéairement. Pour détecter les étapes qui pourraient se produire dans l'évolution de la similarité cosinus, nous approximons sa dérivée en calculant la différence de premier ordre. Les "sauts" élevées devraient ainsi se traduire par des pics (par exemple, figure 5.4b). Nous calculons l'évolution de la similarité cosinus pour les 80 mots cibles codés avec le meilleur modèle entraîné.

En moyenne, il y a 1,25 pic par mot pour le modèle entraîné contre 0,1 pic par mot pour notre condition de base, montrant que l'évolution du cosinus est linéaire dans le second, mais pas dans le premier. Par conséquent, dans notre modèle entraîné, certains vecteurs MFCC sont plus déterminants pour la représentation finale que d'autres. En effet, certains vecteurs MFCC déclenchent un "saut" élevé dans l'évolution du cosinus, ce qui suggère que l'embedding se rapproche soudainement de sa valeur finale.

6.3 Compétition lexicale

Certains modèles psycholinguistiques (voir section 1.4) supposent que le premier phonème d'un mot active tous les mots commençant par le même phonème. Le mot que le locuteur veut prononcer et qu'il prononce progressivement est appelé le mot "cible". Les mots qui sont activés mais qui ne correspondent pas au mot cible sont appelés "concurrents". Lorsque l'auditeur perçoit de plus en plus le mot cible, certains concurrents sont désactivés. Cela signifie qu'ils ne sont plus considérés comme le mot potentiel, car ils ne correspondent pas à ce qui est perçu.

6.3.1 Méthodologie

Nous avons sélectionné 12 paires de mots commençant par la même séquence de phonèmes et avons testé si elles étaient activées par un processus de compétition.[1]

Pour chaque paire de mots, nous avons sélectionné un des mots que nous considérons comme le mot cible, en laissant progressivement le réseau voir de plus en plus de vecteurs MFCC composant ce mot. À chaque pas de temps, le réseau produit un vecteur, que nous utilisons pour classer les images de l'image la plus proche à l'image la moins proche. Ensuite, pour les 50 images les plus proches, nous vérifions si au moins une des légendes contient soit le mot cible, soit le concurrent.

Comme le concurrent et le mot cible commencent par les mêmes phonèmes, nous nous attendons à ce que le réseau produise un vecteur qui active à la fois le concurrent et le mot cible au début, puis, lorsque le signal acoustique ne correspond plus au concurrent, nous

[1] wii/window, cat/cattle, cat/cow, cat/catcher, cattle/catcher, floor/flower, fridge/frisbee, kid/kitchen, player/plate, tree/train, meat/meter, and train/truck.

nous attendons à ce que le réseau soit capable d'activer uniquement le mot cible. Pour chaque paire de mots, chaque mot est utilisé alternativement comme mot cible.

6.3.2 Résultats

Trois résultats sont présentés pour 3 paires de mots : "Train/Truck" (5.6) pour laquelle une forte compétition existe, "Cat/Cattle" (5.5) pour laquelle une compétition modérée existe, et "Frigde/Frisbee" (5.7) pour laquelle aucun phénomène de compétition n'a pu être mis en évidence. Les modèles de la COHORTE (Marslen-Wilson & Welsh 1978, Marslen-Wilson 1987*b*) et TRACE (McClelland & Elman 1986*b*) affirment tout deux que les mots concurrents sont tous activés en même temps, c'est-à-dire lorsque le premier phonème est perçu. Même si certains exemples sont conformes à cette affirmation — comme celui présenté dans la figure 5.6b — ce n'est pas le cas de toutes les paires de mots. Dans certains cas, la compétition est minime car la représentation des deux mots est activée de manière séquentielle (comme dans la figure 5.5) et dans d'autres cas, il n'y a absolument aucune compétition entre les deux mots (comme dans la figure 5.7) bien que les deux mots commencent de manière similaire. Par conséquent, le comportement du réseau semble très peu clair. Il semble toutefois que le réseau active le mot qui est le plus courant. Par exemple, il y a beaucoup plus de photos de trains que de photos où les camions constituent l'objet principal de l'image. Ainsi, le réseau semble activer préférentiellement la représentation de l'objet qui est le plus fréquent dans les images et les légendes.

6.4 Conclusion

Dans ce chapitre, nous avons montré qu'un modèle PVC est capable de faire correspondre des mots individuels à leurs référents visuels bien qu'il ait été entraîné sur des légendes complètes. van Zon (1997, p. 8) note que dans les modèles COHORT et TRACE "la segmentation est le résultat de la reconnaissance". Nous pensons que la reconnaissance de mots individuels montre que le modèle a implicitement segmenté ses entrées en sous-unités, et proposons la formulation inverse : "la reconnaissance est la preuve de la segmentation". Un point important que nous avons souligné dans ce chapitre est cependant que la segmentation résultante peut ne pas toujours correspondre à des mots graphiques.

Dans ce chapitre, nous avons introduit plusieurs méthodologies pour analyser comment la représentation des mots individuels se construit au cours du temps. Notamment, nous avons adapté le paradigme de *gating* de Grosjean (1980) afin d'analyser comment les mots sont activés par le réseau. Nous avons observé que la représentation des mots n'est pas construite linéairement par le réseau et que la reconnaissance peut se produire avec une entrée partielle, corroborant ainsi que "la reconnaissance se produit souvent avant la fin du mot" (van Zon 1997, p. 8). Nous avons montré que pour pouvoir activer la représentation correcte d'un mot donné, le réseau doit avoir accès aux premiers phonèmes de ce mot, car lorsqu'ils sont supprimés, le réseau est incapable d'activer la représentation correcte. Ainsi, lorsque la reconnaissance de mots est observée avec une entrée partielle, c'est uniquement lorsque l'entrée partielle englobe la première partie du mot, mais pas lorsqu'elle n'englobe que la partie finale du mot.

Finalement, nous avons cherché à étudier la compétition lexicale sans avoir pu parvenir à mettre en avant une systématicité de ce phénomène dans notre réseau.

7 Impact de l'introduction d'information linguistiques

Dans les chapitres précédents, nous avons montré que les modèles de parole visuellement contextualisée basés sur des RNN utilisent leur mécanisme d'attention pour mettre en évidence les mots qui sont pertinents pour retrouver l'image correcte, et que ces modèles segmentent implicitement l'entrée parlée en sous-unités. Dans ce chapitre, au lieu de comprendre quels types d'unités ont été implicitement segmentés par le réseau, nous abordons le problème dans l'autre sens et nous posons la question suivante : quelle segmentation maximise la performance (rappel@k) d'un modèle PVC si la parole devait être segmentée ? Afin de répondre à cette question, nous explorons *comment* l'information de frontière peut être intégrée, *quel* type de frontière est le plus efficace (soit le phone, la syllabe ou le mot), et enfin *où* – c'est-à-dire, à quelle couche – une telle frontière devrait être introduite dans le réseau. Enfin, nous explorons également des modèles hiérarchiques pour lesquels nous fournissons plusieurs niveaux de segmentation en même temps afin de comprendre l'effet de la modélisation explicite de la nature hiérarchique de la parole.

7.1 Information de frontières

7.1.1 Types de frontières

Comme indiqué précédemment, nous souhaitons donner des informations linguistiques à notre réseau, et plus particulièrement des informations sur les frontières des segments. Dans ce chapitre, nous définissons un *segment* comme étant soit un phone, soit une syllabe, soit un mot. Les frontières des segments ont été dérivées des métadonnées de l'alignement forcé (voir § 3.2.4) afin de savoir quel vecteur MFCC constitue une frontière ou non.

Nous considérons deux types différents de syllabes dans ce travail : en effet, lorsque nous parlons, les mots ne sont pas prononcés les uns après les autres de manière déconnectée, mais sont plutôt connectés par un processus appelé "resyllabification". En anglais, ce phénomène est visible lorsqu'un mot se terminant par une consonne est suivi d'un mot commençant par une voyelle. Dans ce cas, la consonne finale du premier mot tend à s'en détacher et à se rattacher au mot suivant, franchissant ainsi la frontière du mot. Les deux types de syllabes que nous considérons dans ce travail sont les suivants : "syllabe-mot" qui fait référence aux syllabes qui résultent d'une segmentation qui ne prend pas en compte la resyllabification, et "syllabe-connectée" qui fait référence aux syllabes qui résultent d'une segmentation qui prend en compte la resyllabification.

Ainsi, pour chaque légende, nous avons une séquence X de longueur T de vecteurs acoustiques de dimension d : $X = \left[x_1^d, x_2^d, ..., x_T^d\right]$; et une séquence correspondante de scalaires B de longueur T représentant les frontières $B = [b_1, b_2, ..., b_T]$, $b_t \in \{0, 1\}$, où $b_t \triangleq 1$ si x_t est une limite de segment, 0 sinon.

7.1.2 Intégrer des frontières de segments

Afin d'intégrer des informations de frontière dans notre réseau, nous tirons parti de sa conception, et plus particulièrement des cellules récurrentes et de la manière dont ces cellules calculent leur sortie, qui peut être formalisée comme suit :

$$h_t = f\left(h_{t-1}, x_t\ ; \theta\right) \tag{D.10}$$

où l'état caché au temps t, noté h_t, est une fonction f du vecteur précédent h_{t-1} et du vecteur d'entrée actuel x_t, et où θ est un ensemble de paramètres entraînables de la fonction f.

Notre approche pour intégrer l'information de frontière dans les couches récurrentes de notre réseau peut être formalisée comme suit :

$$h_t = \begin{cases} f\left(h_0, x_t\ ; \theta\right), \text{ if } b_{t-1} = 1 \\ \\ Nf\left(h_{t-1}, x_t\ ; \theta\right), \text{ otherwise} \end{cases} \tag{D.11}$$

Dans notre approche, h_t ne dépend du vecteur précédent h_{t-1} que si le vecteur précédent n'est pas un vecteur acoustique correspondant à la frontière d'un segment ($b_{t-1} \neq 0$). Si le pas de temps précédent correspond à la frontière d'un segment ($b_{t-1} = 1$), nous réinitialisons l'état caché pour qu'il soit égal à h_0. Ainsi, les vecteurs d'un même segment sont temporellement dépendants, mais les vecteurs appartenant à deux segments différents ne le sont pas. Les GRU qui utilisent ce schéma de calcul seront dorénavant appelés $\text{GRU}_{\text{PACK.}}$, car les vecteurs appartenant au même segment sont "regroupés" (*packed*) ensemble.

7.1.3 Condition ALL et KEEP

À partir de cette configuration initiale de $\text{GRU}_{\text{PACK.}}$, nous en avons proposé deux versions différentes : ALL et KEEP. Dans la version ALL (voir Figure 6.1b), tous les vecteurs appartenant à un segment sont transmis à la couche suivante. Dans la version KEEP, seul le dernier vecteur de chaque segment est transmis à la couche suivante (voir Figure 6.1c). La longueur de la séquence de sortie et d'entrée reste la même dans la condition ALL mais dans la condition KEEP, la longueur de la séquence de sortie est plus courte que la séquence d'entrée.

7.2 Méthodologie

Afin de comprendre où les informations de frontière doivent être introduites (c'est-à-dire à quel niveau de l'architecture), nous entraînons autant de modèles que le nombre de couches récurrentes, où chaque fois une couche de GRU est remplacée par une couche $\text{GRU}_{\text{PACK.}}$. Par exemple, "$\text{GRU}_{\text{PACK.}}$–3 " fait référence à un modèle où la troisième couche de GRU est une couche $\text{GRU}_{\text{PACK.}}$ et les autres couches (1^e,2^e,4^e, et 5^e) sont des couches de GRU classiques.

Afin de comprendre si l'introduction d'informations sur les frontières aide le réseau dans sa tâche, nous comparons les performances des modèles utilisant des informations sur les frontières avec un modèle de base qui n'en utilise aucune (ainsi, toutes les couches récurrentes de l'architecture de base sont des couches de GRU normaux). Ce modèle sera dorénavant appelé BASELINE. Nous introduisons également une autre condition, où, au lieu d'entraîner des modèles avec des limites de segment réelles (que l'on appellera désormais TRUE), nous entraînons des modèles avec des limites aléatoires (que l'on appellera désormais RANDOM). Les frontières aléatoires ont été générées en mélangeant simplement la position des frontières réelles, ce qui donne autant de frontières positionnées aléatoirement que de frontières réelles.

Les modèles sont évalués en termes de Rappel@k (R@k). Pour une requête orale, le R@k évalue la capacité des modèles à classer l'image jumelée cible parmi les k meilleures images. Afin d'évaluer si les résultats observés dans nos différentes conditions expérimentales (TRUE-ALL, TRUE-KEEP, RANDOM-ALL, RANDOM-KEEP) sont différents les uns des autres et de la condition BASELINE, nous avons utilisé un Z-test de proportion binomiale. Ce test est utilisé pour évaluer s'il existe une différence statistique entre deux proportions indépendantes. Dans notre cas, la proportion que nous testons est le nombre de succès sur le nombre

d'essais (qui correspond au nombre de paires légende/image différentes dans l'ensemble de test).

7.3 Résultats

Dans l'ensemble, nos paramètres expérimentaux ont conduit à l'entraînement de 81 modèles différents par ensemble de données. Les résultats de base sont présentés dans le tableau 6.1. Les résultats pour les conditions TRUE/RANDOM obtenus sur les jeux de données COCO et FLICKR8K sont présentés dans le Tableau 6.2 et 6.3 respectivement. Nous obtenons des résultats plus faibles sur FLICKR8K que sur COCO, ce qui montre la difficulté de la tâche sur la parole naturelle.

Dans l'ensemble, les modèles entraînés sur FLICKR8K avec les frontières TRUE ont un R@1 significativement meilleur que leurs homologues de base et les modèles entraînés avec les frontières RANDOM, ce qui indique que les modèles ont utilisé efficacement l'information sur les frontières. Pour COCO, nous observons que certains modèles entraînés avec des frontières aléatoires ont des scores significativement meilleurs que BASELINE (en particulier lors de l'utilisation de frontières de phone dans la condition KEEP), cependant cet effet disparaît lors de l'utilisation d'unités plus grandes, telles que des mots.

Il existe un effet d'interaction entre l'utilisation de frontières TRUE et RANDOM, que ce soit dans la condition ALL ou KEEP. En effet, dans la condition RANDOM-ALL, aucun résultat n'est statistiquement meilleur que BASELINE, alors que dans la condition RANDOM-KEEP, les résultats sont statistiquement moins bons que BASELINE. Par conséquent, l'utilisation de frontières aléatoires qui ne délimitent pas d'unités linguistiques significatives nuit réellement aux performances du réseau. Cependant, dans la condition TRUE-KEEP, les résultats sont meilleurs que BASELINE. Par conséquent, la condition KEEP contraint le réseau à apprendre de meilleures représentations, alors que dans la condition ALL, les informations sur les frontières sont diluées par les vecteurs voisins, ce qui conduit à une utilisation sous-optimale de ces informations.

Dans nos expériences, nous avons utilisé quatre types différents de segments correspondant à deux types différents d'unités linguistiques : les phones, les syllabes-connectées, les syllabes-mots et les mots. Nous nous attendons à ce que les segments de type mot (ou les segments qui préservent les frontières du mot et qui portent une quantité substantielle d'informations sémantiques) obtiennent de meilleurs résultats. Les unités de mots obtiennent en effet des résultats statistiquement meilleurs que BASELINE pour FLICKR8K et COCO (R@1 = 5.4, +1.1pp et R@1 = 11.3, +2.3pp respectivement). Les syllabes-mots apportent également une amélioration significative (R@1 = 5.3 pour FLICKR8K et R@1 = 10.9 pour COCO), toutefois légèrement inférieure à celle obtenue avec les unités de mots.

Nos résultats montrent clairement que l'introduction d'informations sur les frontières à différentes couches a un impact substantiel sur les résultats : l'utilisation de ces informations à la première ou à la cinquième couche est inutile, car nous remarquons qu'elle donne des résultats similaires à ceux de BASELINE ($\text{GRU}_{\text{PACK.}}$–1) ou qu'elle détériore les résultats quel que soit le type de frontière utilisé ($\text{GRU}_{\text{PACK.}}$–5), alors que les résultats sont meilleurs lorsque ces informations sont introduites dans les couches intermédiaires.

7.4 Informations hiérarchiques

Dans les expériences susmentionnées, nous fournissons au réseau un seul type de frontière (soit phone, syllabe ou mot) mais pas plusieurs en même temps, comme si plusieurs unités ne pouvaient pas coexister en même temps. Pourtant, de multiples unités parlées existent

en même temps, et elles sont structurées de manière hiérarchique : les mots peuvent être décomposés en syllabes, qui peuvent à leur tour être décomposées en phones. Afin de modéliser cette nature hiérarchique de la parole, nous pouvons empiler autant de $\text{GRU}_{\text{PACK.}}$ que nous le souhaitons, où une couche gère un type de segment (par exemple, les phones) et la couche suivante du $\text{GRU}_{\text{PACK.}}$ gère un autre type de segment, qui est hiérarchiquement au-dessus du précédent (par exemple, des mots, voir Figure 6.2).

Nous explorons l'effet de l'utilisation d'une architecture hiérarchique sur l'ensemble de données Flickr8k. Contrairement à nos expériences précédentes, nous ne considérons que les architectures hiérarchiques qui utilisent des couches $\text{GRU}_{\text{PACK.}}$–keep. Nous testons plusieurs architectures qui gèrent différents types de frontières simultanément dont deux des cinq couches récurrentes sont des couches $\text{GRU}_{\text{PACK.}}$, puis dans l'expérience suivante, dont trois des cinq couches sont $\text{GRU}_{\text{PACK.}}$. Dans les deux cas, nous testons toutes les positions possibles ainsi que le type de frontière pour comprendre quelle est la meilleure combinaison.

7.5 2 couches GRU Packager

7.5.1 Phones et mots

En utilisant les phones et les mots ensemble, les résultats (Tableau 6.4) sont supérieurs à ceux de l'architecture de base et à ceux de l'architecture $\text{GRU}_{\text{PACK.}}$ à couche unique. En effet, nous obtenons un R@1 maximal de 8.2% lorsque nous utilisons $\text{GRU}_{\text{PACK.}}$ au niveau des couches 2 et 3, ce qui représente +3.9pp par rapport à l'architecture de base et +2.8pp par rapport à une architecture monocouche. Nous remarquons également que l'utilisation d'une architecture hiérarchique nous permet d'entraîner des réseaux moins profonds tout en améliorant les résultats par rapport à notre architecture de base.

7.5.2 Phones et syllabes

Les résultats que nous obtenons en utilisant des phones et des syllabes sont présentés dans le tableau 6.6. Ici aussi, nous remarquons que les résultats sont meilleurs que ceux de la baseline (sans aucun $\text{GRU}_{\text{PACK.}}$) et également meilleurs que lorsque nous utilisons un seul $\text{GRU}_{\text{PACK.}}$: $R@1 = 7.9$, +2.5pp. Contrairement à l'expérience précédente, l'architecture à 4 couches converge mieux que l'architecture à 5 couches. Néanmoins, les résultats sont inférieurs à ceux obtenus en utilisant conjointement les phones et les mots (−0.3pp), ce qui indique que l'utilisation conjointe des phones et des syllabes n'est pas la combinaison idéale.

7.5.3 Syllabes et mots

Les résultats que nous obtenons en utilisant des syllabes et des mots sont présentés dans le tableau 6.7. Une fois de plus, nous remarquons que R@1 est supérieur à une architecture à une couche $\text{GRU}_{\text{PACK.}}$ ($R@1 = 7.6, +2,2$pp) mais pire qu'une architecture à deux couches traitant les phones et les mots : $-0,6$pp. Le meilleur résultat est également inférieur au meilleur résultat obtenu en utilisant des phones et des syllabes : −0.3pp. Cependant, nous observons que l'architecture à deux couches est plus performante que la baseline (+1.0pp) et plus performante que l'architecture à deux couches qui utilise conjointement les phones et les syllabes (+0.7pp).

7.5.4 Conclusion

Nos expériences montrent que le meilleur résultat en utilisant deux couches $\text{GRU}_{\text{PACK.}}$ est obtenu en utilisant conjointement les frontières des phones et des mots, lorsque les

couches $\text{GRU}_{\text{PACK.}}$ sont placées au milieu de la pile de cellules récurrentes. Nous observons également que, globalement, nos résultats atteignent leur maximum lorsque les couches se suivent. Enfin, cette série d'expériences nous a permis de montrer que le réseau converge mieux lorsque les segments de bas niveau (phones) et les segments de haut niveau (mots) sont utilisés conjointement. Nous expliquons cela par le fait que cela permet au modèle d'apprendre des représentations robustes pour les phones, tout en ayant des unités de haut niveau qui portent beaucoup d'informations sémantiques. Par conséquent, l'utilisation de segments intermédiaires tels que les syllabes n'est pas utile car elles sont trop longues pour apprendre une représentation cohérente tout en étant trop courtes par rapport à la quantité d'informations sémantiques qu'elles portent.

7.6 3 couches GRU Packager

Enfin, nous avons intégré trois niveaux de segments dans un seul modèle. Comme dans nos expériences précédentes, nous expérimentons avec un nombre différent de couches (de 3 à 5), avec à chaque fois 3 couches $\text{GRU}_{\text{PACK.}}$ à chaque position possible. Les résultats de cette expérience sont présentés dans le tableau 6.8.

Nous observons que le meilleur résultat obtenu avec cette architecture ($R@1 = 9.6$) est bien meilleur que la baseline (+5.3pp), supérieur au meilleur résultat d'une architecture à une seule couche (+4.2pp) mais aussi supérieur au meilleur résultat d'une architecture à deux couches (+1.4pp sur l'architecture phonème-mot). Notre meilleur résultat est obtenu par une architecture à cinq couches avec $\text{GRU}_{\text{PACK.}}$ en position 1, 3 et 4. Bien que nous remarquions que le R@1 s'est amélioré lors de l'ajout d'un autre niveau de frontière dans cette dernière expérience, nous observons également que le saut dans les résultats n'est pas aussi important que ce que nous avons observé auparavant. En utilisant un seul $\text{GRU}_{\text{PACK.}}$, nous observons une amélioration de +1.1pp dans R@1 par rapport à notre architecture de base. En utilisant deux $\text{GRU}_{\text{PACK.}}$, nous observons une amélioration de +2.8pp par rapport à l'utilisation d'un $\text{GRU}_{\text{PACK.}}$. Enfin, en utilisant un troisième $\text{GRU}_{\text{PACK.}}$, nous avons observé une amélioration de +1.4pp par rapport à l'utilisation de deux $\text{GRU}_{\text{PACK.}}$. Ainsi, même si l'introduction de plus de structure dans le réseau est bénéfique, nous observons également que certains niveaux sont plus critiques que d'autres et ont un effet plus important sur le résultat final.

7.7 Conclusion du chapitre

Dans ce chapitre, nous avons étudié l'impact de la segmentation préalable de la parole dans un modèle PVC. Nous avons présenté une méthode simple pour introduire des informations de frontière dans n'importe quelle couche récurrente de notre architecture. Pour ce faire, il suffit de réinitialiser l'historique du RNN chaque fois qu'il y a une limite de segment. Nous avons constaté que la segmentation de la parole en sous-unités est utile : la segmentation en mots donne de meilleurs résultats que la segmentation en phonèmes, mais nous observons également que la segmentation en syllabes donne des résultats similaires à la segmentation en mots. Néanmoins, la segmentation en mots semble être une représentation plus robuste. Nous avons observé des résultats différents selon le niveau auquel les informations de segmentation sont introduites. Nous avons observé des effets négatifs si l'information de frontière est introduite trop tard (dernière couche de notre architecture).

Néanmoins, même si l'introduction d'informations sur les frontières est utile, elle n'améliore que légèrement les performances du réseau. Ce n'est que lorsque différents niveaux sont combinés que les performances du réseau atteignent leur maximum. Nos expériences révèlent qu'une structure qui utilise des segments de bas niveau (c'est-à-dire

des phones) conjointement avec des segments de haut niveau (c'est-à-dire des mots) est meilleure que l'utilisation de segments qui sont plus ou moins de la même taille (c'est-à-dire l'utilisation conjointe de syllabes et de mots).

8 Conclusion

8.1 Résumé des contributions

Dans cette thèse, nous avons étudié un modèle récurrent de parole visuellement contextualisée. Notre objectif était d'analyser les représentations apprises par notre modèle afin de mieux comprendre quelles connaissances linguistiques les modèles neuronaux sont capables d'acquérir de manière non supervisée. Nous avons comparé ces représentations à ce qui est connu du traitement de la parole humaine. Plus spécifiquement, nous nous sommes concentrés sur l'acquisition lexicale et avons trouvé des points communs entre les processus à l'œuvre dans les modèles que nous avons étudiés et les processus rapportés dans la littérature sur l'acquisition du langage chez l'enfant.

Plus précisément, les principales contributions de cette thèse sont :

- **Le jeu de données image/parole STAIR.** Nous avons introduit le jeu de données image/parole STAIR qui est basé sur le jeu de données image/texte STAIR (Yoshikawa et al. 2017). Cet ensemble de données constitue l'équivalent japonais de l'ensemble de données "Synthetically spoken COCO" (Chrupała et al. 2017*a*) pour l'anglais et permet d'entraîner des modèles de parole visuellement contextualisée.

- **Analyse de l'attention.** Nous avons montré que les modèles basés sur les RNN sont capables de détecter des motifs récurrents spécifiques dans l'entrée acoustique. Les modèles que nous avons entraînés se concentrent spécifiquement sur les mots concrets tels que les noms, car ils se réfèrent à des objets qui sont particulièrement saillants dans les images. Nous avons observé ce comportement pour deux langues typologiquement distinctes, l'anglais et le japonais, montrant ainsi que ce comportement ne dépend pas d'une langue particulière. Nous avons conclu que les modèles que nous avons étudiés présentaient une préférence pour les noms, comme celle que l'on trouve également chez les humains au cours du processus d'acquisition lexicale. Nous avons mis en évidence que les modèles japonais ont appris à détecter et à mettre en évidence des particules (comme la particule "ga"), et montrent ainsi un comportement spécifique à la langue. Ce faisant, les modèles ont adopté le même comportement que les enfants japonais afin de segmenter le flux de parole. La capacité des réseaux à mettre en évidence les noms — et les particules pour le japonais — est un comportement que les réseaux ont appris à adopter avec très peu d'exemples — moins de 500 paires image/légende — montrant que le réseau apprend rapidement quelles sont les parties les plus importantes des légendes.

- **Analyse de la connaissance des mots isolés.** Nous avons montré que le réseau était capable d'associer des noms individuels à leurs référents visuels corrects. Cela suggère que le réseau segmente implicitement son entrée en sous-unités, puis les associe à un contexte visuel. Cependant, nous avons observé que le réseau n'était pas capable d'associer tous les noms isolés à leurs référents visuels de manière égale ce qui suggère que le lexique du modèle est limité à un ensemble spécifique de mots. Nous avons observé que ce phénomène était principalement dû à la fréquence du mot dans la légende : plus une forme de mot est fréquente, mieux le modèle la relie à son référent visuel.

  Nous avons ensuite étudié comment le réseau active un mot isolé et l'avons comparé aux modèles d'activation et de reconnaissance des mots chez l'homme. En utilisant un équivalent algorithmique du paradigme du *gating* de Grosjean (1980, 1985), une méthodologie issue de la psycholinguistique, nous avons pu observer que l'activation

des mots ne se produit pas de manière linéaire, mais évolue plutôt par étapes. Ceci nous a permis de conclure que le modèle était capable de reconnaître un mot à partir d'une entrée partielle, c'est-à-dire avant la fin de celui-ci. Nous avons montré que le réseau devait nécessairement avoir accès au premier phonème d'un mot pour pouvoir activer la représentation du mot cible. Ce résultat est similaire à ce qui est postulé dans le modèle COHORT de reconnaissance de la parole, où le début des mots est d'une importance particulière pour activer et reconnaître un mot. Nous avons cherché à savoir si la reconnaissance des mots s'effectuait par un processus d'activation simultanée d'une cohorte de mots, qui seraient ensuite en compétition pour la reconnaissance. Nous avons trouvé des preuves d'activations simultanées pour certains mots et de compétition entre eux, cependant, ce processus semble loin d'être systématique, car nous avons montré que certains mots sont activés sans entrer en compétition avec des mots à consonance similaire.

- **Introduction d'informations linguistiques préalables.** Enfin, nous avons cherché à savoir si l'introduction d'informations linguistiques préalables sous la forme d'informations sur les frontières de segments était bénéfique. Nous avons en effet constaté que le réseau pouvait utiliser de manière adéquate de telles informations, en particulier lorsque le réseau recevait des frontières de mots. Le réseau utilise de manière adéquate les informations sur les limites des phonèmes et des syllabes, mais moins que les frontières des mots. Plus important encore, nous avons constaté que la prise en compte de la nature hiérarchique de la parole, en donnant simultanément au réseau les frontières des phonèmes, des syllabes et des mots, donnait des résultats encore meilleurs.

Dans son livre, Bloom (2002, p. 60) soutient que "les enfants apprennent le sens des mots par la théorie de l'esprit. Si cela est vrai, alors une mise en œuvre connexionniste directe de l'apprentissage des mots, dans laquelle les sons sont associés à des percepts, est irréalisable. (Et cela exclut toutes les théories connexionnistes de l'apprentissage des mots dont [il] a connaissance)". Nous pensons que les expériences que nous avons menées dans cette thèse, ainsi que les travaux précédents de Harwath et al. (2016), Harwath & Glass (2017), Chrupała et al. (2017*a*), Merkx et al. (2019) (entre autres) montrent que les modèles purement connexionnistes sont capables d'associer directement les sons aux percepts, ici sous la forme de représentations vectorielles de stimuli visuels. Par conséquent, les modèles connexionnistes sont capables d'apprendre des mots dans une certaine mesure. Bien entendu, nous ne prétendons pas que l'acquisition lexicale chez l'enfant se fait uniquement *via* un mécanisme purement associatif, mais il se pourrait qu'un mécanisme d'apprentissage purement associatif permette d'amorcer l'acquisition lexicale chez l'enfant. Un argument contre ce fait pourrait être que les approches connexionnistes nécessitent de grandes quantités de données pour être entraînées efficacement. Cependant, nos expériences montrent que nos modèles ont appris à se concentrer sur des noms spécifiques avec un très petit nombre d'exemples, ce qui suggère que l'amorçage associationniste constitue un mécanisme viable pour acquérir un lexique.

Les expériences que nous avons menées dans cette thèse nous permettent de conclure que les modèles PVC segmentent implicitement leur entrée en sous-unités et associent ces sous-unités à leur référent visuel. Ce processus ne semble émerger que comme un sous-produit de leur tâche principale qui est de minimiser la distance entre un stimulus acoustique et un stimulus visuel correspondant. Notre conclusion est en accord avec le travail très récent de Khorrami & Räsänen (2021).

8.2 Futurs travaux

Compte tenu des travaux que nous avons menés dans cette thèse, plusieurs travaux futurs pourraient être réalisés :

- **Activation des mots dans les modèles basés sur les CNN**. Nous avons étudié comment les modèles basés sur les RNN stockent les unités lexicales et activent la représentation des mots individuels en utilisant le paradigme de *gating*. Cette méthodologie pourrait également être appliquée pour analyser les représentations apprises par les modèles basés sur les CNN (tel que celui de Harwath & Glass (2017)) afin de comprendre comment la représentation d'un mot donné est activée dans ces modèles. Un modèle CNN tel que Harwath & Glass (2017) nous permettrait de reproduire directement les expériences linguistiques qui mesurent la reconnaissance et la compétition des mots à l'aide de dispositifs d'oculométrie tels que celui de Huettig & McQueen (2007).

- **Image-To-Speech.** Nous aimerions également étudier comment des réseaux Image-To-Speech (tels que ceux de Hsu et al. 2020 et Wang et al. 2020), qui produisent de la parole à partir d'une image d'entrée, apprennent progressivement à produire leurs premières phrases, et comparer cette évolution à celle des enfants. En effet, lorsque les enfants apprennent à parler, ils ne commencent pas par prononcer des phrases complètes, mais plutôt par produire des mots isolés. Plus tard, ils produisent des phrases de deux mots et, à partir de ce moment, ils commencent à produire des phrases complètes. Il serait intéressant d'étudier si un modèle Image-To-Speech passe par les mêmes étapes que les enfants et, dans le cas contraire, d'en étudier les raisons. De même, une telle expérience permettrait de mieux comprendre le développement linguistique des enfants en étudiant les représentations apprises par un tel modèle.

- **Segmentation discrète.** Les expériences menées dans le chapitre 6 révèlent que le fait de donner une segmentation explicite améliore la capacité des modèles à faire correspondre correctement une légende parlée à son contexte visuel. Il serait donc souhaitable que le réseau apprenne à segmenter *explicitement* l'entrée parlée en sous-unités. Plusieurs options, telles que celles proposées par Kreutzer & Sokolov (2018), Chen et al. (2019) et Shain & Elsner (2020) sont à explorer.

En conclusion, les modèles neuronaux de parole à base visuelle offrent des opportunités inestimables pour étudier et tester des hypothèses sur l'acquisition du langage chez l'enfant, grâce à leur capacité à modéliser des interactions complexes à travers plusieurs modalités. De nouveaux ensembles de données, tels que l'ensemble de données SEEDLingS (Bergelson & Aslin 2017) ou l'ensemble de données récemment collecté par Tsutsui et al. (2020), qui sont des enregistrements à grande échelle recueillis dans des environnements écologiques (Dupoux 2018), permettront aux chercheurs de simuler l'acquisition du langage avec des données plus réalistes que jamais.