Acquiring sounds and meaning jointly in early word learning

Abdellah Fourtassi

ENS

Thèse de Doctorat

En vue de l'obtention du grade de

DOCTEUR
DE L'ÉCOLE NORMALE SUPÉRIEURE

École doctorale

Cerveau Cognition Comportement (ED 158)

Discipline ou spécialité :
Sciences Cognitives

Présentée et soutenue par :
Abdellah Fourtassi

le 12/12/2015

Titre

Acquiring Sound and Meaning Jointly in Early Language Learning

---

Unité de recherche LSCP, ENS Ulm

Thèse dirigée par Emmanuel Dupoux

Membres du jury

Henny Yeung (rapporteur)

Robert Daland (rapporteur)

Janet Werker

Michael Frank

Marco Baroni

Christian Lorenzi

Numéro identifiant de la Thèse :

Contents

List of Tables **vi**

List of Figures **vii**

Introduction **xii**

I State of the Art **1**

1 The mechanisms of phoneme learning **3**

1.1 What is a phoneme? 4

1.2 Infants learn phonemes 6

1.2.1 Perceptual approach 6

1.2.2 Functional approach 7

1.3 The mechanisms of phoneme learning 9

1.3.1 Minimal-pair-based learning 9

1.3.2 Distributional learning 10

1.3.3 Word-form based mechanism 12

1.3.4 Complementary distribution 14

1.4 This study 15

1.4.1 Semantic cues 15

1.4.2 Statistical learning and early semantic representation 17
1.4.3 Word co-occurrence as a proxy for the general context 19
1.4.4 The distributional hypothesis 21
1.4.5 The proposed learning mechanism 22

II Computational experiments **25**

2 Datasets and Representations **27**

2.1 Datasets and the phonemic representation 28
2.1.1 The Corpus of Spontaneous Japanese (CSJ) 29
2.1.2 The Buckeye Corpus of Spontaneous Speech 32
2.2 Input representations 34
2.2.1 Early phonetic representation 34
Random allophones 37
HMM-based allophones 38
2.2.2 Early word-form representation 41
2.2.3 Early semantic representation 46

3 Phoneme learning 1: modeling perceptual reorganization **50**

3.1 Cues to phonemicity 52
3.1.1 Acoustic similarity 52
3.1.2 Word-form similarity cue 52
3.1.3 Semantic similarity cues 54
3.2 Task and Evaluation 55
3.3 Experiment 1 57
3.3.1 Results 58
3.3.2 Discussion 59

3.4 The scope of top-down cues 60
3.4.1 Invisible pairs and the phonemic status 61
3.4.2 The proportion of invisible pairs in natural speech 63
3.5 Experiment 2 64
3.5.1 Results 65
3.5.2 Discussion 66
3.6 General discussion 68

4 Phoneme learning 2: the number of categories 71
4.1 Problem specification 72
4.1.1 Learning as optimization 72
4.1.2 Semantic Consistency 74
4.2 Representations 75
4.2.1 Levels of phonological analysis 75
4.2.2 Induced lexicon 76
4.2.3 Semantic Consistency 77
4.3 Experiments 78
4.3.1 Experiment 1: Random partitioning 78
4.3.2 Experiment 2: nested-category-based partitioning 82
4.4 General discussion 87

III Human experiments 91

5 Learning semantic similarity through word co-occurrence 93
5.1 Introduction 94
5.2 Word co-occurrence and semantic similarity 95
5.3 Zero-shot learning 96

5.4 Method 97
5.5 Results and Analysis 102
5.6 Discussion 103

6 Semantic similarity modulates the phonemic status of phones 107

6.1 Introduction 108
6.2 Semantic similarity and the phonemic status 109
6.3 Method 111
6.4 Results and Analysis 113
6.5 Discussion 116

7 Conclusion and Future work 120

A Pre-processing of the CSJ corpus 124

B HTK phonetic decision tree 126

C Allophones and sensitivity to frequency and variation 129

D How HTK-based allophones affect the phonemic classification 132

E Levels of phonetic clustering 136

F Instructions to participants 139

G Extension to infant studies 140

H Whyisenglishsoeasytosegment? 144

I A corpus-based evaluation method for DSMs 145

References 146

List of Tables

2.1 Characteristics of phonemically transcribed corpora in English and Japanese 36
2.2 the number of word types in the allophonic representation relative to the phonemic representation, and the corresponding average number of tokens per type, as a function the average number of allophones per phoneme. . . . 44

3.1 The AUROC of the acoustic cue as a function of the average number of allophones per phoneme. 58
3.2 The AUROC of the word-form similarity cues as a function of the average number of allophones per phoneme, using both artificial allophones and HTK-based allophones, and using the gold segmentation. 58
3.3 The AUROC of the semantic cues as a function of the average number of allophones per phoneme, using both artificial allophones and HTK-based allophones, and using the gold segmentation. 59
3.4 Proportion (in %) of invisible allophonic contrasts out of the total number of allophonic contrasts. 61
3.5 The proportion (in %) of invisible pairs as a function of the average number of allophone per phoneme, generated by HTK. 63
3.6 The proportion (in %) of visible pairs as a function of the size of the corpus, in the case of 2 allophones/phoneme in average (generated using HTK) . . 63

B.1 Questions on contexts used in HTK's decision-tree state-tying procedure. Japanese data 127
B.2 Questions on contexts used in HTK's decision-tree state-tying procedure. English data. 128

List of Figures

0.1 In the "received timeline" of language learning, sounds are learned first, and meaning next. Sounds can possibly influence word meaning learning, but not the other way around. In a more realistic timeline, sounds and meanings are learned in parallel, and can influence each other throughout the learning process. xiii

1.1 Pitta-Patta language contains three vowels. Two features are therefore required to distinguish them (±high, ±round). Capital letters represent vowels that are specified only for minimally contrastive features. Remaining features required for pronunciation are supplied by a set of phonetic realization rule. From Dresher (2011) 5
1.2 Data representing exemplar distribution of three hypothetical categories. From Pierrehumbert (2003) 5
1.3 Upon familiarization with either a Bimodal or a Unimodal distribution along the continuum of [da]-[ta], infants of 6 months of age were able to discriminate this contrast in the first case but not in the second one. From Maye et al. (2002) 10
1.4 Vowel distribution in a formant space. Vowels values were collected from 46 men, 48 women and 46 children. Notice the high degree of overlap between vowel categories. From Hillenbrand et al. (1995) 11
1.5 Both Russian and Korean production data show a bimodal distribution along the VOT dimension. This bimodal distribution corresponds to two phonemic categories in Russian, and to one phonemic category in Korean. From Kazanina et al. (2006) 12
1.6 Upon familiarization with a non-native contrast paired with two objects, only infants in the consistent pairing group succeeded in the subsequent discrimination test. From Werker et al. (2012) 16
1.7 An illustration of a putative early semantic representation of the word "kitchen", including co-occurring objects, words and time periods. 19

1.8 schematic illustration of the proposed learning mechanism. Learning is understood to be the outcome of an interaction between three levels of linguistic representation. The phonetic level is composed of fine-grained categories. This information percolates into the lexical level where word-forms are extracted and stored along the granularity of the phonetic level. The semantic level represents the distribution of these word-forms according to their co-occurrence in conversations. Through a feedback loop, the semantic level readjusts the phonetic level along the relevant (i.e., phonemic) dimensions . 22
1.9 Allophonic rule of the French uvular fricative 23

2.1 Schematic description of the computational part of the dissertation. It includes a description of the way the representations are derived, and a global indication of how the proposed algorithm operates on these representations. 29
2.2 The consonants of Japanese as used in this study. The symbol to the right represents the voiced version of the consonant. 31
2.3 The vowels of Japanese as used in this study. 31
2.4 Frequency distribution of phonemes in the CSJ Corpus. 32
2.5 The consonants of English, as used in the Buckeye corpus. The symbol to the right represents the voiced version of the consonant. The symbol between parentheses represents the syllabic version of the consonant (analysed as phonemes in the Buckeye corpus). The ARPA transcription is used for ease of reading. 33
2.6 The vowels of English, as used in the Buckeye corpus. Each vowel V is accompanied with a lexical CVC example. The diphthongs are analysed as single phonemes. The ARPA transcription is used for ease of reading. . . . 34
2.7 Frequency distribution of phonemes in the Buckeye Corpus 35
2.8 An allophonic rule of the French uvular fricative 36
2.9 Example of rule application on the utterance “atama ga itai” (my head hurts), using random rules which assign each phoneme one of two allophones. (from Martin et al. 2013) 38
2.10 Example of a phonetic decision tree used in the state-tying procedure of HTK. From young et al. (2006) 40
2.11 Schematic description of modeling early word segmentation. For ease of presentation, I only show the case of the phonemic representation. The allophonic representation at different levels of complexity is obtained by replacing each phoneme with the allophone that fits the particular context where this phoneme occurs 42
2.12 Negative log posterior probability (lower is better), as a function of iteration, for corpora at different levels of allophonic complexity. I used a collocation adaptor grammar with Pitman-Yor adaptors and an incremental initialization. 43
2.13 Token F-score (higher is better) as as a function of iteration, for corpora at different levels of allophonic complexity. I used a collocation adaptor grammar with Pitman-Yor adaptors and an incremental initialization. . . . 43

2.14 F-scores of optimal segmentations as a function of the average number of allophones per phoneme for English and Japanese data, using a collocation adaptor grammar model and a random segmenation as a control. 45
2.15 LSA takes as input a matrix consisting of rows representing word types, and columns representing contexts. The values correspond to the number of times a word is uttered in a given context. A matrix reduction operation (Singular Value Decomposition) is performed to obtain a compact semantic space. The semantic distance of two words in the resulting space is given by the angle formed by their vectors. 46
2.16 SDT-ρ as a function of the number of utterances to be taken as a unit of context, averaging over values of semantic dimensions ranging from 5 to 500. 49
2.17 SDT-ρ as a function of the number of semantic dimensions, averaging over values of context size ranging from 5 utterances to 500 49

3.1 An illustration of a typical distribution of allophonic/phonemic contrasts, according to a given indicator of phonemicity. The evaluation of the classification depends on where the threshold is set. On the left, we have a rather "conservative" threshold, and on the right we have a rather "liberal" threshold. 56
3.2 Three binary classifiers that vary in their quality from bad (left) to great (right) 57
3.3 The ROC curves corresponding to the three classifiers above. The hatched area represents the Area Under the ROC curve (AUROC) of the 'bad' model. Adapted from Weiss (2008). 57
3.4 Example of an allophonic rule 62
3.5 The AUROC of top down cues as a function of the average number of allophones per phoneme, and as a function of the quality of the segmentation: ideal, unsupervised and random 65
3.6 The AUROC of top down cues as a function of the size of data available to the learner, in the case of 2 allophones per phoneme, and as a function of the quality of the segmentation: ideal, unsupervised and random 67
3.7 The AUROC of top down cues as a function of the size of data available to the learner, in the case of 4 allophones per phoneme, and as a function of the quality of the segmentation: ideal, unsupervised and random. 67

4.1 Upon hearing the sound "cat", the English-learning infant can a priori represent it through phonetic categories at different resolutions (different phonological analyses). The challenge is to select the "optimal" level of representation 76
4.2 Token F-score of optimal segmentations of English and Japanese corpora transcribed with inventories of different sizes, using a collocation adaptor grammar model. The red color refers to the phonemic inventory. 77
4.3 Schematic description of sub-type derivation and SC score computation, using random partitioning. 79

4.4 Semantic Consistency scores across different phonetic inventories and different levels of word segmentation, using Random Partitioning. The white points and error bars show the means and standard errors over different parameter settings. The black points refers to the phonemic inventory of each language 80
4.5 Histogram of the SC score peaks across different parameter settings, using Random Partitioning. The red arrows point towards the phonemic inventory. 81
4.6 Semantic Consistency score of lexicons represented with different phonetic inventories, and for corpus sizes ranging from 100% (about 50.000 utterances) to 0.1% (50 utterances). The SC score is computed using random partitioning, and the ideal word segmentation. The white points refer to the individual scores using different parameter settings, the red points to the phonemic inventory, and the blue to the means of the individual scores. . . 83
4.7 Schematic description of sub-types' derivation and Semantic Consistency score computation, using nested-category-based partitioning. 84
4.8 Semantic Consistency score of lexicons represented with different phonetic inventories and ideal segmentation, using nested-category partitioning. The points and error bars show the means and standard errors over different parameter settings. The red colors refers to the phonemic inventory of each language. 85
4.9 Histogram of the SC-score peaks across different parameter settings, using nested-category partitioning. The red arrows point towards the phonemic inventory. 86
4.10 Semantic Consistency score of lexicons represented with different phonetic inventories, and for corpus sizes ranging from 100% (about 50.000 utterances) to 0.1% (50 utterances). The SC score is computed using nested-category partitioning, and the ideal word segmentation. The white points refer to the individual scores using different parameter settings, the red points to the phonemic inventory, and the blue to the means of the individual scores . . . 88

5.1 Referential familiarization. Participants are presented with multiple series of word-objects pairings. The objects belong to the category of animals or the category of vehicles. 98
5.2 Learning consolidation. Two-Alternative Forced Choice paradigm (2AFC), with feedback. 99
5.3 Distributional familiarization. Sequences of words are presented with no visual referents. Two new words ("guta" and "lita") are introduced and co-occur consistently with the words corresponding to one of the two semantic categories ("romu" and "komi" for the category of animals, and "nulo" and "pibu" for the category of vehicles) 100
5.4 Order of exposure of the experimental settings. Participants are trained referentially once (part 1 and part 2), distributionally twice (part 3). They are tested in three sessions (part 4): before and after each block of distributional learning 101

5.5 proportion of correct answers in filler condition (known words) and target condition (new words), before any distributional exposure (session 0) and after the first and second block of exposure (session 1 and 2) 103

6.1 Word learning as a mapping between a phonological category and a meaning category. Green circles refer to correct generalizations: the referents have a high semantic relatedness, therefore, variation is analyzed as allophonic. The red circles refer to wrong generalizations: the referents have a low semantic relatedness, therefore, variation is analyzed as phonemic 110
6.2 target word-object association between the minimal pair ("gutah"/"gutaw") and a pair of referents with different levels of semantic similarity. The first referent represents a cow, the second referent represents, respectively, another cow, a buffalo, a deer, a bird, and a car. 112
6.3 filler word-object pairings. "pibu" and "komi" were paired with a picture representing a house and a book. The pairing was kept the same across all groups. 113
6.4 Proportion of 'same' answers on same-trials of both the minimal pair used in the training (gutah- gutah, gutaw-gutaw) and the new minimal pair (litah-litah, litaw- litaw), as a function of the similarity of the two referents in training. The semantic similarity ranges from 1 (the most similar) to 5 (the least similar). The dotted line represents chance. 114
6.5 Proportion of 'different' on different-trials including both the minimal pair used in the training ("gutah" vs. "gutaw"), and the new minimal pair ("litah" vs. "litaw"), before and after the pairing with two referents with various degrees of semantic similarity, ranging from 1 (the most similar) to 5 (the least similar). The dotted line represents chance level 115
6.6 Proportion of 'different' answers on different-trials including both the minimal pair used in the training ("gutah" vs. "gutaw") and the new minimal pair ("litah" vs. "litaw"), as a function of the similarity of the two referents in training. The semantic similarity ranges from 1 (the most similar) to 5 (the least similar). The dotted line represents chance level 116

D.1 example of an allophonic rule 132
D.2 The number of allophones for each phoneme in English data (left) and Japanese data (right), in the case of 2 allophones per phoneme in average. 135

E.1 Hierarchical clustering of Japanese phonemes. 137
E.2 Hierarchical clustering of English phonemes. 138

G.1 An illustration of the proposed experimental dessign. 142

Introduction

Research in early language acquisition has—whether implicitly or explicitly—treated the processes of learning sounds and of learning meanings as a succession of two steps. According to the received view, babies first master the phonetic categories of their native language: they learn to ignore irrelevant variations in pronunciation (such as differences in talker, speech rate, emotion, and linguistic context) (Kuhl, 2004). Only then, according to this view, can they map sounds to meaning (Bloom, 2000).

The goal of this dissertation is, however, to investigate the possibility of interactions in learning phonemes and semantics in the early stages of language learning. In questioning the received view, we are supported by recent experimental evidence and computational studies. Developmental data shows, on the one hand, that infants do not wait to have completed the acquisition of phonemes to start learning meanings (Tincoff & Jusczyk, 1999; Bergelson & Swingley, 2012). On the other hand, the phonological representation continues developing beyond the age of perceptual attunement (Stager & Werker, 1997). Thus, the developmental trajectory of phonology and meaning overlaps. Infants do not wait to have completed one to start the other, rather, they learn the sound system and word meanings in a parallel fashion. This change in perspective, from sequential to parallel, suggests that phonology and semantics influence each other throughout the learning process (Figure 0.1).

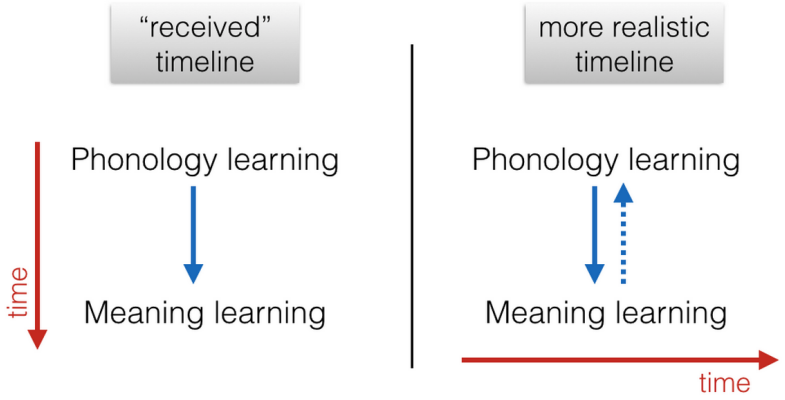

Figure 0.1: In the "received timeline" of language learning, sounds are learned first, and meaning next. Sounds can possibly influence word meaning learning, but not the other way around. In a more realistic timeline, sounds and meanings are learned in parallel, and can influence each other throughout the learning process.

Another reason why the received sequential approach is problematic, is that the speech input exhibits high variability and leads to massive ambiguity in the phonetic space (Hillenbrand, Getty, Clark, & Wheeler, 1995). When purely bottom-up clustering algorithms—that is, algorithms that try to find sounds using only the raw speech signal—are tested on realistic input, they systematically fail to learn the right categories (e.g., Varadarajan, Khudanpur, & Dupoux, 2008). In this study, I describe a new mechanism in which the process of learning sounds benefits from the top-down constraint of early semantics, albeit ambiguous and rudimentary.

This mechanism will be investigated through two complementary levels of analysis:

- Computational analysis of the **input**: the learning mechanism will be implemented and tested on data extracted from corpora of natural speech in two typologically different languages. This will allow us to quantify how much learning is *a priori* possible, based on cues available in the signal.

- Experimental test of the **learner**: having a rich input does not guarantee that

learners, with their inherent limitations and biases, are actually making use of this input. This part of the study will thus provide evidence for the *cognitive plausibility* of the proposed mechanism.

The structure of the dissertation

This study includes three parts: a review of the literature, a modeling study and an experimental test of human subjects. **The first part**, which corresponds to the first chapter, introduces the scientific question and situates the present work in the context of an ongoing research program in developmental psychology. Besides it states the main proposal of this dissertation: a new top-down mechanism to phoneme learning. **The second part** deals mainly with the computational level of analysis, i.e., the extent to which the input offers sufficient cues to support the proposed learning mechanism. It consists of the second, third and fourth chapters. **The third part** complements the second one, in that it provides evidence for the cognitive plausibility of the mechanism I propose, through testing humans with a controlled input. It consists of the fifth and sixth chapters.

Part I

State of the Art

This part of the dissertation corresponds to the first chapter. It reviews work on the acquisition of phonemes and early word meaning. The learning of these two linguistic levels have typically been considered separately, usually under the assumption that meaning cannot be fully mastered until phonemes are learned. Here I describe a learning mechanism that shows how even an early and ambiguous semantic representation can help with the acquisition of phonemes.

Chapter 1

The mechanisms of phoneme learning

1.1 What is a phoneme?

The concept of 'phoneme' corresponds to the intuition of a "smallest contrastive linguistic unit which may bring about a change of meaning" (Gimson, 1962). For example, difference in meaning between the word *right* and the word *light* results from the minimal exchange of the phoneme /r/ for the phoneme /l/. Despite the apparent simplicity of this notion, there are actually differing views within linguistics on the exact way phonemes should be defined, and controversies about their attributes, content, or even their ontological status (Dresher, 2011). In fact, while some scholars such as Jones (1967) see the phoneme as a physical entity, a 'family' of sounds that "count for practical purposes as if they were one and the same", others considers it, rather, as a psychological unit (or in modern terms, a 'mental representation'), distinct from the sounds that it represents.(e.g., Trubetzkoy, 1939; Sapir, 1933). A different group of researchers (e.g., Twaddell, 1935) thinks that, though it is admittedly a useful tool, the phoneme remains a fictitious unit with no underlying reality.

Sapir (1925) proposed to characterize phonemes in term of their contrastive properties. This idea was carried further by phonologists of the Prague school, leading to the notion of phonemic make-up (Jakobson) and phonemic content (Trubetzkoy), both of which describe the contrastive features necessary to distinguish a phoneme from others in the same linguistic system. Crucially, the contrastive representation of the phoneme is underspecified, and any physical realization of the phoneme requires the specification of additional features. An example of this representation is shown in Figure 1.1. Other researchers, e.g., 'Exemplar Theorists', posit, in contrast, that instances of speech sounds are stored in great detail, forming exemplar "clouds" of phonemes (e.g., K. Johnson, 1997; Bybee, 2001; Pierrehumbert, 2001, 2003) (Figure 1.2).

Researchers may disagree on whether to locate the phoneme in the signal or in the psyche. They may also disagree on the level of abstractness most appropriate to characterize the phoneme. However, it is undeniable that this unit enables us to efficiently describe a large

Underlying representations

/I/	/A/	/U/
[+high, −round]	[−high]	[+high, +round]

Some realization rules

i. [ ] → [+tense] / __ in an open syllable

ii. [ ] → [−tense] / __ in a closed syllable

Figure 1.1: Pitta-Patta language contains three vowels. Two features are therefore required to distinguish them (±high, ±round). Capital letters represent vowels that are specified only for minimally contrastive features. Remaining features required for pronunciation are supplied by a set of phonetic realization rule. From Dresher (2011)

Figure 1.2: Data representing exemplar distribution of three hypothetical categories. From Pierrehumbert (2003)

number of words with a small set of segments. For instance, English language contains more than 170.000 lexical items according to the second edition of the 20-volume Oxford English Dictionary. These words can be generated with an inventory of about 40 phonemes, only. World languages contain around 800 phonemes in total (Ladefoged, 2001), and each makes use of a subset (e.g., there are only 11 phonemes in Piraha, and about 140 in !Xũ).

1.2 Infants learn phonemes

Acquiring their native language requires the learners to narrow down on the relevant subset of sounds. English learners, for instance, have to learn the distinction between the phoneme /l/ and the phoneme /r/ to differentiate minimal pairs such as *light* and *right*. In contrast, Japanese learners need not differentiate these sounds, which do not bring about difference in word meaning in their language. Similarly, English learners need not differentiate the aspirated and unaspirated allophones of the phoneme /p/, which correspond to the sounds occurring, respectively, in the first segment of the word *pin* (phonetically noted as [p]), and the second segment of the word *spin* (phonetically noted as [p^h]). Thai and Korean learners, in contrast, have to pay attention to this contrast since it might change the meaning of words. This can be exemplified by the Korean lexical minimal pair /p^hul/ ("grass") vs. /pul/ ("fire"). Psycholinguistic studies show, indeed, that Japanese-speaking adults have difficulties discriminating the English [r] vs. [l] contrast (Miyawaki et al., 1975), and English-speaking adults have difficulties perceiving the difference between [p] and [p^h] sounds used in Thai (Lisker & Abramson, 1970).

Language acquisition research has devoted a lot of effort to study phoneme learning, since it is believed to lie the foundation for later stages of linguistic development (such as word learning). It was investigated through at least two main approaches, as will be explained in the following.

1.2.1 Perceptual approach

The first line of research tries to characterize phoneme learning through the evolution of infants' perceptual sensitivity to native vs. non-native contrasts. It has been initiated by the seminal work of J. Werker and Tees (1984), through the use of the Conditioned Head-turn Procedure. In an experiment of this kind, infants hear repeatedly one stimulus (e.g., the syllable *ba*). Every four to twenty repetitions, a different stimulus (e.g., *da*) is

presented. The occurrence of the new sound *da* is associated with the activation of a little toy animal. Babies are thus conditioned to turn their head to see the toy perform when they detect a change in the stimulus presentation. Using such an experimental setting, J. Werker and Tees (1984) compared the head-turn behavior of English babies from 6 to 12 months of age, as a reaction to the contrast /ba/-/da/, used in both English and Hindi languages, and the contrast /ta/-/ʈa/ (voiceless dental vs. retroflex stop), only used in Hindi language. All subjects succeeded in discriminating the /ba/-/da/ contrast. However, discrimination of the contrast /ta/-/ʈa/ varied as a function of age. 6 to 8-month old English-learning babies succeeded in discriminating the non-native contrast, but this discrimination declined by 10-12 months of age. Hindi Infants, in contrast, maintained the discrimination of the /ta/-/ʈa/ contrast. Many studies have replicated this finding (see Gervain and Mehler (2010) for a review).

These results were often interpreted as the proof that babies start with a universal phonetic sensitivity, and by the end of their first birthday, this sensitivity is maintained for native contrasts only. Recent findings show, nonetheless, that perceptual attunement is a more complicated process. For instance, some native contrasts are not just maintained, but are further enhanced by 1 year of age (Kuhl et al., 2006). Moreover, some difficult native contrasts require language exposure to be successfully discriminated (Narayan, Werker, & Beddor, 2010). Besides, the process of perceptual attunement continues well beyond the first year of life (Sundara, Polka, & Genesee, 2006).

1.2.2 Functional approach

The second line of research characterizes phoneme learning not through the babies' perception of a contrast, but through their ability to use this contrast. In fact, Stager and Werker (1997) showed that being able to discriminate sounds is not necessarily equivalent to being able (or willing) to use these sounds to learn words. This was first shown using an

experimental paradigm which has come to be called the "switch" task. In this paradigm, babies are first familiarized with two word-object pairings (referred to, hereafter, by the labels A and B) for many trials until their looking time drops to a habituation criterion. In the test phase, babies are presented with two types of trials. The 'same' trial consists of a correct pairing between a word and an object as in the familiarization phase (e.g., Word A with Object A), whereas the 'switch' trial consists of a wrong pairing (e.g., Word A with Object B). If subjects have correctly learned the association during the familiarization, they are supposed to be surprised by the 'switch' trial and not by the 'same' trial. This fact is quantified through measuring the looking time to the 'switch' relative to that of 'same'. J. Werker, Fennell, C.T., Corcoran, and Stager (2002) showed that babies under 17 months old are not able to notice the 'switch' when Word A and Word B are minimal pairs (e.g., *bih* vs. *dih*), even though they can perfectly discriminate the contrast on a purely perceptual level! More recent experiments showed, nonetheless, that younger infants do succeed in learning when the relevant phonetic dimension is highlighted during the familiarization (Thiessen, 2007; Rost & McMurray, 2009), when a salient contrast is used (Curtin, Fennell, & Escudero, 2009), when the referential context is made explicit (Fennell & Waxman, 2010), or when a more fine-grained testing paradigm is used (Yoshida, Fennell, Swingley, & Werker, 2009). Interestingly, success in the switch task was shown to correlate, within subjects, with the vocabulary size (J. Werker et al., 2002). This provides support for the hypothesis that learning the function of phonemes and their use in language requires experience with word learning, and not just passive exposure to meaningless speech (see J. Werker & Curtin, 2005). The functional approach has been particularly useful when perception was not, by itself, a sufficient indicator of learning, e.g., when the non-native contrast was perceptually salient. For example, Dietrich, Swingley, and Werker (2007) tested 18 month-old English- and Dutch-learning toddlers on a contrast phonemic in both languages (vowel quality), and a salient contrast that was phonemic only in Dutch (vowel length). They found that both

groups noticed the switch when the first contrast was used to differentiate the labeling words (*tam* vs. *tem*), but only Dutch babies looked significantly longer at the switch when the second contrast was used (*tam* vs. *taam*).

1.3 The mechanisms of phoneme learning

Even though the timeline of phoneme learning has been documented in detail over the past 30 years of research, little is known about the cognitive mechanisms involved in this learning. In what follows, I review the main theoretical proposals and discuss their scope and limitations.

1.3.1 Minimal-pair-based learning

The most intuitive learning mechanism consists in making use of lexical minimal pairs. For instance, the presence of pairs such as '*light* and *right*, which have different meanings, points towards the phonemic status of the contrast /r/ vs. /l/. This mechanism is indeed the standard method used in field linguistics to determine the segmental inventory of a given language (Pike, 1947). It has been proposed as a developmental mechanism to various degrees in the work of researchers such as Jakobson (1966), MacKain (1982), and Best (1993). However, inspection of developmental data shows that infants' lexicon contains very few, if any, minimal pairs during the time period of phoneme development (generally within the first year and a half).

The babies' lexicon is generally probed through the MacArthur-Bates communicative Development inventory (CDI), which consists of parental/caregiver report on the words that infants understand and simultaneously produce (Dale, 1996). Based on a cross-linguistic analysis of CDI from English- and Italian-learning babies between 8 and 16 months of age, Caselli et al. (1995) found that the first 50 words that English-learning babies are most likely to understand contain no minimal pairs. For Italian babies, they found only two minimal

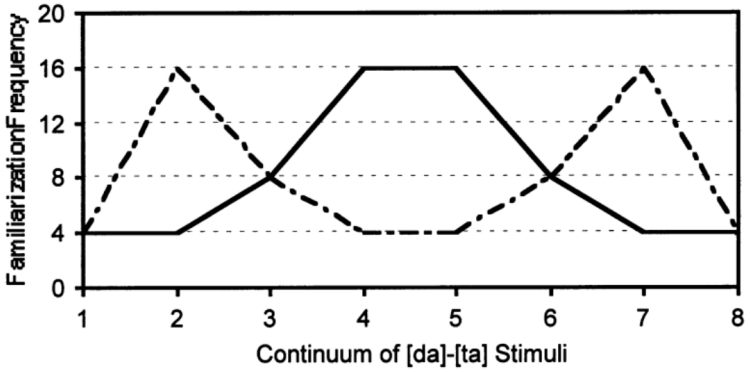

Figure 1.3: Upon familiarization with either a Bimodal or a Unimodal distribution along the continuum of [da]-[ta], infants of 6 months of age were able to discriminate this contrast in the first case but not in the second one. From Maye et al. (2002)

pairs, *nonno* ("grandpa") vs. *nonna* ("grandma"), and *nonna* vs *nanna* ("sleep/bedtime"). A similar analysis of CDI from 18-month-old Dutch-learning babies conducted by Swingley and Aslin (2007) revealed that about 78.6% of the words in their receptive vocabulary have no phonological neighbors.

1.3.2 Distributional learning

Researchers have therefore directed their attention towards bottom-up mechanisms which do not necessarily require substantial lexical knowledge. The most studied bottom-up mechanism is based on the observation that speech input shows distributional patterns that tend to correlate with the phonemic status. For instance, J. Werker et al. (2007) analysed child directed speech of both English and Japanese mothers, and found that vowels' distributions have a bimodal shape along the dimension of 'duration' in Japanese, and along the dimension of 'color' in English. These bimodal distributions correspond to the correct phonemic split in each language. If infants could track down such statistical distributions, they would be able to infer when a sound contrast is phonemic (bimodal distribution), and when it is not (unimodal distribution). In fact, this was shown to be true in a laboratory experiment conducted by Maye, Werker, and Gerken (2002), as illustrated in Figure 1.3.

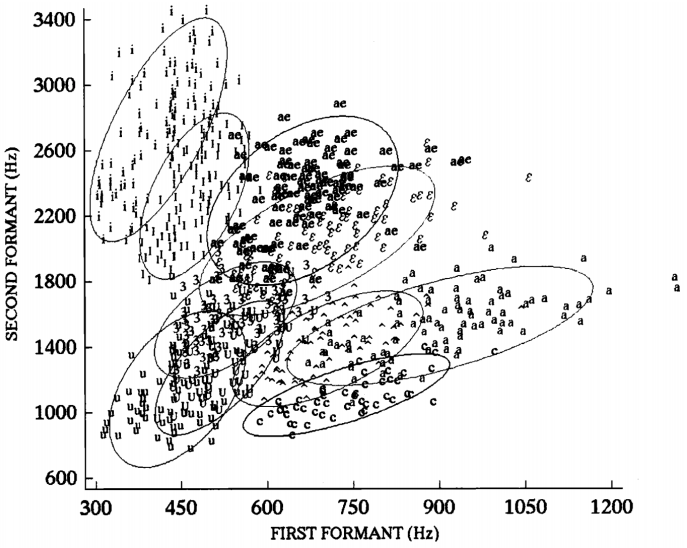

Figure 1.4: Vowel distribution in a formant space. Vowels values were collected from 46 men, 48 women and 46 children. Notice the high degree of overlap between vowel categories. From Hillenbrand et al. (1995)

Distributional learning cannot, however, account for the entire acquisition process. In fact, although phonemes that have sufficient separation in the acoustic space can be successfully recovered in a distributional fashion (e.g., McMurray, Aslin, & Toscano, 2009; Vallabha, McClelland, Pons, Werker, & Amano, 2007), some segments show substantial overlap between categories (Hillenbrand et al., 1995) and pose a significant challenge to this bottom-up mechanism, since it becomes difficult to tease apart unimodal from bimodal distributions in a reliable fashion (Figure 1.4). Another serious, and more fundamental, limitation of distributional learning is that the peaks or modes of distribution do not necessarily signal a phonemic split. For example, in both Russian and Korean, production data show a bimodal distribution of instances of [t] and [d] along the VOT dimension (Figure 1.5). However, only in russian does this contrast signal difference in word meaning as witnessed by the minimal pair *dom* ("house") vs. *tom* ("volume"). Korean speakers, in contrast, never use these sounds to contrast meaning, since they occur in a complementary distribution. Interestingly, Kazanina, Phillips, and Idsardi (2006) demonstrated that Rus-

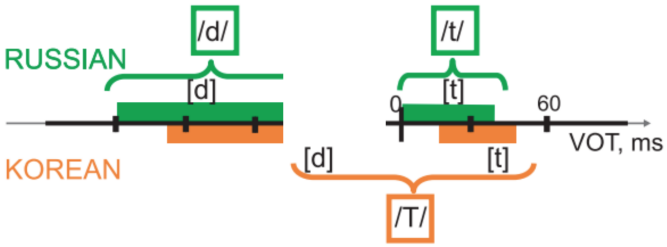

Figure 1.5: Both Russian and Korean production data show a bimodal distribution along the VOT dimension. This bimodal distribution corresponds to two phonemic categories in Russian, and to one phonemic category in Korean. From Kazanina et al. (2006)

sian adults showed a magnetic mismatch negativity (MMNm) upon hearing a sequence of [da] syllables followed by a [ta] syllable, whereas Korean adults did not show any significant perceptual discrimination of the [da]-[ta] contrast.

1.3.3 Word-form based mechanism

The limitations of the distributional approach as well as the minimal-pair-based approach have urged researchers to seek a middle way between, on the one hand, purely top-down mechanisms that rely on too advanced referential knowledge, and on the other hand, purely bottom-up mechanisms that suffer from ambiguity and lack functional insight. The use of the word-form (as opposed to word meaning) has been suggested to play the role of the missing link (e.g., Swingley, 2009). In fact, we know that infants begin to segment word-forms from continuous speech by as early as 6 months of age (Bortfeld, Golinkoff, & Rathbun, 2005), and there is evidence that 11-month-old babies store in their memory high frequent chunks that may or may not correspond to meaningful words (Ngon, Martin, Dupoux, Cabrol, & Peperkamp, 2013).

The argument behind word-form based mechanism can be articulated as follows. If, because of insufficient semantic knowledge, the learner cannot tell whether two similar word-forms are same or different, she can, in contrast, more easily infer that different word-forms represent different lexical items. For example, upon hearing the words *bat* and *bet*,

the learner might not be able to decide if *bet* is a variation (or mispronunciation) of *bat*, or if it is a distinct word. But she is more likely to decide that *bat* and *egg* are different words. This knowledge can be used to discriminate similar phonemes. Imagine, for instance, that tokens of the vowel categories /ae/ and /ɛ/ overlap in the acoustic space. The learner cannot decide a priori if they correspond to one or two categories, but since they occur in different word contexts, i.e., "b[ae]t" and "[ɛ]gg", the learner will be able to tease them apart. Indeed, a couple of laboratory experiments backs this learning mechanism. Thiessen (2007) exposed 15-month-old toddlers to phonemes embedded in either similar words or dissimilar words. Children who experienced phonemes in the latter condition were more successful in using the contrast between the phonemes in a word referent learning task. In another experiment that did not involve referential learning, Feldman, Myers, White, Griffiths, and Morgan (2013) familiarized 8-month-old infants to two phones that occur in either minimal pairs or non-minimal pairs. They showed that infants are sensitive to this word level information, since they were able to discriminate the phones only when they occur in a non-minimal pair context.

The word-form based mechanism suffers, nonetheless, from the following limitation. While in controlled laboratory experiments babies are exposed to phones exclusively in similar or dissimilar words, natural speech to which learners are exposed in a realistic learning environment includes, indistinguishably, both similar and dissimilar words. More precisely, lab experiments show that (non-)words like *dawgoo* and *tawgoo* (Thiessen, 2007) or *gutah* and *gutaw* (Feldman, Myers, et al., 2013) do not facilitate discrimination (or even impairs it, as suggested in the work of Martin, Peperkamp, and Dupoux (2013) and Fourtassi and Dupoux (2014)), whereas words like *dawboo* and *tawgoo* or *gutah* and *litaw* do facilitate discrimination. However, a more realistic setting would include *dawgoo* and *tawgoo* as well as *dawboo* and *tawgoo* in the case of Thiessen (2007), and *gutah*, *gutaw*, as well as *litah* and *litaw* in the case of Feldman, Myers, et al. (2013). Interestingly, discrimination was

reported to be significantly lower in this situation (Feldman, Myers, et al., 2013). The reason behind this effect (or absence of effect) is that minimal pairs tend to counteract the effect of maximal pairs. In fact, while the learners tend to differentiate two similar phonemes when they hear them in a non-minimal pair context, they will also tend to collapse them in one category when they hear them in a minimal pair context! In a computational study, Feldman, Griffiths, Goldwater, and Morgan (2013) found that, while the word-form mechanism greatly improves learning compared to the distributional learning mechanism, it also tends to produce errors in the presence of minimal pairs.

1.3.4 Complementary distribution

Peperkamp, Le Calvez, Nadal, and Dupoux (2006) proposed a cue to phonemicity based on a well documented fact in phonology, which states that allophones tend to be in complementary distributions. For example, the English vowel /æ/ is nasalised before nasals (e.g., [mæ̃n]) and realized as oral in all other situations (e.g., [mæd]). Therefore the allophones [æ] and [æ̃] are in complementary distribution. According to Peperkamp et al., the degree of distributional complementarity of any pair of phones correlates with the probability of them being allophones. They quantified the degree of complementarity between two phones through measuring the dissimilarity between their context distributions, using the classic information theory measure called Kullback-Leibler divergence (hereinafter, KL).

Peperkamp et al. (2006) found that this cue allows for a relatively successful learning when tested with simplified artificial corpora. However, when real corpora with realistic linguistic allophones were used, it performed badly. Similar results were reported in Boruta (2012) and Martin et al. (2013) where the performance of the KL measure degraded as the allophonic complexity increased, eventually approaching chance level. The main reason behind this failure is the fact that segments in real languages can be in complementary (or near complementary) distribution even if they are not allophones. This could be due to

constraints linked to the syllabic structure or the phonotactics. For example, in English, the phonemes /h/ and /ŋ/ occur in different syllable positions and, therefore, are in complementary distributions although they are not considered allophones in any phonological theory.

As regards the cognitive plausibility of the mechanism, White, Peperkamp, Kirk, and Morgan (2008) familiarized infants with consonant alternation in words (e.g., *bevi* vs *pevi*) with either a consistent triggering context (e.g., *bevi* only after *na*, and *pevi* only after *rot*), or an inconsistent context (*bevi* and *pevi* after both *na* and *rot*). They found a difference in looking time to voicing contrast depending on the condition, and they attributed this difference to the fact that infants in the first case grouped in one category alternating voiced and voiceless consonants. One problem with this interpretation, however, is that it interferes with the one made in Feldman, Myers, et al. (2013), based on the word-form similarity mechanism. In fact, while both mechanisms function in a rather similar fashion, they make opposing predictions. For instance, learners that hear two similar sounds consistently in two different word-forms could a priori either group them in one category according to the complementary distribution mechanism, or in two categories according to the word-form based mechanism.

1.4 This study

1.4.1 Semantic cues

The above review shows that we are still a long way from understanding the entire phenomenon of phoneme acquisition. In fact, none of the reviewed mechanisms gives a sufficient account of the learning process. More research is needed to test other potential learning strategies, taking into account the timeline of acquisition and its inherent set of constraints.

Yeung and Werker (2009) had the idea of testing semantic cues in phoneme learning

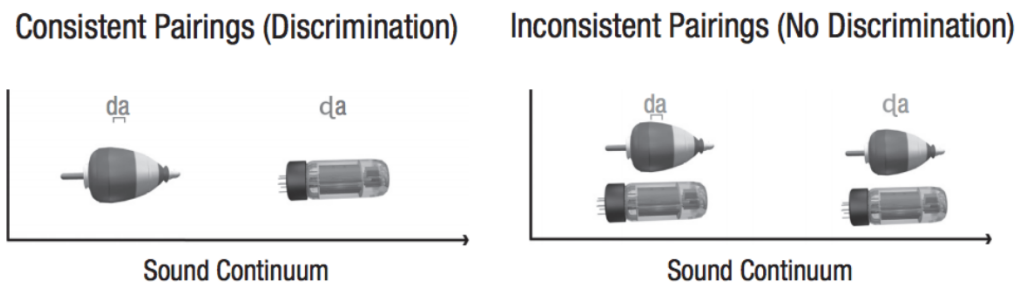

Figure 1.6: Upon familiarization with a non-native contrast paired with two objects, only infants in the consistent pairing group succeeded in the subsequent discrimination test. From Werker et al. (2012)

(Figure 1.6). They first familiarized 9-month-old English-learning infants with a non-native hindi contrast [d̪a] vs. [ɖa]) paired with two objects in a consistent or inconsistent fashion. Then they tested the infants' perceptual discrimination of the contrast, and found that only those who received a consistent training succeeded in this discrimination. This result is surprising because we know that infants at this age are *not* able to learn the referential mapping in typical word-object association tasks (Stager & Werker, 1997), not to mention learning the mapping of minimal pairs with two different objects. The latter was shown to be possible only for older babies (14 months, when tested with the looking-while-listening paradigm (Yoshida et al., 2009), and 17 months, when tested with the 'switch' task, J. Werker et al. (2002)). Nonetheless, infants in this experiment seemed to have benefited from the contextual co-occurrence of a minimal pair with two different objects to refine their sensitivity of the phonetic contrast. Indeed, Yeung and Werker (2009) concluded that this effect is more likely to be the outcome of a general perceptual learning mechanism known as acquired distinctiveness (Hall, 1991), rather than the outcome of a referential semantic learning.

This experiment allows us to see the old minimal-pair-based mechanism in a different light. As we mentioned above, the bulk of the criticism aimed at this mechanism was based on the argument that babies' vocabulary does not contain sufficient lexical minimal pairs. The argument was put forward by many researchers such as Maye et al. (2002); Swingley

and Aslin (2007); E. Thiessen and Saffran (2007); Feldman, Griffiths, et al. (2013). Here is a well articulated quote from Feldman, Griffiths, et al. (2013) that represents this position:

> The role of minimal pairs in phonetic category acquisition therefore critically depends on the extent to which young infants have access to associations between form and meaning. Children do appear to know some minimal pairs at a young age, but may not have sufficient vocabulary knowledge to support large-scale minimal-pair-based learning, making it unlikely that early sound category acquisition relies primarily on information from minimal pairs.

The criticism assumes that the use of minimal pairs is conditioned upon learning their referential semantic representation. However, as one can conclude from the experiment of Yeung and Werker (2009), infants need not necessarily learn the exact referential meaning of a word to benefit from its top-down constraint. In fact, the learner need only associate two words with different objects or events to home in the contrastive element. In fact, since the aim is to decide whether or not a contrast is phonemic, one only need to know whether this contrast signals two different words or a mere variation within a single word.

But how can babies know whether two word-forms represent one or two lexical items, if they dont know their referential meanings? In order to answer this question, I will first show evidence that a rudimentary semantic representation is accessible to babies from the early stages of their development. Second, I will show that this early semantic representation contains distributional cues as to how word-forms relate to each other. Finally, I will introduce the main proposal of this dissertation, a learning mechanism based on the distributional properties of this early semantic representation.

1.4.2 Statistical learning and early semantic representation

As rightly argued by Feldman, Griffiths, et al. (2013), infants lack sufficient referential knowledge during the process of acquiring phonemes. Nonetheless, they are in possession of a highly sophisticated apparatus for tracking co-occurrence statistics of different sorts,

and at different linguistic levels. For example, babies at an early age can learn how sounds are structured within words, i.e., phonotactic patterns (e.g., J. Saffran & Thiessen, 2003). They also learn how syllables are combined to form words, based on their co-occurrence statistics (J. R. Saffran, Aslin, & Newport, 1996), and how words can be combined into phrases and sentences (e.g., J. R. Saffran & Wilson, 2003; Gómez & Gerken, 1999). Besides their ability to track sequential co-occurrences, babies around 15 months of age are able to learn non-adjacent co-occurrence probabilities[1] (Gómez & Maye, 2005). More importantly, babies are able to learn the co-occurrence of objects and their referents. L. Smith and Yu (2008) showed that starting from 12 months of age, babies demonstrate a significant sensitivity to the correspondence between words and the context in which they are uttered. More precisely, they are able to integrate information gathered across many ambiguous situations to settle on the most probable referent (i.e., 'cross-situational learning').

This powerful ability to track co-occurrence statistics—especially the gradual learning of word-object associations across multiple contexts of exposure—suggests that infants' semantic representation does not function in a black-and-white mode. The failure to recognize the semantic referent of a given word at a certain point in development does not, a priori, preclude the presence of a rudimentary semantic representation for this word. In fact, according to J. Werker and Curtin (2005):

> The transition from recognizing word-forms to full referential understanding likely involves several steps (see Hirsh-Pasek & Golinkoff, 1996; Nazzi & Bertoncini, 2003; Werker & Tees, 1999). One of the earliest steps is learning arbitrary associative links between words and objects or events in the world. Although this kind of "goes with" understanding falls short, on a number of dimensions, of full referential understanding (Bloom, 1999; Merriman & Bowman, 1989), it is an essential step toward meaning.

I will illustrate this idea with a concrete example. Before the baby learns that the word "kitchen" refers to the concept of a room where food is prepared and/or eaten, she might at

[1] Non-adjacent patterns are needed, for instance, to learn grammatical dependencies marking tense, as in the relationship between the auxiliary "is" and the inflection "-ing", which are necessarily separated by a verb.

Figure 1.7: An illustration of a putative early semantic representation of the word "kitchen", including co-occurring objects, words and time periods.

first, as in typical experimental setting of cross-situational learning, associate it with the general spatial context involving other co-occurring objects (KITCHEN, APPLE, SPOON,...). Actually, the real situation is more complicated than the typical laboratory setting. In fact, the baby sees many objects at different periods of the day, and hears many words in the same conversation. Therefore, it is likely that the the infant's early semantic representation of the word "kitchen" involves not only the co-occurring objects located in space (KITCHEN, APPLE, SPOON,...) but also the corresponding time of the day (MORNING, MIDDAY, EVENING,...), as well as the co-occurring words in the conversation ("kitchen", "apple", "spoon",...) as illustrated in Figure 1.7.

1.4.3 Word co-occurrence as a proxy for the general context

Roy, Frank, DeCamp, and Roy (2015) conducted a large scale, longitudinal study of a childs daily life from 9 to 24 months of age. They recorded approximately 10 hours/day

during this period, capturing about 70% of the childs waking hours. Their primary goal was to predict the childs age of production of the first 679 words, based on how the parents' utterance of these words were distributed across the three dimensions we mentioned previously, i.e., the location in physical space where it is spoken, the time of the day at which it is spoken, and the other words that appear nearby it in the conversation. Once they had created context distributions for each dimension, they computed the 'distinctiveness' of words along that dimension, a measure that captures the distance between the contextual distribution of the word and that of language more generally. For example, content words like "fish" or "breakfast" have more distinct spatial, temporal and linguistic distributions than function words like "with" or "the". They showed that all these dimensions were accurate in predicting the age at which words were first produced. But more striking was the finding that, despite the radically different data they were derived from (videos, time of the day of each utterance, and the conversation transcripts), the three distinctiveness variables showed strong correlations with one another. This lead the authors to suggest that each of the three investigated dimensions represents a proxy for a single underlying pattern of word distinctiveness.

Following this work, I propose to approximate the early semantic representation of a word by the set of co-occurring words. This assumption captures only one of the three above-mentioned dimensions. However, as was noted by Roy et al. (2015), information related to contextual distribution tends to be redundant across these dimensions, thus making word co-occurrence a reasonable proxy to the general context. Moreover, as I will explain below, this approximation will allow us to take advantage of a useful linguistic property known as the 'distributional hypothesis'.

1.4.4 The distributional hypothesis

The early semantic representation—as approximated now by the set of co-occurring words—is deemed highly ambiguous by the typical framing of meaning learning (e.g., Quine, 1960),. In fact, characterizing the referent of a word (e.g., "kitchen") by the neighboring words in the conversation ("breakfast", "apple", "spoon",..) is ambiguous and does not correspond to our mature and precise intuition of this referent. Nonetheless, this characterization contains considerable distributional information as to how words are related to each other. In fact, linguists such as Harris (1954) and Firth (1957) noted that words' similarity can be derived from their distribution in natural speech, since semantically related words tend to co-occur more often in conversations than unrelated words. For example, as the word "cat" occurs more often with "dog", than with, say, "school", it is natural to expect "cat" and "dog" to be more semantically related than "cat" and "school". This seemingly simple property probably plays a crucial role in learning the lower units of speech. Imagine that the learner hears the word "cat" in an ambiguous context that includes many potential referents that co-occur consistently with a cat, such as "dog", "chair", "table",... And suppose the learner also hears the word "cab" in a context that consistently includes "car", "building", "stranger",... etc. Our learner may not have the words "cat" and "cab" in her vocabulary, i.e., she may not be able to look or point at the correct referent when the word is uttered, and therefore these words would not show up in a CDI-like repertoire. Nonetheless, she will probably be able to decide that "cat" and "cab" refer to different things, since they occur consistently in different contexts. This knowledge is sufficient to decide, for instance, that [t]-[b] is a phonemic contrast.

In the next section, this intuition will be developed in a formal sense, thus introducing the central learning mechanism of this dissertation.

1.4.5 The proposed learning mechanism

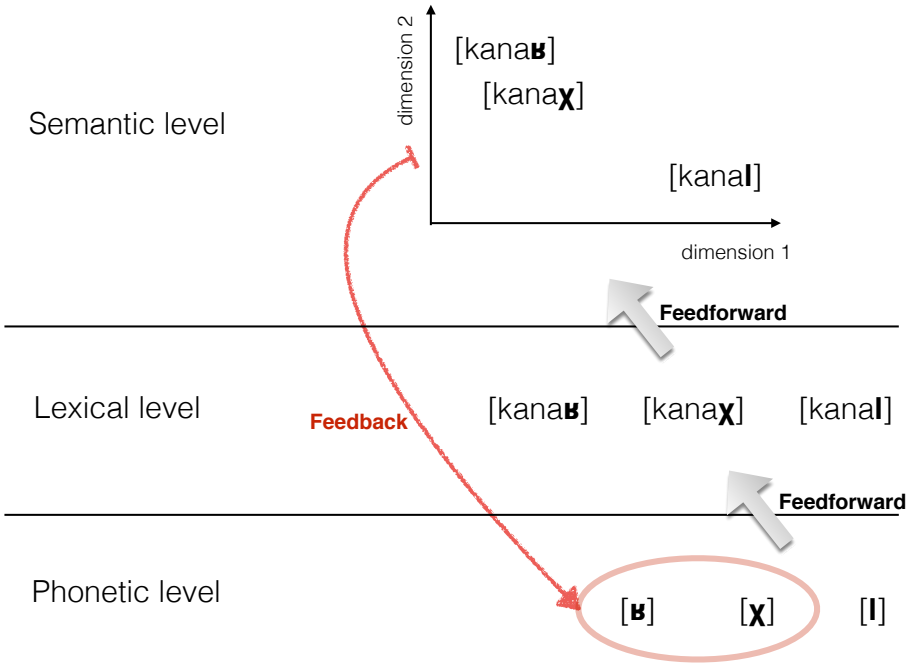

Figure 1.8: schematic illustration of the proposed learning mechanism. Learning is understood to be the outcome of an interaction between three levels of linguistic representation. The phonetic level is composed of fine-grained categories. This information percolates into the lexical level where word-forms are extracted and stored along the granularity of the phonetic level. The semantic level represents the distribution of these word-forms according to their co-occurrence in conversations. Through a feedback loop, the semantic level readjusts the phonetic level along the relevant (i.e., phonemic) dimensions

The main theoretical proposal of the present study is that learners need not have a fully developed high level representation (e.g., semantic representation) to make use of it as a top down constraint to learn the low level representations (e.g., phonemes). I explore the hypothesis according to which infants learn approximate, provisional linguistic representations in parallel, and that these approximate representations are subsequently used to improve each other. More precisely, I make four assumptions:

1. **Fine-grained categories**: infants start by paying attention to fine-grained variation

in the acoustic input, thus constructing perceptual phonetic categories that are not phonemes, but segments encoding fine grained phonetic details.

2. **Proto-lexicon**: these units enable infants to segment proto-words from continuous speech and store them in this detailed format.

3. **Early semantic similarity**: infants can use this imperfect lexicon to acquire a sense of semantic similarity, according to the distributional hypothesis.

4. **Feedback**: as their exposure to language develops, infants reorganize the initial phonetic categories along the relevant dimensions of their native language based on cues derived from this early sense of semantic similarity.

I will illustrate this mechanism in the light of a concrete example (Figure 1.8). In French, the uvular fricative can take a voiced or voiceless surface form according to the voicing of the following segment. This can be summarized in the allophonic rule shown in Figure 1.9.

$$/ʁ/ \rightarrow \begin{cases} [χ] & \text{before a voiceless consonant} \\ [ʁ] & \text{elsewhere} \end{cases}$$

Figure 1.9: Allophonic rule of the French uvular fricative

Imagine, as in typical experimental framing of phoneme acquisition, that a french-learning baby has to decide whether the sounds [ʁ] and [χ] correspond to instances of the same phonemic category or instances of two different categories. We assume that, regardless of their knowledge about the phonemic status of these phones, learners can use them to segment acoustically detailed word-forms, such as [kanaʁ] and [kanaχ] (both are surface forms of the French word *canard*, meaning "duck"). This brings the initial problem to the lexical level. Instead of deciding if two sounds belong to one or two phonemic categories, the learner has now to decide if two word-forms correspond to one or two lexical items. Crucially,

we suppose that the learner does not learn these detailed word-forms in a vacuum, but in a linguistic and conversational context where they co-occur with other related words. According to the distributional hypothesis, this general linguistic context offers a sense of semantic similarity which can be used as a top down feedback. In fact, if two similar word-forms have similar distributions (i.e., occur with similar words) as is the case with the pair [kanaʁ] vs. [kanaχ], they are more likely to be merged in the same lexical category, which means that the corresponding phones [ʁ] and [χ] will be categorized in the same phonemic category. Conversely, if two similar word-forms have dissimilar distributions (i.e., co-occur with different words) as is the case with [kanaʁ] vs. [kanal] ("channel"), they are more likely to be categorized in two different lexical categories, which means that the corresponding phones [ʁ] and [l] will be categorized in two different phonemic categories.

In the second part of this dissertation, I will tackle the computational aspect of this learning mechanism, that is, starting from a realistic input and abstracting away from the cognitive limitations of the learner, I will explore if sufficient cues to learning are available to support the proposed mechanism. In the third part, I will study the complementary question, i.e., starting from a controlled input, I will explore if human learners are cognitively equipped to learn according to the mechanism I propose.

Part II

Computational experiments

This part of the dissertation deals with the computational level of analysis, i.e., the extent to which the input offers sufficient cues to support the learning mechanism proposed in Section 1.4.5. It includes the second, third and fourth chapters. In the second chapter, I will present the datasets used and explain how the input will be represented at different linguistic levels. In the third chapter, I will test bottom-up and top-down cues that are believed to shape the learners' early sensitivity to phonemic contrasts. I will compare their performance to that of two implementations of the top-down mechanism that I propose. While the third chapter explores rather *continuous* cues to phonemicity, the fourth chapters models the 'hard decision' of phoneme acquisition, i.e., the fact that babies should be able, not only to tell how phonemic a pair of phone is, but also to learn how much phonemicity is *contrastive* in their native language. Here again, I will show that an early and ambiguous semantic representation allows for successful learning.

Chapter 2

Datasets and Representations

Figure 2.1 shows the general scheme of the computational part of this dissertation. It combines both a description of the way the linguistic representations were derived (the raw input and softwares used to generate them) and a global indication of how the learning algorithm I propose will operate on these representations. The present chapter deals rather with the first aspect, that is, the derivation of the linguistic representations. This includes both the end-state representation (i.e., phonemes), and the input representation (the format in which information is extracted from the environment and putatively represented in the mind before learning). Crucially, it is assumed here that the input representation is not limited to phonetic information (contextual allophones), it also includes information from higher linguistic levels such as the lexicon and the semantics.

In Section 2.1, I will present the datasets used in this study, and I explain how the phonemic representation (end-state of learning) were derived from them. In Section 2.2, I will describe the derivation of the input representations at the phonetic, lexical and semantic level.

2.1 Datasets and the phonemic representation

Ideally, the dataset should consist of speech recordings from the natural environment of infants. However, it is not easy to have access to corpora of Infant Directed Speech (hereafter, IDS) which are well annotated and time-aligned at the phonetic level. Moreover, most of the available IDS corpora do not come with a sound quality high enough to allow the use of speech recognition techniques, and do not contain data large enough to support the use of tools from computational linguistics and information retrieval. Here I made a pragmatic choice by relying on existing corpora of rather adult direct speech, representing mostly spontaneous conversations and monologues. I selected corpora from English and Japanese, two languages that differ typologically along several phonological dimensions

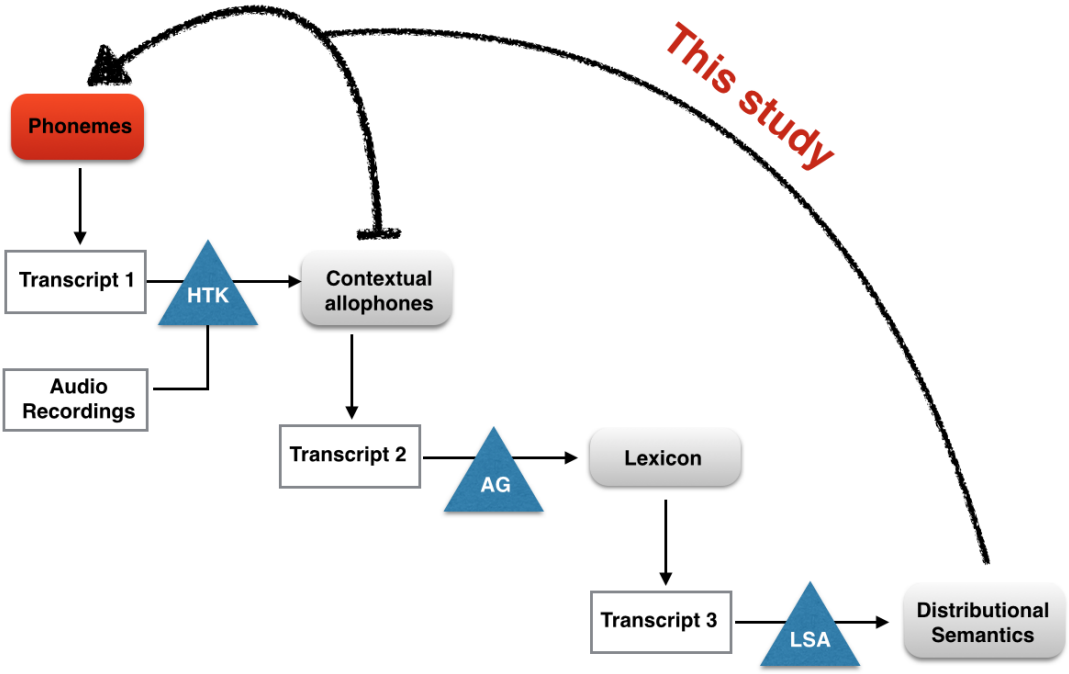

Figure 2.1: Schematic description of the computational part of the dissertation. It includes a description of the way the representations are derived, and a global indication of how the proposed algorithm operates on these representations.

such as the phonotactic constraints and the rhythmic structure. In the following, I present the corpora used, and a summary of the main pre-processing steps made to derive the phonemic representation.

2.1.1 The Corpus of Spontaneous Japanese (CSJ)

The Corpus of Spontaneous Japanese (hereafter, CSJ) (Maekawa, Koiso, Furui, & Isahara, 2000), is a large-scale annotated corpus of spontaneous Japanese. The whole CSJ contains about 650 hours of spontaneous speech that correspond to about 7 million words. All these speech material are recorded using head-worn close-talking microphones and digital audio tapes, and down-sampled to 16 kHz, 16 bit accuracy (CSJ website, 2015). In this study, we only used the subset of CSJ called the 'Core', which contains about 500 thousands

words or about 45 hours of speech. It is the part of CSJ to which the cost of annotation is concentrated. As explained in Maekawa et al. (2000), the speech recorded for CSJ is so-called common, or standard, Japanese, a variety shared widely by educated people and used in more or less public circumstances. Most of the speech material is devoted to spontaneous monologues, of which the two main types are academic presentation speech done in various academic societies, and simulated public speaking, in which laymen talk about everyday topics like 'my most delightful memory' or 'If I live in a deserted island'. The age and sex of the speakers were approximately balanced.

The corpus is annotated for various linguistic levels (phonetics, phonology, morphology, syntax, prosody,...). I pre-processed the original XML files to extract information about phonemes, words and utterances. In order to obtain the phonemes, I ignored the phonetic details annotated below what the CSJs documentation refers to as the 'phonemic level'. I also ignored sub-segmental events of various sorts (release of the stop closure, voicing of vowels,...). Yet, some of the remaining segments do not represent real contrastive elements. I performed further pre-processing in order to obtain a true phonemic inventory (See Appendix A). The resulting segmental inventory consists of 25 phonemes, 15 consonants represented in Figure 2.2, and 10 vowels represented in Figure 2.3. This final phonemic inventory is identical to the one used in Boruta (2012). It is also almost identical to the inventory proposed by Okada (1999) (who posited an additional phoneme, /ts/) and to the one described in the collaborative encyclopedia (Wikipedia, 2015), in which the only difference is that the abstract moraic obstruent /Q/ is considered as an independent segment. Figure 2.4 shows the frequency of the final phonemes observed in the CSJ corpus. We can see that the frequency distribution mimics a Zipfian trend, i.e., very few elements have high frequencies and most elements have low frequencies.

In order to prepare the data for the derivation/evaluation of the higher linguistic levels (lexicon and semantics), I also extracted higher units of speech: words and utterances.

	bilabial	alveolar	palatal	velar	Uvular	glottal
Nasal	**m**	**n**			**N**	
Stop	**p b**	**t d**		**k g**		
Fricative		**s z**				**h**
Flap		**r**				
Approximant			**y**	**w**		

Figure 2.2: The consonants of Japanese as used in this study. The symbol to the right represents the voiced version of the consonant.

	Front		central		Back	
	long	short	long	short	long	short
close	**/i:/**	**/i/**			**/u:/**	**/u/**
mid	**/e:/**	**/e/**			**/o:/**	**/o/**
open			**/a:/**	**/a/**		

Figure 2.3: The vowels of Japanese as used in this study.

CSJ is annotated for "short-unit words" (SUW) and "long-unit words" (LUW). I chose the short-units to be the words since, according to the CSJ's documentation, they correspond to approximate items of ordinary Japanese dictionaries, whereas long units correspond to compounds made up of more than two SUWs. For example, the phonemic sequence /toHkyoHkoHgyoHdaigaku/ is analyzed into three short units: /tokyoH/ (Tokyo), /ko-HgyoH/ (technology), and /daigaku/ (university), but constitutes a single compound noun, (i.e., LUW) translated as "Tokyo Institute of Technology".

An utterance is defined in CSJ as a speech unit bounded by pauses of more than 200 ms. In the pre-processing, I discarded utterances that were not at least 100 ms long (these represent less than 1% of the corpus).

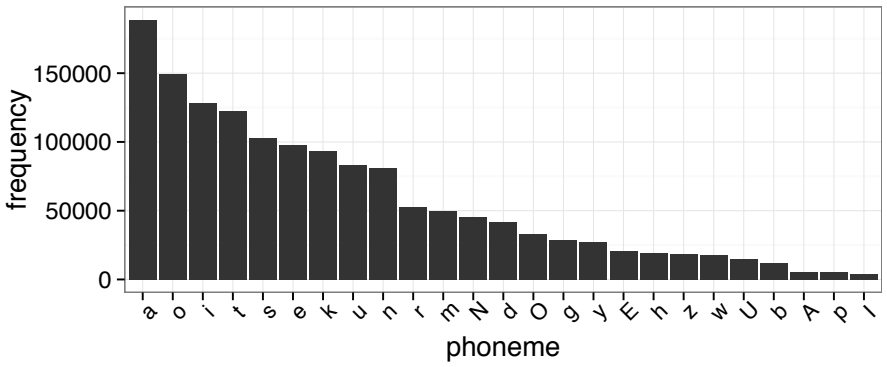

Figure 2.4: Frequency distribution of phonemes in the CSJ Corpus.

2.1.2 The Buckeye Corpus of Spontaneous Speech

The Buckeye Corpus of Spontaneous Speech (Pitt et al., 2007) is a database of approximately 300,000 words or about 40 hours of conversational speech by native central Ohio speakers. It consists of interviews with 40 speakers balanced for age (over 40, under 40) and gender. As explained in Pitt et al. (2007), talkers were told that the research team was interested in how people express their opinions. They were invited to come to the Ohio State University campus to have a conversation about everyday topics such as politics, sports, traffic and schools. Speech was digitally recorded in a quiet room with a close-talking head-mounted microphone, which allowed freedom of movement. The microphone was fed to a digital audio tape recorder.

The corpus is annotated orthographically, phonemically and phonetically. The phonetic transcription is fine-grained, and encodes various kinds of phonological and speaker-dependent variations. As I am interested in the phonemes, I naturally chose the phonemic level of transcription. This allows me to use directly the original coding scheme of the corpus, and avoid reprocessing a rich phonetic annotation into a reduced set of contrastive sounds (as I did with Japanese data). The phonemic inventory used in the Buckeye corpus is composed of 42 segments, 27 consonants summarized in Figure 2.5, and 15 vowels

summarize in Figure 2.6, using lexical examples for each vowel. The number of vowels was eventually reduced to 14, after tokens of the segments [ɑ] and [ɔ] were collapsed into the same category. In fact, The vowel ɔ occurred only 5 times in the corpus, compared to 17984 instances of the vowel ɑ[1]. Figure 2.7 shows the frequency of the final phonemes observed in the Buckeye corpus. As in Japanese data, we see a power-law-like frequency distribution.

	bilabial	labio-dental	dental	alveolar	post-alveolar	palatal	velar	glottal
Nasal	**m (em)**			**n (en)**			**ng**	
Stop	**p b**			**t d**			**k g**	
Affricate					**ch jh**			
Fricative		**f v**	**th dh**	**s z**	**sh zh**			**hh**
Approximant				**r**		**y**	**w**	
Lateral				**l (el)**				

Figure 2.5: The consonants of English, as used in the Buckeye corpus. The symbol to the right represents the voiced version of the consonant. The symbol between parentheses represents the syllabic version of the consonant (analysed as phonemes in the Buckeye corpus). The ARPA transcription is used for ease of reading.

The corpus was annotated for word boundaries, but not for utterance boundaries. I took as utterances sequences of words bounded by non-speech events such as silence (labeled <SIL>), laughter (<LAUGH>), hesitation (<HES>), noise (<VOCNOISE>), or change of speaker (<IVER>). Table 2.1 gives some useful statistics that characterize English and Japanese data, when represented with phonemes.

[1]My decision to merge these vowels in the same category was motivated by statistical considerations. It is worth mentioning, however, that it corresponds to a real psycholinguistic phenomenon known to linguists as the *cot-caught* merger (Labov, 1991). It has been established in many regions such as New England and Canada, and it appears to be spreading rapidly in many parts of the United States, including Ohio, where the Buckeye corpus was recorded.

	Front		central	Back	
	long	short		long	short
close	**iy** (beat)	**ih** (bit)		**uw** (boot)	**uh** (book)
mid		**eh** (bet)	**er** (bird)	**ao** (caught)	
open		**ae** (bat)	**ah** (butt)	**aa** (cot)	
Diphthongs	**ay** (bite), **aw** (now), **oy** (boy), **ow** (boat), **ey** (bait)				

Figure 2.6: The vowels of English, as used in the Buckeye corpus. Each vowel V is accompanied with a lexical CVC example. The diphthongs are analysed as single phonemes. The ARPA transcription is used for ease of reading.

2.2 Input representations

In the previous section, I characterized the end-state of learning (i.e., the phonemes). In this part, I will characterize the early speech representations at the phonetic, lexical and semantic level (Figure 2.1). Note that these early representations are supposed to be available to the learners before they come to the task of phoneme learning. The methods and softwares used to generate these representations are not considered part of the learning mechanism I propose, but only tools to approximate the learners' prior knowledge.

2.2.1 Early phonetic representation

Even if languages make use of a finite inventory of segments (or phonemes) to form words and utterances, the physical realization of these segments is not constant. In fact, instances of the same phoneme can take various forms depending on both the properties of the speaker (such as identity, sex, age, mood and spech rate), and the phonology/phonetics of the language. A lot of studies in the modeling literature have dealt with random, low level variation that results mainly from idiosyncratic properties of the speaker [2]. In this

[2] often relying on the seminal work of Hillenbrand et al. (1995), who derived empirical estimates of acoustic parameters for the English vowel category means and covariances, based on production data by men, women and children

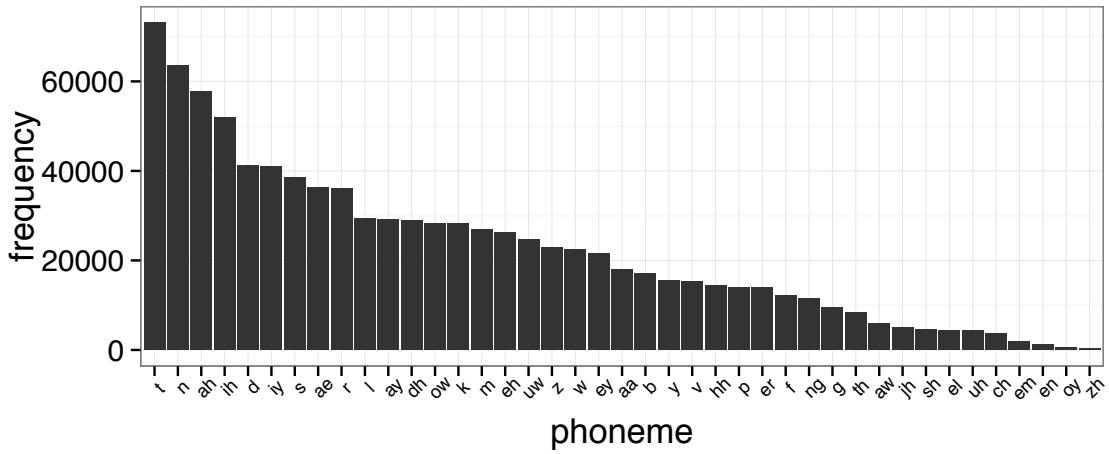

Figure 2.7: Frequency distribution of phonemes in the Buckeye Corpus

study, I chose, instead, to focus on a relatively under-studied source of variation, which is, rather, intrinsic to the linguistic system.

We can enumerate three main aspects of language-dependent variation. The first one is **free variation**, that is, when a phoneme takes two or more forms in the same phonetic environment without a change in meaning. An example of free variation in English is the glottalization of voiceless stops in word-final position (e.g., the word *stop* can be pronounced with a plain unaspirated [p], or with with a glottalized [p$^{?}$]). The second source is **positional variation**, which could be described as a convention within a speech community to use certain allophones in certain contexts. For example, English voiceless stops are aspirated when they are word-initial or begin a stressed syllable, as in *pill* or *kill*, and they are unaspirated when following word-initial *s*, as in *spill* and *skill*. The third source of variation, and arguably the most pervasive one, is **contextual variation**, which is mainly driven by the phenomenon of 'coarticulation', that is, when vocal tract gestures for one sound overlap with gestures for another. For example, the English vowel /æ/ is nasalised before nasals (e.g., [mæ̃n]) and realized as oral in all other situations (e.g., [mæd]) [3]. The present study

[3]The second and third sources of variation are not necessarily orthogonal to one another. Conventions

	English	Japanese
Tokens		
utterances	49,626	53,870
words	283,129	436,493
phonemes	2,263,683	2,890,624
Types		
words	9,288	17,429
phonemes	41	25
Averages		
words/utterance	5.7	8.1
phonemes/word	8.0	6.6
phonemes/utterance	45.6	53.6

Table 2.1: Characteristics of phonemically transcribed corpora in English and Japanese

$$/ʁ/ \rightarrow \begin{cases} [χ] & \text{before a voiceless consonant} \\ [ʁ] & \text{elsewhere} \end{cases}$$

Figure 2.8: An allophonic rule of the French uvular fricative

deals mainly with this third source of variation, and we will call "allophones" of a phoneme, the realization of the latter in different phonetic contexts.

Another example of contextual allophony (which we saw in Chapter 1), is the case of the French uvular fricative phoneme (Figure 2.8). If the following segment is a voiced obstruent, like /ʒ/, it surfaces as a voiced sound (i.e., [ʁ]) like in [kanaʁ ʒon] ("canard jaune", yellow duck). If, instead, the following segment is voiceless, like /f/, it surfaces as a voiceless sound (i.e., [χ]) like in [kanaχ flotan] ("canard flottant", floating duck). The pair of phones [ʁ] and [χ] might sound almost indistinguishable to French adults, but they represent, a priori, two different sounds to French-learning infants. The problem of phoneme

about positional allophones can well be driven by coarticulatory considerations.

acquisition can be understood as learning when two sounds belong to the same abstract category, and when they belong to different categories, signaling change in meaning, as in [kanaʁ] (duck) vs. [kanal] (canal or channel). Before studying the learning mechanism per se (Chapters 3 and 4), I will investigate possible ways to code the corpus into fine-grained context-dependent units (which we now call allophones). They will be taken as the initial phonetic representation putatively processed by infants before they learn the relevant phonemic categories of their language. Previous work in this line of research generated allophonic variation using either random rules as in Martin et al. (2013) or Hidden Markov Models (HMMs) trained on audio recordings as in Schatz & Dupoux (unpublished) and Boruta (2012). Interestingly, choosing one or the other has been reported to make a huge difference in terms of the quality of the subsequent learning (Boruta, 2012).

Random allophones

Martin et al. (2013) generated allophonic variation through a random partitioning of the linguistic contexts. That is, for a given phoneme (e.g., /p/), the set of all possible contexts is randomly partitioned into a fixed number N of subsets. In the transcription, the phoneme /p/ is converted into one of its allophones (p_1,p_2,..,p_N) depending on the subset to which the context belongs. Note that this procedure does not take into account the fact that contexts normally belong to natural classes (e.g., voiced vs. voiceless or obstruent vs. sonorant). Figure 2.9 provides an example from Martin et al. (2013) that illustrates this procedure on a concrete example in Japanese. In this example, each phoneme has two allophones, and the notation is read as follows: a rule of the form X → Y / __ {A, B, C} states that phoneme X is realized as allophone Y when followed by A, B, or C.

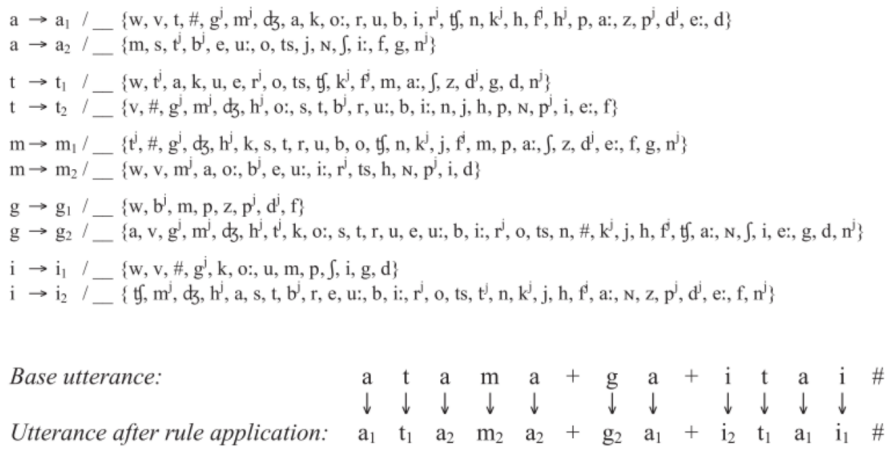

Figure 2.9: Example of rule application on the utterance "atama ga itai" (my head hurts), using random rules which assign each phoneme one of two allophones. (from Martin et al. 2013)

HMM-based allophones

Schatz & Dupoux (unpublished) used the speech recognition software HTK (Young et al., 2006) to generate allophonic rules that are linguistically and acoustically more realistic than the random rules used in Martin et al. (2013). HTK applies a standard Hidden Markov Model (HMM) phoneme recognizer with a three-state-per-phone architecture to the signal. We can summarize the process in the three following steps [4]:

- **Feature extraction**: the raw speech waveform is converted into successive vectors of Mel Frequency Cepstrum Coefficients (hereafter, MFCCs), computed over 25 ms windows, using a period of 10 ms (the windows overlap). 12 MFCC coefficients are used, plus the energy, plus the first and second order derivatives, yielding 39 dimensions per frame.

- **Modeling the phonemes**: each phoneme in the inventory is modeled by a three-state HMM. The first state models the first part of the sound, the second state models the middle part, and the third, the end-part of the sound unit. The HMM is trained

[4] but see Young et al. (2006) for more details

on all the attested realizations of the phoneme in the corpus (as represented by their MFCC features). This is possible because speech utterances are time-aligned with their phonemic transcription.

- **Modeling the allophones**: the model of each phoneme is cloned into context-dependent triphone, for each context in which the phoneme actually occurs. For example, if the phoneme /o/ occurs in the context [d–o–g] as in the word "dog", a triphone unit is created (labeled in HTK as "d–o+g" with a "–" sign standing for following and a "+" sign standing for preceding). The triphone inherits the HMM model of the original phoneme (/o/ in our example). Next, it is retrained on only the subset of the data corresponding to the given triphone context. These detailed models are clustered in a procedure called 'state-tying'. In brief, the i^{th} state ($i = 1$, 2 or 3) of all allophonic HMMs of a given phoneme are first put in a single cluster, and then split into finer grained categories, gathering allophones with phonetically similar preceding (or following) contexts. This partitioning is performed iteratively according to a structure resembling a decision tree (Figure 2.10). At each new branch, different ways of partitioning are tried, ranging from broad natural classes (e.g., does the left (or right) context belong to the category of consonants?) to very specific questions (i.e., is the left (or right) context a glottal fricative?) (See Appendix B for the list of questions I designed for the purpose of this study). The algorithm selects the question that maximizes the likelihood of the data. This procedure is repeated at every new branch until a global threshold is reached. I chose as a termination criterion for this procedure the global number of desired allophones, representing the complexity of the initial sound inventory. There is a debate on how far a phonological theory should go in describing the details of contextual variation (e.g., Cohn, 2006). I got around this debate by varying the global number of allophones in the range of interest. A preliminary investigation showed that varying this number from twice to 16 times the

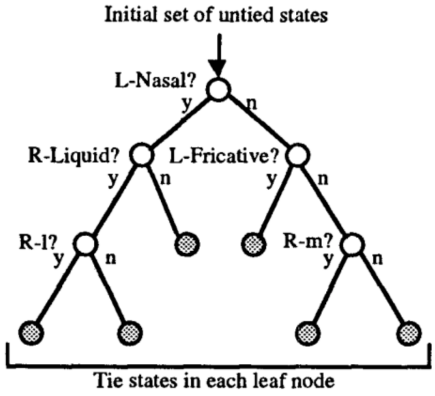

Figure 2.10: Example of a phonetic decision tree used in the state-tying procedure of HTK. From young et al. (2006)

size of the phonemic inventory was sufficient to study the behavior of our learning algorithm [5].

In Appendix C, I statistically characterize HTK-allophones in terms of sensitivity to variation and frequency. Random allophones, in contrast, were found to be sensitive to variation but not to frequency. Data from perceptual learning literature show that, in reality, both sensitivity to variation and sensitivity to frequency are crucial to representing categories. While the former is necessary to differentiate items in the perceptual space, the latter is necessary to avoid the spurious proliferation of unrepresentative categories. For example, Maye et al. (2002) showed that exposure to a bimodal frequency distribution lead to the representation of two categories along a VOT continuum, whereas exposure to a unimodal frequency distribution lead to the formation of only one category, although, crucially, items from each point in the continuum were heard.

[5] In fact, as will be shown in the next chapter, the effect tends to vanish starting from an allophonic complexity equal to 8 times the size of the phonemic inventory

2.2.2 Early word-form representation

Developmental studies show that babies do not wait to have mastered phonemes to start segmenting words from connected speech (e.g. Jusczyk & Aslin, 1995). How can infants learn word-forms at an age when they still lack a robust phonemic representation? J. Werker and Curtin (2005) suggested that infants use a finer-grained phonetic representation to code speech and segment word-forms. This claim is supported by the fact that infants' initial word-form representation is itself acoustically detailed (Houston & Jusczyk, 2000; Curtin, Mintz, & Byrd, 2001). I operationalized the assumption of J. Werker and Curtin (2005) through using early phonetic categories (i.e., allophones) to encode speech in a detailed format, i.e., two word-forms that belong to the same lexical category but differ in their allophonic representation (e.g., [kæt^{h}], [kæʔ]) will be considered as two different items.
In addition to phonetic variability, infant's early lexicon is affected by segmentation errors. It contains under-segmented word-forms such as "get-a", "put-a" and "want-to" (Brown, 1973), and over-segmented words, which can be illustrated by the example reported in Peters (1983): when an adult tells a child that she "must behave", the child responds: "I am [heɪv]!", indicating that she analyzed "behave" as "be [heɪv]". More generally, Ngon et al. (2013) showed that infants tend to store in their memory frequent chunks, whether they match correct words or not.

In order to approximate infants' word segmentation, I use a state-of-the-art unsupervised word segmentation model, called Adaptor Grammar (AG, M. Johnson, Griffiths, & Goldwater, 2007). It makes use of a stochastic process known as Pitman-Yor (Pitman & Yor, 1997). The algorithm takes as input multiple sequences of unsegmented utterances, looks for recurring chunks, and reuses them to parse the corpus anew. It converges over many such iterations towards an overall segmentation of the corpus where words tend to be distributed according to Zipf's law, thus simulating the distribution observed in natural speech (Goldwater, Griffiths, & Johnson, 2011). Figure 2.11 provides a schematic description of

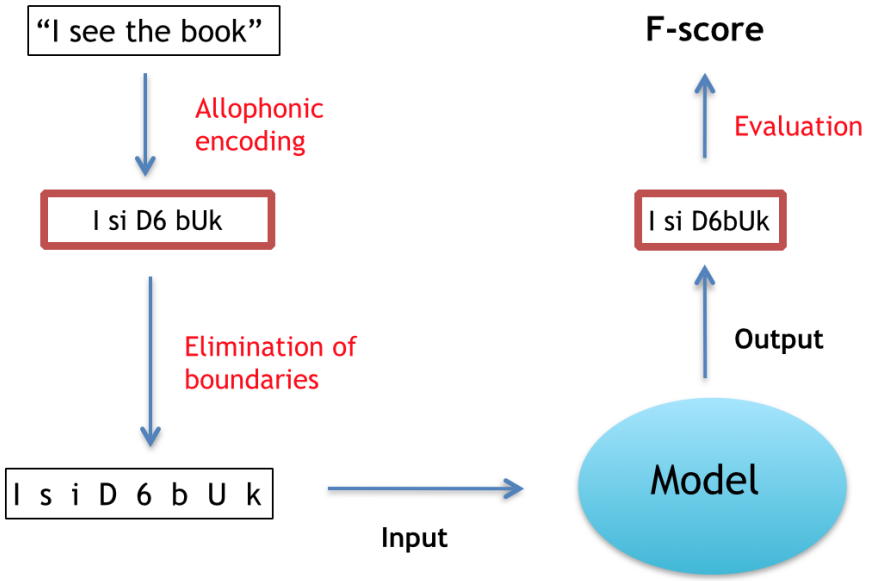

Figure 2.11: Schematic description of modeling early word segmentation. For ease of presentation, I only show the case of the phonemic representation. The allophonic representation at different levels of complexity is obtained by replacing each phoneme with the allophone that fits the particular context where this phoneme occurs

the way I model and evaluate early word segmentation. The corpus is initially transcribed using the phonemic inventory, then each phoneme is replaced with the allophone that fits the particular context where this phoneme occurs. Next, information about word boundaries is removed in each utterance. The unsegmented utterances are then given as input to the segmentation model. The model tries to put the boundaries back, in a way that optimizes the posterior probability of the data, according to a Bayesian generative model (Johnson et al. 2007). Finally, the output of the model is evaluated by comparison to the initial transcription, and assigned an F-score (F), defined as the harmonic mean of Precision (P) and Recall (R):

$$F = \frac{2 * P * R}{P + R}$$

where P is defined as the number of correct items found, out of all items posited, and R is the number of correct items found, out of all items in the ideal segmentation.

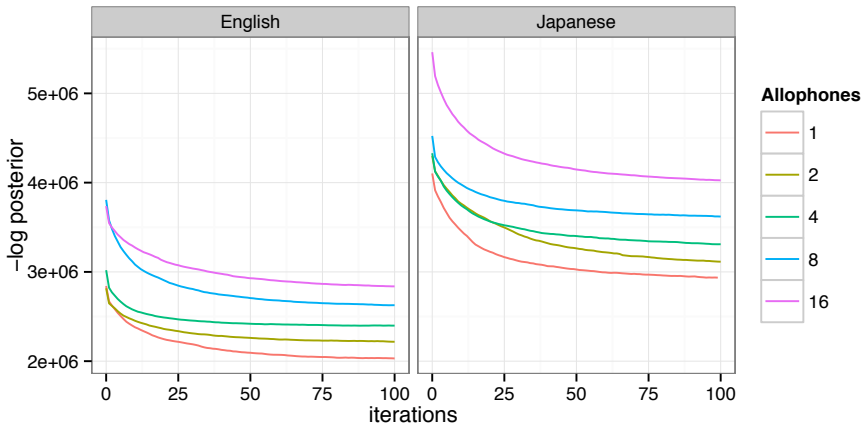

Figure 2.12: Negative log posterior probability (lower is better), as a function of iteration, for corpora at different levels of allophonic complexity. I used a collocation adaptor grammar with Pitman-Yor adaptors and an incremental initialization.

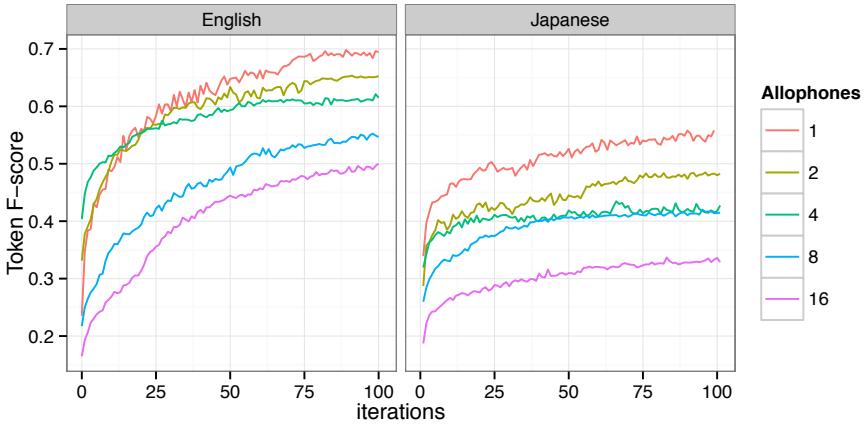

Figure 2.13: Token F-score (higher is better) as as a function of iteration, for corpora at different levels of allophonic complexity. I used a collocation adaptor grammar with Pitman-Yor adaptors and an incremental initialization.

For each level of allophonic complexity, I ran a Markov Chain Monte Carlo sampler for 100 iterations (Figure 2.12), a number around which the log posterior probability tends to level out. We observe a clear difference between English and Japanese data, the former having a higher posterior probability. Moreover, within each language, we observe a monotonic increase in the probability as a function of the average number of allophones per

phoneme. As expected, the F-score mirrors closely the behavior of the segmentation probability (Figure 2.13). The higher the probability, the better the quality of the segmentation. The buckeye corpus leads to an overall better segmentation, and generally more allophones lead to lower segmentation scores.

In order to derive a single score for each corpus, I collected sample segmentations for evaluation after a burn-in of 80 iterations, and I performed a Minimum Bayes Risk decoding over these samples. Figure 2.14 shows the results of the optimal segmentation F-score for each corpus, both in terms of word tokens and word types. In addition to the output of the segmentation model, I considered a random segmentation as a control. The F-scores of the optimal segmentations show, again, that more allophones lead to a lower quality segmentation. This pattern can be attributed to the fact that increasing the number of allophones increases the number of word types (remember that different ways of pronouncing the same word are considered independent word types). These spurious words occur therefore with less frequency than the original ones, making them harder to find in continuous speech. Table 2.2 shows the increase of the number of word types in the allophonic representation relative to the phonemic representation, and the corresponding decrease in the average number of tokens per type.

	English		**Japanese**	
Allo./phoneme	w.form/word	tokens/w.form	w.form/word	tokens/w.form
1	1	30.48	1	25.04
2	1.46	20.84	1.12	22.27
4	2	15.21	1.42	17.64
8	2.49	12.23	2.02	12.39
16	3.06	9.96	2.46	10.17

Table 2.2: the number of word types in the allophonic representation relative to the phonemic representation, and the corresponding average number of tokens per type, as a function the average number of allophones per phoneme.

The other salient effect in Figure 2.14 is that English tends to fare better than Japanese

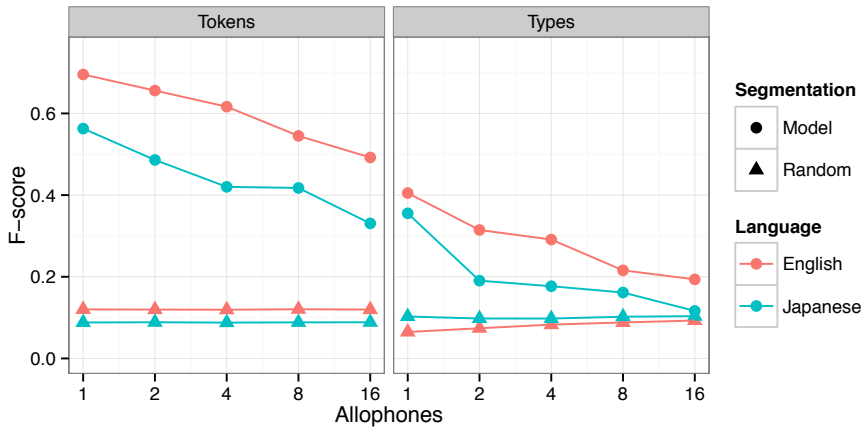

Figure 2.14: F-scores of optimal segmentations as a function of the average number of allophones per phoneme for English and Japanese data, using a collocation adaptor grammar model and a random segmenation as a control.

at the segmentation task. Unlike the previous pattern, the effect of language cannot simply be attributed to a difference in statistical evidence (i.e., frequency). In fact, Table 2.2 shows that, while the average number of tokens in Japanese vs. English varies (Japanese starts lower: around 25, compared to about 30 in English, and becomes higher starting from 8 allophones), the quality of English segmentation remains higher.

In Fourtassi, Benjamin Borschinger, and Dupoux (2013) (paper in Appendix H), we conducted an investigation to understand this phenomenon. We first reproduced the difference between English and Japanese segmentation across different corpora of both adult and child directed speech, showing that it cannot be reduced to some idiosyncratic features of the corpus used, but that it is more probably due to an intrinsic property of the language. Second, we introduced a quantity we named Normalized Segmentation Ambiguity (NSE), which quantified how ambiguous speech utterances were. For example, the utterance "I S K R E M" is ambiguous because it can be segmented as either "I SKREM" (I scream) or "IS KREM" (Ice cream). We showed that, in general, Japanese utterances tend to be more ambiguous than English ones (mainly because Japanese words contain more syllables), and that this ambiguity leads to more errors, since many wrong segmented utterances end up

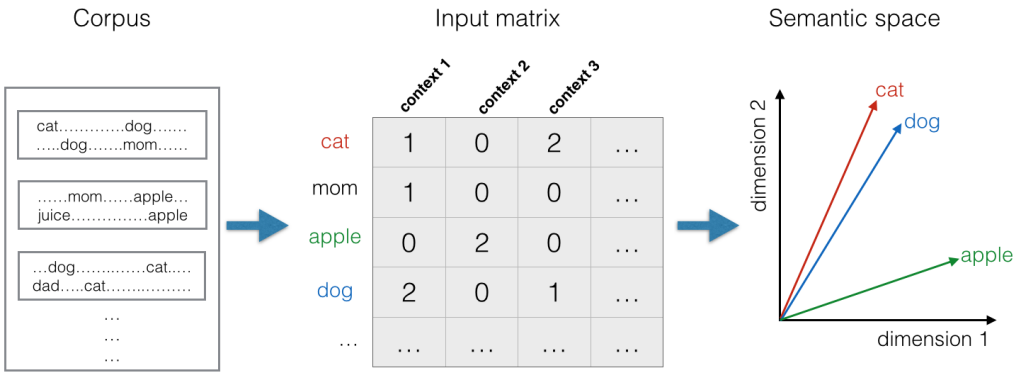

Figure 2.15: LSA takes as input a matrix consisting of rows representing word types, and columns representing contexts. The values correspond to the number of times a word is uttered in a given context. A matrix reduction operation (Singular Value Decomposition) is performed to obtain a compact semantic space. The semantic distance of two words in the resulting space is given by the angle formed by their vectors.

having a higher probability than the correct ones.

2.2.3 Early semantic representation

Recent developmental studies show that infants not only start segmenting words from continuous speech early in their development, they also begin to learn their meaning. For example, infant can recognize highly frequent word-forms like their own names, by as early as 4 months of age (Mandel, Jusczyk, & Pisoni, 1995). They also know the meaning of many words linked to food and body parts by as early as 6 months (Tincoff & Jusczyk, 1999; Bergelson & Swingley, 2012). It is true that the lexicon remains relatively small till around 18 months of age, when word learning speeds up (a phenomenon known as the 'vocabulary spurt'). However, I already argued in Chapter 1 that babies' semantic representation does not necessarily function in a black-and-white mode, and that a rudimentary semantic representation can be learned very early in development through tracking the co-occurrence statistics of words (i.e., the distributional hypothesis).

I model this early semantic representation through a tool borrowed from the field of information retrieval called Latent Semantic Analysis (LSA, Landauer & Dumais, 1997).

This model has proved to be very effective in modeling human similarity judgement in general. For instance, Griffiths, Steyvers, and Tenenbaum (2007) showed that it predicted word ranks in the word association norms collected by Nelson, McEvoy, and Schreiber (1998). The results of Griffiths et al. (2007) were replicated in Fourtassi and Dupoux (2013) using a different learning dataset. In priming studies, Parviz, Johnson, Johnson, and Brock (2011) showed that LSA predicted the strength of the neurological N400 signal (sensitive to semantic relatedness). In this model, a word is represented by a vector. Each cell in this vector represents a context and the value corresponds to the number of times the word is uttered in this context. Thus, the model takes as input a matrix consisting of rows representing word types and columns representing contexts in which tokens of the word type occur (Figure 2.15).

Context

The context is defined as a time window that should be aligned, in principle, with a set of utterances that form a meaningful event. However, there is no obvious way to decide what constitutes a coherent and independent event. It can vary in size depending on the level of specificity and detail required. Moreover, there is no unequivocal way to detect the event's boundaries. Very often, spontaneous discussions (as the ones recorded in the corpora) move from topic to topic in a rather continuous way, and speakers may talk about many things at the same time or talk about the same thing for a long period of time. Rather than segmenting the corpora into events whose significance may vary subjectively, I followed Roy et al. (2015) in setting a fixed contextual window. Here I define this window as a fixed number of utterances, and I take this number as a parameter in the model.

Semantic dimension

As the input matrix is usually sparse and noisy, LSA makes the assumption that there is a set of few underlying variables that encompass the meaning that can be expressed in a given language. Thus, the set of initial contexts is reduced to a fewer set of semantic dimensions over which all words' meanings are distributed. This is achieved through a matrix algebra technique that consists in factoring the initial matrix (Singular Value Decomposition), re-ranking the dimensions according to the values of the resulting diagonal matrix (Singular values), and truncating them to the desired number of semantic dimensions (Landauer & Dumais, 1997). The number of these semantic dimensions is considered as another parameter in this study. The semantic similarity of two words can be obtained in the reduced semantic space by computing the cosine of the angle formed by their vectors (Figure 2.15).

SDT-ρ

In Fourtassi and Dupoux (2013) (paper in Appendix I), we conducted a comprehensive study of both parameters (size of the context and the number of semantic dimensions) and their effect on learning word similarity. More importantly, we developed a technique to set these parameters in an unsupervised way. The method consists in deriving a corpus-based quantity (named SDT-ρ), which measures the stability of the semantic space under different parameter settings. We showed that it predicted accurately the outcome of two human-data-based evaluation methods (the TOEFL synonym test, and word ranks in the word association norms collected by Nelson et al. (1998)). Figure 2.16 shows the evolution SDT-ρ as a function of the number of utterances to be taken as a unit of context, averaging over values of semantic dimensions ranging from 5 to 500. Conversely, Figure 2.17 shows the evolution of SDT-ρ as a function of the number of semantic dimensions, averaging over values of context size ranging from 5 utterances to 500. From these figures, we see that both corpora show an optimal behaviour with a context size of about 50 utterances and a

semantic dimension of 100. Unless otherwise mentioned, we will use these values of LSA to study the properties of our data in the next chapter.

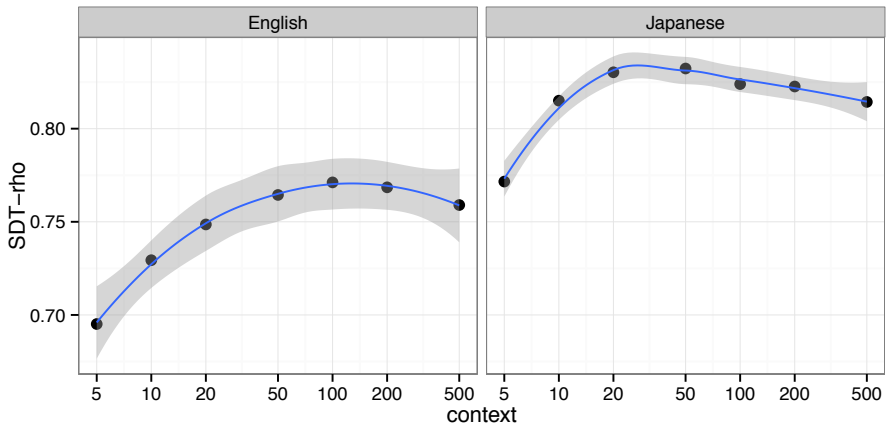

Figure 2.16: SDT-ρ as a function of the number of utterances to be taken as a unit of context, averaging over values of semantic dimensions ranging from 5 to 500.

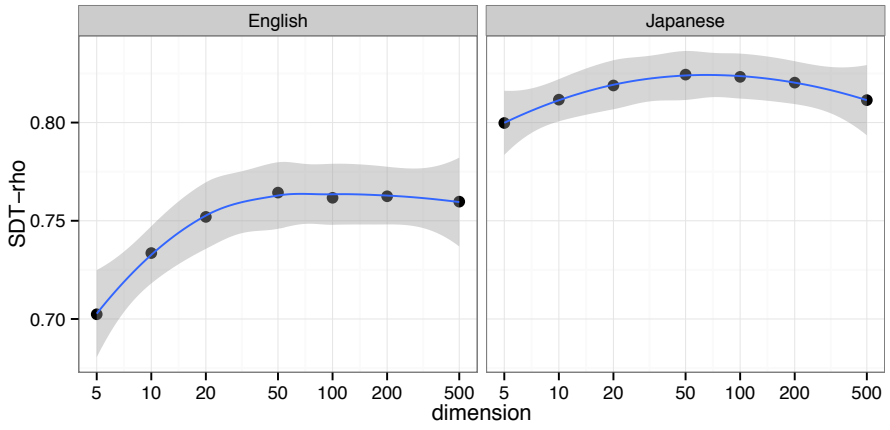

Figure 2.17: SDT-ρ as a function of the number of semantic dimensions, averaging over values of context size ranging from 5 utterances to 500

Chapter 3

Phoneme learning 1: modeling perceptual reorganization

To learn the phonemes of their native language, infants have to undo the irrelevant variation in the input, i.e., the variation that does not cue meaning. Variation could be due to a wide variety of sources such as low level acoustic noise, different properties of the speaker that affect the signal (such as identity, age, sex, mood, or speed), and systematic variation governed by various linguistic rules. As I mentioned in the previous chapter, the present work focuses on systematic variation due to contextual allophones.

In infant studies, learning phonetic categories is experimentally probed through perceptual sensitivity to sound contrasts. Difference in sounds that instantiates two different phonemes (e.g., *ta*-*da*) is perceived better than difference that falls under the same phonemic category (e.g., *ta*-*ta*). Using such a testing paradigm, it has been shown that babies start with a quasi-universal ability to perceive sounds, and, by their first birthday, become less sensitive to the sounds that do not exist in their native language (J. Werker and Tees (1984); but see Narayan et al. (2010) and Sundara et al. (2006)). Moreover, studies found that, around the same age, infants become less sensitive to sound contrasts that *do* exist in their native language, but do not cue meaning (i.e., allophones) (see Seidl and Cristia (2012) for a review). This phenomenon (sometimes referred to as perceptual reorganization) depends naturally on the native language of the baby. In fact, as we saw earlier, sound contrasts that are allophonic in a given language might be phonemic in another one. For example, ([l] vs. [r]) is allophonic in Japanese but phonemic in English, and ([p] vs. [p^h]) is allophonic in English but phonemic in Thai. It follows that the babies' perceptual reorganization depends on aspects of the input that are specific to their native language.
In what follows, I identify some of those aspects, which I call 'cues to phonemicity' (but see Chapter 1 for a more comprehensive list). I chose to test, as a bottom-up mechanism, the acoustic similarity, and as a top-down mechanism, the word-form similarity cue. I will also introduce two implementations of the semantic-similarity-based mechanism proposed in Section 1.4.5.

3.1 Cues to phonemicity

3.1.1 Acoustic similarity

The acoustic cue is based on the assumption that instances of the same phoneme (i.e., allophones) are likely to be acoustically more similar than instances of different phonemes. Linguists such as Trubetzkoy (1939) noted that acoustic similarity is not always a valid cue to phonemicity. In fact, there are cases where very similar phones are judged by native speakers as belonging to different phonemes (as in the case of unreleased voiced/voiceless stops in English, e.g., "cap" vs. "cab"). Conversely, there are cases where very dissimilar phones are felt by native speakers as belonging to the same phonemes (such as [t] and [ʔ] in English). Nonetheless, it is generally accepted that the acoustic cue plays a significant role in early phonetic category learning (Maye et al. (2002); Vallabha et al. (2007); McMurray et al. (2009) to cite but a few). In the particular case of allophonic variation, acoustic salience was proposed as a cue which biases infants' attention towards phonemic contrasts, and away from allophonic contrasts (Cristia & Seidl, 2014).

Acoustic similarity will be tested in Experiment 1. As our input contains phones trained with acoustic models (see Chapter 2), I will take as a measure of similarity the distance between the acoustic models of each pair of phones. In technical terms, the 3-state HMMs of a given pair of allophones are aligned with Dynamic Time Warping (DTW), using Kullback-Leibler divergence as a local distance between pairs of emitting states (where each state is approximated by a single non-diagonal Gaussian). The perceptual distance is defined as the sum of local KL measures between states that correspond to the optimal path provided by DTW.

3.1.2 Word-form similarity cue

The word-form similarity cue is based on the work of Martin et al. (2013). It rests on the intuition that true lexical minimal pairs (/bæt/ vs /pæt/) are not very frequent in

languages, as compared to minimal pairs due to mere allophonic variation ([bæt] vs [bæʔ]). In fact, allophones create alternants of the same lexical items systematically. For instance, the surface form of the first and final phoneme of a word is conditioned by the adjacent sounds. For example, since the English /t/ has many allophones such as [t^h], [ʔ] (glottal stop) and [ɾ] (flap), the word "bat" may take many forms (i.e., minimal variants) such as [bæt^h], [bæʔ] and [baɾ] depending on the phonetic properties of the following segment.

Learning based on the lexical cue can proceed as follows. If learners find a minimal pair of words differing in the first or last segments (e.g., [bæt^h] and [bæʔ]) then they can consider these two segments (e.g., [t^h], [ʔ]) as allophones. Conversely, if a pair of phones is not forming any minimal pair, then they can consider it as phonemic. This binary strategy clearly gives rise to false alarms in the case of true minimal pairs like "bat" and "pat", where the phones /b/ and /p/ would be mistakenly considered as allophonic to each other. However, Martin et al.'s work suggests that their strategy yields good results because correct decisions outnumbers false ones in corpora of natural speech, especially when the number of allophones is relatively high. In order to mitigate the problem of false alarms, one solution was proposed by Boruta (2011). It consists of a continuous version of the binary measure introduced in Martin et al. (2013). For each pair of phones, instead of looking for the presence or absence of a minimal pair, the learner counts the total number of those minimal pairs. The higher this number, the more the pair of phones is likely to be considered as allophonic. More precisely, for a pair of phones, say X and Y, the new cue is defined as the number of lexical minimal pairs that vary on the first segment (XA, YA) or the last segment (AX, AY), where A stands for the rest of the word. Using the word-form similarity cue in learning phonetic categories is consistent with experimental findings. For example Feldman, Myers, et al. (2013) showed that 8 month-old infants pay attention to word level information, and demonstrated that they do not discriminate between sound contrasts that occur in minimal pairs (as suggested by the word-form similarity cue), and, conversely,

discriminate contrasts that occur in non-minimal pairs.

3.1.3 Semantic similarity cues

The semantic similarity cue is based on the intuition that true minimal pairs (e.g., /bat/ and /pat/) are associated with different semantic events, whereas allophonic alternants of the same word (e.g., [bæt^{h}] and [bæʔ]) are expected to co-occur with similar events. As I explained in Chapters 1 and 2, semantic similarity need not rest on a fully developed referential mapping between a word and a unique referent. It can be derived even from an ambiguous situation where words are characterized by features of the general context. For instance, according to the distributional hypothesis, people can develop a sense of semantic similarity of words through keeping track of their co-occurring words.

There are two 'intuitive' way one can define a semantic similarity cue, building on the word-form cue:

- **Semantic similarity cue 1**: one way consists in mapping each pair of phones to the average semantic similarity of the lexical minimal pairs it forms. We know that the vectorial representation of a given word (e.g., w_1), allows us to measure its semantic similarity to any other word (e.g., w_2) by computing the cosine of the angle formed by their respective vectors[1]. For each pair of phones (X,Y), instead of counting the number of minimal pairs of the form (AX,AY) or (XA,YA) as in the word-form similarity cue, we simply compute the average semantic similarity of these pairs. The higher the average semantic similarity, the more the model is likely to classify (X,Y) as allophonic.

- **Semantic similarity cue 2**: alternatively, we can define a semantic cue through computing the sum of minimal pairs (as in the lexical cue), but "weighted" with their

[1]This is equivalent to performing a normalized dot product since we have $Cos(\widehat{\vec{w_1}, \vec{w_2}}) = \frac{\vec{w_1}.\vec{w_2}}{|w_1||w_2|}$

semantic similarity values, as follows:

$$Sem(X,Y) = \sum cos(\widehat{\overrightarrow{AX}, \overrightarrow{AY}}) + \sum cos(\widehat{\overrightarrow{XA}, \overrightarrow{YA}})$$

For every minimal pair, the lexical cue is incremented by one, whereas the semantic cue is incremented by one times the cosine of the angle formed by this pair.

Learning based on the semantic similarity cue could be based on the perceptual learning mechanism known as 'acquired distinctiveness/equivalence'. In fact, it was observed that pairing two target stimuli with different events enhances their perceptual differentiation (acquired distinctiveness), whereas pairing two target stimuli with similar events, impairs their subsequent differentiation (acquired equivalence) (Lawrence, 1949; Hall, 1991). As I mentioned in Chapter 1, this mechanism has been tested in the context of phonetic category learning by Yeung and Werker (2009). In Chapter 6, I will provide a generalization to this experiment, in which learning will not be limited to 'same'/'different' targets, but extended to a graded semantic similarity scale.

3.2 Task and Evaluation

For each corpus, I first list all possible combinations of pairs of allophones. Some of these pairs are allophones of the same phoneme and are labeled "**0**" (allophonic), and others are allophones of different phonemes and are labeled "**1**" (phonemic). Second, each pair of allophones is given a score from the cue that is being tested. The scores are normalized, and vary continuously from **0** to **1**. Figure 3.1 gives an illustration of a typical distribution of allophonic and phonemic contrasts, according to a given cue. The evaluation is a bit tricky. Even though the task is clearly a binary classification, the cues are continuous. We cannot evaluate the classifier in the usual sense (i.e., count the number of hits and false alarms) unless we set a threshold. Nonetheless, there is no purely objective way to set such a threshold because it depends on whether we are statistically "conservative" or

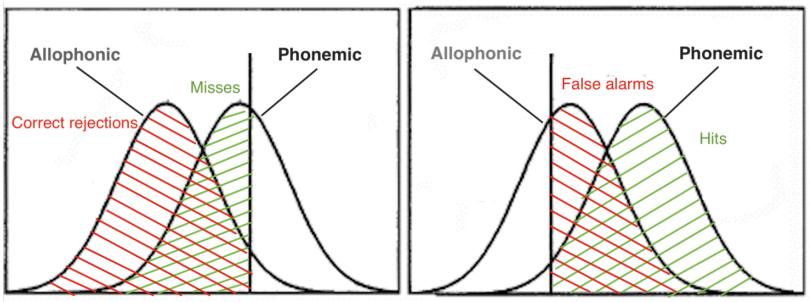

Figure 3.1: An illustration of a typical distribution of allophonic/phonemic contrasts, according to a given indicator of phonemicity. The evaluation of the classification depends on where the threshold is set. On the left, we have a rather "conservative" threshold, and on the right we have a rather "liberal" threshold.

"liberal". In the first case, we tend to have as many 'correct rejections' as possible. We will set the threshold more on the right, but we will also have many 'misses' (Figure 3.1, left). In the second case, we tend to have as many 'hits' as possible. We will set the threshold more on the left and, consequently, end up with many 'false alarms' (Figure 3.1, right). In order to get around this somewhat arbitrary decision, I use a standard technique in signal detection theory called the Receiver Operating Characteristic (ROC). This technique consists of a graph that illustrates the evolution of the proportion of hits as a function of the proportion of false alarms, when we vary the discrimination threshold. Figure 3.2 gives an example of three different classifiers, and Figure 3.3, their corresponding ROC curves. We see that the better the classifier, the less overlap there is between the two distributions, and, interestingly, the further its ROC curve is from the diagonal. In fact, the diagonal represents the ROC curve of a random classifier: the proportion of hits is in nowhere superior or inferior to that of false alarms. This fact is quantified in signal detection theory through measuring the Area Under the ROC curve (AUROC). When using normalized units, the random classifier has an AUROC of 0.5 (as the curve is identical to the diagonal), and the ideal classifier has an AUROC of 1 (as the curve is the furthest from the diagonal). It turn out AUROC has a probabilistic interpretation: it is equal to the probability that a classifier will rank a randomly chosen instance from the right distribution higher than a randomly

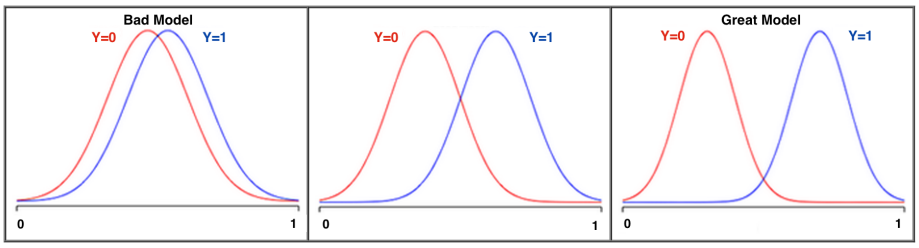

Figure 3.2: Three binary classifiers that vary in their quality from bad (left) to great (right)

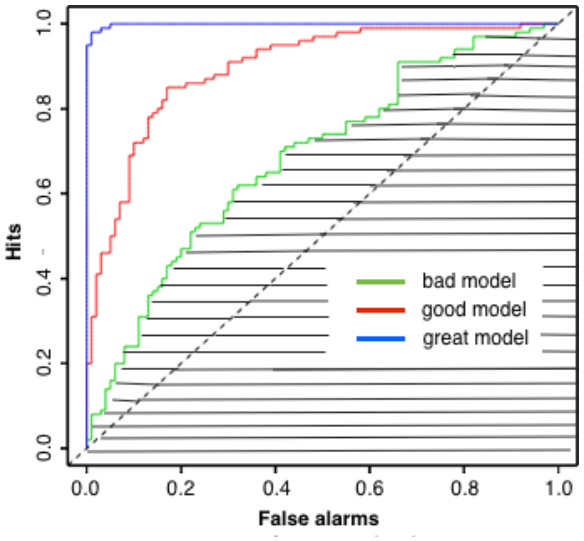

Figure 3.3: The ROC curves corresponding to the three classifiers above. The hatched area represents the Area Under the ROC curve (AUROC) of the 'bad' model. Adapted from Weiss (2008).

chosen instance from the left distribution. In our case, the AUROC reflects the probability that a given cue ranks a random allophonic pair higher than a random phonemic pair.

3.3 Experiment 1

In this first experiment, we test the cues introduced above with the English and Japanese data, and we report the scores obtained using the evaluation procedure described above.

3.3.1 Results

In Table 3.1, I report the scores obtained with the acoustic cue. For obvious reasons, the acoustic cue cannot be computed with artificial allophones, but only with the allophones generated using trained acoustic models (HMMs).

Allo./phoneme	**English**	**Japanese**
2	0.919	0.918
4	0.912	0.886
8	0.878	0.866
16	0.947	0.881

Table 3.1: The AUROC of the acoustic cue as a function of the average number of allophones per phoneme.

I report in Table 3.2 the scores obtained with the word-form similarity cues: the original one introduced in Martin et al. (2013), and the improved continuous version proposed by Boruta (2011). We tested both the lexicon transcribed using artificial allophones, and the lexicon transcribed using HTK based allophonic categories. Similarly, Table 3.3 shows the results obtained with the semantic similarity cues. We tested both versions of the cue, the one based on average semantic similarity, and the one based on semantic weighing of the word-form similarity cue.

	Random Allophones				**HTK Allophones**			
	Martin et al.		Boruta		Martin et al.		Boruta	
Allo./phoneme	**Eng.**	**Japa.**	**Eng.**	**Japa.**	**Eng.**	**Japa.**	**Eng.**	**Japa.**
2	0.753	0.540	0.954	0.986	0.555	0.404	0.609	0.441
4	0.803	0.583	0.964	0.981	0.578	0.486	0.608	0.515
8	0.848	0.676	0.965	0.955	0.576	0.523	0.583	0.538
16	0.900	0.774	0.967	0.910	0.572	0.542	0.573	0.547

Table 3.2: The AUROC of the word-form similarity cues as a function of the average number of allophones per phoneme, using both artificial allophones and HTK-based allophones, and using the gold segmentation.

	Random Allophones				**HTK Allophones**			
	Semantic 1		Semantic 2		Semantic 1		Semantic 2	
Allo./phoneme	**Eng.**	**Japa.**	**Eng.**	**Japa.**	**Eng.**	**Japa.**	**Eng.**	**Japa.**
2	0.925	0.977	0.803	0.560	0.638	0.488	0.649	0.472
4	0.897	0.930	0.786	0.574	0.629	0.546	0.638	0.547
8	0.911	0.900	0.749	0.647	0.593	0.558	0.595	0.560
16	0.922	0.881	0.712	0.655	0.567	0.559	0.567	0.560

Table 3.3: The AUROC of the semantic cues as a function of the average number of allophones per phoneme, using both artificial allophones and HTK-based allophones, and using the gold segmentation.

3.3.2 Discussion

The acoustic score is very accurate (all the scores are around 0.9) for both languages and quite robust to variation. The case of top down cues presents an interesting asymmetry between random and HTK allophones. For the word-form similarity cue, I found, in the case of artificial allophones, a pattern similar to that reported in Martin et al. (2013) using Japanese and Dutch corpora. In fact, the cue's performance starts low, and improves as we increase the allophonic complexity. I also replicated the finding reported in Boruta (2011) where the continuous version of the cue was shown to outperform the original binary one. However, the case of HTK allophones is very different, with both cues performing almost at chance level! Boruta (2012) reported similar low values when using the HTK phones and the CSJ corpus. Here we confirm the same surprising result with a different corpus in a different language. The results obtained with the semantic cues follow generally a similar pattern: an almost chance performance with HTK allophones, and a relatively good performance with Random allophones. As it is consistent across corpora, this asymmetry is unlikely to be the result of an idiosyncratic property of the corpus, as suggested by Boruta (2012). It is now clear that the observed discrepancy is rather due to the way the allophones are generated. This will be discussed in the next section.

3.4 The scope of top-down cues

Top down cues represent the feedback loop in the interactive scenario sketched in Chapter 1, Figure 1.8. However, the feedback is viable only to the extent that low-level information percolates properly into the high level. For instance, if we want to determine the phonemic status of a sound contrast, we might want to know how this contrast affects the lexicon, i.e., whether it forms a minimal pair (e.g [kæt] vs [bæt] for the contrast [k] / [b]), or an allomorphic pair (e.g [kæt^{h}] vs [kæʔ] for the contrast [t^{h}] / [ʔ]). But what happens when a sound contrast does not surface at all as a lexical information, that is, when there is no evidence of this contrast forming a minimal or an allomorphic pair? An example of this situation is the English sounds [h] and [ŋ], which occur in different syllable positions. Should we consider them as allophones of the same phoneme, or as two different phonemes? All phonological theories put them in separate categories, but based on purely low-level considerations, such as the acoustic similarity criterion. We can also find the opposite situation where an allophonic contrast that occurs in complementary distribution (e.g., the English sounds [p] and [p^{h}]) does not necessarily generate an allomorphic pair (later in this section, I will give two reasons this might occur). In such a situation, and contrary to the first example, linguists would consider the sounds as allophones of the same phoneme, based, again, on the phonetic proximity of the contrast, which is a bottom-up criterion. It follows that top-down cues do not cover all cues, but only a subset of the data that percolates into high linguistic levels. The rest of the pairs, which generate neither minimal nor allomorphic pairs are fundamentally beyond their direct[2] scope, and are probably better learned through bottom-up means. I will refer to those as "invisible" pairs, since they are invisible to the lexicon in the sense I just explained.

[2]I said "direct" since I am not considering here the possibility of generalization across phonetic features, but see Maye et al. (2008) and Calamaro & Jarosz (2015).

3.4.1 Invisible pairs and the phonemic status

	Random Allophones		HTK Allophones	
Allo./phoneme	English	Japanese	English	Japanese
2	02.4	00.0	54.5	60.0
4	02.5	00.0	61.5	72.1
8	03.5	00.5	75.8	78.9
16	04.2	07.4	81.8	84.0

Table 3.4: Proportion (in %) of invisible allophonic contrasts out of the total number of allophonic contrasts.

In top-down cues defined above and in previous studies (Martin et al., 2013; Boruta, 2011, 2012), invisible pairs are automatically set to 1, meaning that they are necessary phonemic (because 1 is the highest possible value in the interval [0,1]). This is a correct decision for ([h] vs [ŋ]), but not for an allophonic invisible pair (e.g., [p] vs [p^h], assuming they dont emerge as an allomorphic lexical pair). This assumption in the way lexical top down cues were defined was made (usually in an implicit way) based on statistical evidence from the corpora tested, namely the fact that the number of invisible pairs (especially allophonic ones) is negligible. To what extent does this assumption hold for Random and HTK allophones? In table 4, I show the proportion of invisible allophonic contrasts (e.g., [p] vs [p^h]) out of the total number of allophonic contrasts. In the case of artificial allophones, this proportion varies between 2.4% and 4.2% in English, and between 0% and 7.4% in Japanese. These proportions of bad decisions (in signal detection terms, they represent 'false alarms') can indeed be considered negligible. The same thing cannot be said about HTK allophones where the number of invisible allophonic pairs varies between 54% and 82 % in English, and between 60% and 84% in Japanese. These high proportions mean that the overwhelming majority of allophonic contrasts will necessarily be labeled as phonemic, which is a mistake!

There are basically two factors that could explain why an allophonic contrast would be

invisible to the lexicon. The first factor is of a statistical nature, and consists in the fact that the edges of the word with the underlying phoneme do not appear in enough contexts to generate the corresponding allomorphs. This happens, for instance, when the corpus is so small that no word ending with, say, the French /ʁ/ (e.g. /Knanaʁ/), appears in both voiced and voiceless contexts, to trigger both the voiced allophone [ʁ] and the voiceless one [χ] as in the allomorphic pair [Knanaʁ] vs. [Knanaχ]. However, I could not find any trivial reason why this would affect differently corpora based on Random vs. HTK-based allophones. The Second factor consists in the fact that some allophones are triggered on maximally different contexts (on the right and the left) as illustrated in the allophonic rule of Figure 3.4. When the set of contexts A doesnt overlap with C, and B does not overlap

$$/p/ \rightarrow \begin{cases} [p_1] \,/\, A\_\_B \\ [p_2] \,/\, C\_\_D \end{cases}$$

Figure 3.4: Example of an allophonic rule

with D, it becomes impossible for the contrast ([p_1], [p_2]) to surface as an allomorphic pair of the form ([Xp_1] vs. [Xp_2]) (1), or the form ([p_1X] vs. [p_2X]) (2) (where X stands for the rest of the word). The reason is simply because allophones have to share at least one triggering context to be able to form allomorphic variants of the same word. The shared context should be that of the penultimate segment of the word in the case of an allomorphic pair of the form (1), and the second segment in the case of an allomorphic pair of the form (2). In Appendix D, I explain, using a concrete toy example, why this is more likely to happen in a realistic procedure that makes use of acoustic/phonetic similarity (such as HTK software).

3.4.2 The proportion of invisible pairs in natural speech

In Table 3.5, I computed the proportion of invisible pairs (out of the total number of pairs) as a function of the average number of allophones per phoneme. As we can see, this proportion increases when we increase the number of allophones. To test if this proportion is constrained by the size of the corpora used in this study, I varied the amount of data available to the learner from 1 hour of speech to 40 hours. The observed monotonic decrease in Table 3.6 suggests that invisible pairs can, by extrapolation, be reduced beyond the numbers obtained in Table 3.5, where speech corpora in both English and Japanese are limited to 40 hours of speech.

Allo./phoneme	**English**	**Japanese**
2	65.3	41.6
4	76.4	69.7
8	90.2	83.3
16	95.4	91.9

Table 3.5: The proportion (in %) of invisible pairs as a function of the average number of allophone per phoneme, generated by HTK.

Corpus size (hours)	**English**	**Japanese**
1	77.8	68.7
2	76.9	62.8
5	74.1	56.9
10	70.6	53.0
20	68.5	46.3
40	65.3	41.6
∞	58.9	19.0

Table 3.6: The proportion (in %) of visible pairs as a function of the size of the corpus, in the case of 2 allophones/phoneme in average (generated using HTK)

Is there an inferior limit to this subset of pairs, or can we reduce them (with more data) to zero? To answer this question, I generated an artificial corpus that uses the same lex-

icon in both languages, but with all possible word orders so as to maximize the contexts for words' edges. This artificial corpus decreases the proportion of invisible pairs down to about 58.9% in English and 19 % in Japanese. This non-zero limit is consistent with the explanation provided in the previous subsection and Appendix D. In fact, even under all possible triggering contexts, there is still an irreducible set of invisible pairs whose context triggering sets are not overlapping.

To sum up, I investigated in this section the systematic discrepancy in performance when the cues to phonemicity are tested with either Random allophones or HTK-based allophones. The investigation dealt with the nature of the interface formed by the phonetic representation and the corresponding lexicon. To this end, I introduced the notion of 'invisible pairs', a notion that translates the fact that some phone pairs generate a minimal or allomorphic pair (visible), and other phone pairs do not (invisible). I argued for the fact that top-down cues cannot be used universally to learn the phonemic status of all sound contrasts. In particular, the approximation that consists in considering all invisible pairs as phonemic, does not scale-up to realistic input (i.e., HTK-based allophones), causing systematic classification errors. In experiment 2, I will test a prediction that follows from this analysis. In fact, if top-down cues are mainly compromised by invisible pairs, then they should perform well on the set of visible pairs.

3.5 Experiment 2

In this experiment I apply top down cues to the subset of visible pairs. I test the effect of the average number of allophones per phoneme, in the case where top-down cues are computed using the ideal segmentation, the output of unsupervised segmentation, and a completely random segmentation. Note that, in this new framing, Martin et al.'s cue is completely uninformative since it assigns the same value to all visible pairs, it was therefore

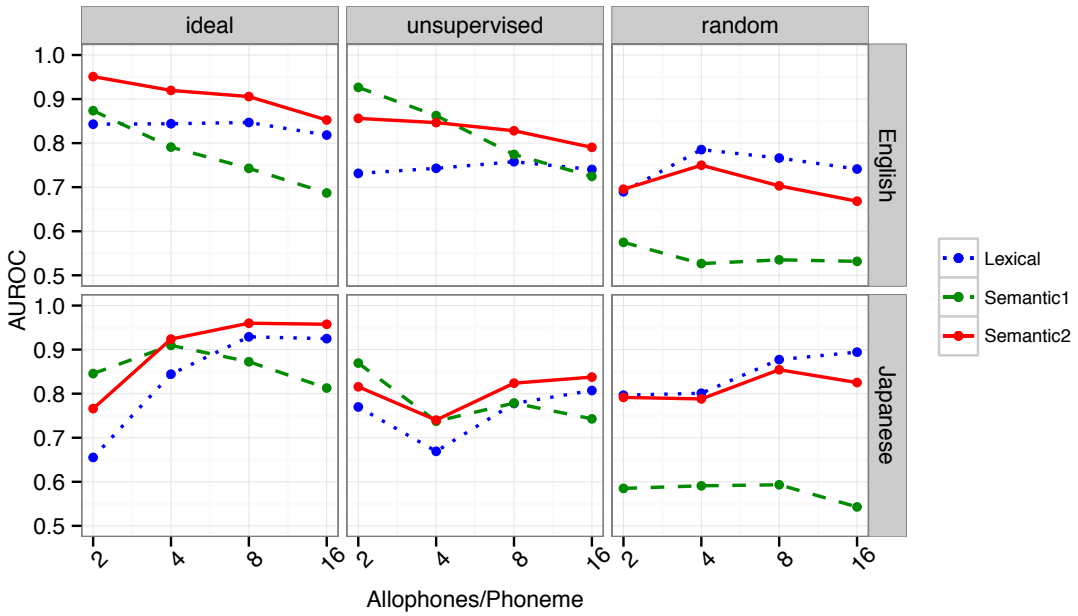

Figure 3.5: The AUROC of top down cues as a function of the average number of allophones per phoneme, and as a function of the quality of the segmentation: ideal, unsupervised and random

omitted from the analysis. The remaining cues are the continuous word-form similarity cue (Boruta, 2011, 2012), and the semantic similarity cues (this study).

3.5.1 Results

As predicted by the hypothesis developed in Section 3.4, the overall accuracy of the cues on the subset of visible pairs is quite high, even in the case of unsupervised and random segmentation (Figure 3.5). The word-form similarity cue is robust to extreme variation and to segmentation errors. It performs pretty well on high allophonic complexities and almost regardless of the quality of the segmentation. The associated AUROC generally remains above the decent value of 0.7 (chance level is 0.5). In contrast, the semantic similarity cue 1 gets better when the allophonic complexity decreases. It performs relatively well on an intermediate quality segmentation (unsupervised), but falls down to around chance in the

case of random segmentation. Finally, the semantic similarity cue 2 tends to combine the advantages of the word-form similarity cue and the semantic similarity cue 1, at both the level of allophonic variation and the segmentation quality. In order to see how the cues perform with different amount of learning data, I show in Figure 3.6 and Figure 3.7 the scores for portions of the corpora inferior to 40 hours of speech (the approximate original size in both languages). I focused on the two interesting cases where the semantic cue 1 performs better than the word-form cue, i.e., the case of 2 and 4 allophones per phoneme in average, respectively. The results tend to mimic the trends obtained with allophonic complexity. The lexical cue is, again, highly robust to the scarcity of the data, the semantic cue 1 tends to do better when data increases, and the semantic cue 2 does well whenever one of the other cues shows a high performance.

3.5.2 Discussion

Although not perfect, the lexical cue is robust to high allophonic complexity, extreme segmentation errors (random), and to the scarcity of the data. This, in part, replicates the finding of Martin et al. (2013) and Boruta (2011) where the word-form similarity cue was shown to perform well with high allophonic complexity and relatively bad word segmentation (N-grams). The semantic cue 1, in contrast, depends on all the above dimensions. In fact, it only becomes interesting (i.e., trumps the word-form cue) when data is sufficient, reasonably segmented and presenting relatively low variation. In fact, these are all cases where word co-occurrence statistics becomes consistent enough to accurately indicate word similarity. When the allophonic complexity is high, the frequency with which each word type appears is low, let alone the frequency of its co-occurrence with other words. Similarly when the amount of data is small, words do not co-occur in enough contexts to lead to a reliable word similarity pattern, and when the segmentation is random, it is obvious that the resulting words would not co-occur in a consistent fashion. The pattern of the semantic cue 2 is very

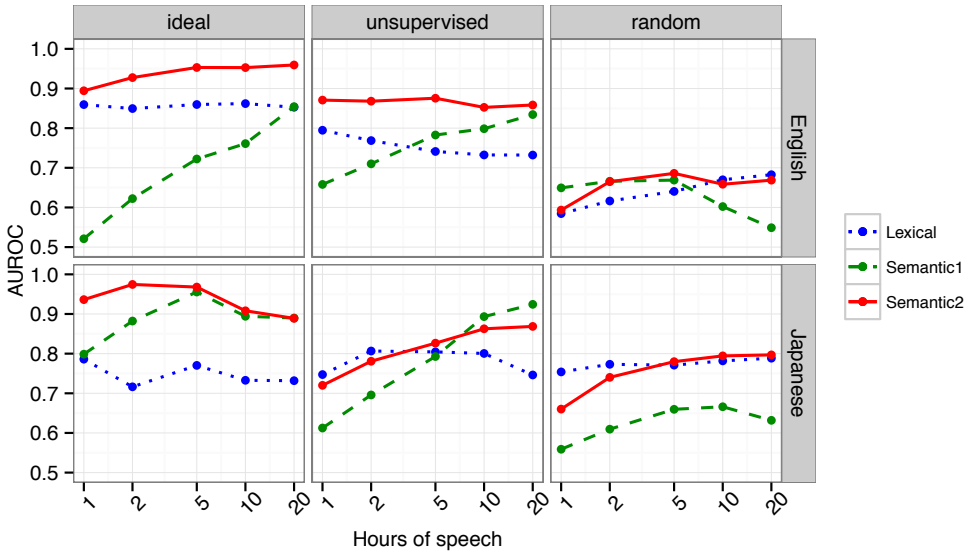

Figure 3.6: The AUROC of top down cues as a function of the size of data available to the learner, in the case of 2 allophones per phoneme, and as a function of the quality of the segmentation: ideal, unsupervised and random

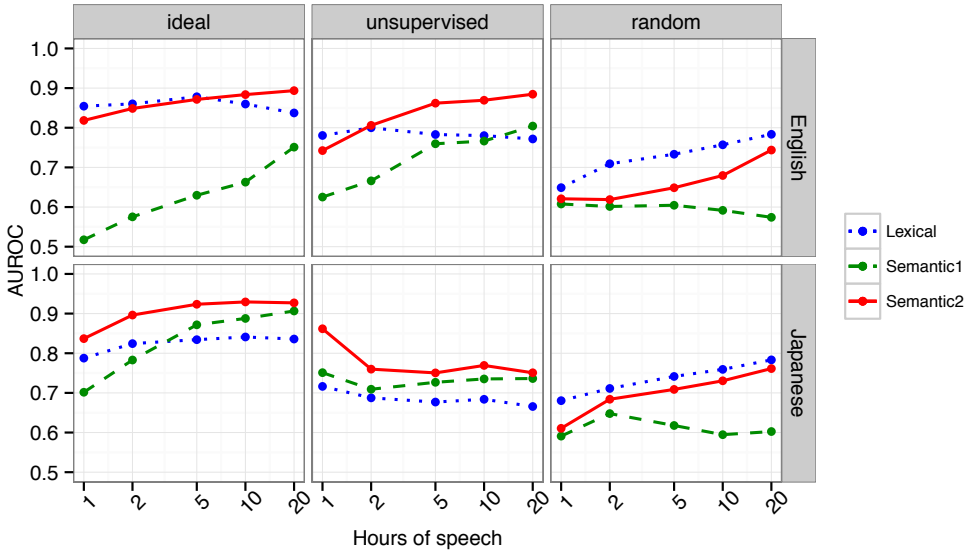

Figure 3.7: The AUROC of top down cues as a function of the size of data available to the learner, in the case of 4 allophones per phoneme, and as a function of the quality of the segmentation: ideal, unsupervised and random.

interesting. It seems to maintain the robustness of the lexical cue to extreme situations, while showing performance generally higher than both the word-form cue and semantic cue 1. This naturally follows from the definition of the cue, which could be seen as an intuitive combination of the word-form cue and the semantic cue 1.

3.6 General discussion

This chapter dealt with the question of phoneme learning, as characterized by the learner's differentiated sensitivity to phonemic vs. allophonic contrasts. It investigated various aspects of the input that might bias the learner's attention towards phonemic contrasts, and away from allophonic contrasts. As a bottom-up cue, I tested the acoustic similarity between phone pairs. The cue gives high scores, allowing the model to distinguish accurately between relevant and irrelevant variations. In top down cues, I distinguished between the word-form similarity cue introduced in (Martin et al., 2013), and the two semantic cues introduced in this study. The word-form cue fares relatively well under extreme conditions, whereas the semantic cue 1 requires some degree of intelligibility in the input to trump the word-form cue. In fact, the semantic cue 1 performs well under the conditions that allow the model to learn a consistent pattern of word co-occurrence. Finally, the semantic cue 2 could be seen as a form of combination between the word-form cue and semantic cue 1. It benefits from the strengths of both cues, in that it resists to extreme conditions, and gives generally very high scores. Nonetheless, the scope of top-down cues is restricted to pairs of sound that percolate into the lexical level, giving rise to either a minimal pair or an allomorphic pair. In fact, top-down information boils down to knowing whether two word-forms are instances of the same lexical item (variation is considered allophonic), or to two different lexical items (variation is then considered phonemic). Phonetic variation that, for some reason, does not cause lexical variation (such as the English [h] vs. [ŋ]) is, in this particular sense, "invisible" to the lexicon, and is better learned through bottom-up means.

Such form of 'division of labour' between top-down and bottom-up cues is not unusual in phonological theories. For instance, although they both occur in complementary distributions and do not generate lexical minimal pairs, the English pair ([h] vs. [ŋ]) is treated differently from the pair ([p] vs. [p^h]), based on a bottom-up criterion. The present study makes the suggestion that this division of labour also makes sense in the perspective of a learning theory.

The chapter also addressed the chicken-and-egg problem of phoneme learning (Fourtassi & Dupoux, 2014). In fact, phonemes are classically defined by their ability to contrast word meanings (Trubetzkoy, 1939), and therefore, require semantic top-down information. However, since the quality of the semantic representation depends on the quality of the phonemic representation that is used to build the lexicon, we face a circularity problem. In this chapter, I proposed a way to break the circularity by building approximate representation at different linguistic levels. The infants' initial attunement to language specific categories was represented in a way that mirrors the linguistic and statistical properties of the speech. I showed in Chapter 2 that this detailed (proto-phonemic) inventory enabled word segmentation from continuous transcribed speech, but resulted in a low quality lexicon 2.14. The poorly segmented corpus was used here to derive a semantic similarity metrics between pairs of words, based on their co-occurrence statistics. The results showed that information from the derived lexicon and semantics, albeit very rudimentary, helped discriminate between allophonic and phonemic contrasts, with a high degree of accuracy. Thus, this work strongly supports the claim that the lexicon and semantics play a role in the refinement of the phonemic inventory (Feldman, Myers, et al., 2013; Frank, Feldman, & Goldwater, 2014), and, more importantly, that this role remains functional under more realistic assumptions (i.e., unsupervised word segmentation and ambiguous semantics). We also found that lexical and semantic information were not redundant and could be usefully combined, as we can see from the performance of the semantic cue 2. That being said,

this work relies on the assumption that infants start with initial perceptual categories (allophones), but I did not show how such categories could be constructed from raw speech. More work is needed to explore the robustness of the model when these units are learned in an unsupervised fashion (Lee & Glass, 2012; Huijbregts, McLaren, & van Leeuwen, 2011; Jansen & Church, 2011; Varadarajan et al., 2008).

Finally, this work could be seen as a proof of principle for a learning mechanism, whereby phonemes emerge from the interaction of low level perceptual categories, word-forms, and semantics (see J. Werker and Curtin (2005) for a similar theoretical proposition). One question remained unanswered though. In fact, even if we know how phonetic categories are organized in the perceptual space, we still need to know how many categories are relevant in a particular language (i.e., where to stop the categorization process). The next chapter proposes a solution to this problem, through the notion of Semantic Consistency. In brief, it is suggested that the optimal level of clustering is also a level that globally optimizes the semantic consistency of lexical categories. Too broad allophonic categories result in many incoherent homophones, but too detailed allophonic categories result in unnecessary synonymous. Somewhere in the middle, the optimal number of phonemes optimizes the consistency and the parsimony of the lexicon.

Chapter 4

Phoneme learning 2: the number of categories

The previous chapter dealt with perceptual reorganization of sounds. I identified cues in the input that purport to account for the fact that phonemic sound contrasts end up being perceived better than allophonic contrasts. The derivation and evaluation of these cues were designed with the perspective of a rather 'soft decision' model. That is, for a given pair of phones, the cues provided a continuous measure that indicated the degree of phonemicity of this pair. At the evaluation level, even if the cues were compared against a binary gold standard (**0** for allophonic, **1** for phonemic), I did not fix a particular threshold of "contrastiveness". Rather, I used the AUROC technique that spans all possible thresholds and provides a kind of summary evaluation of the cues under different thresholds. Nonetheless, to acquire the sound inventory of their native language, infants should also learn the 'hard decision'. They should be able, not only to tell how phonemic a contrast is, but also to learn how much phonemicity is contrastive.

If we approach the problem of phoneme learning through clustering analysis, then we can draw an interesting parallel between, on the one hand, the continuous cues and the 'distance function', and, on the other hand, the threshold of contrastiveness and the 'number of clusters'. In clustering analysis we usually need both the distance function and the number of clusters to perform the task successfully. I assume, similarly, that phoneme acquisition requires learning an additional parameter that specifies the number of contrastive units in the sound inventory.

4.1 Problem specification

4.1.1 Learning as optimization

How do learners converge on the number of phonemes in their native language? Suppose we start with a continuous measure that reflects the degree of phonemicity of various sound contrasts. This continuous information is not enough to tell us where to draw the functional

boundary. One thing we can do, instead, is to form a hierarchy of clusters, i.e., to group phones into varying numbers of nested categories. Then learning can be characterized as an 'optimization problem': which level of clustering is the optimal one, given the set of possible clusterings? To answer this question, we first need to specify the sense in which the phonemic level is supposed to be optimal. In machine learning terms, we need to define a real valued 'objective function' f, whose optimum is achieved by phonemes. In a formal sense, if C refers to the set of potential clusterings, and c_{phm} stands for the phonemic level of clustering, then we should have $f(c) \leq f(c_{phm})$, for all c in C.

Framed in this way, the phonemic representation becomes a special case that results from a particular choice of the objective function. Different objective functions can, in principle, be defined for various representations of the data. For example, a coarse-grained inventory composed of consonants and vowels (2 categories) can result from an objective function linked to the task of (re)syllabification, and a fine-grained representation composed of allophones can result from an objective function linked to the task of perception/production of allophonic rules[1]. Here we are looking for an objective function that results in an intermediate level of clustering (phonemes), and whose task is to represent and contrast lexical meanings.

In order to simulate the learning situation of the baby, the objective function that leads to the phonemic representation should not be endowed with any prior knowledge about the lexicon. In machine learning terms, the procedure should be *unsupervised*. Nonetheless, given that phonemes are supposed to represent lexical meaning, how is it possible to learn them without supervised information that specifies, for instance, whether two word-forms represent one or two lexical items? In the modeling literature, many of the proposed strategies fall under one of two extremes. In the first one, we suppose a staged approach to language learning and assume that the learner settles on the sound inventory before

[1]This framing is compatible with developmental data showing that babies are able to access different representations of the input depending on the task (Werker and Curtin, 2005)

learning the words, or the words' meanings. According to this view, no top-down feedback information is possible. In the other extreme, we usually assume a fully developed lexicon, and use it to learn the lower units of speech. Obviously, neither of these extremes can be taken as a realistic learning mechanism (See Chapter 1 for a detailed discussion). In the following, I propose an intermediate solution, which neither precludes the effect of words, nor presupposes their full development. More precisely, I make the assumption that a candidate representation at the low level (i.e., a phonological analysis) can percolate into the high level (i.e., lexicon). Thus, the top-down effect can take place through evaluating the induced lexicons, and using this evaluation to select the optimal phonological interpretation of the input.

4.1.2 Semantic Consistency

A candidate phonological analysis vary in theory from coarse-grained (e.g., consonants/vowels) to fine-grained (e.g., detailed allophones). When used to represent words, they give rise to lexicons with different 'resolutions'. Suppose, for example, that the learner hears the following acoustically detailed word instances: [kæʔ], [kæt^{h}], [bæʔ] and [bæt^{h}]. The learners can, a priori, identify these instances as four lexical items (i.e., /kæʔ/, /kæt^{h}/, /bæʔ/, /bæt^{h}/), if they choose a fine phonological interpretation. They can also identify them as two lexical items (i.e., /kæt/, /bæt/) if they choose an intermediate level of phonological analysis, or one item (i.e., /CVC/) if they choose a coarse analysis. Crucially, these lexicons will have different distributional and semantic properties, which I propose to characterize as follows. Suppose that each word token w is associated with a semantic representation $SR(w)$, and suppose we have a similarity function Sim defined over the set of pairs of semantic representations. The Semantic Consistency of a word type W (hereafter, SC(W)) can be defined as the average similarity of all pairs of word tokens (w, w') that belong to the word category

W, according to a given phonological analysis. It is computed as follows:

$$SC(W) = \frac{1}{|W|^2} \sum_{(w,w') \in W^2} Sim(SR(w), SR(w'))$$

A word whose instances share consistently similar semantic feature has a high Semantic Consistency value. For example, if most tokens of the word "breakfast" are uttered in the kitchen, in the morning, and co-occur consistently with words such as "food", "eat", "bread"... then its SC score will be high. If, in contrast, the tokens of the word occur in many locations, at different times of the day, and with unrelated words, then its SC score will be low (e.g., a function word like "the").

In the following section, I explain how key concepts (such as 'phonological analysis', 'induced lexicon' and 'semantic consistency') will be implemented and tested in our modeling scheme.

4.2 Representations

4.2.1 Levels of phonological analysis

We generate different levels of phonological analyses for English and Japanese data, starting from the ideal (i.e., phonemic) inventory. To generate categories coarser than the phonemes, I collapsed the segments in English from 41 phonemes to 19, then to 10, 4 and 2. Similarly, I collapsed the segments in Japanese from 25 to 13, 8, 4 and 2 (see Annexe E for the details of this hierarchical clustering). To generate categories finer than the phonemes, I consider contextual allophones. As explained in Chapter 2, a given phoneme is split into possibly several allophones as a function of its left and/or right phonetic context. In order to generate these allophones in a phonetically and acoustically controlled fashion, I followed the HTK procedure (see Chapter 2). I generated inventories of various sizes (from 2 to 16 times the size of the phonemic inventory). Note that the size of an inventory increases as a

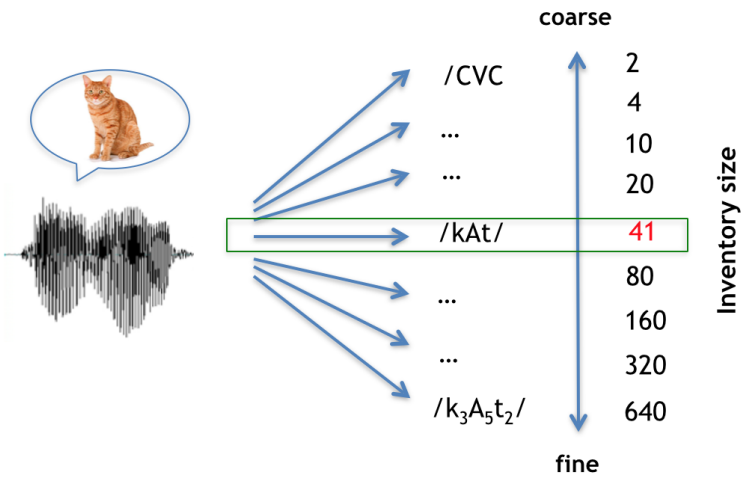

Figure 4.1: Upon hearing the sound "cat", the English-learning infant can a priori represent it through phonetic categories at different resolutions (different phonological analyses). The challenge is to select the "optimal" level of representation

function of the phonetic detail considered.

4.2.2 Induced lexicon

I transcribed the corpora using each of these alternate inventories. The resulting lexicon has, therefore, a representation that varies along the phonological analysis on which it is based. For example, the lexical item "cat" can have the following representations: /CVC/, /kæt/, or /kæt^{h}/ (Figure 4.1) Moreover, in order to approximate infants' early segmentation of words, I, similarly, followed the procedure described in Chapter 2, and used the state-of-the-art unsupervised word segmentation model called Adaptor Grammar (M. Johnson et al., 2007) to segment the corpora transcribed with different inventories. Figure 4.2 shows the token F-score of word segmentation in each case. The F-score is computed by comparing the segmentation under a given inventory with the ideal segmentation under the same inventory. It shows that the segmentation is optimal for the phonemic inventory in the case of Japanese, and with the slightly coarser grained inventory in the case of English. The segmentation performance drops for both finer- and coarser-grained inventories. The

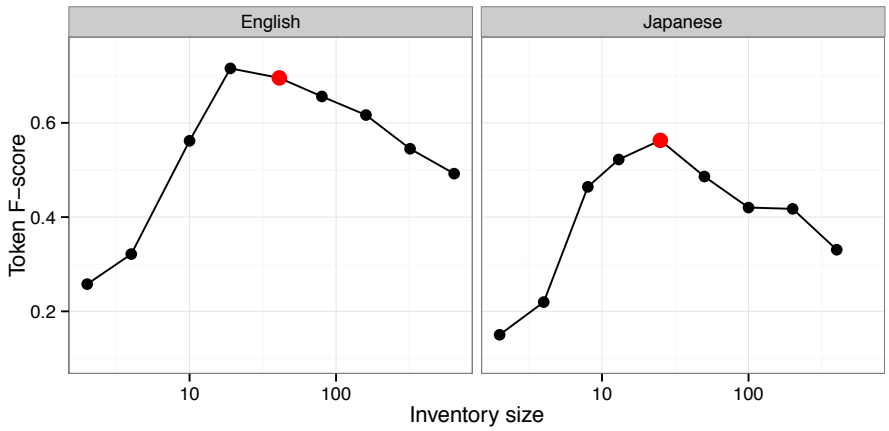

Figure 4.2: Token F-score of optimal segmentations of English and Japanese corpora transcribed with inventories of different sizes, using a collocation adaptor grammar model. The red color refers to the phonemic inventory.

fact that the phonemic inventory leads to an almost optimal segmentation is an encouraging result. However, information about the segmentation performance was based on the comparison with the ideal segmentation, to which, the learner does not have access. My objective, in what follows, will be to obtain the phonemic inventory as an optimal value of a rather unsupervised evaluation metrics (based on the notion of semantic consistency).

4.2.3 Semantic Consistency

As explained earlier in this dissertation, I use word co-occurrence as a proxy for the general multi-modal semantic context, and Latent Semantic Analysis as a modeling framework. Remember that in LSA, a word is represented by its frequency distribution over different contexts (time windows). LSA allows us to compute the semantic similarity between words (the cosine between word vectors in the semantic space). Two words have a high semantic similarity if they have similar distributions, i.e., if they co-occur in most contexts.

The idea, as formalized in Subsection 4.1.2, is to probe the semantic consistency of word types by measuring the extent to which their tokens are semantically similar to each other. In our modeling framework (i.e., LSA), it is easier to derive representations for word types,

than to derive representations for word tokens. Therefore, I chose to replace the notion of a token by the more general (and more flexible) concept of a *sub-type*. As its naming suggests, a sub-type of a word type is a subset of its tokens. For example, the type /kæt/ has [kæʔ] and [kæt^{h}] as possible sub-types. In the following section, I describe two ways one can derive sub-types from the vectorial representation of a type, and I test their subsequent effect on learning the phonemic inventory.

4.3 Experiments

4.3.1 Experiment 1: Random partitioning

In this experiment, I propose to derive sub-types through partitioning the set of tokens in two random subsets. It is illustrated schematically in Figure 4.3, and it is computed as follows. For each level of phonetic clustering, a corpus is generated by transcribing its utterances according to the phonetic inventory at hand. From this original corpus I derive a "sub-corpus", where each word type is randomly replaced by one of two lexical variants. For example, the label of the word 'cat' is replaced in the sub-corpus by two different labels: 'cat_1' or 'cat_2'. Thus, each word that occurs at least twice is duplicated, and each variant (or sub-type) appears with roughly half of the frequency of the original word. After applying LSA to the derived corpus, we obtain for each word type vector (e.g., $\vec{cat}$) two sub-type vectors ($\vec{cat}_1$ and $\vec{cat}_2$). The semantic consistency of a word type is simply the cosine of the angle formed by the two sub-vectors in the semantic space.

In order to obtain a semantic consistency score for the entire lexicon, I test how the measure of semantic consistency allows us to distinguish sub-types from random pairs of words. To this end, I use the Receiver Operating Characteristic curve to compare the distribution of cosine distances across the entire list of sub-types to that of an equivalent number of random pairs. From this curve, I derive the Area Under the ROC curve (AUROC). The

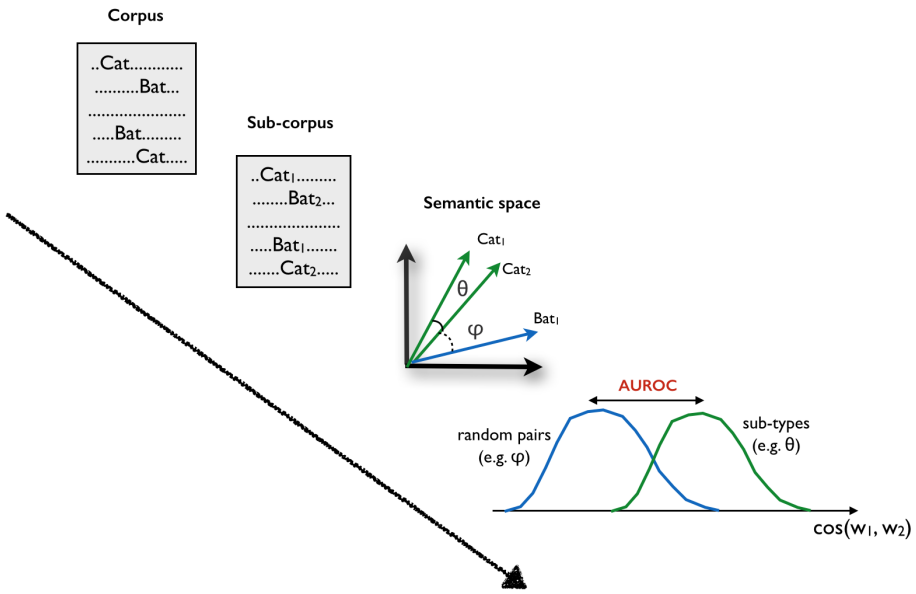

Figure 4.3: Schematic description of sub-type derivation and SC score computation, using random partitioning.

resulting score can be interpreted as the probability that, given two pairs of words, of which one is a sub-type pair, the pairs are correctly identified based on semantic similarity. A value of 0.5 represents pure chance, and a value of 1 represents perfect performance. Note that the SC score depends on the LSA parameters: the size of the context and the dimensions of the semantic space. We thus test the robustness of the score when we vary these parameters. For each representation of the lexicon, we compute different SC scores for values of context size ranging from 5 and 500 utterances, and for semantic space dimensions ranging from 5 to 500 dimensions. In Chapter 2 (Figure 2.16 and Figure 2.17), I showed that these values cover a reasonable portion of the parameter space (STD-ρ tends to decrease at both endpoints). Finally, we distinguish three kinds of segmentations: ideal, unsupervised and random.

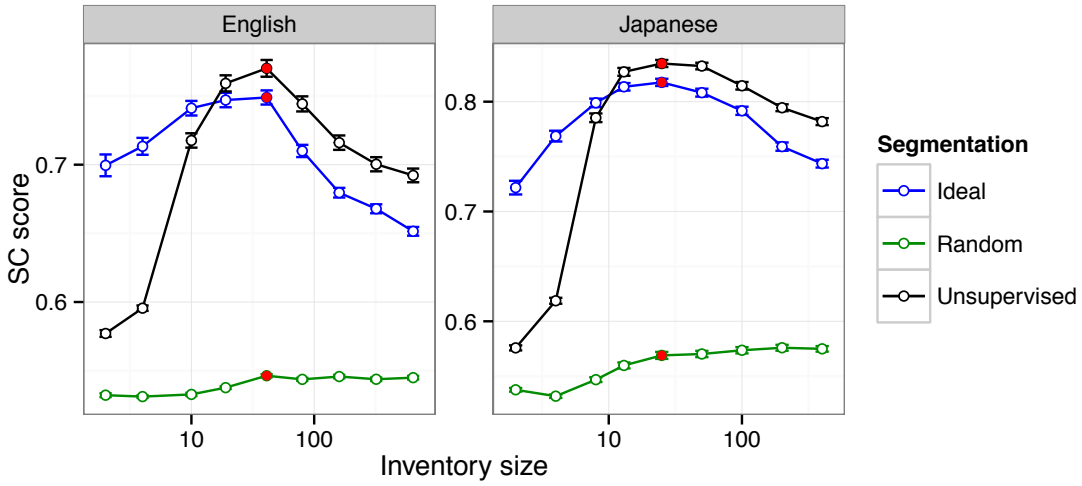

Figure 4.4: Semantic Consistency scores across different phonetic inventories and different levels of word segmentation, using Random Partitioning. The white points and error bars show the means and standard errors over different parameter settings. The black points refers to the phonemic inventory of each language

Results and discussion

Results are shown in Figure 4.4. As expected, for unsupervised and ideal segmentation, the score peaks at the phonemic inventory of each language (41 in English and 25 in Japanese). The absence of such peak in the random segmentation demonstrates that the result is not a mere artifact of the way the phonetic inventories were generated, but, rather, a consequence of the way this variation affects the semantic representation of the lexicon. When the inventory is small, the lexicon is less semantically coherent, since it has more homophones. For example, in an inventory composed of coarse-grained natural classes, two words that have rather orthogonal semantics, like /kæt/ and /bæg/, will be treated as tokens of the same type, since all the consonants belong to the class of stops. This type will not have a consistent distribution, since it occurs in contexts that are not necessarily semantically related.

Fine-grained inventories, on the other hand, increase the number of types, which therefore occur with a lower frequency. This makes the contextual representation statistically sparse.

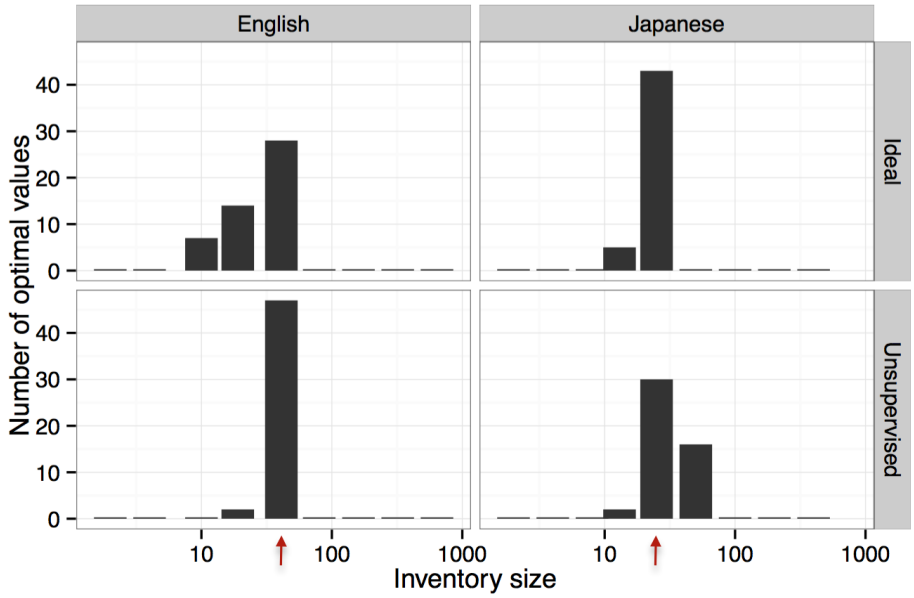

Figure 4.5: Histogram of the SC score peaks across different parameter settings, using Random Partitioning. The red arrows point towards the phonemic inventory.

For example, the case of maximal variation leads to a token/type ratio inferior to 3 in the English corpus (compared to 30 in the phonemic inventory) and a ratio inferior to 6 in the Japanese corpus (compared to 33 in the phonemic inventory). Such level of detail eventually affects the semantic coherence of types, since the surrounding context will tend to differ from token to token.

For a given inventory, the SC score distinguishes between ideal and supervised segmentations on the one hand, and random segmentation on the other. The reason is apparent: the distribution of a type across contexts will not be consistent in the case of a random segmentation. However, the SC score does not distinguish consistently between ideal and unsupervised segmentations. Figure 4.4 also indicates that the utility of the SC score in picking out the best representation for the lexicon is relatively independent of the parameter settings (the error bars are over runs with different parameters). Figure 4.5 shows the histograms of the SC score peaks across different parameters. That is, for each value of context size and semantic dimension, I select the inventory at which the SC score peaks,

and increment it by one. The resulting histograms show that, indeed, the SC score enables us to select the right inventory without any parameter tuning.

In order to test the effect of the data size on learning, I derived scores for different portions of the data, ranging from 100% (about 50.000 utterances) to 0.1% (about 50 utterances). Results are shown in Figure 4.6. The model succeeds in learning the phonemic inventory starting from 5.000 utterances, and becomes completely uninformative with 50 utterances, a case in which values tend to fluctuate around chance level. The case of 500 utterances is interesting. Though the scores are not around chance, they do not peak at the phonemic inventory, but at the next coarse-grained inventory in both English and Japanese. This problem will be explained and addressed in the discussion of the next experiment.

4.3.2 Experiment 2: nested-category-based partitioning

In this experiment, we use another method to derive sub-types. This method consists in basing the partitioning of the set of tokens on nested phonetic representations. An illustration is given in Figure 4.7, and it proceeds as follows.

A corpus is generated for each level of phonetic clustering, as in the previous method. Now instead of deriving a sub-corpus, I simply consider the corpus generated by the directly lower phonetic level, consisting of a finer grained inventory. For a given word type at the high level (e.g., "cat"), its occurrences are matched with their equivalents at the low level (e.g., "ca[ʔ]" or "ca[t^h]"). Thus, each word type at one level is now mapped to its sub-types on the lower levels. The number of sub-types can vary from one to many, depending on the segments composing the word, and more precisely, on how many contextual allophones they have. The semantic consistency of a word type is the average cosine distance of all possible pairs of sub-types. In order to characterize the semantic consistency of the entire induced lexicon, I compare the distribution of sub-type pairs sampled from word types, to that of an equivalent number of random pairs of words. I finally compute the resulting AUROC

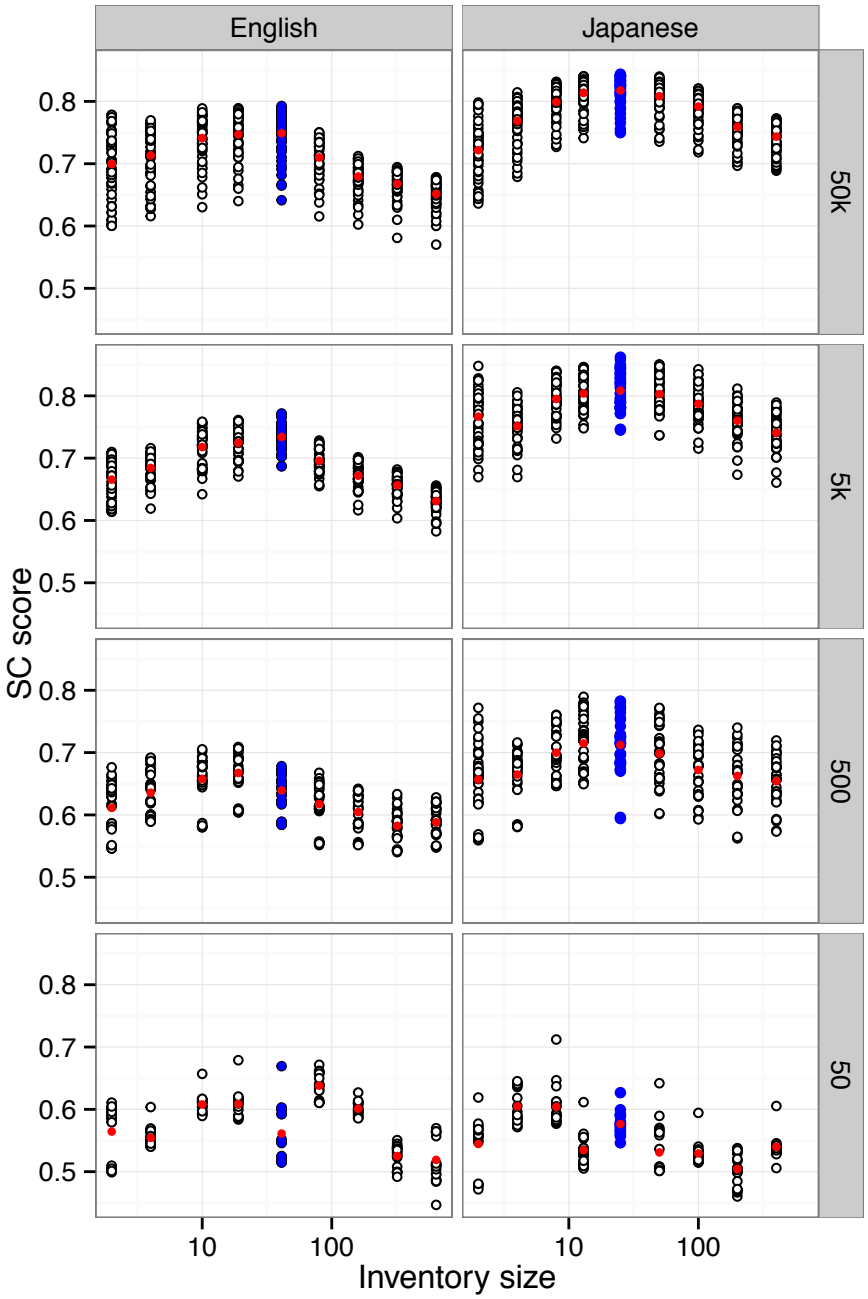

Figure 4.6: Semantic Consistency score of lexicons represented with different phonetic inventories, and for corpus sizes ranging from 100% (about 50.000 utterances) to 0.1% (50 utterances). The SC score is computed using random partitioning, and the ideal word segmentation. The white points refer to the individual scores using different parameter settings, the red points to the phonemic inventory, and the blue to the means of the individual scores.

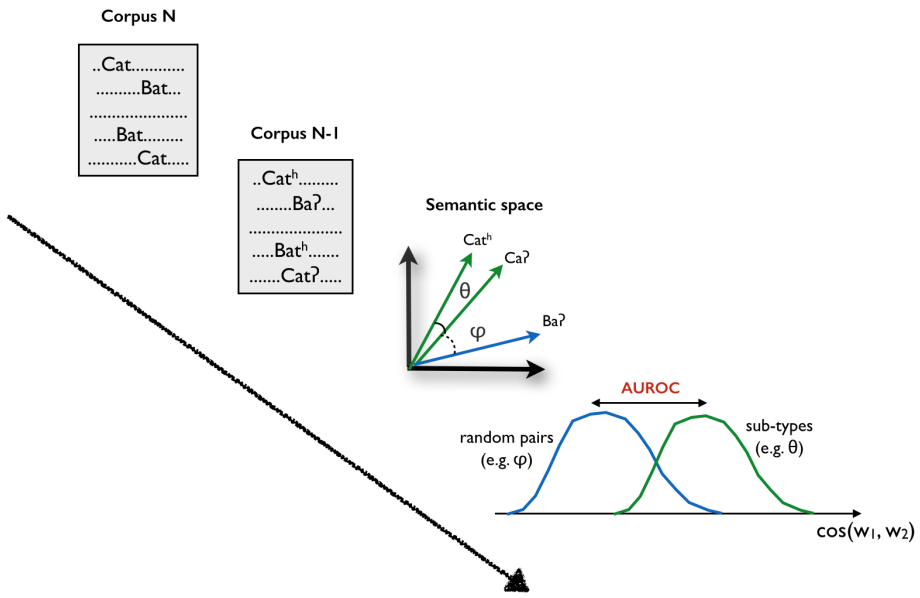

Figure 4.7: Schematic description of sub-types' derivation and Semantic Consistency score computation, using nested-category-based partitioning.

that I take as the final SC score of the induced lexicon. As in the previous experiment, for each representation of the lexicon we compute different SC scores for values of context size ranging from 5 and 500 utterances, and for semantic space dimensions ranging, similarly, from 5 to 500 dimensions. As for word segmentation, the nested-category-based partitioning requires an exact match between tokens at different hierarchical levels, i.e., the segmentation should remain the same. So, unlike the previous experiment, the SC scores will be computed only for the ideal segmentation.

Results and discussion

Results are shown in Figure 4.8. The SC score peaks at the phonemic inventory in the case of English, and at both the phonemic and the slightly finer grained (2 allophones per phoneme) inventories in the case of Japanese. Similar to the previous experiment, we show in Figure 4.9 the histograms of the peak values across different parameter settings. In the case of English, the histogram selects unequivocally the phonemic inventory. In the case of

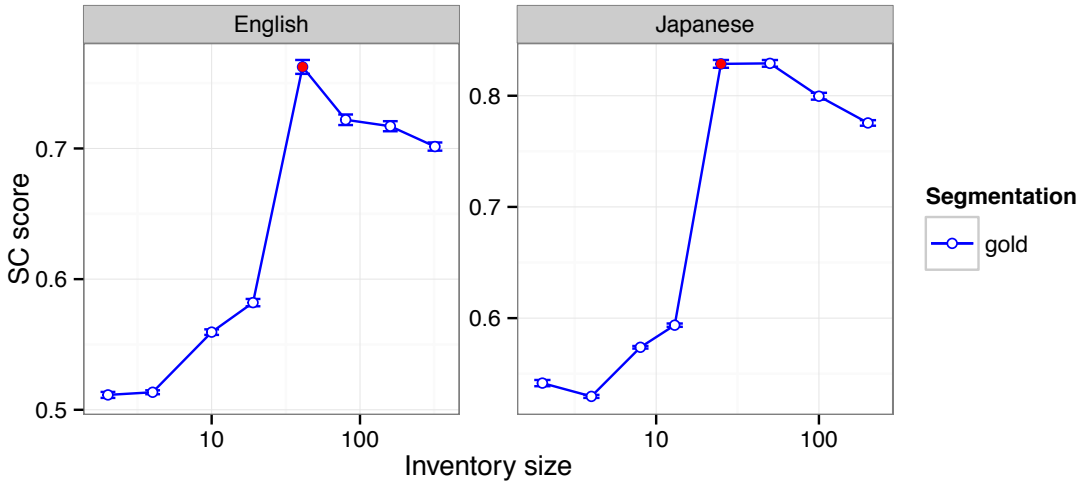

Figure 4.8: Semantic Consistency score of lexicons represented with different phonetic inventories and ideal segmentation, using nested-category partitioning. The points and error bars show the means and standard errors over different parameter settings. The red colors refers to the phonemic inventory of each language.

Japanese, there is a slight preference for phonemes over the next finer grained level (we will discuss this case later in this section).

As in the previous experiment, the semantic consistency tends to decrease as the inventory size exceeds that of phonemes, which is, similarly, due to an increasingly higher statistical sparsity of sub-types. In contrast, the small inventories do not decrease continuously from the peak, as was observed with the random partitioning. In fact, they show here an interesting asymmetry with the big inventories, being all almost at chance level (below 0.6). In order to understand why we get the asymmetry in the case of nested-category-based partitioning and not in the case of random partitioning, we examine the case of coarse representations in a deeper way than we did in the analysis of Experiment 1. Take the example of the words "cat" and "bag", which have orthogonal meanings, and suppose we are evaluating a coarse phonological analysis where both words have the same representation (e.g., /SVS/, where S refers to the category of stop consonants, and V to vowels). When splitting the set of tokens of the word type SVS in a random fashion, both sub-types

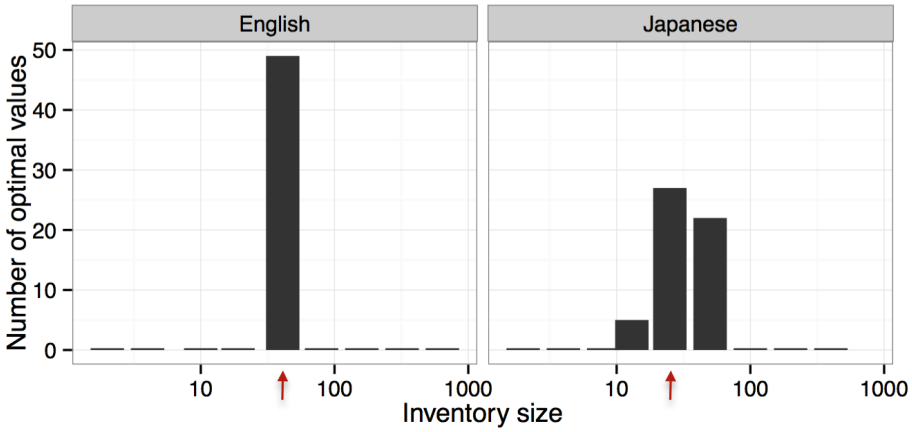

Figure 4.9: Histogram of the SC-score peaks across different parameter settings, using nested-category partitioning. The red arrows point towards the phonemic inventory.

(SVS_1 and SVS_2) will end up having a mix of "cat"s and "bags"s. The semantic similarity of these sub-types will be different from that of two random word types (e.g., SVS_1 and $SVSV_1$), but because they have a less differentiated representation than the original words, this difference will be a bit smaller, leading to a slightly lower SC score. The coarser the representation, the more semantically unrelated words will be contained in a given word type, and the less differentiated the resulting sub-types will get from a random pair, leading to the observed continuously decreasing SC score.

In contrast, if the splitting is performed based on nested representations, the sub-types will be orthogonally different words, the moment we get down from the phonemes. If we take our previous example, the sub-types of the word SVS will not consist of a mix of both "cat"s and "bat"s as in the random partitioning case, they will, in contrast, be exactly SVS_1="cat" and SVS_2= "bag" (or the other way around), which represents basically a random pair of words. This explains why the SC score is almost at chance level starting from the first inventory below the phonemic one, hence the asymmetry. As we can see from the results, this property is interesting since it allows for a clearer distinction between coarse and phonemic analyses. Moreover, as we can see in Figure 4.10, the asymmetry helps

in addressing the problem we faced with small portions of data in Experiment 1. In fact, when we derive scores for different portions of the corpora, we get, similarly, a successful learning with 5.000 utterances, and random fluctuation with 50 utterances. However, unlike Experiment 1, learning becomes possible even with 500 utterances.

Now we get back to the particular case of Japanese data where the SC score peaks at both the phonemic and the next finer grained level. Here, thanks again to the asymmetry, phonemes stand out as the optimal choice when we take into account both semantic consistency *and* economy in representational resources. In fact, the learner can be understood to be seeking the least number of categories that maximizes the consistency of the lexicon.

4.4 General discussion

In this chapter, I investigated a learning mechanism that learners might use to make the 'hard decision' of phoneme acquisition, i.e., deciding on the threshold that results in the adequate representational resolution. I framed this learning as an optimization problem: learners are understood as trying out many hypotheses and selecting the optimal one, i.e., the one with the most semantically consistent lexicon. The mechanism does not purport to account for *how* a representation is constructed (I assumed that the cues to phonemicity, which play here the role of the distant function, are perfect), but rather for how it performs, compared to other possible representations. Such framing is common in language acquisition research. For example, Chomsky (1965) suggested that the mind is equipped with a Language Acquisition Device (LAD), which has "a method for selecting one of the (presumably, infinitely many) hypotheses that are allowed, and that are compatible with the primary linguistic data". Although they generally disagree with Chomsky on the degree of innateness and learning, Bayesian psychologists frame the question is a similar fashion. In fact, we can read in Chater and Manning (2006) the following:

> From a Bayesian standpoint, each candidate grammar is associated with a

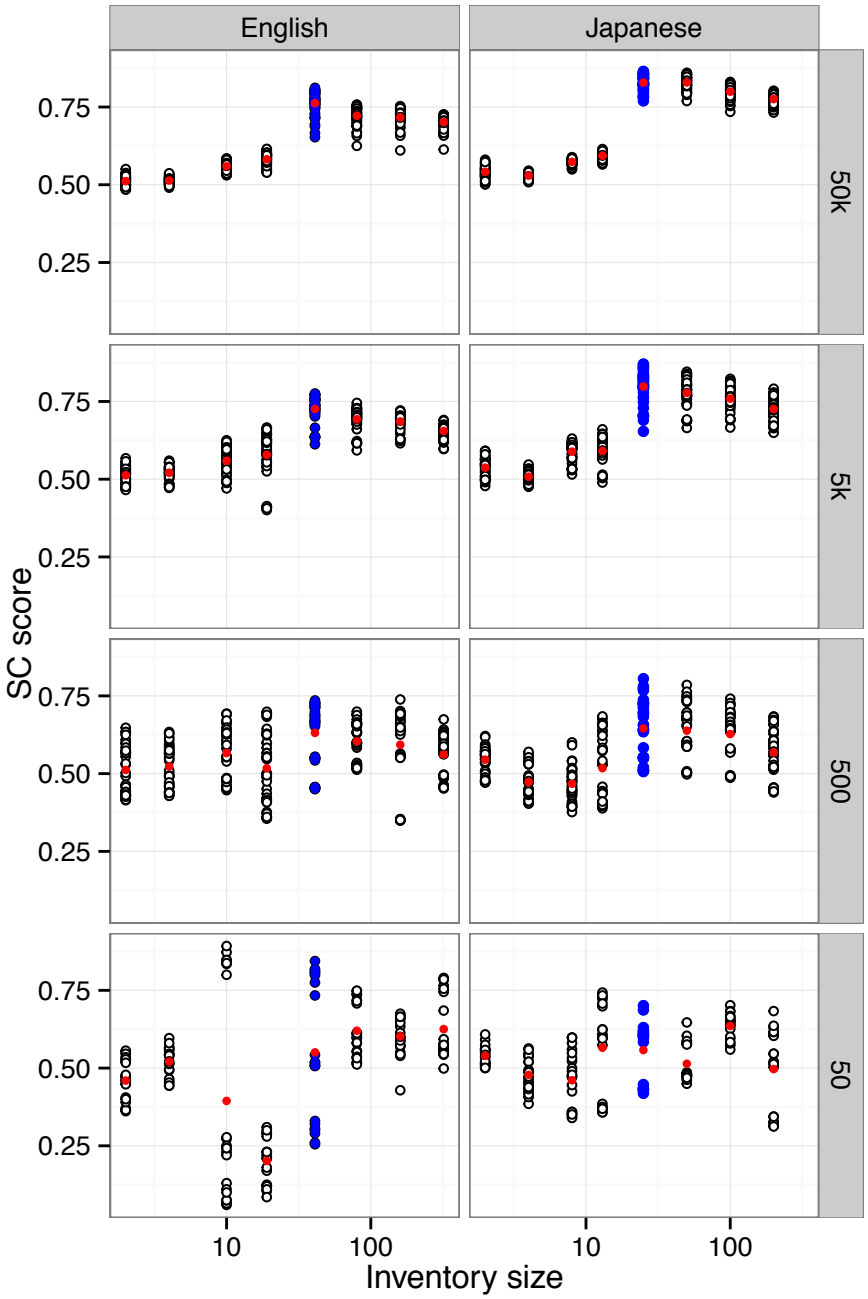

Figure 4.10: Semantic Consistency score of lexicons represented with different phonetic inventories, and for corpus sizes ranging from 100% (about 50.000 utterances) to 0.1% (50 utterances). The SC score is computed using nested-category partitioning, and the ideal word segmentation. The white points refer to the individual scores using different parameter settings, the red points to the phonemic inventory, and the blue to the means of the individual scores

> prior probability; and these probabilities will be modified by experience using Bayesian updating [...]. The learner will presumably choose a language with high, and perhaps the highest, posterior probability.

By analogy to the Bayesian framework, the Semantic Consistency score can be seen as the equivalent of the posterior probability, operating in a space of hypotheses, where a hypothesis is defined as a particular phonological analysis of the input, associated with its corresponding distribution over semantic contexts.

The philosophy behind the SC score is that infants are learning and optimizing an entire system, rather than learning different sub-levels in isolation (learning phonemes, then word-forms, then semantics). In particular, I assume that phoneme learning is driven by the need to make sense of the input, the selection pressure coming from the process of extracting meaning. The quality of a phonological analysis is measured by the semantic consistency of the lexicon that it induces. Crucially, though it makes use of higher linguistic levels, the SC score does not require them to be fully developed, tolerating to a relatively high extent, segmentation errors and semantic ambiguity.

I characterized the semantic consistency of a word by the distributional consistency of its tokens, and I implemented this idea through using the more general notion of "sub-type", defined as a subset of the word's tokens. I tested a learning mechanism wherein sub-types are obtained based on random partitioning in Experiment 1, and on nested phonetic representations in Experiment 2. Both methods allow for the correct selection of the phonemic analysis, when the learning data is relatively big (around 5.000 utterances). However, the interesting asymmetry found in Experiment 2 (between inventories smaller and bigger than the phonemes) allows for successful learning even with a small dataset (around 500 utterances). Deriving sub-types according to nested categories is, arguably, more compatible with developmental data, showing that infants can access different levels of acoustic/phonetic representations (McMurray & Aslin, 2005; White & Morgan, 2008). In simple terms, this means that babies are more likely to test phonological analyses that

fall under phonetically-informed boundaries (e.g., [kæʔ] and [kæt^{h}]), than to test randomly picked representations (e.g., two sub-types composed, each, of a mix of both [kæʔ] and [kæt^{h}]).

Last but not least, we found the utility of the SC score in picking out the best representation to be independent of the parameter setting to a large extent, and to operate with minimal, if any, external supervision. This contrasts with learning models based on the Expectation Maximization (EM) categorization algorithm, where the number of categories is specified in advance (de Boer & Kuhl, 2003), and with Bayesian models based on the Dirichlet Process (Feldman, Griffiths, et al., 2013), which require the specification of a hyper-parameter (concentration parameter) that monitors the number of categories. It is also worth mentioning that Vallabha et al. (2007) proposed a learning algorithm similar to EM, which provides an automatic way to infer the number of categories: the models starts with a high number of phonetic categories, and eliminates, over learning, the ones whose frequency drops below a predefined threshold. Note, however, that these comparisons are undertaken here only qualitatively. No precise quantitative comparison between the learning mechanism I propose and these models can be made, since they differ on the type of variation considered in learning.

Part III

Human experiments

While the computational part showed how the learning mechanism scales up to a realistic *input*, this part will provide support for the *cognitive plausibility* of the mechanism by testing human learners in a controlled setting. In Chapter 1, I mentioned that the mechanism is based on 4 basic assumptions. The first one assumes that infants pay attention to fine grained phonetic categories (Werker & Tees, 1984; White & Morgan, 2008; McMurray & Aslin, 2005 to cite but a few). The second assumption supposes that learners rely on their fine-grained perception to segment and store lexical items (see for instance, Houston & Jusczyk, 2000 and Curtin et al. 2001). The third assumption posits that learners are able to infer a sense of semantic similarity from co-occurrence statistics, and the fourth, that this semantic similarity can help with determining the phonemic status of phones.

If the first two assumptions have received significant empirical support, the last two have not. Chapter 5 and 6 try to fill this gap in the literature by providing experimental support for, respectively, the third and fourth assumptions.

Chapter 5

Learning semantic similarity through word co-occurrence

5.1 Introduction

How do children learn the meanings of words in their native language? This question has intrigued a lot of scholars studying human language acquisition. Quine (1960) famously noted the difficulty of this process. In fact, every naming situation is ambiguous. For example, if I utter the word *gavagai* and point to a rabbit, you may possibly infer that I mean the rabbit, the rabbit's ear, or its tail or color,...etc. A popular proposal in the language acquisition literature suggests that, even if one naming situation is ambiguous, being exposed to many situations allows the learner to narrow down, over time, the set of possible word-object mappings (e.g., Pinker, 1989). This proposed learning mechanism has come to be called Cross-Situational Learning (hereafter, XSL). Laboratory experiments have shown that humans are cognitively equipped to learn in this way. For example, L. Smith and Yu (2008) presented adults with trials that simulated real world uncertainty: each trial was composed of a set of words and a set of objects, in such a way that no single trial had enough information about the precise mappings. However, after being exposed to many of such trials, participants were eventually able to name the objects with a better-than-chance performance. Many experiments replicated this effect with adults, children and infants (Suanda, Mugwanya, & Namy, 2014; Vlach & Johnson, 2013; Yu & Smith, 2007). Subsequent research tried to characterize the algorithmic underpinnings of XSL. Some experiments suggested that learners accumulate in a parallel fashion all statistical regularities about word-object co-occurrences, and they use them to gradually reduce ambiguity across learning situations (McMurray, Horst, & Samuelson, 2012; Vouloumanos, 2008; Yurovsky, Yu, & Smith, 2013). Other experiments suggested that learners maintain, instead, a single hypothesis about the referent of a given word. New evidence either corroborate this hypothesis or contradict it (Medina, Snedeker, Trueswell, & Gleitman, 2011; Trueswell, Medina, Hafri, & Gleitman, 2013). Yurovsky and Frank (2015) proposed a synthesis of both accounts, whereby the learner choice's to adopt one of the two learning strategies depend on

the complexity of the learning situation.

This being said, XSL is unlikely to be the unique mechanism of word learning at work. First, real learning situations are much more ambiguous than typical simulated situations used in laboratory experiments. When subjects are tested in a more realistic learning context, the load on memory increases and, therefore, the ability to make use of the available visual information diminishes (Medina et al., 2011; Yurovsky & Frank, 2015).

Second, XSL assumes a perfect covariance between words and their referents. This assumption does not take into account the fact that words –in real situations– are sometimes uttered in the absence of their referents (e.g. when talking about past events, "remember that cat?"). In this experiment, I propose a statistical learning mechanism that purports to complements XSL, through relying on cues from the concomitant linguistic information, and more precisely on word co-occurrence.

5.2 Word co-occurrence and semantic similarity

Typical XSL settings assume that words occur in isolation. In real learning contexts, however, words are embedded in natural speech, and have various distributional properties. In particular, semantically similar words tend to co-occur more often than semantically unrelated words. For example, the word "ball" and "play" tend to co-occur more often than "ball" and "eat". This fact is documented in linguistics under the name of the 'distributional hypothesis' (hereafter, DH) (Harris, 1954), and has been popularized by Firth's famous quote "You shall know a word by the company it keeps" (Firth, 1957). The distributional hypothesis is also the basis for distributional semantics, the sub-field of computational linguistics that aims at characterizing words' similarity, based on their distributional properties in large text corpora. Tools from the field of distributional semantics such as Latent Semantic Analysis, (Landauer & Dumais, 1997), Topic Models (Blei, Ng, & Jordan, 2003), or more recently Neural Networks (Mikolov, Karafiát, Burget, Cernocký, & Khudan-

pur, 2010) have proved to be very effective in modeling human word similarity judgement (Griffiths et al., 2007; Baroni & Lenci, 2010; Fourtassi & Dupoux, 2013; Parviz et al., 2011). Moreover, we saw in the computational part of this dissertation that distributional semantic models can be useful in modeling phoneme learning.

5.3 Zero-shot learning

Learning through DH typically require a large corpus, especially if nothing is known about the language. Here, we explore the case where some words are already known and only one word is learned through DH. This corresponds to the so-called 'zero-shot learning' situation.

An interesting example of this situation has been given by Socher, Ganjoo, Manning, and Ng (2013). They built a model that can map a label to a picture even if this mapping was not seen in training! More precisely, using the CIFAR-10 dataset, the model was first trained to map 8 out of the 10 labels ("automobile, "airplane, "ship, "horse, "bird, "dog, "deer, "frog) in the dataset, to their visual instances. The remaining labels ("cat and "truck) were omitted and reserved for the zero-shot analysis. Second, they used a distributional semantic model (based on Neural Networks) to obtain vector representations for the entire set of labels (i.e., including "cat and "track) based on their co-occurrence statistics in a large text corpus (Wikipedia text). When tested on it ability to classify a new picture (a cat or a truck) under either the label of "truck" or "cat", the model performed with a high accuracy, using only the patterns of co-occurrence among labels, and the semantic similarity between the new and old pictures. For example, when presented with the picture of a cat, the model has to classify it as "cat" or "truck". The models makes the link between the picture of the cat and that of a similar picture (e.g. dog), and chooses the label that is more related to the label of this similar picture, i.e., "cat". In fact, "cat" co-occurs more often with "dog" than with, say, "airplane". Therefore the label "cat" is favored over the alternative label (i.e., "truck").

The conditions of zero-shot learning are often met in the context of word acquisition. For instance, this corresponds to the (rather ubiquitous) situation where an unknown word is heard in the absence of its visual referent. Therefore, I suggest that the learner can go about it in a way that mimics the mechanism of zero-shot learning. In the following, we test this hypothesis with adults, following closely the spirit of the model developed by Socher et al. (2013).

5.4 Method

The experiment consists of 4 parts:

1. Referential familiarization
2. Learning consolidation
3. Distributional familiarization
4. Semantic generalization

The referential familiarization and consolidation consists in explicitly teaching subjects the association between words in an artificial language and their referents. In the distributional familiarization, participants hear 'sentences' made of words from this artificial language without visual referents; some of these words have already been introduced in the referential familiarization, and some are new words. Crucially, the new words co-occur consistently with words of the same semantic category. Finally, the semantic generalization phase tests whether subjects can rely on distributional information *alone* to infer the semantic category of the new words, without any prior informative referential situation. Below is a detailed description of each part of the experimental procedure.

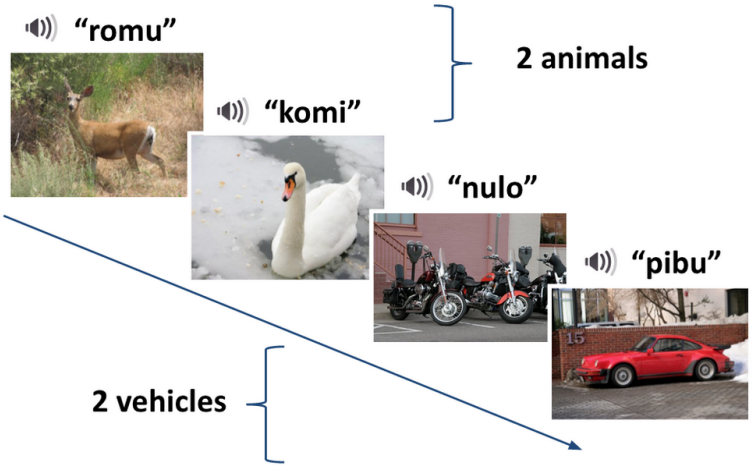

Figure 5.1: Referential familiarization. Participants are presented with multiple series of word-objects pairings. The objects belong to the category of animals or the category of vehicles.

Part 1: Referential familiarization

In this phase of the experiment(Figure 5.1), participants are taught the pairing of 4 words in an artificial language[1] with 4 objects. The objects belong to either the category of vehicles (car, motorcycle) or the category of animals (deer, swan). Participants see a picture of the referent on the screen and hear its label simultaneously. There are 3 trials, each consists of a randomized presentation of the series of 4 pairings.

Part 2: Learning consolidation

The purpose of this phase is to consolidate and strengthen the participants' knowledge about the 4 word-object pairings (Figure 5.2). Participants are tested using a Two Alternative Forced Choice paradigm (2AFC). They are presented with a series of trials where they hear a label (*pibu*, *nulo*, *romu* or *komi*) and are shown two objects; one of which is the correct referent, and the other belongs to other semantic category. Crucially, after they have made a choice, they get a feedback on their answers ("correct"/"wrong"). Participants are

[1]The audio stimuli were graciously provided by Naomi Feldman

presented with 16 questions of this sort, which correspond to the combinatorial possibilities of forming pairs of items from one semantic category with items from the other category (4 cases), in conjunction with the order of the visual presentation of the referents (4×2 cases) and the item being labeled ($4 \times 2 \times 2 = 16$ cases in total).

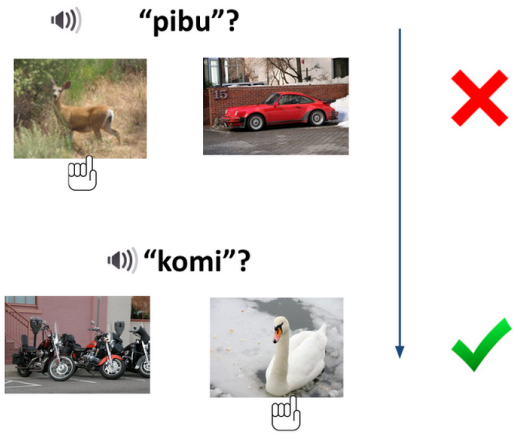

Figure 5.2: Learning consolidation. Two-Alternative Forced Choice paradigm (2AFC), with feedback.

Part 3: Distributional familiarization

Distributional familiarization follows the referential training and consolidation. Participants listen to 'sentences' made of words from this artificial language without any visual referent. As explained in Figure 5.3, each sentence consists of 3 words. Two of which are known words from one semantic category, i.e., either *romu* and *komi* (animals) or *pibu* and *nulo* (vehicles). The third word is a new artificial word that consistently co-occurs with them. The new words are *guta* and *lita*. The way *guta*/*lita* are distributed with either (*romu*, *komi*) or (*pibu*, *nulo*) was counterbalanced across participants so as to avoid different sorts of linguistic and perceptual biases that may arise from the way the stimulus is organized. There is a 750 ms pause between words, and 2500 ms pause between sentences. There are 16 sentences in total, 8 for each semantic context; (*romu*, *komi*) and (*pibu*, *nulo*).

Words within sentences are randomized and the semantic context is alternated during the exposure.

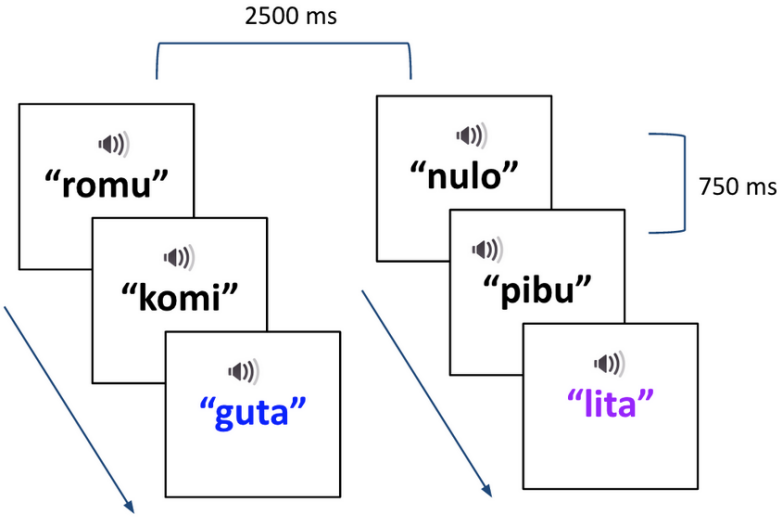

Figure 5.3: Distributional familiarization. Sequences of words are presented with no visual referents. Two new words ("guta" and "lita") are introduced and co-occur consistently with the words corresponding to one of the two semantic categories ("romu" and "komi" for the category of animals, and "nulo" and "pibu" for the category of vehicles)

Part 4: Testing semantic generalization

Participants are presented again with a two alternative forced choice. As explained previously in the learning consolidation phase, they hear a label and they are asked to choose between two objects, but here participants do not get feedback on their answers. We are particularly interested in how participants respond in the situation where they hear the new labels (*guta* or *lita*) and are presented with two new objects that represent a new animal (squirrel) and a new vehicle (trolley). Participants have never been shown the referential mapping of the new words, so their answer would reveal whether distributional learning alone had helped them infer semantic knowledge about the word (i.e., the semantic category of the referent). This test phase is composed of 4 questions about the new labels/objects, varying the visual order of the objects (1×2) and the object being named ($1 \times 2 \times 2 = 4$

cases in total), in addition to 4 selected filler questions about the old words/objects used in the referential training. I separated trials of the new objects from trials of the old objects so as to avoid any form of cross-situational learning during the test phase.

Procedure

As shown in Figure 5.4, participants are first trained on the pairing of 4 artificial words with their referents (part 1 and 2). Then they are exposed to 2 blocks of distributional familiarization that combine both old and new words in a distributionally coherent fashion (part 3), and they are tested 3 times (part 4): before any exposure to distributional information and after each block of exposure.

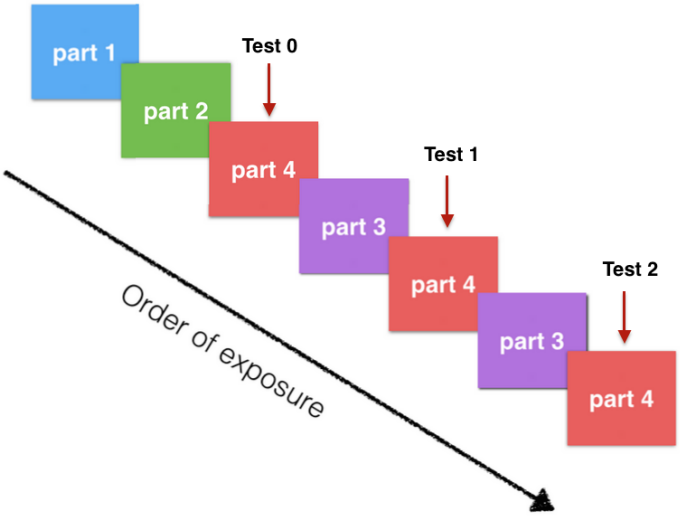

Figure 5.4: Order of exposure of the experimental settings. Participants are trained referentially once (part 1 and part 2), distributionally twice (part 3). They are tested in three sessions (part 4): before and after each block of distributional learning

Population and rejection criterion

50 Participants were recruited on-line through Amazon Mechanical Turk. I included in the analysis participants whose total score on the filler questions during the testing phases (i.e.,

part 4) were above chance level. This is a way to select only subjects who paid attention during the training parts. 2 participants was excluded based on this criterion.

5.5 Results and Analysis

Figure 5.5 shows the proportion of correct answers on both filler and target questions, as a function of the testing session. In the filler condition, answers were almost perfect in the three conditions (before exposure, after one block, and after two blocks of exposure to part 3). This shows that participants have reliably learned the association between words and their referents during the training phase, and that this learning was not affected by subsequent exposure to distributional information. In the target condition (new words), and before distributional training (i.e., session 0), subjects were at chance level ($M = 50.5\%$ of correct answers). In fact, a one sample t-test comparing the mean against chance (i.e, 50%) gives a $t(47) = 0.083$ with p-value $= 0.93$. The absence of learning is a predictable result since participants had no prior cue about the relevant object mapping. However, after one and two blocks of distributional training, subjects were significantly above chance level. A one sample t-test gives, respectively, for session 1 an average of correct answers $M = 72.4\%$, with $t(47) = 3.942$ ($p < 0.001$), and for session 2, an average of $M = 68.2\%$, with $t(47) = 2.852$ ($p = 0.006$). In order to compare the behaviour of the participants before and after distributional training, I performed a paired t-test. For session 0 vs. session 1, there is a significant change, the difference mean is equal to $M = 0.218$, with $t(47) = 2.99$ ($p < 0.01$). Similarly, for session 0 vs. session 2, the difference mean is $M = 0.177$, with $t(47) = 2.238$ ($p = 0.029$). However, between session 1 and session 2, the difference mean $M = 0.041$ is not significant, $t(47) = 0.662$, $p = 0.51$. This shows that most of the learning occurs during the first block of distributional exposure. Additional training does not significantly improve learning (if anything, it seems to slightly decrease the average of correct responses).

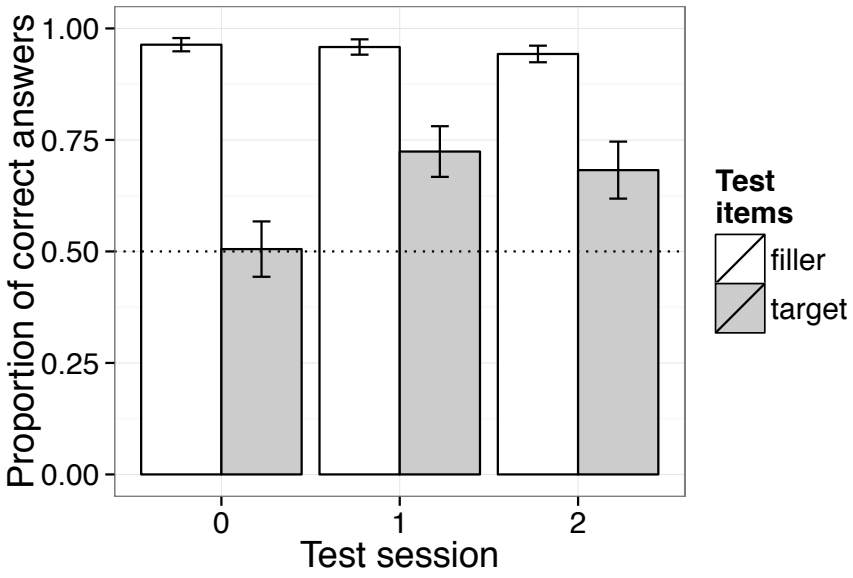

Figure 5.5: proportion of correct answers in filler condition (known words) and target condition (new words), before any distributional exposure (session 0) and after the first and second block of exposure (session 1 and 2)

5.6 Discussion

The results show that, when learning the meaning of words, people are sensitive, not only to the co-occurrence of words and objects (as suggested in XSL), but also to co-occurrence statistics between words themselves (i.e., DH). More importantly, I showed that these two sensitivities interact in a way that mimics a machine learning mechanism called zero-shot learning. In fact, participants in our experiment were able to guess the semantic category of a new word through the semantic properties of the words with which it co-occurred consistently. Participants knew beforehand that they would be introduced to an artificial language and that they would have to learn the meaning of words in this language, but they were not explicitly instructed about the fact that words that co-occur in the same sentences are supposed to have similar meanings. Participants have spontaneously turned to co-occurrence in order to cue semantic similarity, and infer the category of the ambiguous words.

Although we used an artificial language whose 'sentences' fall short, on many aspects, of real speech, this work provides evidence for the *cognitive plausibility* of this learning mechanism, much in the spirit of the statistical learning literature (e.g., L. Smith & Yu, 2008; J. R. Saffran et al., 1996). If it scales up to real languages, this word-word co-occurrence mechanism would prove crucial in complementing word-object co-occurrence mechanisms. In fact, most word-object co-occurrence learning strategies (e.g. XSL) assume that words covary perfectly with their referents. This assumption is not always correct. For example, when talking about a past event, the conversation may not match the immediate visual context. In contrast, words used in a given conversation, be it about present, past or future events, normally co-occur in a coherent fashion. The learner can rely on this intrinsic property of speech to bring about robustness to the learning process. For example, suppose the learner, while at home, hears a discussion about the last visit to the "zoo". XSL learning, if operating alone, would be confusing. In contrast, if XSL operates in concert with DH, the learner would tend, if in doubt, to link a new word (e.g., "zoo") not to some surrounding object, but to other co-occurring words, which are likely to be zoo-related words (such as "animals", "bird" and "monkey"). Further work is needed to characterize the precise conditions under which learners would rather switch to the word-word co-occurrence cue to infer meaning.

Moreover, the proposed mechanism can help the learners develop an early semantic representation for words with a rather abstract meaning. Abstract words (e.g., verbs) are learned later in development than words with salient concrete referents (e.g., "ball") (e.g., Bergelson & Swingley, 2013). They are considered as "harder" to learn because there is no obvious or/and lasting correspondence between the word and the physical environment (Gleitman, Cassidy, Papafragou, Nappa, & Trueswell, 2005), and because they probably require advanced social and intention-reading skills (Tomasello, 2001), or the development of syntactic structures (Gleitman et al., 2005). Nonetheless, this experiment suggests that,

before babies become efficient at acquiring abstract words, they have what it takes to begin characterizing them through concrete (and more easily learnable) words with which they co-occur. Suppose the learner hears the word "play" occur consistently with known concrete words such as "toy" and "ball", and the word "eat" co-occur with words such as "spoon" and "apple". Even though, at an early age, they might lack a fully developed representation for these abstract words, they can a priori infer that one of them is more related to the context of play, and the other, to the context of eating. It is true that, in these particular examples, verbs have a significant concrete (or 'observable') dimension, and can therefore be characterized by visual input as well. For instance, the word "play" can be associated with both the co-occurring concrete words (e.g., "toy") and the corresponding visual input (e.g., a toy). However, for verbs with a high degree of abstractness ("think", "believe", "remember",..) learners are generally more sensitive to linguistic cues (Papafragou, Cassidy, & Gleitman, 2005). More work is needed to investigate the extent to which infants relay on co-occurrence information to develop an early semantic representation for abstract verbs.

Finally, during the write-up of this paper, it came to our knowledge that Ouyang, Boroditsky, and Frank (in press) conducted an experiment that shared many similarities with ours. However, it also presented interesting differences both in terms of the experimental setup and the results. Ouyand at al. exposed adult participants to auditory sentences from a MNPQ language. It is an artificial language where sentences take the form of "M and N" or "P and Q". Ms and Ps are used as context words, whereas Ns and Qs are target words. We believe there are two crucial differences between the two experiments. First, the context words (M and P) were composed of a mix of various proportions of real English words or non-words. In our experiment, they were all non-words. Second and more important, Ouyang et al. (in press) followed the spirit of MNPQ's paradigm in keeping constant the order of the words in the sentences, that is, M and P always occurring first in the sentence, and N and Q always occurring last. Our experiment was more faithful to the hypothesis of

bag-of-words, which is crucial in distributional semantic models: order within a particular semantic context (e.g., a sentence) is irrelevant. It was therefore randomized across trials. Interestingly, although none of the context words we used were known words, we obtained a high learning rate. In contrast, Ouyang et al. (in press) obtained successful learning only when most of the context words were familiar English words. A plausible explanation for this difference is that, in the case of MNPQ language, participants have two possible learning dimensions: learning the positional patterns (what word comes first, and what words comes last) and learning the co-occurrence patterns (what couple of words co-occurred with each other). In fact, it has been shown that when both positional and co-occurrence cues are available, participant tend focus on the first ones (K. Smith, 1966). By using familiar words, Ouyang et al. (in press) showed that participants were more likely to learn co-occurrence patterns, probably through alleviating part of the memory constraint. In our case, the positional patterns was random, which left participants with only one learning dimension (i.e., co-occurrence pattern).

To conclude, this chapter provided a cognitive proof of principle to the third assumption of the dissertation, according to which an early sense of semantic similarity can be learned through sensitivity to word co-occurrence in speech. In the next chapter, I will deal with the other major assumption, which posits a close relationship between semantic similarity judgement and phonemic categorization.

Chapter 6

Semantic similarity modulates the phonemic status of phones

6.1 Introduction

Word learning can be characterized as a mapping between two categories: The phonological category and the meaning category. Learning the phonological category of a word consists in acquiring its phonemic representation, i.e., being able to recognize different realizations of a word-form as the same, ignoring irrelevant variation (such as difference in talker, speech rate, emotion, and linguistic context) (Kuhl, 2004). For example, in English one has to interpret the forms [kæt^{h}] and [kæʔ] as variation of the same word /kæt/, and the forms [kæʔ] and [bæʔ] as instances of different words, i.e., /kæt/ and /bæt/. Learning the meaning category requires, upon hearing a few examples of a word paired with an object, to determine an accurate semantic category for the word. For example, one has to learn that /kæt/ can refer to cats of different colors and sizes, but not to dogs (Bloom, 2000).

The classic view of early word learning assumes that the processes of learning phonology and of learning meanings can be studied independently. This division follows from the assumption that infants master the phonological properties of the word-form much earlier in development than when they start mapping these forms to meanings (i.e., semantics). According to this view, word-form can influence the later development of meaning, but not the other way around. Indeed, there is a wealth of studies documenting this one-way influence (see Vouloumanos and Waxman (2014) for a review). For instance, Fulkerson and Waxman (2007) showed that infants as young as 6 months of age were able to form a 'meaning' category related to dinosaurs when they were familiarized with pictures of different dinosaurs paired with the same word ("Look at the toma!"). If, instead, the pictures were paired with a sequence of tones (non-speech like form), babies did not succeed in the categorization task. word-forms act therefore as an "invitation to form [semantic] categories" (Waxman & Markow, 1995). However, as I mentioned earlier, recent findings have started to challenge the received timeline of word learning. On the one hand, infants start mapping form to meaning very early in life (Tincoff & Jusczyk, 1999; Bergelson &

Swingley, 2012), even before their perception becomes attuned to the sounds of their native language (J. Werker & Tees, 1984). On the other hand, the phonological representation of word-forms continues developing beyond the critical age of perceptual attunement (Stager & Werker, 1997). Thus, the developmental trajectory of phonology and meaning overlaps. Infants do not wait to have completed one to start the other, rather, they learn form and meaning in a parallel fashion. This suggests that phonology and semantics influence each other throughout the learning process (Figure 0.1).

6.2 Semantic similarity and the phonemic status

Yeung and Werker (2009) provided experimental evidence for a top down effect where semantics does influence phonology in infants by as early as 9 months of age (see Chapter 1 for a detailed description of the experiment). The experiment suggests that babies are sensitive to semantic cues, and are willing to adapt their phonological analysis of word-forms accordingly. Here we investigate whether the top-down effect shown in Yeung and Werker (2009) is modulated by the similarity of the referents, as was in fact suggested in the computational part of this dissertation. In Chapter 3, I assumed that the phonemic status of phones varies as a function of the semantic relatedness of the corresponding minimal words, and in Chapter 4, I posited that candidate phonological analyses are assessed through the semantic coherence they induce at the lexical level. In both situations, indeed, I assumed an underlying close relationship between the semantic and the phonemic spaces: semantic similarity is understood as modulating the phonemic status of phones.

In the light of this, I test the hypothesis according to which, the willingness to adopt a phonological analysis of a given word-form is closely related to the willingness to extend the meaning of its referent. Imagine, for instance, that the learner is in a situation where she has to decide if a pair of similar word-forms [X] vs. [Y] (e.g., [kæt^{h}], [kæʔ]) correspond to one or two lexical items (i.e., if the variation is allophonic or phonemic), and suppose that

[X] and [Y] are respectively paired with objects A and B. I suggest that the phonological analysis of [X]/[Y] is modulated by the propensity of the learner to treat A and B as possible members of the same meaning category. The more A and B are semantically related, the more they can form a valid meaning category, and the more the phonological interpretation of [X]/[Y] tends towards allophony. Conversely, the more A and B are semantically distant, the more it becomes difficult to put them in one single meaning category, and the more the phonological analysis of [X]/[Y] tends towards phonemicity (Figure 6.1).

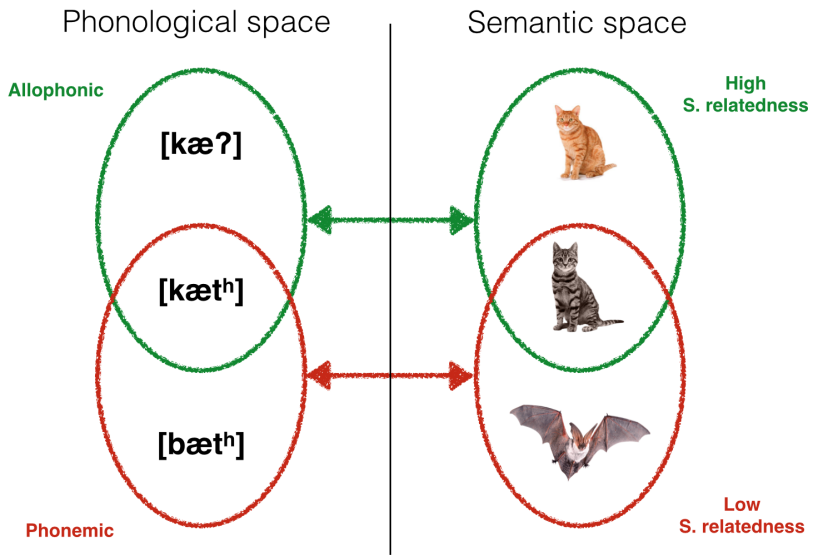

Figure 6.1: Word learning as a mapping between a phonological category and a meaning category. Green circles refer to correct generalizations: the referents have a high semantic relatedness, therefore, variation is analyzed as allophonic. The red circles refer to wrong generalizations: the referents have a low semantic relatedness, therefore, variation is analyzed as phonemic

In order to investigate this hypothesis, I test adults in an experimental paradigm similar to that of Yeung and Werker (2009). However, here the semantic relatedness of the referents will be varied in a continuous way.

6.3 Method

The experiment consists of a training phase and a testing phase. First, Participants are trained to learn the pairing between a minimal pair in an artificial language and objects with various degrees of semantic similarity. Second, they are tested over these words in a same-different task. The audio stimuli used in this experiment are identical to the ones used in Feldman, Myers, et al. (2013), and consists of (*gutah*, *gutaw*) a minimal pair that constitutes the target of the training phase, (*litah*, *litaw*) a minimal pair that varies along the same vowel contrast. It is used in the test phase to probe generalization. Finally *pibu* and *komi* are used as filler words to monitor the subjects' attendance to the task. Note that the minimal pairs vary along a vowel contrast that is neither too acoustically similar, nor too different. In fact, depending on the dialect, these two vowels can be treated by English native speakers as belonging to one or two categories (Labov, 1991). This is supposed to put the participants in a rather flexible situation where they can switch between phonological interpretations on the basis of the properties of the input.

In the training phase, the target words (*gutah*/*gutaw*) were paired with two objects whose semantic similarity was varied across 5 groups (Figure 6.2). In all groups, one member of the minimal pair (e.g., *gutah*) was paired with a picture of a 'cow'. The second member (i.e., *gutah*) was paired with a referent whose semantic similarity with the first referent varies on a five-step scale, from very similar, in the first group (another token of the same category, i.e., a different cow), to very different (a car), in the fifth group. The filler words were paired in all groups with the pictures of a house and a book (Figure 6.3).

During the training phase, participants hear the word and see the corresponding object simultaneously. They were exposed to 3 series composed each of a randomized presentation of the 4 word-object pairings (targets and fillers). In the test phase, participants hear a series of trials composed of two word tokens, and are asked to judge if these tokens correspond to different words in this artificial language (phonemic interpretation), or if they represent

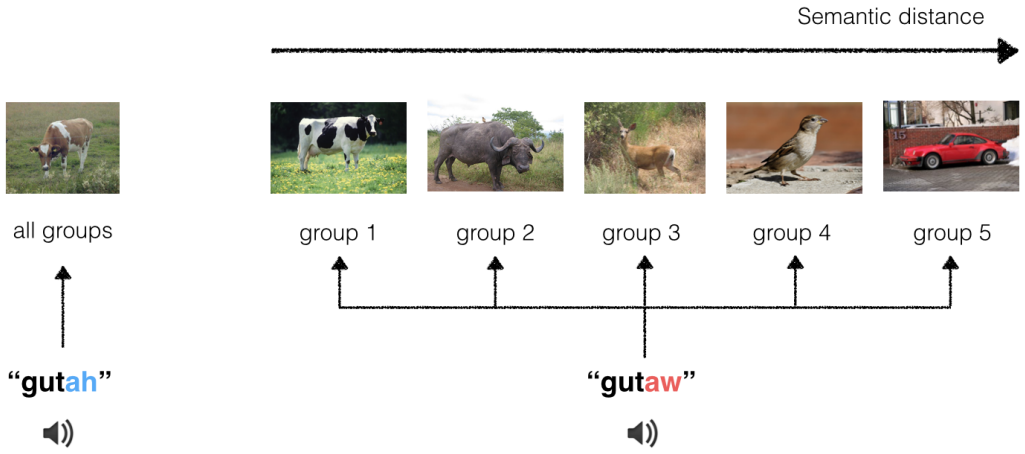

Figure 6.2: target word-object association between the minimal pair (“gutah”/“gutaw”) and a pair of referents with different levels of semantic similarity. The first referent represents a cow, the second referent represents, respectively, another cow, a buffalo, a deer, a bird, and a car.

a mere phonetic variation of the same word (allophonic interpretation). Participants were tested on 3 kinds of word pairs: the fillers (*komi* and *pibu*), the minimal pair used in the training phase (*gutah* and *gutaw*), and a new minimal pair that has not been heard by the participants before, but varies along the same vowel contrast (*litah* and *litaw*). Half of the trials consisted of exactly same words (e.g., *komi-komi*, *gutah-gutah*, *litaw-litaw*) and the other half, of different words (e.g., *komi-pibu*, *gutah-gutaw*, *litah-litaw*). For both same and different trials, participants have to answer by ‘same’ or ‘different’, according to their own phonological analysis. The order of the presentation was randomized between same and different. Participants were tested twice, once before the training phase and once after the training.

Participants and the rejection criterion:

150 Participants were recruited online through Amazon Mechanical Turk (30/ condition). We included in the analysis participants whose score on the filler questions were above chance level. This is a way to select only subjects who paid attention during the training

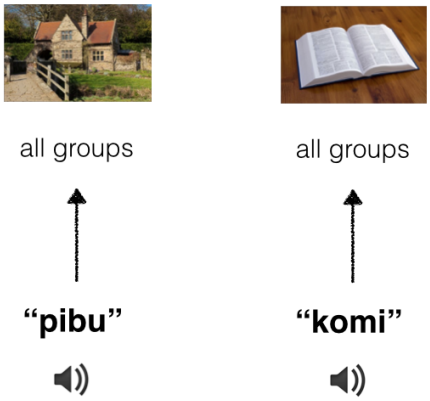

Figure 6.3: filler word-object pairings. “pibu” and “komi” were paired with a picture representing a house and a book. The pairing was kept the same across all groups.

phase. Based on this criterion, 3 subjects were excluded (2 from group 3, and 1 from group 5).

6.4 Results and Analysis

Figure 6.4 shows the proportion of time participants answered ‘same’ on same-trials (e.g., *gutah-gutah*, *litaw-litaw*), both before and after training. As expected, participants were almost at ceiling, i.e., they answered ‘same’ almost systematically when they heard a pair of identical word tokens. I analyse the more interesting case of different-trials (e.g., *gutah-gutaw*, *litaw-litah*), which probes the participants’ subjective judgement of whether a pair of slightly different word tokens belong to one or two lexical items (Figure 6.5). A 2 $\times$ 5 testing session (before vs. after) $\times$ condition (semantic similarity ranging from 1 to 5) mixed design ANOVA was conducted. I obtained a main effect of condition in both the case of learning (*gutah* vs. *gutaw*) ($F(4, 139) = 3.092$, $p = 0.017$), and the case of generalization (*litah* vs. *litaw*) ($F(4, 139) = 5.104$, $p = 0.029$). I also obtained a main effect of testing session in the case of learning ($F(4, 139) = 5.10$, $p = 0.025$), but not in the case of generalization. Moreover, a significant session by condition interaction was obtained in

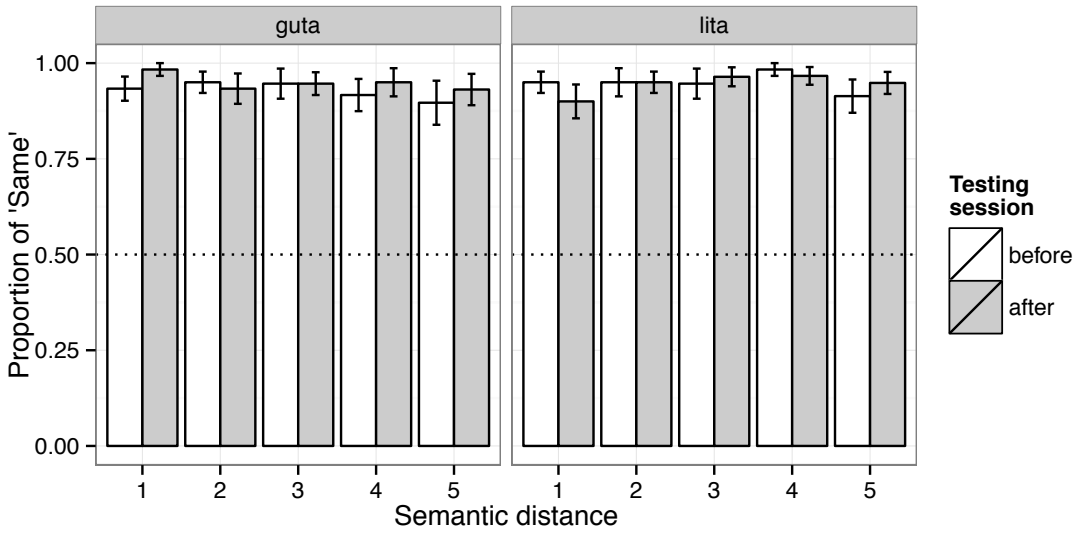

Figure 6.4: Proportion of 'same' answers on same-trials of both the minimal pair used in the training (gutah- gutah, gutaw-gutaw) and the new minimal pair (litah-litah, litaw- litaw), as a function of the similarity of the two referents in training. The semantic similarity ranges from 1 (the most similar) to 5 (the least similar). The dotted line represents chance.

learning ($F(4, 139) = 4.52$, $p = 0.001$) but not in generalization. Paired t-tests of the simple effects of testing session at each level of semantic similarity reveals, in the case of learning, a significant decrease in 'different' answers after training in condition 1 ($t(29) = 2.53$, $p = 0.016$), and a significant increase in condition 4 and 5, with, respectively, $t(29) = 3.12$ ($p = 0.004$) and $t(28) = 2.11$ ($p < 0.05$). In the case of generalization however, none of the semantic conditions present a significant difference (although the mean differences follow, overall, a similar trend to that of learning) (see Figure 6.5).

Now I examine the effect of condition after training (obviously, there is no effect of semantic condition before training) (Figure 6.6). First, we compared responses to chance level. Responses below chance translate the fact that participants picked an 'allophonic' analysis of the phonetic variation. Conversely, responses above chance mean that participants chose a 'phonemic' interpretation of the phonetic variation. In the case of learning, responses were significantly below chance level in group 1 with $t(29) = 3.12$ ($p = 0.004$),

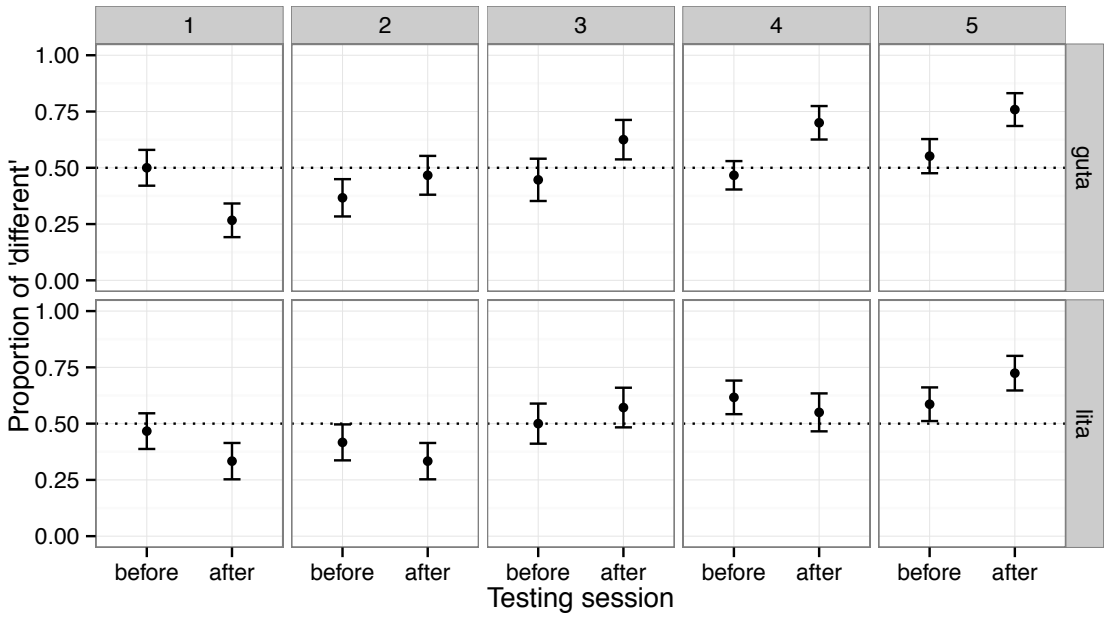

Figure 6.5: Proportion of 'different' on different-trials including both the minimal pair used in the training ("gutah" vs. "gutaw"), and the new minimal pair ("litah" vs. "litaw"), before and after the pairing with two referents with various degrees of semantic similarity, ranging from 1 (the most similar) to 5 (the least similar). The dotted line represents chance level

and significantly above chance in group 4 and 5, with respectively, $t(29) = 2.69$ ($p = 0.01$) and $t(29) = 3.55$ ($p = 0.001$). There is a significant difference between group 1 and groups 3, 4 and 5, with, respectively, $t(54) = 3.12$ ($p = 0.003$), $t(58) = 4.1$ ($p < 0.001$) and $t(57) = 4.71$ ($p < 0.001$), and a difference between group 2 and groups 4 and 5, with respectively $t(57) = 2.05$ ($p < 0.05$) and $t(56) = 2.58$ ($p = 0.01$). In the case of generalization, participants were below chance in group 1 and 2 with respectively $t(29) = 2.06$ ($p < 0.05$) and $t(29) = 2.06$ ($p < 0.05$). They were above chance in group 5 with $t(28) = 2.91$ ($p < 0.007$). There is a borderline significant difference between group 1 and groups 3 and 4 with respectively, $t(55) = 2$ ($p = 0.051$) and $t(58) = 1.85$ ($p = 0.068$), and a significant difference with group 5 with $t(57) = 3.5$ ($p < 0.001$). A similar difference was found between group 2, and groups 3 and 4 on the one hand, and group 5 on the other hand.

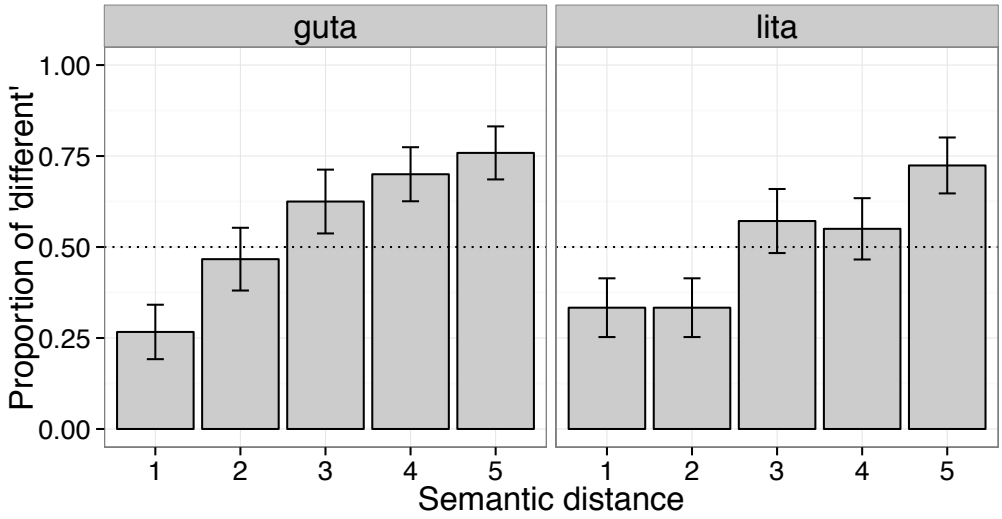

Figure 6.6: Proportion of 'different' answers on different-trials including both the minimal pair used in the training ("gutah" vs. "gutaw") and the new minimal pair ("litah" vs. "litaw"), as a function of the similarity of the two referents in training. The semantic similarity ranges from 1 (the most similar) to 5 (the least similar). The dotted line represents chance level

Second I tested the effect of semantic similarity gradient on the phonemic interpretation. To this end, I fitted a linear regression on groups 2 to 5 (excluding group 1 where the referents have the same label)[1]. We get significant scores both in the case of learning $F(1, 115) = 7.2$ ($p = 0.008$ and $R^2 = 0.06$), and the case of generalization $F(1, 115) = 9.916$ ($p = 0.002$ and $R^2 = 0.08$).

6.5 Discussion

The results of this experiment provide empirical evidence for the fourth assumption of the dissertation, according to which, the words' semantic similarity (and not just words' identity, as in Yeung and Werker (2009)) shape the phonetic space according to the relevant (i.e., phonemic) dimensions. In fact, participants were more likely to choose an allophonic

[1]I excluded the case of tokens belonging to the same category to show that the effect is not majorly driven by the label's identity (i.e., 'cow'). Note that a regression including group 1 also yielded statistically significant scores.

analysis of a minimal word-form pair when the corresponding semantic referents were more semantically related, and a phonemic analysis when the referents were less semantically related.

Comparing the participants' responses before and after training revealed a significant interaction with the semantic condition in the case of learning, which translated into significant decrease in discrimination when the referents were more semantically related (group 1), and significant increase when the referents were more semantically distant (group 4 and 5). In order to explain these results, two mechanisms can be put forward. The first one is similar to the mechanism of *acquired distinctiveness*, according to which, pairing two target stimuli with distinct events enhances their differentiation. The second one is akin to *acquired equivalence*, which expresses the converse of the first mechanism, i.e., pairing two target stimuli with similar events impairs the participants' subsequent differentiation (Lawrence, 1949; Hall, 1991). The experiment of Yeung and Werker (2009) provided evidence for the first mechanism in learning phonemes. The present work suggests that, actually, both mechanisms can take place. In fact, the results of the linear regressions go even further than this dichotomy between acquired distinctiveness/equivalence. They suggest that a more complex interaction is at work, involving a richer scale of semantic similarity (not jute the binary distinction of 'same'/ 'different'), which affects in a seemingly continuous way the corresponding sale of phonological analysis. Note, however, that testing does not consist of a purely perceptual task, as in Yeung and Werker (2009) with infant learners. Here, I followed Feldman, Myers, et al. (2013) in asking the adult participants explicitly to determine if the phonetic variation corresponds to relevant or irrelevant distinction in this particular artificial language. The reason was to avoid any ambiguity as to how participants were supposed to answer. In the usual same-different task, some participants may understand 'same' as 'exactly the same', and others as 'functionally the same', and the same thing can be said for the case of 'different'. I chose to specify to the participants

that their answers should be based on functional similarity, rather than perceptual one. The former is, arguably, more relevant to the task of phoneme learning. In fact, many experiments have shown that people are able to access fine-grained variation even within phonemic categories (McMurray, Tanenhaus, & Aslin, 2002; McMurray & Aslin, 2005) (See Appendix F for the instructions given to participants in the testing phase).

The effect of semantic similarity on the phonological interpretation was tested for both the pair used in training, and the new minimal pair that varies along the same vowel contrast, used to probe the participants' ability to generalize. The effect was, nonetheless, more robust in learning than in generalization. Although the mean difference follows a trend similar to that of learning, the interaction did not reach significance. Moreover, despite the fact that the linear regression on responses after training was significant for both learning and generalization, we note, however, that semantic relatedness affects the phonological analysis in a finer way in learning. In fact, there is an increase in mean response with every increase in semantic similarity, whereas generalization lags behind (there is an increase in mean response every other step on the semantic scale). This difference in the rate of learning and generalization cannot be explained by theories proposing a purely abstract representational system (e.g., featural theories of representation such as Chomsky and Halle (1968)), otherwise learners would have been able to perform similarly regardless of the specific word context in which the phonetic variation occurred, i.e., the first syllable of the word (*gu* or *li*). Although some degree of abstraction is necessary to explain the presence of the effect in the case of generalization, the difference observed between learning and generalization shows that even abstract representations, e.g., phonemes, can incorporate contextual detail such as the word context in which the phoneme occurred (see E. D. Thiessen and Yee (2010) for a similar conclusion).

On a last note, realize that in order to study the effect of semantic relatedness on the phonological analysis, I assumed, for the seek of simplicity, a perfect semantic representation

(every word-form was unambiguously mapped to a unique referent). It remains to be shown that the effect holds even when the semantic representation is ambiguous. Escudero, Mulak, and Vlach (2010) showed that human learners can encode fine phonetic details while tracking word-referent co-occurrence statistics. Following this work, future research will test whether the phonological analysis of a minimal word-form pair is modulated by the semantic relatedness of their statistical distributions, using either visual stimuli of potential referents (as in typical cross-situational learning) or, more importantly, only the co-occurring words (as in Chapter 5).

Chapter 7

Conclusion and Future work

The present dissertation explored the interaction between phoneme learning and word meaning learning. These learning problems have typically been considered independently, but new findings have pointed out that this assumption is problematic in many regards (Bergelson & Swingley, 2012; Stager & Werker, 1997; Varadarajan et al., 2008). I proposed a new mechanism, in which the process of learning phonemes benefits from the top-down constraint of early semantics, albeit ambiguous and rudimentary.

In the **computational part**, I selected two cues that are believed to shape the learners' early sensitivity to phonemic contrast: a bottom-up cue (acoustic similarity) and a top-down cue (word-form similarity). I compared their performance to that of two implementations of the semantic-similarity-based mechanism. The results showed that these cues can play complementary roles. The word-form cue fares relatively well under extreme conditions, whereas the first semantic cue requires a reasonable degree of intelligibility in the input. Information from both cues combine in the second semantic cue, which was shown to be resistant to extreme conditions while achieving a high degree of accuracy. Phonetic contrasts that do not cause lexical variation (such as the English [h] vs. [ŋ]) remain out of the top-down cues' reach, and are, therefore, better learned through bottom-up means

(acoustic similarity). These bottom-up and top-down cues provide continuous measure of phonemicity, and are a priori insufficient to acquire the phonemic inventory. In fact, the learners should be able, not only to tell how phonemic a pair of phone is, but also to decide whether this amount of phonemicity is contrastive in their native language. This question was addressed through another implementation of the learning mechanism, in which different candidate phonological analyses (which varied from fine- to coarse-grained) were assessed through the quality of the lexicon they induced, and more precisely, through the degree of semantic coherence of this induced lexicon. The mechanism enabled us to pick out the correct (i.e., phonemic) analysis of the input, while operating with minimal, if any, external supervision.

In sum, the computational part dealt with the problem of phoneme learning by isolating two different, but complementary, aspects. In Chapter 3, I modeled the perceptual reorganization in terms of continuous cues to phonemicity, abstracting away from the question of drawing the functional (i.e., contrastive) boundary. In Chapter 4, I showed that the right phonological interpretation (defining the functional boundaries and, for instance, the number of categories) can be derived when perceptual organization is assumed to be perfect. In a future work, I will bring together both aspects of phoneme learning in one computational model, which can proceed as follows. Starting from a large number of allophones that model the diversity of sounds present in a given language, a phone-phone similarity matrix will be computed, integrating cues from acoustics, word-forms and semantics (based on evidence from chapter 3). Categories of these phones will be made hierarchically based on this global similarity metrics, and the right level will be selected according to the intrinsic semantic consistency metrics (SC score defined in chapter 4). The model will aim at providing a self-sufficient account of phoneme learning starting from a contextual variation. I will, in addition, explore ways the model could be extended to include unsupervised discovery of contextual allophones from raw speech.

The computational part showed how the proposed mechanism can scale up to a realistic input, but it abstracted away from the cognitive limitations of the learner (memory constraints, limited attentional resources, cognitive biases,..). The **behavioral part** aimed, therefore, at providing support for the psychological plausibility of the main assumptions made in the learning mechanism. These assumptions (which were used throughout the computational part) can be summarized in the following. First, I assumed that infants pay attention to fine grained phonetic categories (J. Werker & Tees, 1984; White & Morgan, 2008; McMurray & Aslin, 2005). Second, I supposed that learners rely on their fine-grained perception to segment and store lexical items (Houston & Jusczyk, 2000; Curtin et al., 2001). Third, I assumed that learners are able to infer a sense of semantic similarity from co-occurrence statistics, and fourth, that this semantic similarity helps with determining the phonemic status of phones. If the first two assumptions have received significant empirical support, the last two have not. In Chapter 5, I provided evidence for the fact that human subjects are capable of developing a sense of semantic similarity for word-forms, without any informative word-object mapping situation. Participants were able to infer the semantic category of a new word through the semantic properties of the words with which it co-occurred consistently. In Chapter 6, I showed that human subjects demonstrated a graded sensitive to the semantic similarity of the referents, and crucially, that this graded sensitivity affected continuously the corresponding phonological analyses. Participants were more likely to choose an allophonic analysis of a minimal word-form pair when the corresponding semantic referents were more semantically related, and a phonemic analysis when the referents were less semantically related. However, for the seek of simplicity, I assumed a perfect semantic representation (every word-form was unambiguously mapped to a unique referent). In a future work, I will explore ways to bring together the experimental paradigm of Chapter 5 and that of Chapter 6 in one task, which would probe whether semantic similarity modulates the phonological analysis even when the semantic representation is

ambiguous, and particularly, when it is derived in a rather distributional fashion, as in Chapter 5. Moreover, I will explore ways this work could to extended to infant studies. For instance, Appendix G proposes a detailed description of how the experiment in Chapter 6 could be adapted to infant testing paradigms.

Appendix A

Pre-processing of the CSJ corpus

Some of the segments included in the CSJ phonetic annotation do not represent real contrastive elements, others are noisy events or abstract placeholders that are inappropriate for the purpose of acoustic model training. Therefore, I performed further pre-processing in order to obtain a true phonemic inventory that we can use in our modeling work, following, in this regard, mostly the steps of Boruta (2012).

- CSJ denotes phonetic palatalization with distinct labels (e.g., [kj]). As this is just an allophonic variation, I mapped the palatalized segments to their corresponding phonemes (e.g., /k/).

- Besides phonetic palatalization, there is what is called in the CSJ documentation phonological palatalization, which is phonemic as exemplified by the minimal pair: /ko/ ("child") and /kyo/ ("hugeness"). In this case, the cluster (e.g., [ky]) was mapped onto two phonemes consisting of the plain version of the consonant (e.g., /k/) and the yod (relabeled /j/). Similarly, voiceless alveolar sibilant affricate (i.e., /ts/) was mapped onto the two corresponding phonemes.

- Vowel length is phonemic in Japanese. It has five vowel qualities ([i], [a], [u], [o], and [e]) and CSJ uses an abstract moraic chroneme [H] to denote the second half of the

long vowels. The long vowels were set apart from the short ones by mapping the sequences [iH], [aH], [uH], [oH], and [eH] onto the atomic labels /iː/, /aː/, /uː/, /oː/ and /eː/.

- Consonant length (gemination) is also phonemic in Japanese. The abstract moraic obstruent [Q] was used in the CSJ corpus to denote the second half of a geminate. Here a geminate consonant (e.g., [ni**Q**poN]; 'Japan') was mapped to to a sequence of two instances of the obstruent at hand (/ni**p**poN/).

Other minor codings:

- The voiceless bilabial fricative [F] and the glottal fricative [h] are allophones in Japanese. They were mapped onto the phoneme /h/.
- Tags at the phonemic level that denote hesitations ([VN]) or unidentifiable consonants and vowels ([?] and [FV]) were collapsed into one single "Noise" label.
- Segments that are heavily underrepresented, and which therefore represent no statistical or distributional interest, were also collapsed into the same "Noise" label. These segments are [kw], [Fy], and [v] with, respectively 1,1, and 2 occurrences.

Appendix B

HTK phonetic decision tree

Questions on contexts used in HTK's decision-tree state-tying procedure. All questions were input to the system in the order specified in this table, and each question was duplicated to account for phonemes preceding and following contexts. Table B.1 shows the questions for Japanese data, and Table B.2 shows the questions for English data.

	Question	**Set of context phonemes**
1	Consonant?	{p, b, m, t, d, n, s, z, r, k, g, y, w, N, h}
2	Vowel?	{a, e, i, o, u, aː, eː,, iː, oː, uː}
3	Voiced consonant ?	{b, m, d, n, z, r, g, y, w, N}
4	Stop consonant?	{p, b, t, d, k, g}
5	Alveolar consonant?	{n, t, d, s, z, r}
6	Short vowel?	{a, e, i, o, u}
7	Long vowel?	{aː, eː,, iː, oː, uː}
8	Voiceless consonant?	{p, t, k, s, h}
9	Close vowel?	{i, u, iː, uː}
10	Mid vowel?	{e, o, eː, oː}
11	Front vowel?	{i, iː, e, eː}
12	Back vowel?	{o, oː, u, uː}
13	Nasal consonant?	{n ,m, N}
14	Bilabial consonant?	{m, p, b}
15	Velar consonant?	{k, g, w}
16	Fricative consonant?	{s, z, h}
17	Approximant consonant?	{y, w}
18	Vowel quality [a]?	{a, aː}
19	Vowel quality [e]?	{e, eː}
20	Vowel quality [i]?	{i, iː}
21	Vowel quality [o]?	{o, oː}
22	Vowel quality [u]?	{u, uː}
23	Flap consonant?	{r}
24	Uvular consonant?	{N}
25	Glottal consonant?	{h}
26	Silence?	{<s>}

Table B.1: Questions on contexts used in HTK's decision-tree state-tying procedure. Japanese data

	Question	**Set of context phonemes**
1	Consonant?	{b, p, d, t, g, k, v, f, dh, th, z, s, zh, sh, jh, ch, m, em, n, en, ng, r, l, el, w, y, hh }
2	Voiced consonant ?	{b, d, g, v, dh, z, zh, jh, m, em, n, en, ng, r, l, el, w, y}
3	Vowel?	{aa, ae, ay, aw, oy, ow, eh, ey, er, ah, uw, uh, ih, iy}
4	Fricative consonant?	{s, sh, z, f, v, zh, th, dh, hh}
5	Alveolar consonant?	{t, d, s, z, n, en, l, el, r}
6	Mid vowel?	{er, ey, eh, ow, oy}
7	Nasal consonant?	{m, n, ng, en, em}
8	Stop consonant?	{p, b, t, d, k, g}
9	Front vowel?	{iy, ih, ey, eh, ae}
10	Open vowel?	{ae, aa, ay, aw, ah}
11	Diphthong vowel?	{ay, aw, oy, ey, ow}
12	Back vowel?	{ow, uw, uh, aa}
13	Post-alveolar consonant?	{sh, zh, r, ch, jh}
14	Bilabial consonant?	{p, b, m, em}
15	Close vowel?	{iy, ih, uw, uh}
16	Velar consonant?	{k, g, ng, w}
17	Approximant consonant?	{w, y, r}
18	Affricate consonant?	{ch, jh}
19	Central vowel?	{ah, er}
20	Dental consonant?	{th, dh}
21	Labio-dental consonant?	{f, v}
22	Palatal consonant?	{y}
23	Glottal consonant?	{hh}
24	Silence?	{<s>}

Table B.2: Questions on contexts used in HTK's decision-tree state-tying procedure. English data.

Appendix C

Allophones and sensitivity to frequency and variation

Here I compare Random and HTK-based allophones with respect to their sensitivity to frequency and variation. In the case of Random allophones, phonemes get split based on random partitioning of contexts. The occurrence of a given phoneme in a linguistic context once, is sufficient to instantiate the corresponding allophonic category. Thus, Random allophones are only sensitive to variation in the linguistic context, but not to the frequency of occurrence in this context. In the case of HTK-allophones, phonemes get split based on a combination of linguistic-based similarity rules and acoustic models. The resulting inventory of allophones reflects an interesting trade-off between linguistic, acoustic and statistical considerations. Thus we suspect the HTK-based allophones to be sensitive to both variation and frequency.

We quantify this observation through testing how variation and frequency fare in predicting the global number of allophones per phoneme in each case. Context variation will be characterized by the total number of attested contexts, and the frequency of occurrence by the global frequency of the phoneme.[1]

[1]This assumption, i.e., characterising variation and frequency of occurrence as, respectively, the global

APPENDIX C.

I ran a multilinear regression model specified as follows:

$$nbrAllo(p) = \alpha + \beta_1 freq(p) + \beta_2 nbrCont(p)$$

Where *nbrAll* stands for the number of allophones that a phoneme p gets, $freq(p)$ is the frequency of the phoneme in the corpus, and $nbrCont(p)$ is the number of linguistic contexts in which the phoneme occurs. Remember that a linguistic context is defined as the pair composed of the preceding and the following segments. For example, in the utterance "look!" (phonemically represented as l:uh:k) the phoneme /uh/ occurs in the context l_k.

As the number of attested contexts depends obviously on the frequency, the predictive variables must present some degree of collinearity. In fact the Pearson coefficient gives a value of $R = 0.69$ in the case of the Buckeye corpus and $R = 0.53$ in the case of the CSJ corpus. Nonetheless, such amount of collinearity is unlikely to affect adversely the estimation of regression statistics. In fact, Tolerance to collinearity is equal to $1 - R^2 = 0.53$ in English and 0.70 in Japanese, a number higher than the thresholds used in the literature, which varies from 0.1 (e.g., Tabachnick & Fidell, 2001) to 0.25 (Huber & Stephens, 1993).

After running the model in the case of HTK allophones, using an allophonic complexity equal to 32 times the size of the phonemic inventory, we get for the CSJ corpus: $F(2, 22) = 94.45$ ($p = 1.591\text{e-}11$ and $R^2 = 0.895$), indicating that we should clearly reject the null hypothesis that the variable $freq$ and *nbrCont* have no effect on *nbrAllo*. The results also show that the variable $freq$ is significantly controlling for the variable *nbrCont* ($p = 1.46e-09$), and *nbrCont* is significantly controlling for $freq$ ($p < 0.0113$). For the Buckeye corpus we get a similar pattern: $F(2, 38) = 35.86$ ($p = 1.782\text{e-}09$ and $R^2 = 0.653$) (thus rejecting the null hypothesis). As in Japanese, $freq$ is significantly controlling for the *nbrCont* ($p = 0.0001$), and *nbrCont* is significantly controlling for $freq$ ($p = 0.0297$).

number of attested contexts and the global frequency of the phoneme, is not perfect. A phoneme may occur in different contexts which can still be similar to each other, or occur with a high global frequency, but in only a few linguistic contexts. However, we consider this global variables to be reasonable proxies to our notions of variation and frequency of occurrence.

When we run the model on the artificial allophones, we get in the case of CSJ corpus: $F(2, 22) = 2.248$ ($p = 0.1293$ and $R^2 = 0.169$), which means that we cannot reject the null hypothesis according to which frequency and the number of attested contexts have no effect on the resulting number of allophones. When we run two separate linear models, using either $freq$ or $nbrCont$, we get a significant score $F(1, 23) = 4.66$ ($p = 0.041$ and $R^2 = 0.168$) in the case of $nbrCont$, but not in the case of $freq$, where the score is $F(1, 23) = 0.88$ ($p = 0.357$). As for the artificial allophones of the Buckeye corpus, the model gives a significant score $F(2, 38) = 8.966$ ($p <$ 6.46e-04 and $R^2 = 0.32$), the $nbrCont$ variable had a coefficient significantly different from zero ($p = 0.00406$), but not the coefficient associated with the variable $freq$ ($p \approx 1$), indicating that the effect is attributed mainly to the number of attested contexts.

To sum up, analysis of the HTK-based allophones shows that the number of allophones per phoneme is reliably predicted by both the frequency of the phoneme and the number of contexts in which this phoneme occurs. In contrast, analysis of artificial allophones shows that the number of allophones depends on the number of attested contexts, but not on the frequency of the phoneme.

Appendix D

How HTK-based allophones affect the phonemic classification

One factor that explains why an allophonic contrast would be invisible to the lexicon consists in the fact that some allophones are triggered on maximally different contexts (on the right and the left) as illustrated in the allophonic rule of Figure D.1.

$$/p/ \rightarrow \begin{cases} [p_1] \,/\, A\_\_B \\ [p_2] \,/\, C\_\_D \end{cases}$$

Figure D.1: example of an allophonic rule

When the set of contexts A doesnt overlap with C, and B does not overlap with D, it becomes impossible for the contrast ([p_1], [p_2]) to surface as an allomorphic pair of the form ([Xp_1] vs. [Xp_2]) (1), or the form ([p_1X] vs. [p_2X]) (2), where X refers to the rest of the word. The reason is simply because allophones have to share at least one triggering context to be able to form allomorphic variants of the same word. The shared context should be that of the penultimate segment of the word in the case of an allomorphic pair of the form (1), and the second segment in the case of an allomorphic pair of the form (2).

When asked to split the set of contexts in two distinct categories that trigger either [p_1] or [p_2] (this corresponds to the sets A__B and C__D in the example above), the random procedure will often make A overlap with C and B overlap with D, since this procedure is completely oblivious to the acoustic/phonetic similarity. This makes it always possible for a pair of allophones to generate a pair of lexical allomorphs.

To illustrate this, here is a simple example. Suppose we have a toy language composed of the following phonemic inventory:

$$P = \{a, b, c, d\}$$

The set of possible contexts corresponds to the set of all possible pairs of phonemes, i.e., the left context and the right context. We have 16 pairs, detailed in the following set:

$$C = \{a\_\_a, a\_\_b, a\_\_c, a\_\_d, b\_\_a, b\_\_b, b\_\_c, b\_\_d, c\_\_a, c\_\_b, c\_\_c, c\_\_d, d\_\_a, d\_\_b, d\_\_c, d\_\_d\}$$

A random partitioning in two triggering sets of contexts gives, for example, the two following sets:

$$A\_\_B = \{a\_\_a, b\_\_b, b\_\_c, c\_\_a, c\_\_b, c\_\_d, d\_\_c, d\_\_d\}$$

$$C\_\_D = \{a\_\_b, a\_\_c, a\_\_d, b\_\_a, b\_\_d, c\_\_a, d\_\_a, d\_\_b\}$$

As you can see, A__B and C__D present overlaps on both sides. For instance, all phonemes (a, b, c and d) occur in both A and C, and in both B and D. We can, in principle, find allomorphic pairs of the form [Xp_1)] vs. [Xp_2] or the form [p_1X] vs. [p_2X] (where the phoneme /p/ belongs to the set {a,b,c,d}). For example, we can have the allomorphic pair [abcp_1] / [abcp_2], since the pair (p_1,p_2) shares the context 'c' on the left (i.e., c $\in$ A $\cap$

C). Similarly, we can have a pair of the second form [$p_1$bca] / [$p_2$bca], since the allophones share the context 'b' on the right (i.e., b $\in$ B $\cap$ D).

The case of HTK allophones is different. The procedure of context splitting is not random, but performed based on phonetic and acoustic considerations (see Chapter 2). This procedure, contrary to the random one, tends to maximize within-category similarity, and maximize between-category distance, resulting in less overlaps. If we take our previous example, a possible context splitting can take the following form:

$$A\_\_B = \{a\_a, a\_b, a\_c, a\_d\}$$

$$C\_\_D = \{b\_a, b\_b, b\_c, b\_d, c\_a, c\_b, c\_c, c\_d, d\_a, d\_b, d\_c, d\_d\}$$

This happens, for instance, when the phoneme 'a', as a left context, triggers instances of the phoneme /p/ whose acoustic models are so similar to each other, and so different from the rest of the instances, that HTK will put the corresponding contexts into one category (here A), and the rest of contexts in a separate category. Thus, there is no overlap between the left triggering context of the allophones (A $\cap$ C = $\emptyset$). The probability of finding an allomorphic pair of the form [Xp_1] vs [Xp_2] is, therefore, zero. Note however, that in this toy example, there is still the possibility of finding an allomorphic pair of the form [p_1X] vs [p_2X]. Nonetheless, in real HTK splitting, phonemes tend either to have many allophones (the frequent ones), or to have none (the infrequent ones) even with inventory sizes as small as 2 allophones/phoneme in average (Figure D.2). In fact, the more allophones a phoneme has, the bigger the chance is to end up with non-overlapping triggering categories, and consequently, the more invisible allophonic pairs we will have. This general trends can also be observed in Table 3.5 where the number of invisible pairs increases as a function of the average number of allophones per phoneme.

To sum up, one factor explaining the discrepancy in performance between random and

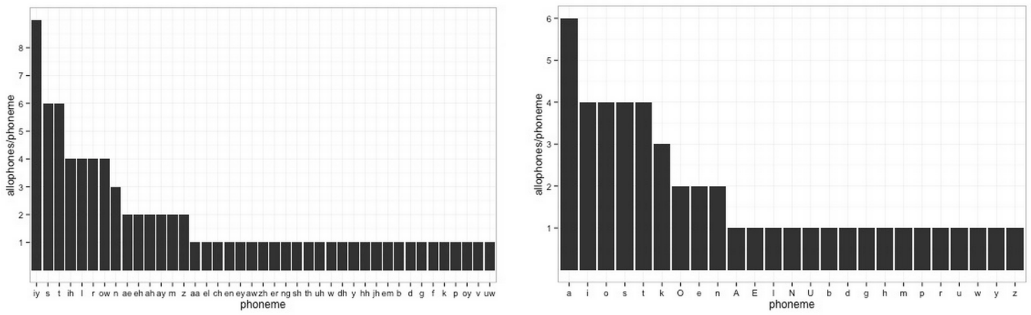

Figure D.2: The number of allophones for each phoneme in English data (left) and Japanese data (right), in the case of 2 allophones per phoneme in average.

HTK allophones lies in the fact that invisible allophonic pairs are more frequent in the latter, since it takes into account phonetic/acoustic similarity. The invisible pairs are automatically set to **0** in top down cues, which translates systematically into false alarms in the classification task.

Appendix E

Levels of phonetic clustering

To generate categories coarser than the phonemes, I collapsed the segments in Japanese from 25 to 13, 8, 4 and 2 (Figure E.1). Similarly, I collapsed the segments in English from 41 phonemes to 19, then to 10, 4 and 2 (Figure E.2).

Inventory (25)	H1 (13)	H2 (8)	H3 (4)	H4 (2)
a	A	Open	Vowel	Vowel
a:				
e	E	Mid		
e:				
o	O			
o:				
i	I	Close		
i:				
u	U			
u:				
m	Nasals	Nasal	Nasal	Consonant
n				
N				
b	B	Stop	Obstruent	
p				
d	D			
t				
g	G			
k				
s	Z	Fricative		
z				
h	H			
r	Flap	Flap	Sonorant	
y	Approximant	Approximant		
w				

Figure E.1: Hierarchical clustering of Japanese phonemes.

Inventory (41)	H (19)	H (10)	H(4)	H(2)
ae	A	Open	Vowel	Vowel
ah				
aa				
eh	E	Mid		
er				
iy	I	Close		
ih				
uw	U			
uh				
ay	Di	Di		
aw				
oy				
ow				
ey				
m	M	Nasal	Nasal	Consonant
em				
n	N			
en				
ng	ng			
b	B	Stops	Obstruent	
p				
d	D			
t				
g	G			
k				
ch	CH	Affricate		
jh				
f	F	Fricative		
v				
th	Th			
dh				
s	Z			
z				
sh	Zh			
zh				
hh	hh			
r	Approximant	Approximant	Sonorant	
w				
y				
l	L	Lateral		
el				

Figure E.2: Hierarchical clustering of English phonemes.

Appendix F

Instructions to participants

Below are the instruction givens to participants in the testing phase of the experiment of Chapter 6.

"You will listen to pairs of words from an artificial language. You should decide if they are same or different. The words can be different in the language even if they are similar (like the pair CAP-GAP in the English language). Conversely, they can be same (ex: GAP-GAP in the English language) even if they are pronounced slightly differently!"

Appendix G

Extension to infant studies

In this appendix, I propose a detailed description of how the experiment in Chapter 6 could be adapted to infant testing paradigms. The mechanism suggests that different phonological analyses of the form are assessed based on the coherence of the semantic categories they induce. We can break this down into three basic psychological assumptions:

1. Children can access different possible phonological analyses of the form (regardless of whether or not these analyses are relevant to identify words).

2. Children can assess meaning coherence (the category of *cats* is more semantically coherent that the category of both *cats* and *bats*, which is more coherent than the category of *cats*, *cows* and *cars*).

3. Each phonological analysis accessed thanks to (1) is assessed top-down, through (2).

The first assumption has already been given empirical support (J. Werker & Tees, 1984; McMurray & Aslin, 2005). The second and the third have yet to be tested. In what follows, I will propose a way to test if babies are sensitive to different levels of semantic coherence (**Experiment 1**), and whether this sensitivity is used to identify words in an ambiguous learning situation (**Experiment 2**).

APPENDIX G.

Participants In each experimental condition, we can test a group of 40 18 month-old infants (32 + 20% drop-out). Infants at this age were shown to be at crucial transition in phonological learning, interpreting phonological variation appropriately in some contexts but not others (Swingley & Aslin, 2007). Interestingly, this age also corresponds to qualitative and quantitative changes in semantic learning (Nazzi & Bertoncini, 2003).

Experiment 1 : sensitivity to semantic coherence

Upon hearing a few examples of a word paired with an object, children are generally able to determine an accurate semantic category for the word, extending the meaning from the single object they see (for example, they learn that /kæt/ can refer to cats of different colors and sizes, but not to dogs). Researchers have identified different semantic features that children bring to this task, two of which have received a lot of attention: the **shape bias** (Landau, Smith, & Jones, 1988) and the **taxonomic bias** (Markman, 1991).

I will use these documented features to manipulate the semantic coherence of a pair of objects at 3 levels. The first pair will consist of two similar objects (e.g., two *cats*), the second and third will be composed of two objects that vary, respectively, along the dimension of shape (e.g., a *cat* and a *cow*), and along both shape and taxonomy (e.g., a *cat* and a *car*). I propose to probe infants' graded sensitivity to semantic coherence in a naming extension task, using **looking time** as a behavioral index. This method has been successfully used before as a fine-grained behavioral tool to investigate sensitivity to different degrees of phonetic mismatch; e.g., McMurray and Aslin (2005) and White et al. (2008).

Experimental paradigm and procedure I will use an adapted version of the switch paradigm (Werker, Cohen, Lloyd, Casasola, & Stager, 1998). Babies are first habituated to the mapping between a non familiar word-form (e.g., *zem*) and a novel object. In the test phase, they are presented with different trials where they hear the same word-form

and they see objects that differ from the initial object in a graded way, along both shape and taxonomy. If children exhibit graded sensitivity to the degree of semantic mismatch, **we expect them to show increases in looking time as the distance between the original and new object increases**. Figure G.1 provides an illustration of the experimental design (here we used familiar items for ease of explanation, the real experiment will consist of novel referents).

PHASE	Habituation	Test		
WORD-FORM	"zem"	"zem"	"zem"	"zem"
REFERENT				
MISMATCH		Same	1 feature (shape)	2 features (shape+taxonomy)

Figure G.1: An illustration of the proposed experimental dessign.

Experiment 2: coherence modulates the phonological interpretation

Experiment 1 tests sensitivity to semantic coherence in young children; experiment 2 will test if this sensitivity modulates the phonological interpretation, as predicted by our mechanism. To this end, we can use a task similar to that of Werker et al. (1998). Infants are tested on their ability to map a minimal pair (e.g., "bin"/"din") to two different objects. Success at this task requires that infants be able to identify the difference between the two sounds as lexically contrastive, rather than considering it as mere variation of the same word. I vary the semantic coherence of the objects as explained in Experiment 1, and I predict that the more semantically distant the referents are, the less likely infants are to consider them as a plausible single meaning for one word; their phonological processing will, therefore, favor the two word interpretation.

APPENDIX G.

Experimental paradigm and procedure we can use the looking-while-listening paradigm (Fernald, Zangl, Portillo, & Marchman, 2008), since it has been shown to provide a more precise and fine-grained test of learning than the paradigm originally introduced by Werker et al. (1998). In the habituation phase, children will be taught the pairing between two words that differ minimally ("bin"/"din") and two objects. I vary the semantic coherence of that pair of objects as described in Experiment 2. One condition will have a pair of referents with one degree of semantic mismatch, and the other condition will have a pair of referents with 2 degrees of mismatch. In the test phase, children in each condition will hear one label ("bin" or "din") and see the two objects simultaneously. If babies have learned correctly the mapping between the words and the objects, they are expected to look longer to the correct object when they hear its label. I expect children's looking time at the correct object to correlate with the semantic coherence of the referents. This would provide support for the hypothesis that semantic coherence modulates the phonological interpretation of word-forms.

Appendix H

Whyisenglishsoeasytosegment?

Whyisenglishsoeasytosegment?

Abdellah Fourtassi[1], Benjamin Börschinger[2,3]
Mark Johnson[3] and Emmanuel Dupoux[1]
(1) Laboratoire de Sciences Cognitives et Psycholinguistique, ENS/EHESS/CNRS, Paris
(2) Department of Computing, Macquarie University
(3) Department of Computational Linguistics, Heidelberg University
{abdellah.fourtassi, emmanuel.dupoux}@gmail.com , {benjamin.borschinger, mark.johnson}@mq.edu.au

Abstract

Cross-linguistic studies on unsupervised word segmentation have consistently shown that English is easier to segment than other languages. In this paper, we propose an explanation of this finding based on the notion of segmentation ambiguity. We show that English has a very low segmentation ambiguity compared to Japanese and that this difference correlates with the segmentation performance in a unigram model. We suggest that segmentation ambiguity is linked to a trade-off between syllable structure complexity and word length distribution.

1 Introduction

During the course of language acquisition, infants must learn to segment words from continuous speech. Experimental studies show that they start doing so from around 7.5 months of age (Jusczyk and Aslin, 1995). Further studies indicate that infants are sensitive to a number of word boundary cues, like prosody (Jusczyk et al., 1999; Mattys et al., 1999), transition probabilities (Saffran et al., 1996; Pelucchi et al., 2009), phonotactics (Mattys et al., 2001), coarticulation (Johnson and Jusczyk, 2001) and combine these cues with different weights (Weiss et al., 2010).

Computational models of word segmentation have played a major role in assessing the relevance and reliability of different statistical cues present in the speech input. Some of these models focus mainly on *boundary detection*, and assess different strategies to identify them (Christiansen et al., 1998; Xanthos, 2004; Swingley, 2005; Daland and Pierrehumbert, 2011). Other models, sometimes called *lexicon-building algorithms*, learn the lexicon and the segmentation at the same time and use knowledge about the extracted lexicon to segment novel utterances. State-of-the-art lexicon-building segmentation algorithms are typically reported to yield better performance than word boundary detection algorithms (Brent, 1999; Venkataraman, 2001; Batchelder, 2002; Goldwater, 2007; Johnson, 2008b; Fleck, 2008; Blanchard et al., 2010).

As seen in Table 1, however, the performance varies considerably across languages with English winning by a high margin. This raises a generalizability issue for NLP applications, but also for the modeling of language acquisition since, obviously, it is not the case that in some languages, infants fail to acquire an adult lexicon. Are these performance differences only due to the fact that the algorithms might be optimized for English? Or do they also reflect some intrinsic linguistic differences between languages?

Lang.	F-score	Model	Reference
English	0.89	AG	Johnson (2009)
Chinese	0.77	AG	Johnson (2010)
Spanish	0.58	DP Bigram	Fleck (2008)
Arabic	0.56	WordEnds	Fleck (2008)
Sesotho	0.55	AG	Johnson (2008)
Japanese	0.55	BootLex	Batchelder (2002)
French	0.54	NGS-u	Boruta (2011)

Table 1: State-of-the-art unsupervised segmentation scores for eight languages.

The aim of the present work is to understand why English usually scores better than other languages, as far as unsupervised segmentation is concerned. As a comparison point, we chose Japanese because it is among the languages that have given the poorest word segmentation scores. In fact, Boruta et al. (2011) found an F-score around 0.41 using both Brent (1999)'s MBDP-1 and Venkataraman (2001)'s NGS-u models, and Batchelder (2002) found an F-score that goes from 0.40 to 0.55 depending on the corpus used. Japanese also differs typologically from English along several phonological dimensions such as

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 1–10,
Sofia, Bulgaria, August 8, 2013.

number of syllabic types, phonotactic constraints and rhythmic structure. Although most lexicon-building segmentation algorithms do not attempt to model these dimensions, they still might be relevant to speech segmentation and help explain the performance difference.

The structure of the paper is as follows. First, we present the class of lexical-building segmentation algorithm that we use in this paper (Adaptor Grammar), and our English and Japanese corpora. We then present data replicating the basic finding that segmentation performance is better for English than for Japanese. We then explore the hypothesis that this finding is due to an intrinsic difference in segmentation ambiguity in the two languages, and suggest that the source of this difference rests in the structure of the phonological lexicon in the two languages. Finally, we use these insights to try and reduce the gap between Japanese and English segmentation through a modification of the Unigram model where multiple linguistic levels are learned jointly.

2 Computational Framework and Corpora

2.1 Adaptor Grammar

In this study, we use the Adaptor Grammar framework (Johnson et al., 2007) to test different models of word segmentation on English and Japanese Corpora. This framework makes it possible to express a class of hierarchical non-parametric Bayesian models using an extension of probabilistic context-free grammars called Adaptor Grammar (AG). It allows one to easily define models that incorporate different assumptions about linguistic structure and is therefore a useful practical tool for exploring different hypotheses about word segmentation (Johnson, 2008b; Johnson, 2008a; Johnson et al., 2010; Börschinger et al., 2012).

For mathematical details and a description of the inference procedure for AGs, we refer the reader to Johnson et al. (2007). Briefly, AG uses the non-parametric Pitman-Yor-Process (Pitman and Yor, 1997) which, as in Minimum Description lengths models, finds a compact representation of the input by re-using frequent structures (here, words).

2.2 Corpora

In the present study, we used both Child Directed Speech (CDS) and Adult Directed Speech (ADS) corpora. English CDS was derived from the Bernstein-Ratner corpus (Bernstein-Ratner, 1987), which consists in transcribed verbal interaction of parents with nine children between 1 and 2 years of age. We used the 9,790 utterances that were phonemically transcribed by Brent and Cartwright (1996). Japanese CDS consists in the first 10, 000 utterances of the Hamasaki corpus (Hamasaki, 2002). It provides a phonemic transcript of spontaneous speech to a single child collected from when the child was 2 up to when it was 3.5 years old. Both CDS corpora are available from the CHILDES database (MacWhinney, 2000).

As for English ADS, we used the first 10,000 utterances of the Buckeye Speech Corpus (Pitt et al., 2007) which consists in spontaneous conversations with 40 speakers in American English. To make it comparable to the other corpora in this paper, we only used the idealized phonemic transcription. Finally, for Japanese ADS, we used the first 10,000 utterances of a phonemic transcription of the Corpus of Spontaneous Japanese (Maekawa et al., 2000). It consists of recorded spontaneous conversations, or public speeches in different fields ranging from engineering to humanities. For each corpus, we present elementary statistics in Table 2.

3 Unsupervised segmentation with the Unigram Model

3.1 Setup

In this experiment we used the Adaptor Grammar framework to implement a Unigram model of word segmentation (Johnson et al., 2007). This model has been shown to be equivalent to the original MBDP-1 segmentation model (see Goldwater (2007)). The model is defined as:

$$Utterance \rightarrow \underline{Word}^{+}$$
$$\underline{Word} \rightarrow Phoneme^{+}$$

In the AG framework, an underlined non-terminal indicates that this non-terminal is adapted, i.e. that the AG will cache (and learn probabilities for) entire sub-trees rooted in this non-terminal. Here, *Word* is the only unit that the model effectively learns, and there are no dependencies between the words to be learned. This grammar states that an utterance must be analyzed in terms of one or more Words, where a Word is a

Corpus	Child Directed Speech		Adult Directed Speech	
	English	Japanese	English	Japanese
Tokens				
Utterances	9,790	10,000	10,000	10,000
Words	33,399	27,362	57,185	87,156
Phonemes	95,809	108,427	183,196	289,264
Types				
Words	1,321	2,389	3,708	4,206
Phonemes	50	30	44	25
Average Lengths				
Words per utterance	3.41	2.74	5.72	8.72
Phonemes per utterance	9.79	10.84	18.32	28.93
Phonemes per word	2.87	3.96	3.20	3.32

Table 2 : Characteristics of phonemically transcribed corpora

sequence of Phonemes.

We ran the model twice on each corpus for 2,000 iterations with hyper-parameter sampling and we collected samples throughout the process, following the methodology of Johnson and Goldwater (2009)[1]. For evaluation, we performed their Minimum Bayes Risk decoding using the collected samples to get a single score.

3.2 Evaluation

For the evaluation, we used the same measures as Brent (1999), Venkataraman (2001) and Goldwater (2007), namely token Precision (P), Recall (R) and F-score (F). Precision is defined as the number of correct word tokens found out of all tokens posited. Recall is the number of correct word tokens found out of all tokens in the gold standard. The F-score is defined as the harmonic mean of Precision and Recall , $F = \frac{2*P*R}{P+R}$.

We will refer to these scores as the *segmentation* scores. In addition, we define similar measures for word *boundaries* and word types in the *lexicon*.

3.3 Results and discussion

The results are shown in Table 3. As expected, the model yields substantially better scores in English than Japanese, for both CDS and ADS. In addition, we found that in both languages, ADS yields slightly worse results than CDS. This is to be expected because ADS uses between 60% and 300% longer utterances than CDS, and as a result presents the learner with a more difficult segmentation problem. Moreover, ADS includes between 70% and 280% more word types than CDS, making it a more difficult lexical learning problem. Note, however, that despite these large differences in corpus statistics, the difference in segmentation performance between ADS and CDS are small compared to the differences between Japanese and English.

An error analysis on English data shows that most errors come from the Unigram model mistaking high frequency collocations for single words (see also Goldwater (2007)). This leads to an under-segmentation of chunks like "a boy" or "is it" [2]. Yet, the model also tends to break off frequent morphological affixes, especially "-ing" and "-s" , leading to an over-segmentation of words like "talk ing" or "black s".

Similarly, Japanese data shows both over- and under-segmentation errors. However, over-segmentation is more severe than for English, as it does not only affect affixes, but surfaces as breaking apart multi-syllabic words. In addition, Japanese segmentation faces another kind of error which acts across word boundaries. For example, "ni kashite" is segmented as "nika shite" and "nurete inakatta" as "nure tei na katta". This leads to an output lexicon that, on the one hand, allows for a more compact analysis of the corpus than the true lexicon: the number of word types drops from 2,389 to 1,463 in CDS and from 4,206 to 2,372 in ADS although the average token length – and consequently, overall number of tokens – does not change as dramatically, dropping from 3.96 to

[1] We used incremental initialization

[2] For ease of presentation, we use orthography to present examples although all experiments are run on phonemic transcripts.

	Child Directed Speech						Adult Directed Speech					
	English			Japanese			English			Japanese		
	F	P	R	F	P	R	F	P	R	F	P	R
Segmentation	**0.77**	0.76	0.77	**0.55**	0.51	0.61	**0.69**	0.66	0.73	**0.50**	0.48	0.52
Boundaries	0.87	0.87	0.88	0.72	0.63	0.83	0.86	0.81	0.91	0.76	0.74	0.79
Lexicon	0.62	0.65	0.59	0.33	0.43	0.26	0.41	0.48	0.36	0.30	0.42	0.23

Table 3 : Word segmentation scores of the Unigram model

3.31 for CDS and from 3.32 to 3.12 in ADS. On the other hand, however, most of the output lexicon items are not valid Japanese words and this leads to the bad lexicon F-scores. This, in turn, leads to the bad overall segmentation performance.

In brief, we have shown that, across two different corpora, English yields consistently better segmentation results than Japanese for the Unigram model. This confirms and extends the results of Boruta et al. (2011) and Batchelder (2002). It strongly suggests that the difference is neither due to a specific choice of model nor to particularities of the corpora, but reflects a fundamental property of these two languages.

In the following section, we introduce the notion of *segmentation ambiguity*, it to English and Japanese data, and show that it correlates with segmentation performance.

4 Intrinsic Segmentation Ambiguity

Lexicon-based segmentation algorithms like MBDP-1, NGS-u and the AG Unigram model learn the lexicon and the segmentation at the same time. This makes it difficult, in case of poor performance, to see whether the problem comes from the intrinsic segmentability of the language or from the quality of the extracted lexicon. Our claim is that Japanese is intrinsically more difficult to segment than English, even when a good lexicon is already assumed. We explore this hypothesis by studying segmentation alone, assuming a perfect (Gold) lexicon.

4.1 Segmentation ambiguity

Without any information, a string of N phonemes could be segmented in 2^{N-1} ways. When a lexicon is provided, the set of possible segmentations is reduced to a smaller number. To illustrate this, suppose we have to segment the input utterance: /ay s k r iy m/ [3], and that the lexicon contains the following words : /ay/ (I), /s k r iy m/ (scream), /ay s/ (ice), /k r iy m/ (cream). Only two segmentations are possible : /ay skriym/ (I scream) and /ays kriym/ (ice cream).

We are interested in the ambiguity generated by the different possible parses that result from such a supervised segmentation. In order to quantify this idea in general, we define a *Normalized Segmentation Entropy*. To do this, we need to assign a probability to every possible segmentation. To this end, we use a unigram model where the probability of a lexical item is its normalized frequency in the corpus and the probability of a parse is the product of the probabilities of its terms. In order to obtain a measure that does not depend on the utterance length, we normalize by the number of possible boundaries in the utterance. So for an utterance of length N, the Normalized Segmentation Entropy (NSE) is computed using Shannon formula (Shannon, 1948) as follows:

$$NSE = -\sum_i P_i log_2(P_i)/(N-1)$$

where P_i is the probability of the parse i .

For CDS data we found Normalized Segmentation Entropies of 0.0021 bits for English and 0.0156 bits for Japanese. In ADS data we found similar results with 0.0032 bits for English and 0.0275 bits for Japanese. This means that Japanese needs between 7 and 8 times more bits than English to encode segmentation information. This is a very large difference, which is of the same magnitude in CDS and ADS. These differences clearly show that intrinsically, Japanese is more ambiguous than English with regards to segmentation.

One can refine this analysis by distinguishing two sources of ambiguity: ambiguity *across word boundaries*, as in "ice cream / [ay s] [k r iy m]"

[3] We use ARPABET notation to represent phonemic input.

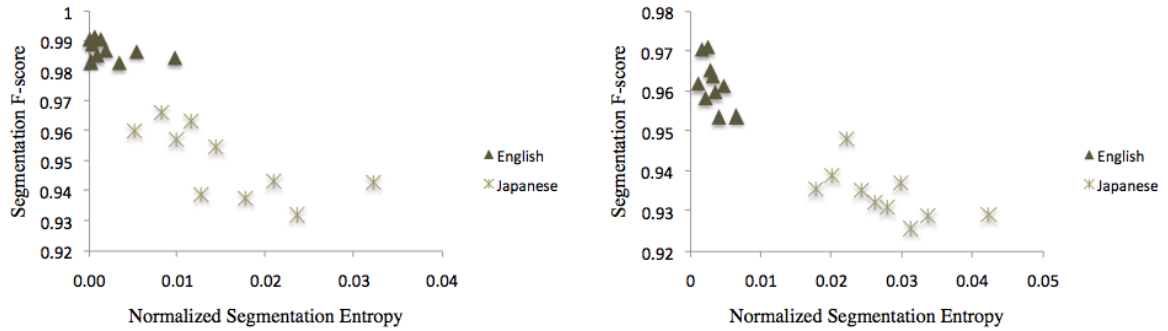

Figure 1 : Correlation between Normalized Segmentation Entropy (in bits) and the segmentation F-score for CDS (left) and ADS (Right)

vs "I scream / [ay] [s k r iy m]". And ambiguity *within the lexicon*, that occurs when a lexical item is composed of two or more sub-words (like in "Butterfly").

Since we are mainly investigating lexicon-building models, it is important to measure the ambiguity within the lexicon itself, in the ideal case where this lexicon is perfect. To this end, we computed the average number of segmentations for a lexicon item. For example, the word "butterfly" has two possible segmentations : the original word "butterfly" and a segmentation comprising the two sub-words : "butter" and "fly". For English tokens, we found an average of 1.039 in CDS and 1.057 in ADS. For Japanese tokens, we found an average of 1.811 in CDS and 1.978 in ADS. English's averages are close to 1, indicating that it doesn't exhibit lexicon ambiguity. Japanese, however, has averages close to 2 which means that lexical ambiguity is quite systematic in both CDS and ADS.

4.2 Segmentation ambiguity and supervised segmentation

The intrinsic ambiguity in Japanese only shows that a given sentence has multiple possible segmentations. What remains to be demonstrated is that these multiple segmentations result in systematic segmentation errors. To do this we propose a supervised segmentation algorithm that enumerates all possible segmentations of an utterance based on the gold lexicon, and selects the segmentation with the highest probability. In CDS data, this algorithm yields a segmentation F-score equal to 0.99 for English and 0.95 for Japanese. In ADS we find an F-score of 0.96 for English and 0.93 for Japanese. These results show that lexical information alone plus word frequency eliminates almost all segmentation errors in English, especially for CDS. As for Japanese, even if the scores remain impressively high, the lexicon alone is not sufficient to eliminate all the errors. In other words, even with a gold lexicon, English remains easier to segment than Japanese.

To quantify the link between segmentation entropy and segmentation errors, we binned the sentences of our corpus in 10 bins according to the Normalized Segmentation Entropy, and correlate this with the average segmentation F-score for each bin. As shown Figure 1, we found significant correlations: ($R = -0.86$, $p < 0.001$) for CDS and ($R = -0.93$, $p < 0.001$) for ADS, showing that segmentation ambiguity has a strong effect even on supervised segmentation scores. The correlation within language was also significant but only in the Japanese data : $R = -0.70$ for CDS and $R = -0.62$ for ADS.

Next, we explore one possible reason for this structural difference between Japanese and English, especially at the level of the lexicon.

4.3 Syllable structure and lexical composition of Japanese and English

One of the most salient differences between English and Japanese phonology concerns their syllable structure. This is illustrated in Figure 2 (above), where we plotted the frequency of the different syllabic structures of monosyllabic tokens in English and Japanese CDS. The statistics show that English has a very rich syllabic composition where a diversity of consonant clusters is allowed, whereas Japanese syllable structure is quite simple and mostly composed of the default CV type. This difference is bound to have an effect on the structure of the lexicon. Indeed, Japanese has to use

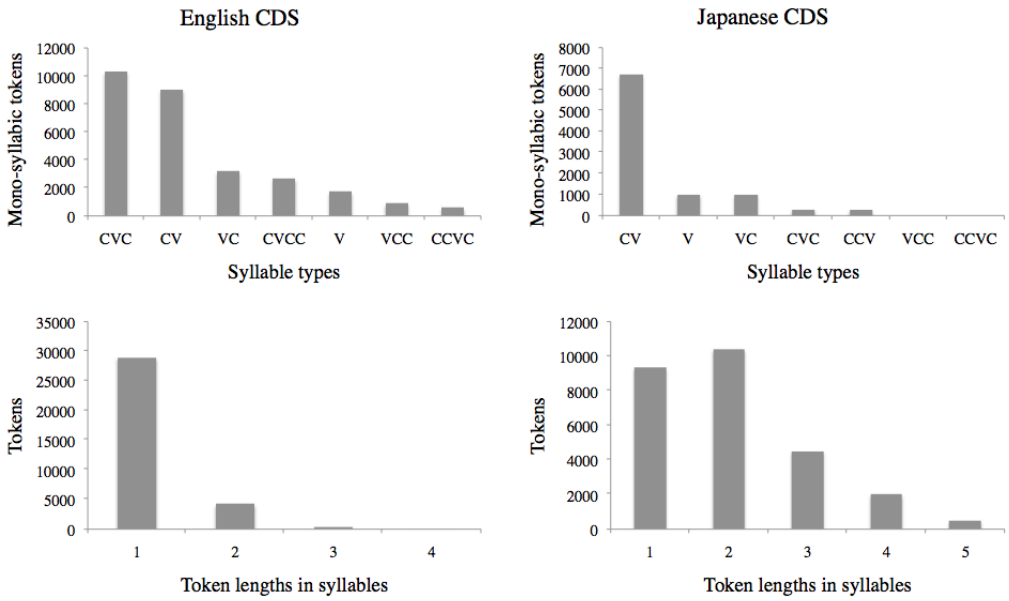

Figure 2 : Trade-off between the complexity of syllable structure (above) and the word token length in terms of syllables (below) for English and Japanese CDS.

multisyllabic words in order to achieve a large size lexicon, whereas, in principle, English could use mostly monosyllables. In Figure 2 (below) we display the distribution of word length as measured in syllables in the two languages for the CDS corpora. The English data is indeed mostly composed of mono-syllabic words whereas the Japanese one is made of words of more varied lengths. Overall, we have documented a trade-off between the diversity of syllable structure on the one hand, and the diversity of word lengths on the other (see Table 4 for a summary of this tradeoff expressed in terms of entropy).

	CDS		ADS	
	Eng.	Jap.	Eng.	Jap.
Syllable types	2.40	1.38	2.58	1.03
Token lengths	0.62	2.04	0.99	1.69

Table 4 : Entropies of syllable types and token lengths in terms of syllables (in bits)

We suggest that this trade-off is responsible for the difference in the lexicon ambiguity across the two languages. Specifically, the combination of a small number of syllable types and, as a consequence, the tendency for multi-syllabic word types in Japanese makes it likely that a long word will be composed of smaller ones. This cannot happen very often in English, since most words are monosyllabic, and words smaller than a syllable are not allowed.

5 Improving Japanese unsupervised segmentation

We showed in the previous section that ambiguity impacts segmentation even with a gold lexicon, mainly because the lexicon itself could be ambiguous. In an unsupervised segmentation setting, the problem is worse because ambiguity within and across word boundaries leads to a bad lexicon, which in turn results in more segmentation errors. In this section, we explore the possibility of mitigating some of these negative consequences.

In section 3, we saw that when the Unigram model tries to learn Japanese words, it produces an output lexicon composed of both over- and under-segmented words in addition to words that result from a segmentation across word boundaries. One way to address this is by learning multiple kinds of units jointly, rather than just words; indeed, previous work has shown that richer models with multiple levels improve segmentation for English (Johnson, 2008a; Johnson and Goldwater, 2009).

5.1 Two dependency levels

As a first step, we will allow the model to not just learn words but to also memorize sequences of words. Johnson (2008a) introduced these units as "collocations" but we choose to use the more neutral notion of *level* for reasons that become clear shortly. Concretely, the grammar is:

	CDS						ADS					
	English			Japanese			English			Japanese		
	F	P	R	F	P	R	F	P	R	F	P	R
Level 1												
Segmentation	**0.81**	0.77	0.86	0.42	0.33	0.55	**0.70**	0.63	0.78	0.42	0.35	0.50
Boundaries	0.91	0.84	0.98	0.63	0.47	0.96	0.86	0.76	0.98	0.73	0.61	0.90
Lexicon	0.64	0.79	0.54	0.18	0.55	0.10	0.36	0.56	0.26	0.15	0.68	0.08
Level 2												
Segmentation	0.33	0.45	0.26	**0.59**	0.65	0.53	0.50	0.60	0.43	**0.45**	0.54	0.38
Boundaries	0.56	0.98	0.40	0.71	0.87	0.60	0.76	0.95	0.64	0.73	0.92	0.60
Lexicon	0.36	0.25	0.59	0.47	0.44	0.49	0.46	0.38	0.56	0.43	0.37	0.50

Table 5 : Word segmentation scores of the two levels model

$$Utterance \rightarrow \underline{level2}^{+}$$
$$\underline{level2} \rightarrow \underline{level1}^{+}$$
$$\underline{level1} \rightarrow Phoneme^{+}$$

We run this model under the same conditions as the Unigram model but evaluate two different situations. The model has no inductive bias that would force it to equate *level1* with words, rather than *level2*. Consequently, we evaluate the segmentation that is the result of taking there to be a boundary between every *level1* constituent (Level 1 in Table 5) and between every *level2* constituent (Level 2 in Table 5). From these results , we see that English data has better scores when the lower level represents the Word unit and when the higher level captures regularities above the word. However, Japanese data is best segmented when the higher level is the Word unit and the lower level captures sub-word regularities.

Level 1 generally tends to over-segment utterances as can be seen by comparing the Boundary Recall and Precision scores (Goldwater, 2007). In fact when the Recall is much higher than the Precision, we can say that the model has a tendency to over-segment. Conversely, we see that Level 2 tends to under-segment utterances as the Boundary Precision is higher than the Recall.

Over-segmentation at Level 1 seems to benefit English since it counteracts the tendency of the Unigram model to cluster high frequency collocations. As far as segmentation is concerned, this effect seems to outweigh the negative effect of breaking words apart (especially in CDS), as English words are mostly monosyllabic.

For Japanese, under-segmentation at Level 2 seems to be slightly less harmful than over-segmentation at Level 1, as it prevents, to some extent, multi-syllabic words to be split. However, the scores are not very different from the ones we had with the Unigram model and slightly worse for the ADS. What seems to be missing is an intermediate level where over- and under-segmentation would counteract one another.

5.2 Three dependency levels

We add a third dependency level to our model as follows :

$$Utterance \rightarrow \underline{level3}^{+}$$
$$\underline{level3} \rightarrow \underline{level2}^{+}$$
$$\underline{level2} \rightarrow \underline{level1}^{+}$$
$$\underline{level1} \rightarrow Phoneme^{+}$$

As with the previous model, we test each of the three levels as the word unit, the results are shown in Table 6.

Except for English CDS, all the corpora have their best scores with this intermediate level. Level 1 tends to over-segment Japanese utterances into syllables and English utterances into morphemes. Level 3, however, tends to highly under-segment both languages. English CDS seems to be already under-segmented at Level 2, very likely caused by the large number of word collocations like ”is-it” and ”what-is”, an observation also made by Börschinger et al. (2012) using different English CDS corpora. English ADS is quantitatively more sensitive to over-segmentation than CDS mainly because it has a richer morphological structure and relatively longer words in terms of syllables (Table 4).

	CDS						ADS					
	English			Japanese			English			Japanese		
	F	P	R	F	P	R	F	P	R	F	P	R
Level 1												
Segmentation	**0.79**	0.74	0.85	0.27	0.20	0.41	0.35	0.28	0.48	0.37	0.30	0.47
Boundaries	0.89	0.81	0.99	0.56	0.39	0.99	0.68	0.52	0.99	0.70	0.57	0.93
Lexicon	0.58	0.76	0.46	0.10	0.47	0.05	0.13	0.39	0.07	0.10	0.70	0.05
Level 2												
Segmentation	0.49	0.60	0.42	**0.70**	0.70	0.70	**0.77**	0.76	0.79	**0.60**	0.65	0.55
Boundaries	0.71	0.97	0.56	0.81	0.82	0.81	0.90	0.88	0.92	0.81	0.90	0.74
Lexicon	0.51	0.41	0.64	0.53	0.59	0.47	0.58	0.69	0.50	0.51	0.57	0.46
Level 3												
Segmentation	0.18	0.31	0.12	0.39	0.53	0.30	0.43	0.55	0.36	0.28	0.42	0.21
Boundaries	0.26	0.99	0.15	0.46	0.93	0.31	0.71	0.98	0.55	0.59	0.96	0.43
Lexicon	0.17	0.10	0.38	0.32	0.25	0.41	0.37	0.28	0.51	0.27	0.20	0.42

Table 6 : Word segmentation scores of the three levels model

6 Conclusion

In this paper we identified a property of language, *segmentation ambiguity*, which we quantified through Normalized Segmentation Entropy. We showed that this quantity predicts performance in a supervised segmentation task.

With this tool we found that English was intrinsically less ambiguous than Japanese, accounting for the systematic difference found in this paper. More generally, we suspect that Segmentation Ambiguity would, to some extent, explain much of the difference observed across languages (Table 1). Further work needs to be carried out to test the robustness of this hypothesis on a larger scale.

We showed that allowing the system to learn at multiple levels of structure generally improves performance, and compensates partially for the negative effect of segmentation ambiguity on unsupervised segmentation (where a bad lexicon amplifies the effect of segmentation ambiguity). Yet, we end up with a situation where the best level of structure may not be the same across corpora or languages, which raises the question as to how to determine which level is the correct lexical level, i.e., the level that can sustain successful grammatical and semantic learning. Further research is needed to answer this question.

Generally speaking, ambiguity is a challenge in many speech and language processing tasks: for example part-of-speech tagging and word sense disambiguation tackle lexical ambiguity, probabilistic parsing deals with syntactic ambiguity and speech act interpretation deals with pragmatic ambiguities. However, to our knowledge, ambiguity has rarely been considered as a serious problem in word segmentation tasks.

As we have shown, the lexicon-based approach does not completely solve the segmentation ambiguity problem since the lexicon itself could be more or less ambiguous depending on the language. Evidently, however, infants in all languages manage to overcome this ambiguity. It has to be the case, therefore, that they solve this problem through the use of alternative strategies, for instance by relying on sub-lexical cues (see Jarosz and Johnson (2013)) or by incorporating semantic or syntactic constraints (Johnson et al., 2010). It remains a major challenge to integrate these strategies within a common model that can learn with comparable performance across typologically distinct languages.

Acknowledgements

The research leading to these results has received funding from the European Research Council (FP/2007-2013) / ERC Grant Agreement n. ERC-2011-AdG-295810 BOOTPHON, from the Agence Nationale pour la Recherche (ANR-2010-BLAN-1901-1 BOOTLANG, ANR-11-0001-02 PSL* and ANR-10-LABX-0087) and the Fondation de France. This research was also supported under the Australian Research Council's Discovery Projects funding scheme (project numbers DP110102506 and DP110102593).

References

Eleanor Olds Batchelder. 2002. Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, 83(2):167–206.

N. Bernstein-Ratner. 1987. The phonology of parent-child speech. In K. Nelson and A. van Kleeck, editors, *Children's Language*, volume 6. Erlbaum, Hillsdale, NJ.

Daniel Blanchard, Jeffrey Heinz, and Roberta Golinkoff. 2010. Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language*, 37(3):487–511.

Benjamin Börschinger, Katherine Demuth, and Mark Johnson. 2012. Studying the effect of input size for Bayesian word segmentation on the Providence corpus. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, pages 325–340, Mumbai, India. Coling 2012 Organizing Committee.

Luc Boruta, Sharon Peperkamp, Benoît Crabbé, and Emmanuel Dupoux. 2011. Testing the robustness of online word segmentation: Effects of linguistic diversity and phonetic variation. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–9, Portland, Oregon, USA, June. Association for Computational Linguistics.

M. Brent and T. Cartwright. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61:93–125.

M. Brent. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105.

Morten H Christiansen, Joseph Allen, and Mark S Seidenberg. 1998. Learning to segment speech using multiple cues: A connectionist model. *Language and cognitive processes*, 13(2-3):221–268.

Robert Daland and Janet B Pierrehumbert. 2011. Learning diphone-based segmentation. *Cognitive Science*, 35(1):119–155.

Margaret M. Fleck. 2008. Lexicalized phonotactic word segmentation. In *Proceedings of ACL-08: HLT*, pages 130–138, Columbus, Ohio, June. Association for Computational Linguistics.

Sharon Goldwater. 2007. *Nonparametric Bayesian Models of Lexical Acquisition*. Ph.D. thesis, Brown University.

Naomi Hamasaki. 2002. The timing shift of two-year-olds responses to caretakers yes/no questions. In *Studies in language sciences (2)Papers from the 2nd Annual Conference of the Japanese Society for Language Sciences*, pages 193–206.

Gaja Jarosz and J Alex Johnson. 2013. The richness of distributional cues to word boundaries in speech to young children. *Language Learning and Development*, (ahead-of-print):1–36.

Mark Johnson and Katherine Demuth. 2010. Unsupervised phonemic Chinese word segmentation using Adaptor Grammars. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 528–536, Beijing, China, August. Coling 2010 Organizing Committee.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 317–325, Boulder, Colorado, June. Association for Computational Linguistics.

Elizabeth K. Johnson and Peter W. Jusczyk. 2001. Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44:1–20.

Mark Johnson, Thomas Griffiths, and Sharon Goldwater. 2007. Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 139–146, Rochester, New York. Association for Computational Linguistics.

Mark Johnson, Katherine Demuth, Michael Frank, and Bevan Jones. 2010. Synergies in learning words and their referents. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1018–1026.

Mark Johnson. 2008a. Unsupervised word segmentation for Sesotho using Adaptor Grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27, Columbus, Ohio, June. Association for Computational Linguistics.

Mark Johnson. 2008b. Using Adaptor Grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*, pages 398–406, Columbus, Ohio. Association for Computational Linguistics.

Peter W Jusczyk and Richard N Aslin. 1995. Infants detection of the sound patterns of words in fluent speech. *Cognitive psychology*, 29(1):1–23.

Peter W. Jusczyk, E. A. Hohne, and A. Bauman. 1999. Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics*, 61:1465–1476.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk. Transcription, format and programs*, volume 1. Lawrence Erlbaum.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of japanese. In *proc. LREC*, volume 2, pages 947–952.

Sven L Mattys, Peter W Jusczyk, Paul A Luce, James L Morgan, et al. 1999. Phonotactic and prosodic effects on word segmentation in infants. *Cognitive psychology*, 38(4):465–494.

Sven L Mattys, Peter W Jusczyk, et al. 2001. Do infants segment words or recurring contiguous patterns? *Journal of experimental psychology, human perception and performance*, 27(3):644–655.

Bruna Pelucchi, Jessica F Hay, and Jenny R Saffran. 2009. Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, 113(2):244–247.

J. Pitman and M. Yor. 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900.

M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and Fosler-Lussier. 2007. Buckeye corpus of conversational speech.

J. Saffran, R. Aslin, and E. Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274:1926–1928.

Claude Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Daniel Swingley. 2005. Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50:86–132.

A. Venkataraman. 2001. A statistical model for word discovery in transcribed speech. *Computational Linguistics*, 27(3):351–372.

Daniel J Weiss, Chip Gerfen, and Aaron D Mitchel. 2010. Colliding cues in word segmentation: the role of cue strength and general cognitive processes. *Language and Cognitive Processes*, 25(3):402–422.

Aris Xanthos. 2004. Combining utterance-boundary and predictability approaches to speech segmentation. In *First Workshop on Psycho-computational Models of Human Language Acquisition*, page 93.

Appendix I

A corpus-based evaluation method for DSMs

A corpus-based evaluation method for Distributional Semantic Models

Abdellah Fourtassi[1,2]
abdellah.fourtassi@gmail.com

Emmanuel Dupoux[2,3]
emmanuel.dupoux@gmail.com

[1]Institut d'Etudes Cognitives, Ecole Normale Superieure, Paris
[2]Laboratoire de Sciences Cognitives et Psycholinguistique, CNRS, Paris
[3]Ecole des Hautes Etudes en Sciences Sociales, Paris

Abstract

Evaluation methods for Distributional Semantic Models typically rely on behaviorally derived gold standards. These methods are difficult to deploy in languages with scarce linguistic/behavioral resources. We introduce a corpus-based measure that evaluates the stability of the lexical semantic similarity space using a pseudo-synonym same-different detection task and no external resources. We show that it enables to predict two behavior-based measures across a range of parameters in a Latent Semantic Analysis model.

1 Introduction

Distributional Semantic Models (DSM) can be traced back to the hypothesis proposed by Harris (1954) whereby the meaning of a word can be inferred from its context. Several implementations of Harris's hypothesis have been proposed in the last two decades (see Turney and Pantel (2010) for a review), but comparatively little has been done to develop reliable evaluation tools for these implementations. Models evaluation is however an issue of crucial importance for practical applications, i.e., when trying to optimally set the model's parameters for a given task, and for theoretical reasons, i.e., when using such models to approximate semantic knowledge.

Some evaluation techniques involve assigning probabilities to different models given the observed corpus and applying maximum likelihood estimation (Lewandowsky and Farrell, 2011). However, computational complexity may prevent the application of such techniques, besides these probabilities may not be the best predictor for the model performance on a specific task (Blei, 2012). Other commonly used methods evaluate DSMs by comparing their semantic representation to a behaviorally derived gold standard. Some standards are derived from the TOEFL synonym test (Landauer and Dumais, 1997), or the Nelson word associations norms (Nelson et al., 1998). Others use results from semantic priming experiments (Hutchison et al., 2008) or lexical substitutions errors (Andrews et al., 2009). Baroni and Lenci (2011) set up a more refined gold standard for English specifying different kinds of semantic relationship based on dictionary resources (like WordNet and ConceptNet).

These behavior-based evaluation methods are all resource intensive, requiring either linguistic expertise or human-generated data. Such methods might not always be available, especially in languages with fewer resources than English. In this situation, researchers usually select a small set of high-frequency target words and examine their nearest neighbors (the most similar to the target) using their own intuition. This is used in particular to set the model parameters. However, this rather informal method represents a "cherry picking" risk (Kievit-Kylar and Jones, 2012), besides it is only possible for languages that the researcher speaks.

Here we introduce a method that aims at providing a rapid and quantitative evaluation for DSMs using an internal gold standard and requiring no external resources. It is based on a simple same-different task which detects pseudo-synonyms randomly introduced in the corpus. We claim that this measure evaluates the intrinsic ability of the model to capture lexical semantic similarity. We validate it against two behavior-based evaluations (Free association norms and the TOEFL synonym test) on semantic representations extracted from a Wikipedia corpus using one of the most commonly used distributional semantic models : the Latent Semantic Analysis (LSA, Landauer and Dumais (1997)).

In this model, we construct a word-document matrix. Each word is represented by a row, and

Proceedings of the ACL Student Research Workshop, pages 165–171,
Sofia, Bulgaria, August 4-9 2013.

each document is represented by a column. Each matrix cell indicates the occurrence frequency of a given word in a given context. Singular value decomposition (a kind of matrix factorization) is used to extract a reduced representation by truncating the matrix to a certain size (which we call the semantic dimension of the model). The cosine of the angle between vectors of the resulting space is used to measure the semantic similarity between words. Two words end up with similar vectors if they co-occur multiple times in similar contexts.

2 Experiment

We constructed three successively larger corpora of 1, 2 and 4 million words by randomly selecting articles from the original "Wikicorpus" made freely available on the internet by Reese et al. (2010). Wikicorpus is itself based on articles from the collaborative encyclopedia Wikipedia. We selected the upper bound of 4 M words to be comparable with the typical corpus size used in theoretical studies on LSA (see for instance Landauer and Dumais (1997) and Griffiths et al. (2007)). For each corpus, we kept only words that occurred at least 10 times and we excluded a stop list of high frequency words with no conceptual content such as: the, of, to, and ... This left us with a vocabulary of 8 643, 14 147 and 23 130 words respectively. For the simulations, we used the free software Gensim (Řehůřek and Sojka, 2010) that provides an online Python implementation of LSA.

We first reproduced the results of Griffiths et al. (2007), from which we derived the behavior-based measure. Then, we computed our corpus-based measure with the same models.

2.1 The behavior-based measure

Following Griffiths et al. (2007), we used the free association norms collected by Nelson et al. (1998) as a gold standard to study the psychological relevance of the LSA semantic representation. The norms were constructed by asking more than 6000 participants to produce the first word that came to mind in response to a cue word. The participants were presented with 5,019 stimulus words and the responses (word associates) were ordered by the frequency with which they were named. The overlap between the words used in the norms and the vocabulary of our smallest corpus was 1093 words. We used only this restricted overlap in our experiment.

In order to evaluate the performance of LSA models in reproducing these human generated data, we used the same measure as in Griffiths et al. (2007): the median rank of the first associates of a word in the semantic space. This was done in three steps : 1) for each word cue W_c, we sorted the list of the remaining words W_i in the overlap set, based on their LSA cosine similarity with that cue: $cos(LSA(W_c), LSA(W_i))$, with highest cosine ranked first. 2) We found the ranks of the first three associates for that cue in that list. 3) We applied 1) and 2) to all words in the overlap set and we computed the median rank for each of the first three associates.

Griffiths et al. (2007) tested a set of semantic dimensions going from 100 to 700. We extended the range of dimensions by testing the following set : [2,5,10,20,30,40,50,100, 200, 300,400,500,600,700,800,1000]. We also manipulated the number of successive sentences to be taken as defining the context of a given word (document size), which we varied from 1 to 100.

In Figure 1 we show the results for the 4 M size corpus with 10 sentences long documents.

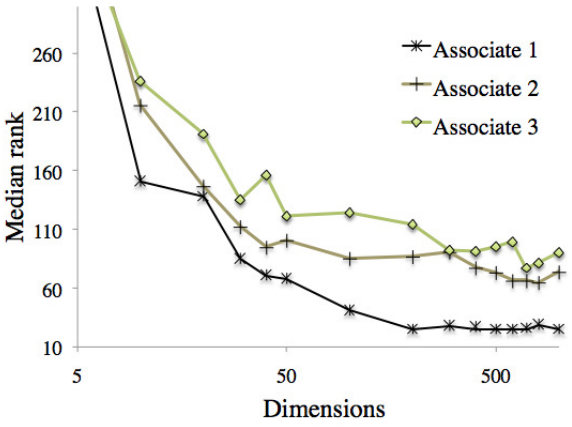

Figure 1 : The median rank of the three associates as a function of the semantic dimensions (lower is better)

For the smaller corpora we found similar results as we can see from Table 1 where the scores represent the median rank averaged over the set of dimensions ranging from 10 to 1000. As found in Griffiths et al. (2007), the median rank measure predicts the order of the first three associates in the norms.

In the rest of the article, we will need to characterize the semantic model by a single value. Instead of taking the median rank of only one of the

Size	associate 1	associate 2	associate 3
1 M	78.21	152.18	169.07
2 M	57.38	114.57	131
4 M	54.57	96.5	121.57

Table 1 : The median rank of the first three associates for different sizes

associates, we will consider a more reliable measure by averaging over the median ranks of the three associates across the overlap set. We will call this measure the Median Rank.

2.2 The Pseudo-synonym detection task

The measure we introduce in this part is based on a Same-Different Task (SDT). It is described schematically in Figure 2, and is computed as follows: for each corpus, we generate a Pseudo-Synonym-corpus (PS-corpus) where each word in the overlap set is randomly replaced by one of two lexical variants. For example, the word "Art" is replaced in the PS-corpus by "Art_1" or "Art_2". In the derived corpus, therefore, the overlap lexicon is twice as big, because each word is duplicated and each variant appears roughly with half of the frequency of the original word.

The Same-Different Task is set up as follows: a pair of words is selected at random in the derived corpus, and the task is to decide whether they are variants of one another or not, only based on their cosine distances. Using standard signal detection techniques, it is possible to use the distribution of cosine distances across the entire list of word pairs in the overlap set to compute a Receiver Operating Characteristic Curve (Fawcett, 2006), from which one derives the area under the curve. We will call this measure : SDT-ρ. It can be interpreted as the probability that given two pairs of words, of which only one is a pseudo-synonym pair, the pairs are correctly identified based on cosine distance only. A value of 0.5 represents pure chance and a value of 1 represents perfect performance.

It is worth mentioning that the idea of generating pseudo-synonyms could be seen as the opposite of the "pseudo-word" task used in evaluating word sense disambiguation models (see for instance Gale et al. (1992) and Dagan et al. (1997)). In this task, two different words w_1 and w_2 are combined to form one ambiguous pseudo-word $W_{12} = \{w_1, w_2\}$ which replaces both w_1 and w_2 in the test set.

We now have two measures evaluating the quality of a given semantic representation: The Median Rank (behavior-based) and the SDT-ρ (corpus-based). Can we use the latter to predict the former? To answer this question, we compared the performance of both measures across different semantic models, document lengths and corpus sizes.

3 Results

In Figure 3 (left), we show the results of the behavior-based Median Rank measure, obtained from the three corpora across a number of semantic dimensions. The best results are obtained with a few hundred dimensions. It is important to highlight the fact that small differences between high dimensional models do not necessarily reflect a difference in the quality of the semantic representation. In this regard, Landauer and Dumais (1997) argued that very small changes in computed cosines can in some cases alter the LSA ordering of the words and hence affect the performance score. Therefore only big differences in the Median Ranks could be explained as a real difference in the overall quality of the models. The global trend we obtained is consistent with the results in Griffiths et al. (2007) and with the findings in Landauer and Dumais (1997) where maximum performance for a different task (TOEFL synonym test) was obtained over a broad region around 300 dimensions.

Besides the effect of dimensionality, Figure 3 (left) indicates that performance gets better as we increase the corpus size.

In Figure 3 (right) we show the corresponding results for the corpus-based SDT-ρ measure. We can see that SDT-ρ shows a parallel set of results and correctly predicts both the effect of dimensionality and the effect of corpus size. Indeed, the general trend is quite similar to the one described with the Median Rank in that the best performance is obtained for a few hundred dimensions and the three curves show a better score for large corpora.

Figure 4 shows the effect of document length on the Median Rank and SDT-ρ. For both measures, we computed these scores and averaged them over the three corpora and the range of dimensions going from 100 to 1000. As we can see, SDT-ρ predicts the psychological optimal document length,

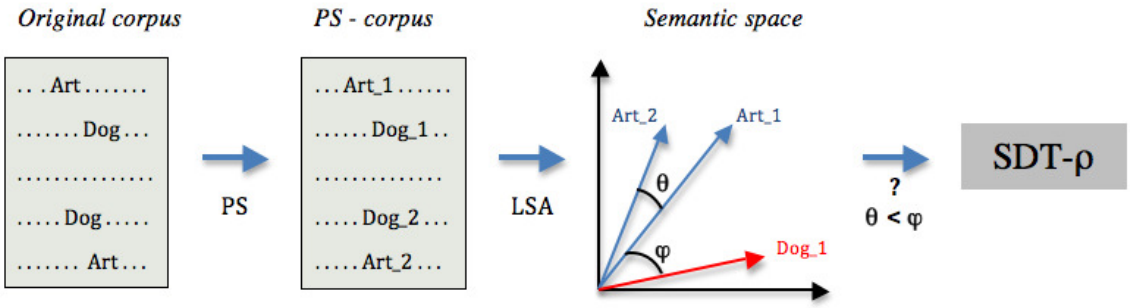

Figure 2 : Schematic description of the Same-Different Task used.

which is about 10 sentences per document. In the corpus we used, this gives on average of about 170 words/document. This value confirms the intuition of Landauer and Dumais (1997) who used a paragraph of about 150 word/document in their model.

Finally, Figure 5 (left) summarizes the entire set of results. It shows the overall correlation between SDT-ρ and the Median Rank. One point in the graph corresponds to a particular choice of semantic dimension, document length and corpus size. To measure the correlation, we use the Maximal Information Coefficient (MIC) recently introduced by Reshef et al. (2011). This measure captures a wide range of dependencies between two variables both functional and not. For functional and non-linear associations it gives a score that roughly equals the coefficient of determination (R^2) of the data relative to the regression function. For our data this correlation measure yields a score of $MIC = 0.677$ with ($p < 10^{-6}$).

In order to see how the SDT-ρ measure would correlate with another human-generated benchmark, we ran an additional experiment using the TOEFL synonym test (Landauer and Dumais, 1997) as gold standard. It contains a list of 80 questions consisting of a probe word and four answers (only one of which is defined as the correct synonym). We tested the effect of semantic dimensionality on a 6 M word sized Wikipedia corpus where documents contained respectively 2, 10 and 100 sentences for each series of runs. We kept only the questions for which the probes and the 4 answers all appeared in the corpus vocabulary. This left us with a set of 43 questions. We computed the response of the model on a probe word by selecting the answer word with which it had the smallest cosine angle. The best performance (65.1% correct) was obtained with 600 dimensions. This is similar to the result reported in Landauer and Dumais (1997) where the best performance obtained was 64.4% (compared to 64.5% produced by non-native English speakers applying to US colleges). The correlation with SDT-ρ is shown in Figure 5 (right). Here again, our corpus-based measure predicts the general trend of the behavior-based measure: higher values of SDT-ρ correspond to higher percentage of correct answers. The correlation yields a score of $MIC = 0.675$ with ($p < 10^{-6}$).

In both experiments, we used the overlap set of the gold standard with the Wikicorpus to compute the SDT-ρ measure. However, as the main idea is to apply this evaluation method to corpora for which there is no available human-generated gold standards, we computed new correlations using a SDT-ρ measure computed, this time, over a set of randomly selected words. For this purpose we used the 4M corpus with 10 sentences long documents and we varied the semantic dimensions. We used the Median Rank computed with the Free association norms as a behavior-based measure.

We tested both the effect of frequency and size: we varied the set size from 100 to 1000 words which we randomly selected from three frequency ranges : higher than 400, between 40 and 400 and between 40 and 1. We chose the limit of 400 so that we can have at least 1000 words in the first range. On the other hand, we did not consider words which occur only once because the SDT-ρ requires at least two instances of a word to generate a pseudo-synonym.

The correlation scores are shown in Table 2. Based on the MIC correlation measure, mid-

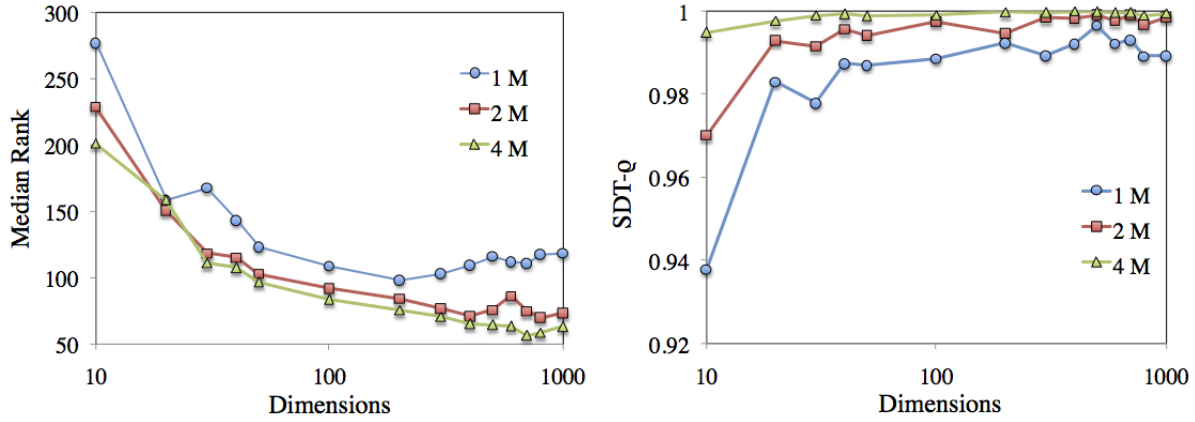

Figure 3 : The Median rank (left) and SDT-ρ (right) as a function of a number of dimensions and corpus sizes. Document size is 10 sentences.

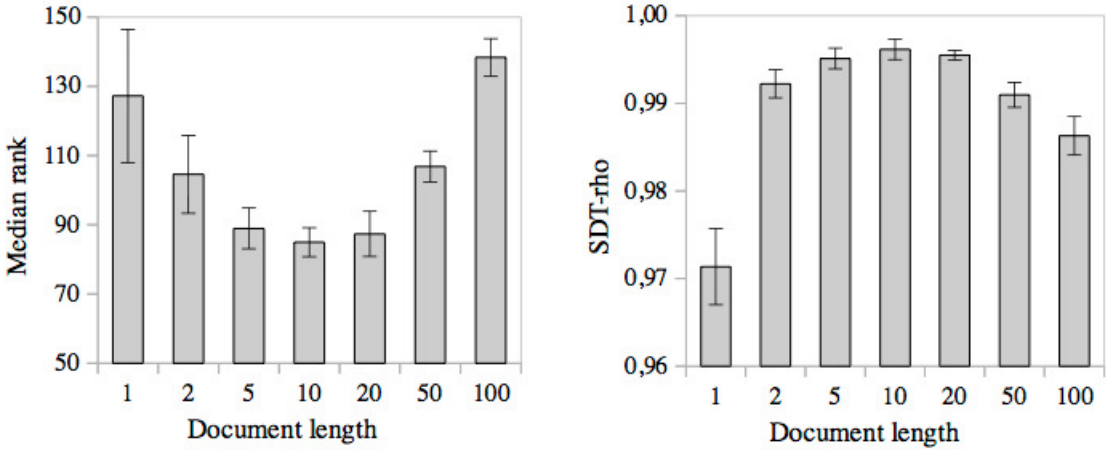

Figure 4 : The Median rank (left) and SDT-ρ (right) as a function of document length (number of sentences). Both measures are averaged over the three corpora and over the range of dimensions going from 100 to 1000.

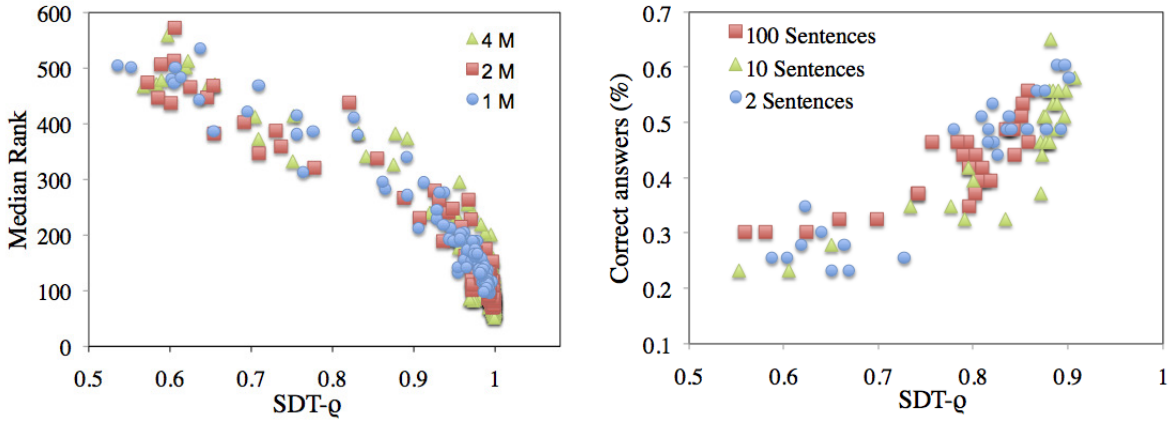

Figure 5 : Overall correlation between Median Rank and SDT-ρ (left) and between Correct answers in TOEFL synonym test and SDT-ρ (right) for all the runs. .

Freq. x	$1 < x < 40$			$40 < x < 400$			$x > 400$			**All**	**Overlap**
Size	100	500	1000	100	500	1000	100	500	1000	~ 4 M	1093
MIC	0.311	0.219	0.549*	0.549*	**0.717***	**0.717***	0.311	0.205	0.419	0.549*	**0.717***

* : $p < 0.05$

Table 2 : Correlation scores of the Median Rank with the SDT-ρ measure computed over randomly selected words from the corpus, the whole lexicon and the overlap with the free association norms. We test the effect of frequency and set size.

frequency words yield better scores. The correlations are as high as the one computed with the overlap even with a half size set (500 words). The overlap is itself mostly composed of mid-frequency words, but we made sure that the random test sets have no more than 10% of their words in the overlap. Mid-frequency words are known to be the best predictors of the conceptual content of a corpus, very common and very rare terms have a weaker discriminating or "resolving" power (Luhn, 1958).

4 Discussion

We found that SDT-ρ enables to predict the outcome of behavior-based evaluation methods with reasonable accuracy across a range of parameters of a LSA model. It could therefore be used as a proxy when human-generated data are not available. When faced with a new corpus and a task involving similarity between words, one could implement this rather straightforward method in order, for instance, to set the semantic model parameters.

The method could also be used to compare the performance of different distributional semantic models, because it does not depend on a particular format for semantic representation. All that is required is the existence of a semantic similarity measure between pairs of words. However, further work is needed to evaluate the robustness of this measure in models other than LSA.

It is important to keep in mind that the correlation of our measure with the behavior-based methods only indicates that SDT-ρ can be trusted, to some extent, in evaluating these semantic tasks. It does not necessarily validate its ability to assess the entire semantic structure of a distributional model. Indeed, the behavior-based methods are dependent on particular tasks (i.e., generating associates, or responding to a multiple choice synonym test) hence they represent only an indirect evaluation of a model, viewed through these specific tasks.

It is worth mentioning that Baroni and Lenci (2011) introduced a comprehensive technique that tries to assess simultaneously a variety of semantic relations like meronymy, hypernymy and coordination. Our measure does not enable us to assess these relations, but it could provide a valuable tool to explore other fine-grained features of the semantic structure. Indeed, while we introduced SDT-ρ as a global measure over a set of test words, it can also be computed word by word. Indeed, we can compute how well a given semantic model can detect that "Art_1" and "Art_2" are the same word, by comparing their semantic distance to that of random pairs of words. Such a word-specific measure could assess the semantic stability of different parts of the lexicon such as concrete vs. abstract word categories, or the distribution properties of different linguistic categories (verb, adjectives, ..). Future work is needed to assess the extent to which the SDT-ρ measure and its word-level variant provide a general framework for DSMs evaluation without external resources.

Finally, one concern that could be raised by our method is the fact that splitting words may affect the semantic structure of the model we want to assess because it may alter the lexical distribution in the corpus, resulting in unnaturally sparse statistics. There is in fact evidence that corpus attributes can have a big effect on the extracted model (Sridharan and Murphy, 2012; Lindsey et al., 2007). However, as shown by the high correlation scores, the introduced pseudo-synonyms do not seem to have a dramatic effect on the model, at least as far as the derived SDT-ρ measure and its predictive power is concerned. Moreover, we showed that in order to apply the method, we do not need to use the whole lexicon, on the contrary, a small test set of about 500 random mid-frequency words (which represents less than 2.5 % of the total vocabulary) was shown to lead to better results. However, even if the results are not directly affected in our case, future work needs to investigate the exact effect word splitting may have on the semantic model.

References

Andrews, M., G. Vigliocco, and D. Vinson (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review 116*, 463–498.

Baroni, M. and A. Lenci (2011). How we BLESSed distributional semantic evaluation. In *Proceedings of the EMNLP 2011 Geometrical Models for Natural Language Semantics (GEMS 2011) Workshop*, East Stroudsburg PA: ACL, pp. 1–10.

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM 55*(4), 77–84.

Dagan, I., L. Lee, and F. Pereira (1997). Similarity-based methods for word sense disambiguation. In *Proceedings of the 35th ACL/8th EACL*, pp. 56–63.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters 27*(8), 861–874.

Gale, W., K. Church, and D. Yarowsky (1992). Work on statistical methods for word sense disambiguation. *Workings notes, AAAI Fall Symposium Series, Probabilistic Approaches to Natural Language*, 54–60.

Griffiths, T., M. Steyvers, and J. Tenenbaum (2007). Topics in semantic representation. *Psychological Review 114*, 114–244.

Harris, Z. (1954). Distributional structure. *Word 10*(23), 146–162.

Hutchison, K., D. Balota, M. Cortese, and J. Watson (2008). Predicting semantic priming at the item level. *Quarterly Journal of Experimental Psychology 61*(7), 1036–1066.

Kievit-Kylar, B. and M. N. Jones (2012). Visualizing multiple word similarity measures. *Behavior Research Methods 44*(3), 656–674.

Landauer, T. and S. Dumais (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*(2), 211–240.

Lewandowsky, S. and S. Farrell (2011). *Computational modeling in cognition : principles and practice*. Thousand Oaks, Calif. : Sage Publications.

Lindsey, R., V. Veksler, and A. G. andWayne Gray (2007). Be wary of what your computer reads: The effects of corpus selection on measuring semantic relatedness. In *Proceedings of the Eighth International Conference on Cognitive Modeling*, pp. 279–284.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development 2*(2), 157–165.

Nelson, D., C. McEvoy, and T. Schreiber (1998). The university of south florida word association, rhyme, and word fragment norms.

Reese, S., G. Boleda, M. Cuadros, L. Padro, and G. Rigau (2010). Wikicorpus: A word-sense disambiguated multilingual wikipedia corpus. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC'10)*, La Valleta, Malta.

Řehůřek, R. and P. Sojka (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, pp. 45–50.

Reshef, D., Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti (2011). Detecting novel associations in large datasets. *Science 334*(6062), 1518–1524.

Sridharan, S. and B. Murphy (2012). Modeling word meaning: distributional semantics and the sorpus quality-quantity trade-off. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, COLING 2012, Mumbai, pp. 53–68.

Turney, P. D. and P. Pantel (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research 37*, 141–188.

References

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*.

Bergelson, E., & Swingley, D. (2012). At 6 to 9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9).

Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition*, *127*.

Best, C. T. (1993). Emergence of language-specific constraints in perception of non-native speech: A window on early phonological development. In B. de Boysson-Bardies & S. de Schonen (Ed.), *Developmental neurocognition: Speech and face processing in the first year of life.* Norwell, Kluwer Academic Publishers.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Bloom, P. (2000). *How children learn the meanings of words.* Cambridge, MA. MIT Press.

Bortfeld, J., H. an Morgan, Golinkoff, R., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech stream segmentation. *Psychological Science*, *16*.

Boruta, L. (2011). Combining Indicators of Allophony. In *Proceedings ACL-SRW* (p. 88-93).

Boruta, L. (2012). *Indicateurs d'allophonie et de phonémicité* (Doctoral dissertation). Université Paris-Diderot - Paris VII.

References

Brown, R. (1973). *A first language: The early stages.* Harvard U. Press.

Bybee, J. L. (2001). *Phonology and language use.* Cambridge, Cambridge University Press.

Caselli, M., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Weir, J. (1995). A cross-linguistic study of early lexical development. *Cognitive Development*, *10*(2), 159–199. doi: 10.1016/0885-2014(95)90008-x

Chater, N., & Manning, C. (2006). Probabilistic models of language processing and acquisition. *Trends in cognitive sciences*, *10*.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MIT Press.

Chomsky, N., & Halle, M. (1968). *The sound pattern of english.* New York, Harper and Row.

Cohn, A. C. (2006). Is there gradient phonology? In R. V. G. Faneslow C. Fery & M. Schlesewsky (Eds.), *Gradience in grammar: Generative perspectives.* Oxford, OUP.

Cristia, A., & Seidl, A. (2014). The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language*, *4*.

Curtin, S., Fennell, C., & Escudero, P. (2009). Weighting of acoustic cues explains patterns of word-object associative learning. *Developmental Science*, *12*.

Curtin, S., Mintz, T. H., & 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.*

Dale, P. S. (1996). *Parent report assessment of language and communication.* Paul H Brookes Publishing.

de Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*.

Dietrich, C., Swingley, D., & Werker, J. (2007). Native language governs interpretation of salient speech sound differences at 18 months. *Proceedings of the National Academy of Sciences*, *104*.

References

Dresher, B. E. (2011). The phoneme. In E. H. Marc van Oostendorp Colin J. Ewen & K. Rice (Eds.), *The blackwell companion to phonology.* Oxford, Wiley-Blackwell.

Escudero, P., Mulak, K., & Vlach, H. (2010). Cross-situational learning of minimal word pairs. *Cognitive Science*.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*.

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J. L. (2013). Word-level information influences phonetic learning in adults and infants. *Cognition*, *127*.

Fennell, & Waxman. (2010). What paradox? referential cues allow for infant use of phonetic detail in word learning. *Child Development*, *81*.

Fernald, A., Zangl, R., Portillo, A., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In I. Sekerina, E. Fernandez, & H. Clahsen (Eds.), .

Firth, J. R. (1957). A synopsis of linguistic theory 19301955. In *Studies in linguistic analysis.* Oxford, Blackwell.

Fourtassi, A., Benjamin Borschinger, M. J., & Dupoux, E. (2013). Why is English so easy to segment? In *Proceedings of CMCL.*

Fourtassi, A., & Dupoux, E. (2013). A corpus-based evaluation method for distributional semantic models. In *Proceedings of ACL.*

Fourtassi, A., & Dupoux, E. (2014). A rudimentary lexicon and semantics help bootstrap phoneme acquisition. In *Proceedings of CoNLL.*

Frank, S., Feldman, N., & Goldwater, S. (2014). Weak semantic context helps phonetic learning in a model of infant language acquisition. In *Proceedings of the 52nd annual meeting of the association of computational linguistics.*

Fulkerson, A. L., & Waxman, S. R. (2007). Words (but not tones) facilitate object catego-

rization: Evidence from 6- and 12-month-olds. *Cognition*, *150*.

Gervain, J., & Mehler, J. (2010). Speech perception and language acquisition in the first year of life. *Annual Review of Psychology*, *61*, 191-218.

Gimson, A. C. (1962). *An introduction to the pronunciation of english.* London, Edward Arnold.

Gleitman, L., Cassidy, K., Papafragou, A., Nappa, R., & Trueswell, J. (2005). Hard word. *Journal of Language Learning and Development*, *1*.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2011). Producing power-law distributions and damping word frequencies with two-stage language models. *Journal of Machine Learning Research*, *12*, 2335-2382.

Gómez, R., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*, 109–135.

Gómez, R., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, *7*.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211-244.

Hall, G. F. (1991). *Perceptual and associative learning.* Oxford, UK: Clarendon Press.

Harris, Z. (1954). Distributional structure. *Word*, *10*(23), 146-162.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of american english vowels. *Journal of the Acoustical Society of America*, *97*.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*.

Huijbregts, M., McLaren, M., & van Leeuwen, D. (2011). Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection. In *Acoustics, speech and signal processing (ICASSP).*

References

Jakobson, R. (1966). Beitrag zur allgemeinen. In J. Bybee & P. Hopper (Eds.), *Readings in linguistics ii.* Chicago, University of Chicago Press.

Jansen, A., & Church, K. (2011). Towards unsupervised training of speaker independent acoustic models. In *Proceedings of INTERSPEECH.*

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In J. . Mullennix (Ed.), *Talker variability in speech processing.* San Diego, Academic Press. pp. 145-165.

Johnson, M., Griffiths, T., & Goldwater, S. (2007). Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Human language technologies 2007: The conference of the north american chapter of the association for computational linguistics; proceedings of the main conference* (p. 139-146). Rochester, New York: Association for Computational Linguistics.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants detection of the sound patterns of words in fluent speech. *Cognitive psychology*, *29*(1), 1-23.

Kazanina, N., Phillips, C., & Idsardi, W. (2006). The influence of meaning on the perception of speech sounds. *Proceedings of the National Academy of Sciences*, *103*.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, *5*.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*.

Labov, W. (1991). The three dialects of english. In P. Eckert (Ed.), *New ways of analyzing sound change.* New York, Academic Press.

Ladefoged, P. (2001). *Vowels and consonants: an introduction to the sounds of languages.* Maldon, Mass. & Oxford, Blackwell Publishers.

Landau, B., Smith, L., & Jones, S. (1988). The importance of shape in early lexical learning.

Cognitive Development, *3*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*(2), 211-240.

Lawrence, D. (1949). Acquired distinctiveness of cues: I. transfer between discriminations on the basis of familiarity with the stimulus. *Journal of Experimental Psychology*, *39*(6), 770784.

Lee, C., & Glass, J. (2012). A nonparametric bayesian approach to acoustic model discovery. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers-volume 1* (p. 40-49).

Lisker, L., & Abramson, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the 6th International Congress of Phonetic Sciences.*

MacKain, K. S. (1982). Assessing the role of experience on infants speech discrimination. *Journal of Child Language*, *9*.

Maekawa, K., Koiso, H., Furui, S., & Isahara, H. (2000). Spontaneous speech corpus of japanese. In *LREC* (pp. 947–952). Athens, Greece.

Mandel, D., Jusczyk, P. W., & Pisoni, D. B. (1995). Infants' recognition of the sound patterns of their own names. *Psychological Science*, *5*(6), 314–317.

Markman, E. M. (1991). The whole-object, taxonomic, and mutual exclusivity assumptions as initial constraints on word meanings. In S. A. Gelman & J. P. Byrnes (Eds.), .

Martin, A., Peperkamp, S., & Dupoux, E. (2013). earning phonemes with a proto-lexicon. *Cognitive Science*, *37*.

Maye, J., Werker, J., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*.

McMurray, B., Aslin, R., & Toscano, J. (2009). Statistical learning of phonetic categories:

insights from a computational approach. *Developmental science*, *12*(3), 369.

McMurray, B., & Aslin, R. N. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition*, *95*.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*.

McMurray, B., Tanenhaus, M., & Aslin, R. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*.

Medina, T., Snedeker, J., Trueswell, J., & Gleitman, L. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014.

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of INTERSPEECH.*

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of japanese and english. *Perception & Psychophysics*, *18*.

Narayan, C., Werker, J., & Beddor, P. (2010). The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental Science*, *13*.

Nazzi, T., & Bertoncini, J. (2003). Before and after the vocabulary spurt: Two modes of word acquisition? *Developmental Science*, *6*.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The university of south florida word association, rhyme, and word fragment norms.* [http://www.usf.edu/FreeAssociation].

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., & Peperkamp, S. (2013). Nonwords, nonwords, nonwords: Evidence for a proto-lexicon during the first year of life. *Devel-*

opmental Science, *16*.

Okada, H. (1999). Japanese. In *Handbook of the international phonetic association: A guide to the usage of the international phonetic alphabet.* Cambridge, Cambridge University Press.

Ouyang, L., Boroditsky, L., & Frank, M. C. (in press). Semantic coherence facilitates distributional learning of word meaning. *Cognitive Science*.

Papafragou, A., Cassidy, K., & Gleitman, L. (2005). When we think about thinking: The acquisition of belief verbs. *Cognition*, *105*.

Parviz, M., Johnson, M., Johnson, B., & Brock, J. (2011). Using language models and latent semantic analysis to characterise the n400m neural response. In *Proceedings of the Australasian Language Technology Association Workshop.*

Peperkamp, S., Le Calvez, R., Nadal, J., & Dupoux, E. (2006). The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, *101*.

Peters, A. (1983). *The units of language acquisition.* Cambridge, Cambridge University Press.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of lexical structure.* Amsterdam, John Benjamins.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, *46*.

Pike, K. L. (1947). *Phonemics: a technique for reducing languages to writing.* Ann Arbor, University of Michigan Press.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure.* Cambridge, MA: MIT press.

Pitman, J., & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, *25*, 855-900.

References

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier. (2007). *Buckeye corpus of conversational speech* (Second ed.). [www.buckeyecorpus.osu.edu].

Quine, W. (1960). *Word and object*. The MIT Press.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*.

Roy, B. C., Frank, M. C., DeCamp, M. M., P., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*.

Saffran, J., & Thiessen, E. (2003). Pattern induction by infant language learners. *Developmental Psychology*, *39*.

Saffran, J. R., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926.

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multi-level statistical learning by 12-month-old infants. *Infancy*, *4*.

Sapir, E. (1925). Sound patterns in language. *Language*, *1*.

Sapir, E. (1933). La réalité psychologique des phonèmes. *Journal de Psychologie Normale et Pathologique*, *30*.

Seidl, A., & Cristia, A. (2012). Infants' learning of phonological status. *Frontiers in Psychology*, *3*.

Smith, K. (1966). Grammatical intrusions in the recall of structured letter pairs: mediated transfer or position learning? *Journal of Experimental Psychology*, *72*, 580–588.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Socher, R., Ganjoo, M., Manning, C. D., & Ng, A. Y. (2013). Zero-Shot Learning Through Cross-Modal Transfer. In *Proceedings of Conference on Neural Information Processing Systems (NIPS).*

References

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640).

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *Journal of Experimental Child Psychology*, *126*.

Sundara, M., Polka, L., & Genesee, F. (2006). Language-experience facilitates discrimination of /d-d/ in monolingual and bilingual acquisition of english. *Cognition*, *100*.

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B*, *364*.

Swingley, D., & Aslin, R. (2007). Lexical competition in young children's word learning. *Cognitive Psychology*, *54*.

Thiessen. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, *56*.

Thiessen, E., & Saffran, J. (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development*, *3*(1), 73–100.

Thiessen, E. D., & Yee, M. N. (2010). Dogs, bogs, labs, and lads: What phonemic generalizations indicate about the nature of children's early word-form representations. *Child Development*, *81*.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*.

Tomasello, M. (2001). Perceiving intentions and learning words in the second year of life. In *Language acquisition and conceptual development* (pp. 132–159). New York: Cambridge University Press.

Trubetzkoy, N. S. (1939). *Gründzuge der phonologie*. Göttingen, van der Hoeck and Ruprecht.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational learning. *Cognitive Psychology*, *66*.

References

Twaddell, W. F. (1935). *On defining the phoneme*. Baltimore, Waverly Press.

Vallabha, G., McClelland, J., Pons, F., Werker, J., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*(33), 13273.

Varadarajan, B., Khudanpur, S., & Dupoux, E. (2008). Unsupervised learning of acoustic sub-word units. In *Proceedings of the association for computational linguistics.*

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition*, *127*.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*(2), 729–742.

Vouloumanos, A., & Waxman, S. R. (2014). Listen up! speech is for thinking during infancy. *Trends in cognitive sciences*, *18*(12), 642–646.

Waxman, S., & Markow, D. (1995). Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302.

Werker, Cohen, Lloyd, Casasola, & Stager. (1998). Acquisition of word-object associations by 14-month-old infants. *Developmental Psychology*, *34*(6).

Werker, J., & Curtin, S. (2005). Primir: A developmental framework of infant speech processing. *Language learning and development*, *1*.

Werker, J., Fennell, C.T., Corcoran, K., & Stager, C. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy*, *3*.

Werker, J., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in english and japanese. *Cognition*, *103*.

Werker, J., & Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*.

White, K., & Morgan, J. (2008). Sub-segmental detail in early lexical representations.

Journal of Memory and Language, *59*.

White, K., Peperkamp, S., Kirk, C., & Morgan, J. (2008). Rapid acquisition of phonological alternations by infants. *Cognition*, *107*.

Yeung, H., & Werker, J. (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition*, *113*, 234-243.

Yoshida, K., Fennell, C., Swingley, D., & Werker, J. (2009). 14-month-olds learn similar-sounding words. *Developmental Science*, *12*.

Young, S. J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (2006). *The HTK Book Version 3.4*. Cambridge University Press.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.

Yurovsky, D., & Frank, M. C. (2015). An Integrative Account of Constraints on Cross-Situational Learning. *Cognition*, *145*.

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, *37*.