Phoneme learning is influenced by the taxonomic organization of the semantic referents

Abdellah Fourtassi (afourtas@stanford.edu)
Department of Psychology, Stanford University, USA

Emmanuel Dupoux (emmanuel.dupoux@gmail.com)
ENS/CNRS/EHESS/INRIA/PSL Research University, France

Abstract

Word learning relies on the ability to master the sound contrasts that are phonemic (i.e., signal meaning difference) in a given language. Though the timeline of phoneme development has been studied extensively over the past few decades, the mechanism of this development is poorly understood. Previous work has shown that human learners rely on referential information to differentiate similar sounds, but largely ignored the problem of taxonomic ambiguity at the semantic level (two different objects may be described by one or two words depending on how abstract the meaning intended by the speaker is). In this study, we varied the taxonomic distance of pairs of objects and tested how adult learners judged the phonemic status of the sound contrast associated with each of these pairs. We found that judgments were sensitive to gradients in the taxonomic structure, suggesting that learners use probabilistic information at the semantic level to optimize the accuracy of their judgements at the phonological level. The findings provide evidence for an interaction between phonological learning and meaning generalization, raising important questions about how these two important processes of language acquisition are related.

Keywords: language acquisition; phonological development; word learning; speech perception.

A crucial part of language acquisition is the mastery of the sound inventory, i.e., the set of atomic sounds of which words are made. The sound inventory is language-specific. English speakers, for instance, have to learn the distinction between the sounds /l/ and /r/ to differentiate minimal pairs such as *glass* and *grass*. In contrast, Japanese learners need not differentiate these sounds, which do not bring about difference in word meaning in their language. Crucially, even within the same language, learners have to distinguish the sounds that contrast word meaning (phonemic contrasts) from the sounds that do not (non-phonemic contrasts). For example, the aspirated and unaspirated versions of /p/ (which occur, respectively, in the first segment of the word *pin*, and the second segment of the word *spin*) belong to the same phonemic category. Another example, is the *cot-caught* merger whereby the vowels [ɑ]-[ɔ] have come to be treated by some English speakers as non-phonemic variations of the same sounds (Labov, 1991).

How do people learn when a sound contrast is phonemic and when it is a phonetic variation of the same sound category? Children start to show sensitivity to their native sounds at a very early age (e.g., Werker & Tees, 1984). Throughout development, they also learn to distinguish the subset of the native sounds that cue meaning (Dietrich, Swingley, & Werker, 2007; Seidl, Cristi, Onishi, & Bernard, 2009; Kazanina, Phillips, & Idsardi, 2006). These developmental facts have been documented in detail over the past few decades, but the mechanism of this learning is still poorly understood.

Most research has focused on exploring mechanisms which operate on the speech signal without any referential input (Peperkamp, Le Calvez, Nadal, & Dupoux, 2006; Maye, Werker, & Gerken, 2002; Vallabha, McClelland, Pons, Werker, & Amano, 2007; Swingley, 2009; Martin, Peperkamp, & Dupoux, 2013; Feldman, Myers, White, Griffiths, & Morgan, 2013; Dillon, Dunbar, & Idsardi, 2013). These mechanisms have been tested successfully with simplified input. However, they were not as successful when tested on more realistic acoustic data which is highly variable and noisy (e.g., Varadarajan, Khudanpur, & Dupoux, 2008; Fourtassi, Schatz, Varadarajan, & Dupoux, 2014; Jansen et al., 2013). Thus, though these mechanisms may play an important role, they are unlikely to account for the entire process of learning and refinement.

Learners are exposed to more than the speech signal. In particular, they usually have access to multimodal input which co-occur with speech. For example, the words *glass* and *grass* in English are typically associated with different visual input. Experimental data has shown that both children and adults can leverage such semantic/visual information to discriminate ambiguous sounds (Teinonen, Aslin, Alku, & Csibra, 2008; Yeung & Werker, 2009; Hayes-Harb, 2007).

Nevertheless, previous research has generally assumed–whether implicitly or explicitly–that learners have access, not only to the immediate visual input, but also to the entire meaning category intended by the speaker, e.g., the meaning of the word ‘cow’ is not limited to one specific cow–it includes cows of all shapes and colors, and it excludes instances of another category such as deer. Knowing the meaning's extension and boundary of a given word is crucial to the task of phoneme learning: If an ambiguous sound contrast is associated with two different objects (e.g., a cow and a deer), then in order to decide whether or not this contrast is phonemic, the learner has to determine first if the speakers' target meanings are two specific categories (cow and deer) or one broad category (e.g., animal). The contrast is phonemic in the former case, and non-phonemic in the latter.

Often, however, learners in the early stages of acquiring their first or second language, do not yet know the full meaning extensions of the words they hear around them. Work in the word learning literature suggests that humans spontaneously restrict the set of possible extensions to taxonomic classes (Markman, 1989). For example, upon hearing the word ‘cow’ with an instance of, say, a brown cow, humans are unlikely to consider an extension that includes ‘milk’ or ‘brown rice’. Though the taxonomic assumption simplifies the task, it still leaves a great deal of ambiguity regarding the level of generalization intended by the speaker. For example, the word ‘cow’ could have meant more abstract categories such as “mammal” or “animal”.

The current work aims at studying how learners behave in a situation where there is uncertainty at both the phonological and semantic levels. We associate minimally different non-sense words (along the ambiguous sound contrast ɑ-ɔ) with pairs of objects that vary in their taxonomic proximity (Figure 1). Crucially, mere exposure to instances of the sound-object pairings is not enough to determine the exact meaning extension, leaving the participants in a situation of uncertainty similar to that faced by learners in the early stages of language acquisition. We are interested in the participants' subsequent judgment about the phonemic status of the pair of sounds.

There are several possible scenarios. For instance, participants may not be sensitive to the degree of taxonomic distance, treating all visual differences as equally indicative of a phonemic status. It is also possible that participants treat degrees of taxonomic distance in a categorical way, i.e., they may treat pairs of objects up to a certain taxonomic level as equally indicative of non-phonemicity, whereas they treat pairs of object beyond that level as equally indicative of phonemicity. Finally, participants may be sensitive to each gradient of taxonomic distance in their phonemic learning, in which case their judgements should be graded as well.

In what follows, we test these predictions with adults learning an alien language. In Experiment 1, we parametrize a subset of the semantic space, creating an evenly-spaced taxonomic scale, and we use this scale to explore the effect of different gradients of taxonomic distance on the phonemic status of the ɑ-ɔ contrast. In Experiment 2, we test whether results of Experiment 1 are due to interference from existing lexicalized categories in the first language. Finally, we discuss the implication of the findings on phoneme learning in the context of early language acquisition.

Experiment 1

The goal of this first experiment is to use a parameterized subset of the semantic space to test the effect of each gradient of taxonomic distance on learning the phonemic status of an ambiguous sound contrast. We use a between-subject design to avoid carry-over effects in the sound judgements.

Participants

152 Participants in total were recruited online through Amazon Mechanical Turk, restricting the pool to the United States residents. At the end of the experiment, participants were asked to rate the overall quality of the audio-visual stimuli on their local software/hardware. We excluded participants who judged this quality as medium or bad (N=26), keeping only those who rated the quality as good, that is, those for whom the experiment functioned correctly. We also excluded participants who took the experiment more than once (N=3), and participants who obtained less than 50% correct answers on the obvious filler questions (e.g., are the words ”komi” and ”pibu” different?) (N=5). We ended up with a sample size of 115 participants split across 5 groups.

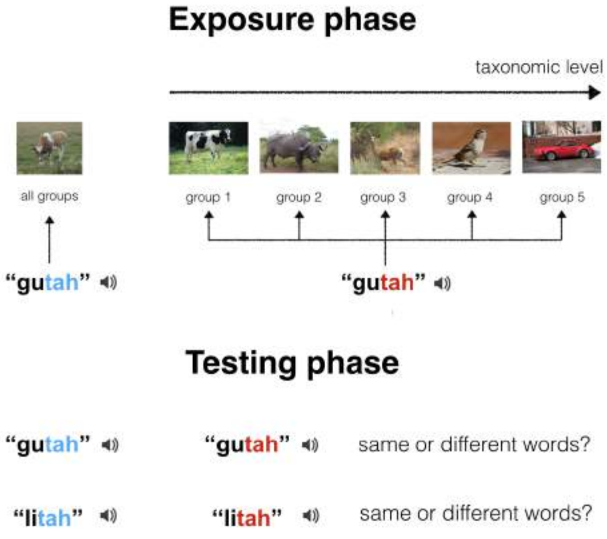

Figure 1: Overview of the task.

Stimuli

Objects The stimuli consist of a reference object (a cow), and five other objects which varied in their similarity to this reference. These objects were, in this order, another cow (with a different color), a buffalo, a deer, a bird and a car (Figure 2). To parmaterize the object stimuli in the taxonomic space, we recruited an additional N=30 participants online (through Amazon Mechanical Turk), and we asked them to rate the similarity of a series of pairs of objects in a 9 point-scale, 1 being “very similar” and 9 being “very different”. The pairs of objects were formed by the pairwise combination of all six items described above. The order of trials was randomized across participants. We computed the average rating for each pair, which gave us a distance matrix.

Figure 2 (left) shows the taxonomic organization of the object stimuli, which we obtained via hierarchical clustering (using average linking) applied to participants' similarity data. Height indicates the average similarity within clusters at each hierarchical/taxonomic level. Figure 2 (right) shows a different visualization of the same data using bi-dimensional scaling. Both representations show that the way objects are organized around the reference (i.e. cow) corresponds to graded differences in the semantic space, and that these gra-

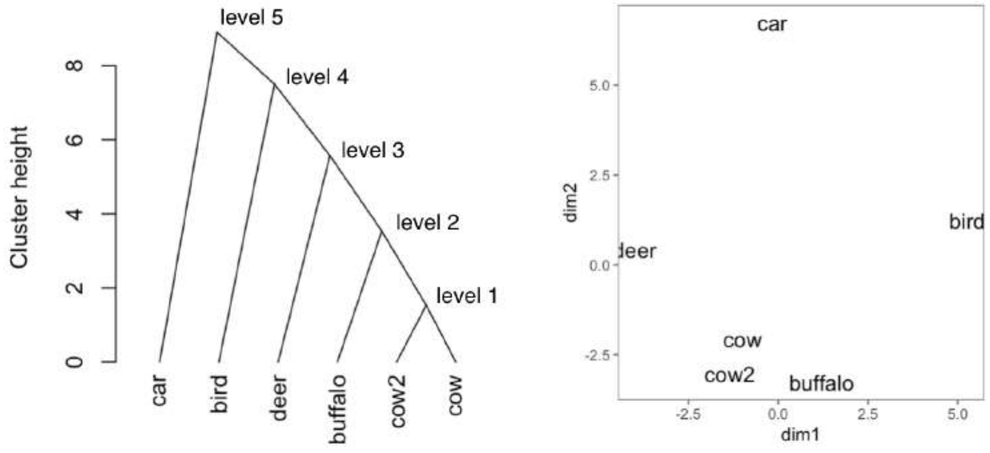

Figure 2: The graph on the left shows the taxonomic organization of the object stimuli obtained via hierarchical clustering of the participants' similarity ratings. The graph on the right shows another way of visualizing the same data via bi-dimensional scaling. Both representations show that the object stimuli lead to a graded and evenly-spaced semantic scale.

dients are quite evenly-spaced.

Sounds We followed Feldman et al. (2013) in using minimal pairs that vary along the vowel contrast [ɑ]-[ɔ]. This contrast is neither too acoustically similar, nor too different. In fact, depending on the dialect, these two vowels can be treated by English speakers as belonging to one or two categories (Labov, 1991). We chose such an ambiguous contrast in order to put the participants in a rather flexible situation where they can switch between phonological interpretations depending on the context. Two minimal pairs were constructed by concatenating two context syllables ([gu] and [li]) produced by a female native speaker of American English, with a target syllable contrast ([tɑ]-[tɔ]) produced by the same speaker. The resulting minimal pairs were [gutɑ]-[gutɔ] and [litɑ]-[litɔ]. For ease of presentation, we will refer to these minimal pairs by *gutah*/*gutaw* and *litah*/*litaw*. In addition, we used two artificial filler words [pibu], [komi] which we obtained by concatenating four vowels produced separately by the same speaker.[1]

Procedure

In order to avoid any carry-over effect on sound judgements, we used a between-subject design, i.e., each group of participants were exposed to only one degree of taxonomic distance. The minimal pair was paired with two objects whose similarity varied across five groups of participants (Figure 1). In all these groups, one member of the minimal pair (e.g., *gutah*) was paired with picture of a cow. The second member (i.e., *gutaw*) was paired with a referent whose similarity with the first referent varied on the five-step taxonomic scale.

The experiment had an exposure and a testing phase. In the exposure phase, participants heard a novel word in an alien language and saw the corresponding object simultaneously. In this phase subjects did not have to perform any specific task, but they were encouraged to listen carefully and try to learn the words. They were exposed to 3 series composed each of a randomized presentation of 4 word-object pairings: 2 target words (*gutah*/*gutaw*) whose referents similarity varied across groups of participants (Figure 2), and 2 filler words (*pibu* and *komi*) mapped invariably to two different objects. There were 12 trials in total, with each presentation lasting around 850 ms (i.e., the time it took the bi-syllabic word audio to complete).

In the testing phase, participants heard a series of trials composed of two word tokens, and were asked to judge if these tokens corresponded to different words in this artificial language, or if they represented a mere phonetic variation of the same word. We used a wording similar to the one used by Feldman et al. (2013)[2]. In this testing phase, subjects were encouraged to follow their intuition and think carefully before answering.

Half of the testing trials contained identical sounds ('same trials'), and the other half contained different sounds ('diff. trials') and were presented in a random order. The diff. trials were composed of the minimal pair used during the exposure phase (*gutah*/*gutaw*) plus a novel minimal pair containing the same syllable contrast (*litah*/*litaw*), and which we used only in the testing phase to investigate the ability of participants to generalize across the lexicon. There were 12 test trials in total: 4 for for the exposure word (2 same and 2 different), 4 for the generalization word (idem), and 4 for the fillers komi/pibu (idem). Participants were tested twice, once before exposure to referential data and once after the exposure.

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

[2] "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. Conversely, they can be same even if they are pronounced slightly differently."

Results

Figure 3 shows the proportion of times participants judged the minimal pairs as phonologically different as a function of group (i.e., taxonomic level), and as a function of the testing session, i.e., before or after exposure to referential data.[3] We show results for both diff. trials (e.g., "gutah"/"gutaw") and same trials ("gutah"/"gutah").

Note, first, that the proportion of 'different' answers for same trials was close to zero across groups, showing that participants were almost perfect in detecting same pairs. However, the proportion of 'different' on the diff. trials varied across groups, and this proportion was 50% in the 'before' session. These initial observations confirm our choice of the sound contrast, which was supposed to be perceptually distinguishable, but ambiguous in terms of its phonemic status, allowing participants to adjust their phonological interpretation depending on the referential context.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.07	0.41	-2.59	0.01
group	0.32	0.12	2.65	0.01
session	-0.86	0.35	-2.46	0.01
trial	-3.59	0.19	-18.55	<0.01
group:session	0.28	0.10	2.85	<0.01

Table 1: A mixed-effects logistic regression predicting participants' responses in a same-different task.

After exposure to referential data, we observed a graded effect of the objects' taxonomic distance on phonological judgment: Participants were more likely to judge the phonologically ambiguous contrast as different when this contrast corresponded to higher taxonomic levels (Figure 3). We fit a mixed-effect logistic regression which predicted the participants' response by the group (i.e., taxonomic distance), the session (before or after exposure the referential data), and the trial (same or diff. pairs). The model was specified as follows: `response ∼ group * session + trial + (1|Subj) + (1|item)`. The estimates are summarized in Table 1. Confirming our qualitative observations, the model shows that the type of the trial predicted the participants' responses (i.e., answering more 'same' on same pairs). Crucially, we also found an interaction between group and session, indicating that exposure to referential data influenced the participants' responses.

To further examine the influence of exposure on learning and generalization, we fit two simple mixed-effects logistic models to the diff. trials after exposure predicting responses as a function of group. We found an effect of object semantic distance on phonological judgment in both the exposure word ($\beta = 4.55$, $SE = 1.13$, $p < 0.01$) and the generalization item ($\beta = 2.58$, $SE = 1.00$, $p < 0.01$).

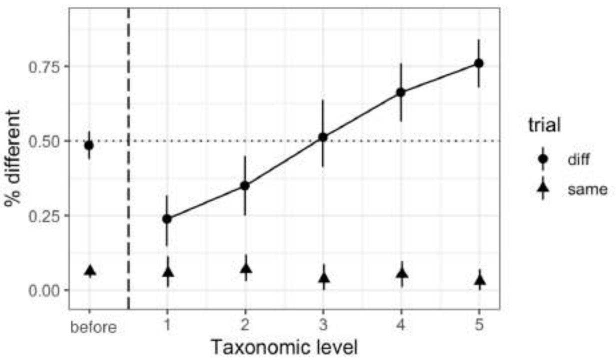

Figure 3: The points are the proportion of times participants judged the pair of sounds as 'different'. The triangles represent the judgments for exactly same pairs (e.g., gutah-gutah). The circles represent the judgments for different pairs (e.g., gutah-gutaw). Data on the left side of the vertical dashed line show the average responses before exposure to the referential data. The horizontal dotted line represents chance. Error bars represent 95% confidence intervals.

Discussion

Experiment 1 tested how gradients in the semantic space influence judgments about the phonemic status. It is possible, however, that participants relied, not on the taxonomic distance in the semantic space, but on available lexicalized concepts in their first language. Indeed, one could imagine that the more a common label is easily accessible for a pair of objects, the more participants answer "same" in the phonemic task. For example, if it is easier to access a common label in the case of *cow* and *deer* (e.g., "mammals"), than it is in the case of *cow* and *car* (e.g., "things"), then this difference may explain why participants judged the sound contrast more as phonemic in the latter. In Experiment 2 we explore whether such an account could explain the findings.

Experiment 2

We asked participants to provide common labels in English for each of the objects pairs used in Experiment 1, and we quantified the difficulty they had in generating these labels. If the phonemic judgements are driven by common labels in the first language, then the difficulty in accessing these labels should mimic closely the phonemic judgments (Figure 3).

Participants

40 participants were recruited online through Amazon Mechanical Turk, restricting the pool to the United States residents.

Stimuli

The same object stimuli used in Experiment 1.

Procedure

Participants were presented with pairs of objects, and were asked to type in, as fast as they could, the most specific la-

[3] Since answers do not vary across groups prior to the exposure phase, in the 'before' session we only show the average results where data were collapsed across groups.

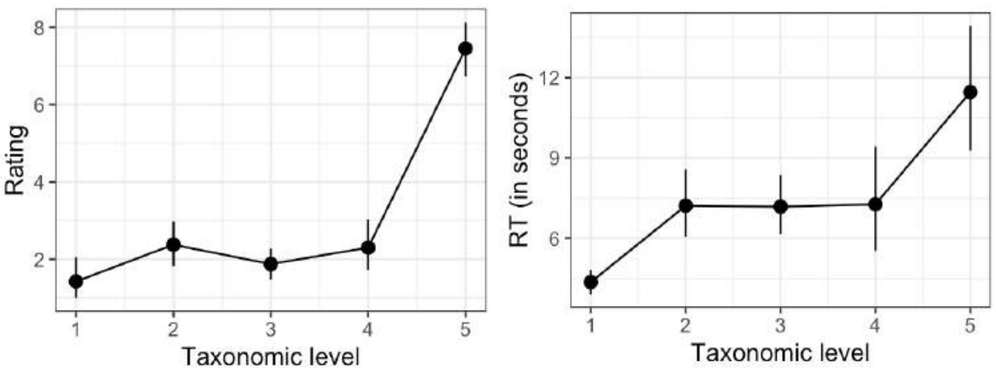

Figure 4: The graph on the left shows the average self-reported difficulty in generating a common English label for pairs of objects at different taxonomic levels (on a scale from 1 to 9). The graph on the right shows the average time it took participants to generate these labels.

bel in English that describes both objects (or "none" in the case they could not find a common label). We obtained both the labels and the reaction times, i.e., the time it took them from seeing the object to confirming their answer. Besides, participants were asked to evaluate the difficulty they had in generating the label on a scale from 1 to 9. The pairs were randomized across participants. To avoid carry-over effects, each participants saw the pairs only once.

Results

Results are shown in Figure 4. Overall, participants were faster and found it easier to generate a common labels when both objects were cows (the most frequent response was 'cow'). They were slower and found it difficult to generate a common labels when the pair was cow/car (most participants did not find a common label and typed 'none'). That said, they did not show any noticeable difference (neither in reaction times nor in subjective evaluation) for the intermediate cases. For all these cases, the most frequent response was 'animal'. Thus, though common labels in the first language may explain limit cases, it does not account for the entire pattern of graded responses obtained in Experiment 1.

General discussion

Previous research has suggested that semantic information can help with phoneme acquisition (Yeung & Werker, 2009; Hayes-Harb, 2007; Werker & Curtin, 2005). Nevertheless, learners often have to learn the phonemic status of the sounds they hear around them before they have determined the exact extension of the meaning intended by the speaker (e.g., a cow and a deer can be described by one or two words depending on the speaker's target level of taxonomy). The current work studied how the process of phoneme learning is influenced by such uncertainty at the semantic level.

More precisely, this study explored the effect of taxonomic distance on phonemic judgments. We associated minimally different word-forms with two semantic referents whose taxonomic distance varied across groups, and we asked participants in each group to judge the phonemic status of the corresponding sound contrast. We found that increasing the taxonomic distance induced graded judgments on the phonemic status, suggesting that learners are sensitive to the taxonomy of the referents when acquiring phonemes.

According to work in the word learning literature, humans have a bias towards extending the meaning of novel words to objects of similar kinds (Markman, 1989; Xu & Tenenbaum, 2007). In our case, this bias may have prompted participants to treat objects that were taxonomically similar (i.e., two cows with different colors, or cow/buffalo) as instances of the same meaning category, thus judging the sound variation as non-phonemic. In contrast, they may have treated objects of different kinds (cow/bird, or cow/car) as instances of different meaning categories, thus judging the corresponding sound variation as phonemic. Besides, the fact that participants provided graded–rather than stepwise–pattern of judgments mirroring the graded taxonomic distance suggests that they make use of probabilistic information at the semantic level to optimize the accuracy of their inference at the phonological level (see also Fourtassi & Frank, 2017).

How could the obtained relationship between taxonomic distance and phonemic judgements inform our understanding of development? First, this relationship may allow learners to collapse non-phonemic but perceivable sounds into the same phonemic category. This is crucial since the majority of sound contrasts in natural input consists of different pronunciations of the same word, rather than words that differ minimally (see Martin et al., 2013). Instances of the same words are likely to be associated with similar semantic information (i.e., at a similar taxonomic level), thus inducing a non-phonemic judgment for the corresponding contrasts.

As for true phonemic contrasts (e.g., *glass* vs. *grass*), sensitivity to the taxonomic structure will favor differentiation to the extent that minimal pairs have distant taxonomic distance in natural languages. Some research suggests that words that are similar phonologically tend to be similar semantically as well (Dautriche, Mahowald, Gibson, & Piantadosi, 2017).

However this research measured semantic similarity using a distributional model which relies on co-occurrence in a large corpus of text. It is possible that the type of semantic relationship that the model derived was thematic, rather than taxonomic. In fact, thematically related words can be taxonomically different (e.g., *cow* and *milk*). It has been shown that the nature of the semantic relationship depends on the model's parameter setting (Lenci, 2018). Further work on the semantic organization of minimal pairs is needed to elucidate this point.

The current study has some limitations. First, we only used the taxonomy of a subset of the conceptual space. To test the generality of the findings, future work will use different scales spanning several conceptual domains. Second, we only used familiar stimuli (real world objects and a native sound contrast). To completely rule out interference from categories in the native language, future work will seek to replicate the findings with non-native contrasts and with novel object stimuli.

To conclude, the current work showed that different degrees of taxonomic distance in the semantic space influence the acquisition of the phonemic status of sound contrasts. The findings show that learners make use of probabilistic information at the semantic level to optimize the accuracy of their phonemic judgments. More generally, this work suggests there to be an interaction between sound learning (phonemic judgment) and word learning (meaning generalization). Further work should aim at characterizing precisely this interaction and exploring its implications for both phonological and semantic development, two aspects of language development which have largely been studied separately.

All data and code for these analyses are available at
`https://github.com/afourtassi/top-down`

Acknowledgements

This work was supported by a post-doctoral grant from the Fyssen Foundation.

References

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. (2017). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, *41*(8).

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*.

Dillon, B., Dunbar, E., & Idsardi, W. (2013). A single-stage approach to learning phonological categories: Insights from inuktitut. *Cognitive science*, *37*(2), 344–377.

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*.

Fourtassi, A., & Frank, M. C. (2017). Word identification under multimodal uncertainty. In *Proceedings of the 39th annual meeting of the Cognitive Science Society.*

Fourtassi, A., Schatz, T., Varadarajan, B., & Dupoux, E. (2014). Exploring the Relative Role of Bottom-up and Top-down Information in Phoneme Learning. In *Proceedings of ACL.*

Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research*, *23*(1).

Jansen, A., Dupoux, E., Goldwater, S., Johnson, M., Khudanpur, S., Church, K., ... others (2013). A summary of the 2012 jhu clsp workshop on zero resource speech technologies and models of early language acquisition. In *2013 ieee international conference on acoustics, speech and signal processing* (pp. 8111–8115).

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*.

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

Lenci, A. (2018). Distributional models of word meaning. *Annual review of Linguistics*, *4*.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. The MIT Press.

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*.

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

Seidl, A., Cristi, A., Onishi, K., & Bernard, A. (2009). Allophonic and phonemic contrasts in infants' learning of sound patterns. *Language Learning and Development*, *5*.

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

Teinonen, T., Aslin, R., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, *108*.

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.*

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

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

Xu, F., & Tenenbaum, J. (2007). Word Learning as Bayesian Inference. *Psychological Review*, *114*, 245.

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.