Self-Consistency as an Inductive Bias in Early Language Acquisition

Abdellah Fourtassi (abdellah.fourtassi@gmail.com)
Ewan Dunbar (emd@umd.edu)
Emmanuel Dupoux (emmanuel.dupoux@gmail.com)
Laboratoire de Sciences Cognitives et Psycholinguistique, ENS/EHESS/CNRS
Paris 75005, France

Abstract

In this paper we introduce an inductive bias for language acquisition under a view where learning of the various levels of linguistic structure takes place interactively. The bias encourages the learner to choose sound systems that lead to more "semantically coherent" lexicons. We quantify this coherence using an intrinsic and unsupervised measure of predictiveness called "self-consistency." We found self-consistency to be optimal under the true phonemic inventory and the correct word segmentation in English and Japanese.

Keywords: Language acquisition, inductive bias, phonemes, word segmentation, semantics.

Introduction

In learning their native language, infants need to make sense of the sounds they are hearing. For the segmental inventory, they need to decide how much of the detail present in the signal matters, and how much of the detail they should ignore. The inventories that human lexicons make use of are somewhere in between maximally coarse and maximally fine-grained. For word segmentation, learners need to decide what to take as a lexical unit of speech: this could in principle be anywhere from a single segment up to an entire utterance, but, in reality, the result is somewhere in between.

Whether learning is seen from a nativist or empiricist perspective, it cannot happen without some kind of learning bias (whether domain specific or domain general) which delimits the hypothesis space, however broadly, and favors one representation over another, however weakly (see Pearl and Goldwater (in press) for a review).

In this paper we propose a novel learning bias and show that it aids in picking out the right level of granularity for both the segmental inventory and lexical segmentation. It makes use of the synergy between different levels of representation (inventory, lexicon, semantics). It takes a systemic approach to language acquisition, whereby infants are understood as trying to build and optimize a coherent system with compatible levels of representation.

Recent developmental studies have indeed begun to suggest that infants start learning both the sound system and the lexicon of their native language at the same time, around 6 months (see Gervain and Mehler (2010) for a review). This paper proposes that these two levels crucially interact in learning.

The bias towards global coherence is coded by a measure we call the *self-consistency* score (SC-score). It is used to evaluate a phonetic inventory and a word segmentation, as a function of the predictiveness of the lexicon they induce. The lexicon should be one in which words are highly predictive of other (neighboring) words. This can be seen as guiding the learner towards a more "semantically coherent" lexicon. We show, using English and Japanese corpora, that the SC-score picks out the correct (ideal) inventory and word boundaries. We also show that, although the SC-score has some free parameters, it is largely independent of the way these parameters are set.

The paper is organized as follows. We begin by setting the framework of our experiment (modeling of phonetic variation, word segmentation, and semantics). Then, we introduce our learning bias, the SC-score, and explain how it links these different levels of representation in a coherent and intuitive fashion. Next, we present the results of our simulations on two different speech corpora in English and Japanese.

The framework

In order to acquire language, infants must undo various kinds of sub-phonemic variation present in the phonetics, segment words from continuous utterances, and assign meaning to these words. In this section, we explain how phonetic inventories, word segmentation, and semantics are operationalized in this study.

Corpora

We use two speech corpora: the Buckeye Speech corpus (Pitt, Johnson, Hume, Kiesling, & Raymond, 2005), which consists of 40 hours of spontaneous conversations with 40 speakers of American English, and the core of the Corpus of Spontaneous Japanese (Maekawa, Koiso, Furui, & Isahara, 2000) which consists of 45 hours of recorded spontaneous conversations and public speeches in different fields, ranging from engineering to humanities. Following Boruta (2012), we use an inventory of 25 phonemes for transcribing Japanese. For English, we use the phonemic transcription of Pitt et al. (2005), which consists of a set of 45 phonemes. We take these phonemic transcriptions to give the ideal lexical inventories for the two languages.

Phonetic variation

We generate alternate inventories for English and Japanese by modifying the phonetic transcription of each corpus, starting from the ideal (i.e., phonemic) transcription.

To generate inventories smaller than the true inventory, we collapse the segments into 9 natural classes: stops, fricatives, affricates, nasals, liquids, glides, high vowels, mid vowels and low vowels; then, into 4 coarser-grained classes: obstruents,

nasals, sonorants and vowels; and, finally, into only two segmental categories: consonants and vowels. We then rewrite the corpus transcription using each of these alternate inventories.

To generate inventories larger than the true inventory (i.e., with a finer grain than the phoneme), we use the same logic as in Peperkamp, Le Calvez, Nadal, and Dupoux (2006) and Martin, Peperkamp, and Dupoux (2013), and consider contextual allophones. That is, a given segment is split into possibly several allophones as a function of its left and/or right context as in Figure 1.

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

Figure 1: Allophonic variation of French /ʁ/

In order to generate these allophones in a phonetically controlled fashion, we follow Fourtassi, Schatz, Varadarajan, and Dupoux (2014) in using Hidden Markov Models (HMM).

We convert the raw speech waveform of the corpora into successive vectors of Mel Frequency Cepstrum Coefficients (MFCC), computed over 25 ms windows, using a period of 10 ms (the windows overlap). We use 12 MFCC coefficients, plus the energy, plus the first and second order derivatives, yielding 39 dimensions per frame. Each state is modeled by a mixture of 17 diagonal Gaussians.

The HMM training starts with one three-state model per (true) phoneme. Then, each phoneme model is cloned into context-dependent triphone models, for each context in which the phoneme actually occurs (for example, the phoneme /ɑ/ occurs in the context [d–ɑ–g] as in the word /dɑg/ ("dog"). The triphone models were then retrained on only the relevant subset of the data, corresponding to the given triphone. Finally, these detailed models were clustered back into artificial inventories of various sizes (from 2 to 8 times the size of the phonemic inventory) using a linguistic feature-based decision tree. The HMM states of linguistically similar triphones were tied together so as to maximize the likelihood of the data (Young et al., 2006).

Word segmentation

In the word segmentation literature, we can distinguish two major types of algorithms, modeling two strategies infants might use to segment words from continuous speech. The first is boundary detection using transition probabilities (TP) between pairs of phones. For example, the sequence [pd] occurs almost nowhere in the English lexicon, so the TP of [p] and [d] is very low; [pd] thus likely signals a word boundary. Empirical studies have shown that infants can use TP statistics in word segmentation (Saffran, Aslin, & Newport, 1996).

The second strategy is lexicon building. Unlike the previous strategy, where words are obtained as a mere byproduct of boundaries, this strategy looks explicitly for reoccurring chunks in the input, and uses them to parse novel utterances. Ngon et al. (2013) have shown that infants indeed recognize highly frequent n-grams (both words and non-words).

For this study, we use state-of-the-art algorithms from each of these two families. On the boundary detection side, we use the Diphone-Based Segmentation (DiBS: Daland and Pierrehumbert, 2011); from the lexicon building side, we use an Adaptor Grammar with a Unigram Model (AG: Johnson, Griffiths & Goldwater, 2007). The input to these models consists of a phonetic transcription of the corpus, with boundaries between words eliminated (we vary this transcription to correspond to the different candidate inventories in both experiments below). The models try to reconstruct the boundaries, following their respective strategies.

For the evaluation, we use the same measures as Brent (1999) and Goldwater (2006), 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 ideal segmentation. The F-score is defined as the harmonic mean of Precision and Recall:

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

Semantics

We use as the semantic representation of a word its frequency distribution over different documents (contexts). This simplified way of assigning meaning to words is known as Distributional Semantics. The idea can be traced back to Harris (1954): the meaning of a word can be inferred in part from its context. For us, this is more than a simplifying assumption. The SC-score we propose below uses *only* this contextual representation. It is usable by a learner who has no referential semantic knowledge.

We chose one of the simplest and most commonly used distributional semantic models, Latent Semantic Analysis (LSA: Landauer & Dumais, 1997). The LSA algorithm takes as input a matrix consisting of rows representing word types and columns representing contexts in which tokens of the word type occur. A context is defined as a fixed number of utterances. Singular value decomposition (a kind of matrix factorization) is used to extract a compact representation, in which words and contexts can be represented as vectors smaller than the original matrix (we call this reduced size the *semantic dimension* of the model). The cosine of the angle between vectors in the resulting space is used to measure the semantic similarity between words. Two words have a high semantic similarity if they have similar distributions, i.e., if they co-occur in most contexts. The model has two parameters: the dimension of the semantic space, and the number of utterances taken as defining the context of a given word form.

The self-consistency score

In this section, we introduce the self-consistency score. It takes as input a representation of the lexicon, including dis-

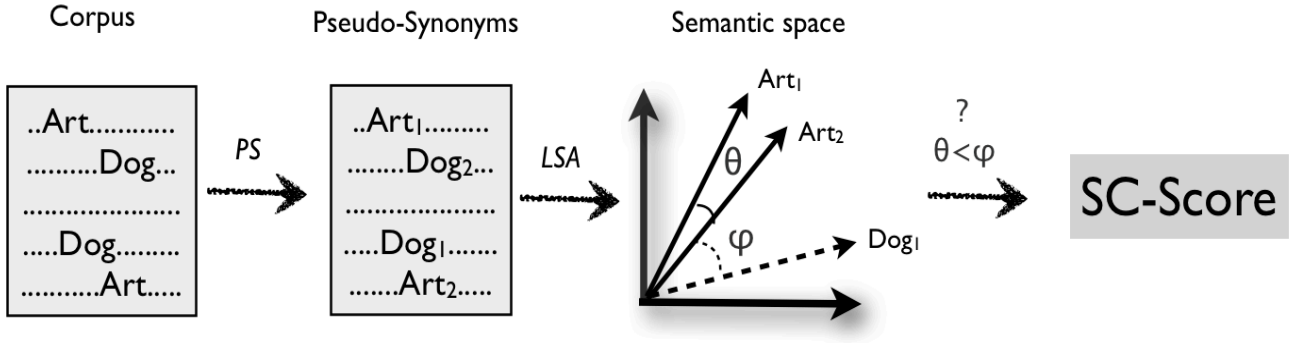

Figure 2: A schematic description of the SC-score computation

tributional semantic information, and outputs a score that reflects the global contextual informativity of the lexicon.

Representation of the lexicon

The representation of the lexicon varies along two dimensions: first, the **segmentation** that defines it. For example, the utterance "the doggie is eating" (represented here orthographically for readability) can lead to the following lexicons (among others): {the, dog, -ie, is, eat, -ing}, {the, doggie, is, eating} (under the ideal segmentation) or {thedoggie, iseating}. Depending on the segmentation strategy, we may end up with an oversegmented or undersegmented lexicon (or a mix of both). Second, the **segmental inventory** on which it is based. For example, the lexical item "cat" can have the following representations: /CVC/, /kæt/, or /$k_2æ_1t_3$/. Depending on how fine-grained the inventory is, some representations will be underspecified and some will be "overspecified." In Experiment 1, we examine how these two dimensions interact.

How the score is computed

Suppose we have a representation of the lexicon, i.e., a combination of an inventory and a segmentation. Each item is, in addition, endowed with a distributional information (a vector representing frequencies over contexts) as explained above. The self-consistency score operates at the distributional semantic level and examines the extent to which the distribution of items over different contexts is consistent. It is illustrated schematically in Figure 2, and it is computed as follows.

First, for each representation, we generate a pseudo-synonym corpus, (PS-corpus), where each word is randomly replaced by one of two lexical variants. For example, the word dog is replaced in the PS-corpus by dog_1 or dog_2. In the derived corpus, each word that occurs at least twice is duplicated, and each variant appears with roughly half of the frequency of the original word.

Second, we perform a same–different task: a pair of words is selected at random from the derived corpus, and the task is to decide whether the two are variants of each other or not based (only) 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 to compute a Receiver Operating Characteristic curve (Fawcett, 2006), from which one derives the area under the curve. The resulting score can be interpreted as the probability that, given two pairs of words, of which one is a pseudo-synonym pair, the pairs are correctly identified based on cosine distance. This is the SC-score. A value of 0.5 represents pure chance, and a value of 1 represents perfect performance.

When we split the tokens of a lexical item in two variants at random, these two variants (pseudo-synonyms) should still have roughly the same distributions, leading to a high distributional semantic similarity. The more consistent the distribution, the higher the similarity between the two pseudo-synonyms, and the easier it gets to distinguish them from random pairs. Intuitively, if a lexicon is coherent, it will have the property that it supports predicting a word from other words in its context.

In Experiment 2, we examine how the SC-score allows us to select the optimal representation of the lexicon.

Experiments and discussion

Experiment 1: Interaction between variation and segmentation

As a prelude to our test of the SC-score, (Experiment 2), we explore how the sound inventory influences the outcome of the segmentation strategies, in order to check whether the general strategy is valid. We want to see whether optimizing one part of the representation can lead to better results for another part. In Figure 3, we show the token F-scores under different inventories. The F-score is computed by comparing the segmentation under a given inventory with the ideal segmentation under the same inventory. It shows that both segmentation strategies are optimal for the phonemic inventory. Their performance drops for both finer- and coarser-grained inventories.

The token F-score, however, penalizes over- and under-segmentation equally. In order to explore the kind of errors made by the segmentation algorithm, we compared the bound-

ary precision (number of correct boundaries found, out of all boundaries posited) and recall (the number of correct boundaries found, out of all boundaries in the ideal segmentation). If the precision is higher than the recall, then the algorithm has a tendency to under-segment; if precision is lower than recall, the algorithm has a tendency to oversegment (Goldwater, 2006). To give an intuitive sense of why this is the case, consider the segmentation of the utterance: /ðə dɑg/ ("the dog"). The extreme oversegmentation corresponds to the case where the algorithm considers there to be a boundary between each pair of phones: /ð ə d ɑ g/. The boundary precision of this segmentation is very low, and the boundary recall is maximal.

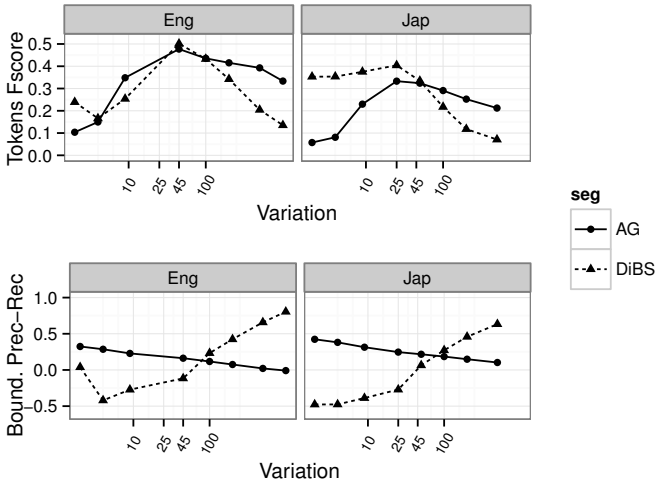

Figure 3: Segmentation scores

We show in Figure 3 the difference between precision and recall (precision − recall) as a function of the level of phonetic variation. We find an interesting interaction between variation and segmentation strategy. Variation seems to cause the AG to undersegment less and less in the range of variation that we are considering, and the general pattern points towards oversegmentation with more variation. For DiBS, on the other hand, the pattern moves from oversegmentation to undersegmentation as soon as the inventory becomes finer than the phonemic representation. The reason for the first pattern is that larger inventories increase the number of word forms, which each occur, therefore, with lower frequency. Consequently, the lexicon building algorithm will posit as "words" smaller chunks, which still occur with reasonable frequency. As for the second pattern, for a given pair of phones, the boundary probability drops as the inventory size increases. Consequently, many pairs that would otherwise be above the boundary threshold, will drop below it, leading to undersegmentation for DiBS.

Experiment 2: Evaluation of the lexicon representations

In the previous experiment, we showed that the quality of the lexicon can be used to choose the right amount of variation at the phonetic level. However, the information about the segmentation performance was based on the comparison with the ideal segmentation. In this experiment, we go a step further in our reasoning: we test whether moving to a higher level of representation can offer an unsupervised alternative.

Each representation of the lexicon corresponds to a corpus transcribed with a phonetic inventory and segmented using one of the segmentation algorithms. To evaluate a representation, we generate a PS-corpus (as described in the previous section) and apply the LSA model to derive the distributional semantic space, in which each word is represented by a vector corresponding to the distribution of its tokens over the relevant dimensions (which could be seen as topics). Next, we derive the matrix of distributional semantic distances between all pairs of words in the lexicon. Finally we compute the SC-score based on this matrix (Figure 2). The LSA was performed using the software Gensim (Řehůřek & Sojka, 2010).

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 10 and 100 utterances, and for semantic space dimensions ranging from 10 to 200 dimensions (Fourtassi and Dupoux (2013) showed that the performance of LSA tends to level out after about 200 dimensions).

In addition to DiBS and AG, we consider a random segmentation and the ideal (gold) segmentation as controls. Figure 4 shows the SC-score as a function of the inventory and the segmentation. For a given segmentation, the score peaks around the phonemic inventory of the language (45 in English and 25 in Japanese). The absence of such a 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 consistent, since it has more homophones. In an inventory composed of coarse-grained natural classes, two words that have 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. The smaller the inventory is, the more homophones will be created, and the less informative word-level context will be.

Larger inventories increase the number of types, which therefore occur with lower frequency. This makes the contextual representation less informative. The case of extreme 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 (as compared to 33 in the phonemic inventory). Ratios of this order of magnitude are evidently not sufficient to build a predictive lexicon.

For a given inventory, the SC-score distinguishes between

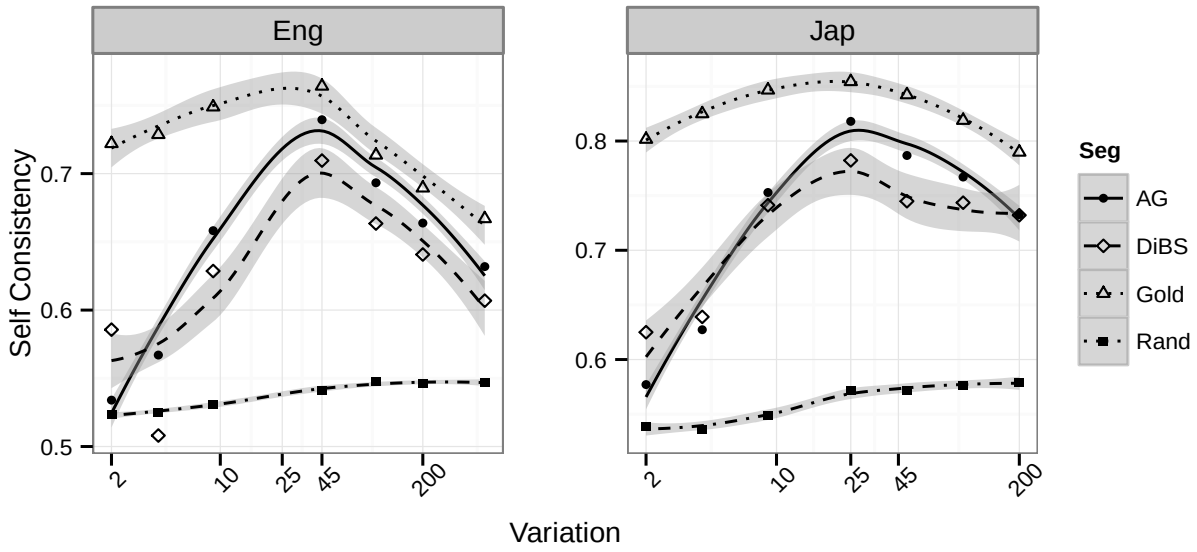

Figure 4: Self-consistency scores across different phonetic inventories and different word segmentations. The points show the mean score over different parameter settings. The lines are smoothed interpolations (local regressions) through the means. The grey band shows a 95% confidence interval.

random, ideal, and intermediate-quality segmentations. Note that we obtain this result without making use of the ideal segmentation as in Experiment 1. For random segmentation the reason is apparent: the distribution of a type across contexts will clearly not be consistent. The reason the ideal segmentation leads to a better score as compared to the output of the segmentation algorithms is that oversegmentation and undersegmentation both change the token/type ratio. The extreme case of undersegmentation corresponds to taking each utterance as a word; the chances of a whole utterance being repeated enough times to lead to an informative and consistent distribution are very small. Conversely, the extreme case of oversegmentation corresponds to taking each phone as a word. As in the case of extreme homophony, this leads to a very small lexicon with technically an uninformative (flat) distributional over contexts.

Figure 4 also indicates that the utility of the SC-score in picking out the best representation for the lexicon is largely independent of the parameter settings (the confidence bands are over runs with different parameter settings). Statistical tests confirm this. For the inventory size, three (generalized) linear models confirm that the ideal (phonemic) inventory is the peak: the ideal inventory runs versus the next most coarse-grained inventory (9 categories, for both English and Japanese); ideal inventory versus the next most fine-grained inventory (100 for English, 50 for Japanese); and the ideal inventory runs versus all the other runs (in that case, refitting the model many times, undersampling the non-ideal runs uniformly each time to get balance in the two groups). A generalized linear model (logit link) on SC-score with ideal-size/non-ideal-size as a fixed effect, and with segmentation model and language as random effects (intercepts and slopes) gives an estimated increase of 0.200 (logit scale) for ideal-sized versus next-coarser ($p = 5 \times 10^{-5}$); of 0.168 for ideal-sized versus next-finer ($p = 5 \times 10^{-10}$); and 0.350 for ideal-sized versus all other runs (mean, $N = 10000$; geometric mean $p = 3 \times 10^{-8}$). Similarly, we confirm improvements for AG versus random segmentation (0.896, $p < 2 \times 10^{-16}$); DiBS versus random (0.728, $p < 2 \times 10^{-16}$); gold versus AG (0.412, $p < 2 \times 10^{-16}$); gold versus DiBS (0.527, $p < 2 \times 10^{-16}$).

Thus, the SC-score enables us to select the right representation (for instance, the size of the segmental inventory and the size of the lexicon) without any hyperparameter tuning.

Conclusion

We have introduced a learning bias that provides a potential guide for infants during language acquisition. The SC-score is not a learning algorithm: it does not account for *how* a representation is built. Here, learning is stated statically, abstracting away from the actual learning procedure, as in the "evaluation measure" approach to language acquisition (Chomsky, 1965). Thus it should be seen as an inductive bias in that it provides a criterion for ranking different candidate representations. The simplifying assumption being made here is that all the representations are consistent with the data from the infant's perspective. As such, the SC-score corresponds to a "prior probability" in the Bayesian framework (Jaynes, 2003), operating in a space of hypotheses, where a hypothesis is defined as a representation of the lexicon (as defined in section

2) associated with a distribution over contexts.

The philosophy of the bias is that infants are learning and optimizing an entire system, rather than optimizing different sub-levels in isolation. Thus, a representation at one level is constrained by the extent to which it is compatible with other levels, like pieces of a puzzle.

We assume that language acquisition 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 representation is measured by the informativeness of context when that representation is used. We have operationalized this using a measure we call self-consistency, which applies to the lexicon. We found our method to disfavor both over-fine and over-coarse hypotheses, based, strikingly, on a purely intrinsic criterion having nothing to do with phonology per se. We found optimal SC-scores for the true phonemic inventories and the ideal word segmentations of two typologically different languages: English and Japanese. We also found the SC-score to be independent of the parameter setting to a large extent, and to operate with minimal, if any, external supervision.

Acknowledgments

This work was supported in part by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the Fondation de France, the Ecole de Neurosciences de Paris, and the Région Ile de France (DIM cerveau et pensée).

References

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

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*(1-3), 71-105.

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

Daland, R., & Pierrehumbert, J. B. (2011). Learning diphone-based segmentation. *Cognitive Science*, *35*(1), 119-155.

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

Fourtassi, A., & Dupoux, E. (2013). A corpus-based evaluation method for distributional semantic models. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop* (pp. 165–171). Sofia, Bulgaria: Association for Computational Linguistics.

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 the 52nd Annual Meeting of the Association for Computational Linguistics*.

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

Goldwater, S. (2006). *Nonparametric bayesian models of lexical acquisition* (Unpublished doctoral dissertation). Brown University.

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

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge University Press.

Johnson, M., Griffiths, T. L., & Goldwater, S. (2007). Adaptor grammars: a framework for specifying compositional nonparametric bayesian models. In B. Sch"olkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems 19* (pp. 641–648). Cambridge, MA: MIT Press.

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

Maekawa, K., Koiso, H., Furui, S., & Isahara, H. (2000). Spontaneous speech corpus of japanese. In *LREC*. Athens, Greece: European Language Resources Association.

Martin, A., Peperkamp, S., & Dupoux, E. (2013). Learning phonemes with a proto-lexicon. *Cognitive Science*, *37*(1), 103-124.

Ngon, C., Martin, A., Dupoux, E., Dominique, C., Dutat, M., & Peperkamp, S. (2013). (non)words, (non)words, (non)words: evidence for a protolexicon during the first year of life. *Developmental Science*, *16*(1), 24–34.

Pearl, L., & Goldwater, S. (in press). Statistical learning, inductive bias, and bayesian inference in language acquisition. In J. Lidz (Ed.), *Oxford handbook of developmental linguistics*. Oxford University Press.

Peperkamp, S., Le Calvez, R., Nadal, J.-P., & Dupoux, E. (2006). The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, *101*(3), B31–B41.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. D. (2005). The buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89-95.

Řehůřek, R., & Sojka, P. (2010, May 22). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

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