Motif discovery in infant- and adult-directed speech

Bogdan Ludusan[1], **Amanda Seidl**[2], **Emmanuel Dupoux**[1], **Alejandrina Cristia**[1]
[1]Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS)
Département d'Études Cognitives, École Normale Supérieure, PSL Research University, France
[2]Department of Speech, Language, and Hearing Sciences
Purdue University, USA
bogdan.ludusan@ens.fr, aseidl@purdue.edu
{emmanuel.dupoux, alecristia}@gmail.com

Abstract

Infant-directed speech (IDS) is thought to play a key role in determining infant language acquisition. It is thus important to describe how computational models of infant language acquisition behave when given an input of IDS, as compared to adult-directed speech (ADS). In this paper, we explore how an acoustic motif discovery algorithm fares when presented with speech from both registers. Results show small but significant differences in performance, with lower recall and lower cluster collocation in IDS than ADS, but a higher cluster purity in IDS. Overall, these results are inconsistent with a view suggesting that IDS is acoustically clearer than ADS in a way that systematically facilitates lexical recognition. Similarities and differences with human infants' word segmentation are discussed.

1 Introduction

The ability to learn words from continuous speech is a crucial skill in language acquisition, since only about 7% of words occur in isolation, and thus infants must be able to segment, i.e. pull out words from running speech. It has been proposed that infant-directed-speech (IDS), the particular register that parents use when addressing their infant, could facilitate word segmentation when compared to adult-directed-speech (ADS) (Singh et al., 2009; Thiessen et al., 2005). Even though a number of acoustic and linguistic studies have documented systematic differences between these registers (Cristia, 2013; Fernald and Morikawa, 1993), there is little computational work assessing how precisely word segmentation performance is affected by these differences. The present report takes one step in this direction.

1.1 Computational model of word segmentation

We model infant word learning using MODIS (Catanese et al., 2013), a computational system which attempts to discover spoken terms from the raw speech signal. We think that this system is cognitively plausible for several reasons. First, the algorithm does not rely on labeled or pre-segmented data. Instead, it takes as input spectral features and looks for repetitions inside of a short signal buffer (which thus resembles a short-term memory). When the first repetition is found, two acoustic stretches that are judged to be matched are stored together as a cluster (represented as a kind of average of the acoustic items it contains) inside the library. Clusters can be thought of as 'lexical entries' in the context of this project and the library as its long-term memory. It then continues parsing the speech looking for matches with respect to the clusters in the long-term memory as well as other close repetitions in the buffer. If a match to an existing cluster is found, the cluster model is updated, in order for it to also contain information about the latest token.

Given its general features, this algorithm appears to be a reasonable approximation of word segmentation strategies used by a naïve learner (a learner who has not yet extracted abstract phonemic categories). It is very likely that infants begin to segment words before they have learned their language's phoneme inventory since, in certain situations, infants as young as 4 months of age can recognize words in fluent speech (Johnson et al., 2014), but there is little evidence that infants this young have converged upon their native phonemes (Tsuji and Cristia, 2014). Moreover, since young infants can more easily recognize word tokens that are similar acoustically, than tokens which are dissimilar (Bortfeld et al., 2005; Singh et al., 2012), it follows that an acoustic motif discovery algorithm is not an unreasonable first approach.

Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning, pages 93–102,
Lisbon, Portugal, 18 September 2015.

It is also very plausible that the patterns infants discover in running speech will be constrained to a short-term memory window, although we do not know of evidence directly addressing this (most work has investigated the limits of long-term memory, e.g. (Houston and Jusczyk, 2003), rather than how close in time two subsequent repetitions must occur to be detectable). Finally, we know that infants can store repeated words in some form of long-term memory because this is precisely the type of design that typical word segmentation studies have, whereby the child is familiarized with a word repeated and later tested with novel instances of those wordforms.

1.2 Influencing factors and general predictions

Properties of IDS compared to ADS	Predictions for word learning
Not tested in this paper	
prosodic boundaries easier	IDS>ADS
clearer referential situation	IDS>ADS
simpler vocabulary	IDS>ADS
more attention grabbing	IDS>ADS
Tested in this paper	
acoustically more variable	IDS<ADS
more repetitions	IDS>ADS
more bursty	IDS>ADS

Table 1: Differences between IDS and ADS and potential effects for word learning.

IDS is characterized by an array of properties (Cristia, 2013, see Table 1), some of which could facilitate or hinder word segmentation. IDS has been reported to contain shorter utterances and clearer *prosodic boundaries* than ADS. To test this would require a learner that extracts and uses prosodic cues from the speech signal, which is not the case in the current implementation of MODIS (see also the Conclusions). The same would also be true for *referential and contextual cues*. The effect of *vocabulary* is neutralized in our experiment, because the corpus used contained the same keywords in both registers, and only these keywords were considered for the evaluation of word learning. Regarding *attention*, since MODIS works by finding acoustic matches in the speech signal, it does not have a cognitive component that models the attention process.

Therefore, none of first 4 differences in Table 1 are tested here. Instead, our corpus and computational model allows us to look at differences in performance related to three other properties of IDS: *acoustic variability*, *repetitions*, and *burstiness*.

First and foremost, mounting evidence suggests that sounds and words are more variable in IDS than ADS. For instance, Martin and colleagues have documented that phonemic categories are significantly *harder* to classify in IDS than in ADS (Martin et al., 2015). This may be due to an increase in variability, which has been documented in several studies (Cristia and Seidl, 2014; Kuhl et al., 1997; McMurray et al., 2013). If the acoustic implementation of phonemes is more variable in IDS than ADS, it is possible that other linguistic levels that build on sounds, such as words, might also be significantly different across the registers.

To our knowledge, there is only one modelling study that partially investigated this question, although it was not a model of word learning, but rather of phoneme learning. Kirchhoff and Schimmel (2005) trained a speech recognizer with human-segmented and labeled tokens of three minimally different target words (sheep, shoe and shop) drawn either from IDS or ADS, and tested the performance on a new set of IDS and ADS tokens. Results revealed a lower performance overall in the IDS-trained classifier, but a smaller generalization cost (i.e., the loss in performance in switching from IDS to ADS was smaller than vice versa). These results are consistent with the idea that words are more variable in IDS, and suggest that there could be learnability differences across the two registers. It remains to be seen whether such effects would also emerge in a model of word learning in which there is no explicit human-obtained segmentation and labels.

It is to be expected that acoustic variability could be problematic to learners who find wordforms using acoustic pattern matching, leading them to posit too many or too few types (e.g., the word *dog* is so variable that the learner posits two different types, dog1 and dog2; or confuses them with similar words such that *dog* and *dock* are clustered together). Laboratory work in infancy demonstrates that early on infants have difficulty matching wordforms that are acoustically variable (Bortfeld et al., 2005; Singh et al., 2012), as if infants create separate lexical entries for e.g., the word *dog* spoken by two different speakers. This

is precisely what occurs with word segmentation models that operate on the basis of acoustic motif discovery, and thus we predict that our the model would perform more poorly in IDS than ADS, because of its greater variability.

The second property of IDS that we address is *repetition*. It has been reported that IDS is more repetitive than ADS (Daland, 2013). We suspect repetition is perceptually relevant to infants because most word segmentation experiments use very repetitive stimuli, and this feature has even been found to draw infants' attention (McRoberts et al., 2009). Some repetition is necessary for our model learner, as this is a condition for incorporating an item into the lexicon (words that are not repeated cannot be found). However, once the second token of the same type is detected, it is unclear whether any benefit is derived from additional repetitions. MODIS will decide whether a pattern encountered matches one in the lexicon, by comparing the new pattern to an average or prototype of all the other patterns in that cluster. Hence, it is possible that additional tokens of the same type will simply compound the negative effects of increased segmental variation.

A third property of IDS that we address is *burstiness*, which characterizes the likelihood of a word to re-appear in the same conversation once it has been used. Thus, registers where one tends to stay longer on a given topic will be more bursty - for instance, news reports are more bursty than spontaneous phone conversations. Daland has hypothesized that burstiness should be higher in IDS than ADS (Daland, 2013), although we know of no systematic investigation of IDS corpora or the effects of burstiness on infant perception. Nonetheless, it is certain that burstiness should improve the word segmentation performance of our learner, since having a higher proportion of repetitions of the same word inside the short memory buffer would translate into a higher chance of detecting that word.

2 Methods

2.1 Corpus

2.1.1 Speakers

The twenty speakers in this study were ten mothers of 4-month-olds (M = 0;4.35, range: 0;3.95 -0;4.99) and ten mothers of 11-month-olds (M = 0;11.40, range: 0;11.120;12.01). The mothers were the child's primary caregiver, and native speakers of American English from a small Midwestern city. Infants were healthy full-terms with typical development and no known personal or familial history of hearing or language impairments, according to parental report.

2.1.2 Recording and human coding procedure

Full details on the corpus can be found on: `https://sites.google.com/site/acrsta/Home/nsf_allophones_corpora`. The key information for the present purposes is the following:

Speakers were provided with a set of objects and photos, each labeled with a target word. They were told that we were interested in how parents talk to their children about objects. The words containing the vowels did not constitute minimal pairs, so as not to make the parents overly conscious of the contrasts under study. The IDS portion was always carried out first, and during it, the caregiver and child were left alone. When the mother had finished going through all items, an experimenter returned accompanied by a confederate adult. The mother then repeated the task with the confederate.

The 20 speakers included in the present work are a subset of 36 mothers whose speech (excluding sections with overlapping noise or speech) had been analyzed in previous work (Cristia and Seidl, 2014). In that study, only one vowel per target word was coded and analyzed. A subset of caregivers was used for the present purposes because their speech had also been coded to investigate whether IDS and ADS differed to similar extents in the weak and strong vowels of bisyllabic and trochaic target word (Wang et al., 2015). This meant that we had access to the temporal location of both strong and weak vowels in some of the target words.

For the current study, since only the first two vowels of some words were coded in the corpus, we will use a proxy for words, which we call a target segment. It is defined as being the stretch of speech between the beginning of the first vowel and the end of the second vowel of a coded word. This definition is illustrated in Figure 1. We then kept the target segments appearing in both speech registers and collapsed all composed words classes into the class containing the first word only (e.g. picnic basket → picnic, peekaboo book → peekaboo), since the coded vowels actually belong to

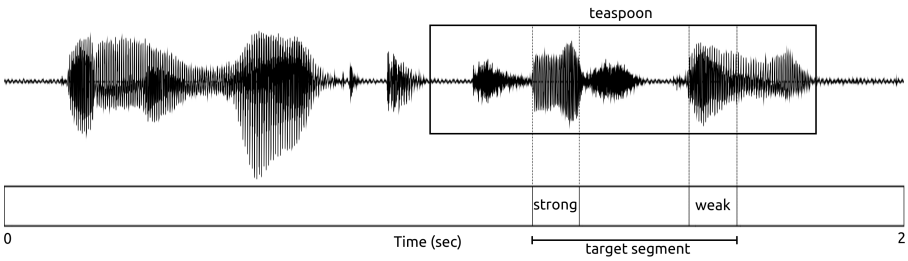

Figure 1: Example of target segment. The waveform and associated annotation of the utterance "Then we have a teaspoon" is illustrated. The annotation codes the position of the two vowels of the word "teaspoon". Below the vowel annotation we illustrate the target segment considered, defined as the stretch of speech between the beginning of the first vowel and the end of the second vowel of a given word. For comparison, the entire word is represented above the waveform.

the first word (e.g. picnic, peekaboo), not to the second one (e.g. basket, book). Composed words whose first word contained only one vowel were kept in their own separate class (e.g. tea-kettle, best-in-show). This gave us a total of 2298 target segments, 1300 in IDS and 998 in ADS. A complete list of the target segments is presented in Appendix A. Note that this coding was used only for evaluation purposes, as there is no training phase for the algorithm.

As can be seen in the example in Figure 1, the target segment only partially covers an actual word. We have estimated this coverage to be between 80-90%, in the case of words starting with a consonant-vowel (CV) sequence and ending in a vowel (e.g. bamboo, pesto) and around 50%, in the case of 4-syllable words (e.g. dictionary, tapioca). Since most of the target segments, both in terms of number of types and number of tokens, belong to the words starting with a CV sequence and ending with a vowel-consonant sequence (e.g. bacon, picnic), we could conclude that the majority of our target segments cover at least 2/3 of the actual word.

2.1.3 Corpus characteristics

In previous analyses comparing IDS and ADS on this subset of the corpus, pitch was found to be higher (particularly in stressed vowels) in IDS and there was also a trend for more peripherality in IDS, but no stable differences in vowel duration were seen (Wang et al., 2015). Also, an analysis of the whole corpus has shown greater variability in acoustic characteristics of stressed vowels in IDS than ADS (weak vowels had not been marked or analyzed) (Cristia and Seidl, 2014). Thus, the corpus represents well the prosodic and segmental characteristics of IDS alluded to in the introduction.

For the purposes of the present project, we further investigated potential differences in repetition. As expected, parents produced more repetitions of the target segments in IDS than ADS (significant according to a Wilcoxon's test, V (19) = 187, $p = 0.001$; mean for IDS = 3.358 repetitions per target segment, SD = 1.358; mean for ADS = 2.147, SD = 0.564).

Besides computing a measure of repetition, we have also attempted to measure differences in burstiness between IDS and ADS. Burstiness was defined as the reciprocal of the average distance (in seconds) between the end of the n^{th} occurrence of a target segment and the beginning of the $n+1^{th}$ occurrence of the same word, provided that these two occurrences are not separated by another target segment. It was computed on a per-speaker basis and only for target segments appearing at least twice, in both the IDS and ADS recordings of the same speaker. About 4.618 seconds elapsed between two consecutive repetitions in ADS, compared to 7.371 in IDS. This meant that the average burst rate was 0.292 (SD = 0.186) in ADS, and 0.15 (SD = 0.065) in ADS. Thus, contrary to our expectations, a higher burstiness was obtained for ADS than for IDS.[1]

[1]We checked whether the difference in burstiness could be explained by the speech rate difference between the two registers. In order estimate speech rate, we calculated the average duration of the target words, all of which were bi-syllables and occurred in both registers.The average duration was .311 s (SD = .042) in ADS and .362 (SD = .068) in IDS, in line with the view that IDS is slower than ADS.The speech rate difference (14%) does not seem to fully explain the difference seen in the burstiness between the two registers (48%). Nonetheless, this measure does not take into account pauses, which are likely to be considerably longer in IDS.

We have seen that the three IDS characteristics that might affect the performance of the model point in different directions. We lay out our predictions once our evaluation metrics have been introduced.

2.2 Algorithm

We used the open-source spoken term discovery system called MODIS (Catanese et al., 2013). Is it based on the seed discovery principle: it searches for matches of a short audio segment, referred to as the seed, in a larger segment, called a buffer. The search is performed by using a segmental variant of the dynamic time warping (DTW) algorithm. Once a match is found (decision taken based on a similarity threshold between the two speech segments), the seed will be extended and the match performed using the longer seed. This process will continue as long as the dis-similarity between the segments stays under the set threshold. When this threshold is reached, the term candidate is checked as to whether is complies with a minimum length requirement and stored in the motif library. An abstraction of the matched segments is stored in the library, represented by their median model. Next, this library of terms is compared against any new seed and only if no match is found in the library will the DTW search explained earlier take place. The match against the library terms employs also a self similarity matrix check. After the entire data set is searched, a post-processing of the obtained terms is performed in order to merge all overlapping segments into one single term.

The algorithm has several important parameters that must be set: the *seed size*, the minimum stretch of speech matched against the buffer, the *minimum term size* the algorithm will find, the *buffer size* in which the seed is searched and the *similarity threshold*, ϵ_{DTW}. Since the latter parameter influences the level of similarity between the members of the same term class, we have varied it in our experiment, while keeping the rest of the parameters constant. The seed length was set to 0.25 s, while the buffer length was set to 90 s, in order to model infants' short-term memory. The minimum term size considered was 0.5 s so as to be able to contain the majority of the target segments.

The variation of the similarity threshold can be seen as follows: When this parameter is low, even small deviances of similarity are rejected, representing a 'conservative' approach. When it is high, even large dissimilarities are accepted, representing a 'lax' approach. Based on previous infant word segmentation research, it appears that young children are conservative early on (Singh et al., 2012) – but how conservative? There is no principled way to set this parameter, as any decision we make would likely not have a clear basis in research. However, in order to restrain the search range of ϵ_{DTW} values on which we will perform our analysis, we ran MODIS on the combined ADS-IDS recordings of one speaker and we decided to take an interval of [2.0, 4.0]. The minimum value was the lowest threshold that returned any term classes, while the maximum value was the threshold value that gave a saturation point for the evaluation metrics measured.

We use as input features for the spoken term discovery system Mel frequency cepstral coefficients, a standard spectral representation used in speech applications. We compute the first 12 cepstral coefficients and the energy in a 20 ms window, every 10 ms, along with their delta (difference) and double delta (acceleration) coefficients.

2.3 Evaluation

As noted in the Introduction, our conceptual goal is to compare performance of this segmentation algorithm between IDS and ADS. We have also drawn several specific predictions. In this section, we explain how these predictions map onto the dependent variables used for the evaluation.

Since the corpus has not been exhaustively coded, we did not penalize the algorithm for clusters that do not include any target segments. Indeed, there may be other words that are repeated in the corpus (e.g., 'baby' or 'mommy') which have not been coded, so clusters could have been formed around these other words. Instead, we inspect only clusters that include at least one token of a target segment.

Given that word edges for target segments are not marked (only vowels), we consider a cluster to include a given token if one of the acoustic stretches included in that cluster covers the region between the beginning of the first vowel of the word and the end of the second vowel of that token. We derive a measure of *recall* as the number of tokens that appear in any given cluster divided by the total number of coded tokens. It is possible

We return to the potential limitations of our implementation of burstiness in the discussion.

that the higher repetition found in IDS will lead to higher coverage in this register as compared to ADS. At the same time, the opposite outcome could be expected if one would take into account the higher burstiness found in ADS. Thus, a clear prediction cannot be made.

As mentioned, there are target segments whose vowels were not coded because they overlap with speech or noise or were not produced with the intended vowel, nonetheless, it is possible for the algorithm to recognize matches for such uncoded words. Therefore, it would be unfair to penalize clusters that include target segments as well as stretches of speech other than the target segments that have not been coded by humans. However, when one cluster contains tokens from two or more different target segments, this will be penalized by our second dependent measure, namely cluster *purity*. It is defined as being the number of different target segments contained in a cluster, divided by the number of target segment classes. On the basis of our arguments above, we cannot make any clear prediction regarding how IDS and ADS will differ for this measure.

Third, we derive a measure that describes the amount of fragmentation of the found motif clusters. It is defined as being the percentage of clusters into which a particular target segment is found, out of the total number of clusters where target segments have been found. We will report the results in terms of *collocation*, defined as being equal to 1 - the amount of fragmentation. We expect IDS, with its greater variability, to yield a lower collocation.

3 Results

Analysis scripts and primary data and results files are available for download from `https://osf.io/y7kfw/`.

When no target segment is found by the algorithm for the speech of one caregiver, this results in missing data, as no recall, purity, or collocation can be calculated in these conditions. Therefore, we excluded from inspection all settings of the similarity threshold that resulted in missing data prior to carrying out statistical analyses. Data was included for settings 2.9-4 (at .1 intervals).

In general terms, we observed that performance is very good in terms of collocation and purity (above .6 for all individual speakers and for both registers), with performance for both of these decreasing and becoming more variable at the individual level as ϵ_{DTW} is set to laxer criteria. In contrast, recall performance is overall lower and more variable, with coverage increasing as laxer criteria are used.

Turning now to our key question, we calculated the difference in performance in IDS and ADS, for each measure and for each speaker. We tested for significant differences across the two registers in two ways: (1) keeping each ϵ_{DTW} value separate, and (2) collapsing across all ϵ_{DTW} values.

To evaluate for significance in the separate case, given that many such tests would have to be carried out (there are 12 levels for the similarity threshold in each evaluation measure), we wanted to control for repeated testing to avoid alpha risk inflation. Therefore, we used a step-down permutation resampling test (N = 10,000) and estimated the p-value for an observed t-statistic (from a one-sample t-test) through the rank of that p-value within the distribution of values for that statistic found under the null hypothesis.[2]

For the analyses collapsing across this threshold, we took the median across all threshold values within each caregiver, and used a Wilcoxon one-sample test to assess whether this average difference score was significantly different from zero for each evaluation dimension separately. We decided to employ the median followed by a Wilcoxon's non-parametric test based on the sum of the signed ranks because there was not clear evidence that such difference scores were normally distributed (the distributions were kurtotic with some outliers).

Both analyses revealed that there were some

[2]In the general permutation procedure, a distribution of a test statistic under the null hypothesis can be generated as follows: the sign of a random number and selection of individual difference scores is flipped (such that what used to indicate higher performance in IDS than ADS becomes the opposite) and the appropriate statistic (in this case, the t from a one-sample t-test, following usual practice van der Laan et al. (2004)) is calculated. The procedure is repeated many times, to generate a distribution of p-values under the null hypothesis. The adjusted p-value is then estimated as the rank of the absolute of the statistic in question against the distribution of absolute values found when the null hypothesis is true. The step-down version of the permutation procedure involves two changes. First, flipping the difference scores is done for all observations associated with the same individual together, which preserves the correlational structure of the data. Second, the distribution under the null is calculated once with all the data, and then repeated removing the strand of data (in our case, all the data associated with a given threshold parameter value) whose adjusted p-value is significant. The procedure stops when the adjusted value exceeds alpha.

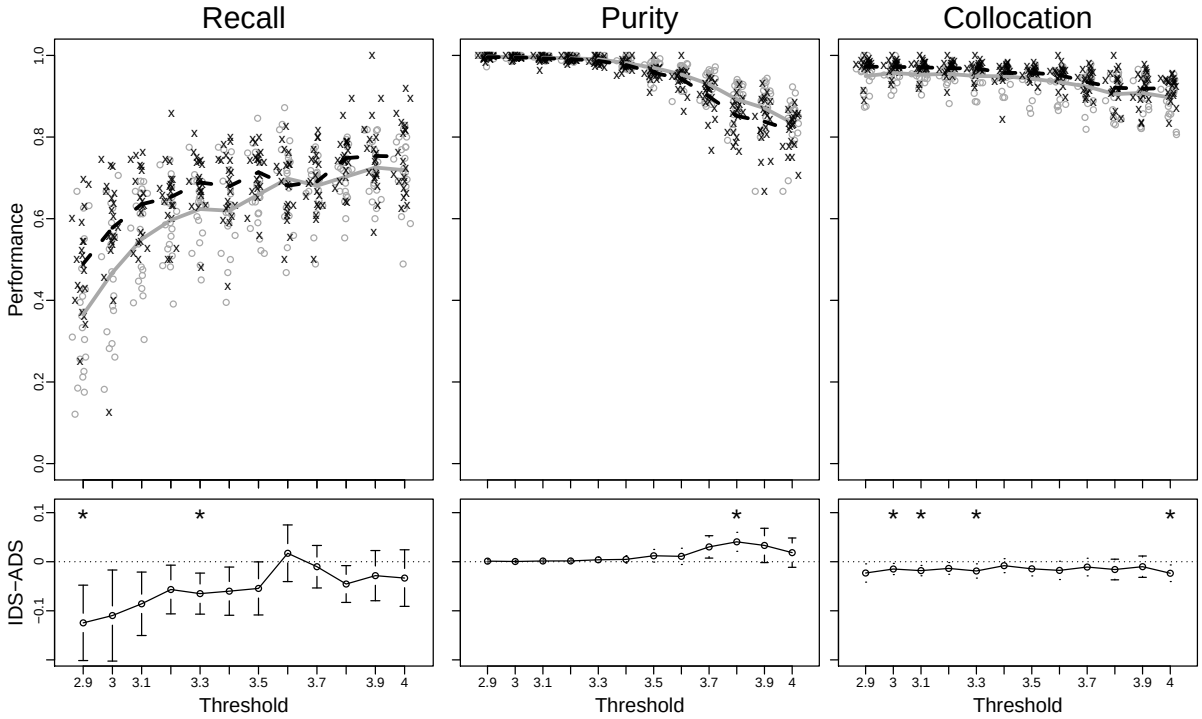

Figure 2: Performance (top panels; IDS in gray and ADS in black) and difference (IDS-ADS, bottom panels) for each of the three evaluation dimensions as a function of the ϵ_{DTW} threshold. Each point in the top panel represents a mother's score, separately for IDS (gray circles) and ADS (black crosses). The difference scores in the bottom represent the average difference and 95% confidence intervals across parents. Stars represent cases where the difference is significant at the $p < .05$ level, corrected for multiple comparison using a step-down permutation resampling test across parents.

significant differences across the registers for all the evaluation metrics computed. As shown in Figure 2, two of the ϵ_{DTW} values (both in the "conservative" region) lead to significantly higher performance in ADS than in IDS in terms of recall. This was replicated in our second analysis (based on the median across all ϵ_{DTW}): V(19) = 9, p=.016, 95% confidence interval (-0.087;-0.019), pseudo-median -0.050. As for purity, there was a trend for better performance in IDS than ADS that was significant for one ϵ_{DTW} value, closer to the liberal end of our threshold continuum. This advantage was replicated when looking at median values: V(19) = 55, p = .006, 95% confidence interval (.005;0.027), pseudo-median 0.016. As for collocation, performance was significantly better in ADS than IDS mostly in the same conservative region as with recall, a result replicated in the Wilcoxon's t-test on median difference scores: V(19) = 1, p=.003, 95% confidence interval (-.012;-0.006), pseudo-median -0.010.

Next, we had wondered whether greater repetition and burstiness would lead to better recall. The overall pattern of results appears to indicate this is not the case because although IDS has more repetitions, it has lower recall – although this could possibly relate to burstiness. As a first approach, we calculated Spearman correlations across speakers between recall performance (averaged across all parameters) and number of repetitions, on the one hand, or rate of burstiness, on the other, within each register separately.

As for repetitions, the estimate was moderate and positive in both registers, albeit significant for IDS r(18) = .549, p = .014, but only marginally in ADS r(18) = .430, p = .060. Thus, there appears to be some relationship between recall and repetition, but the greater number of repetitions in IDS over ADS is not sufficient for there to be a boost in recall in IDS over ADS overall.

Regarding burstiness, estimates were low, non-significant and surprisingly negative: IDS r(18) = -.159, p =.501; ADS r(18) = -.299, p =.199). The negative correlation would indicate that the higher burstiness is, the lower the recall – we return to this issue in the discussion.

4 Discussion

The first conclusion that must be drawn from the results of running our naïve learning algorithm on these data is that the difference in performance with IDS and ADS materials is subtle: Collapsing across threshold parameter values, it only amounts to absolute differences of between 1 and 5%. Nonetheless, these differences are there, since they surface in all three evaluation metrics, both when we use a multiple comparisons correction procedure, and when we average across all reasonable settings of the similarity threshold.

We had stated several predictions based on previous work. We had no clear expectation regarding *recall*, since the two factors that might affect it, repetition and burstiness, seemed to favour different registers. Overall, we observed an ADS advantage of about 5%, concentrated in the conservative regions of the similarity parameter. As for the relationship between repetition and recall, we found that while our IDS was more repetitious than the ADS, recall was lower for the former than the latter. However, the correlations in individual variation within each register were positive. This pattern of results partially supports our intuition: More repetition helps unsupervised motif discovery. However, the data go beyond our hunch in that differences in repetitiousness do not account for register differences. Regarding burstiness, we failed to confirm the prediction that IDS was more bursty, and we further found a negative non-significant correlation with recall. This may indicate that our corpus, elicited in a task where speakers did not have much lexical choice, was not ideal to measure burstiness differences. Additionally, the precise implementation we used may have confounded tempo differences, and an alternative burstiness definition, in terms of number of intervening words, could be more appropriate.

Turning to the second evaluation metric, *purity*, we also had no specific hypothesis, however, we found an overall advantage for IDS, with significant results for only one parameter value (located towards the liberal end of our continuum) as well as in analysis over median scores. Overall, performance with IDS was about 1.6% higher than that for ADS in this metric, this effect being mainly located in the more liberal region of the similarity threshold. This indicates that, at least for those parameter values, clusters tend to straddle over lexical categories slightly more in ADS than IDS, or, put differently, that it is more often the case that two targets are classified into a single motif in ADS than IDS. This is unexpected but interesting, because the target words studied in the present corpus were not necessarily very similar to one another (see Appendix A).

Finally, as we expected, target segments were more often split into multiple clusters (reflected in a lower *collocation* score) in IDS than ADS. This corroborates our suspicion that the acoustic implementation of words is more variable in IDS, which also explains why differences are particularly clear for conservative parameter values. Nonetheless, the difference across registers was small, only about 1%.

We provided results for all the values of the similarity threshold because we believe it can yield some insight into infants' performance at different points of development, since younger infants (7.5-month-olds) have been found to be more conservative than older ones (9-12 months of age, (Singh et al., 2012)). Our computational model suggests that, if they behave like our model learner, younger infants should both fail to recognize words across diverse instantiations (cf. our recall results) and postulate too many lexical entries (cf. our collocation results). In other words, our computational model predicts that signal-related effects of register on word segmentation performance will be greatest, with an IDS disadvantage, for younger rather than older infants. It is possible that our purity results suggest that the IDS disadvantage be reversed in these older ages, who are supposedly more liberal in their acoustic matching. Extant infant work showing IDS advantages has looked at 7- and 8-month-olds (Singh et al., 2009; Thiessen et al., 2005), so future work should test these specific predictions in even younger infants.

Together with (Kirchhoff and Schimmel, 2005), which relied on hand-segmented words, the present results are relevant to the interpretation of infant performance in word segmentation tasks which compare IDS and ADS. Specifically, since neither classification of segmented words, nor motif discovery, are overall more successful in IDS than ADS, then it follows that infants' improved segmentation performance for IDS is not due to words being physically (or segmentally) easier to find or classify in IDS than ADS. Instead, there must be something else in the spoken signal that boosts infant performance in IDS. This other fac-

tor may be attention/arousal: Perhaps infants attend more to IDS stimuli (which is clear in preferential studies (Dunst et al., 2012)). Alternatively, infant performance may reflect a more complex cognitive bias, for instance if they apply different learning strategies when prompted by IDS (as proposed in the Natural Pedagogy framework (Csibra and Gergely, 2009)). Similar explanations have been put forward to explain improved performance for boosts in word-meaning mapping tasks in IDS over ADS (Graf Estes and Hurley, 2013).

There are many open questions that need to be revisited in other research, such as to what extent motif discovery reflects meaningful features of the algorithm that real infant utilize during word segmentation in the lab and in the world, the integration of multimodal information, or the extent to which specific predictions made from MODIS versus competing models are born out by infant data.

5 Conclusions

In this paper, we focused mainly on one documented difference between IDS and ADS, namely phonetic variability, and considered two lexical parameters, repetition and burstiness. We found that performance was affected by register, with an overall trend for lower performance in IDS than ADS when three metrics was considered. The impact of register was greatest when our model learner, which relies on acoustic matching, was conservative. We believe this result suggests that register differences relate to the differences in phonetic variability that have been separately documented, although additional analyses (for instance using regressions to explore individual variation) are needed to confirm this hypothesis. Furthermore, it would be important to repeat these analyses with other corpora, particularly those gathered at home, which may vary more naturally along other dimensions we also intended to explore, such as repetition and burstiness.

Additionally, other models are needed to gain a more holistic understanding of how register features affect learners' performance, since we only explored effects of a few IDS characteristics, and others remain unexplored (see Table 1). For example, IDS contains shorter utterances (Albin and Echols, 1996; Aslin et al., 1996) and is produced with more exaggerated prosodic edge marking than ADS (Fernald and Mazzie, 1991; Kondaurova and Bergeson, 2011). If there are shorter utterances in IDS it means that more words will occur at utterance edges which are, as mentioned above, also marked with increased acoustic salience in IDS. These utterance edges have been shown to be hot-spots for word segmentation (Seidl and Johnson, 2006), so much so that even infants as young as 4 months are able to find words at utterance edges using this strategy (Johnson et al., 2014). Recent work on speech-based spoken-term discovery has shown that the integration of prosodic boundary information in such a system improves segmentation performance (Ludusan et al., 2014). Since this was found in corpora containing ADS, we would like to explore whether the prosodic structure would give a boost in performance when IDS is given as input to MODIS, compared to when ADS is employed.

Acknowledgments

This work was supported by the European Research Council (grant ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (grants ANR-14-CE30-0003 MechELex, ANR-2010-BLAN-1901-1 BOOTLANG, ANR-10-IDEX-0001-02 PSL*, and ANR-10-LABX-0087 IEC), the Fondation de France, and NSF grant number 0843959. Statement of contributions: AS collected the corpus and oversaw coding; BL carried out the MODIS experiments; AC and BL carried out the statistical analyses, with input from all authors; AC and BL wrote a first draft; all authors contributed to the design and writing.

Appendix A. List of target words: *baboon, bacon, bamboo, basil, bassinet, beetle, Benji, best-in-show, dancer, dancing, daycare, decker, dictionary, disney, pansy, paper, pedal, peekaboo, pegboard, pencil, pendant, pepsi, pesto, picnic, piglet, shopping, tambourine, tapioca, tassel, teakettle, teaspoon, teddy* and *tender*.

References

Drema Albin and Catharine Echols. 1996. Stressed and word-final syllables in infant-directed speech. *Infant Behavior and Development*, 19:401418.

Richard Aslin, Julide Woodward, Nicholas LaMendola, and Thomas Bever. 1996. Models of word segmentation in fluent maternal speech to infants. In J. Morgan and K. Demuth, editors, *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. Lawrence Earlbaum Associates, Mahwah, NJ.

Heather Bortfeld, James Morgan, Roberta Golinkoff, and Karen Rathbun. 2005. Mommy and me: Familiar names help launch babies into speech stream segmentation. *Psychological Science*, 16:298–304.

Laurence Catanese, Nathan Souviraà-Labastie, Bingqing Qu, Sebastien Campion, Guillaume Gravier, Emmanuel Vincent, and Frédéric Bimbot. 2013. MODIS: an audio motif discovery software. In *Proceedings of Interspeech*.

Alejandrina Cristia and Amanda Seidl. 2014. The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language*, 41:913–934.

Alejandrina Cristia. 2013. Input to language: The phonetics and perception of infant-directed speech. *Language and Linguistics Compass*, 7:157–170.

Gergely Csibra and György Gergely. 2009. Natural pedagogy. *Trends in cognitive sciences*, 13(4):148–153.

Robert Daland. 2013. Variation in the input: a case study of manner class frequencies. *Journal of child language*, 40(5):1091–1122.

Carl Dunst, Ellen Gorman, and Deborah Hamby. 2012. Preference for infant-directed speech in preverbal young children. *Center for Early Literacy Learning*, 5(1).

Anne Fernald and Claudia Mazzie. 1991. Prosody and focus in speech to infants and adults. *Developmental Psychology*, 27:209–221.

Anne Fernald and Hiromi Morikawa. 1993. Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child development*, 64(3):637–656.

Katharine Graf Estes and Karinna Hurley. 2013. Infant-directed prosody helps infants map sounds to meanings. *Infancy*, 18(5):797–824.

Derek Houston and Peter Jusczyk. 2003. Infants' long-term memory for the sound patterns of words and voices. *Journal of Experimental Psychology: Human Perception and Performance*, 29(6):1143.

Elizabeth Johnson, Amanda Seidl, and Michael Tyler. 2014. The edge factor in early word segmentation: utterance-level prosody enables word form extraction by 6-month-olds. *PLoS ONE*, 9(1):e83546.

Katrin Kirchhoff and Steven Schimmel. 2005. Statistical properties of infant-directed versus adult-directed speech: insights from speech recognition. *The Journal of the Acoustical Society of America*, 117(4):2238–2246.

Maria Kondaurova and Tonya Bergeson. 2011. The effects of age and infant hearing status on maternal use of prosodic cues for clause boundaries in speech. *Journal of Speech Language and Hearing Research*, 54:740–754.

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

Bogdan Ludusan, Guillaume Gravier, and Emmanuel Dupoux. 2014. Incorporating prosodic boundaries in unsupervised term discovery. In *Proceedings of Speech Prosody*, pages 939–943.

Andrew Martin, Thomas Schatz, Maarten Versteegh, Kouki Miyazawa, Reiko Mazuka, Emmanuel Dupoux, and Alejandrina Cristia. 2015. Mothers speak less clearly to infants than to adults a comprehensive test of the hyperarticulation hypothesis. *Psychological science*, 26(3):341–347.

Bob McMurray, Kristine Kovack-Lesh, Dresden Goodwin, and William McEchron. 2013. Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? *Cognition*, 129:362–378.

Gerald McRoberts, Colleen McDonough, and Laura Lakusta. 2009. The role of verbal repetition in the development of infant speech preferences from 4 to 14 months of age. *Infancy*, 14(2):162–194.

Amanda Seidl and Elizabeth Johnson. 2006. Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science*, 9 (6):565–573.

Leher Singh, Sarah Nestor, Chandni Parikh, and Ashley Yull. 2009. Influences of infant-directed speech on early word recognition. *Infancy*, 14(6):654–666.

Leher Singh, Steven Reznick, and Liang Xuehua. 2012. Infant word segmentation and childhood vocabulary development: A longitudinal analysis. *Developmental Science*, 15:482–495.

Erik Thiessen, Emily Hill, and Jenny Saffran. 2005. Infant-directed speech facilitates word segmentation. *Infancy*, 7:53–71.

Sho Tsuji and Alejandrina Cristia. 2014. Perceptual attunement in vowels: A meta-analysis. *Developmental psychobiology*, 56(2):179–191.

Mark J van der Laan, Sandrine Dudoit, and Katherine S Pollard. 2004. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–35.

Yuanyuan Wang, Amanda Seidl, and Alejandrina Cristia. 2015. Acoustic-phonetic differences between infant-and adult-directed speech: the role of stress and utterance position. *Journal of child language*, 42(4):821–842.