# Shared Understanding and Idiosyncratic Expression in Early Vocabularies

Mayor, Julien; Plunkett, Kim

Developmental Science

Developmental Science (2014), pp 1–12 DOI: 10.1111/desc.12130

PAPER

# Shared understanding and idiosyncratic expression in early vocabularies

Julien Mayor[1] and Kim Plunkett[2]

1. *FPSE, University of Geneva, Switzerland*
2. *Department of Experimental Psychology, University of Oxford, UK*

## Abstract

*To what extent do toddlers have shared vocabularies? We examined CDI data collected from 14,607 infants and toddlers in five countries and measured the amount of variability between individual lexicons during development for both comprehension and production. Early lexicons are highly overlapping. However, beyond 100 words, toddlers share more words with other toddlers in comprehension than in production, even when matched for lexicon sizes. This finding points to a structural difference in early comprehension and production: Toddlers are generalists in comprehension but develop a unique, expressive voice. Variability in production decreases after two years of age, suggesting convergence to a common expressive core vocabulary. We discuss potential exogenous and endogenous contributions to the inverted U-shaped development observed in young children's expressive lexical variability.*

## Introduction

Knowing words that other people know is crucial to achieving successful communication. Utterances achieve their impact when there is an alignment between the speaker's expressive vocabulary and the listener's receptive vocabulary. Yet interactions with a restricted number of people early in life may lead infants and young children to develop idiosyncratic vocabulary knowledge. Several studies have highlighted the environmental contribution to vocabulary development (Hart & Risley, 1995; Huttenlocher, 1991; Bradley, Caldwell & Rock, 1988) but evidence is scant as to whether environmental factors influence the composition of comprehension and production vocabularies evenly. Furthermore, relatively little is known about the amount of variability between the composition of individual lexicons, or the role played by environmental factors in modulating this variability. Environmental factors may shape comprehension and production differently. A lag between comprehension and production in early vocabulary development is well attested (Goldin-Meadow, Seligman & Gelman, 1976; Fenson, Dale, Reznick, Bates, Thal & Pethick, 1994) and the proportions of word types used in production differs from that found in comprehension (Benedict, 1979). Pragmatic constraints on communication may create a general pressure towards shared receptive vocabularies while permitting a greater degree of variability in expressive vocabularies. However, to date, it remains unclear whether the variability in the composition of receptive vocabularies differs from the variability in the words infants produce or how this variability evolves during the first few years of life.

Quantitative analyses of vocabulary checklists reveal that comprehension is more stable than production across development: early receptive vocabulary scores are more predictive of later scores than are early expressive scores (Fenson *et al.*, 1994). Recent statistical modelling based on these parental checklists have attempted to provide more accurate estimates of receptive and vocabulary sizes than raw checklist scores permit (Mayor & Plunkett, 2011). These analyses confirm the existence of a substantial developmental lag between comprehension and production reported in earlier studies. However, the results do not indicate the extent to which children share specific vocabulary items,

Address for correspondence: Julien Mayor, FPSE, University of Geneva, Boulevard du Pont d'Arve 40, Geneva 1211, Switzerland; e-mail: julien.mayor@unige.ch

neither in production nor comprehension, despite the fact that the database comprises parents' judgements about their offspring's knowledge of individual words. In the current paper, we extend Mayor and Plunkett's (2011) analyses to estimate the overlap, and hence variability, in the expressive and receptive vocabularies of infants and young children.

## Method

We analyse databases compiled from vocabulary checklists collected in five different countries: The MacArthur-Bates CDI, Words and Gestures (MCDI-WG) and Words and Sentences (MCDI-WS) (Fenson, Marchman, Thal, Dale & Reznick, 2007; Dale & Fenson, 1996), and what we will refer to as 'the Danish CDI' (Bleses, Vach, Wehberg, Kristensen & Madsen, 2007), 'the Norwegian CDI' (Kristoffersen, Simonsen, Eieslan, & Henriksen, 2012), the 'German CDI' (Szagun, Stumper, & Schramm, 2009), all accessed via CLEX (Jorgensen, Dale, Bleses, & Fenson, 2010), and the 'Oxford CDI' (Hamilton, Plunkett, & Schafer, 2000). Details of the databases are provided in Table 1.

### *Sigmoidal analysis*

We apply the method developed for the MacArthur-Bates CDI (Mayor & Plunkett, 2011). Only age groups consisting of at least 50 infants or toddlers for each country are included in the analyses. The method was originally used to provide an estimate of vocabulary size from CDI scores, but can also be used to measure the amount of overlap between vocabularies. The advantages of this method are that (1) no detailed knowledge of individual vocabularies is needed other than the percentage of infants/toddlers knowing a given word and (2) the method can be applied to a wide range of ages and vocabulary scores. The method uses as input the proportion of infants, at a given age, who understand (or produce) a specific word and yields, amongst other measures, an index of vocabulary variability called *b*. Note that parameter *b* is a measure of variability within a group of children, and cannot be applied or interpreted for an individual. As illustrated in Figure 1, a low value for *b* corresponds to low lexical variability ($b = 0$ providing the boundary condition where individual lexicons overlap perfectly) and a high value for *b* corresponds to a lot of variability. In other words, many infants possess a large number of idiosyncratic words. A brief description of the rationale behind the calculation of parameter *b* is provided in Appendix A, and a detailed

**Table 1** *Names and sizes of the databases, and corresponding number of infants for each age group. Superscripts indicate whether comprehension is assessed (c), production (p) or both (c,p). Only age groups consisting of at least 50 infants or toddlers are included in the results*

| Name | Size | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MCDI-WG$^{c,p}$ | 396 | 68 | 89 | 136 | 95 | 157 | 116 | 111 | 94 | 84 | 73 | 66 | | | | | | | | | | | | | | | | | | |
| MCDI-WS$^{p}$ | 680 | | | | | | | | | 74 | 81 | 103 | 98 | 117 | 95 | 90 | 104 | 135 | 107 | 100 | 113 | 84 | 80 | 80 | | | | | | |
| Danish-WG$^{c,p}$ | 410 | 174 | 196 | 182 | 200 | 188 | 216 | 178 | 203 | 175 | 170 | 198 | 156 | 162 | | | | | | | | | | | | | | | | |
| Danish-WS$^{p}$ | 725 | | | | | | | | | 184 | 176 | 160 | 205 | 196 | 236 | 235 | 190 | 216 | 138 | 206 | 176 | 190 | 182 | 173 | 175 | 143 | 151 | 143 | 113 | 126 |
| German$^{p}$ | 588 | | | | | | | | | | | 85 | 79 | 96 | 77 | 95 | 96 | 103 | 96 | 92 | 89 | 93 | 89 | 93 | | | | | | |
| Norwegian-WG$^{c,p}$ | 396 | 1 | | | 2 | | 4 | 3 | 5 | 12 | 11 | 22 | 21 | 38 | 43 | | | | | | | | | | | | | | | |
| Norwegian-WS$^{p}$ | 731 | | | | | | | | | 135 | 169 | 182 | 205 | 189 | 270 | 260 | 211 | 195 | 227 | 207 | 188 | 187 | 196 | 216 | 210 | 193 | 211 | 183 | 211 | 170 |
| Oxford CDI$^{c,p}$ | 416 | | | | 1 | 25 | 28 | 4 | 27 | 57 | 13 | 95 | 118 | 7 | 19 | 68 | 17 | 72 | 103 | 7 | 4 | 2 | | | 78 | | | | | |

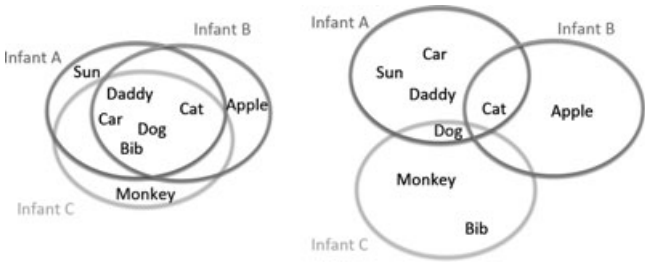


**Figure 1** *Illustration of lexical variability. Left panel: lexical variability is low (parameter b is low); many words are shared by several infants. Right panel: lexical variability is high (b is large); few words are shared amongst infants.*

mathematical description is given in Mayor and Plunkett (2011, p. 772).

### *Direct analysis*

When dealing with a more detailed database that includes the words known by each child, as is the case with the Oxford CDI, we can obtain a direct measure of lexical variability. This direct measure is computed by calculating the mean Euclidean distance between individual vocabularies and the mean vocabulary, where each word is either understood/produced (coded as 1 in a vector containing all words on the CDI) or not understood/not produced (coded as 0). As this metric is heavily dependent on the total number of words known on the CDI, these mean Euclidean distances are then normalized by the underlying binomial distribution, produced by measuring the Euclidean distance when vector values are drawn at random. The Normalized Euclidean Distance (NED hereafter) is computed according to the following equation:

$$NED = \frac{\sum_{j=1}^{N}\sum_{i=1}^{W}(x_{ij}-p_i)^2}{\sum_{j=1}^{N}\sum_{i=1}^{W}(y_{ij}-q_i)^2}$$

where $W$ refers to the number of words on the CDI, $N$ the number of infants, $x_{ij}$ is equal to 1 if the word $i$ is understood/produced by infant $j$ and 0 otherwise. $p_i$ corresponds to the fraction of infants that understand/produce word $i$. $y_{ij}$ corresponds to the random assignment to word $i$ in run $j$, where 1 and 0 are assigned randomly so that mean vocabularies match and $q_i$ corresponds to the fraction of runs for which word $i$ is understood/produced. Unfortunately, this calculation cannot be performed on databases other than the Oxford CDI, since only the percentage of infants or toddlers knowing each word can be accessed, the detailed vocabulary of each individual infant or toddler being unavailable.

Correlations between both measures of lexical variability by age group for the Oxford CDI, obtained via the sigmoidal analysis and via NED, confirm a high level of agreement between both methods ($r(5) = 0.83$, $p < .001$). The use of the sigmoidal analyses for databases that do not provide details about individual vocabularies is, therefore, fully justified and can be expected to be reliable and accurate.

## Results

### *Asymmetric development in the first two years of life*

#### Age effects

First, we evaluate the variability between individual vocabularies for different ages and different countries, for both comprehension and production. The top panel of Figure 2 depicts a measure of lexical variability

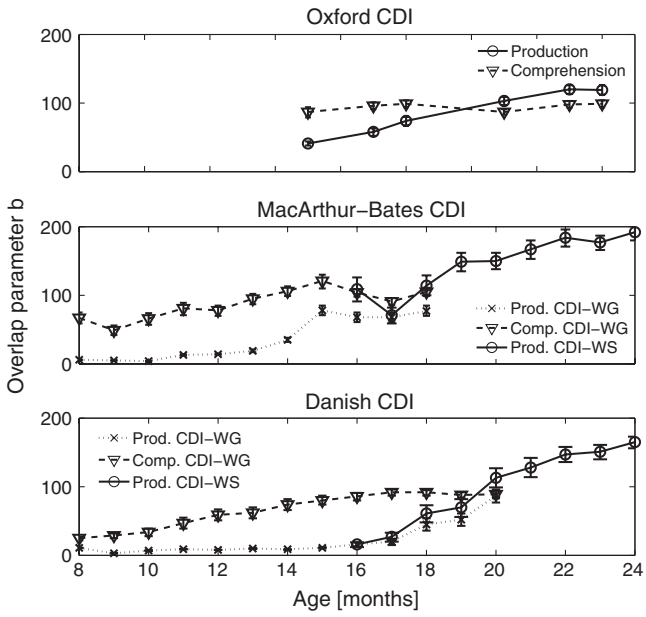


**Figure 2** *Amount of variability between individual vocabularies in comprehension and in production: a low value for parameter b corresponds to low variability whereas a high value means only few words are shared across the population of children. Note that the proportion of shared words is only matched in comprehension and production at around 20 months of age. For any other age, differences in the degree of overlap between individual lexicons are present in comprehension and production. Error bars represent the interval around the optimal value b found by regression, for which the residual error between the data and the fit do not exceed by more than 5% the error associated with the best fit.*

(parameter *b*) for different ages, for infants assessed with the Oxford CDI, from 15 to 24 months of age. This database is valuable because it is the only checklist of words for which *both* comprehension and production are assessed for each infant beyond 20 months of age, and spans a period of development during which we expect to observe rapid changes in production and where a comparison to comprehension is crucial. Correlational analyses of parameter *b* with age between 15 and 24 months reveal a clear developmental trend for variability in production (lexical diversity increases with age; $r(4) = 0.98$, $p < .001$), whereas lexical variability in comprehension does not correlate with age throughout the age range under consideration ($r(4) = -0.33$, $p = .53$). The top panel of Figure 2 also shows that infants assessed with the Oxford CDI have less overlapping lexicons in comprehension than production prior to 20 months of age, whereas there is less overlap in production beyond this age.

A two-way ANOVA (age * mode [comprehension or production]) on measures of lexical diversity obtained by the direct method (Normalized Euclidean Distances) confirmed this observation, revealing a main effect of age ($F(1, 5) = 2.92$, $p = .012$), no main effect of mode ($F(1, 1) = 1.33$, $p = .24$) and a strong interaction between age and mode ($F(2, 5) = 6.69$, $p < .001$).

The middle panel of Figure 2 depicts lexical variability for the MacArthur-Bates CDI in comprehension (CDI-Words and Gestures (CDI-WG, dashed line), from 8 to 18 months of age) and in production (CDI-WG from 8 to 18 months of age (dotted line) and CDI-Words and Sentences (CDI-WS, solid line) from 16 to 30 months of age, data shown up to 25 months of age). The comparison between comprehension and production before 18 months of age reveals that vocabularies possess less variability in production than in comprehension.[1] A similar finding is depicted in the lower panel of Figure 2 for infants and toddlers assessed with the Danish CDI.[2] Paired sample *t*-tests confirmed that lexical diversity is larger in comprehension than in production when matched for age on the MacArthur-Bates CDI ($t(10) = 9.51$, $p < .001$) and on the Danish CDI ($t(12) = 7.04$, $p < .001$). Between 16 m and 24 m, lexical variability in production correlated with age (MacArthur-Bates CDI:$r(7) = 0.92$, $p < .001$ and Danish CDI: $r(7) = 0.98$, $p < .001$), thus confirming the trend observed on the Oxford CDI.

These findings indicate that the overlap between infants' expressive lexicons is not the same as the overlap in their receptive lexicons and that until about 18–20 months of age, infants share more words in their productive vocabulary than in comprehension. Greater overlap in early expressive lexicons is readily understood in terms of the general lag of production behind comprehension: All else equal, fewer words in production yields less variability. However, this explanation does not hold for toddlers older than 20 months of age, where expressive lexicons are more variable than receptive lexicons, despite the continued asymmetry in size between expressive and receptive vocabularies (Fenson *et al.*, 1994). In order to compensate for the numerical imbalance between production and comprehension at any given age, we next perform an analysis of lexical variability in comprehension and production as a function of vocabulary size.

## Vocabulary effects

The top panel of Figure 3 depicts lexical variability assessed on the Oxford CDI as a function of the mean vocabulary score,[3] and confirms the trends observed for the correlation with age: lexical variability is independent of mean vocabulary score in comprehension ($r(4) = -0.21$, $p = .69$) whereas a correlation of lexical variability with mean vocabulary score is observed in production ($r(4) = 0.95$, $p = .004$).[4]

The middle panel (MacArthur-Bates CDI) and lower panel (Danish CDI) of Figure 3 also display a comparison of lexical variability between individual vocabularies in both production and comprehension. Assessment at younger ages reveals that variabilities in comprehension and production follow a similar trajectory as new words enter the infants' lexicons until reaching about 100 words on the CDIs. Thereafter, the amount of variability remains approximately constant in comprehension

[1]Note that Mayor and Plunkett (2011) showed that a regression of parameter *b* with age revealed that the hypothesis that *b* is independent of age cannot be rejected, in comprehension, for infants older than 11 months ($p = .23$) nor in production for children older than 20 months ($p = .15$).

[2]Other databases (the German CDI and the Norwegian CDI) were omitted from the present comparison between comprehension and production either because the database did not include data for comprehension or because the number of infants were insufficient to derive an accurate estimate of lexical variability.

[3]As we do not have access to the individual lexicon sizes for all databases, vocabulary plots and analyses are made to the mean vocabulary corresponding to the age under consideration.

[4]Note that these correlations are closely related to previous analyses since mean vocabulary sizes and age are correlated. As a consequence, and for the sake of readability, correlation analyses on the MacArthur-Bates and Danish CDIs will not be reported (see previous section for age effects).

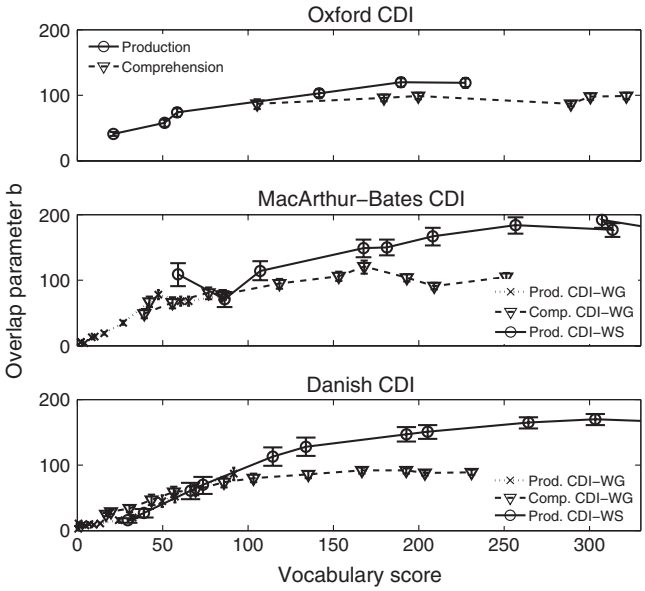


**Figure 3** *Amount of lexical variability as a function of mean vocabulary. Small lexicons exhibit substantial overlap in both comprehension and production. As the number of words increases, the amount of variability increases. However, the amount of lexical variability in production continues to increase beyond the point when variability has stabilised in comprehension. Children become general comprehenders whereas they find their unique way of speaking.*

whereas toddlers show less overlap (more variability) with other toddlers in production.

This finding is substantiated further by direct measurement of lexical variability on the Oxford CDI, where Normalized Euclidean Distance (NED) between *individual* vocabularies can be computed. These direct measures confirm that lexical variability is correlated with mean vocabulary score in production ($r(4) = 0.97$, $p = .002$), but not in comprehension ($r(4) = 0.56$, $p = .22$), thereby providing evidence that the asymmetry between comprehension and production vocabularies is not an artifact of the method used for assessing overlap between individual vocabularies.

Furthermore, when the two types of lexicons, expressive and receptive, are compared after controlling for size, more variability is observed in production than in comprehension during the second half of the second year. For example, at 23 months of age, the *mean* productive vocabulary size (189.5) on the Oxford CDI matches approximately receptive vocabulary at 17 m (171.2 words) and 18 m (195.3 words). Two-sample *t*-tests were carried out and confirmed that lexical variability, assessed by direct measurements of lexical diversity (using the direct measure; the Normalized Euclidean Distance), was higher in production than in comprehension in both cases ($t(165) = 3.10$, $p = .002$ and $t(188) = 3.52$, $p < .001$).

Production does not just lag behind comprehension, it is *more idiosyncratic* than comprehension. Infants begin their apprenticeship with language using a vocabulary common to other infants, both in production and comprehension. As vocabulary grows beyond approximately 100 words, toddlers use an increasing number of words that are not produced by other toddlers, whereas their receptive vocabularies remain aligned with other toddlers despite substantial increases in vocabulary size.

### *Is lexical diversity distributed homogeneously among children?*

Access to a detailed database such as the Oxford CDI allows us to carry out further analyses which cannot be performed on the other databases that do not report the detailed vocabulary of each individual infant or toddler. However, the measure of lexical diversity, as assessed with Mayor and Plunkett's (2011) method, fails to distinguish between cases in which lexical diversity is regularly distributed amongst all infants from cases in which distinct subgroups of infants differ dramatically from each other in terms of lexical composition, despite averaging to the same level of lexical variability overall.[5] In order to evaluate whether lexical diversity is distributed homogeneously among our sample population, we applied several clustering methods to identify the *optimal* (or natural) number of clusters of infants in the Oxford CDI. Consistent results across each test would point to a sub-structure in lexical diversity, whereas any lack of agreement between a battery of clustering algorithms would suggest a relative homogeneity in lexical diversity in each age group. Furthermore, the application of these algorithms to artificial vocabulary datasets that are known to be either structured or randomly generated in their distribution of lexical diversity provides a useful comparison for evaluating the relative homogeneity in the distribution of lexical diversity across the Oxford CDI.

Table 2 reports the natural number of clusters at 21 m of age in comprehension and in production, as obtained from the following algorithms; Silhouette (Rousseeuw & Kaufman, 1990), Davies-Boudin (Davies & Bouldin, 1979), Krzanowski-Lai (Krzanowski & Lai, 1988), Hartigan (Hartigan, 1975) and Dunn (Dunn, 1973). There is a clear lack of agreement in the number of clusters found by the different clustering methods.

[5]We thank an anonymous reviewer for raising this issue.

**Table 2** *Natural number of clusters of 21-month-old infants, classified in terms of lexical overlap. Lack of agreement between the different clustering methods suggests that lexical diversity is relatively homogeneously distributed among infants, for both comprehension and production. Clustering results for artificially randomized and structured vocabularies are also reported for baseline comparisons*

| | Silhouette | Davies-Bouldin | Krzanowski-Lai | Hartigan | Dunn |
|---|---|---|---|---|---|
| Comprehension | 2 | 4 | 2 | 1 | 3 |
| Production | 2 | 3 | 3 | 2 | 4 |
| Randomized Comp. | 2 | 4 | 3 | 1 | 3 |
| Randomized Prod. | 2 | 5 | 4 | 1 | 3 |
| Artificial data (3 clusters) | 3 | 3 | 3 | 3 | 3 |

However, when these clustering techniques are applied to artificial vocabulary data containing three clusters of lexical diversity, all methods converge on the same result, thereby confirming that if a clustering existed, all of them would find it. When vocabulary data are randomized, any clustering disappears and, again, clustering methods produce different results from each other, confirming the absence of subgroups in the data. This pattern of results suggests that lexical diversity is relatively homogeneously distributed among the infants reported in the Oxford CDI. Consequently, these clustering simulations indicate that the sigmoidal analysis as a tool for investigating lexical diversity during development is appropriate for this database.

### *Inverted U-shaped pattern in the overlap of vocabularies in production*

The proportion of words shared by a population of infants or toddlers is different in comprehension and production. For lexicons larger than 100 words (as assessed by CDIs), lexical variability in comprehension remains stable, whereas in production variability increases monotonically with vocabulary size, so that individual expressive lexicons share proportionally fewer words. The previous analysis focused on toddlers up to 24 months of age. We now include older toddlers to evaluate if this trend is observed after 2 years of age. The analysis focuses exclusively on expressive vocabularies as comprehension is not typically assessed with CDIs in these older age groups.

We measure lexical variability in production by analysing CDIs for toddlers using the McArthur-Bates CDI, the Danish CDI, the Norwegian CDI and the German CDI, over an age range of 16 to 36 months. As shown in the left panel of Figure 4, the additional corpora confirm the increasing variability of productive vocabulary (an increase in parameter $b$) throughout the vocabulary spurt (from 18 to 24 months of age).

After their second birthday, toddlers from all four language corpora show a *decrease* in the amount of variability between individual vocabularies (see Figure 4, left panel). This inverted U-shaped developmental profile of variability in production reveals that infants start off with a common core of vocabulary. With age and a growing vocabulary, they become increasingly idiosyncratic in their choice of words. Around their second birthday, this trend reverses so that succeeding months herald an increased overlap in the word choices comprising toddlers' utterances. Notably, this reversal takes place around the same age for toddlers of all the languages surveyed. The right panel of Figure 4 depicts lexical diversity as a function of vocabulary score. The same inverted U-shaped profile in variability is observed; lexical diversity increases in production until toddlers possess about 300–400 words in their lexicons, and then decreases as more words enter their productive lexicons. Pairwise correlational analyses between the language corpora, taken two-by-two, are all highly significant (all $r$s > .91 and all $p$s < .001), confirming that this effect takes place for all languages tested. Furthermore, a paired $t$-test between measures of lexical diversity at 24 months of age and at 30 months of age[6] confirm the significance of the reduction in lexical diversity after two years of age ($t(3) = 3.83$, $p = .03$). Additional cross-validation tests and comparisons of the sigmoidal method with the direct method using Normalized Euclidean Distance, described in Appendix B, confirm the robustness of the effect, and indicate that the observed inverted U-shaped curve in lexical diversity is unlikely to be an artifact of the analyses applied, or due to the limited size of the vocabulary checklists. These results indicate that by the age of 3 children have reverted to establishing a common core in their expressive vocabularies. Potential explanations of the underlying phenomena are discussed in the next section.

[6]Thirty months of age is the oldest age for which at least four independent databases assessed toddlers.

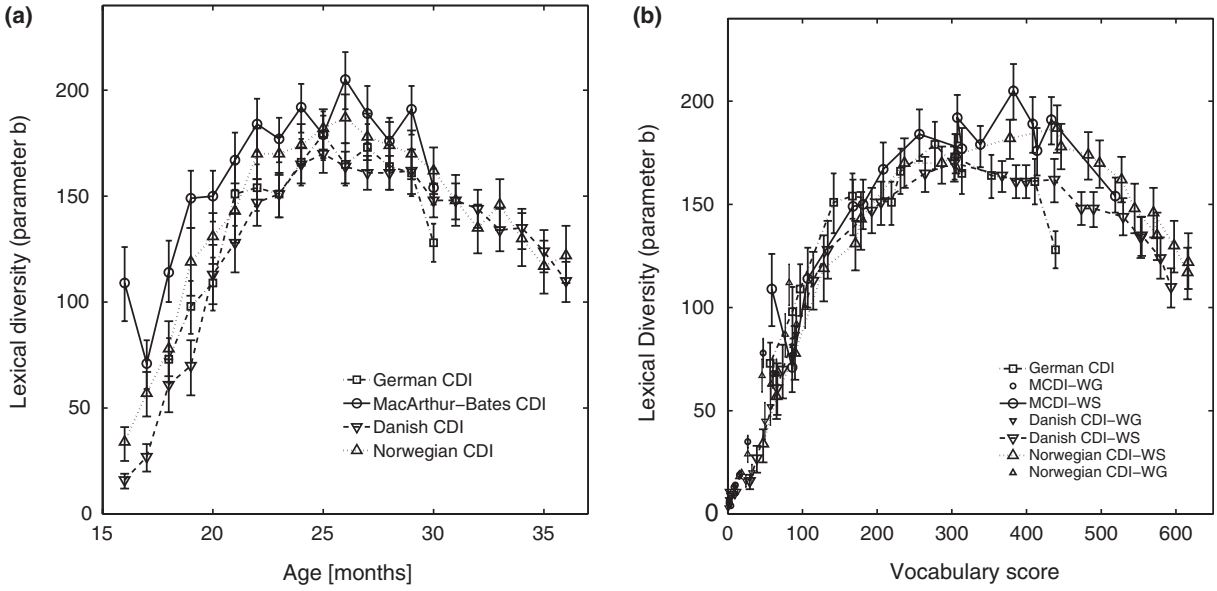


**Figure 4** *Left: Lexical variability in production as a function of age. Young speakers use an increasing proportion of idiosyncratic words in the first 2 years of life, until approximately 2 years of age, when lexical variability between individual productive lexicons decreases to more uniform levels. Right: Lexical variability in production as a function of vocabulary score. Variability increases until infants have reached about 300–400 words in their productive lexicon. Lexical diversity decreases as vocabulary size keeps on increasing.*

## Discussion

Comparisons between the expressive vocabularies of infants and toddlers, and between their receptive vocabularies, indexed by parental report, indicate that early in development their lexicons share a common core. As their vocabularies expand, lexical variation between infants increases as a result of their personal experience and the inherent statistical variability associated with larger numbers. However, beyond a vocabulary size of 100 words, the overlap of receptive vocabularies between infants remains stable, whereas the variability in expressive vocabularies continues to increase, even when the two vocabularies are matched for size.

Stability in overlap between receptive vocabularies is not an artefact of the checklist method of assessing vocabulary knowledge. At first sight, a limited sample of items might have introduced a ceiling effect, thus artificially reducing measures of lexical diversity in comprehension. We can, however, rule out this explanation, since stability in the variability of comprehension between infants is achieved well before ceiling effects can have an impact, i.e. from 100 words onwards with a possible total score of 416 for the Oxford CDI. Moreover, variability in production, assessed over the same set of items, continues to increase despite numbers lagging behind comprehension. When receptive and expressive vocabularies are matched in size, the asymmetry in overlap remains apparent throughout the period spanning the vocabulary expansion from 100 to 300 words (see Figure 3), further corroborating the claim that the observed differences in variability between comprehension and production are not an artefact of lexicon size.

The lower variability in infants' comprehension vocabularies compared to their productive vocabularies offers concrete evidence for the supposition that pragmatic constraints exert a greater pressure on shared understanding than on expressive repertoires. Our findings are consistent with the view that there is a greater communicative need for infants and toddlers to ensure that they understand what is said by others than to guarantee that they express messages composed of the same words produced by others. It is clear that production is not just a delayed version of comprehension. These analyses demonstrate that the well-known quantitative asymmetry between early comprehension and production is also manifest as a structural difference. Apparently, the pressure to achieve effective communication does not apply evenly to the development of comprehension and production.

Individuals possess unique ways of expressing themselves. Simultaneously, they need to understand many speakers, and to be understood by many interlocutors. Adults possess a much larger lexicon than infants. As

long as infants' primary interlocutors are adults, infants will be understood whatever words they use. There is no incentive for the infant to adapt production to the interlocutor and production may become more idiosyncratic. In comprehension, however, infants need to be able to understand different caregivers; infants will become general comprehenders.

Once infants start communicating with other infants or their young peers, they can no longer assume that their interlocutor will know all the words they use. Communicative pressure will enforce the requirement that a sufficient subset of the words used in production by an infant will be part of other infants' receptive vocabularies. We expect, then, to witness a convergence between the variability of production and comprehension; lexical diversity between productive vocabularies should decrease. This is exactly what is observed when analysing lexical diversity in production after two years of age (see Figure 4). In all the four datasets reported for ages over 24 months (MacArthur-Bates CDI-WG, Danish CDI, German CDI, and Norwegian CDI), variability in expressive vocabulary *decreases* in the period extending from 25–30/36 months of age, apparently converging with the levels of variability observed in comprehension.[7] This is an age during which we expect infants to experience a wider range of environments beyond their home as they develop cognitively and motorically, and in western cultures at least, are likely to be introduced to nursery school. For example, about 50% of Australian 1-year-olds whose mother is employed go to formal child care – a percentage that rises to 70% by two years of age. Australian infants whose mother is not employed follow a similar pattern, rising from 21% to 39% over the same time period (Baxter, 2011, Figures 1 and 2).

In Nordic countries, a similar trend is observed even though more infants attend daycare (rising from 58.4% at 12 m to 82.3% at 24 m, Nordic Council of Ministers, 2010). In the United States, attendance is somehow lower, rising from 18.3% between 1 and 2 years of age, to 35.4% between 2 and 3 years of age (US. Census Bureau, 2010). Furthermore, Eckerman, Whatley and Kutz (1975) report that by their second birthday, young children's social play exceeds their solitary play and their social partner is more often a child of the same age, rather than an adult. (See Howes, 1985, 1988 for further discussion of the emergence of social play early in the third year.)

Infants' developing socio-cognitive understanding may also make them increasingly sensitive to the limited vocabulary of their peers. For example, improved performance on the mirror self-recognition test around the age of 2 (Amsterdam, 1972) indicates an increased sensitivity to the *self–other* distinction that would be important for an appreciation that other toddlers may differ in their vocabulary repertoire. Similarly, 2-year olds begin to demonstrate an understanding of their parent's knowledge (or lack thereof) in gestural communication tasks (O'Neill, 1996). This sensitivity to another's knowledge state may well generalize to their infant interlocutors, permitting an appropriate adjustment in their dialogue.[8]

Of course, the limited size of the lexicon of some interlocutors need not be the only factor playing a role in shaping the developing lexicon after 2 years of age. The number and types of interlocutors is also important. As they grow older, children communicate with more people and are thus afforded the opportunity to fine-tune their utterances to a broader range of individuals, adding further communicative pressure to standardize their utterances to a societally determined norm.

Furthermore, if infants are raised in environments that are similar to each other (in terms of frequently named items, words used, etc.) such as in daycare centres, we would expect the shared environment to constrain lexical diversity; the correlation between individual experience would increase. Thus, the decrease in the amount of lexical diversity after 2 years of age may also be linked to the educational role daycares provide, offering another example of the role daycares have on language development in children (see also McCartney, 1984).

In addition, around their second birthday, toddlers start forming multiple-word utterances (Miller & Chapman, 1981). It has been suggested that the rate of lexical acquisition may be temporarily subdued during this period (see for example, Roy, Frank & Roy, 2009). Infants may settle into a comfort zone and use more familiar words when experimenting with two- or three-word sentences. These words are likely to be known to other children as a result of their early integration into the lexicon. Consequently, beyond a slowing-down in the rate of word learning during this transition, the degree of lexical diversity in production may also decrease.

Finally, an alternative explanation for the inverted U-shaped development observed in young children's expressive lexical variability can be given in terms of infant's phonological abilities.[9] There is evidence that infants start their vocabulary development with a

[7]Although we do not have data on lexical diversity in comprehension beyond 24 months, Figure 4 indicates that parameter *b* is reducing to levels compatible with the earlier levels observed in comprehension depicted in Figure 3.

[8]Thanks to Ted Ruffman for bringing this literature to our attention.

[9]Many thanks to an anonymous reviewer for pointing out this alternative explanation.

restricted number of phonological templates, templates that define a syllable structure (Boysson-Bardies & Vihman, 1991) or a specific consonant-vowel pair (Levelt, Schiller & Levelt, 2000) that an infant is able to use productively. These templates are initially idiosyncratic. When children's expressive vocabularies begin to expand, they are constrained to utter only words that fit these templates. Therefore, the more words an infant adds to her expressive vocabulary, the more her expressive vocabulary reflects the initial idiosyncratic starting templates and the more it diverges from the vocabularies of infants with other phonological templates. When infants become more proficient speakers, they loosen these templatic constraints. This permits them to start using words that they previously avoided and so converge towards other infants' expressive vocabularies. It is assumed that phonological templates apply only to production, thereby explaining why the lexical overlap between receptive vocabularies is relatively constant. Further research will help distinguish between different potential explanations of these novel findings.

## Acknowledgements

This work is supported by the Swiss National Science Foundation grant 131700 awarded to Julien Mayor and by the Economic and Social Research Council Grant RES-062-23-0194 awarded to Kim Plunkett.

## References

Amsterdam, B. (1972). Mirror self-image reactions before age two. *Developmental Psychobiology*, **5** (4), 297–305.

Baxter, J. (2011). *Parents working out work*. Available from www.aifs.gov.au/institute/pubs/factssheets/2013/familytrends/aft1/

Benedict, H. (1979). Early lexical development: Comprehension and production. *Journal of Child Language*, **6**, 183–200.

Bleses, D., Vach, W., Wehberg, S., Kristensen, K., & Madsen, T. (2007). *Tidlig kommunikativ udvikling: Et værktøj til beskrivelse af sprogtilegnelse baseret på CDI forældrerapport-undersøgelser af danske normalthørende og hørehæmmede børn*. Syddansk Universitetsforlag Odense.

de Boysson-Bardies, B., & Vihman, M.M. (1991). Adaptation to language: evidence from babbling and first words in four languages. *Language*, **67**, 297–319.

Bradley, R., Caldwell, B., & Rock, S. (1988). Home environment and school performance: a ten-year follow-up and examination of three models of environmental action. *Child Development*, **59** (4), 852–867.

Dale, P., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments & Computers*, **28** (1), 125–127.

Davies, D., & Bouldin, D. (1979). A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **2**, 224–227.

Dunn, J. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, **3**, 32–57.

Eckerman, C.O., Whatley, J.L., & Kutz, S.L. (1975). Growth of social play with peers during the second year of life. *Developmental Psychology*, **11** (1), 42.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). *Variability in early communicative development* (Vol. 59) (No. 5). Monographs of the Society for Research in Child Development.

Fenson, L., Marchman, V., Thal, D., Dale, P., & Reznick, J. (2007). *MacArthur-Bates Communicative Development Inventories: user's guide and technical manual* (2nd edn.). Baltimore, MD: Paul H. Brookes Pub. Co.

Goldin-Meadow, S., Seligman, M., & Gelman, R. (1976). Language in the two-year-old. *Cognition*, **4**, 189–202.

Hamilton, A., Plunkett, K., & Schafer, G. (2000). Infant vocabulary development assessed with a British communicative development inventory. *Journal of Child Language*, **27**, 689–705.

Hart, B., & Risley, T. (1995). *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore, MD: Brookes Publishing Company Inc.

Hartigan, J. (1975). *Clustering algorithms*. New York: John Wiley & Sons Inc.

Howes, C. (1985). Sharing fantasy: social pretend play in toddlers. *Child Development*, **56**, 1253–1258.

Howes, C. (1988). Peer interaction of young children. *Monographs of the Society for Research in Child Development*, **53**, 1–92.

Huttenlocher, J. (1991). Early vocabulary growth: relation to language input and gender. *Developmental Psychology*, **27** (2), 236–248.

Jorgensen, R., Dale, P., Bleses, D., & Fenson, L. (2010). Clex: A cross-linguistic lexical norms database. *Journal of Child Language*, **37** (2), 419–428.

Kristoffersen, K.E., Simonsen, H.G., Eiesland, E.A., & Henriksen, L.Y. (2012). Utvikling og variasjon i kommunikative ferdigheter hos barn som lrer norsk en cdi-basert studie. *Norsk tidsskrift for logopedi*, **58**, 34–43.

Krzanowski, W., & Lai, Y. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, **44**, 23–34.

Levelt, C.C., Schiller, N.O., & Levelt, W.J. (2000). The acquisition of syllable types. *Language Acquisition*, **8** (3), 237–264.

MacWhinney, B. (1991). *The CHILDES Project: Tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mayor, J., & Plunkett, K. (2011). A statistical estimate of infant and toddler vocabulary size from CDI analysis. *Developmental Science*, **14** (4), 769–785.

McCartney, K. (1984). Effect of quality of day care environment on children's language development. *Developmental Psychology*, **20** (2), 244–260.

Miller, J.F., & Chapman, R.S. (1981). The relation between age and mean length of utterance in morphemes. *Journal*

*of Speech, Language and Hearing Research*, **24** (2), 154–161.

Nordic Council of Ministers (2010). *Nordic statistical yearbook 2010* (Vol. 48). (ISBN 978-92-893-2137-2)

O'Neill, D.K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child Development*, **67** (2), 659–677.

Rousseeuw, P., & Kaufman, L. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: John Wiley.

Roy, B., Frank, M., & Roy, D. (2009). Exploring word learning in a high-density longitudinal corpus. In Proceedings of the 31st Annual Cognitive Science Conference.

Szagun, G., Stumper, B., & Schramm, S. (2009). *FRAKIS: Fragebogen zur frühkindlichen Sprachentwicklung (FRAKIS) und FRAKIS-K (Kurzform)*. Frankfurt: Pearson.

US Census Bureau (2010). *Survey of Income and Program Participation (SIPP)*, 2008 Panel, Wave 5. Available from http://www.census.gov/hhes/childcare/data/sipp/2010/tables.html

## Appendix A
## Derivation of the measure of vocabulary diversity

Parental reports are tabulated into a list in order to compute the proportion of infants at a given age that are reported to understand and/or produce any given word on the list. After sorting words in the list, for a given age group, according to the proportion of infants that know the words, we model the resulting distribution of word knowledge using a standard sigmoidal function, providing a mathematically well-defined probability distribution that a word is known given its rank among other words.

The sigmoidal function (of the form $y = 1 - 1/(e^{-(x-a)/b})$ provides an intuitively satisfying fit of this distribution with values close to 100% for highly ranked words (very common words, known by every infant) and

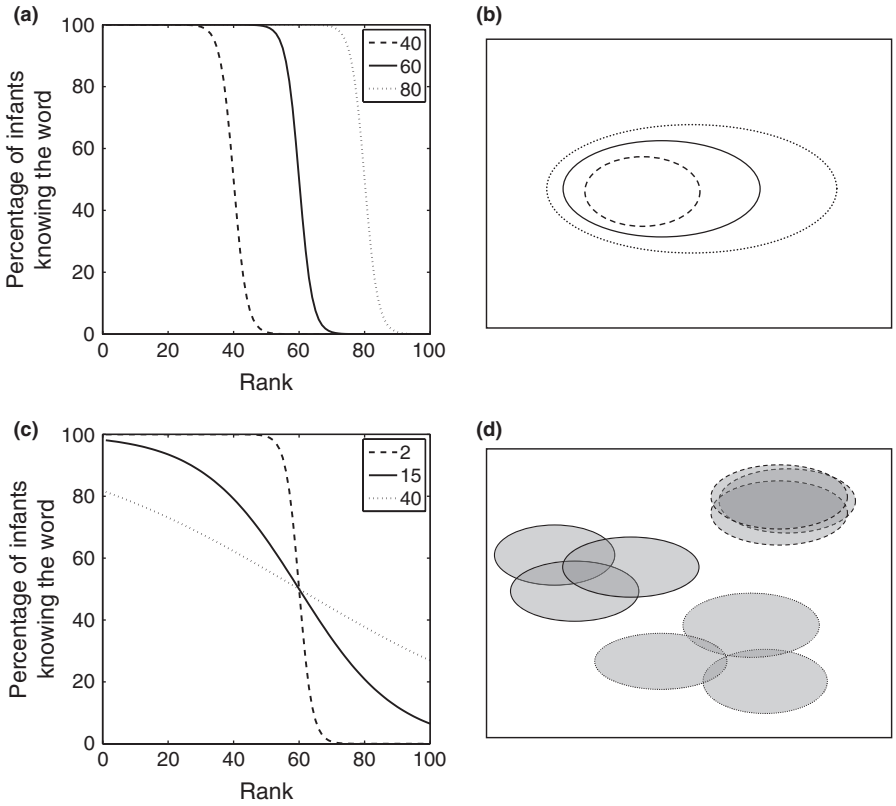


**Figure A1** *Examples of curves describing the proportion of infants knowing a word given its rank among other words. The parameter a regulates the overall vocabulary size (top panels) whereas parameter b describes the structure of vocabulary knowledge, i.e., the amount of knowledge overlap in the infant population (bottom panels). Figure (a). Impact of parameter a on the distribution; (b) Corresponding structures in vocabulary space (c) Impact of parameter b on the distribution (d) Corresponding structures in vocabulary space*

values closer to 0% for low ranked words, known to only a very small subset of the population. Furthermore, sigmoidal functions have only two free parameters. The first of these parameters, *a*, determines the location of the nonlinearity in a sigmoidal curve. For current purposes, this first free parameter determines the rank of the word that is known to 50% of the infants; it is an index of overall vocabulary size. The second parameter, *b*, determines the steepness of the nonlinearity in the sigmoidal curve. In the present model, this second free parameter determines the overlap of word knowledge across the population of infants at a given age, in other words, lexical diversity. A very low value for *b* corresponds to a steep probability distribution, whereas a high value yields a shallow distribution. Shallow distributions correspond to low overlap of individual vocabularies, whereas low values correspond to high overlap. A precise mathematical treatment for the determination of *b* is provided in Mayor and Plunkett (2011). Figure A1 provides a graphical illustration of the impact of varying the two parameters *a* and *b* (reproduced from Figure 2 of Mayor & Plunkett, 2011, p. 773).

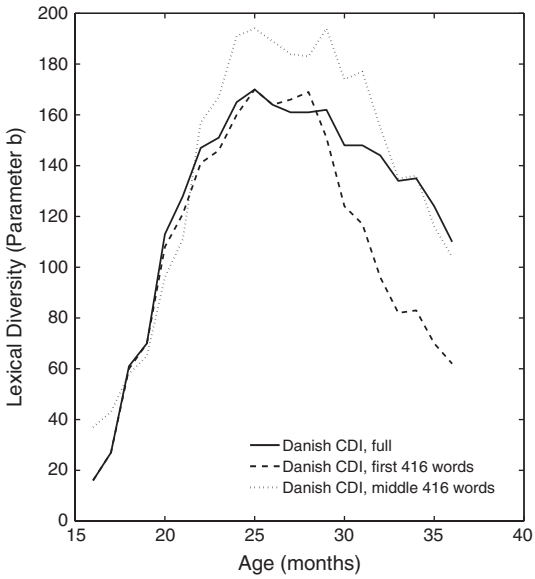


**Figure B1** *Lexical variability in production as a function of age. Different selection of words (full Danish CDI, solid line, the 416 most-known words, dashed line, and 416 words centrally distributed, dashed line) lead to similar measure of lexical diversity. Ceiling effects and edges of the sigmoidal distribution (low- and high-ranked words) do not seem to alter the qualitative behavior of lexical diversity.*

## Appendix B
## Additional tests confirming the robustness of the U-shaped trajectory of lexical diversity

We carried out two additional tests, in order to evaluate the robustness of the sigmoidal method for assessing lexical variability. First, we manipulated the composition of the densest corpus, the Danish CDI, by reducing the total number of words monitored. We selected either the 416 words with the highest rank (a number corresponding to the Oxford CDI) or the 416 words that are ranked in the center of the distribution of word knowledge, for each age group. The aim is to test (1) whether having a limited CDI size affects the trajectory of lexical diversity and (2) determine if a ceiling effect could give rise to a non-monotonic trajectory of lexical diversity. It is possible that when vocabulary size becomes large, words that are known to few infants only would bias the evaluation of lexical diversity by flattening the sigmoidal fit, thereby increasing the optimized value for parameter *b*. The selection of only the first 416 words (the most known) among the 725 ensured that rare words did not bias the estimate of lexical variability. In addition, high frequency words may poorly reflect lexical diversity as a whole. A selection of the 416 words that are centrally distributed in terms of proportion of infants/toddlers knowing these words was also used to evaluate lexical variability.

Figure B1 depicts lexical variability as assessed on the full Danish CDI (solid line), as assessed from only the first 416 words (dashed line) and assessed with the middle 416 words. For all analyses, lexical diversity follows a similar inverted U-shaped trajectory; from an initial high level of overlap, diversity peaks at around 25 to 29 months and then reduces to lower levels. Even though the exact values attained for measures of lexical diversity vary by small amounts, as does the exact location for which lexical diversity peaks, the good agreement between the three conditions suggests that high- and low-ranked words do not bias the sigmoidal method for assessing lexical diversity and that ceiling effects are unlikely to explain its non-monotonic trajectory during development.

Second, we applied both methods for evaluating lexical diversity, the Normalized Euclidean Distance (NED) and the sigmoidal method, in order to evaluate the consistency between both methods when evaluating the non-monotonic trajectory of lexical diversity. Figure B2 depicts the different methods for evaluating lexical diversity on the Oxford CDI (the only database for which we know the exact composition of each child, thus allowing us to apply the Euclidean metrics). In the

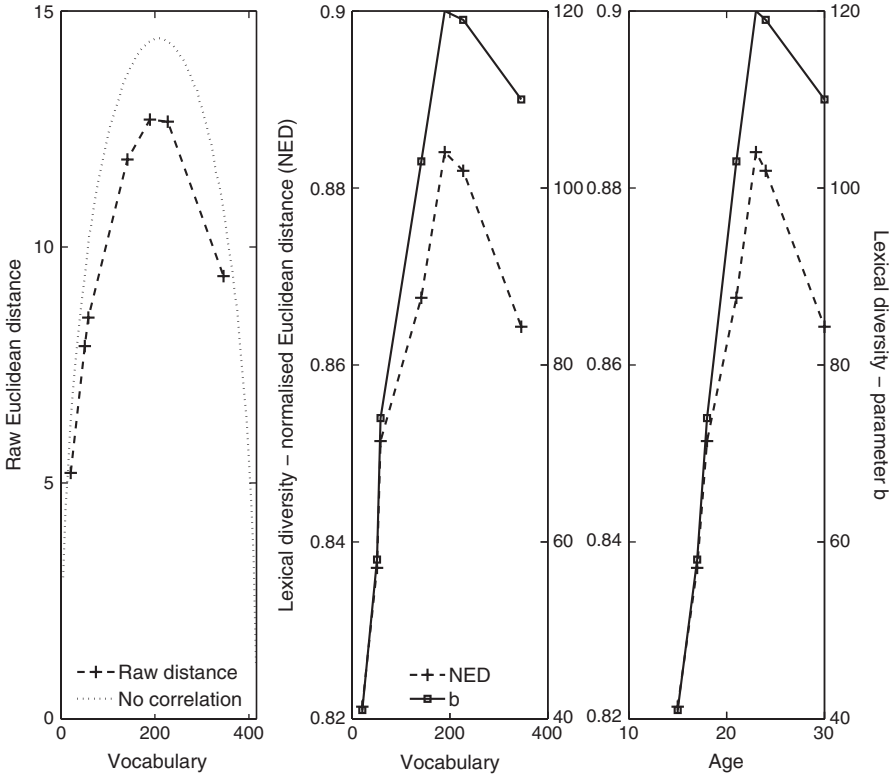


**Figure B2** *Lexical variability in production, for the Oxford CDI. Left, raw Euclidean distance are reported, along with the baseline (dotted line). Middle panel, NED are reported as a function of mean vocabulary, along with measure of parameter b, obtained from the sigmoidal method. The good agreement between both methods is also visible when both measures of lexical diversity are plotted as a function of age (right panel).*

left panel, raw Euclidean distances are reported along with the baseline obtained when no correlations among words and infants are present (where vocabulary data are randomized, similar to the cluster analysis made in the main text). The baseline highlights the strong dependency of the Euclidean metrics on total vocabulary size and also illustrates why raw assessments of lexical diversity based on raw production are not feasible; when comparisons over a fixed number of words is not possible, such as transcripts in CHILDES (MacWhinney, 1991), an appropriate normalization cannot be applied. The middle panel of Figure 2B depicts both the NED and parameter *b* (assessed by the application of the sigmoidal method) as a function of mean vocabulary size. Note the strong agreement between both methods, with both yielding a non-monotonic, U-shaped trajectory of development. The right panel of Figure 4 depicts the trajectory of lexical diversity as a function of age for NED and parameter *b*, and shows that lexical diversity peaks at around 2 years of age and diminishes thereafter.