Finding developmental groups in acquisition data: variability-based neighbor clustering

Stefan Th. Gries*
University of California, Santa Barbara

Sabine Stoll
MPI for Evolutionary Anthropology, Leipzig

Title for running heads

Finding developmental groups in acquisition data

Abstract

This paper introduces a quantitative, data-driven method to identify clusters of groups of data points in longitudinal data. We illustrate this method with examples from first language acquisition research. First, we discuss a variety of shortcomings of current practices in the identification and handling of stages in studies of language acquisition. Second, we explain and exemplify our method, which we refer to as variability-based neighbor clustering, on the basis of mean length of utterance (MLU) values and lexical growth in two different corpora. Third, we discuss the method's advantages and briefly point to further applications both in language acquisition and in diachronic linguistics.

Acknowledgments

The larger portion of this paper was written during the first author's stays at the Department of Psychology and the Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology. We thank Michael Tomasello, Elena Lieven and Bernard Comrie for providing enormously stimulating working environments as well as Balthasar Bickel and Patricia M. Clancy for helpful comments and discussion. The usual disclaimers apply.

Key words

stages of acquisition, clustering, MLU, syntactic development, lexical development

Document statistics

approx. 8500 words, one table, eight figures

* Corresponding author's address
Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
Santa Barbara, CA 93106-3100
United States of America
Email: <stgries@linguistics.ucsb.edu>
Fax (dept.): +1-805-893-7769

Much research in linguistics involves longitudinal quantitative data. To name just two examples, both diachronic linguists and language acquisition researchers monitor how particular distributional patterns change over time. In both domains, it is often assumed that such longitudinal development can best be characterized in terms of stages. In the domain of language acquisition, on which we focus in this paper, there are several approaches to how stages are used, and the purpose for which they are posited may influence how stages are determined.

The most prominent approach in the study of language acquisition is based on the assumption that there are specific stages of development every child goes through and in which there is coherence across specific domains. In the vast majority of cases, the parameter underlying such developmental stages is a child's age, cognitive development (Piaget, 1935, 1937) or, more frequently in language acquisition studies, an index representing the child's linguistic development such as mean length of utterance (MLU) in words or morphemes (Brown, 1973), mean syntactic length (MSL) (Klee, 1989, 1992), and in some cases the score on the Index of Productive Syntax (Scarborough, Wyckoff, & Davidson, 1986; Scarborough, 1990). The reason why many studies have relied on one of these latter indices is the assumption that they are better predictors of children's syntactic knowledge than age given the large age variation found in children's acquisition of all kinds of linguistic features. This approach is based on the assumption that the general stages arrived at by the analysis of, for instance MLU, allow us to make predictions about the development of other domains such as morphology or syntax. The most famous example of this approach to stages is Brown's (1973) groundbreaking study of the grammatical development of three children: Eve made the same grammatical progress from 1;7 to 2;3 that Adam and Sarah made from 2;2 to 3;6. It is yet unclear, though, to what degree MLU stages correlate with age: De Villiers and de Villiers (1973) as well as Miller and Chapman

(1981) found strong correlations between age and MLU (0.78 and 0.88 respectively) a finding, however, which has been difficult to replicate. For example, Klee and Fitzgerald (1985) found no significant correlation (especially for the age range between 24 and 48 months). The stages that are usually assumed are represented in Table 1.

Table 1: MLU stages according to Brown (1973)

The second main use of stages focuses on single domains; using stages merely as a technique for aggregating enough data for analysis. An example of this kind of approach is Klima and Bellugi's (1966) study of questions in which they used MLU stages for extracting questions .

Whether stages are used to make predictions of the development in other domains or whether they are used for single domains, two different ways of using such indices stand out in particular. First, these values are used punctually. For example, MLU values are often given for (parts of) a particular corpus sample under investigation or as a characteristic of children having participated in an experiment with the intention to provide critical information about the child's grammatical development. Also, MLU values are used to match normally-developing children to linguistically-impaired children. Second, they are used longitudinally, i.e., in order for example to reflect the development of a single child.

It is probably fair to say that both of these strategies are quasi-standard in contemporary language acquisition studies. However, ever since the publication of Brown's (1973) seminal work, it is also well known that MLU values and, to a considerable extent, other comparable quantitative indices come with some difficulties as will be discussed below. In addition, the grouping of any quantitative data into different stages also comes with a few risks and potential problems which we focus on below. Given the central role that stages have played for many acquisition or diachronic longitudinal studies in the past, we propose a method of how to group data into stages in a way that circumvents many of these issues. More specifically, we propose a statistical method that can be applied to observational acquisition data in order to identify groups in successive recordings for which MLU values or any other parameters of interest are available. The key characteristic of the method is that it operates in a bottom-up manner, i.e., the categorization is performed on the basis of the data and the parameters of interest themselves rather than on the basis of theoretical preconceptions or on the basis of data from other children or other phenomena.

However, although the present study proposes and exemplifies a method to identify developmental stages in acquisition data, we are not arguing that one should always or mostly use developmental stages. Thus, this paper neither argues in favor of stage-based research in language acquisition nor does it attempt to discuss the overall relevance of stage-based work. What we do argue is that if stages are assumed they need to be calculated in a statistically appropriate way. The goal of this paper is to provide a method to detect the quantitatively most promising candidates for qualitatively interesting changes and stages within a domain based on statistics rather than on intuitions of any one researcher.

In the following section, we first focus on the major problems that come with some ways of using quantitative indices such as MLU values in language acquisition. We then justify and outline the statistical approach underlying our method. Next, we exemplify the method in a case study involving MLU values from a corpus of Russian language acquisition and another case study involving the growth of the lexicon in English acquisition. Lastly, we discuss how this approach can be applied to other quantitative parameters in developmental studies.

Problems with stages of temporally-ordered data

Before we outline the statistical approach to be introduced here, let us briefly exemplify why we think such an approach would in fact benefit the analysis by briefly recapitulating a few problems of MLU values and MLU-based stages in language acquisition research. Note, first, that we will only be concerned with challenges that arise once one has obtained utterance lengths and their means – we will not discuss issues of how these MLU values are arrived at to begin with(cf. Crystal, 1974:295-9). Many of the former kind of problems we will look at are well-known but in order to understand our approach, it is instructive to briefly recapitulate them before we present our alternative way of arriving at stages. Note also that, while this section is largely based on language acquisition research, all of the issues discussed also apply to diachronic studies.

Relevance Problem

The relevance problem is concerned with the fact that in the analysis of a particular phenomenon there may often be no *a priori* reason to use stages based on MLU values or IPSyn values rather than stages defined on the basis of the phenomenon one is actually interested in. In most first language acquisition research, data are grouped into MLU-based stages (e.g. for a study on aspect, Shirai & Andersen, 1995). However, the use of stages based on something other than the variable of interest (such as, the development of tense) can then bias the results in unpredictable directions. Thus, one should avoid this problem in deriving stages on the basis of the variable of interest (cf. Aksu-Koç, 1998 on the acquisition of tense/aspect in Turkish). Aksu-

Koç also groups her data into stages, but on the basis of the phenomenon of interest, the occurrence of particular tense-aspect morphemes. Note in passing that the relevance problem is also relevant to similar analysis of diachronic data such as when researchers are considering to combine data from different historical periods.

Variability Problem

The other problems we would like to raise in this paper can be exemplified best on the basis of an actual example from our data. Consider Figure 1, in whose upper half we plot MLU values (in words) of one Russian child, Child 5 of the Stoll-corpus (Stoll, unpublished data) on the left *y*-axis against the child's age on the *x*-axis. As for the MLU values, they are based on all of 66 recordings of that child, each of which covers approx. one hour during which the child interacted with the mother and other family members in an uncontrolled setting. Since there is no standard procedure of how to deal with potentially repetitive utterances, all utterances were included into the analysis. Given the early stages covered here, the number of multi-word repetitive utterances is likely to be small and does not affect the general methodological point made in this paper. As for the age of the child, it is expressed in decimal format such that, e.g., 2;6.0 / 2 years, six months is expressed as 2.5. Finally, in the lower part of the plot, the dashed line plots the sizes of the standard errors of the MLU values at a higher resolution against the right *y*-axis.

Figure 1: MLUs of 66 recordings of Child 5 between 1;11.28 and 4;03.12

Figure 1 suggests the following: Impressionistically, there is the expected increase of

MLU over time, which can be characterized well with the type of correlation often used in language acquisition studies (Pearson's *r*=.71; *F*(1,64)=64.52; *p*<.001; regression function: MLU=1.46+0.24·AGE; polynomials of degree two and three did not provide significantly better fits).[1] The increase is also clearly reflected by the nonparametric smoothing line shown in the graph (Cleveland's (1979) locally-weighted robust regression). More importantly, the data exhibit what we call the *variability problem:* the data are very variable in terms of both how the mean values increase and decrease again over time and how the sizes of the error bars associated with individual mean values vary. Thus, MLU values are often notoriously unstable and variable, which is often not discussed explicitly although providing measures of central tendencies (e.g., means or medians) without an indication of their dispersion is hardly ever useful. Griffiths (1974:113, n. 1) makes this point in an early review, and Klee and Fitzgerald (1985:259) show that children may be grouped into one of up to three stages, depending on which 100-utterance sample is chosen.[2]

A subproblem of this variability problem is what we call the *developmental problem*. This arises in the longitudinal use of MLU values, i.e., when one tries to characterize the development of a child in terms of MLU stages. It is immediately obvious that one cannot simply lump together all utterances with a particular MLU value because this procedure would be completely blind to the order of elements and the developmental implications this may have. Thus in the example of Figure 1 without taking the order of recordings into account one would end up merging, e.g., the recording at age 2;01.12 (with the MLU value of 1.91) with the sixth recording at age 2;10.06 (with the MLU value of 1.9), effectively merging data that are separated by nine months of age and characterized by widely different neighboring values. An analysis allowing for such wide gaps between to-be-merged values would not only transcend intermediate

MLU stages on just two data points, it would also lump together data points separated by 15 months of age and is thus useless from the perspective of a developmental psychologist. Thus, this problem is more of an academic nature and concerns merely computational issues but since it will be important for the algorithm below we mention it here.

In addition, this variability gives rise to what we will refer to as the data-sparsity problem and the arbitrariness problem.

Data-sparsity problem

The *data-sparsity problem* is concerned with what we have referred to above as the punctual use of MLU values. Especially in cross-sectional approaches, the large variability strongly increases the risk that arbitrarily picking out an isolated MLU value to represent the status of grammatical development at some point of time may result in very different values. In this particular example, note how often even adjacent MLU values can be so different from each other that they would in fact belong to a different stage if stages were used as defined by Brown (1973) or others; cf. Crystal (1974) for some early critique of other aspects that increase the variability of MLU values from different studies. Unfortunately, there is little one can do about this state of affairs other than always trying to maximize the amount of available data and, of course, bearing in mind the danger that comes with this punctual use of MLU values. Given the variability exemplified in Figure 1, it also follows that MLU data from different studies are actually much more difficult to compare than one might assume.

Arbitrariness Problem

The final problems to be mentioned here are what we will refer to as arbitrariness problem . The *arbitrariness problem* is concerned with the fact that the boundaries between different MLU stages are completely arbitrary in the sense that there is really no motivation for why the mean MLU of the first stage should be 1.75 rather than 1.74, 1.69, or 1.81 etc., an issue that Brown (1973:58) himself mentioned and that is reflected in the fact that other scholars have revised Brown's stages into equidistant groups (cf., e.g., de Villiers and de Villiers, 1973). Again, the same problem can surface in diachronic studies. Hilpert (2006) investigates the historical development of the English auxiliary verb *shall* by looking at its verbal complements in two different corpora. Crucially, his corpora cover six successive 70-year periods from 1500 to 1920, but to arrive at larger sample sizes for his statistical tests, Hilpert collapsed these into three consecutive 140-year periods without testing whether this merging of data is in fact warranted (cf. Gries & Hilpert, to appear).

This problem even arises if researchers determine stages on the basis of their data instead of relying on the traditional stages introduced by Brown (1973). This is exemplified by the extreme improbability that different researchers asked to classify the recordings represented in Figure 1 into different groups while simultaneously avoiding to group together widely disparate ages would agree on the exact groups or even, more modestly, identical numbers of groups. With 66 consecutive recordings as variable as those represented in Figure 1, one can devise very many different groupings, and for most of these there are many different ways to arrive at them. This lack of a common procedure to determine stages can pose serious challenges not only within individual studies but also for the comparison of results from different studies . If, for example, researchers are interested in using only a selection of recordings of an individual study for future coding of other variables and the grouping is arbitrary, the validity of all further results is threatened. Unfortunately, these difficulties can also arise with the in general more sophisticated moving average techniques: On the one hand, means obtained from a window span of, for instance, three adjacent MLU values may also result in means that do belong to a MLU stage different from those of the surrounding values. On the other hand, the choice of any window span is arbitrary to begin with.

Some of these issues may appear to be more of an academic point of critique rather than a substantive issue. For example, to a developmental psychologist, a grouping that lumps together recordings as disparate in age as the examples just discussed may seem utterly absurd. However, characterizations of how MLU-based groups are arrived at often leave open questions as to how important methodological decisions were taken – it is not enough to simply state that one attempted to create groups with high between-group variability and small within-group variability if the exact methodological choices are not made explicit. Questions that are often left open include the following: Did the authors use means? Or, did they use medians, following Griffith (1974:113, n. 1)? How was the inevitable variation of the means figured into the recognition of the groups? How were the differences between stages arrived at? Did the grouping of the files take into consideration the possibility that temporally, and thus developmentally, widely disparate recordings have virtually identical MLU values?

Some authors explicitly mention if some sessions are grouped differently then their MLU would suggest (e.g. Bloom, Lifter, and Hafitz (1980:388), but it is usually not explained how the decision on a particular grouping was made and even though this must ultimately be a quantitatively-informed decision, no statistical method has been proposed that allows such groupings on objective grounds. In the following we will propose such a method.

The method and two case studies

A solution that addresses all problems at the same time involves two major changes from the currently dominant practice. First, we suggest to sample the relevant recordings on the basis of the phenomenon of interest – as opposed to MLU values or some other general parameter. Second, we suggest to use a principled bottom-up algorithm – as opposed to a subjective grouping – such that one could choose from groups which are relatively homogeneous internally and relatively heterogeneous when compared to other groups.

One of the methods most widely used on similar occasions is hierarchical agglomerative clustering (cf., e.g., Rousseeuw & Kaufman, 1990 for an overview). While there is a multitude of different clustering algorithms available, the underlying logic of most of them is the one represented Algorithm 1 in pseudo-code. While this representation is not yet particularly frequent in linguistic circles, it has the advantage that it does not require readers to know any programming language but is more explicit than many of the usual characterizations.

Algorithm 1. *Pseudo-code of many hierarchical agglomerative clustering algorithms*

```
1 compute a distance or a similarity matrix,[3] which provides the (dis-)
  similarity of all elements to each other on the basis of some distance
  measure
2 repeat
3     identify the two elements that are most similar to each other
      (in the case of ties, choose one pair randomly);
4     merge the two elements that are most similar to each other and compute
      new distances on the basis of this merger
5 until the number of elements is one
6 draw a dendrogram that summarizes the groupings arrived at in steps 1 to 5
```

Distance measures that are often used (in line 1) are the Euclidean distance, the

Manhattan or City-Block metric, or the cosine; amalgamation rules one often finds (in line 4) are average linkage or Ward's method. For reasons that will become apparent below, it is useful to briefly explain how Ward's method handles the crucial steps in lines 3 and 4. Ward's linking rule involves the reiteration of two steps. First, the algorithm tests all possible fusions of two elements to determine which fusion minimizes the resulting sums of squares. Second, the two elements thus identified are then merged into a new cluster that takes on the weighted average of the original values; then the whole process is repeated.

While agglomerative clustering is often a very revealing procedure, in the present context it suffers from the problem that this kind of cluster analysis is blind to the order of elements. What is needed, thus, is a method that takes the order of elements into account and such a method will be introduced and exemplified presently.

The general algorithm

The method, which we refer to as variability-based neighbor clustering (VNC), starts out from similar data as those underlying Figure 1 and, as it will become evident below, is conceptually similar to agglomerative clustering using Ward's method. However, a crucial difference is that, in addition to the raw MLU values represented in Figure 1, the method by definition also takes into consideration the variability of the data in the recordings. Let us first introduce the VNC algorithm abstractly by means of the pseudo-code in Algorithm 2.

Algorithm 2. *Pseudo-code of general variability-based neighbor clustering*

```
given a set of n recordings where each recording (i) is identified by a
different age and (ii) comes with one or more statistics regarding a
phenomenon in question …
```

```
repeat
    for all groups of recordings named $age_x$ and all recordings named after the next higher $age_{x+1}$
        compute and store some measure of variability for the combined data of all recordings named $age_x$ or $age_{x+1}$
    identify the smallest of all n-1 measures of variability, which is called minvar
    merge the data from all recordings of $age_{minvar}$ or $age_{minvar+1}$
    change the age names of all recordings of $age_{minvar}$ or $age_{minvar+1}$ to the weighted mean of their combined ages
until all recordings have the same age name
```

In order to render this algorithm easier to understand let us see how it works on the basis of two examples.

A first case study: MLUs in Russian acquisition

The first case study is concerned with MLUs in the first language acquisition of Russian and based on the data from a child from the Stoll corpus of Russian. The case study is based on similar data that were used to generate Figure 1. We analyzed all utterance lengths (in words) from 123 recordings of Child 3 (age 1;03.26 and 4;09.30) from the Stoll corpus. To obtain these data, we retrieved all character sequences separated by whitespace from all utterances of all files of this particular child. Accordingly, we obtained all utterance lengths of the child in *n*=123 recordings and, thus, had the right input for Algorithm 2. We now propose to operationalize the similarity of two adjacent recordings as the absolute difference of their two MLU values (in words, MLUw) divided by the standard deviation of all utterance lengths in the two adjacent recordings. In Algorithm 3, we render the general algorithm more precise.

Algorithm 3. *Pseudo-code of variability-based neighbor clustering 1*

```
repeat
    for all groups of recordings named $age_x$ and all recordings named after
    the next higher $age_{x+1}$
        compute the MLUs of all recordings named $age_x$ or $age_{x=1}$
        compute the absolute difference $diff_{x,\ x+1}$, abs($MLU_x$-$MLU_{x+1}$)
        compute the standard deviation $sd_{x,\ x+1}$ of the combined data from
        all recordings named $age_x$ or $age_{x+1}$ and add a small constant
        compute the quotient $diff_{x,\ x+1}$ divided by $sd_{x,\ x+1}$
        store this quotient for the set of recordings named $age_x$ or $age_{x+1}$
    identify the smallest of all n-1 quotients of variability, which is
    called minvar
    merge the data of $recording_{minvar}$ and $recording_{minvar+1}$ into a new recording
    change the age names of all recordings of $age_{minvar}$ or $age_{minvar+1}$ to the
    weighted mean of their combined ages
    for each recording
        compute the MLU and its standard error
        plot the MLU against the (original or already changed mean) age
        at the MLU's recording time
        add vertical bars representing the standard error of the MLU
        add horizontal error bars encompassing the range of original
        recordings that are summarized in the data point
    for future evaluation, store in an extra output file the elements
    clustered and the distance (the relevant quotient)
stop repeating all this when all recordings have been merged
```

Figure 2 plots 123 MLU values of Child 3 against the ages of this child at each recording time; the vertical error bars represent one standard error for each recording and the dotted line summarizes the overall tendency of the MLU values using a nonparametric smoothing technique (cf. Cleveland, 1979).[4]

Figure 2: MLUs of 123 recordings of Child 3 between 1;03.26 and 4;09.30: before amalgamation starts

When the algorithm goes through lines 2 to 5 the first time, it computes the MLU of all fifteen utterances of the first recording (which was made at age 1;03.26, which we display on the *x*-axis in Figure 2 as a decimal: $1+{}^{3}/_{12}+{}^{26}/_{365}=1.32123$) and all 238 utterances of the second recording (which was made at age 1;04.03, which in the same decimal format is 1.34155). All utterance lengths are 1, which is why (i) both MLUs amount to one and the absolute difference is zero and (ii) the standard deviation of all utterance lengths from both recordings is 0, to which a small constant is added in line 5 to rule out divisions by zero. The quotient computed and stored in lines 6 and 7 is thus ${}^{0}/_{0.00001}\approx 0$, i.e. the minimal possible value. The algorithm then performs these steps for all adjacent recordings (i.e., 2 and 3, 3 and 4, …, 122 and 123). This is the equivalent to computing a distance matrix in regular clustering, but in this case the algorithm only computes a distance vector because it does not compare utterance lengths of all recordings to every other one but only to those of the temporally neighboring ones.

The algorithm then determines in line 8 that the first quotient of all 122 is one out of several minimal ones and, in line 9, merges the data of the first and the second recording into one new recording, which now comprises 15+238=253 utterances. In line 10, this new, merged, recording gets as a name the mean of the two original recordings (i.e., 1.33139); this merely serves the purpose of always being able to identify to which (merged) recording each MLU value belongs.[5] Lines 11 to 15 produce a new plot, which is represented in Figure 3. Line 16 stores all the information resulting from this merging into an interim/output file. Line 17 ends the algorithm when all recordings have been merged into just one: just like all agglomerative clustering approaches, our amalgamation process ends when all initially distinct data points have been amalgamated into one big cluster. This must not be misunderstood as meaning that all data sets must be interpreted as constituting one single cluster. It means that any algorithm will ultimately amalgamate all data points into one cluster and that – as with all exploratory methods – it is the researcher who will have to decide on the desired resolution of the results.

Figure 3: MLUs of 123 recordings of Child 3 between 1;03.26 and 4;09.30: step 1

The number of required computations is substantial. It is therefore not feasible to perform the computations of 7,500+ standard deviations (the cumulative sum of 1 to 123) as well as all 122 mergings etc. by hand. We used an R script written by the first author, who can be contacted for details and the actual code.

Now what is the result of this whole process? The result is a stepwise amalgamation process just like in hierarchical agglomerative clustering. The successive amalgamation of all 123 recordings shows that the differences between any two successive graphs is minimal, but the overall clustering process can be checked either on the basis of individual graphs such as those in the above figures or, for a more intermediate step, Figure 4.

Figure 4: MLUs of 123 recordings of Child 3 between 1;03.26 and 4;09.30: step 80

More rewarding, however, is the inspection of graphs near the end of the amalgamation process. For example, Figure 4 suggests that the MLUs of this child be summarized as making up several different stages with ascending MLU values and one 'outlier' at the age of approximately 4;0.24 (which is 4.07 in decimal format). It is worth noting that the definition of outliers here is not arbitrary – rather, the user can rely on the data-driven results for identifying outliers. For example, one could define outliers by noting that a particular value violates the developmental tendency of all or nearly all other values in the figure as this one does. In addition, outliers could be identified on the basis of the distance that has to be bridged for amalgamation, which is given in the caption of each figure and also logged in an additional output file (cf. line 16 of the algorithm in Algorithm 2). We will mention a third way below.

Note several important features of this type of analysis. First, just as would be expected the analysis avoids the developmental problem, i.e., the amalgamation of non-adjacent recordings and the disadvantages that come with this otherwise standard approach to clustering because only recordings with adjacent age names are compared and can be merged. Second, just like regular clustering algorithms, this method also does not only indicate the most useful groupings of data points – it also indicates the size of the clusters and, thus, the lengths of the developmental stages that are most strongly supported. Third and more importantly, as the implementation presented above is modeled on the Ward algorithm, it takes into consideration not only the means as such, but also the dispersion of the values by always including the standard errors of all merged recordings.

Fourth, for those who are more accustomed to interpreting such results on the basis of traditional tree diagrams, the data can be transformed into a dendrogram, too, where the ages are plotted on the *x*-axis while the distances on the *y*-axis correspond to the sums of quotients computed in line 8 and 9 of the algorithm. This way, the more intuitive representation in Figure 5 is available.

Figure 5: A dendrogram-like representation of the amalgamation of the 123 recordings of Child 3 between 1;03.26 and 4;09.30

This dendrogram allows for the third way of identifying outliers: An outlier could be defined as a single data point that is merged very late and, if desirable, the researcher could even specify a particular threshold value on the *y*-axis to define what is meant by 'late'; as an example consider again the above mentioned data point at age 4.07, which is merged at a sum-of-differences point as late as at 8.42 out of a maximum sum of differences of 10.68 for this child. Just as with other agglomerative cluster analyses, there are three different ways to approach the output of our algorithm.

First, one can inspect the dendrogram on an intuitive, eye-balling basis. On that basis, this dendrogram suggests, for example, one big cluster from the earliest recording until about age 2;03, followed by a smaller cluster of nine recordings, then three clusters etc. as indicated by the boxes.

Second, one can resort to more rigorous bottom-up analysis. For example, it is possible to plot the difference quotient obtained in line 8/9 against the amalgamation steps (as an inverse scree plot) to determine where natural numbers of clusters emerge because dissimilarity increases after a few amalgamation steps. Such a graph is represented in Figure 6, which suggests to adopt five or twelve clusters because the fifth and the twelfth data point are local minima. That means, when there are only twelve clusters left and one wants to merge elements again, then one has to bridge a huge distance before the next recording can be merged with one of the existing twelve clusters, which reflects the fact that this amalgamation is costly in terms of resulting in a large increase in heterogeneity; the same applied to the fifth cluster.

Figure 6: Difference coefficients across all amalgamation steps

Third, one can approach the dendrogram with an *a priori* defined number of clusters in mind and determine the sizes and locations of these clusters. For example, if a researcher required a *n*-cluster solution for, say, a particular sampling scheme, then the analyst can inspect the dendrogram for a solution with a number of clusters closer to the required number.

It is worth pointing out in this connection what this implies and relate this to some general caveats with regard to cluster-analytic approaches. On the one hand, this means that the results can be used flexibly, but this does not also mean that the researcher can lump together any recordings because (i) the clusters that were arrived at on a principled bottom-up basis constrain the ranges of possible groups and (ii) the distances between clusters make some groupings more likely than others. This approach, thus, avoids the arbitrariness problem but at the same time allows to identify structure where other statistical/representational formats do not (as in Figure 1). On the other hand, the algorithm does not deterministically decide on the number of clusters that are observed in the data. Any cluster analysis will detect structure in data – it is up to the researcher to decide which grouping is theoretically most meaningful and/or practically most desirable. It should be emphazised that VNC is a heuristic tool and the algorithm we propose needs to be backed by theoretically informed or practically necessary decisions .

Given all these advantages, we suggest that the proposed algorithm and its diagrammatic representations allow for a more objective and data-driven identification of relevant groups in the data than a mere eye-balling of the distribution of the MLU values. In the following section, we present a second case study to provide some further support for this claim.

A second case study: lexical development in English

The second case study examines the lexical development in a corpus of of one English child (Adam) from the Brown corpus (1973) taken from the CHILDES database (MacWhinney, 2000). We can expect a positive correlation between the age of the child and the size of his lexicon. In order to represent approximately what this correlation looks like, we analyzed all 53 files of Adam. In order to determine how Adam's lexicon grew over time, we used an R script implementing Algorithm 4.

Algorithm 4. *Pseudo-code to estimate the growth of Adam's lexicon*

```
for each recording$_x$
    read all utterances by Adam
    split each line into words at sequences of characters that were not letters, hyphens, or plus-signs
    store all types that occur in each recording in a list
define a vector lexicon
store all types from the first recording into the vector lexicon
define a vector lexicon.size
store the number of types in the first recording in lexicon.size$_1$
for each recording$_y$ (where y iterates from 2 to n)
    determine which of the types in recording$_y$ is not yet in the vector lexicon
    add these types to the vector lexicon
```

```
store the new length of the vector lexicon in lexicon.size_y
```

As a result of this, we obtain a vector `lexicon.size`, which contains for each recording the number of types produced at least once until and including this recording. The development is summarized in Figure 7.

Figure 7: Frequencies of all word types produced by Adam between 2;3.04 and 5;02.12

For illustration purposes we chose a simplistic approach to measuring the size of Adam's lexicon. As a more refined method, for example, one could include a word only after it has been produced a second time or after it has been produced in at least two different contexts. Our point, however, is that whatever approach one adopts, one would also end up with some such vector containing type frequencies and would have to take some decisions as to how to group the data. In other words, the approach we propose here is independent of how the original vector to which it is applied has been arrived at.

It is obvious that any researcher who wished to classify these data into groups would probably run the considerable risk of ending up with a rather subjective classification that many other researchers might not want to subscribe to. Also, there is the additional problem that there is just one observation per case, viz. the number of words that the child is assumed to know. This is untypical because the usual kind of application of a cluster analysis requires that the elements to be clustered are described on the basis of $c>1$ criteria, i.e., each element is characterized on the basis of a vector with more than one element so that measures of vector similarity such as correlations, cosines etc. can be applied. We solve this problem by choosing an appropriate similarity measure namely the variation coefficient: That is, the similarity of lexicon sizes at two points of time is measured by the quotient of the mean of their joint type frequencies divided by the standard deviation of their joint type frequencies, which is less dependent on the size of the mean as a regular standard deviation would be; cf. Algorithm 5.

Algorithm 5. *Pseudo-code of variability-based neighbor clustering 2*

```
01 repeat
02     for all groups of recordings named $age_x$ and all recordings named after the next higher $age_{x+1}$
03         compute the variation coefficient of all these recordings named $age_x$ or $age_{x=1}$
04         store this variation coefficient for the set of recordings named $age_x$ or $age_{x+1}$
05     identify the smallest of all n-1 variation coefficients, which is called minvar
06     merge the data of $recording_{minvar}$ and $recording_{minvar+1}$ into a new recording
07     change the age names of all recordings of $age_{minvar}$ or $age_{minvar+1}$ to the weighted mean of their combined ages
08     store the new age names and the variation coefficient of the recordings just merged
09 stop repeating all this when there is just one recording left
10 for all mergers just stored
11     plot the sizes of the variation coefficients on the y-axis against the ages on the x-axis
```

When the algorithm goes through lines 1 to 3 the first time, it computes the variation coefficient of the two cumulative type frequencies of the first two recordings, which are 397 word types and 592 word types respectively, which amounts to 0.2788. This is then done for all adjacent recordings (i.e., 2 and 3, 3 and 4, …, 52 and 53). The algorithm then determines in line

5 that the variation coefficient for the recordings at age 4;06.24 (i.e., 4.586) and at age 4;07.01 (i.e., 4.6628) is the smallest (namely 0.0054), and, in line 6, merges the data of these two recordings into one new recording, which now comprises the values of 397 and 592. In line 7, this new, merged, recording gets as a name the mean of the two original recordings (i.e., 4.6244). Lastly, this smallest variation coefficient is stored for later plotting and the algorithm is repeated until all recordings have been merged. The resulting dendrogram is shown in Figure 8.

Figure 8: A dendrogram-like representation of the amalgamation of the lexical growth, 55 recordings of Adam between 2;3.04 and 5;02.12

As the dendrogram shows, there is quite some structure in the development that seemed so difficult to characterize in Figure 7. Depending on one's needs or on more detailed analysis of the actual types involved, the analysis strongly suggests that it is best to choose five, or more liberally between four and up to seven, clusters of consecutive recordings. Also, the dendrogram considerably constrains the range of possible groupings that Figure 7 would still have allowed for and all the other advantages discussed in above apply as well. We thus again submit that the overall VNC approach – i.e., regardless of the exact statistics involved – allows to identify structure in seemingly messy and continuous data sets on a principled, quantitative, bottom-up basis.

Conclusion

We started out by pointing out a variety of problems of studies using temporally-ordered data in language acquisition research: the problem that many groupings are not performed on the basis of the phenomenon of interest but on the basis of the more convenient approach using MLU values as well as several problems that have to do with the exact way of how groupings are established. While we do not necessarily believe that developmental stages are always necessary, we suggest that, *if* stages are used, then variability-based neighbor clustering allows to identify the number of groups that is suggested by the dataset itself in a data-driven, quantitative and replicable way. Crucially, these stages are based on the phenomenon one is actually interested in rather than on age or MLU.

It is worth emphasizing again that nothing hinges on the particular operationalizations we have chosen here. This is true on two relevant dimensions. First, with regard to the cluster-analytic statistics: the measure of similarity and the amalgamation rule; second with regard to the statistic on which the clustering is performed. As to the former, for example, in the first case study the measure of similarity we used was the standard deviation while in the second it was the variation coefficient. In addition, we have used an amalgamation strategy modeled on the well-known method proposed by Ward. We do not intend to suggest that these are *a priori* the best methods – other researchers might prefer other measures of variability or similarity such as standard errors, entropy values, cosines between vectors …, or other amalgamation methods. We suggest, however, that a data-driven, bottom-up approach of the above kind is useful for determining stages independent of the kind of developmental data investigated., and we believe that the flexibility of the general algorithm is a virtue in the sense that researchers can tailor it to their particular needs.

As to the latter, in the first case study we used MLU values (because they are the most widely used statistic for arriving at stages), and in the second we used lexical growth. Again, other researchers, however, will want to apply the method to yet other data. It is especially in this respect that we think that this approach has a lot to offer. The general algorithm allows to group recordings on the basis of *any* quantitative measure irrespective of whether each recording is characterized by many values (as in case study 1 with the individual utterance lengths per recording) or just a single value (as in case study 2 with a single type frequency per recording). Thus, the stage-wise development of any linguistic feature can now be described completely in its terms and without reference to predefined MLU stages or other potentially irrelevant parameters. The same holds for all other studies in which different recordings/files are associated with quantitative data, opening up new areas of exploration also in, for instance, diachronic studies. For example, in the domain of historical linguistics, Gries and Hilpert (to appear) discuss how VNC can be used to determine different historical stages in the development of the English auxiliary *shall* and the English present perfect.

Two final remarks. First, we deliberately avoid the addition of significance tests or additional mathematical modeling to the current method. Neither does this rule out the addition of significance testing to the algorithm (by, for example, employing resampling methods; cf. Suzuki 2006) nor does it preclude additional mathematical modeling of the results. Our focus in introducing this method here, however, is that we intend VNC to be primarily used on an exploratory basis just like most other clustering methods. Second, we explicitly argue against blindly applying some implementation of VNC to longitudinal data. Different cluster solutions need to be explored and checked for similarity or consistency (using, say, Fowlkes and Mallow's 1983 measure or similar statistics); the impact of removing potential outliers then should be tested in ways similar to model criticism on the basis of adjusted R^2 or *AIC* values.

In sum, we introduced a general kind of clustering method for longitudinal data, one that is comprehensive, rigorous, replicable, data-driven and bottom-up. We believe that linguistics as a field has much to gain from at least exploring such methods in more detail to ultimately arrive at more objective ways of categorizing the various kinds of longitudinal data we regularly encounter.

References

Aksu-Koç, A. (1998). The role of input vs. universal predispositions in the emergence of tense-aspect morphology: evidence from Turkish. *First Language*, *18*, 255-280.

Baayen, R.H. (2004). Statistics in psycholinguistics: a critique of some current gold standards. *Mental Lexicon Working Papers*, *1*, 1-45.

Bloom, L., Lifter, K., & Hafitz, J. (1980). Semantics of verbs and the development of verb inflection in child language. *Language*, *56*, 386-412.

Bondal, J.A., Ghiotto, M., Bredart, S., & Bachelet, J.-F. (1987). Age-relation, reliability and grammatical validity of measures of utterance length. *Journal of Child Language*, *14*, 433-446.

Brown, R. (1973). *A first language: the early stages*. Cambridge, MA: Harvard University Press.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical* Association, *74*, 829-836.

Crystal, D. (1974). Review of Roger Brown, A first language. *Journal of Child Language*, *1*, 289-307.

Fowlkes, E.B., & Mallows, C.L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, ***78,*** (383), 553-569.

Gries, St.Th. (2006). Exploring variability within and between corpora: some methodological considerations. *Corpora*, *1*, 109-51.

Gries, St.Th., and Hilpert, M. (to appear). The identification of stages in diachronic data: variability-based neighbor clustering. *Corpora*, *3*.

Griffiths, P. (1974). Review of Melissa Bowerman, Early Syntactic Development: a cross-linguistic study with special reference to Finnish. *Journal of Child Language*, ***1,*** 111-122.

Klee, T., & Fitzgerald, M.D. (1985). The relation between grammatical development and mean length of utterance in morphemes. *Journal of Child Language*, *12*, 251-269.

Klima, E.S., & Bellugi, U. (1966). Syntactic regularities in the speech of children. In: J.R. Lyons & R.J. Wales (Eds.), *Psycholinguistic papers: the proceedings of the 1966 Edinburgh Conference* (pp. 183-208). Edinburgh: Edinburgh University Press.

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Miller, J.F., & Chapman, R.S. (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech, Language, & Hearing Research*, *24*, 154-161.

Parker, M.D., & Brorson, K. (2005). A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). *First Language*, *25*, 365-376.

Piaget, J. (1935/1952). *The origins of intelligence in children*. New York: Norton.

Piaget, J. (1937/1954). *The construction of reality in the child*. New York: Basic Books.

R Development Core Team (2006). *R: a language and environment for statistical computing*. R Foundation for Statistical. Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

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

Sagae, K., Lavie, A., & MacWhinney, B. (2005). Automatic measurement of syntactic development in child language. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 197-204.

Scarborough, H.S., Wyckoff, J., Davidson, R. (1986). A reconsideration of the relation between age and mean utterance length. *Journal of Speech, Language, & Hearing Research*, *29*, 394-399.

Scarborough, H.S. (1990). Index of productive syntax. *Applied Psycholinguistics*, ***11,*** 1-22.

Shirai, Y., & Andersen, R.W. (1995). The acquisition of tense-aspect morphology: a prototype account. *Language*, *71*, 743-762.

Stoll, S., & Gries, St.Th. (under revision). The acquisition of tense and aspect in Russian: an association strength approach.

Suzuki, R. 2006. pvclust 1.2-0. A package for R. URL: <http://www.is.titech.ac.jp/ %7Eshimo/prog/pvclust>, last accessed Nov 8, 2007.

de Villiers, J.G., & de Villiers, P.A. (1973). A cross-sectional study of the acquisition of grammatical morphemes in child speech. *Journal of Psycholinguistic Research*, *2*, 267-278.

Table 1: MLU stages according to Brown (1973)

Stage	Average age	Mean MLU	MLU range
I	15-30	1.75	1-2
II	28-36	2.25	2-2.5
III	36-42	2.75	2.5-3
IV	40-46	3.5	3-3.7
V	42-52+	4	3.7-4.5

Figure 1: MLUs of 66 recordings of Child 4 between 1;11.28 and 4;03.12

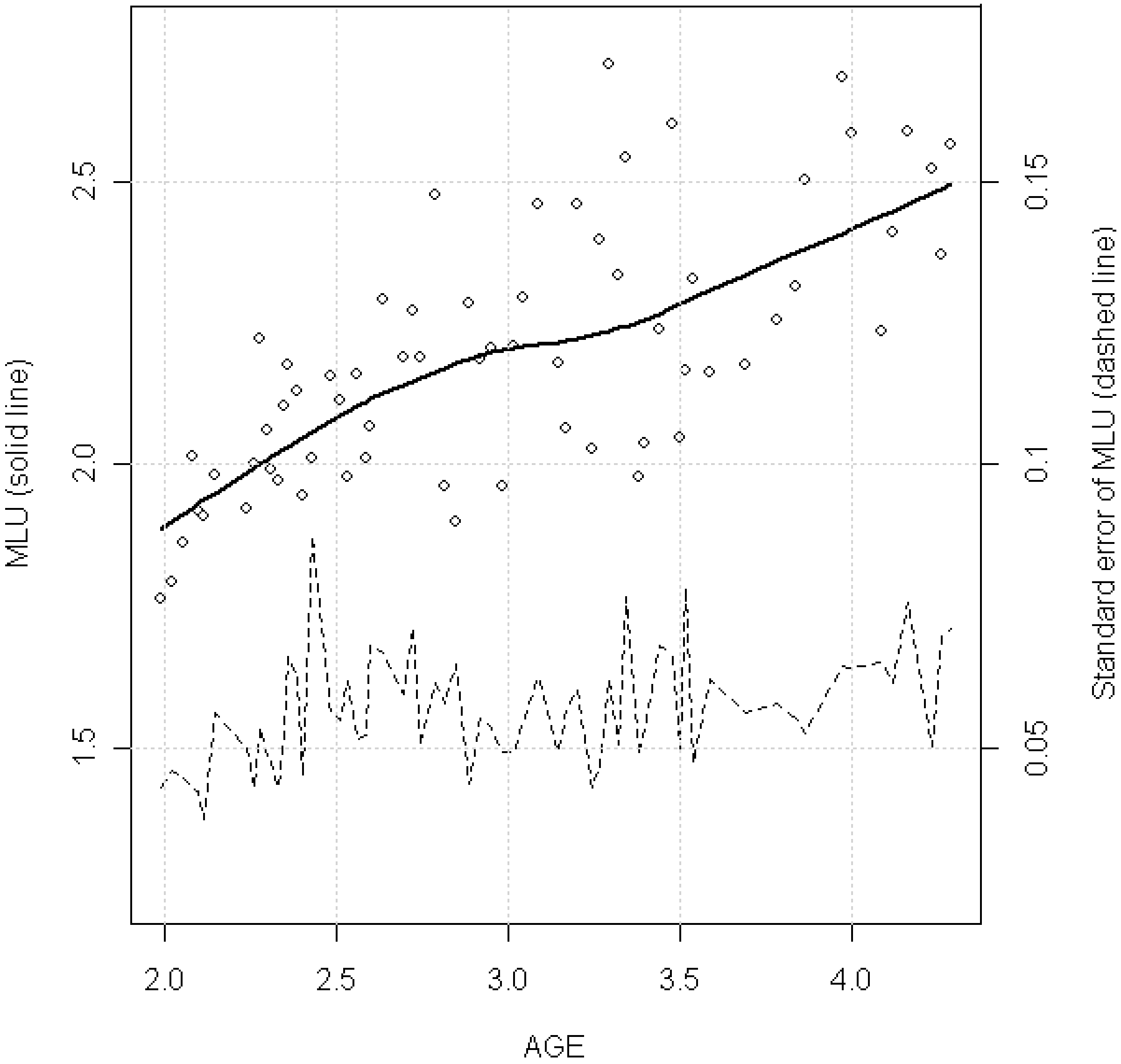


Figure 2*: MLUs of 123 recordings of Child 2 between 1;03.26 and 4;09.30: before amalgamation starts*

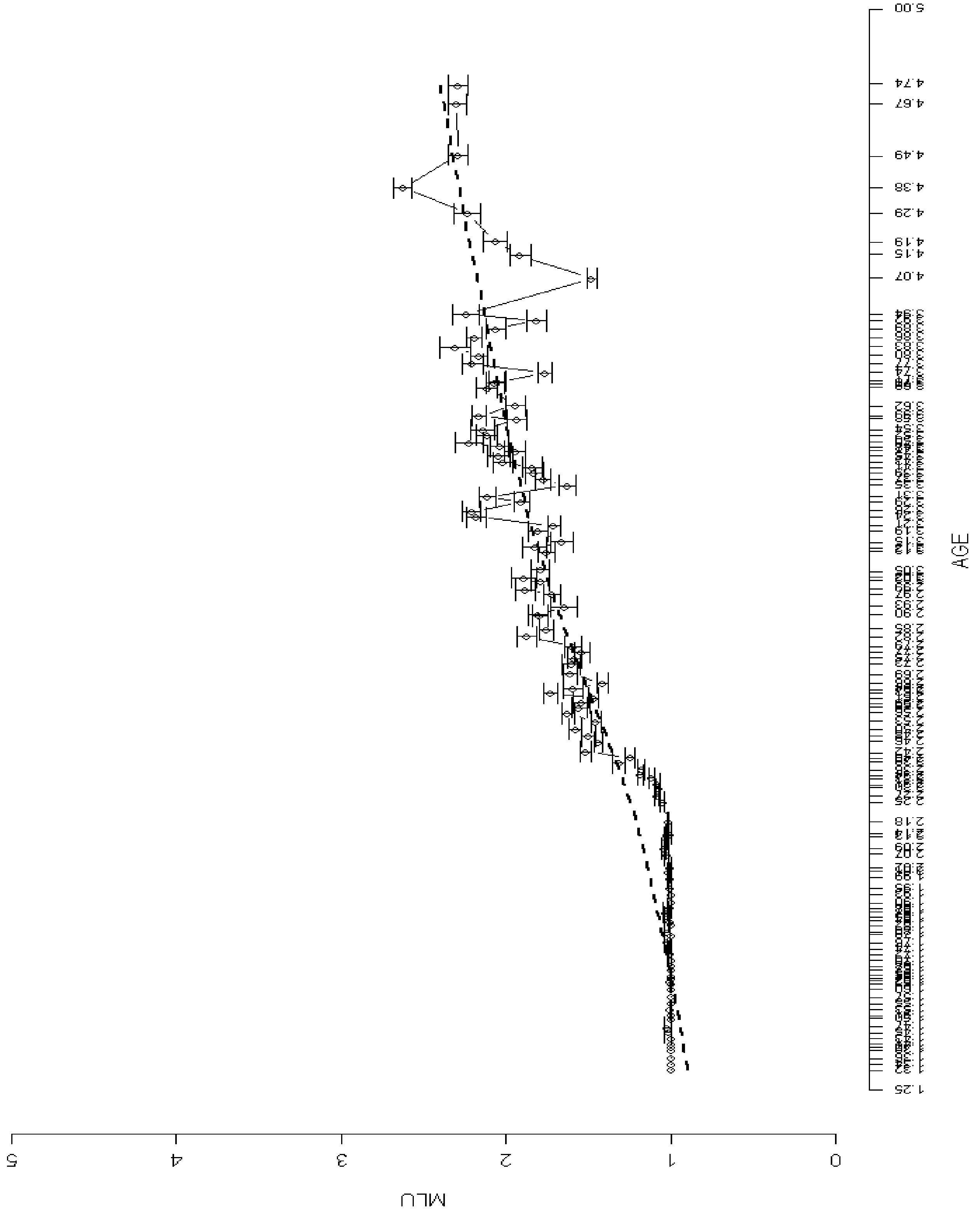


Figure 3: MLUs of 123 recordings of a Child 2 between 1;03.26 and 4;09.30: step 1

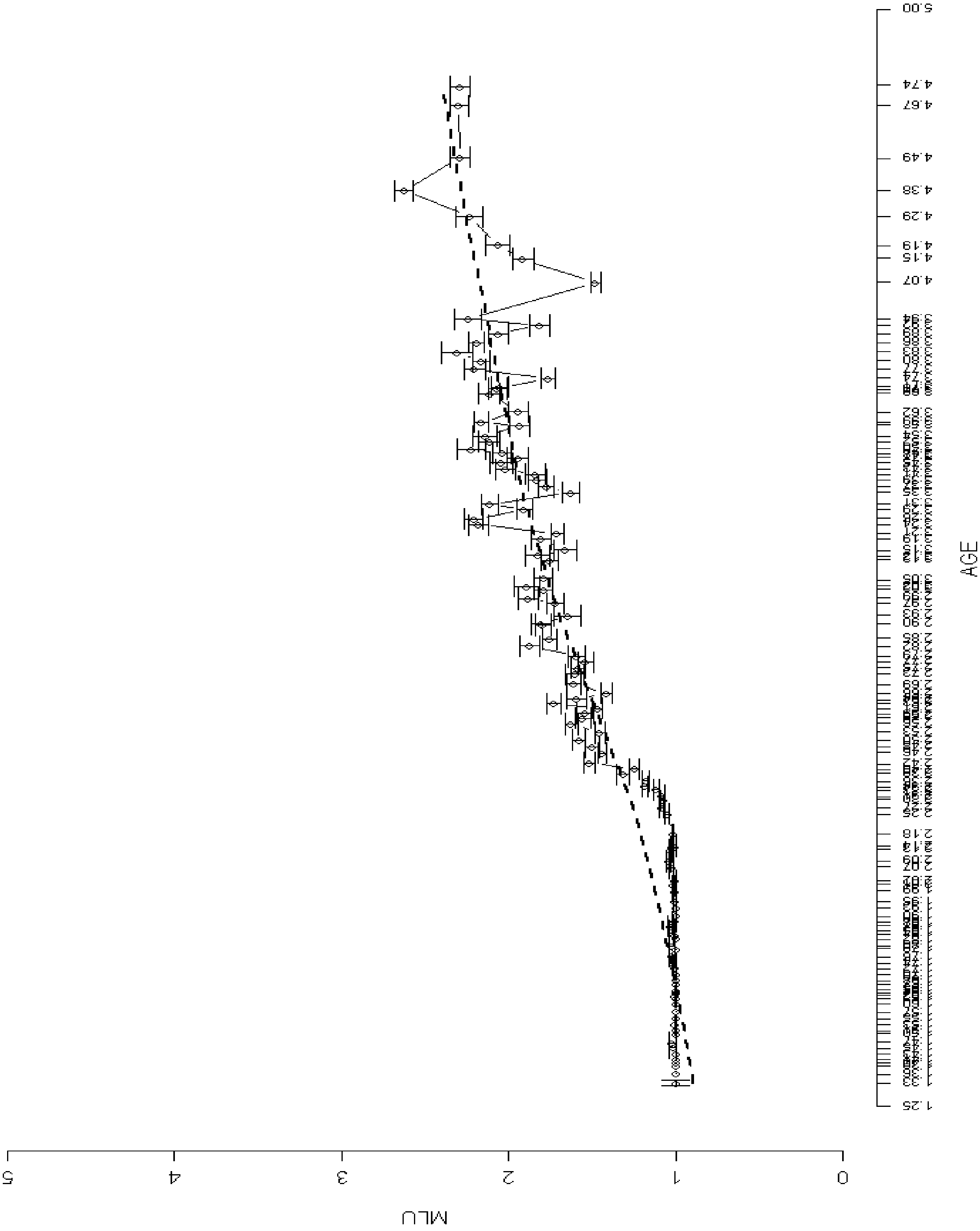


Figure 4: MLUs of 123 recordings of Child 2 between 1;03.26 and 4;09.30: step 80

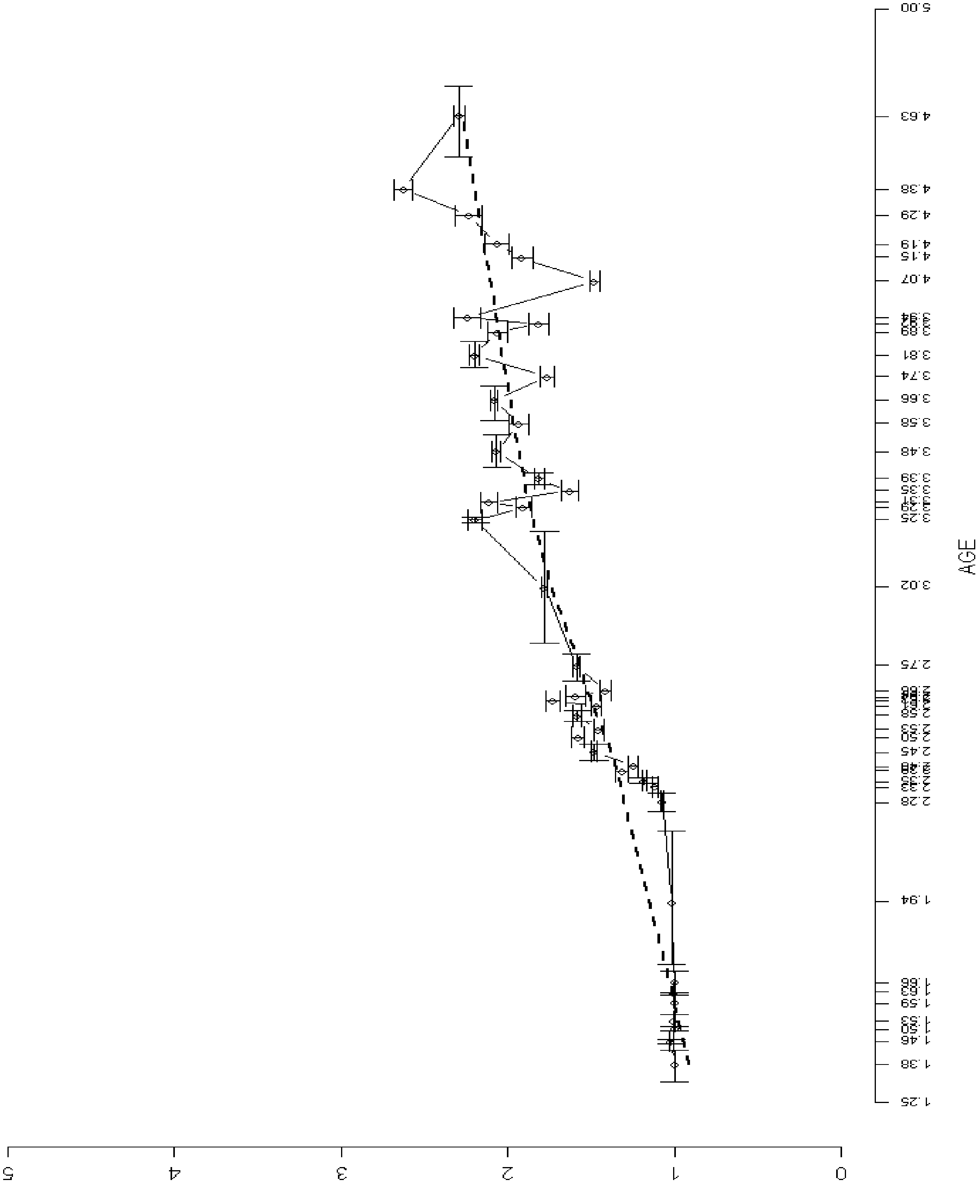


Figure 5: A dendrogram-like representation of the amalgamation of the 123 recordings of Child 2 between 1;03.26 and 4;09.30

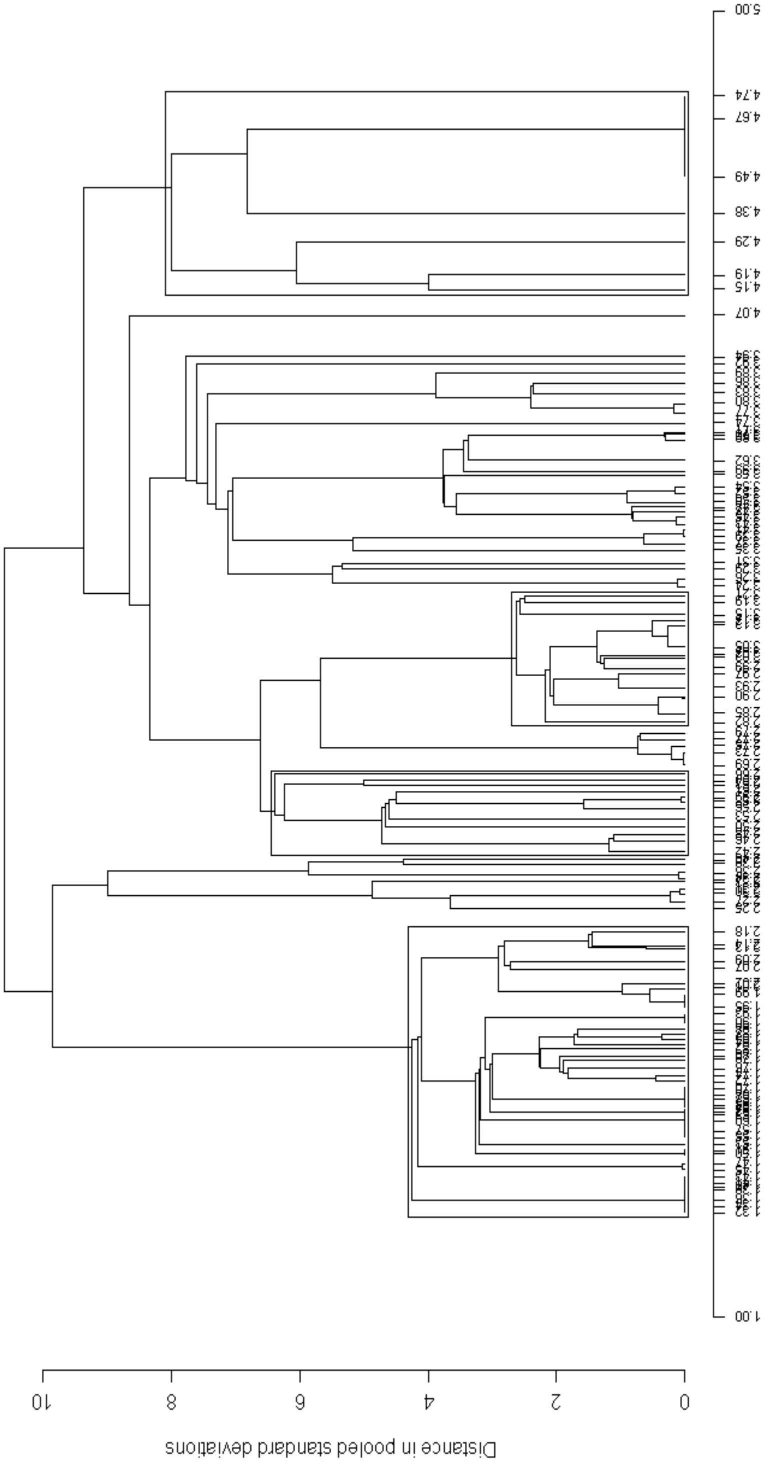


Figure 6: Difference coefficients across all amalgamation steps

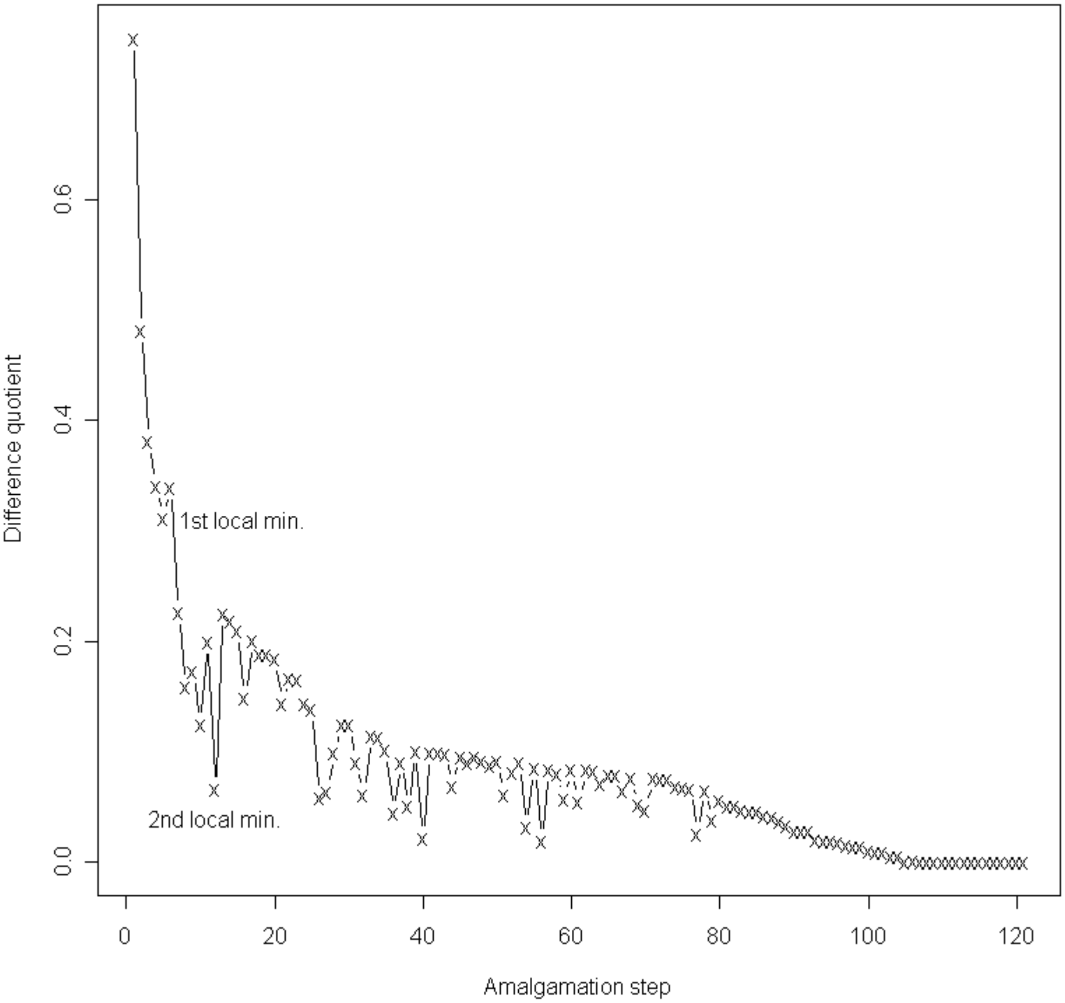


Figure 7: Frequencies of all word types produced by Adam between 2;3.04 and 5;02.12

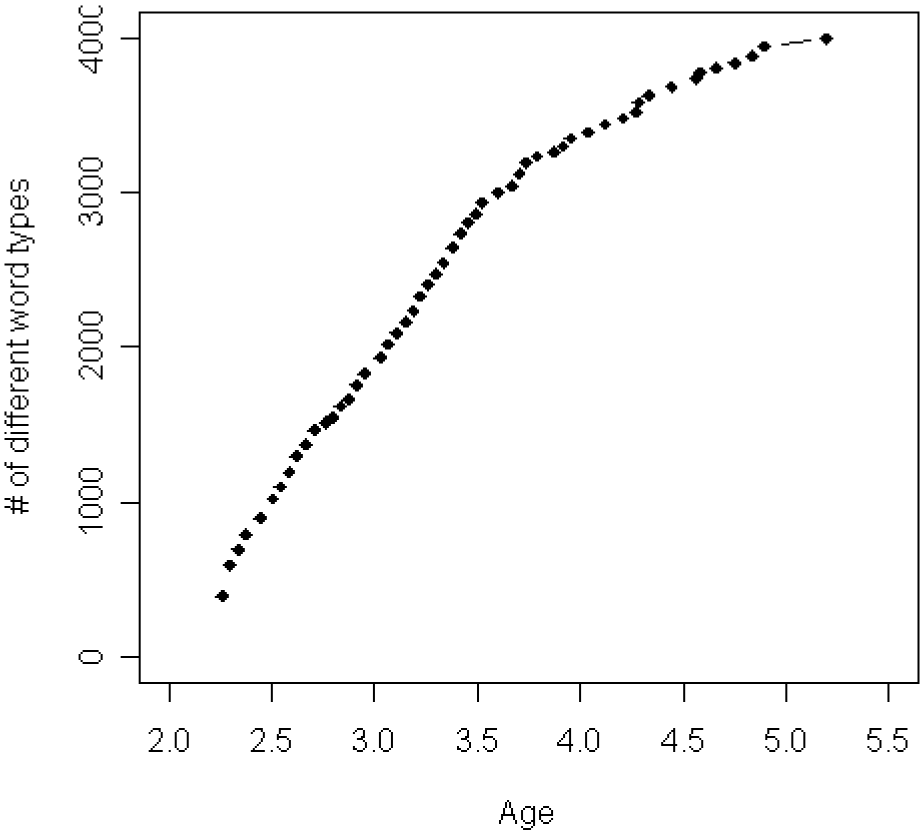
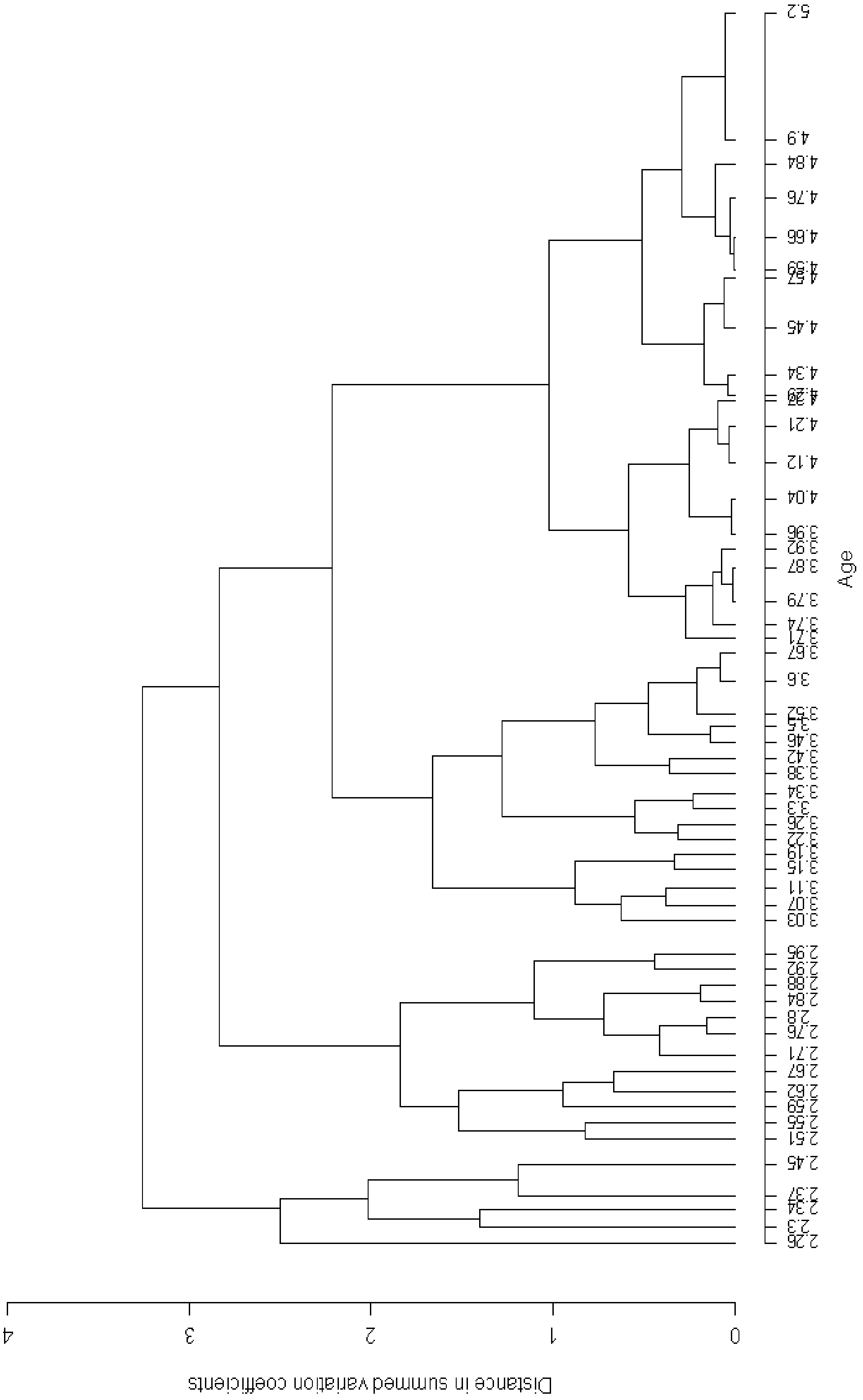


Figure 8: A dendrogram-like representation of the amalgamation of the lexical growth 55 recordings of Adam between 2;3.04 and 5;02.12



[1] We know that a linear regression is not really possible here since, e.g., the data points violate the assumptions of homoscedasticity and normality of errors, but following the tendency in the literature to report linear regressions results, we provide the relevant statistics for the sake of comparability.

[2] Klee and Fitzgerald's claim must be taken with a grain of salt since their argument is based on confidence intervals computed from standard errors, which are problematic since the data are certainly not normally distributed; the same holds of course for Bondal et al.'s (1987) replication. Again, for the sake of comparability, we will also use standard deviations below, the ideal method would involve bootstrapping or even a permutational approaches to compute a more precise range of MLU values from the samples (cf. Gries, 2006 for exemplification).

[3] Depending on the particular measure that is used, distance matrices and similarity matrices can often be thought as derivative of each other; we will use the terms *distance matrix* and *distance measure* but nothing here hinges on this terminological choice.

[4] We performed all computations and generate all graphics in this paper using R for Windows 2.4; cf. R Development Core Team (2006).

[5] Note that if at a later stage the data for these two recordings are merged with the next one (at age 1;04.11, 1.36347), then the new mean is weighted to yield $(2\cdot1.33139+1.36347)/3=1.342083$.