Quantifying the development of inflectional diversity

[Running headline: Quantifying inflectional development]

Aris Xanthos, *University of Lausanne*
Steven Gillis, *University of Antwerp*

Please address correspondence to:

Aris Xanthos
Department of computer science and mathematical methods
University of Lausanne
Anthropole
CH-1015 Lausanne

Phone: +41 21 692 30 25
Fax: +41 21 692 30 45

Email: Aris.Xanthos@unil.ch

Abstract

This study introduces a new metric for assessing the inflectional diversity of morphologically analyzed language transcripts. The proposed metric is based on the intuitive notion of *mean size of paradigm* (MSP) and makes extensive use of random sampling procedures for normalization purposes. This approach is systematically evaluated on the basis of large sets of Dutch acquisition corpora, including both child speech and child-directed speech. It is shown to be an efficient way of controlling for sample size in the measurement of inflectional diversity, as well as a suitable method for assessing inflectional development in longitudinal data. MSP is compared with ID ('Inflectional Diversity') introduced by Malvern, Richards, Chipere and Durán (2004).

Keywords

Developmental variables; inflectional diversity; longitudinal corpora; mean size of paradigm; random sampling.

Quantifying the development of inflectional diversity

Introduction

Many developmental measures are designed to assess the increasing complexity or richness of a child's productions as reflected in longitudinal speech transcripts. Such indexes can be broadly classified according to two criteria: the aspect of linguistic structure they address (e.g. phonology, grammar, lexicon, etc.) and the dimension of linguistic richness they capture (either syntagmatic or paradigmatic). Mean length of utterance (MLU) and type-token ratio (TTR) are classical examples of this. MLU is the average number of words or morphemes per utterance in a sample (see Brown, 1973), and thus it is a measure of syntagmatic richness in the domain of utterances. TTR is the ratio of the number of distinct words (types) to the total number of words (tokens) in a sample (see e.g. Malvern, Richards, Chipere, & Durán, 2004), and as such it constitutes a measure of paradigmatic richness at the lexical level, or *lexical diversity*.

Lexical diversity has been the topic of an impressive number of studies in quantitative and corpus-based linguistics, as reviewed by Malvern et al. (2004). By contrast, little or no research has systematically addressed the evaluation of *morphological* or *inflectional diversity* until recently. As far as child language studies are concerned, this methodological gap is surprising, considering the distinct interest of many researchers in charting morphological development (see e.g. Aksu-Koç, 1998; Ravid & Farah, 1999; Thordardottir, Weismer, & Evans, 2002; Bassano, Laaha, Maillochon & Dressler, 2004), as well as the wide availability of computer tools for morphological analysis (MacWhinney, 2000; Goldsmith, 2001).

In this paper we will be concerned with quantifying morphological diversity, or in other words, quantifying the diversity of *wordforms* that are related to a given *lemma*. For instance, the English verb lemma BREAK is related to inflected wordforms such as 'I *break*' or 'he *breaks*', derived wordforms such as the adjective *breakable*, or compounds such as *breakneck*.[1] In particular, we aim to propose a specific measure for quantifying *inflectional diversity*, leaving aside other aspects of morphology (derivation, compounding) that also contribute to the overall diversity of the morphological system of a language.

Measures of morphological diversity

To the best of our knowledge, there have been very few proposals regarding the quantitative assessment of morphological diversity. Since the present paper focuses on *paradigmatic* morphological richness, we explicitly leave aside measures of *syntagmatic* morphological richness in this review, such as Greenberg's index of synthesis, defined as the average number of morphemes per word in a corpus (Greenberg, 1954), or Nichols' morphological complexity, defined as the maximum number of positions where an inflectional morpheme may occur in a sentence (Nichols, 1992). We also exclude the compression-based approach of Juola (1998), on the grounds that it mixes paradigmatic and syntagmatic complexity, and therefore does not qualify as a specific index of diversity.

Measures of morphological diversity at the level of individual lemmas are introduced by Schreuder and Baayen (1997) as well as Blanche-Benveniste and Adam (1999). They use indexes that characterize the size of a specific inflectional or derivational paradigm, and this approach is extended to account for the entropy (see Appendix) of a paradigm by Moscoso del Prado Martín, Kostíc, and Baayen (2004).

However, no attempt is made to generalize these indexes to *sets* of lemmas, which is necessary in order to provide a global evaluation of the diversity of a given morphological system or for assessing diversity within corpora.

Inflectional diversity as measured by the average number of inflected wordforms per lemma is used by i.a. Stephany (1985), Küntay and Slobin (1996), Laaha (2004), Ogura, Dale, Yamashita, Murase, and Mahieu (2006). This way of measuring inflectional diversity is intuitively appealing, since it captures the idea that a child who uses the forms *walk*, *walks*, *walked* and *walking* of the English verb lemma WALK, has a richer inflectional system as compared to a child that only uses the form *walk* – all other things being equal. Averaging over the entire set of lemmas in a representative sample, it is possible to get an estimate of the inflectional richness of the child's language production. This average is the basic metric that we propose to call *mean size of paradigm* (MSP) in what follows. However none of the authors mentioned above take into account the crucial issue of the dependence of MSP on size of the language sample studied. Indeed, it is well established that measures of lexical diversity are dependent on the size of the sample used for their estimation (Tweedie & Baayen, 1998; Malvern et al., 2004; Tomasello & Stahl, 2004) and this holds for measures of morphological and inflectional diversity as well, as we will show in section *MSP and sample size*.

Inflectional diversity (ID)

Malvern et al. (2004) propose a measure of inflectional diversity which they call ID.[2] It is the only measure in the literature that explicitly takes into account the relation between diversity and sample size, and as such it deserves particular attention in this review.

In order to understand the definition of ID, it is necessary to consider first the definition of D, a measure of *lexical* diversity that has been proposed by the same authors (Malvern et al., 2004). The rationale behind the definition of D is based on the observation that type-token ratio (TTR), the most widely used measure of lexical diversity, is dependent on sample size: to be more precise, TTR tends to *decrease* when sample size increases. Indeed, it is at least intuitively clear that as a language sample grows, the chance that a particular word is repeated increases given the set of words of the language. On these grounds, Malvern et al. argue that merely reporting the TTR value of a given sample is not a proper assessment of lexical diversity. Such an assessment should rather aim at characterizing the entire *curve* of TTR as a function of sample size, for the population (in a statistical sense) from which the sample is drawn.

When trying to implement this idea, the first problem is that in principle, we do not have access to the population in question, but only to a single sample containing a fixed number of tokens. Obviously, there is no way of *observing* the TTR values corresponding to larger sample sizes than what is actually available. On the other hand, it is possible to estimate the TTR values corresponding to *smaller* sizes by randomly drawing subsamples from the original sample and reporting their TTR values. In statistics, this way of resampling the data is referred to as *bootstrapping* (Efron, 1979; Baayen, 2008). For example, given a sample of 100 tokens, we may estimate the TTR value corresponding to a sample of 50 tokens (drawn from the same population) by constructing a number of random subsamples of size 50 (drawn from the original sample), then measuring the TTR value of each subsample and finally computing their average. By repeating this procedure for an arbitrary range of sizes

(35 to 50 tokens in the case of D), we may produce an empirical estimate of the curve of TTR relative to sample size or, to be precise, a segment of such a curve.

Thus, bootstrapping procedures make it possible to study the behavior of TTR over a range of sample sizes – even though the data consist of a single sample. The second step in the methodology proposed by Malvern et al. (2004) aims at describing the empirical curve of TTR by means of a single numeric value, which is the value of measure D for the sample in question. Clearly, identifying a curve, i.e. a series of pairs of numbers, to a single number involves a loss of information. This is comparable to reporting the average of a variable instead of its entire distribution: several distinct distributions have the same average, and similarly, several distinct TTR curves will be associated with the same value of D. On the other hand, provided that this value is calculated in an appropriate way, the greater ease with which it can be appreciated and manipulated (as opposed to a curve segment) should counterbalance the loss of information.

In order to determine the value of D associated with a given empirical TTR curve, Malvern et al. (2004) use a method that relies on an important assumption: that the relationship between TTR and sample size (N) can be described by a particular equation with a single free parameter D:

(1)

Regardless of the mathematical details, the main implication of this assumption is that the TTR value associated with any sample of size N depends only on the value of parameter D. In other words, once the value of D is known, we may let the sample size N vary over an arbitrary range of values and calculate the corresponding values

of TTR, which amounts to constructing a curve of TTR as a function of sample size. Thus, each possible value of *D* defines a TTR curve – a curve which may be referred to as *theoretical* (as opposed to empirical), in the sense that it does not stem from observation but from the adoption of the mathematical model expressed in (1).

In this way, determining the value of measure D for a given sample can be seen as an optimization problem. The goal is to find, among all theoretical curves corresponding to possible values of parameter *D* in equation (1), the curve that is most similar to the empirical TTR curve estimated from the sample. Malvern et al. (2004) use least-square fitting to find the optimal theoretical curve, and the value of *D* corresponding to this curve is reported as the value of measure D for the sample in question.

As noted by Malvern et al. (2004), the estimation of D displays considerable variation depending on what counts as a word type. D is highest when the unit of analysis is the inflected wordform (thus the English verb forms *go*, *goes* and *went* count as three types). It is somewhat lower when the unit is the stem (*go* and *goes* count as one type and *went* as another). It is lowest when the unit is the root or lemma (*go*, *goes* and *went* count as a single type). This means that we may distinguish between several versions of the measure: $D^{wordforms}$, D^{stems}, and D^{roots}.

Malvern et al. (2004) propose to view the *difference* between $D^{wordforms}$ and D^{stems}, as well as the difference between $D^{wordforms}$ and D^{roots} as measures of inflectional diversity which they call ID^{stems} and ID^{roots} respectively. Based on evidence from longitudinal corpora, they show that these measures correlate well with age and other developmental variables. Furthermore, comparison of development patterns in

English and Spanish children shows that ID successfully captures the difference in inflectional diversity of these languages (as reflected in samples of child speech).

The need for a new measure

While many arguments speak in favor of the proposal of Malvern et al (2004), there remain several points of concern. An important practical issue is that the unit in which ID is expressed has no meaningful interpretation. Empirical tests with artificially constructed samples show that ID can take values as low as 0 and as high as 1,200 (for a sample of 50 tokens). While it is always possible to compare two ID values, appreciating the meaning of a single ID value in absolute terms proves generally problematic, as will become apparent in the discussion of our results below.

Another shortcoming of ID relates to the adoption of the mathematical model expressed in equation (1). This model embodies a very specific assumption about the relation between the number of types and tokens in language samples. Making this assumption is the price that Malvern et al (2004) are willing to pay in order to be able to summarize a whole segment of TTR curve by means of a single numeric value. They argue that among models that have been proposed in the literature (and more precisely among models with a single free parameter), this particular model yields the best fit with empirical TTR curves for the small sample sizes that are characteristic of child language studies. However, even in this particular research field, the need to work with larger samples cannot be entirely neglected. In this perspective, the adoption of a model that is specifically tailored for small samples appears as an unfortunate loss of generality.

A last and more subtle issue concerns the definition of ID in *subtractive* terms. It seems reasonable that a measure of inflectional diversity should account for the

discrepancy that exists between the lexical diversity in terms of lemmas and the lexical diversity in terms of inflected wordforms. However, it is not clear that inflectional diversity is best (or even appropriately) expressed as the *difference* between these two kinds of lexical diversities. To see why this can be troublesome, consider the calculation of $ID^{roots} = D^{wordforms} - D^{roots}$ for two samples of identical size (so that sample size is not an issue).[3] Suppose further that both samples come from the same language, but for some arbitrary reason (developmental, pragmatic, etc.), one sample contains a larger number of distinct lemmas than the other. On the grounds that there cannot be less inflected wordforms than lemmas in a corpus, we may expect that this corpus will have a higher lexical diversity both in terms of lemmas and in terms of wordforms.

To make matters concrete, suppose that $D^{wordforms}$ is 20 for the first sample and 80 for the second, and that D^{roots} is 10 for the first sample and 50 for the second. The resulting values of ID are $20 - 10 = 10$ for the first sample and $80 - 50 = 30$ for the second. Hence it would seem that inflectional diversity increases together with lexical diversity, yet this is likely to be wrong. Indeed, given that sample size remains constant, any increase in the diversity of lemma is matched by a corresponding decrease in the average frequency of lemmas. As more distinct lemmas occur, each of them has less frequent occurrences, which means less space for deploying the variety of its inflected wordforms. Rarer inflections are thus less likely to appear in the sample, and on average a lemma will tend to have a smaller number of distinct wordforms. Overall, a *decrease* in inflectional diversity should occur as a result of the increase in lexical diversity.

This example shows that in the context of an increase in lexical diversity (and such increases are obviously quite frequent in acquisition data), ID is liable to detect

spurious increases in inflectional diversity – increases that are mere side-effects of the subtractive definition of the measure. Along with the other concerns expressed above, this suggests that ID is not a flawless measure of inflectional diversity.

Present study

The aim of the present study is to introduce a new metric for measuring inflectional diversity, based on the intuitive notion of *mean size of paradigm* (*MSP*). The remainder of this paper is organized as follows. In the next section, we give a conceptual introduction to the basic definition of MSP, and present a random sampling procedure intended to control for sample size. In the Method section, we describe the setup of several experiments designed to evaluate the behavior of the proposed metric with regard to sample size, as well as its adequacy as a developmental variable; the metric is also systematically compared with the ID measure introduced by Malvern et al. (2004). The main outcomes of the experiments are summarized in the Results section, and in the Discussion section we argue that the proposed metric is both a more intuitive and a more reliable measure of inflectional diversity than ID.

Mean size of paradigm (MSP)

Basic definition

In this section, we introduce the simplest version of the measure that we propose to call *mean size of paradigm* (*MSP*); a more formal and general characterization can be found in the Appendix. The definition of MSP relies on a basic model of inflectional morphology. In this conception, a *morphological system* merely characterizes the classification of inflected *wordforms* (or simply *forms*) into *lemmas*. The set of distinct forms corresponding to a given lemma is called a *paradigm*. The set of all

inflected forms in the system is called the *inflected lexicon* and is represented by the symbol *F*. The set of lemmas is called the *root* lexicon and is denoted by *L*. By convention, |*F*| and |*L*| represent the number of wordforms (types) and lemmas (types) respectively.

These values can be easily computed, as in the following example. Consider a sample consisting of *N* = 5 English inflected verb tokens:

(2) *have, are, have, am, are*

The corresponding root lexicon *L* contains |*L*| = 2 lemmas (HAVE and BE), and the inflected lexicon *F* contains |*F*| = 3 wordforms (*have*, *are*, and *am*).

The simplest form of MSP is defined as the ratio of the size of the inflected lexicon to the size of the root lexicon:

(3) $$\text{MSP} := \frac{|F|}{|L|}$$

In effect, this corresponds to the mean number of inflected wordforms per lemma. In our example, we find that MSP = 3/2 = 1.5 forms per lemma. If the two instances of *have* were removed from the sample, the MSP would increase to 2/1 = 2. If the two instances of *are* were removed instead, it would fall to 2/2 = 1. Notice that this version of MSP depends only on type frequencies: it does not matter that BE is more frequent than HAVE, nor that *are* is more frequent than *am*. Two ways of including token frequencies in the calculation of MSP are considered in the Appendix.

By definition, MSP is functionally related to lexical diversity as measured by the size of the inflected lexicon |*F*| and the size of the root lexicon |*L*|. Indeed, inflected lexical diversity is the product of root lexical diversity and inflectional diversity: |*F*| = |*L*| · MSP. This is one of the most important differences between this approach of

inflectional diversity and that of Malvern et al. (2004): the relation between lexical and inflectional diversity is envisioned as a multiplicative one. By contrast, ID relies on an additive conception of the same relation: $D^{wordforms} = D^{roots} + ID^{roots}$.

As a consequence of its relation with lexical diversity, MSP ranges between 1 and $|F|$. It is minimal and equal to 1 if and only if the sample contains only one (possibly repeated) form of each lemma; in this case, root lexical diversity is maximal and accounts for the whole inflected lexical diversity: $MSP = 1 \Leftrightarrow |L| = |F|$. On the other hand, MSP is maximal and equal to $|F|$ if and only if there is only a single (possibly repeated) lemma in the sample; in this case, root lexical diversity is minimal and inflectional diversity accounts for the whole inflected lexical diversity: $MSP = |F| \Leftrightarrow |L| = 1$.

Thus, $|F|$ is the maximal value of both $|L|$ and MSP. Now, the maximal value of $|F|$ itself is set by the size N of the sample. Indeed the following inequalities hold: $1 \leq |L| \leq |F| \leq N$. It follows from this that the maximal value of MSP depends on sample size. In the *Results* section below, we provide empirical evidence supporting the hypothesis that MSP is *generally* dependent on sample size.

Normalized MSP

In order to control for sample size in the measurement of MSP, we propose to use a method inspired by the work of Johnson (1944) and colleagues on the normalization of TTR:

> TTR's for samples of different magnitudes can be made comparable by dividing each sample into like-sized segments of, say, 100 word each, computing the TTR for each segment and then averaging segmental TTR's for each sample. (Johnson, 1944, p.2)

Johnson's approach has been criticized by Malvern et al. (2004) mainly for three reasons:

1. Its adequacy for comparative purposes relies on the use of an arbitrary parameter (the number of tokens per segment, e.g. 100).
2. It characterizes only a single point on the curve of TTR as a function of sample size (as opposed to D, which characterizes the whole curve).
3. Merely splitting a sample into a sequence of segments is not a satisfactory method, because structural features of a segment may affect the corresponding TTR, and because it results in a loss of data when the number of tokens per segment is not a factor of the original sample size.

We believe that the first and second objections do not dismiss the normalization procedure of Johnson (1944) – at least not in comparison to the procedure adopted for the computation of D. As to the first objection, the range of sample sizes over which the empirical TTR curve is estimated in the calculation of D (35 to 50 tokens, see section *Inflectional diversity*) is as arbitrary as the number of tokens per segment in Johnson's approach. We will discuss some arguments that bear on this question after considering the issue of the number of subsamples below. For the time being, we will simply assume that the number of tokens per subsample is set to some arbitrary value *S*.

As to the second objection, we have seen in section *Inflectional diversity* that in order to describe an entire TTR curve with a single numeric value, it is necessary to adopt a particular model of the relation between TTR and sample size, such as the one that is expressed in equation (1). To that extent, a considerable part of the supplementary information brought by D, compared to the approach of Johnson

(1944), lies in a theoretical assumption rather than in the data – and this assumption can only be shown to be a good approximation for a certain set of data.

As to the third objection raised by Malvern et al. (2004), we fully agree with their criticism and with their solution to the problem. Instead of simply segmenting the sample into consecutive segments of equal size as proposed by Johnson (1944), Malvern et al. propose to construct a number of equally sized subsamples of the corpus by random selection (see section *Inflectional diversity*). Thus, for each subsample, a fixed number of tokens is selected from the entire corpus. In constructing a particular subsample, each token of the corpus can be selected only once (i.e., sampling without replacement), though that token can reappear in several different subsamples.

Our implementation of the random sampling procedure differs from that of Malvern et al. (2004) as regards the number of subsamples to be drawn. While this number is constant and equal to 100 in the approach of Malvern et al., we propose to make it a function of the size of the entire corpus and the size of subsamples. In more formal terms, this is accomplished as follows: let N be the total size of the corpus, and S the number of tokens per subsample, then we construct B subsamples, where $B := N / S$ (rounded to the closest integer). The aim of this definition of B is to ensure that, *on average*, a token in the original corpus is sampled only once in the whole set of subsamples. In fact, this is comparable to the approach of Johnson (1944), where the number of "like-sized segments" is implicitly defined as a function of the size of the corpus and the size of the segments.[4]

In light of what precedes, we define the *normalized* MSP (*over S tokens*) of a given sample (of size N) as the average MSP measured on B subsamples of S tokens randomly drawn from the entire sample, with S being an arbitrary parameter and

$B := N / S$ (rounded to the closest integer). In order to alleviate the terminology, we will often refer to this quantity as the value of MSP(*S*) for the sample in question.

The question of the optimal value of parameter *S*, the number of tokens per subsample, is a delicate matter and there may well be no definitive answer to it. An obvious constraint is that *S* must be less than or equal to *N*, the size of the sample on which the computation of MSP(*S*) is performed. In practice, the constraint is even more stringent as *S* must be less than or equal to the size of the *smallest* sample in a *collection* of samples to be compared by means of MSP(*S*). Furthermore, for any sample of size *N*, setting *S* to a smaller value results in a larger value of *B*, the number of subsamples to be drawn, and thus yields a better estimate of the variance of the measure. For these reasons, it is desirable that *S* be set to a value that is small relatively to sample size *N*. On the other hand, a larger value of *S* generally results in a value of MSP(*S*) that is closer to the MSP of the entire sample: by construction, MSP(*N*) is equal to the MSP of the entire sample. Thus, the setting of *S* is the locus of a delicate trade-off between the desire to have a better estimate of the variance of MSP(*S*) and the desire to capture a larger amount of the diversity present in the original data.

Method

This section describes the setup of several experiments conducted in order to give an empirical assessment of the behavior of raw and normalized MSP with regard to sample size, as well as the adequacy of normalized MSP for charting inflectional development. Normalized MSP is also systematically compared with the ID measure proposed by Malvern et al. (2004).

Data

In this study we used a corpus of child directed speech (CDS) and a corpus of children's speech (CS). All corpora consist of longitudinal conversational data of children acquiring Dutch as their first language and are publicly available via the CHILDES database (MacWhinney, 2000). Further details about the children's personal records, about the recordings and the transcription practices can be found in the CHILDES database manual (http://childes.psy.cmu.edu/manuals/).

The corpus of CDS was extracted from the Dutch section of the CHILDES database, and comprised the subcorpora Abel, Daan, Iris, Josse, Laura, Matthijs, Niek, Peter, Sarah, and Tom. The corpus consisted of 1,030,296 word tokens, of which 202,858 were verbs and 138,914 nouns. The transcriptions were tagged for part of speech, morphologically decomposed, and lemmatized. As an additional step in the preprocessing of the data, particle verbs were merged into single lemmas (e.g. forms of AAN#KOMEN 'to arrive', AF#KOMEN 'to descend', BIJ#KOMEN 'to recover', etc. were considered to be forms of KOMEN 'to come'), given the separability of verb prefixes in Dutch. Moreover, homophonous inflected forms were treated as a single form, e.g. no distinction was made between the infinitive and the present plural uses of the form *komen*, as they are formally indistinguishable.

The corpus of children's speech, consisted of the subcorpora Abel, Arnold, Daan, Diederik, Gijs, Joost, Katelijne, Laura, Marie, Matthijs, and Peter of the Dutch section of the CHILDES database, annotated in the same way as the corpus of CDS. In order to allow for a longitudinal comparison of the children's language samples, the selection was restricted to the data between the ages 2;0 and 3;0, since Dutch CHILDES data are only available for a sufficient number of children for this narrow period. Table 1 contains an overview of the subcorpora: for each child, the number of

noun and verb tokens is provided, as well as the range of the MLU values (MLU in words at 2;0 and at 3;0).

[INSERT TABLE 1 ABOUT HERE]

Experiment 1: MSP and sample size

In order to get a sense of how sample size affects raw MSP, a bootstrapping procedure was applied to the Dutch CDS corpus. Since Dutch verbal inflection has a potentially much higher diversity than nominal inflection, verbs and nouns were treated as separate subcorpora in all the experiments, which allows us to monitor the behavior of MSP at different levels of inflectional diversity.

Each subcorpus (verbs and nouns) was processed as follows. We constructed a number of samples of increasing size: 10 tokens, 20 tokens, 40 tokens, …, 40,960 tokens, 80,920 tokens. For each of these 14 sample sizes, 100 samples were constructed, thus yielding 100 samples of 10 tokens, 100 samples of 20 tokens, and so on, randomly drawn from the original subcorpus. Each token of the subcorpus was selected only once while constructing a given sample (i.e., sampling without replacement), though this token could recur in several different samples. Then, for each sample size, we computed the MSP of each of these 100 samples and we report the average and standard deviation of these 100 MSP values.

A (two-tailed) Wilcoxon signed-rank test is used to determine whether, for a given sample size, the average value of raw MSP for nouns is significantly different from the corresponding value for verbs. Because of the limited number of cases in the

comparison (14), the normality assumption is not likely to hold for these data, hence the use of a non-parametric approach.

Experiment 2: normalized MSP and sample size

In order to study the behavior of *normalized* MSP as a function of sample size, a second experiment was run on the corpus of Dutch child-directed speech. The design was similar to the previous experiment. Each subcorpus (nouns and verbs) was processed as follows. First, 100 samples of 50 tokens were randomly drawn using the sampling procedure described in the previous section. For each of these samples, the value of MSP(50) was computed as described in section *Normalized MSP*. The average of these 100 values of MSP(50) is reported along with the corresponding standard deviation. The whole process was repeated with 100 samples of size 100, 200, ..., 25,600, and 51,200. For sizes greater than 500, the average of MSP(500) was also computed.

The decision to use MSP(50) for this experiment was motivated by the desire to make a comparison with ID, the measure proposed by Malvern et al. (2004). In the CLAN implementation provided by CHILDES, the minimum sample size for the computation of ID is 50 tokens, because of the range of sizes used for estimating the empirical TTR curve (35-50 tokens, see section *Inflectional diversity*). Therefore, we have chosen to set the number of tokens per subsample to 50 in the computation of normalized MSP, so that the two measures could be computed over the same range of sample sizes. The results for MSP(500) are reported in order to illustrate the impact of varying the number of tokens per subsample on the value of normalized MSP.

The question whether normalized MSP and ID are affected by sample size is addressed by means of a Spearman correlation test. A Wilcoxon signed-rank test is

used to determine whether MSP(50) differs significantly from MSP(500). Again, the justification for using non-parametric tests lies in the small number of observations: 11 for MSP(50) and ID, 7 for MSP(500).

Experiment 3: normalized MSP and longitudinal data

A last experiment was conducted in order to evaluate the adequacy of normalized MSP as a developmental variable. MSP(50) and ID were computed over the monthly data of each child, separately for verbs and for nouns (excluding proper nouns). Spearman's correlation coefficient between the two metrics and chronological age of the children (in days) was used as a means to test whether the metrics display an increasing trend over time during the observational period. Correlation between the two metrics and another, well-established developmental variable, viz. MLU, was also calculated, as well as the direct correlation between MSP(50) and ID. Spearman's rank correlation was preferred to Pearson's r because it makes no assumption regarding the linearity of the relationship between the variables in question.

The data were further pooled into monthly datasets, and the mean and standard deviation of MSP(50) and ID were computed for each datasets (separately for verbs and for nouns). A Wilcoxon signed-rank test was applied to these values in order to determine whether the 12 monthly values of MSP(50) for verbs were significantly different from the corresponding values of MSP(50) for nouns, and the same procedure was applied to ID.

Results

MSP and sample size

The behavior of raw MSP as a function of sample size is documented in Figure 1. Each curve represents the average MSP value calculated over 100 random samples of 10, 20, ..., 80,920 tokens from a subcorpus of Dutch CDS (verbs and nouns).

[INSERT FIGURE 1 ABOUT HERE]

These data raise two general observations. First, they display a clear non-linear growth of MSP as a function of sample size, irrespective of the potential richness of the inflectional system under consideration. Among simple parametric models (linear regression, with or without logarithmic or power transform), the best fit is obtained by logarithmic regression, with 98% of variance explained for verbs and 91% for nouns. Second, for a given sample size, MSP proves highly sensitive to differences in diversity between verbal and nominal inflection (Wilcoxon test: $z = -3.3$, $p < 0.001$). The main conclusion to be drawn from these results is that raw MSP is dependent on sample size, and hence it is not a valid measure of inflectional diversity unless sample size is controlled for.

Normalized MSP and sample size

The relationship between normalized MSP and sample size is displayed in Figure 2, which shows the curves of MSP(50) and MSP(500) plotted against sample size (on a logarithmic scale) for verbs.[5]

[INSERT FIGURE 2 ABOUT HERE]

While MSP(50) and MSP(500) yield clearly different evaluations of inflectional diversity on average (Wilcoxon test: $z = -2.37$, $p = 0.018$), each of them remains nearly constant over the range of sample sizes that have been studied. This result is confirmed statistically by the absence of a significant correlation between each metric and sample size: for MSP(50), Spearman's $\rho = -0.31$ and $p = 0.36$ ($n = 11$) and for MSP(500), $\rho = -0.36$ and $p = 0.44$ ($n = 7$).

Figure 2 also displays a clear decrease in standard deviation as sample size grows. This accounts for the fact that the *precision* of normalized MSP increases with the size of the sample. Indeed, being able to control for sample size does not mean that there is no advantage in having access to more data.

[INSERT FIGURE 3 ABOUT HERE]

Figure 3 shows the behavior of ID as a function of sample size. It suggests that while ID is reasonably stable for sample sizes ranging from 200 to 51,200 tokens, it tends to produce larger values for smaller sample sizes. This intuition is supported by a highly significant negative correlation between ID and sample size: Spearman's $\rho = -0.83$ and $p = 0.003$ ($n = 11$). Note that, similar to what appears in Figure 2 for normalized MSP, standard deviation clearly decreases as sample size grows.

Normalized MSP and longitudinal data

The development of MSP(50) for nouns and verbs is plotted against the chronological age of children (expressed in days) in Figures 4 and 5. A Spearman correlation test yields a significant increase of MSP(50) for verbs ($\rho = 0.4$, $p < 0.001$, $n = 96$) and a non-significant one for nouns ($\rho = 0.12$, $p = 0.22$, $n = 108$).[6] Hence, for nouns there is hardly any development over time, while the inflectional diversity increases significantly for verbs, which confirms visual inspection of Figures 4 and 5.

[INSERT FIGURES 4 AND 5 ABOUT HERE]

The development of ID for nouns and verbs is displayed in Figures 6 and 7. In marked contrast with the results for MSP(50), ID shows a significant increase in the period studied both for nouns ($\rho = 0.46$, $p < 0.001$, $n = 108$) and for verbs ($\rho = 0.38$, $p < 0.001$, $n = 96$).

[INSERT FIGURES 6 AND 7 ABOUT HERE]

Correlation with a measure of morphosyntactic development, viz. MLU in words, also gives a contrasting picture for MSP(50) and ID. There is a significant positive correlation between MSP(50) for verbs and MLU (Spearman's $\rho = 0.27$, $p = 0.007$, $n = 96$) but not for nouns (Spearman's $\rho = 0.09$, $p = 0.341$, $n = 108$). The situation is the reverse for ID: ID for verbs does not correlate significantly with MLU (Spearman's $\rho = 0.04$, $p = 0.72$, $n = 96$) but ID for nouns does (Spearman's $\rho = 0.27$,

$p = 0.005$, $n = 108$). In spite of these differences, MSP(50) and ID correlate significantly with one another: for nouns, Spearman's $\rho = 0.53$ and $p < 0.001$ ($n = 108$), and for verbs $\rho = 0.48$ and $p < 0.001$ ($n = 96$). Table 2 provides an overview of the correlations between the various metrics.

[INSERT TABLE 2 ABOUT HERE]

Computing the monthly averages of MSP(50) and ID sheds new light on the developmental contrast between the two metrics. Figure 8 displays the monthly averages of MSP(50) over the 11 children in our corpus; in this representation, it is apparent that inflectional diversity is consistently higher for verbs than for nouns (Wilcoxon test: $z = -3.06$, $p = 0.002$). Figure 9 displays the corresponding graph for the development of ID. Here we see almost no difference between the development for nouns and for verbs (Wilcoxon test: $z = -0.63$, $p = 0.53$). Thus MSP(50) reveals a developmental difference for nouns and verbs, while ID does not indicate a different development of both categories.

[INSERT FIGURES 8 AND 9 ABOUT HERE]

Discussion

How to measure inflectional diversity? In this paper we proposed a new metric for that purpose, called normalized Mean Size of Paradigm (MSP). The metric hinges on the intuitively simple idea of computing the average number of inflected wordforms per lemma in a language sample. This idea has already been articulated by various

authors, as reviewed in the introduction. However, it is well-established that measures of diversity are dependent on the size of the sample at hand, and this aspect has never been taken into account in the proposals referred to. Our proposed computation of MSP controls for sample size through a normalization procedure, and it was shown empirically that indeed normalized MSP remains stable across a range of sample sizes (Figure 2), while the basic version does not (Figure 1).

How well does normalized MSP capture the children's actual development of inflectional diversity? When applied to a corpus of 11 Dutch-speaking children between 2;0 and 3;0, MSP(50) for nouns did not correlate significantly with age (in days), while MSP(50) for verbs did correlate significantly with age. So, the question turns up whether this finding can be linked to a qualitative analysis of the development of inflectional complexity in Dutch.

Nominal inflection is quite restricted in Dutch: nouns can only receive number marking. Thus, nouns can appear in their singular form or in their plural form (genitives are extremely rare in spoken Dutch). Recall that simplex nouns and diminutives are treated as separate lemmas. Hence the maximum possible MSP score for Dutch nouns is MSP = 2, which would mean that every noun is found in its singular as well as in its plural form. However, Ravid, Dressler, Nir-Sagiv, Korecky-Kröll, Souman, Rehfeldt, Laaha, Bertl, Basbøll and Gillis (2008: 42) report that in Dutch-speaking children's speech (up to the age of 3;1) only 16% of noun lemmas actually occur in their plural form (and only 7% of all noun tokens are plurals wordforms). Moreover, closer inspection of the cumulative lists of words of the children involved in the present study shows that when we compute the cumulative proportion of noun lemmas with a singular-plural opposition relative to the total number of noun lemmas, the median value is 10.6% (range: 3.9-17.7). This means

that we expect a MSP value close to 1, and indeed, the highest MSP(50) value in Figure 4 is approximately 1.3.

The picture that arises for verbs (Figure 5) shows a significant increase of inflectional diversity between 2;0-3;0. First of all, verbal inflection in Dutch is much richer than its nominal counterpart. Dutch verbs encode the grammatical categories person, number, tense, mood and voice. The categories person (1st, 2nd, 3rd) and number (singular, plural) are expressed synthetically by verbal suffixes, whereas there is no person distinction in the plural, which is formally indistinguishable from the infinitive. The non-finite verb forms include the infinitive (*werk-en* 'to work'), the past participle (*ge-werk-t* 'worked'), and the present participle (*werk-end* 'working'). The category tense is expressed by the present (*ik werk* 'I work', *hij werk-t* 'he works') and the imperfectum (*ik werk-te*, *hij werk-te* 'I/he worked'). These are the only synthetic forms. Thus, there can potentially be quite a few different wordforms for each lemma.[7]

Do (some of) these different forms of a lemma actually show up in children's speech between 2;0 and 3;0? In a case study of the development of verbal paradigm in Dutch, Gillis (2003) found that between 1;5 and 2;5, the child's number of lemmas of main verbs increased from 4 to 95. When looking at the number of lemmas that occurred in the data as 'mini-paradigms' (i.e., at least two non-homophonous wordforms per lemma), it was found that at 1;5 the child had only 4 lemmas and all of them occurred as a single wordform. At 2;0 the number of verb lemmas had increased to 51, 6 two-member mini-paradigms were found as well as 3 mini-paradigms consisting of 3 members. At 2;5, 18 mini-paradigms consisted of 2 members, 7 consisted of 3 members and 3 consisted of 4 members. Hence, the increase of MSP(50) reflects the increase of mini-paradigms reported by Gillis (2003) for a

Dutch-speaking child's development (the child reported in that study develops from MSP(50) = 1 at 1;5 to MSP(50) = 1.24 at 2;0 and further to MSP(50) = 1.39 at 2;5). Overall, the development of MSP matches well the observations resulting from the qualitative analysis of noun and verb early acquisition in Dutch.

In this paper we compared MSP with ID, another measure of inflectional diversity which was introduced by Malvern et al. (2004). MSP is computed as a (normalized) average of the number of inflected wordforms per lemma, while ID is computed as the difference of a (normalized) measure of lexical diversity for wordforms and the corresponding measure for lemmas. The net result of this difference in computation can be readily seen when we compare the actual MSP(50) and ID values. The values of ID range between 0 and 40.68 for nouns and between 0.66 and 23.54 for verbs, while for MSP(50) the values range between 1 and 1.28 for nouns and between 1.08 and 1.97 for verbs. As already indicated, the MSP values have a clear relationship with the actual inflectional paradigms: Dutch noun lemmas can appear in only two forms, viz. the singular and the plural. Thus if each lemma occurs either as a singular or as a plural wordform, then MSP(50) equals exactly 1, and if each lemma occurs as a singular as well as a plural wordform then MSP(50) equals 2. Consequently, a value MSP(50) = 1.28 means that a majority of the nouns attested in the sample appears in only one form (singular or plural), and the remaining noun lemmas occur in two different forms (singular and plural). Such a straightforward interpretation of the values of ID is much more difficult: it is far less transparent what ID = 40.68 means in terms of the nominal inflection in Dutch.

In addition to this matter of transparency, the picture becomes more complicated when we inspect Figures 4 to 7 more closely. For nouns, it appears that MSP(50) does not show a significant increase over time, while ID does show a significant increase.

The development for verbs goes in the same direction for MSP(50) and for ID: they both exhibit a significant increase over time. Thus MSP and ID agree as to the increase of inflectional diversity for verbs, but they disagree as to the increase of the inflectional diversity of nouns. As argued above, MSP(50) appears to be more sensitive in capturing the fact that in Dutch nouns show hardly any inflectional diversity and this is reflected in the fact that children in their third year of life already perform at what appears to be ceiling level, without much further development. The fact that ID shows a significant increase for nouns can arguably be attributed to the subtractive definition of the metric and the growing number of distinct lexical items (as discussed in section *The need for a new measure*).

The difference between MSP(50) and ID appears most clearly when developmental curves are plotted: Figures 8 and 9 show that inflectional diversity as measured by ID is quite similar for nouns and verbs in our corpus of Dutch-speaking children. On the contrary, MSP(50) reveals a significant difference in the development of nouns and verbs. Again, this outcome can be explained in terms of the actual computation of both measures, and it confirms that the development of MSP(50) matches better the results of qualitative analyses.

Finally, the significant correlations between the two measures seem to point at the fact that they both tap the same (or at least similar) underlying morphological complexity phenomena. While this is accomplished directly by MSP (by computing the average number of wordforms per lemma), it is achieved in a more indirect way by ID (by substracting the lexical diversity for lemmas from the lexical diversity for wordforms).

Conclusion

The aim of this paper was to introduce a new methodology for measuring inflectional diversity. We put forth normalized Mean Size of Paradigm (MSP) as a way to quantify the growing inflectional complexity of children's language. MSP is essentially an average number of different wordforms per lemma, and thus captures inflectional diversity. In our formal definition, MSP is further elaborated on as a family of measures that can integrate token frequency as well as type frequency (see Appendix), but this issue is left for future research and operationalization.

In the literature, much concern has been phrased lately about the impact of sample size on the computation of diversity measures. Using a large corpus of child directed speech, we have shown empirically that suitable normalization procedures make it possible to control for sample size when measuring inflectional diversity.

MSP(50) successfully captured the contrasting developmental trends in the inflectional diversity of nouns and verbs in Dutch. Inflectional diversity of Dutch nouns does not increase much over time, thus reflecting the sparsity of nominal marking. In line with what is known about the complexity of the language, MSP(50) for nouns did not show a significant increase over time. The Dutch verbal inflectional system is much more diversified, and this is captured by MSP(50): between 2;0 and 3;0 children start using a more diversified range of verb forms, and this is reflected in a significant increase of the value of MSP(50).

Finally, we compared MSP with another measure of inflectional diversity, viz. ID, proposed by Malvern et al. (2004). Although MSP and ID correlate significantly, ID did not capture the different developmental paths of nominal and verbal inflection in the longitudinal corpus. Hence, although MSP and ID appear to tap approximately the

same morphological developments in a corpus of children's language, MSP appears to be more sensitive to the purely inflectional patterns of development.

Acknowledgements

This research was partly supported by a grant of the Swiss National Science Foundation to the first author, and a grant from the Research Foundation – Flanders to the second author.

We are grateful to a number of colleagues, including François Bavaud, Wolfgang Dressler, Susan Goldin-Meadow, Marianne Kilani-Schoch, Sabine Laaha, Brian Malvern, Dan Slobin, and two anonymous reviewers for stimulating discussions and useful comments on earlier versions of this text.

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, H. (2008). *Analyzing linguistic data*. Cambridge: Cambridge University Press.

Bassano, D., Laaha, S., Maillochon, I., & Dressler, W. U. (2004). Early acquisition of verb grammar and lexical development: Evidence from periphrastic constructions in French and Austrian German. *First Language, 24*, 33-70.

Blanche-Benveniste, C., & Adam, J. P. (1999). La conjugaison des verbes: virtuelle, attestée, défective. *Recherche sur le Français Parlé, 15*, 87-112.

Brown, R. (1973). *A first language: The early stages*. London: George Allen & Unwin Ltd.

De Schutter, G. (1994). Dutch. In E. König & J. van der Auwera (Eds.), *The Germanic Languages* (pp. 439-477). London: Routledge.

Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics, 25*, 220-242.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*, 1-26.

Gillis, S. (2003). A case study of the early acquisition of verbs in Dutch. In D. Bittner, W. U. Dressler & M. Kilani-Schoch (Eds.), *Development of verb inflection in first language acquisition: A cross-linguistic perspective* (pp. 171-203). Berlin: Mouton de Gruyter.

Gillis, S., & De Houwer, A. (Eds.). (1998). *The acquisition of Dutch*. Amsterdam/Philadelphia: Benjamins.

Goldsmith, J. (2001). The unsupervised learning of natural morphology. *Computational Linguistics, 27*, 153-198.

Greenberg, J. (1954). A quantitative approach to morphological typology of language. In R. Spencer (Ed.), *Method and perspective in anthropology* (pp. 192-195). Minneapolis: University of Minnesota Press.

Johnson, W. (1944). Studies in language behaviour: I. A program approach. *Psychological Monographs, 56*, 1-15.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics, 5*, 206-213.

Küntay, A., & Slobin, D. (1996). Listening to a Turkish mother: Some puzzles for acquisition. In D. Slobin, J. Gerhardt, A. Kyratzis & J. Guo (Eds.), *Social interaction, social context, and language* (pp. 265-286). Mahwah: Lawrence Erlbaum.

Laaha, S. (2004). *Développement précoce de la morphologie verbale: une étude comparative sur l'acquisition de l'allemand autrichien et du francais.* Unpublished PhD, University of Vienna - Université Paris 5.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah: Lawrence Erlbaum.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke: Palgrave MacMillan.

Moscoso del Prado Martin, F., Kostic, A., & Baayen, R. (2004). Putting the bits together: An information-theoretical perspective on morphological processing. *Cognition, 94*, 1-18.

Nichols, J. (1992). *Language diversity in space and time*. Chicago: University of Chicago Press.

Ogura, T., Dale, P., Yamashita, Y., Murase, T., & Mahieu, A. (2006). The use of nouns and

verbs by Japanese children and their caregivers in book-reading and toy-play contexts. *Journal of Child Language, 33*, 1-29.

Ravid, D., Dressler, W. U., Nir-Sagiv, B., Korecky-Kröll, K., Souman, A., Rehfelt, K., et al. (2008). Core morphology in child directed speech: Crosslinguistic corpus analyses of noun plurals. In H. Behrens (Ed.), *Corpora in language acquisition research* (pp. 25-60). Amsterdam: Benjamins.

Ravid, D., & Farah, R. (1999). Learning about noun plurals in early Palestinian Arabic. *First Language, 19*, 187-206.

Schreuder, R., & Baayen, R. (1997). How complex simple words can be. *Journal of Memory and Language, 37*, 118-139.

Stephany, U. (1985). *Aspekt, Tempus und Modalität: Zur Entwicklung der Verbalgrammatik in der neugriechischen Kindersprache*. Tübingen: Gunter Narr Verlag.

Thordardottir, E., Ellis Weismer, S., & Evans, J. (2002). Continuity in lexical and morphological development in Icelandic and English-speaking 2 years-old. *First Language, 22*, 3-28.

Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough? *Journal of Child Language*, *31,* 101-121.

Tweedie, F., & Baayen, H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities, 32*, 323-352.

Appendix

The definition of MSP that was given in the Method section is entirely based on *type* frequencies. Indeed, MSP of a given sample was defined as |*F*| / |*L*|, where *F* is the set of wordform types in the sample (the inflected lexicon), *L* is the set of lemma types (the root lexicon), and |*A*| is the number of items in set *A*. In this section, we provide a more general definition of MSP, in which there are two ways of including *token* frequencies: at the level of lemmas and at the level of wordforms. In this perspective, MSP can be conceived as a *family* of measures of inflectional diversity including, among others, the ratio |*F*| / |*L*| that was focused on in the body of the present paper.

The generalized definition of MSP may be written as follows:

$$\text{MSP} := \sum_{l \in L} w(l) \cdot d(l) \tag{4}$$

In this expression, $l \in L$ represents any given lemma in the root lexicon, $d(l)$ stands for some measure of the *diversity* of this lemma's paradigm (i.e. the set of distinct wordforms that correspond to this lemma), and $w(l)$ denotes the *weight* that is assigned to lemma *l* when summing the measure of diversity $d(l)$ over all lemmas in the root lexicon. Thus, MSP is defined as the sum, over all lemmas in the root lexicon, of the diversity of each lemma's paradigm times the weight of this particular lemma.

In this context, the definition of MSP as the ratio |*F*| / |*L*| corresponds to the case where each lemma is assigned the same weight, namely 1 / |*L*|, and the diversity of a lemma's paradigm is simply defined as the number of distinct wordforms in this paradigm, for which we use the notation $\text{variety}(l)$:

(5) $$\frac{|F|}{|L|} = \sum_{l \in L} \frac{1}{|L|} \text{variety}(l)$$

Since all weights are uniform and they sum to 1, (5) defines an unweighted average, which we may call *unweighted variety-based* MSP.

The first way to introduce token frequencies in the definition of MSP is by weighting the contribution of each lemma according to its relative frequency in the sample, which results in the following expression:

(6) $$\sum_{l \in L} \text{frequency}(l) \cdot \text{variety}(l)$$

In a sense, we may say that this *weighted variety-based* version of MSP really measures the inflectional diversity of a given *sample*, whereas its unweighted analogue (5) measures the inflectional diversity of the *lexicon* derived from the sample. As an illustration, we would expect the weighted variant to yield a higher value than the unweighted one when applied to verbal inflection in languages like French or English, for instance. Indeed, weighted MSP gives more importance to frequent lemmas, such as modal and auxiliary verbs, which often have particularly large paradigms in these languages.

The second way to include token frequencies is by using a measure of diversity based on the *entropy* associated to the paradigm of each lemma *l*. Let us introduce the notation $\text{paradigm}(l)$ to represent the set of inflected wordforms associated with *l*. For any wordform *f* in $\text{paradigm}(l)$, let $\text{frequency}(f|l)$ denote the relative frequency of *f within this paradigm*. With these conventions, the entropy of a paradigm can be defined as follows:

(7) $$\text{entropy}(l) := -\sum_{f \in \text{paradigm}(l)} \text{frequency}(f|l) \cdot \ln\left[\text{frequency}(f|l)\right]$$

It follows from definition (7) that the exponential of the entropy of a paradigm, $\exp[\text{entropy}(l)]$ is comprised between 1 and $\text{variety}(l)$. At one extreme, this quantity is equal to 1 when there is only one inflected form in $\text{paradigm}(l)$, with a relative frequency of 1. At the other extreme, it is equal to $\text{variety}(l)$ if all inflected forms in $\text{paradigm}(l)$ have the same relative frequency, namely $1/\text{variety}(l)$. For example, a paradigm containing three forms with relative frequencies 1/3, 1/3, and 1/3 yields $\exp[\text{entropy}(l)] = 3$. For a less uniform paradigm with relative frequencies 3/5, 1/5, and 1/5, $\exp[\text{entropy}(l)] = 2.59$. For a strongly skewed paradigm with relative frequencies 9/10, 1/20, and 1/20, $\exp[\text{entropy}(l)] = 1.48$.

Thus, $\exp[\text{entropy}(l)]$ can be interpreted as a measure of the diversity of a paradigm that accounts for differences in token frequencies between wordforms in this paradigm. Intuitively, the idea is that a paradigm that contains one very frequent form along with two much rarer ones, for instance, could be considered to be more similar to a paradigm with one form than with three forms, in terms of diversity. By defining $d(l)$ as $\exp[\text{entropy}(l)]$ in formula (4), we obtain two *entropy-based* variants of MSP, either unweighted:

(8) $$\sum_{l \in L} \frac{1}{|L|} \exp[\text{entropy}(l)]$$

or weighted:

(9) $$\sum_{l \in L} \text{frequency}(l) \cdot \exp[\text{entropy}(l)]$$

In summary, there are two different ways of including token frequencies in the definition of MSP: either at the level of lemmas, by making MSP a *weighted* average, or at the level of wordforms, by constructing an *entropy-based* version of MSP.

Together, these two independent parameters define the four variants of MSP given in equations (5), (6), (8), and (9). Arguably, the fully unweighted variant (5) and the fully weighted one (9) constitute the most consistent choices.

Figure Captions

Figure 1 Mean Size of Paradigm of Dutch verbs and nouns as a function of sample size (dotted lines represent standard deviations).

Figure 2 MSP(50) and MSP(500) of Dutch verbs as a function of sample size (dotted lines represent standard deviations).

Figure 3 ID (Inflectional Diversity) of Dutch verbs as a function of sample size (dotted lines represent standard deviati

Figure 4 MSP(50) for Dutch nouns as a function of chronological age (in days).

Figure 5 MSP(50) for Dutch verbs as a function of chronological age (in days).

Figure 6 ID for Dutch nouns as a function of chronological age (in days).

Figure 7 ID for Dutch verbs as a function of chronological age (in days).

Figure 8 Development of MSP(50) for nouns and verbs (bars indicate SD).

Figure 9 Development of ID for nouns and verbs (bars indicate SD).

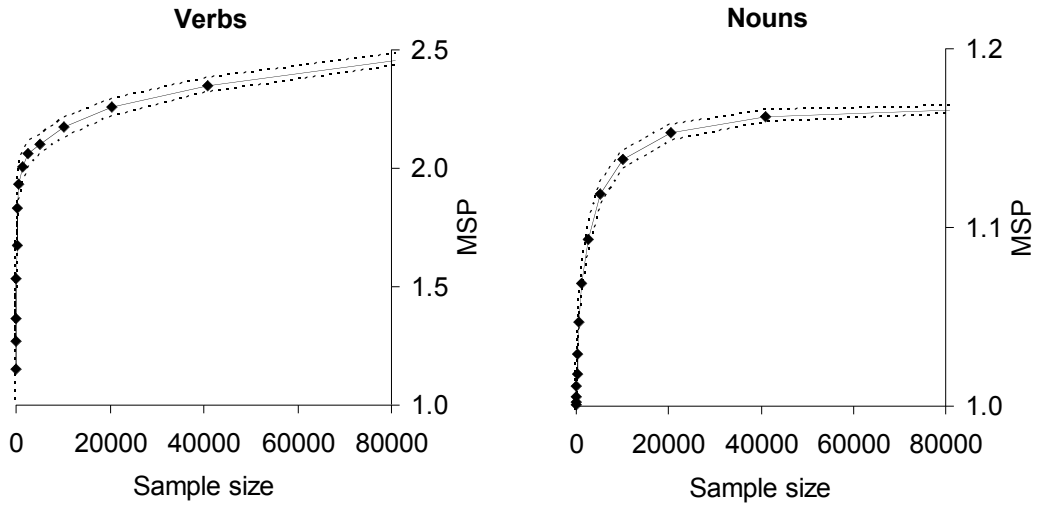

Figure 1 Mean Size of Paradigm of Dutch verbs and nouns as a function of sample size (dotted lines represent standard deviations).

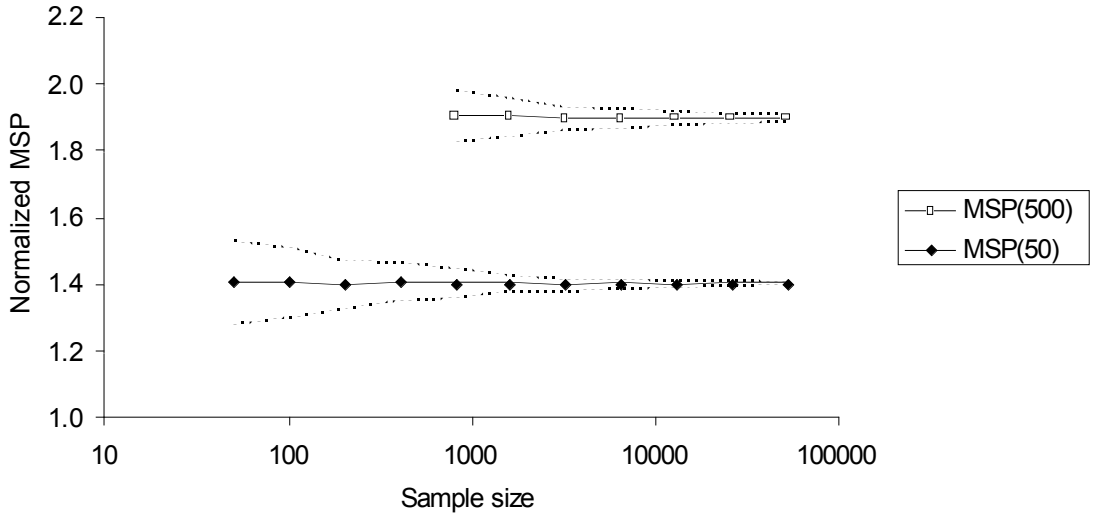

Figure 2 MSP(50) and MSP(500) of Dutch verbs as a function of sample size (dotted lines represent standard deviations).

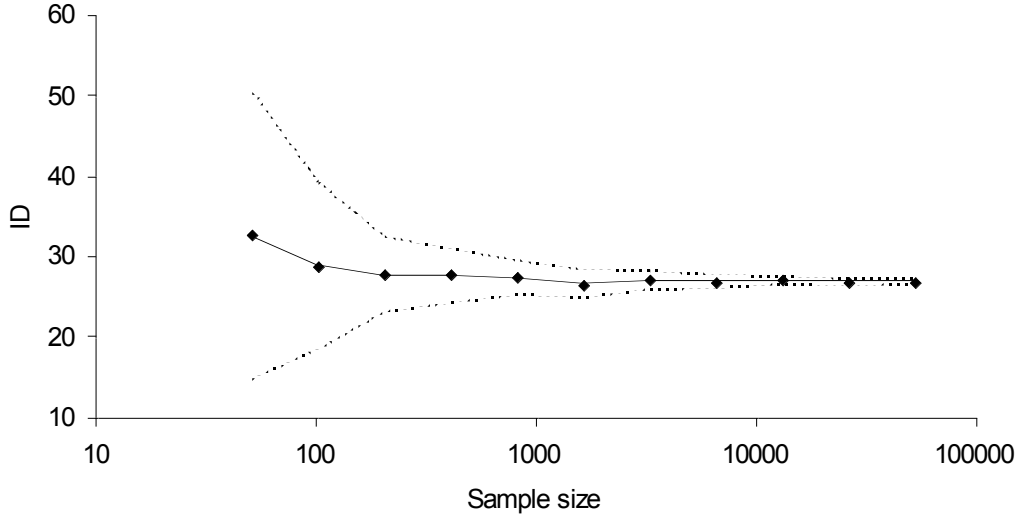

Figure 3 ID (Inflectional Diversity) of Dutch verbs as a function of sample size (dotted lines represent standard deviations).

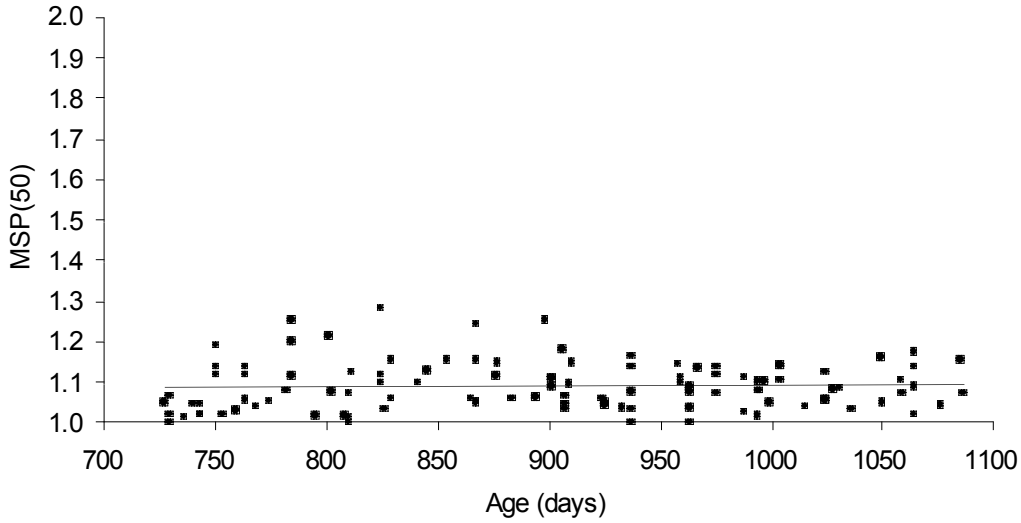

Figure 4 MSP(50) for Dutch nouns as a function of chronological age (in days).

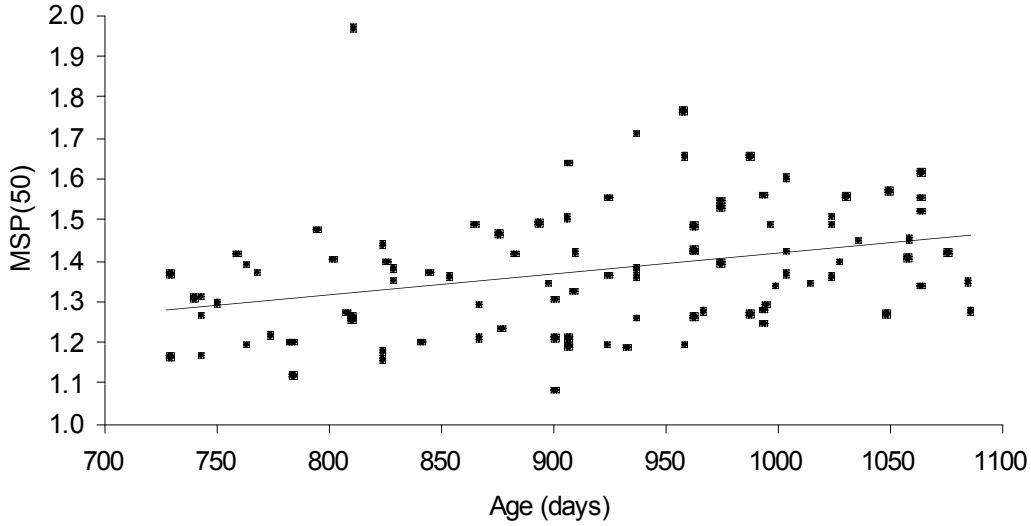

Figure 5 MSP(50) for Dutch verbs as a function of chronological age (in days).

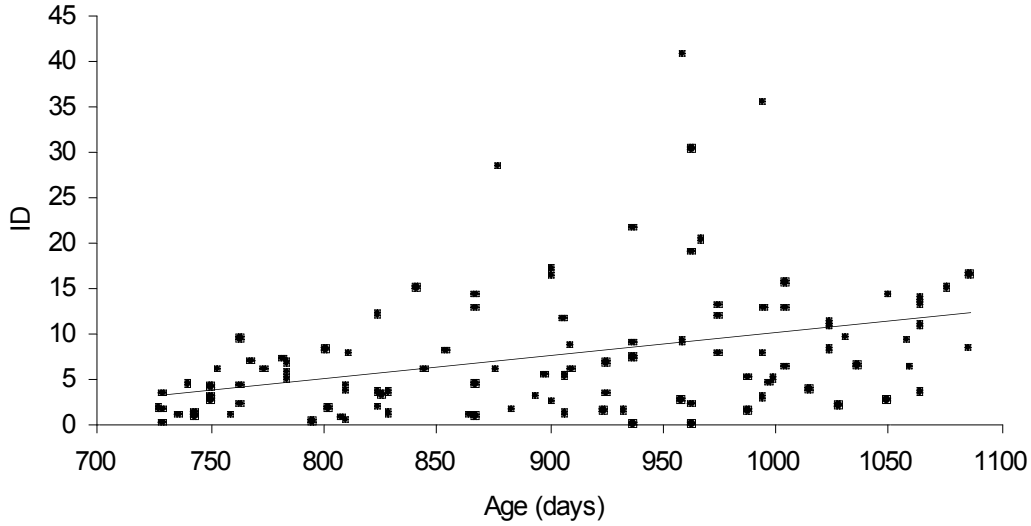

Figure 6 ID for Dutch nouns as a function of chronological age (in days).

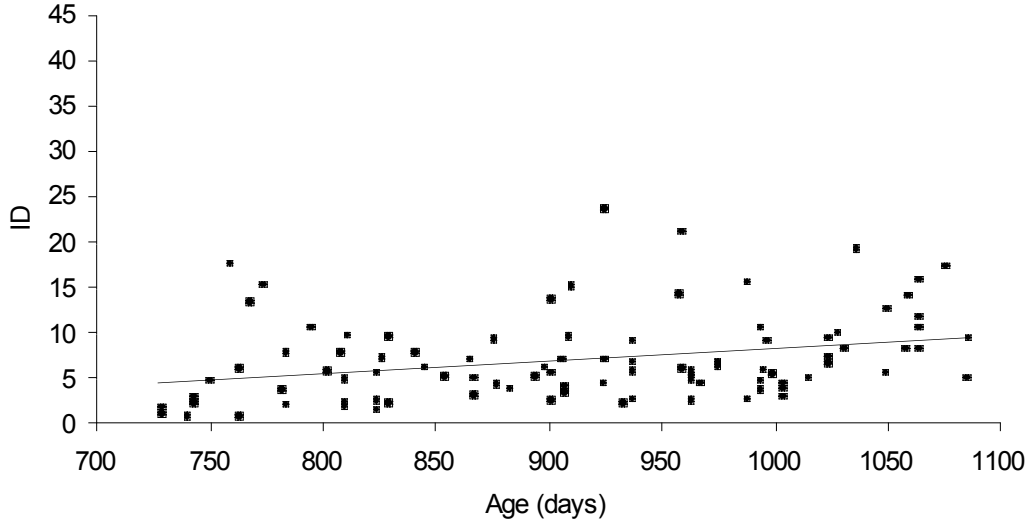

Figure 7 ID for Dutch verbs as a function of chronological age (in days).

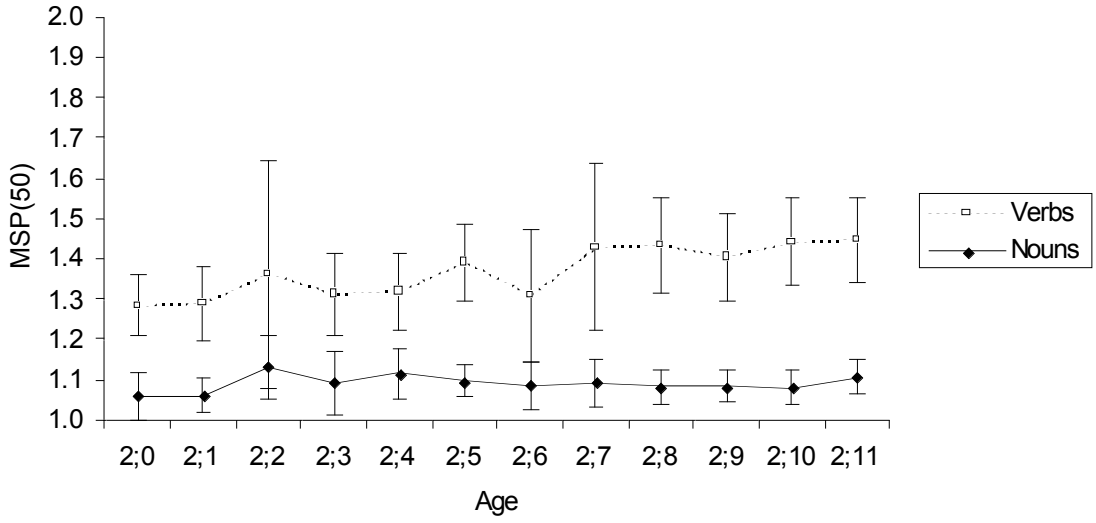

Figure 8 Development of MSP(50) for nouns and verbs (bars indicate SD).

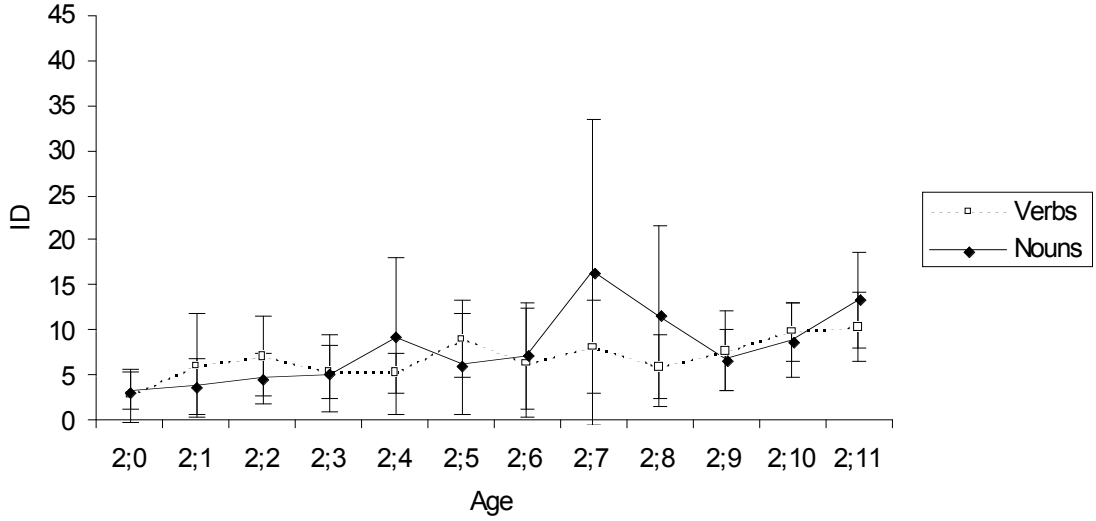

Figure 9 Development of ID for nouns and verbs (bars indicate SD).

Table 1 Overview of the Dutch CHILDES corpora used in the longitudinal experiments.

Child	# Tokens		MLU range
	Nouns	Verbs	
Abel	1,534	2,299	1.330-3.043
Arnold	1,303	982	1.449-4.889
Daan	2,092	2,230	1.101-3.236
Diederik	1,099	865	1.294-4.346
Gijs	1,263	1,358	1.307-5.409
Joost	810	882	1.211-4.096
Katelijne	1,222	1,152	1.107-4.893
Laura	1,835	1,987	1.409-2.709
Marie	1,016	977	1.100-4.545
Matthijs	2,410	1,890	1.457-3.000
Peter	1,850	2,785	1.768-3.408

Table 2 Spearman's correlations between chronological age (in days), MLU in words, ID, and MSP for nouns and verbs.

	MLU	ID Nouns	ID Verbs	MSP(50) Nouns	MSP(50) Verbs
Age	$\rho = 0.65$	$\rho = 0.46$	$\rho = 0.38$	$\rho = 0.12$	$\rho = 0.4$
	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p = 0.22$	$p < 0.001$
MLU		$\rho = 0.27$	$\rho = 0.04$	$\rho = 0.09$	$\rho = 0.27$
		$p = 0.005$	$p = 0.72$	$p = 0.341$	$p = 0.007$
ID Nouns			$\rho = 0.29$	$\rho = 0.53$	$\rho = 0.11$
			$p = 0.005$	$p < 0.001$	$p = 0.299$
ID Verbs				$\rho = 0.16$	$\rho = 0.48$
				$p = 0.113$	$p < 0.001$
MSP(50) Nouns					$\rho = 0.16$
					$p = 0.115$

Notes

[1] By convention, we distinguish lemmas from wordforms by using small caps and italics respectively.

[2] Malvern et al. (2004) actually call their measure *inflectional diversity*. In order to avoid confusions, we will use the acronym ID to refer to it, while reserving the term inflectional diversity for the general notion of paradigmatic inflectional richness.

[3] Note that in this study, we specifically focus on ID^{roots} because the alternative measure that we propose is based on the notion of lemmas (or roots) rather than stems; therefore, we will usually abbreviate ID^{roots} as ID.

[4] We do not dismiss the alternative way of dealing with the number of subsamples, i.e. by setting it to some arbitrarily large constant value. It is the usual practice in the field of bootstrapping methods and it can lead to a considerable reduction of the variance of the measure. Further research will be needed in order to determine the circumstances under which either approach is more appropriate.

[5] The corresponding curves for nouns are not represented here. They are not informative as they remain very close to the minimal value of 1, with a standard deviation nearly equal to zero, regardless of sample size. In effect, this means that the average number of inflected forms per noun lemma in 50 or 500 tokens of our Dutch child-directed speech data is approximately 1.

[6] Both MSP(50) and ID require at least 50 tokens, and while this condition is usually satisfied for nouns, the number of verb tokens is sometimes below that limit, especially at the earliest ages. This explains the discrepancy between the number of observations for verbs (96) and for nouns (108).

[7] For a more elaborate description of Dutch morphology, see De Schutter (1994), and see Gillis and De Houwer (1998) for an overview of the acquisition of Dutch.