Entropy **2010**, *12*, 798-817; doi:10.3390/e12040798

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Article Omission in Dutch Children with SLI: A Processing Approach

Lizet van Ewijk [1,*] and Sergey Avrutin [2]

[1] Universiteit Utrecht, UiL OTS, Janskerkhof 13, 3512 BL, Utrecht, The Netherlands

[2] Universiteit Utrecht, UiL OTS, Trans 10, 3512 JK, Utrecht, The Netherlands; E-Mail: s.avrutin@uu.nl

* Author to whom correspondence should be addressed; E-Mail: e.vanewijk@uu.nl.

Received: 21 December 2009; in revised form: 5 March 2010 / Accepted: 7 April 2010 / Published: 8 April 2010

Abstract: Children with Specific Language Impairment (SLI) show difficulties with grammatical morphology. Based on the data from 12 Dutch children with SLI, an information-theoretical model is proposed in which the noun-article set dependency is modeled as a channel. We propose that reduced capacity of this channel is responsible for article omission. The Kullback-Leibler divergence between input and output distribution of article production provides an index of the channel capacity, which is shown to correlate with the percentage of article omission and to lag behind in SLI development as compared to typically developing children.

Keywords: SLI; processing; articles; Kullback-Leibler divergence

PACS Codes: 89.20. –a, 89.70. –a, 89.70.Cf

1. Introduction

Specific Language Impairment (SLI) is a heterogeneous disorder that impairs language acquisition in children with no obvious cognitive, emotional or social difficulties. Diagnosis of SLI is primarily based on a discrepancy criterion, first suggested by Stark and Tallal [1]. The ICD-10 [2] criteria for example specify that language skills be at least one SD below non-verbal IQ. The incidence of SLI in kindergarten has been estimated to be around 7% [3] and in specialised class units, up to 13% [4].

Although children with SLI show difficulties in all aspects of language, grammatical morphology has often been described as particularly weak in these children (e.g., [5,6]). One of the hallmarks of SLI is the omission or substitution of free and bound grammatical morphemes. Tense marking has been shown to be particularly difficult and for English low levels of accuracy in the set of tense morphemes (-s third person singular, *-ed* regular past, *BE,* and *DO*) has been suggested as a clinical marker for this disorder [7].

Many theories have been proposed to explain the difficulties these children experience with morphosyntax. They can broadly be divided into two categories: linguistic theories claiming that some linguistic information in these children is underspecified or impaired (e.g., Implicit Rule Deficit [8]; Extended Optional Infinitive Account [7]). And processing theories suggesting that the linguistic knowledge for these children is intact, but that they struggle using it due to reduced processing abilities (e.g., Auditory Processing Deficit [9]; Generalised Slowing Hypothesis [10]; Surface Account [11,12]).

In the last two decades, a fair number of studies have been dedicated to a specific subset of grammatical morphemes: the production of articles. Although the pattern and performance on production of articles varies across languages, a weakness in this area is evident for all children with SLI, regardless of the language they are acquiring.

1.1. Article production in SLI

Looking at the production of functional categories in English children with SLI, Leonard [13] found that they produce fewer grammatical elements than language matched younger peers. With regards to article production, they found that all of their ten 3–5 years old children with SLI did produce articles, but significantly less often than the language matched group. In a cross-linguistic study on French and Italian, Le Normand, Leonard and McGregor [14] found that French preschoolers with SLI show similar rates of omissions of articles as their language matched peers and have significantly fewer difficulties than the group of Italian children with SLI. Language matching was based on mean length of utterance (MLU), a standard matching criterion in language acquisition research. Leonard [15] investigated the use of the definite article system in Italian in fifteen preschool children with SLI and found that the SLI children performed significantly worse than their MLU matched peers. Errors of omission were most common.

Various studies have investigated the production of articles in Spanish. Most of these studies found results similar to those found in other languages: children with SLI perform worse than their age-matched peers and errors of omission are by far the most common [16,17]. There is one study on Spanish that reports substitution as the most common error. Restrepo and Gutiérrez-Clellen [18] investigating patterns of article production in school aged children (age 5–7) found that substitutions due to gender difficulties were the most common error produced by their group of SLI children. Overall performance rates, however, were quite high ranging from 77% to 83%.

To provide further insight into the error patterns in article production in Spanish SLI and specifically the role of gender, Anderson and Souto [19] used a set of spontaneous speech samples as well as an elicitation task to study a group of 11 children with SLI (mean age 4 yrs; 10 months). The data consisted of three connected speech samples, as well as an experimental task. The purpose of the

experimental task was to elicit a noun phrase consisting of an article as well as an adjective and noun. Importantly, in Spanish each noun has inherent gender (masculine, feminine and neutral) and both the article and the adjective have to conform to this gender. Adding an adjective to the NP in this experimental setup allowed the researchers to further investigate whether article gender errors are due to lack of gender knowledge of the noun. If substitution errors in the article production are due to lack of grammatical knowledge about noun gender, one would expect the children to make similar errors with the adjectives. The task consisted of a modified barrier game in which the child was asked to describe the order of two pictures of similar objects on a picture card. As each object only differed from the other on one characteristic (e.g., colour) the child had to use an adjective to perform the task accurately. The results of the spontaneous speech analyses showed that the group of SLI children performed significantly worse than their age matched peers. In addition error analyses revealed that omission was the most frequent error (87.2%) followed by substitution of gender (9.5%). On the experimental task the age matched group showed higher accuracy levels (94.5%) than the SLI group (64.3%). Furthermore, errors in the SLI group were predominantly due to omission of the article (78.9%), with some gender substitutions (21.1%). The analysis of the adjective production in this task suggests that the article gender errors that the children produce are neither due to inherent difficulties with gender, nor difficulties with agreement in general. In most cases, when the child produced the wrong gender for the article, the adjective *did* conform to the noun's grammatical gender. In addition each noun was elicited twice and variability was such that one noun was used with the correct article once and an error occurred in the second production. It thus seems that the difficulties that children with SLI experience with the production of articles is not due to lack of grammatical knowledge, but rather that there is a problem with the noun-article connection.

Hansson, Nettelbladt and Leonard [20] investigated the production of definite and indefinite articles in Swedish children with SLI. This study provides further evidence that omission and substitution of articles in children with SLI is not due to lack of grammatical knowledge. In Swedish the indefinite article occurs as a free grammatical morpheme, as in e.g., English, Spanish and Dutch. However, the definite article is marked by a suffix added to the noun. Both indefinite and definite articles furthermore have two phonological forms expressing gender. Definiteness can also be expressed using a free grammatical morpheme, but this only occurs when the noun phrase contains an adjective. In this case, the noun is preceded by a definite article in addition to the suffix. The authors used spontaneous speech as well as an elicitation task to investigate the role of prosody on error patterns in article production. They argue that unstressed syllables occurring in a pre-stress position are more susceptible to omissions and errors than elements immediately following a stressed syllable. Based on this phonological account they therefore hypothesize that indefinite articles will be more problematic to children with SLI than the definite suffix. They tested 13 children with SLI (age range 4; 03 – 5; 07), and MLU control group (age range 2; 09 – 3; 07) and an age-matched group. The age-matched control group performed significantly better than both the SLI and MLU group, whereas the latter two did not differ. Both children with SLI and MLU children used indefinite articles significantly less than definite articles. However, a probe task similar to that used in Anderson and Souto [19] further showed that the children did not have difficulties with indefiniteness per se. On this task the children had to produce article + adjective + noun phrases. As mentioned, in these constructions both definite and indefiniteness are expressed by the use of a free grammatical morpheme. In this probe task, the

children with SLI had difficulties with both types of morphemes. The enhanced performance on definiteness thus only appears when expressed as a suffix. This seems to provide strong evidence for a prosodic explanation of error patterns in article production. However, the results also showed that the SLI group omitted the indefinite neuter article *ett* more often than the neuter indefinite article *en*. This cannot be explained by a prosodic account as both are weak monosyllabic morphemes of cv (consonant–vowel) structure.

In summary, children with SLI have difficulties with the production of articles. The extent of the difficulties seems to be dependent on the language that the child is acquiring. The most common error across all languages is error of omission, although most studies also show some substitution errors. It is unlikely that these errors of substitution are due to lack of grammatical knowledge on the child's part. It has been shown that the children do have knowledge of the noun's gender even when they make gender errors in the article and performance is variable both within and across subjects. Theories assuming that SLI is due to a lack of grammatical knowledge struggle explaining these findings. Furthermore, when compared to typically developing children, SLI children show the same profile across article paradigms. Although the SLI groups in these studies performed significantly worse than their age, and sometimes language matched peers, their pattern of performance was very similar. These factors all seem to indicate that children with SLI do not suffer from a lack of grammatical knowledge, but rather that they are acquiring morphology in the same way as TD children but have a more limited capacity using their knowledge. Although a phonological account proposed by Leonard *et al.* [6] seemed to be able to explain some of the results found in the study on Swedish children, it could not explain all findings. In addition to possible difficulties processing weak phonological information, there thus seems to be another factor contributing to the difficulties of these children to retrieve the (correct) article upon producing the noun.

1.2. Retrieving articles

Activation and retrieval of articles depend on the properties of the noun. Discourse information, semantic information (definiteness/indefiniteness), grammatical information on gender and information on the phonological context that the article will have to be produced in, all play a role in the selection of the correct article. The article system in the lexicon will therefore receive input from different systems at varying times during processing. Alario and Caramazza [21] suggest that determiners are represented by means of a language specific frame with slots for each type of information. These slots must be filled with feature information provided by the noun. For Dutch, it has been shown that article selection and retrieval [22] depend on information on number, gender and definiteness. All slots have to be filled before the correct determiner can be selected, but activation of the determiners occurs as information in the different slots becomes available. Only when all slots are filled and information on gender, definiteness and number are simultaneously active, can the determiner be selected.

Based on the studies discussed so far, we will assume that the children with SLI have all the information from the noun that is necessary to activate the correct article. However, we still find numerous omissions of articles. It appears therefore that the activation of the connections to the article system in these children is insufficient for the (correct) article to be consistently selected and

produced. In short, children with SLI seem to have reduced capacity to select and retrieve the correct article, even when all the required information for selection is there.

Reduced processing capacity has been a much used term in the SLI literature and also a much debated one. One of the main problems with this explanation of SLI is that it lacks specificity and is difficult to measure. In the current paper, we propose a model based on information theory that provides a possibility to measure the degree of complexity, in order to explain article omission patterns and reduced processing capacity in Dutch children with SLI, which (as we show below) can be modeled as a channel capacity.

1.3. Introduction to information theory

Information theory has its roots in two major fields: that of communication engineering and that of statistical theory. In 1949 it was first applied to language and communication theory by Claude Shannon [23]. He investigated the probability of errors occurring when sending messages across a technical channel. The most important discovery he made was that, unlike previously thought, it is not the transmission rate of the information that predicts the number of errors, but the complexity of the encoded information. The complexity measure he created is based on a statistical measure of probability, so that each message has a stated probability of occurrence, or a certain "uncertainty". This uncertainty, or "entropy," is expressed in bits and is calculated as shown in (1):

$$H(x) = -\sum p(x) \log p(x) \quad (1)$$

As the formula shows, the entropy consists of the sum of the informative values of the single elements of the message, in which each of these values is weighed for its share in the total frequency. For each message, a complexity level (H) can thus be calculated. Shannon found that every (technical) channel has a certain capacity. If the amount of entropy of the message that is put into the channel exceeds the channel capacity, the message will get distorted and there will be errors in the output. The channel capacity is thus the amount of entropy a certain channel can cope with in one unit of time [24]. The degree of distortion of the message provides an index of the channel "goodness".

In order to estimate the difference between two probability distributions (and in our case the probability distribution of channel input and output) several measures can be used. Differential entropy defines the mutual information between random variables. Mutual information is a measure of dependence between two variables. It can be calculated by means of the Kullback-Leibler divergence. The mutual information is the KL divergence between the joint probability and the product of the marginal probabilities and is defined as:

$$m_{KL}(P\|Q) = \sum p(x) \log \frac{p(x)}{q(x)} \quad (2)$$

As we will show below, the difference between the input and output probability distribution of articles in SLI speech as measured by the Kullback-Leibler distance, gives a clear prediction of the omission rates.

For our purposes, we do not need to go into all of the technical details of information theory. What is important, however, is that the model is content-independent, that is, the *nature* of the message does not matter for measuring its degree of complexity and for determining the channel capacity. A message

(in its technical sense) can be anything from a single letter to the entire War and Peace; it can be electrical signals sent across a wire, or information shared by DNA molecules. Or, relevant for us, it can be feature information "sent" by a selected noun (which is to be produced) to the set of articles. As we will argue below, it is precisely this process of "sending" the (available) feature information to the article that is weakened in SLI (and in younger typically developing children). Or, in terms of information theory, the channel capacity of the noun-article system in the population is lower than in typically developing children of the same age. Lower in such a way that it allows for making measurable predictions about the distribution and production of articles. Importantly, as an anonymous reviewer points out, a reduction in channel capacity by itself is not necessarily problematic, or, in fact is problematic only if the information is transmitted at maximum rate. Neither the reviewer, nor us think that this is *a priori* the case. What we do want to suggest is that the capacity in SLI is reduced to such an extent that it actually does result in distortion of information flow. At the end of this paper we will also speculate why this might be the case.

1.4. Application of information theory to the processing of language

Although initially the application of information theory on communication focused on technical channels, more recent research has used information theoretical measures to explain online language processing. The first detailed work on information load effects on processing of language focused on inflected morphology and was done by Kostic on Serbian [24,25] and later by others on English [26,27] and Dutch [28]. Although these researchers were not looking at article production (our main focus) but were interested in verbal and nominal morphology, their results did show that entropy as measured by information theoretic tools is a reliable predictor of reaction time of word retrieval and, therefore, of the general processing complexity.

An important measure suggested by Kostic is the individual *information load* of an item. In addition to the frequency of a word, the number of functions a particular inflected word form can perform is taken into account in this measure. Serbian is a richly inflected language and each open class word has an inflectional suffix that specifies its grammatical properties. For nouns case, grammatical number and gender are specified. Some of these cases share the same inflectional suffix. For example, the nominative singular female form of the word grass is *trav-a,* as is the plural genitive female form. Each grammatical form appears with some probability in the language, as does each inflected form. Kostic investigated whether the probability distribution of these inflected nouns could explain variation in reaction time on a lexical decision task. All six inflected forms were presented in a lexical decision experiment. However, Kostic found that frequency of the inflected form alone did not predict the variation in latencies. He discovered that a second important factor determining latencies was the number of syntactic functions and meanings an inflected form could have. He found that while an increase in frequency parallels faster processing, an increase in the number of functions and meanings reduced processing, as the form could be seen as more complex. Kostic suggested that if we divide the frequency of an inflected form by the number of its functions and meaning, we get the average frequency per syntactic function/meaning for a given inflected form. The average frequency per syntactic function/meaning should be expressed relative to the sum of average proportions per function/meaning for other noun forms for a given gender. To calculate the number of bits carried by

each grammatical form, the obtained proportion should undergo a log transform. Thus, if R_e denotes the number of functions and meanings carried by element e and F the frequency of a form, then the weighted amount of information can be expressed as follows:

$$I_e = -\log_2\left(\frac{F_e/R_e}{\sum_e F_e/R_e}\right) \quad (3)$$

This measure is based on the item's frequency as well at the number of functions the word form can have and is described by equation (3). Kostic found that in his lexical decision task processing time variability could almost completely be accounted for by using this measure of individual information load.

De Lange [29] used an information theoretical approach to explain article omission patterns in Dutch, Italian and German typically developing pre-schoolers. She found that the information load for individual articles played a role in the acquisition process of those articles. Articles with higher information loads were used later than those with low information load. In addition to looking at individual articles, she used a modified formula to calculate the entropy for the Dutch, German and Italian article sets. She found that the higher the entropy of these sets, the more articles the children omit and the later they acquire them. More specifically, she found that out of the three languages, the Dutch article set has the highest entropy value. Using longitudinal spontaneous speech data she also found that Dutch children omit articles until a later age than German and Italian children. She argues that young children have limited processing resources and that rather than there being a difference between the brain maturation of Dutch, German and Italian children, the level of complexity of the article systems in the respective language differ. A lower complexity of the article system means that less brain maturation is required to cope with the system. This leads to the finding that Italian children start producing articles at an earlier age than Dutch children.

Ferrer I Cancho and colleagues [30–34] have used several measures based on information theory to investigate a range of linguistic phenomenon. In [30] the authors use mutual information as a measure for the strength of correlation between two words. They argue that strong links between words are a priori harder to establish for high frequency words than lower frequency words. As we will argue below, for the children with SLI this mean that they do not have difficulties with articles and nouns per se, but with the strength of the link between them. As pointed out by an anonymous reviewer, the association between the article and the noun may be weaker than for example the association between noun and adjective (as the results showed in [19]), simply because strong links are more difficult to establish with high frequency words.

Peperkamp *et al.* [35] combine statistical measures from information theory in combination with linguistic constraints to describe the acquisition of allophonic rules in French. One phoneme typically has varying phonetic specifications depending on the context in which the phoneme is produced. Children have to learn which allophones are present in a given language for a given phoneme. Peperkamp *et al.* developed a statistical learning algorithm based on the notion that different allophones of one phoneme generally occur in different contexts. They used Kullback-Leibler

divergence to measure discrepancies in context probabilities for each pair of phonemes. When combined with linguistic filters their algorithm was able to detect allophonic distributions in French.

The above studies show that the human brain is sensitive to complexity of linguistic information and that information with high complexity requires more processing load. Furthermore, the models used in information theory are able to predict and explain variations in processing load by providing individual words or sets of words with a quantitative measure of complexity.

Coming back to the previously mentioned study, De Lange showed that omission patterns of articles in language of typically developing children can be explained as the result of limited processing resources, as shown by cross-linguistic variation. From the studies described above, we know that children with SLI show omission of articles up to a much later age than typically developing (TD) children. In the current paper we would like to propose that the article omission patterns, as well as the distribution patterns of produced articles, can be explained by a model based on information theory. In this model we combine the article selection model by Janssen and Caramazza [22] and the information theory of Shannon [23]. As article selection depends on the information the article set receives from the activated noun, *we represent the linguistic noun–article dependency as Shannon's information channel.*

Our main hypothesis is that children with SLI have a reduced processing capacity and that this leads to distortion in the channel that uses information from the noun to select the correct article. In order to test this hypothesis spontaneous speech samples for a group of Dutch SLI children were analysed. Omission and substitution rates for all articles were noted and the role of sentence position of the article, type of article, sentence type (finite versus infinitive) investigated. Finally, an analysis of the data using and information theoretic approach will be presented.

2. Results and Discussion

2.1. Method

2.1.1. Subjects

For the SLI group, 12 children were selected from the Bol and Kuiken [36] database. In this database children were tested on non-verbal IQ, which fell within normal range for all these children, as reported by Bol and Kuiken. Mean length of utterance measures the mean number of morphemes in an utterance and the verbal utterance score is a measure of how many utterances contain verbs.

Table 1. Age, gender, mean length of utterance (MLU) and verbal utterance (VU) scores. Standard deviation (SD) is provided in brackets for MLU and VU.

Age	Gender	MLU	VU
5; 03 (4; 01–6; 01)	7 boys, 5 girls	3, 5 (2, 2–4, 4)	0, 44 (0, 23–0, 6)

2.1.2. Materials and scoring

Speech samples consisted of spontaneous speech during a free-play session with either a researcher or a speech therapist. For each child at least 100 utterances were used in the analysis. Unintelligible utterances, direct repetitions of an adult utterance, idiomatic expressions, rhymes and songs were excluded from the analyses. For each file the number and type of article used by the child were marked. In addition, all article omissions were marked. All situations in which a noun was produced and in adult speech an article would have been required were counted as omission. Percentages of omission were calculated for each individual article as well as the overall omission rate. These calculations were then used to provide two types of measures:

The *distribution* of article use of the child (*i.e.,* the relative distribution of *de, het* and *een*—the three articles of the Dutch language) and of the required article set were calculated. Note that these measures do not deal with absolute numbers of what is produced or should have been produced. Rather, they provide a measure of the distribution between the three articles. Furthermore, sentence position of the article was noted as well as finiteness of the sentence the article should have been produced in.

2.2. Results

Statistical analyses using Mann-Whitney shows that there is no significant difference between the omission rates of *de* and *het*, *de* and *een* or *een* and *het*.

Table 2. Error analyses: mean overall percentage and SD of omissions and substitutions.

Mean % omission	**Mean % substitution** [37]
36.5 (28)	2.8 (4.1)

Table 3. Percentage of omission and SD per article.

Mean	**De**	**Het**	**Een**
36.5	32.6	53.6	41.8
(28)	(32.5)	(37)	(34.5)

2.2.1. Article omission patterns for *het*

Het omissions were further scrutinised and the nouns divided into two groups: neuter nouns requiring *het* and diminutive nouns that always require *het* regardless of their base noun gender. In Dutch all diminutives take require *het* regardless of their stem. For example *het boek* (the book) becomes *het boekje* (the little book), but *de poes* (the cat) also becomes *het poesje* (the little cat). Mann-Whitney shows that *het* was omitted significantly more often for diminutive nouns than neuter nouns (see Figure 1, $z = -2.024$, $p < 0.05$). We come back to why this may be important in language processing terms in the discussion.

Within the group of diminutive nouns there was no difference between the omission rates of nouns with a common base and those with a neuter base (Figure 2). DP represents nouns that are correctly produced with an article (determiner + noun), in NPs no article was produced.

Figure 1. Production of Determiner Phrases (DP) and Noun Phrases (NP) for neuter nouns and diminutives. *Het* is significantly more often omitted (NP) for diminutives than neuter nouns.

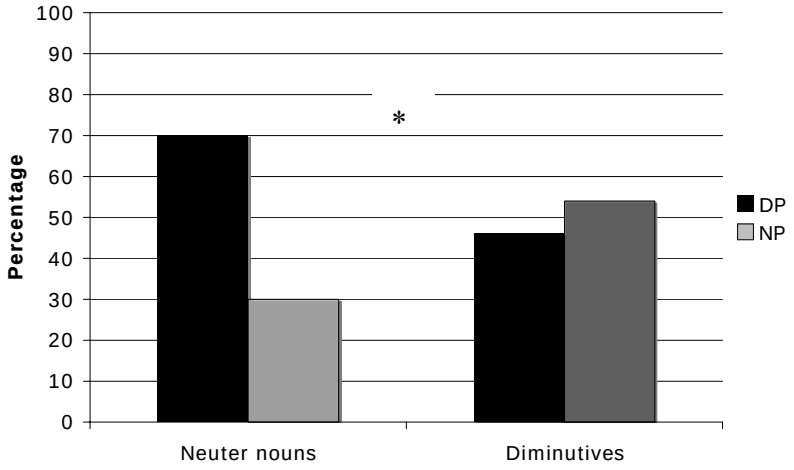
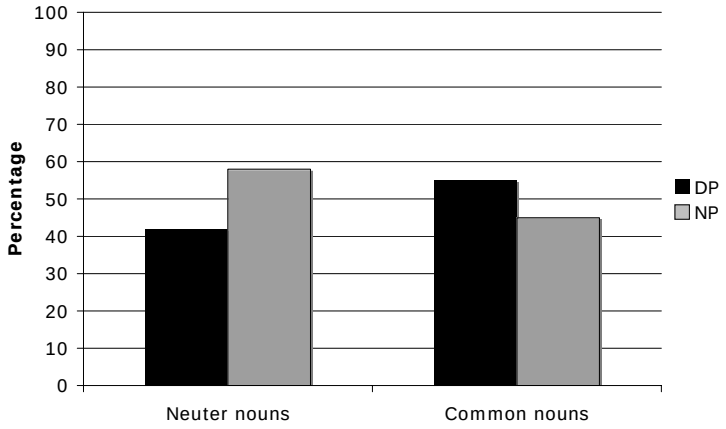


Figure 2. Production of DP and NP for diminutives with a neuter base noun and a common base noun.



2.2.2. Sentence position

The effect of sentence position was also investigated as this has been suggested to play a role in typically developing speech ([38] for German, [39] for Dutch). There was no difference in omission rates for subject or object position.

2.2.3. Information theoretic analysis

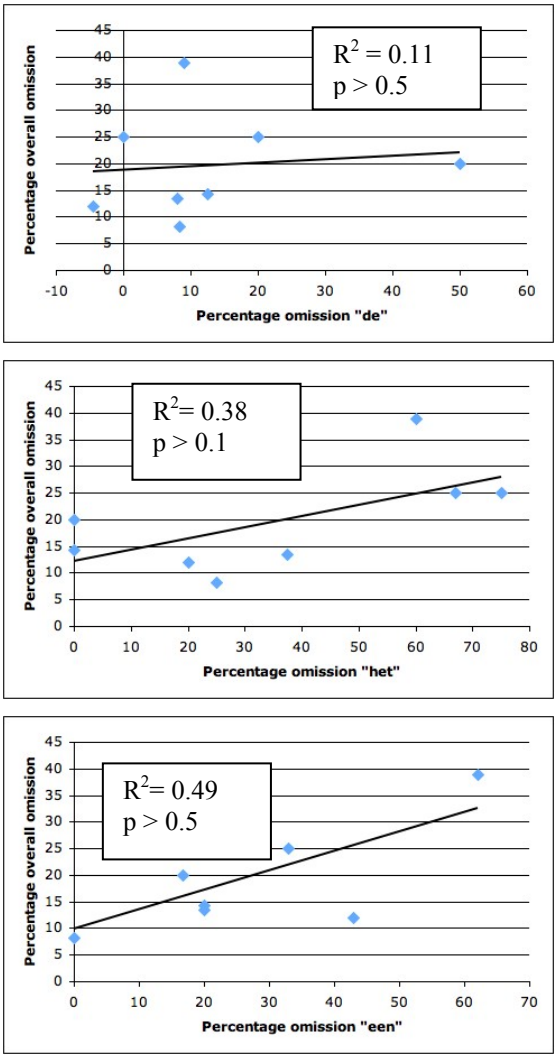
We used only those files where the omission rate was below 50% for information theoretic analyses as suggested by De Lange [29]. This meant that four files were excluded. First of all, the effect of *information load (I)* [24] of the individual articles was investigated. This measure takes into account the number of functions one article can have, as well as its frequency. Frequency measures were taken from the Corpus Gesproken Nederlands (CGN), a collection of approximately nine million spoken Dutch words. The information load I was calculated as described in Equation (3).

Table 4. Frequency, number of functions and information load of individual articles.

	Frequency	Nr functions	I
De	253,210	5	1,33
Het	96,327	3	1.98
Een	179,119	4	1.51

It was then investigated if omission of a certain article can predict the overall omission rate, in other words, if omission rates are due to poor performance on one particular article. We already showed in Section 2.1 that information load of the individual items does not lead to a significant difference in number of omissions.

Figure 3. Correlations between overall omission rates and omissions of (a) *de*, (b) *het* and (c) *een* with R^2 and significance level (p). Each dot represents one child.



However, it could be the case that omission rates of one particular article are responsible for changes of overall omission rate. In other words, it could be that children with reduced processing capacity have more difficulties producing the article with the highest individual information load. Even though this does not show when comparing average omission rates of the individual articles, it

could be that omission of one of the articles explains differences in overall omission rates. Figure 3 however shows that this is not the case.

To investigate whether we can describe the data by representing the process of "noun-article" communication as Shannon's channel, we looked at the probability distribution of the articles. This distribution is calculated on the basis of the nouns produced by the child (with or without an article).

We do acknowledge that the validity of this approach is based on the assumption that the child has acquired the necessary feature specification if the nouns he/she produces. In other words, the assumption is that if a child produces what is in adult speech a singular masculine noun, then it is indeed a singular masculine noun in the child's system as well and that it is singular masculine feature that are to be transmitted to the article set. We believe this is a reasonable assumption given the age of the subjects, a relatively low percentage of substitution, and previous claims in the literature (see above).

We used a Kullback-Leibler divergence to measure the distance between the probability distribution of the article set the child should have produced (q) and that of the article set the child actually produced (p). Thus as measure of the "input message" for the channel we used the distribution of the required articles in a file. For each noun the required article was marked. This provides a distribution of the articles required in a sample and represents the information sent from the noun set to the article set for one child.

To investigate whether this "message" from the noun is accurately transmitted to the article selection system, we also calculated the probability distribution of the output. To this end, we used the probability distribution of the articles that the child actually produced. If the child produced all the required articles, the article distribution would be identical to the input.

We then applied a Kullback-Leibler divergence (2) for each pair of probability distributions (*i.e.*, for each child). This provided us with an index of how similar the article distribution the child uses is to the article distribution the child should have used in that particular conversation, if the feature information from the noun were correctly "transmitted" to the articles set. Table 5 displays probability distributions, KL divergence and overall percentage of omission for each child.

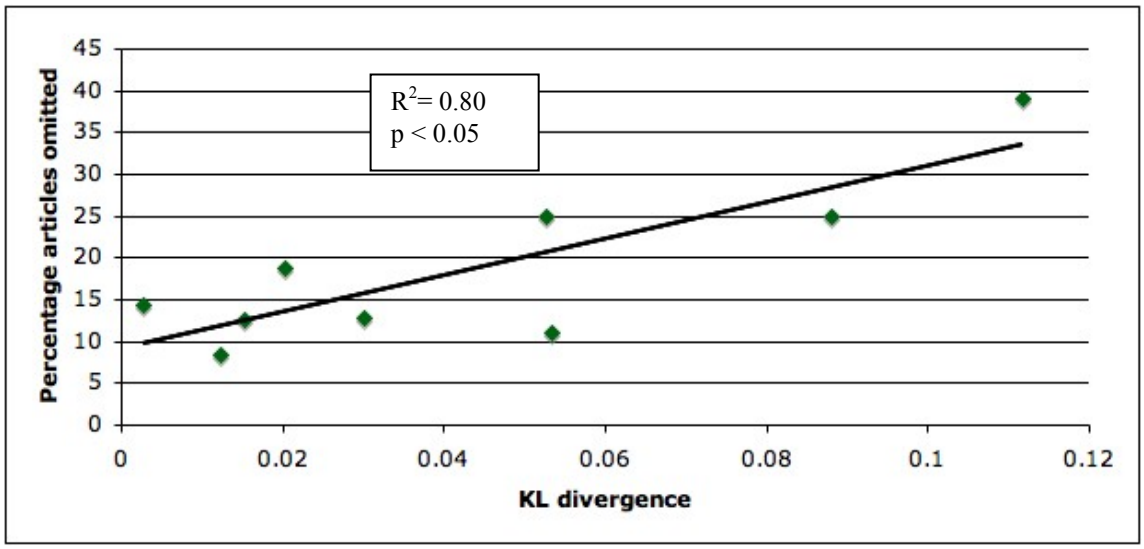
$$m_{KL}(P\|Q) = \sum p(x)\log\frac{p(x)}{q(x)} \tag{2}$$

Table 5. Article distribution of the input (q), output (p), KL divergence and percentage of omission for each child.

Child	Input (q)			Output (p)			KL divergence	% Omission
1	0.62	0.15	0.23	0.8	0.08	0.12	0.112	39
2	0.13	0.12	0.75	0.08	0.15	0.8	0.02	20
3	0.52	0.29	0.19	0.62	0.24	0.14	0.03	13
4	0.42	0.11	0.47	0.44	0.04	0.52	0.053	25
5	0.43	0.18	0.39	0.58	0.08	0.33	0.526	25
6	0.08	0.33	0.58	0.09	0.27	0.64	0.897	8
7	0.57	0.07	0.36	0.58	0.08	0.33	0.003	14
8	0.66	0.21	0.13	0.72	0.16	0.12	0.015	13

Our hypothesis is that the underlying reason for article omission is the underdeveloped capacity of the channel responsible for transmitting the information from the selected noun to article set. Input of the channel is represented by (q) and output by (p). The closer the probability distribution of the input is to the output, the better the channel does its job by correctly transmitting information from the noun to the article set. In other words, it provides an index of the channel's capacity to transmit information. If the hypothesis is correct, there should be a correlation between percentage of omission and the KL divergence, as both measures are determined by the same factor: channel capacity. Figure 4 shows his correlation.

Figure 4. Correlation between KL divergence and overall percentage of article omission.



$R^2 = 0.80$ ($p < 0.05$) for the correlation between KL divergence and percentage of omission. There is thus a strong correlation between KL divergence and the percentage of omission of articles for each child [40].

2.3. Discussion

This study investigated the omission and substitution patterns of articles in spontaneous speech samples for a group of 12 Dutch children with SLI. Our data show that by far the main type of error these children made was error of omission. Mean percentage of omission in our samples was 36 percent, whilst mean percentage of substitutions was only 2.4 percent. This finding is in line with most of the studies discussed in the introduction. Although Restrepo and Gutiérrez-Clellen [18] found substitution to be the most common error to be made by their Spanish cohort of SLI children, the vast majority of studies report omission as the most common error, both in spontaneous speech and experimental tasks. The substitution errors the children did make in our Dutch samples were all *de* for *het* substitutions. This is a common finding in typical language acquisition in Dutch. Van der Velde [41] for example found that in a production task comparing Dutch and French preschoolers, the Dutch children often substitute the common definite article *de* for the neuter definite article *het*. In contrast, a group of French preschoolers in that study did not show this pattern. Zonneveld [42] argues

that Dutch children use *de* as an unmarked, default article for nouns of which they know the meaning but are unsure of the gender.

This line of reasoning, however, seems unlikely for this group of SLI children. If *de* were used as a default article we would expect much higher substitution rates. Furthermore, comparing omission rates of the three articles we find that none of them is significantly more often omitted than the others.

Before we can conclude that the omission patterns are not due to lack of grammatical knowledge, one further phenomenon in Dutch language acquisition should be mentioned, specifically the use of diminutives. In Dutch all diminutive nouns take *het* as their article, regardless of the gender of the base noun from which they are derived. Thus, the neuter base noun *het boek* (the book) becomes *het boekje* (the small book), but the common base noun *de leeuw* (the lion) also becomes *het leeuwtje* (the little lion). Scharlaekens [43] suggests that forming a diminutive is one of the earliest morphological skills in Dutch child language development. Furthermore, diminutive nouns are frequent in child directed speech. Zonneveld [42] further observed high accuracy rates in the production of *het* in young typically developing children when used with diminutive nouns. It could therefore be the case that the accuracy rates for *het* are skewed by a high usage of diminutives in our speech samples. In order to produce the correct article for a diminutive, the child does not need to use their knowledge of the gender of the noun. It could be that these children use a high number of diminutives and that they accurately produce the article for these. Neuter nouns could still have high omission rates, but this could be masked by high accuracy rates on diminutives. If this were the case, children's article production difficulties may still be the result of poor gender knowledge. In order to investigate this possibility, the data for nouns taking *het* were reanalysed and split into two groups: neuter nouns and diminutives. If gender provides a difficulty for these children we would expect higher accuracy rates on the diminutives than the neuter nouns. Error analysis of omission of *het* for our data reveal quite the opposite pattern (see figure 2 above). Children have more difficulties producing *het* when using a diminutive noun than when they use a normal neuter noun. Further inspection shows that they have equal difficulties with those nouns that are derived from common base nouns as those with neuter nouns. The children are not helped by the article of the base noun. They make equal number of errors with nouns such as "*het leeuwtje*" (the little lion) derived from "*de leeuw*" (the lion) as with nouns such as "*het boekje*" (the little book) derived from "*het boek*" (the book). It thus seems that, if anything, gender knowledge *is* present and is actually inhibiting selection of the correct article. The increased uncertainty of competing articles for the base noun and the diminutive form leads to decreased performance. This provides further support for our hypothesis that the difficulty in producing articles is related to the noun-article connection.

Interestingly, similar results have been found in adult language processing. Schiller and Caramazza [44] found longer naming latencies for Dutch adults for diminutive nouns with a common gender base, compared to those with a neuter gender base. They argue that this is due to activation of the common article that goes with the base noun, upon production of the diminutive noun. This leads to competition between the articles and therefore longer latencies. For diminutive nouns that have neuter base gender this competition does not occur, as the base noun will activate the same article as the diminutive form of that noun. The gender feature of the base noun thus seems to be activated in the article selection process. The error pattern of the Dutch SLI children is in line with these findings for adults. Dutch SLI children do have knowledge of the gender of the noun but, paradoxically, this

knowledge can even lead to decreased performance. Competition between the articles leads to increased processing load and a failure in selecting any of the articles.

We also looked at the effect of sentence position on omission patterns. Contrary to some findings in the literature on typically developing children [38] for German and [39] for Dutch, we did not find a sentence position effect.

As lack of grammatical knowledge does not appear to be the cause of omission of article for these children, we then analysed the data using an information theoretic approach. As discussed in the introduction, recent studies have shown that processing time can be accurately described using information-theoretic measures. We first investigated the effect of the individual information load for each article using formula (1). This factor did not seem to determine patterns of omission in the children's output.

We then used the difference between input and output probability distribution of the articles to provide us with an index of the processing capacity of these children. Probability distributions were calculated for the articles that each child should have produced and for what the child actually produced. The first distribution represents the input of the channel: the information that is sent from the nouns to the article set. The second probability distribution represents the output of the channel. KL divergence was used to investigate the difference between these two probability distributions. We hypothesized that when the channel capacity is sufficient for these children, the output will have the same probability distribution as the input. The KL divergence was used as an index of the amount of distortion that occurs in the channel. In other words, the greater the divergence, the more different the output is from the input and the further the child's article distribution is removed from what it should be [45].

KL divergence was then correlated with overall percentage of omission. The correlation between these two variables was over 0.8. This means that the more different the article distribution of the child's production was from what it should have been, the more articles were *also* omitted. Again, it is crucial to realise that the KL divergence measure does not take omissions into account. Rather it provides a measure of how similar the output of a channel is to the input. If the channel capacity is sufficient, output and input will be identical. The difference between output and input provides a measure of the level of distortion that has occurred within the channel. As we observed that the rate of omission correlated with the KL divergence value, *and* as this value is an index of channel capacity to correctly transmit information, it follows that the omissions in SLI speech can be characterised as a consequence of low channel capacity.

This paper has provided more evidence for a processing account of SLI. Unlike previous accounts, we have tried to develop a quantitative measure for the reduced processing in children with SLI. Here we have focused on the production of articles and shown that a measure of channel capacity as described by information theoretical means can provide a model for the reduced processing capacity these children experience. Similar findings have been reported for a group of typically developing Dutch and Italian children [29]. De Lange looked at Dutch preschoolers with a very similar language level to the children with SLI reported on in this paper. Her results show that channel capacity increases with age and the number of omitted articles reduces. This is evidence for the view that children with SLI follow a similar developmental trajectory as typically developing children.

Although this paper has only looked at article production, we can also tentatively speculate that reduced processing may be involved in other aspects of language difficulties these children experience. Inflectional morphology has been shown to be difficult for these children [5,6], with some inflections being more difficult than others. Recent work on inflectional morphology in Dutch [28] has shown that different inflectional forms of verbs have different complexity in terms of information theoretical measures. These have been shown to influence speed of processing in healthy adults. Future research will provide more insight into whether these factors play a role in the difficulties of children with SLI in producing inflectional morphemes.

Speed of processing has also been shown to be reduced in children with SLI [46,47]. Children with SLI are slower on a wide range of (language) tasks, and reducing the input rate of language stimuli can improve performance on comprehension tasks for these children [47]. As channel capacity is a measure of the amount of information the channel can process in *one unit of time*, it follows that in order to perform accurately one can either reduce the amount of information, or increase the amount of time to enhance performance. This is exactly what happens when the speed of input is reduced.

For typically developing children it has been shown that speed of processing increases with age [48]. Within our current model, this means that channel capacity increases with age. Montgomery [47] showed that a similar development occurs for children with SLI, *i.e.*, they show a linear improvement in speed of processing with an increase in age. However, the children's speed of processing remains below that of their age matched- and (in the Montgomery study) language matched peers. In other words, their channel capacity is reduced in comparison to TD children.

If processing capacity is reduced, the obvious next question is whether this reduction is apparent only for language tasks, or whether it is a more general reduction in the processing of information. Despite the requirement of normal non-verbal intelligence for a child to be diagnosed as having SLI, there is a growing amount of research suggesting that the difficulties of children with SLI may not be completely 'language specific'. For example, children with SLI have shown difficulties with spatial processing [49], hierarchical planning tasks [49,50] and hypothesis testing [51,52]. Johnston [53] provides an excellent overview of the literature on the cognitive abilities that have been investigated in SLI and shows that many non-verbal skills lag behind those of typically developing children. Some motor skills have also been shown to be problematic for this group of children (see [54] for an overview). In short, it seems that the reduced processing capacity these children experience may not be language specific. Instead a more general difficulty processing complex information could lead to a more diffuse profile of difficulties in various cognitive domains. Although we can not draw any conclusions on nonverbal skills from our findings, it is interesting to note that information theoretic measure have been applied to non-verbal cognitive processes in healthy adults and that entropy has been shown to be a measure of processing capacity. Rozenholtz [55] for example found that an increase in picture complexity as measured by entropy lead to an increase in reaction time in a visual search task for healthy adults. As entropy provided an accurate measure of processing capacity in our study on article production, it would be very interesting to find out whether the model could also be applied to non-verbal cognitive processing in this group of children.

3. Conclusions

In summary, article omissions in Dutch SLI children seem to be due to a reduced processing capacity, which makes it difficult for children to select the correct article. This can be captured by a model based on information theory that uses probability distributions to calculate channel capacity. We have shown that the level of distortion in the probability distribution from the nouns to the article set provides a measure of the reduction of channel capacity. The more deviant the output probability distribution from the input distribution, the higher the omission rates of articles in a spontaneous speech sample. Furthermore, similarity between older SLI children and younger typically developing (TD) children with regard to omission and distribution of articles suggests that SLI can be characterized, at least in part, as a disorder related to late maturation of the processing capacity which is necessary to realize the available knowledge.

Although the current manuscript only focuses on the production of articles, statistical measures based on information theory could provide promising models of reduced processing capacity in other aspects of language acquisition and more general cognitive problems in SLI.

Acknowledgements

We would like to thank Joke de Lange, Victor Kuperman, Bettina Gruber and Elise de Bree for their helpful suggestions and inspiring discussions.

References and Notes

1. Stark, R.; Tallal, P. Specific language impairment in children. *Adv. Dev. Behav. Pediatr.* **1982**, *2*, 257–271.
2. World Health Organization. *The ICD-10 Classification for Mental and Behavioural Disorders: Diagnostic Criteria for Research*; WHO: Geneva, Switzerland, 1993.
3. Tomblin, J.B.; Records, N.; Buckwalter, P.; Zhang, X.; Smith, E.; O'Brien, M. Prevalence of specific language impairment in kindergarten children. *J. Speech Lang. Hear. Res.* **1997**, *40*, 1245–1260.
4. Archibald, L.M.D.; Gathercole, S.E. Prevalence of SLI in language resource units. *J.Res. Spec. Educ. Needs* **2006**, *6*, 3–10.
5. Bishop, D.V.M. The underlying nature of specific language impairment. *J. Child Psycho. Psychiatry* **1992**, *33*, 1–64.
6. Leonard, L.B.; Bortolini, U. Grammatical morphology and the role of weak syllables in the speech of Italian-speaking children with specific language impairment. *J. Speech Lang. Hear. Res.* **1998**, *41*, 1363–1374.
7. Rice, M.L.; Wexler, K. Toward tense as a clinical marker of specific language impairment in English-speaking children. *J. Speech Lang. Hear. Res.* **1996**, *39*, 1239–1257.
8. Gopnik, M.; Crago, M. Familial aggregation of a developmental language disorder. *Cognition* **1991**, *39*, 1–50.

9. Tallal, P.L.; Miller, S.L.; Bedi, G.; Byma, G.; Wang, X.; Najarajan, S.S.; Schreiner, C.; Jenkins, W.M.; Merzenich, M.M. Language comprehension in language-learning impaired children improved with acoustically modified speech. *Science* **1996**, *271*, 81–84.
10. Kail, R. A method for studying the generalized slowing hypothesis in children with specific language impairment. *J. Speech Hear. Res.* **1994**, *37*, 418–421.
11. Leonard, L.B.; McGregor, K.K.; Allen, G.D. Grammatical morphology and speech perception in children with specific language impairment. *J. Speech Hear. Res.* **1992**, *35*, 1076–1085.
12. Leonard, L.B.; Eyer, J.A.; Bedore, L.M.; Grela, B.G. Three accounts of the grammatical morpheme difficulties of English-speaking children with specific language impairment. *J. Speech Lang. Hear. Res.* **1997**, *40*, 741–753.
13. Leonard, L.B. Functional categories in the grammars of children with specific language impairment. *J. Speech Hear. Res.* **1995**, *38*, 1270–1283.
14. Le Normand, M.T.; Leonard, L.B.; McGregor, K.K. A cross-linguistic study of article use by children with specific language impairment. *Eur. J. Disord. Commun.* **1993**, *28*, 153–163.
15. Leonard, L.B.; Bortolini, U.; Caselli, M.C.; Sabbadini, L. The use of articles by Italian-speaking children with specific language impairment. *Clin. Linguist. Phon.* **1993**, *7*, 19–27.
16. Bedore, L.M.; Leonard, L.B. Grammatical morphology deficits in Spanish-speaking children with SLI. *J. Speech Lang. Hear. Res.* **2001**, *44*, 905–924.
17. Bosch, L.; Serra, M. Grammatical morphology deficits of Spanish-speaking children with specific language impairment. *Amsterdam Ser. Child Lang. Dev.* **1997**, *6*, 33–46.
18. Restrepo, M.A.; Gutierrez-McClellan, V.F. Article use in Spanish-speaking children with SLI. *J. Child Lang.* **2001**, *28*, 433–452.
19. Anderson, R.T.; Souto, S.M. The use of articles by monolingual Puerto Rican Spanish- speaking children with specific language impairment. *Appl. Psycholinguist* **2005**, *26*, 621–647.
20. Hansson, K.; Nettelbladt, U.; Leonard, L.B. Indefinite articles and definite forms in Swedish children with specific language impairment. *First Lang.* **2003**, *23*, 334–362.
21. Alario, F.X.; Caramazza, A. The Production of Determiners: Evidence from French. *Cognition* **2002**, *82*, 179–223.
22. Janssen, N.; Caramazza, A. The selection of closed-class words in noun phrase production: The case of dutch determiners. *J. Mem. Lang.* **2003**, 635–652.
23. Shannon, C.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Champaign, IL, USA, 1949.
24. Kostić, A.; Katz, L. Processing differences between nouns, adjectives and verbs. *Psychol. Res.* **1987**, *49*, 229–336.
25. Kostić, A. The effects of the amount of information on processing of inflected morphology. 2009. mansucript in preparation.
26. Moscoso del Prado Martín, F.M.; Kostić, A.; Baayen, R.H. Putting the bits together: an information theoretical perspective on morphological processing. *Cognition* **2003**, *94*, 1–18.
27. Baayen, R.H.; Moscoso del Prado Martín, F.M. Semantic density and past-tense formation in three Germanic languages. *Language* **2005**, *81*, 666–698.
28. Tabak, W.; Schreuder, R.; Baayen, R.H. Lexical statistics and lexical processing: semantic density, information complexity, sex, and irregularity in Dutch. In *Linguistic Evidence:*

Emperical, Theoretical and Computational Perspectives; Kepser, S., Reis, M., Eds.; Mouton de Gruyer: Berlin, Germany, 2005; pp. 529–555.
29. de Lange, J. *Article Omission in Headlines and Child Language: A Processing Approach*; LOT: Utrecht, The Netherlands, 2008.
30. Ferrer i Cancho, R. Quantifying the semantic contribution of particles. *J. Quant. Linguist.* **2002**, *9*, 35–47.
31. Ferrer i Cancho, R. Euclidean distance between syntactically linked words. *Phys. Rev. E* **2004**, *70*, 056135.
32. Ferrer i Cancho, R. Decoding least effort and scaling in signal frequency distributions. *Physica A* **2005**, *345*, 275–284.
33. Ferrer i Cancho, R. When language breaks into pieces. A conflict between communication through isolated signals and language. *BioSystems* **2005**, *84*, 242–253.
34. Ferrer i Cancho, R. Some word order biases from limited brain resources: a mathematical approach. *Adv. Complex Syst.* **2008**, *11*, 393–414.
35. Peperkamp, S.; Le Calvez, R.; Nadal, J.P.; Dupoux, E. The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition* **2006**, *101*, B31–B41.
36. Bol, G.W.; Kuiken, F. Grammatical analysis of developmental language disorders: A study of the morphosyntax of children with specific language disorders, with hearing impairment and with Down's syndrome. *Clin. Linguist. Phon.* **1990**, *4*, 9.
37. Sustitutions were all *de* for *het*.
38. Schoenenberg, M.; Penner, Z.; Weissenborn, J. Object placement in early german grammar. In Proceedings of 21st Annual Boston University Conference on Language Development, Boston, MA, USA, November 1996; Hughes, A., Hughes, M., Greenhill, A., Eds.; Cascadilla Press: Somerville, MA, USA, 1997.
39. Avrutin, S. Optionality in child and aphasic speech. *Lingue Linguaggio* **2004**, *1*, *65–96.*
40. We also investigated the correlation between entropy as defined by Eq. (1) of the input and percentage of omissions per child and output entropy with percentage of omission. Neither was significant.
41. van der Velde, M. L'acquisition des articles définis en l1. *Acquisition et Interaction en Langue Étrangère (AILE)* **2004**, *21*, 9–46.
42. Zonneveld, W. Het jonge hoofd, De Righthand Head Rule bij kinderen van 4 tot 7 jaar. *Nieu. Taalg.* **1992**, *85*, 37–49.
43. Schaerlakens, A.M. *De Taalontwikkeling van het Kind*; Wolters Noordhoff: Groningen, The Netherlands, 1980.
44. Schiller, N.O.; Caramazza, A. Grammatical feature selection in noun phrase production: evidence from German and Dutch. *J. Memory Lang.* **2003**, *48*, 169–194.
45. It is crucial to note that neither the measure for input entropy nor the measure for output entropy use information on omissions. Input entropy provides a measure of what the article distribution should be like, *i.e.*, the information the nouns sent to the article selection system in that sample, and the output entropy provides a measure of the article distribution the child actually produced. It is important to note that these, in theory, could be identical regardless of the number of omissions the child made. *i.e.*, if the required input consisted of 100 nouns with the probability distribution

of the articles of 0.2, 0.4, and 0.2 and the output consisted of 50 articles with the probability distribution of 0.2, 0.4 and 0.2, input and output entropy would have been identical even though 50% of articles was omitted.

46. Fazio, B.B. The effect of presentation rate on serial memory in young children with specific language impairment. *J. Speech Lang. Hear. Res.* **1998**, *41*, 1375–1383.
47. Montgomery, J.W. Effects of input rate and age on the real-time language processing of children with specific language impairment. *Int. J. Lang. Commun. Disord.* **2005**, *40*, 171–188.
48. Marslen-Wilson, W.D.; Tyler, L.K. Central processes in speech understanding. *Phil.Trans. R. Soc. Lond.* **1981**, *B295*, pp. 317–322.
49. Kamhi, A.G.; Catts, H.W.; Mauer, D.; Apel, K.; Gentry, B.F. Phonological and spatial processing abilities in language- and reading-impaired children. *J. Speech Lang. Hear. Res.* **1988**, *53*, 316–327.
50. Cromer, R. Hierarchical planning disability in drawnings and constructions in a special group of severely aphasic children. *Brain Cogn.* **1983**, *2*, 144–164.
51. Nelson, L.K; Kamhi, A.G.; Apel, K. Cognitive strengths and weaknesses in language impaired children: one more look. *J. Speech Hear. Disord.* **1987**, *52*, 36–43.
52. Ellis Weismer, S. Hypothesis testing abilities of language-impaired children. *J. Speech Lang. Hear. Res.* **1991**, *34*, 1329–1338.
53. Johnston, J. Cognitive deficits in specific language impairments: decisions in spite of uncertainty. *J. Speech-Lang. Pathol. Audiol.* **1999**, *23*, 165–172.
54. Hill, E.L. Non-specific nature of specific language impairment: A review of the literature with regard to concomitant motor impairments. *Int. J. Lang. Commun. Disord.* **2001**, *36*, 149–171.
55. Rosenholtz, R.; Li, Y.; Nakano, L. Measuring visual clutter. *J. Vis.* **2007**, *7*, 1–22.