Learning to categorize verbs and nouns: studies on Dutch
Erkelens, M.A.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
3 Production studies of categorization by Dutch children

This chapter reports two studies that test the predictions from Chapter 2 on child language production data. The learnability of the verbal and nominal representations proposed are pivotal for a viable theory. The theories make two different general predictions for child language production data: either children make a considerable number of categorization errors, or they show adult-like production patterns from the start. In this chapter the predictions are tested by studying Dutch children’s production data. The methods used to analyze these production data will be evaluated with respect to their efficacy in testing the predictions from Chapter 2, and conclusions will be drawn on the basis of the results with respect to the compatibility of the theories with these data.

3.1. Introduction

In the studies presented in this chapter the focus will be on child language production data. The predictions from Chapter 2 will be investigated to decide which theory is most compatible with such data. From the theories in Marantz (1997) and Borer (2003), it is expected that children will make categorization errors at an early stage of language production because their vocabulary development lags behind their syntactic abilities (see §2.2). To test this prediction a small study of categorization errors in Dutch children’s spontaneous speech was conducted and is reported in §3.2.

The other theories described in Chapter 2 all predict that children will show an adult-like use of words in syntactic patterns. Baker (2003) assumes that the categories verb and noun are innate and according to my interpretation of this assumption (§2.3.2), children therefore immediately categorize words as verbs or nouns. Hengeveld (1992b) predicts that children are adult-like, or even more conservative than adults, in their use of verbs and nouns, because they follow the implicational hierarchy. This prediction implies that children start from the assumption that there is only a single class of content words for predication and only gradually expand this assumption towards a word class for reference. Croft (1991, 2000) and Goldberg (1995, 2006) assume that children build their categories gradually based on the input language, and as a consequence restrict themselves to the adult pattern at an early stage of language production. To test the prediction that
children will show adult-like categorization patterns at an early stage of production, a larger study was conducted that took all content words into account, instead of only the possible errors. This longitudinal corpus study conducted on Dutch children’s spontaneous speech is presented in §3.3.

The outcomes of the studies suggest that Dutch children hardly make any errors, instead showing an adult-like pattern from early on. Based on these outcomes, §3.4 discusses the success of the research strategy using early production. Another research strategy is presented that potentially provides more insight into the process of learning to categorize verbs and nouns in Dutch.

3.2. **Categorization errors in Dutch children’s spontaneous speech**

The analysis of the errors children make in their spontaneous speech production is a common method of investigating grammatical development (e.g., Bowerman, 1974; 1976). For example, in morphological development the overgeneralization of the regular past tense form, as in *I eated an apple*, has often been studied (Marcus et al., 1992; Maratsos, 2000; Maslen, Theakston, Lieven, & Tomasello, 2004). These overgeneralization errors are indicative of the grammatical representation of past tense in the child’s grammar; the child’s grammar contains a rule for past tense inflection. The points where the child diverges from the adult patterns and produces non-adult-like language can be observed. In the process of learning to categorize, children are expected to make errors in categorization that reveal the underlying structure of their early categories with respect to those of adults, at least to some extent. Marantz and Borer explicitly predict categorization errors. The study reported here tests this prediction against Dutch children’s spontaneous speech production. However, there are many methodological problems in identifying and interpreting categorization errors. These difficulties will first be discussed in general and will then be taken into account in the analysis of this study.

3.2.1. **Interpreting categorization errors**

A categorization error is attested once a child uses a word form with the morphosyntactic properties of a category in which it cannot be used in adult language, just as a morphological error is attested once a child uses the irregular verb eat with a regular past tense inflection. However, recognizing a categorization error is more

---

9 Although not all scholars agree on the fact that the mental grammar contains rules (see discussion in e.g., Rumelhart & McClelland, 1994; Pinker, 2001), there is at least some kind of abstraction for past tense in the child’s grammar.
difficult than recognizing a morphological error. The regular past tense inflection in English always has the surface form –ed. Furthermore, expressing the regular past tense is one of the few functions of the morpheme –ed in English. This means that once an –ed ending is attested on a form that cannot have this ending, it can be identified as a morphological error of past tense inflection. For categorization errors, there is no such one-to-one relationship between a morpheme and a category. Although verbs often take tense inflection and nouns often take number inflection, this is not necessarily the case (e.g., irregular verbs like ‘know’, ‘leave’, or ‘buy’ do not take –ed in the past tense; mass nouns like ‘sand’, ‘water’, or ‘mud’ are usually not inflected for number). Furthermore, all properties that are indicative of category membership have some other function as well; although tense inflection is indicative of verbhood, the main function of a tense morpheme is to express tense, and not to express category membership. The problem with recognizing categorization errors is that category is indicated by morphemes that have other grammatical functions.

If morphemes cause a problem of interpretation, this problem could in principle easily be solved by only looking at forms without grammatical markers. However, if the content words are not grammatically marked, we cannot decide if word forms have been categorized correctly. This fact is clearly stated by Eve Clark:

“The point is a methodological one: children in the earlier stages of acquisition may not make the same assignments to word classes that adults do, so we cannot automatically assume adult-like assignments until children’s terms appear with enough grammatical information to be certain.” (Clark, 1993: 39)

It is very hard – if not impossible – to recognize categorization errors if there is no grammatical information in the utterances. Since grammatical information is needed in order to decide whether an error has occurred, and since children generally start to produce grammatical structure at around 2 years of age, we can only look for categorization errors in children older than 2;0 (years;months). Aside from the fact that this is quite late for detecting the development we are interested in, there are other developments at this age that cause some challenges for the interpretation of these errors.

Children typically start to use nonexistent innovative words (‘errors’ in the sense of non-adult-like) after 2;0 when they develop more advanced lexical and morphological skills. Such errors should not be labeled categorization errors, since adults can invent new words in a comparable way (Clark & Clark, 1979). In §2.2.2, I mentioned the English example *Mummy trousers me* (Clark, 1982: 406) as an example of evidence for Borer’s (2004) prediction of categorization errors.
However, the form *trouser* with the verbal third person singular morpheme *–s* could be either an instance of a categorization error (the child’s grammar allows *trouser* to be spelled out with a verbal morpheme) or, equally plausible, an innovation created to fill the lexical gap for ‘putting on trousers’ in the adult lexicon by means of a morphological zero-derivation of the nominal *trouser*. Due to the fact that the child lexicon is still developing, such innovations might also be used by children for concepts for which there is a word in the adult lexicon that the child has not learned yet. In these cases, the innovation resulting from a morphological operation is created to fill a lexical gap in the child’s own lexicon that is not actually a lexical gap in the adult language. Distinguishing lexical errors resulting from a morphological operation from categorization errors is difficult because their surface form is exactly the same. This is not to say that such errors are not informative about the acquisition process, because they indicate aspects of lexical and morphological development. However, they are not necessarily indicative of the development of categorization. In adult English a substantial number of words can be used as both nouns and verbs (see also §1.1). If children hear the same word form being used with properties of different categories, they may assume that this is a possibility for all word forms. For example, the earliest reported candidate for a category overgeneralization in English in Clark (1993) is of 2;4-year-old Damon saying *I’m sanding* while grinding pepper, and afterwards referring to the result (the pepper grains) saying *look at the sand!* (Clark, 1993: 200, Table 11-1). The form *sand* can be both a verb and a noun in English, so using *sand* with the verbal ending *–ing* is not a categorization error per se. However, the meaning of the verb *sand* in adult language is different from the Damon’s meaning (namely ‘to smooth wood with sandpaper’). The question is whether Damon is making a categorization error here, or a semantic error concerning the meaning of the adult verb *sand*, or a morphological error concerning the derivational possibilities of *sand*.

In summary, there is an important problem to be dealt with in the interpretation of categorization errors. The morphemes that are indicative of category membership have other functions, such as the expression of tense or number. As a consequence, the surface form of categorization errors is identical to the surface form of morphological errors. If a word form can occur in multiple categories in adult language, the apparent categorization error can also be an overgeneralization of morphological possibilities, or even of meaning. In the next section I will investigate categorization errors in child speech bearing in mind that they may be morphological or semantic errors.
3.2.2. Analysis of categorization errors in Dutch

A corpus study of categorization errors in Dutch was conducted to investigate the compatibility of child language production data with the theories by Marantz and Borer. In the discussion of the results, special attention will be paid to whether the problems of interpreting these errors (see §3.2.1) can be overcome. The recordings of four Dutch children between about 2;6 and 3;6 years of age (Matthijs and Tomas, Wijnen, 1993; Laura and Sarah, Van Kampen, 1997), as archived in the Child Language Data Exchange System (MacWhinney, 1991), were selected for analysis. All child utterances in these recordings were checked by hand to see if there were any possible categorization errors. The working definition of a categorization error used for this analysis was the use of a Dutch word in a morpho-syntactic context in which it cannot be used in adult Dutch.

All observed candidates for category overgeneralizations in these files are listed in Table 3.1. The main finding of this study is that in all the data studied there are only 14 candidates for categorization errors. This is a very small number. To provide an idea of the ratio, 78 recordings were checked, approximately 200,000 tokens of child speech, so the 14 candidates for categorization error constituted about 0.007% of all tokens in the analyzed corpora. This implies that the overwhelming majority of child utterances appear adult-like with respect to category. This impression will be systematically investigated in the longitudinal corpus study reported in the next section. In this stage of child language production, between 2;6 and 3;6, categorization errors are definitely not typical. The predictions based on proposals of Marantz and Borer is not borne out by these data. If children make categorization errors at an early stage of development, it is before 2;6.

Although the number of errors found is very small, it is worthwhile taking a closer look at them to see whether they can be interpreted as actual categorization errors. In order to determine whether these errors are indeed categorization errors and not semantic or morphological errors, all the candidates for overgeneralization error from Table 3.1 were analyzed with respect to the following three questions: (1) does the non-adult-like use of the word form fill a lexical gap in the adult lexicon?, (2) does the non-adult-like use of the word form fill a lexical gap in the child’s lexicon?, and (3) does the word form occur in the category of the non-adult-like utterance in adult speech (with a different meaning)? If the answer to all these three questions is negative, the attested non-adult-like utterance can be interpreted as a categorization error. However, if the answer to one or more of the questions is positive, it is not straightforward how the error can be used to distinguish the child’s categorization abilities from the child’s morphological and semantic abilities.
The questions of whether the non-adult-like utterance filled a lexical gap in the adult’s lexicon and whether the word occurred in the category of the non-adult-like utterance in adult speech were answered based on native speaker intuitions of the researcher. A lexical gap in the child’s lexicon was established based on all available spontaneous speech transcripts of the child, including those where the child was younger than 2;6 and older than 3;6. If the presumably intended adult word did not occur in one of the child’s transcripts when it was required for the context, it was assumed that the child did not have this word in her active lexicon.10

Table 3.2 shows that most candidates for categorization errors could also be morphological or semantic errors. Only one of the examples (pannen) cannot be analyzed as a lexical gap or multiple-category item. This might be due to the fact that it is not clear from the context what the child intends to say here. When the intended meaning is not clear, it is impossible to decide upon the appropriate label for it in adult language. Without the equivalent adult word, it could not be established whether this utterance fills a lexical gap. In support of the idea that these examples could equally well be overgeneralizations of meaning as categorization errors is the fact that in 13 of the 14 cases, the child also uses the lexical item in the adult category in one of the recordings (e.g., in the same transcript where Matthijs produces veeg, he also produces the correct derived form veger ‘brush’). This indicates a certain flexible use of word forms that is not necessarily evidence for different representation of categories in the child grammar compared to adults. They could also be the result of creative analogy with zero-derived adult word forms such as fiets\textsubscript{v} / fiets\textsubscript{n}, ‘bike’ and drinken\textsubscript{v} / drinken\textsubscript{n}, ‘drink’, which are also used correctly in multiple categories from an early age onwards. Although these data provide interesting insights into the lexical and morphological abilities of these children, they do not contribute to a better understanding of their categorization skills.

This study showed that Dutch children produce a very small number of categorization errors. Furthermore, the problems of interpretation presented as a challenge in §3.3.1 could not be overcome in this study. Two conclusions can be drawn at this point: the theories of Marantz and Borer are not compatible with the language production data of these four Dutch children between 2;6 and 3;6, and the analysis of errors as a method to study categorization faces challenges that seem impossible to overcome. The size and method of this study does not allow us to draw firm conclusions about the adult-like use of words by Dutch children in general. Therefore, a more elaborate study of the categorization of verbs and nouns by Dutch children was conducted that also included data from slightly younger

10 Although the transcripts of course only contain a very small percentage of the children’s total speech production, this is the only way in which the children’s vocabulary could be assessed empirically.
Table 3.1. Overview of all candidates for categorization errors from a corpus study of four Dutch children between 2;6 and 3;6, structured according to the category of the adult word form from which the error is derived: nouns, verbs, and adjectives.

<table>
<thead>
<tr>
<th>Non-adult-like utterance*</th>
<th>English translation (and intended meaning)</th>
<th>Adult word form</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NOUNS PRODUCED WITH VERBAL INFINITIVE MARKER –E(N)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>hij gaat zo</em> voeten</td>
<td>‘he goes to foot like that’ (walk the way a crab walks)</td>
<td>voet ‘foot’</td>
<td>2;9</td>
</tr>
<tr>
<td><em>dupele</em></td>
<td>‘to duplo’ (play with duplo)</td>
<td>duplo ‘duplo’</td>
<td>2;11</td>
</tr>
<tr>
<td><em>hameren</em></td>
<td>‘to hammer’</td>
<td>hamer ‘hammer’</td>
<td>3;1</td>
</tr>
<tr>
<td><em>met de deksel pannen</em></td>
<td>‘to pan with the lid’ (?)</td>
<td>pan ‘pan’</td>
<td>3;1</td>
</tr>
<tr>
<td><em>nee, ik wil mee</em> gieteren</td>
<td>‘no, I want to watering-can with that’ (pour water with watering-can)</td>
<td>gieter ‘watering-can’</td>
<td>3;2</td>
</tr>
<tr>
<td><em>ikke wil zalve</em></td>
<td>‘I want to ointment’ (to put on ointment)</td>
<td>zalf ‘ointment’</td>
<td>2;7</td>
</tr>
<tr>
<td><strong>VERBS PRODUCED WITH NOMINAL MARKERS DE / HET / EEN</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>die moet ook niet bij de zwemmen</td>
<td>‘that one must not with the swims, either’ (those that swim)</td>
<td>zwemmen ‘to swim’</td>
<td>2;10</td>
</tr>
<tr>
<td>een veeg!</td>
<td>‘a sweep!’ (brush)</td>
<td>vegen ‘to sweep’</td>
<td>2;7</td>
</tr>
<tr>
<td>dit is het tafeldek</td>
<td>‘this is the lay-table’ (tablecloth)</td>
<td>tafeldekken ‘to lay table’</td>
<td>3;6</td>
</tr>
<tr>
<td>ik wil de hoor</td>
<td>‘I want the hear’ (headphone)</td>
<td>horen ‘to hear’</td>
<td>2;11</td>
</tr>
<tr>
<td><strong>ADJECTIVES PRODUCED WITH VERBAL MARKERS 3RD PERSON SINGULAR –T AND PAST PARTICIPLE MARKER GE-</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>neusje viest</td>
<td>‘nosie drities’ (is dirty)</td>
<td>vies ‘dirty’</td>
<td>3;7</td>
</tr>
<tr>
<td>kijk, gestuk, stuk dat</td>
<td>‘look, made-broken, broken that’ (has been broken)</td>
<td>stuk ‘broken’</td>
<td>2;6</td>
</tr>
<tr>
<td><strong>ADJECTIVES PRODUCED WITH NOMINAL MARKERS (EE)N / DE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>is n heel mooi?</em></td>
<td>‘is a very beautiful?’ (very beautiful one)</td>
<td>mooi ‘beautiful’</td>
<td>2;8</td>
</tr>
<tr>
<td><em>die gaan ook op de snel</em></td>
<td>‘those also go on the fast’ (highway)</td>
<td>snel ‘fast’</td>
<td>2;10</td>
</tr>
</tbody>
</table>

*Category-indicating morpho-syntactic markers are underlined.
children (2;0), to see whether the errors predicted by Marantz and Borer would occur at an earlier stage.

Table 3.2. Analysis of all candidates for categorization errors according to the possibility that they represent innovations (filling lexical gaps) or word forms that can occur in multiple categories in adult speech (see Table 3.1 for English translation). + stands for a positive answer to the statement at the top of the column, - stands for a negative answer.

<table>
<thead>
<tr>
<th>Categorization error from Table 3.1</th>
<th>No word exists in adult language with the intended meaning</th>
<th>Adult equivalent word exists, but was not found in child transcript</th>
<th>Word occurs in this category in adult speech, but with different meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>voeten</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>dupele</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>hameren</td>
<td>-</td>
<td>+ (timmeren)</td>
<td>+ (figuratively)</td>
</tr>
<tr>
<td>pannen</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>gieteren</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>zalve</td>
<td>+</td>
<td>-</td>
<td>+ (archaic)</td>
</tr>
<tr>
<td>zwemmen</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>veeg</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>slaap</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>tafeldek</td>
<td>-</td>
<td>+ (tafelkleed)</td>
<td>-</td>
</tr>
<tr>
<td>hoor</td>
<td>-</td>
<td>+ (koptelefoon)</td>
<td>-</td>
</tr>
<tr>
<td>viest</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>gestuk</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>mooi</td>
<td>-</td>
<td>+ (mooie)</td>
<td>-</td>
</tr>
<tr>
<td>snel</td>
<td>-</td>
<td>+ (snelweg)</td>
<td>-</td>
</tr>
</tbody>
</table>

3.3. A longitudinal corpus study of Dutch children’s spontaneous speech

It is predicted by most of the theories discussed in Chapter 2 that children will show an adult-like, or an even more conservative than adult-like, use of words in syntactic patterns. To test this prediction in more detail than was done in the previous study of categorization errors, a longitudinal corpus study was conducted. The study was
designed to test the specific predictions made by Hengeveld (§2.4) and therefore uses his terminology. However, the results are relevant for all the theories other than those of Marantz and Borer, since the overall prediction is similar (§2.6).

As discussed in §2.4.2, Hengeveld’s (1992b) functional theory of categories predicts that children learn to use word forms in syntactic functions in accordance with the implicational hierarchy presented in (14) in §2.4.2. Since the lexical expression of the syntactic function HP precedes the expression of HR, and both HP and HR precede the modifying functions in this hierarchy, it is expected that children will start using content words predicatively, and in a second stage also referentially. It is also expected that the verb-noun distinction will be made at some point early in development, before separate classes of adjectives and adverbs are present. In this section an analysis of spontaneous speech corpora of four Dutch children and four Dutch adults is reported (see also Erkelens, 2006) to test the general prediction that children will show an adult-like use of verbal and nominal forms in syntactic functions.

3.3.1. Method
In this corpus study every content word (as opposed to function word) used in a number of child and adult transcripts was analyzed. The transcripts used came from four children in CHILDES (MacWhinney, 1991) and four adults in the Corpus Gesproken Nederlands ‘Spoken Dutch Corpus’ (Oostdijk, 2000). The adult data were included in the study to set a baseline of how content words are used in Dutch. Adult-to-adult speech was used instead of adult-to-child speech because this would provide a better baseline measure. Child-directed speech often consists of shorter and less complex sentences and hence probably does not contain the full array of the syntactic possibilities of Dutch. The child transcripts are from the children Daan and Matthijs (Dutch Groningen corpus, Wijnen, 1993) and Sarah and Laura (Dutch Van Kampen corpus, Van Kampen, 2004). A selection of the transcripts was made based on the Mean Length of Utterance (henceforth MLU - Brown, 1973) measured in morphemes. In early acquisition MLU is a better indicator of the stage of linguistic development than age because children differ in the speed of acquisition. As discussed earlier, word classes cannot be assigned to single words produced by children, because there is no grammatical information available to tell what the assignment is. Therefore, a minimal syntactic context (i.e., more than one word) is necessary for any analysis. A minimum MLU of 1.8 was set since it is likely that the majority of the child’s utterances at this point are two words or longer. Analysis was continued until the children’s MLU was about 3.6 based on the availability of data. The age of the children ranged from 2;0:17 (years;months;days) to 3;7:25. The adult
transcripts used were two randomly chosen face-to-face conversations involving two adults. These conversations were divided into four separate transcripts, one per adult. The total number of utterances per adult is much smaller than that of the children. As the adult data serve as a baseline for comparison, a small amount of data is sufficient if the analyzed transcripts of the different adults show a comparable pattern. This was the case. Of the 83 transcripts (79 child and 4 adult transcripts), every content word was coded for both its word class in adult Dutch and for the syntactic function in which it was used.

The syntactic functions coded in the corpora are those reported in §2.4.1. They are taken from Functional Grammar: Head of a Predicate phrase (henceforth HP), Head of a Referential phrase (henceforth HR), Modifier in a Referential phrase (henceforth MR), and Modifier in a Predicate phrase (henceforth MP). These syntactic functions are more or less comparable to the categories more commonly known from generative grammar as head of VP, head of NP, head of AP, and head of AdvP. However, since the syntactic functions were coded independently of the lexical items used in them, the terms V, N, A, Adv were reserved for word classes in order to keep the terminology clear. The implication of this separation of syntactic functions and word classes in Functional Grammar is that the head of the predicate phrase does not have to be a verb, whereas the head of the VP does. For example, in the sentence ‘John is president’ (example (12) in §2.4.1), the HP is ‘president’, whereas ‘president’ can never be the head of a VP, since it is not a verb. The following criteria were used for coding a syntactic function:

- HP: the lexical item qualifies a present utterance or object, without being the term for that object.
- HR: the lexical item refers to an object or concept for which it is the term.
- MR: the lexical item is a modification of the HR (the part that refers to an object or concept for which it is the term).
- MP: the lexical item is a modification of the HP (the part that qualifies a present utterance or object, without being the term for that object).

It is important that the coding of the syntactic functions was conducted with as little interpretation of nonexistent information as possible. Child utterances are often syntactically and morphologically incomplete and therefore hard to analyze. In a rich interpretation the researcher makes assumptions about the missing structure. This is not the kind of interpretation that was used in this study. For the initial coding the child’s utterance was only given credit for the word forms that were actually produced. Even if the syntactic function of a word form would have been different if a presumably omitted function word had been present, the syntactic function coded was based on the child’s actual utterance. That is, the position of the
word in the child utterance and the produced morphological markers were taken as indicators of the syntactic function. We will see in the discussion of the results that for some syntactic functions a reanalysis with limited assumptions about omitted structure is necessary.

An unclear (Unclear) option was added since the syntactic function of a word cannot always be determined on the basis of the limited set of data available in the transcript. One-word utterances, for example, are almost always unclear as regards the syntactic function since there is no overt syntactic structure in the utterance itself. Words can only be used predicatively, referentially, or as modifiers in relation to other words in the utterance. Wherever possible, the context of an utterance was used to make inferences about the most likely syntactic function of the word used. All content words that did not clearly express one of the four syntactic functions were coded as having an unclear syntactic function.

The coding of word class was based on the intuition of native speaker coders as outlined above. The assigned content word classes are: verb (V), noun (N), adjective (A), and proper name (PrN). Not all lexical words fall into one of these four categories. Therefore, a category ‘Other’ was included, which contained adverbs, adpositions, and words that were ambiguous between verb and noun. In the discussion of the results below, the precise contents of the ‘Other’ category are further analyzed when relevant. For the assignment of the word class labels, the coders were naïve with respect to the context and the syntactic category in which the lexical items were used. The word class that was most strongly associated with an item was assigned based on intuition, irrespective of the context or the more peripheral categorization possibilities of the word. For example, the English content word sand would have been coded as a noun since this is the most strongly associated word class for adult speakers of English. However, the earlier examples of a child’s lexical innovations showed that this word can be used in a verbal syntactic context as in I’m sanding (see §3.2.1). The word sand also has a verbal meaning in English, as in He sands the door. Only if the coder really felt that the word was equally likely to be a verb or a noun was the label V/N applied. In all other cases where one word class label stood out as the most natural, this label was assigned. For example, zalve in ikke wil zalve ‘I want to ointment’ (see Table 3.1 in §3.2.2) was labeled as a noun, although according to the online Van Dale it can also be a verb. Even for words with a verbal use in modern Dutch, such as the Dutch huiz-en ‘house-INF’, the word was coded as a noun in the utterance ditte huisje ‘this

---

house-DIM’ because noun is the word class most strongly associated with the form *huis* ‘house’. By basing the word class label on adult speakers’ intuitions, the word class is kept as an independent variable in both children and adults. Note that this coding for word class does not imply that the child has the same word class label for the word in her lexicon. Rather, it functions as an independent measure of the kind of lexical item that is used. For a subset of the data (about 15%), two or more coders labelled the word classes. The intercoder reliability for this subset was always above 90%. The codings of the principal coder were taken in cases of disagreement.

3.3.2. Results

According to Hengeveld (1992b), languages differ from each other in their number of word classes. Dutch is characterized as a language with three word classes: verbs, nouns, and adjectives. Dutch verbs function as the HP, nouns can function as the HR, and adjectives can function as both the MR and the MP. The two modifying syntactic functions are thus expressed by a single word class, as shown in example (1).

(1) \[Een\ [mooi]_{\text{MR}}\ [meisje]_{\text{HR}}\]_{\text{Ref. phrase}} \[\text{zing-t}_{\text{HP}}\ [mooi]_{\text{MP}}\]_{\text{Pred. phrase}}

A beautiful girl sing-3SG beautiful

‘A beautiful girl sings beautifully’

As we will see in the data, nouns and adjectives can also be used predicatively. The prediction from the hierarchy of syntactic functions is that the modifying functions will be expressed lexically later in development than the referential and predicative functions. The coded child language corpora were analyzed to investigate how Dutch children learn the different possibilities of content word classes for expressing syntactic functions. First, the analyses of how the syntactic functions are expressed lexically over time will be presented and then the word classes used for each syntactic function will be discussed.

Figure 3.1 shows which syntactic functions are expressed with content words by the children and the adults. The lines represent the percentage of content words used for a specific syntactic function and the data points each stand for a certain MLU group, as indicated on the x-axis. The data presented in this graph show that even the youngest children, who have an average MLU below 2.0, express three syntactic functions lexically: HP, HR, and MR. Examples from this youngest group are provided in (2) and (3).
Figure 3.1. Percentages of the syntactic functions HP, HR, MR, and MP expressed with a content word by Dutch children at different MLU points and by a control group of Dutch adults.

Even in their earliest word combinations, children express at least three different syntactic functions by means of content words. The finding in these data that children do not express the syntactic function MP with content words until they have an MLU of above 2.5 can be due to the relatively low number of words used in MP in everyday speech overall. Even the adults express the MP function in only 2.65 percent of their lexical expressions. The results per syntactic function are provided to establish whether there is a development in the word classes used in each function.

Let us start with the syntactic function HP. Figure 3.2 shows the word classes used by both children and adults to express this function. The raw numbers of content words used in each syntactic function are provided in Appendix 3.1. The
bars represent the total number of content words used to express the HP. Each bar represents an MLU group, indicated on the X-axis. Each of the bar segments represents the percentage of words from a certain word class. Children and adults use the same kind of word classes to express the HP: all word classes are present in every graph bar. This means that children are aware that not only verbs but also nouns and adjectives can be used to express the HP. There are, however, two important quantitative differences between the children and the adults that might point to categorization errors without morphological marking (recall from the earlier in §3.2.1 that errors seemed only observable if morphological marking is present). The upper graph bar segments are larger for the children than for the adults, which indicates that they use relatively more words from word classes other than verb, noun, and adjective to express the HP. Closer inspection of this ‘other’ category reveals that almost all of these words are adpositions, as in the example in (4).

\[(4) \quad \text{pepernoten in } \text{gingernuts in}\]  
(Matthijs, 2:8,05 – MLU=2.3)

Apparently children use more adpositions than adults do to express the head of a predicate phrase. Dutch has a considerable number of verbs that combine with an adposition-like element (so-called ‘particle verbs’). Children seem to first assume that these particles, or adpositions, can be used without the verb. Adult speakers, to a lesser extent, show the same behavior: all cases of ‘Other’ uses by the adults involve also adpositions.12

From Figure 3.2 we see that the younger children use a fairly large proportion of nouns as HP. This percentage drops as children become older and is minimal in the adults. Between the MLU groups 2.5–3.0 and 3.0–3.6 (that is, at an MLU of around 3.0), the proportion of nouns used as HP becomes adult-like. However, closer inspection of the nouns coded as HP in the speech of the lower MLU groups reveals two interpretation problems arising from the lack of morpho-syntactic material with nouns in the predicate phrase.

The first problem relates to the coding criteria used for HP. The syntactic function HP is coded if the content word qualifies a present utterance or object, without being the term for that object. That is, if a child says something similar to example (5), the syntactic function of \textit{trekker} ‘tractor’ can be coded as HP because it predicates over \textit{dat} ‘that’. It qualifies \textit{dat}, without being its term.

---

12 One possible explanation for the over-use by the children is that the adpositions at an early stage are actually Root Infinitives of particle verbs, as proposed by Krikhaar & Van Dijk (1999).
(5)  \([\text{dat}]_{\text{HR, Ref. phrase}} \quad [\text{is}]_{\text{VP pred. phrase}} \quad [\text{trekker}]_{\text{NP pred. phrase}}\) (Matthijs 2:9;15 – MLU=2.3)  
that is tractor

Figure 3.2. Percentages of content words per word class used by children at different MLU points and a control group of adults to express the syntactic function HP (A = adjective, N = noun, V = verb).

Note that in a generative syntactic analysis the noun is not the head in a copular construction (Baker, 2003). The head of the predicate phrase would be either the auxiliary ‘is’ (6) or an empty predicating head (7).

(6)

(7)
The present study followed the functional syntactic analysis of these phrases proposed by Hengeveld (1992b) according to which the lexical items used predicatively in a copular construction are themselves the head of the predicate phrase, e.g., *trekker* in (5). This is a consequence of the centrality of the notion predication in Functional Grammar: every lexical item has a predicative variable and can be characterized as the head of a predicate phrase. The copula is only a language-specific consequence of the fact that a nominal item is used predicatively; it has no syntactic function in itself. The functional analysis of the utterance ‘*dat is trekker*’ according to FG-notation is provided in (8).

\[
(8) \quad (p_i: [(e_i: [(f_i: \textit{trekker}) (prox x_i :\textless\text{-anim}\textgreater )]) (e_i)]) (p_i)
\]

This says that there is a propositional content (p_i), such that there is a state-of-affairs (e_i), such that the property (f_i), *trekker*, is predicated over a proximate, inanimate object (x_i). The copula plays no role in the underlying functional structure of the utterance. The analysis would have been exactly the same if the child had omitted the copula. The coding followed the FG interpretation, so the noun *trekker* in (5) was coded as HP. However, it is equally valid to code this as HR since it is also the term for the object ‘tractor’. The lexical item *trekker* would then in and of itself be an entire referential phrase and that phrase would be the head of the predicate. The interpretation of such utterances is decisive for the coding applied to the syntactic function for which the word is used. If the coder interpreted the content word not as referring to the object but rather as qualifying ‘that’, the syntactic function was HP. However, had the child expressed this utterance with a determiner such as *een* in (9), the interpretation could only be that the content word itself was the term for the object, since it can no longer be predicating ‘that’ by itself. The entire referential phrase might still be interpreted as expressing the HP according to the criteria we used for syntactic coding, but the lexical item *trekker* can only be head of the referential phrase that as a whole is predicating *dat* (10).

\[
(9) \quad [\textit{dat}]_{HR, \text{Ref. phrase}} [\textit{is}]_{\text{HR, Ref. phrase}} [\textit{een}]_{\text{HP, Ref. phrase}} [\textit{trekker}]_{\text{HR, Pred. phrase}} \quad \text{that is a tractor}
\]

\[
(10) \quad (p_i: [(e_i: [(f_i: [(x_i: (f_k: \textit{trekker}) (f_k)) (prox x_j)]) (e_i)]) (p_i))
\]

In the higher MLU groups and with the adults, the obligatory determiners are almost always present if nouns are used.\(^{13}\) Nouns with explicit determiners were never

\[^{13}\text{It is possible in Dutch to use nouns predicatively without a determiner, but only with a restricted set of profession-indicating predicates (e.g., *hij is bakker* ‘he is baker’).}\]
coded as HP in the files of older children and adults. Since the point at which the children show a decrease in the number of nouns used as HP coincides with the point at which Dutch children are known to acquire the obligatory use of determiners (Van der Velde, 2003; Rozendaal, 2008), the apparent over-use of nouns in HP is most likely a consequence of determiner omission and not of the acquisition of categories, but the interpretation remains ambiguous.

Another problem of interpretation concerns children’s two-word-combinations of pronoun and noun, or noun and noun, as in *I tea* or *daddy cookie*. These examples are comparable to the utterance *dat trekker* discussed above. Again, the syntactic function of *tea* and *cookie* could be HP, because they predicate *I* and *daddy*. However, *tea* and *cookie* can also be used as an argument with an omitted verb such as ‘want(s)’. Without morphological marking and detailed information of the context of the utterance, it is hard to decide which coding is the most appropriate.

Figure 3.3. Percentages of content words per word class used by children at different MLU points and a control group of adults to express the syntactic function HR (PrN = proper noun, N = noun, V = verb).

Parallel to Figure 3.2 for the HP, Figure 3.3 presents the results for the syntactic function HR. There is hardly any development in the data for this syntactic function, especially if proper names are taken as part of the noun category. All child groups and the adults almost exclusively use nouns or proper names to express the HR. The most important differences are observable in the relative proportions of proper names versus nouns. Children with a lower MLU tend to use more proper names.
than children with a higher MLU and adults. Although this issue was not analyzed further because it is beyond the scope of this study, it is conceivable that children in the early stages of language development use fewer functional words to refer to persons than adults because they are pragmatically difficult (e.g., third person pronouns). As a consequence, they tend to use the proper name of the person or animal itself to refer to them.

The third syntactic function analyzed is the modifier in a referential phrase (MR). In Figure 3.4, the word classes to express this function are again visualized as percentages of the total number of content words used to express this function. In the syntactic function MR, for the first time a qualitative difference between the children and the adults can be observed. All children use a word class that adults never use to express the MR, namely proper names. Examples of these proper name uses are provided in (11) and (12).

(11) \textit{op Evelien schoot-je} \hfill (Matthijs, 2:07;23 – MLU = 2.8) \\
\textit{on Evelien lap-DIM} \\
‘on Evelien’s lap’

(12) \textit{dis papa huis} \hfill (Laura, 2:05;00 – MLU = 2.2) \\
\textit{this.is daddy house} \\
‘this is daddy’s house’

All four children showed similar uses of a proper name indicating the possessor in the MR function without addition of a genitive marker. Dutch has two genitive markers: the \textit{s} ending similar to English, and the possessive pronoun (e.g., \textit{op Evelien d’r schootje} ‘on Evelien her lap’, \textit{dis papa z’n huis} ‘this is daddy his house’). The data of the children did contain the possessive \textit{s} ending in other utterances of the same transcripts, but it is not rare for children to show optionality in the omission of grammatical markers. Another explanation for the fact that they omit the ending here is that they omit the prosodically weak possessive pronouns \textit{d’r} ‘her’ and \textit{z’n} ‘his’. Using a combination of a proper name and such a weak possessive pronoun is a very common way to express modification in spoken Dutch. Example (12) would have been adult-like if Laura had said \textit{dis papa z’n huis} ‘this is daddy his house’. In that case the MR is actually not expressed by the proper name, but by the possessive pronoun \textit{z’n}. An alternative interpretation could be that the children use the proper names as the first part of a compound (i.e., not a \textit{poppenhuis} ‘doll’s house’, but a \textit{papahuis} ‘daddy’s house’). Although in adult Dutch a compound requires a determiner in this context, we see in Figure 3.4 that the use of proper names in MR is the highest in those MLU groups (< 2.5) where children still
omit the determiners. Both the omission of possessive pronouns and the omission of determiners with compounds are problems in the acquisition of morphology, and not problems of category.

Figure 3.4. Percentages of content words per word class used by children at different MLU points and a control group of adults to express the syntactic function MR (PrN = proper noun, A = adjective, N = noun).

The last syntactic function studied is the MP. Since the number of content words used to express MP is very low across all transcripts, these data are not presented as percentages but as average numbers. Note that the length of the bars as a whole is not very informative, since the groups differ greatly in the number and size of the transcripts. The average number of words from a certain word class used in MP are presented in Figure 3.5. The few content words used in MP are almost all adjectives. The instances of ‘Other’ word classes used to express this function are mainly adverbs that cannot be used attributively such as *zachtjes* ‘softly’, or *erg* ‘very’. All children only used lexical items in this syntactic function that adults would also use; in this respect they are adult-like from the start.
3.3.3. Conclusion

The longitudinal corpus study of Dutch children’s and adult’s spontaneous speech reported here shows that Dutch children perform almost like adults with respect to the use of verbs, nouns, and adjectives across syntactic functions. The differences between children and adults shown by the codings for the syntactic function HP could be attributed to the omission of determiners and pronouns by the younger children. A re-analysis was made of the first non-rich interpretation of the material. Since we know from independent evidence that children in the lowest MLU groups leave out the determiners in obligatory contexts systematically, it seems legitimate to assume syntactic structures in which a determiner would have been produced. If this assumption is made about the nature of the material that was omitted, we do not find any clear developmental stages in the development of word class use over syntactic functions. The children used verbs only predicatively; they used nouns in diverse syntactic functions mostly comparable to adult use; and they used adjectives both predicatively and attributively.

The results from the small corpus study reported in §3.2 on categorization errors are confirmed by this detailed longitudinal corpus study on categorization in general. From the moment children start to combine words, they use these words syntactically according to the adult pattern. This outcome is compatible with

Figure 3.5. Number of content words per word class used by children at different MLU points and a control group of adults to express the syntactic function MP (A = adjective).
different interpretations of the underlying verbal and nominal representations, including those of Baker, Hengeveld, Croft, and Goldberg are all compatible with these child language production data. Dutch two-year-olds categorize their words in a way that results in adult-like use. Whether the grammars of these children contain the same categories as the grammars of Dutch adults remains an open question.

3.4. Conclusions

The studies presented in this chapter tested which of the theories discussed in Chapter 2 are compatible with Dutch children’s production data. The analyses of spontaneous speech corpora revealed that categorization can only be studied in production if children produce sentences (i.e., more than one-word utterances) with grammatical information. Once they do this, they use the majority of their words according to the adult-like syntactic possibilities. Children have at least initial categories of words from the moment that word combinations can be observed.

This outcome seems incompatible with the theories that predicted that categorization errors would occur in an early stage of development. These errors were hardly attested in the children’s spontaneous speech. But the predicted error stage might occur before the age at which production data can be investigated. So, in fact, the adult-like production patterns are at this point compatible with all of the radically different ideas about the representation of verbs and nouns in the grammar. The production studies in this chapter have shown that children already categorize their words correctly from the moment categorization can be studied in their spontaneous speech. The nature of these child categories cannot be assessed with the methods used in this chapter. Since the difference between the theories of verbs and nouns revolves mainly around the nature of their representations, these data cannot distinguish between the different proposals.

The intriguing question now is whether it is possible to study the categorization abilities of younger children. If we can determine how children arrive at the categorization abilities they have at two years of age, we can shed some light on the nature of the categories they have at this age. The next chapter will discuss methods for studying younger children, and the properties of the input speech that are indicative of the categories verb and noun.