Child language cuts capers
Dijk, Marijn Wouter Geertje van

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2004

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, 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), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 6

Ambiguity in Child Language

The problem of inter-observer reliability in ambiguous observation data


Abstract

In language acquisition, inter-transcriber agreement over linguistic categories assigned to recorded utterances is conceived as a measure of observer reliability. We argue that disagreement is not merely a reflection of observer errors or noisy data, but can be a reflection of the genuine ambiguity of early speech. Disagreement arises from the fact that the child is still building linguistic categories, and therefore, from the fact that the language is truly ambiguous. This ambiguity can be quantified by applying concepts from Fuzzy Logic, for which we demonstrate in a case study. After presenting an index of agreement and a Monte-Carlo procedure for calculating the probability of chance agreement, we introduce an index of ambiguity, based on the fuzzy logic notion of degree-of-membership.
6.1. Introduction

6.1.1 Transcription and reliability

Inter-observer reliability is an important issue in many branches of psychology. For instance, the handbook on developmental psychology by Shaffer (1989), a widely used textbook, states (p. 16) that reliability is one of the most important things to consider if naturalistic observation is employed:

Reliability is most often measured by asking a second person to observe the same events that the first observer witnessed and then comparing the observational records of the two observers. If independent observers largely agree on what occurred the observational records are reliable. A lack of agreement indicates that the observation scheme is unreliable and needs to be revised.”(pp. 16).

Although these statements sound simple and easy to apply, they convey a central assumption about the nature of observational data. This is that there are true observable values or categories of behavior that can be identified as such, with more or less difficulty, by the independent observers. This assumption stems from signal-transmission theory, where a signal (the true category) is being transmitted while ‘noise’ is added to the signal (e.g. Green & Swets, 1966). A lack of reliability is caused either by the observers themselves, or by things related to the observation procedure, such as the quality of the recordings. The suggestion to adjust the observation scheme seems obvious, but does not recognize the fact that developing behavior is often ambiguous by nature.

This ambiguity is particularly characteristic of early language development. Often, it remains unclear to which linguistic categories the individual words in the utterances belong, even with a perfect voice-recording system and perfectly trained observers (more specifically transcribers or coders). We will argue that differences between independent observers are, under these circumstances, not merely caused by observer mistakes or a vague measurement design, but can also be a reflection of the genuine ambiguity in the actual developmental data. In the second part of this article, we will also suggest a measure of ambiguity that is based on fuzzy logic principles.

In the field of language acquisition research, it is common practice to study various aspects of child language by means of naturalistic recordings of spontaneous speech. Examples are: how does the child use verbs, how often and in which syntactical constructions does it use prepositions, how many different words are used, etc. To begin with, the observers must transcribe the child’s speech into a written form. One of the widely used standards for the transcription of language is provided by the Chilides-project (MacWhinney, 1991). In the preface of the Chilides handbook, MacWhinney points at the dangers that arise from the need to compress a complex set of verbal expressions into the narrow channel of written language. He considers the greatest danger facing the transcriber in the tendency to treat spoken language as if it were written language. The decision to write out stretches of vocal material using the forms of written language involves ‘major theoretical commitment’, MacWhinney states. ‘The most difficult bias to overcome is the tendency to map every spoken form by a learner – be it a child, an aphasic or a second language learner – onto a set of standard lexical items in the adult language’ (p. 2). To illustrate this bias, Fletcher (1985) noted that both adults and children produce ‘have’ as /uv/ before main verbs. As a result, forms like “might have gone” assimilate to “mightuv gone”. The question is whether children who produce such an utterance have already
acquired the full auxiliary ‘have’. Another example, in French, is the fact that the various endings of the verb are distinguished in spelling but homophonous in speech. If a child says /mAnZ/ ‘eat’, are we going to transcribe the first person singular mange, as a second person singular manges, or as the imperative mange (example from MacWhinney, 1991)?

Childes offers two solutions: the first is to transcribe the utterance phonetically, the second is to use a standardized set of non-standard lexical forms that the Childes system recognizes (such as ‘mightuv’). Although these solutions may to some degree reduce inter-observer differences, there still is plenty of room for disagreement among observers. For instance, consider a Dutch child that says /mEis/ repeatedly. Although it resembles mij (me), and mijn (mine), and the child stretches its arms to reach something, the question remains if the child intends to use a personal pronoun, a possessive pronoun, or something completely different. In which form should it be transcribed? Using a neutral transcription solves little. In the case of the ‘mightuv’-construction, we know that at some point in time the child will acquire the full auxiliary ‘have’, but at what point in time is a transcriber going to replace ‘mightuv’ with ‘might have’? Bear in mind that adults also use the ‘mightuv’ construction, and this would probably be transcribed as ‘might have’.

In a review of the Childes handbook, Serratrice (2000) states that there are nontrivial problems in the classification of utterances into the different linguistic categories. She ascertains a lack of sufficiently clear guidelines for the coding of problematic borderline cases such as the treatment of certain adjectives and past particles, or particles, prepositions and adverbs. She states: ‘... there are no criteria for complex cases in which the coder is called upon to make a sometimes difficult choice’ (pp. 335), and “It would be extremely helpful if some criteria were provided in the manual to ensure that coders make consistent choices’ (pp. 336).

6.1.2 Reliability in language acquisition research

6.1.2.1 The use of reliability indices in language acquisition research

Although the procedure of categorizing observed non-verbal behavior has many points in common with that of the transcription of language recordings, there are various ways of approaching the issue of inter-observer (inter-transcriber) reliability that are characteristic in the literature of language and language acquisition in particular.

Whereas several studies do not mention reliability, other studies report a measure to indicate the overlap between two observers. If distinct transcribers are involved (which is not always the case), authors do not always make explicit how the transcribers co-operated in bringing about the categorization (independent scoring or not). Sometimes it is – implicitly or explicitly – indicated that transcriptions were not independent. For instance, a description of the reliability procedure might be: The principal investigator also listened to the speech samples in order to check his agreement with the interviewers.[..] Instances of disagreement were resolved through repeated listening and analysis, resulting in a perfect agreement in the irregularities. (example taken from Yairi, 1982)

In order to obtain a general impression of how the inter-observer reliability issue is dealt with in recent publications, we examined randomly selected volumes of the Journal of Child Language (Volume 27,
In the Journal of Child Language, 11 studies out of 27 articles use spontaneous speech samples of young children (the majority below the age of four) that were transcribed and/or categorized. Of these 11, four articles do not mention the inter-observer reliability (Grinstead, 2000; Huang, 2000; Prat-Sala, Shillcock & Sorace, 2000; Rowland & Pine, 2000). It should be noted that two of these studies are based on corpus data: Grinstead used the Linaza corpus; Rowland used the Brown (1973) corpus. Two further studies address the reliability issue in an indirect manner (Maratsos, 2000; Parisse & Le Normand, 2000). Parisse and Le Normand make use of what they call an ‘automatic tagger’ on the Le Veile corpus. This automatic tagger takes three linguistic components at a time (morphology, syntax and distributional analysis), mapping the child’s language into an adult structure, using adult interpretation. The level of ambiguity of the two-year-old language is quite high, ranging between 1.15 and 2.30 possible lexical categories for each word. Maratsos used transcripts made from the Kuczaj corpus, but when discussing these transcripts with the original transcriber, there were frequent disagreements (as high as 40%). Furthermore, two articles that acknowledge reliability have used observers who are not independent (Rescorla, Dahlgaaard & Roberts, 2000; Veneziano & Sinclair, 2000). Three further studies use explicit, standard measures based on independent scoring (Bornstein, Haynes, Painter & Genevro, 2000; Goldfield, 2000; Winsler, Carlton & Barry, 2000). Bornstein et al. report an intra-class correlation, ranging from 0.943 to 0.998 for different variables (total number of utterances, number of child word roots, and MLU) in two-year-olds. Winsler et al. reported percentages of agreement and Cohen’s Kappas for independent observers, which ranged from 78% (Kappa 0.66) to 96% (Kappa 0.90) for the different variables (activity, speech and social context in three- and four-year-olds). Goldfield (2000) reports a Cohen’s Kappa of 0.72 (for the frequency of nouns and verbs in infants age 1:8).

In the First Language volume, 10 out of 12 articles analyze spontaneous speech material. Of these 10 articles, 6 mention the fact that there was independent scoring of two (or more) transcribers (Bouldin, Bavin & Pratt, 2002; Guidetti, 2002; Kunnari, 2002; Lacroix, Pomerleau & Malcuit, 2002; O’Neill & Holmes, 2002; Schiff, 2002). Furthermore, 5 studies report a measure of agreement between observers (O’Neill & Holmes, 2002; Kunnari, 2002; Lacroix, Pomerleau & Malcuit, 2002; Bouldin et al. 2002; Guidetti, 2002). These range from 99% to 81%, depending on the type of languages that had to be coded and the age of the subjects. O’Neill & Holmes report an overlap of 99% (character speech and gesture types in 3 to 4 year olds). Kunnari’s percentages range from 99.2% (number of syllables in a word in subjects age 0;11–1;7) to 86% (syllable structure). Lacroix et al. analyze maternal utterances (adult language) in a free play session and report an overlap of 86%. Bouldin et al. reach an overlap ratio of 28 over 30 in a group assignment task (language had to be assigned to a group of subjects that had an imaginary companion and a group who did not). The Guidetti study transcribes early child speech (age 1;4–2;6) and reports an overlap of 81%. Five studies that concern spontaneous child speech do not report inter-observer reliability (Courtney, 2002; Farrar & Maag, 2002; Joseph, Serratrice & Conti-Ramsden, 2002; Leonard, Caselli & Devescovi, 2002; Schiff 2002). In summary, although we have covered only a small sample of the literature, it has been made clear that the issue of inter-observer reliability is addressed in a wide variety of ways. Whereas a large group of studies does not explicitly mention reliability, other studies report detailed analyses. Of these
studies, the concept of overlap proportion or overlap percentage is often reported, sometimes in combination with a Cohen’s Kappa.

6.1.2.2 Problems with commonly used indices of reliability
The fact that reliability indices are reported that are based on inter-rater agreement, is in itself not sufficient. The index used must be informative about the kind of agreement that really matters for the topic at issue and, most importantly, the index used must be based on assumptions that are valid for the context in which the index is used. The widely used Kappa index, for instance, is now increasingly criticized. Its underlying assumptions often do not match the properties of the rating process to which the index is applied. For instance, its correction for chance agreement implies, first, that the marginal frequencies (frequencies of categories assigned) are fixed and, second, that the correction must be equal to the average of all possible agreements that arise from randomly assigning the fixed number of categories to the items (utterances, observations). However, the assumption of such random assignment is not in agreement with the assumption of fixed marginal frequencies. Random assignment implies the absence of any possible constraint on the link between an item and an assigned category. However, if, by implication, no such link exists, the frequencies of the assigned categories must also be completely random (note that all this applies under the null hypothesis that the agreement is based on chance alone and that the actual agreement must be corrected for chance).

Moreover, the chance correction results in Kappa values that are crucially dependent on the size of the marginal frequencies (e.g., many instances of category A, few of category B), which implies that such values are often not comparable across rated samples and studies. Currently, there is a rather extensive literature on the problems associated with the use of Kappa and many alternative indices have been suggested (for a particularly informative overview and extensive bibliography, see Uebersax, 2001; see further: Cicchetti & Feinstein, 1990; Cook, 1998; Feinstein & Cicchetti, 1990; Guggenmoos-Holzmann, 1996; Gwet, 2001, 2002; Uebersax, 1987, 1988; Zwick, 1988). Uebersax (2001) makes the recommendation that researchers should use very simple indices of agreement (simple percentages of agreement, overlap ratios, etc.); these reveal more about agreement than most of the more complicated measures, which he sees as complements and not substitutes for the simpler methods. A problem with such simple indices is that it is not clear to what extent they capitalize on chance. In order to solve this problem, the researcher should specify a model of chance rating that matches the rating procedure at issue and then run a simple simulation study based on this model to determine the probability that a given rating is indeed based on chance (we will demonstrate this with our overlap ratio).

However, the main point that we want to make in this article is not that current inter-observer agreement indices should be replaced by better ones. Our main goal is to argue for a change of view on the meaning of inter-rater agreement. Currently, agreement indices are seen as indicators of the quality of the rating and hence, of the competence of the raters (assuming, for simplicity, that the recording itself is impeccable). Our message is that inter-rater agreement which is based on the work of competent raters should be seen as information telling us something interesting about the nature of
the rated phenomenon. In order to arrive at this conclusion, we shall first introduce our case study and then proceed to the notion of fuzzy sets and ambiguity.

6.2 The case study: filtering language for prepositions

6.2.1 Overview of the study
We encountered the issue of inter-observer reliability in our own case study on the acquisition of spatial prepositions in young children. Two observers had the task of filtering spatial prepositions out of videotaped samples of spontaneous speech. The total set of spatial prepositions consisted of in, uit, op, af, voor, achter, tussen, over, bij, naar, onder, boven, binnen, buiten, door (approximate translations are in, out, on, off, before/in front of, after/behind, between, over, near (to)/at, to, under, above, in/inside, out/outside, through). The two children were each recorded twice for 60 minutes, around their second birthday. The task of the observers was to watch these four 60-minute videotapes (two per subject) and only transcribe the utterances with spatial prepositions. The observers thus had to judge each utterance for the presence of a preposition, and transcribe the positively identified full utterance. This task is highly similar to the procedure in the study of Stockman and Vaughn-Cooke (1992) in which speech samples were filtered for possible locative expressions. Locative expressions included either a locative word or/and a verb that takes a locative argument.

In our study, the total number of utterances varied for each tape of 60 minutes. Since not all utterances were transcribe, we cannot be sure about the total number of utterances in each recording, but our pilot study demonstrated a minimum of 100 utterances. Furthermore, we also learned from this study that the relative occurrence of utterances that contain a preposition varied from 0.8% to 18% of the total number of utterances recorded as such. The observers have formed their expectations on the basis of these data, and consequently are likely to expect a maximum of 20% preposition utterances.

By selecting, for instance, 20 utterances as utterances containing a preposition, the observers implicitly agree on the remaining ones as utterances that do not contain a preposition. However, the exact number of the additional set is unknown. If we had used a fully transcribed corpus, we would have known the exact total number of utterances and thus also the number of non-preposition utterances over which the observers agree.

Thus, the dataset we acquired has two typical characteristics that complicate the application of Cohen’s Kappa (see the preceding discussion of the Kappa). First, the marginal frequencies are highly skewed (the utterances without prepositions greatly outweigh the utterances with prepositions in number), and second, we have no knowledge of the exact frequency of these non-preposition utterances. Therefore, we calculate an index of agreement – more precisely an overlap percentage – based only on the positive cases (i.e., the utterances explicitly selected by the observers as containing a preposition). At this point, it is useful to elaborate on the concept of this overlap percentage based on the positive cases.
6.2.2 Overlap index

The idea behind this index is as follows. Let \( n_1 \) be the number of preposition utterances selected from the corpus by observer 1 (e.g., 10) and \( n_2 \) the number selected by observer 2 (e.g., 15). The total number of utterances selected by the pair of observers is \( n_1 + n_2 = m \) (e.g., 10 + 15 = 25). From the \( n_1 \) utterances selected by observer 1, \( s_1 \) (e.g., 8) are similar to utterances also selected by observer 2 (and from the \( n_2 \) utterances selected by observer 2, \( s_2 \) are identical to utterances also selected by observer 1; by definition, \( s_1 = s_2 \)). From the first observer’s \( n_1 \) selected utterances, a number equal to \( (n_1 - s_1) = d_1 \) are not identical to utterances selected by observer 2, and, the other way round, from the second observer’s \( n_2 \) selected utterances, \( (n_2 - s_2) = d_2 \) are different from utterances selected by observer 1. For simplicity, we shall call \( s_1 \) – and by definition also \( s_2 \) which is equal to \( s_1 \) – the overlap, \( d_1 \) the non-overlap of observer 1 to observer 2, and \( d_2 \) the non-overlap of observer 2 to observer 1. We define our measure of agreement – or, more correctly, our overlap – as the proportion of items in the set of \( n_1 + n_2 \) items that have the property ‘identical’ over the total number of items in that set. We can express this measure in terms of the numbers of overlapping and non-overlapping items as follows:

\[
\text{Overlap ratio} = \frac{s_1 + s_2}{s_1 + d_1 + s_2 + d_2}
\]

In somewhat different terms, the equation reads as follows

\[
\text{Overlap} = \frac{2 \times \text{overlap}}{(2 \times \text{overlap}) + (\text{non-overlap Obs1:Obs2}) + (\text{non-overlap Obs2:Obs1})}
\]

It should be noted that this definition is highly similar to the definition used by Guidetti (2002). In the literature, we have encountered definitions that are somewhat different, dependent on the task and the available data. Table 6.1 shows the reliability results, in terms of overlap ratio, of two subjects in which the transcripts of the two observers were compared utterance by utterance. The first column, ‘Overlap’, displays the total number of prepositional utterances that both observers agreed on (\( s_1 \) or \( s_2 \), since \( s_1 = s_2 \)). The second and third columns respectively state the number of utterances that were transcribed by observer 1 (Obs1) as a prepositional utterance and not transcribed by observer 2 (Obs2), and vice versa. In the final column, the overlap is calculated according to the equation specified above. In summary, the overlap percentage specifies the agreement between observers 1 and 2 based solely on the items of interest, namely the utterances that contain – effectively or potentially – a preposition.

6.2.3 Probability that the observed overlap is based on chance

As can be seen in Table 6.1, the overlap ratios are both around 0.87. It should be noted that these values were achieved with recordings of good quality, and after extensive training of the observers. The question is what these numbers signify. First, we tested the probability that the positive overlap-ratios found in our case study were based on accidental agreement.
Table 6.1
Reliability results in the study of spatial prepositions (sum of two 60-minute recordings per subject)

<table>
<thead>
<tr>
<th></th>
<th>Overlap</th>
<th>Non-overlap</th>
<th>Overlap ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OBS1-OBS2</td>
<td>OBS2-OBS1</td>
<td></td>
</tr>
<tr>
<td>Subject L.</td>
<td>98</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>Subject B.</td>
<td>17</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

We did this by carrying out a resampling procedure (more specifically a Monte Carlo simulation), using Poptools (Hood, 2001), and found that the probability (p-value) that the observed overlap is based on chance is very small (p < 0.01), a finding which is probably not very surprising. This means that although the value of 0.87 is considered moderately high by psychological standards, it differs highly significantly from a chance model. Second, the values we have reached in this case study are in the same order of magnitude as the value reported by Guidetti (2002) and Kunnari (2002), while the tasks or ages in both studies are highly similar. Thus, although the values of 0.87 are not very high, they do

12 The criterion value was the overlap-percentage generated with a random model (two blind observers categorizing 100 utterances on the basis of the 20:80 % distribution). With this technique we investigated the question what the chances are that our empirically obtained percentage of agreement is based on a categorization procedure carried out by two ‘blind’ observers. The total number of utterances was 100, a minimum we choose conservatively, because higher total numbers of utterances will only decrease the probability of accidental overlap and thus will increase significance. We tested the model by randomly permuting the values 1000 times.

13 Although Monte Carlo procedures are relatively easy to perform with standard software (e.g. Microsoft's Excel extended with a freely available add-in such as Greg Hood's Poptools or our own Functions), many researchers are unfamiliar with them. However, the 95% boundary of agreement by chance can also be approximated by means of the following equation

\[ 95\% \text{ boundary} = 0.68 - 1.04^p \cdot N^{0.05 + 0.15 \ln(p)} \]

for p the expected proportion of prepositions (or any other relevant category) and N the estimated - or known - total number of utterances in the sample. We obtained this equation by first running Monte Carlo procedures for calculating the 95% boundaries for a matrix of expected proportions of prepositions (ranging from 0.5 to 30%) with expected numbers of utterances in the sample (ranging from 10 to 230). We then fitted a power model to the set of 95% boundary values and estimated the parameters of the power model as functions of N (estimated number of utterances) and p (expected proportion of prepositions over non-prepositions). It goes without saying that the equation can be used with any type of syntactic categorization, based on a yes/no decision in a language sample with a known or estimated number of utterances. Note that the statistically relevant 95% boundary value is the value calculated for the largest possible expected p and the smallest possible expected N.
not only differ significantly from chance, but are also not uncommon in the literature on early child speech.

**6.3 Discussion: ambiguity and fuzzy sets**

**6.3.1 Development and fuzzy categories**

We argue that in the case of early child language, there might be a ceiling effect on the reliability values that is not only caused by ‘mechanical’ issues, but also stems from the intrinsic ambiguity of early child speech. This means that even with two infallible observers and excellent recordings there will still be disagreement among observers. In the case of filtering out one or more language variables, observers have the task of deciding whether specific utterances belong to a particular linguistic category or not. We argue that early child language is intrinsically ambiguous because it has to be judged against categories that are not yet fully developed in the child language system. Note that this problem is closely linked to the theoretical assumptions of the researchers. In the full competence approach (posed by Chomsky, for instance 1986), the child has either set the parameter or not, and intermediate positions are theoretically not possible. This approach does not allow a fuzzy definition of the objects and categories of language acquisition. However, dynamic systems theory (Thelen & Smith, 1994) claims that behavior, more specifically language, is constructed in real time, by development that operates as a self-organizing system. While the child is still building linguistic categories, the language output is truly ambiguous. If the recording and transcription procedure were carried out under ideal circumstances, the resulting lack of inter-observer reliability is likely to be the direct effect of this ambiguity. If the acquisition of a particular linguistic category is constructed in real time (as stated by dynamic systems theory), and the child is still in the building-process, it may in some instances be impossible to decide whether or not a specific utterance belongs to this category-under-construction. Another possibility is that our categories do not exhaust the empirical possibilities, i.e., that there exist many forms that belong to two different categories at the same time. These forms are intrinsically ambiguous.

There is still one issue that needs to be addressed in relation to ambiguity in child utterances: it could be argued that only the child’s output is ambiguous, and not the linguistic categories themselves. This would mean that the child’s intentions are not ambiguous, but only the expressions. This form of ambiguity could be called ‘ambiguity-in-expression’, while ‘ambiguity-in-existence’, on the other hand, implies that the acquired linguistic categories are ambiguous themselves. In the case of ambiguity-in-expression, the transcriber faces difficulties because the observable language lacks specific information about which linguistic categories the individual words belong to. Beneath the surface, the child has fully acquired the linguistic category at issue, only it is not clearly observable. In itself, such a theoretical position is defensible. However, one would still have to explain the discrepancy between the observable output and the child’s intentions. In this article we merely want to point at the theoretical possibility of ambiguity-in-existence, which means that the observed categories are
ambiguous themselves. Moreover, we argue that both types of ambiguity can be handled with the aid of fuzzy logic.

6.3.2 Fuzzy Logic and the notion of degree-of-membership
The concept of ambiguous categories is closely related to fuzzy logic, a mathematical theory that can be used to program computers to 'make decisions' based on imprecise data and complex situations (McNeill & Freiberger, 1993). It is a theory of fuzzy sets, sets that calibrate vagueness and ambiguity. Fuzzy logic rests on the idea that all things admit of degree. Temperature, distance, length, friendliness, all things come on a sliding scale. In contrast, most of traditional logic, set theory and philosophy have prescribed sharp distinctions. They have forced us to draw lines in the sand. Something belongs either to a category, or not. For instance, a novel might have 90 pages or more, a novella less than 90. Thus, a 91-page work would be a novel, an 89-page a novella. If the printer reset the novella in a larger type, it would become a novel. Fuzzy logic avoids such absurdities.

Lukasievicz (1878–1955) took the first step towards a formal model of vague classes. He invented the substructure of fuzzy sets, an early logic based on more values than true or false. In this substructure, 1 stood for true and 0 for false. But, in addition, 0.5 stood for possible. The third value acts as a wedge, cracking the traditional true/false division apart. This sliding scale yields greater precision and can quantify degrees of truth. In modern fuzzy logic, ‘objects’ (observations, objects, properties, etc.) are always assigned a degree-of-membership (for a particularly clear technical introduction, see Ross, 1995; McNeill & Freiberger, 1993 provide a highly accessible introduction; see further: Kosko, 1993; 1997; Nguyen & Walker, 1997; von Altrock, 1995). In mutually exhaustive classes, objects have a degree-of-membership of either 1 or 0 (for instance, a book is either a novel or a novella). In fuzzy logic, a book can have any degree of membership between 0 and 1. For instance, a book of 100 pages has a degree of membership of 1 to the category ‘novel’, whereas a book of 90 pages has a degree of membership of 0.5 and a book of 85 pages has one of 0.2. According to Ross (1995: pp. 13), fuzziness describes the ambiguity of an event or object. Maximal ambiguity arises if an object has a degree of membership equal to 0.5. In this case the object has a position at the exact mid-point between the two extremes. When this is the case, its ambiguity is equal to 1. If the degree of membership is either 0 or 1, ambiguity is 0.

6.3.3 Developing linguistic categories as fuzzy sets
The suggestion to treat linguistic categories as fuzzy sets is quite uncommon, to say the least. However, we suggest that this theoretical application is not far-fetched. For instance, the research dealing with ‘filler syllables’ (also called Prefixed Additional Elements, PAE) encounters a similar issue when discussing the status of this phenomenon (see, for instance, Veneziano & Sinclair, 2000). The question is whether these filler syllables can be considered grammatical elements or not. Filler syllables appear in the period of early language acquisition, and their possible sources and functions have long been a subject of interest for many researchers. Researchers have been warning against
crediting the child with early grammatical knowledge on the basis of the appearance of these elements. For instance, Braine (1963/1973, pp. 415) concludes in his study on subject Steven: “while it is quite likely that these elements are an interesting distillate of the unstressed and phonetically often obscure English articles, prepositions and auxiliary verbs, there is no basis for giving them the morphemic status at this stage in Steven’s development”.

One group of authors (e.g. Dolitsky, 1983; Kilani-Schoch & Dressler, 2000; Peters, 1990; Scarpa, 1993; Simmonsen, 1993; Veneziano et al., 1990) linked filler syllables more specifically to the child’s development of grammatical morphemes, considering them ‘an intermediate form on the way to grammatical morphemes’ (Veneziano & Sinclair, 2000: pp. 463). Thus, although the filler syllables are not yet considered to be articles, prepositions or auxiliaries, they take an intermediary position. In fuzzy logic terms, they might acquire the value of 0.5 on the dimension ‘grammatical’, where the value of 0 stands for ‘not-grammatical’ and 1 for ‘grammatical’. In the case of the task of filtering language for prepositions, subject L in our case study produced the utterance ‘stok gaat onder de door’ (stick goes under /the N?(through)/). The adult form of this utterance probably is ‘stok gaat er onderdoor’ (stick goes under it). However, the child places the article ‘de’ before ‘door’ and thus suggesting a noun ‘door’ ((the) through). The question is how many prepositions there are in the utterance, one (just ‘onder’ or the combined form ‘onderdoor’), or two (both ‘onder’ and ‘door’)?

Fuzziness is a characteristic of patterns or categories that are inductively formed by network devices (see, for instance, Jagielska, Matthews & Whitfort, 1999; von Altrock, 1995). Since the brain can be conceived of as such a network device, it is understandable why categories such as prepositions take the form of fuzzy classes, particularly during the period that the category is not yet fully consolidated.

In figure 6.1, we compare a classical with a fuzzy logic approach to the issue of linguistic category membership. Assume that we ask ourselves whether words such as ‘on’ or ‘in’ that a child is using over the course of several months – for instance between months 20 and 32 – are prepositions. This question amounts to the issue of whether or not the child possesses the preposition category. Figure 6.1A shows the classical developmental or emergence answer: somewhere along the time scale the preposition category emerges (in this case around month 26). Since a word is either a preposition or not (has a degree of membership that is either 0 or 1), the preposition category must emerge in a discontinuous fashion. Figure 6.1B represents the classical competence view: the category preposition has been there all along and just waits to be filled with actual words (the degree-of-membership is 1 all the time). Figure 6.1C shows how the preposition category gradually emerges and becomes fully established around 28 months (this age by way of example). This case represents the theoretical viewpoint of ambiguity-in-existence that we presented above. For instance, around the age of 25 months, words such as ‘in’ and ‘on’ have a degree of membership of about 0.5. This degree of membership could be qualified as ‘they are preposition-like, but are still not real prepositions’, or any other qualitative specification that expresses the fuzzy nature of those words during this stage of the child’s language development. Figure 6.1D shows a simplification of the continuous membership function in the form of a linear increase. The latter representation, which causes only minimal loss of information in comparison to the S-shaped function, is customary in fuzzy logic (for instance, see
Kosko, 1993, 1997; Nguyen & Walker, 1997; Ross, 1995; von Altrock, 1995). Finally, Fig. 6.1E shows that the degree of membership to the preposition class of words such as in and on is the complement of their degree of membership to the class of non-prepositions.

Figure 6.1. Classical and fuzzy logic representation of class membership. The class membership applies to a set of words (on, in, to, ...) and to the category “preposition”. Changes in class membership are represented over time in months. Figure 1A shows a classical, discrete emergence approach: prepositions emerge at the age of 26 months. Figure 1B shows the discrete competence approach: the preposition category has been present all the time. Figure 1C represents the fuzzy logic approach: class membership increases continuously, until it reaches full membership around the age of 29 months. At the age of 25 months, words such as on, in, etc. have a preposition class membership of about 0.5. Figure 1D is an alternative representation of the fuzzy membership curve: the straight line provides a standard simplification of the continuous membership curve. Figure 1E compares the membership curve of the class “preposition” with the complementary membership curve of the class “non-preposition”.

122
6.4 Inter-observer agreement and intrinsic ambiguity

6.4.1 High agreement does not necessarily indicate competent rating

In principle, a high level of agreement between observers is viewed as an indicator of the ‘goodness’ of the ratings made (see our discussion about problems with indices of agreement). Thus, the best rating and the best raters should produce an agreement of a 100%. However, the high agreement is only an indicator of a good rating if the following assumptions are accepted. First, the rated phenomena must correspond with ‘crisp’ categories, i.e., they must belong to a category or not. Stated otherwise, it is assumed that the phenomena are not ambiguous. Second, it is assumed that the high agreement is the result of genuine rating quality, carried out by correctly trained, competent observers. However, the high agreement could just as well have resulted from shared errors and biases among the observers. Thus, high agreement is not automatically a guarantee for high quality rating. It is only a guarantee if we accept that the rating has been carried out in the most competent possible way. As scientists, we usually accept this because we trust that our colleagues know their trade. In principle, researchers should justify the quality of the ratings presented by showing that they started from an explicit definition of the rated categories, by making explicit which empirical indicators have been used and where, eventually, cut-off points occurred. They should tell us whether the raters were trained and how this training took place, whether the raters were novices (students for instance) or experts and how disagreement has been treated. If we, the readers of the scientific articles, are informed about these issues and can conclude that the best possible care has been given, we should trust that the ratings have been done in a competent way and are, in that sense, maximally reliable, irrespective of whether the level of agreement is high, intermediate or low.

If the phenomena themselves are truly ambiguous, truly ‘fuzzy’, high quality rating (in which raters have to decide about whether an item does or does not belong to a category) should result in a considerably lower agreement than in the case of basically unambiguous items. In this case, high agreement would be an indicator of low quality rating, for instance rating based on common errors and shared biases. If the quality of the rating has been justified in the way described above, we may trust that the rating is reliable, irrespective of the inter-observer agreement.

This conclusion radically shifts the meaning of the level of inter-observer agreement. Whereas inter-observer agreement is traditionally interpreted as an indicator of the quality of the rating, we must in fact see it as an indicator of the ambiguity of the data. Part of the ambiguity will most likely be of an extrinsic nature. For instance, a car drives past and the noise prevents the observer from understanding what the child exactly says. Another possibility is that the observer may have paid less attention than necessary and may have made an erroneous interpretation. In general, it must be possible to determine how often such problems may have occurred in the worst (reasonable) case.\footnote{The point is not that we should try to find the ‘real’ amount of such cases, because we will probably lack the information to do that. Instead, we should try to determine a reasonable limit, for instance by asking the raters (always provided they are trained professionals) how confident they are that they made a real error, e.g. by being distracted, in 20% of the utterances, 10%, 5%, etc. Somewhere along this sliding scale we can establish a}
Chapter 6

The remaining part of the ambiguity, which in the case of good recordings and competent observers should be the largest part, is intrinsic by nature, i.e., it reflects the intrinsic ambiguity of the data. The ambiguity refers to the categories to which the data are assigned. If those categories are artificial (‘human made’), ambiguity may result from badly specified criteria for distinguishing them. In this case, the ambiguity reflects the inadequate definition of those categories. In the much more interesting case of ‘natural’ categories (whatever that may mean exactly), intrinsic ambiguity refers to the ambiguity of the categories itself, for instance, a child’s early preposition that is only a preposition to a certain extent (we have called this form ambiguity-in-existence earlier in this article).

In case we wish to use the observer disagreement as an indicator of intrinsic agreement, we must have an idea about how much of the observed disagreement is indeed intrinsic, i.e., how much is in fact extrinsic. Although we do not know how much is extrinsic, we probably have an informed idea of the limits between which this extrinsic ambiguity will vary. At best, nothing is extrinsically ambiguous, at worst, a certain percentage is extrinsic (which we should estimate, given that we know the quality of our data). In this case, we recommend the use of a systematic what-if-procedure (see van Dijk & van Geert, submitted, chapter 5 of this thesis), which calculates an index of interest (in this case, an index of ambiguity) based on an informed worst-case scenario.

6.4.2 (Dis)agreement as an indicator of ambiguity

How can the level of (dis)agreement between observers be turned into an ambiguity index? Assume, to begin with, that there is no extrinsic ambiguity in the set (this is only a simplifying assumption, implying that we have not made transcription errors, that all utterances are clearly understandable, etc.; see below for a practical procedure). Assume, next, that in a set of utterances, the majority of the utterances can be unambiguously interpreted as either containing or not containing a preposition. Let us assume that ‘most’ means 90% here. Thus, for 90% of the utterances, an utterance either has a degree-of-membership of 1 or has a degree-of-membership of 0 to the category ‘contains-a-preposition’. Consequently, the remaining 10% of utterances have a degree-of-membership that is neither 0 nor 1, i.e., a degree-of-membership that is somewhere between 0 and 1. Let us also assume that, of this remaining set of utterances, none of the possible degrees-of-membership is privileged, e.g., that degrees-of-membership of 0.1, 0.5 or 0.9, to name just a few, are equally probable. If this is a reasonable assumption, we can proceed as follows (if it is not a reasonable assumption, i.e., if we have reasons to believe that degrees-of-membership are not evenly distributed, we will need to follow a line of reasoning that is equal in terms of underlying principles to the one that we will continue here, but that differs in terms of numbers).

If every possible degree of membership has the same probability of occurring in the sample, we can conclude that the average degree-of-membership (average DoM) to the category of prepositions is 0.5 (which is exactly the middle value in the three-valued logic). We can also assume that the probability that the raters place an utterance in a particular category (e.g., prepositions) is proportional to the

‘reasonable’ limit. It is recommended that such limit should be on the ‘safe side’ and that no discussion should be devoted to the question whether it is 1% or 2%, for instance, because that difference is not meaningful.
degree-of-membership. If the degree-of-membership is 0.5, the utterance is maximally ambiguous and the probability that it is placed in the preposition category is 0.5 (if the DoM is 0.8, the probability would be 80%). Thus, the ambiguous cases will be randomly categorized as either P (preposition) or N (not-preposition) and will occur in the following combinations over the two raters: PP, NP, PN, NN. If all ambiguous prepositions had a degree-of-membership of 0.5, each combination would have equal probabilities. However, since the degrees-of-membership vary linearly between 0 and 1 (according to our assumption of even distribution), the probabilities of occurrence of the cases are different, namely $\frac{1}{3}$ for the PP combination, $\frac{1}{3}$ for the NN-combination and $\frac{1}{3}$ for a combination of not identical categories (i.e. PN or NP).

It may come as a surprise that the number of PN or NP cases is $\frac{1}{3}$ and not $\frac{1}{2}$. If all utterances had a DoM of 0.5, and thus a probability of 0.5 to be categorized as either P or N, the distribution of combinations would be $\frac{1}{4}$ PP, $\frac{1}{4}$ NN, $\frac{1}{4}$ PN and $\frac{1}{4}$ NP, and thus $\frac{1}{2}$ mixed (PN and NP). However, because the DoMs vary between 0 and 1, with all DoMs between 0 and 1 equally probable, the statistical distribution of the cases is different from the former cases: it is $\frac{1}{3}$ PP, $\frac{1}{3}$ NN and $\frac{1}{3}$ mixed (NP or PN). Note that, for all cases belonging to the ambiguous set, the raters agree, on average, in two-thirds of the cases, but this agreement is truly chance agreement, since the utterances that were rated are truly ambiguous (the ambiguity pertains to the utterances in the present example, and does not entail any form of empirical statement). Thus, if all utterances in the ambiguous set have degrees-of-membership that vary evenly between 0 and 1 (basic assumption) and if n ratings are mixed (PN or NP), then on average n ratings will amount to an accidental agreement about the P status and another n ratings will amount to an accidental agreement about the N status. This reasoning follows directly from our observation that the number of PP, NN and mixed (PN or NP) cases is equal, namely each is $\frac{1}{3}$ of the total number of ambiguous cases.

We will now use this simple set of proportions to specify an index of ambiguity, based on disagreement between two competent observers.

### 6.4.3 An index of ambiguity

Since we are studying prepositions, we wish to have an index of the ambiguity of the set of prepositions (and assumed prepositions) that occur in our corpus. Thus, we calculate the ambiguity for the set of utterances that have a degree of membership to the category preposition that is greater than 0. In the framework of fuzzy logic, ambiguity is expressed in terms of the degree of membership of an object (an utterance in this case) or a collection of such objects. From the previous section, we know that if raters disagree about n utterances and agree about p utterances (which they both identify as prepositional utterances), the total number of ambiguous utterances in the sample will on average be $3n$. The agreement set (the set for which both observers agree that the utterance contains a preposition) contains, on average, n utterances that are ambiguous (have a DoM smaller than 1).

Let us take Subject L (see Table 6.2) as an example. The observers agreed on 98 utterances as utterances that contained a preposition. There was a non-overlap of 12 utterances for observer 1 compared to observer 2 and a non-overlap of 18 utterances for observer 2 compared to observer 1. Thus, the number of utterances for which the observers disagreed was $12+18 = 30$. With 30 items in
the disagreement set, the average number of ambiguous items will be 30*3 = 90. Of these 90 ambiguous items, 30 will, on average, have ended up in the agreed-on preposition set (and another 30 will be in the agreed-on non-preposition set, but these are not at stake here). Thus, of the 98 items in the agreed-on preposition set, 30 are, on average ambiguous and, thus 98–30 = 68 are non-ambiguous, i.e., they have a degree-of-membership of 1. Note that, logically, there cannot be more ambiguous items than there are items in the set, and thus, that the result of the subtraction must at minimum be 0\(^{15}\). Thus, for the set of utterances with a degree of membership bigger than 0, 68 have a DoM of 1 and 90 have DoMs that linearly vary between 0 and 1. Their average DoM is 0.5. The total average DoM is \((68*1+90*0.5) / (68+90) = 0.72\). The average DoM of subject B with 17 items in the agreement set and 4 and 1 in the non-overlap sets is also 0.72.

<table>
<thead>
<tr>
<th></th>
<th>agreement set</th>
<th>non-overlap</th>
<th>non-overlap</th>
<th>non-ambiguous</th>
<th>total set</th>
<th>average DoM</th>
<th>Level of ambiguity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obs1·Obs2</td>
<td>Obs1·Obs2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject L.</td>
<td>98</td>
<td>12</td>
<td>18</td>
<td>68</td>
<td>90</td>
<td>158</td>
<td>0.72</td>
</tr>
<tr>
<td>Subject B.</td>
<td>17</td>
<td>4</td>
<td>1</td>
<td>12</td>
<td>15</td>
<td>27</td>
<td>0.72</td>
</tr>
</tbody>
</table>

The equation for the average DoM of all items in the set of interest (in this case, the set of prepositions, whether they are ambiguous or not) is as follows. Let, in accordance with our overlap ratio equation, \(s\) be the overlap set, i.e., set of items for which both observers agree, \(d1\) the non-overlap of observer 1 compared to observer 2, and \(d2\) the non-overlap of observer 2 compared to observer 1.

\[
\text{DoM} = \frac{s - (d_1 + d_2) + 3 \cdot (d_1 + d_2) \cdot 0.5}{s - (d_1 + d_2) + 3 \cdot (d_1 + d_2)}
\]

\(^{15}\) Since the number of ambiguous items is an estimated number, namely the mean of the distribution of the possible number of ambiguous items, this estimated number could at times be bigger than the number of items in the agreement set (although empirically it seems unlikely that this should occur). If it occurs, however, the result of the subtraction is set to 0.
which simplifies to

\[
\text{DoM} = \frac{s + 0.5 \cdot (d_1 + d_2)}{s + 2 \cdot (d_1' + d_2')}
\]

The degree-of-membership gives an indication of the average ambiguity of the utterances in the sets of utterances that were identified as preposition utterances by at least one of the observers. If the average degree-of-membership is 1, the set is entirely non-ambiguous. If the average DoM is close to 0.5, the set is maximally ambiguous. We can use this principle to transform the average degree-of-membership into a direct measure of ambiguity \(a\). Let maximum ambiguity be equal to 1 and minimal ambiguity equal to 0. The required transformation then equals:

\[
a = 1 - 2 \cdot (\text{DoM} - 0.5).
\]

An average DoM of 0.72 (see Table 6.2) corresponds with a level of ambiguity of 0.56 (somewhere midway complete ambiguity and complete non-ambiguity). This index of ambiguity is approximately equal to the proportion of the number of ambiguous utterances over the total number of utterances in the P-set. For instance, for subject L the total set is the sum of the ambiguous and non-ambiguous items, i.e., \(90 + 68 = 158\). The proportion of ambiguous items over the total set is thus \(90/158 = 0.569\), which approaches the level of ambiguity of 0.56 calculated according to the first method. This proportion is just another, intuitively simple expression of ambiguity. Whatever index one wishes to choose, its meaning comes alive only in relation with psychologically or linguistically meaningful variables. For instance, one may ask whether and how ambiguity changes in the course of the development of the variable at issue (prepositions, for instance) or whether there are differences between variables in terms of ambiguity (prepositions and verbs, for instance).

### 6.5 Conclusion

In the case study we have presented, we demonstrated a reliability procedure that can be used to establish the statistical significance of the overlap ratio (0.87 in the data) against a chance model. As it turned out, these values are way out of the range of a chance model. The values we found also correspond to values found in other studies where observers face a similar task of classifying spontaneous early child speech into adult linguistic categories.

We have claimed that the strength of the inter-observer agreement does not necessarily indicate whether observers need additional training. It goes without saying that inter-observer differences between observers are not caused solely by ambiguity. Naturally, training should be optimized in order to reduce its negative effect on the reliability measure. But even if training is optimal, there might be a ceiling effect: a reliability value the observers cannot exceed, when parts of observed events are ambiguous. Conceiving linguistic categories in early child language as a fuzzy set, does not only explain a possible ceiling effect of the inter-observer reliability, it might also be a more adequate representation of how children acquire these categories.
Two final problems deserve further discussion. The first concerns the fact that in the above calculations we implicitly assumed that all disagreement among observers is due to intrinsic ambiguity of the categories (e.g., preposition versus non-preposition) and none to extrinsic ambiguity (e.g., due to noisy recordings, inadvertence and lack of knowledge from the side of the observers). Stated this absolutely, it is an unrealistic assumption. However, if all cases of notably noisy data have been removed from the sample and the observers are competent and have been adequately trained, the assumption that the remaining disagreement is due to intrinsic ambiguity can be used as a reasonable working hypothesis, with the additional assumption that, for instance, no more than 5% or 10% of the ambiguity is due to observer factors, such as misunderstanding the utterance (see also Endnote viii). We can also expect that the amount of ambiguity due to both extrinsic and intrinsic factors decreases as the child grows older (see van Dijk & van Geert, submitted, see chapter 5 of this thesis). The point is not that one should try to find the exact amount of intrinsic ambiguity, because such an exact amount does probably not exist, but that disagreement between competent observers can be viewed as a reasonable and informative indicator of the intrinsic ambiguity of the observed variable, and that the meaning of an ambiguity index only makes sense in comparisons between ages or variables.

The second problem concerns the fact that the above calculations are based on average expected numbers of utterances in the ambiguous set. In practice, the actual number of items in that set will vary. This is not a problem per se, because a widely used index such as Cohen’s Kappa in fact corrects for accidental agreement on the basis of the mean of the distribution of statistically possible levels of accidental agreement. However, instead of using the average, it is more informative to look for an interval, for instance a 95% confidence interval. For any given number of utterances in the disagreement set (the sum of the non-overlap sets), an interval can be computed that covers the range between the smallest possible set of ambiguous items at the 2.5% level and the biggest possible set of ambiguous items at the 97.5% level.

Based on the claims we have formulated above, we have two further predictions. The first prediction concerns the use of fuzzy concepts in the study of inter-observer reliability. Instead of employing the two crisp categories ‘preposition-present’ and ‘no-preposition-present’, the observer might be able to apply the category ‘preposition-possible’ or ‘ambiguous-preposition’. Instead of classifying child utterances in one of these two extremes, the utterances can take an intermediate position. In the example, these utterances belong neither to the category preposition-present nor to the category preposition-not-present but belong to the category preposition-possible or ambiguous-preposition. We might expect that using such an intermediate category will improve the overlap percentages to some degree, since the ambiguity is acknowledged in the coding system. The advantage of such an approach is that results can be reported dependent on these degrees of certainty. For instance, a graph can be plotted with a single line that reports all prepositions (or any other linguistic category) in which the observers are certain and a line that incorporates all ‘possible’ instances. Comparing both lines expresses to what degree the results are sensitive to the uncertainty of the observers (for a generalization of this line of reasoning, see van Dijk and van Geert, submitted, chapter 5 of this thesis).
However, although it would certainly be informative to ask raters how sure they are about their rating, this procedure might also only replace the problem. Instead of having an assumption of two mutually exclusive (‘crisp’) categories, i.e., preposition and not-preposition, we now have the assumption of three mutually exclusive categories, namely preposition, ambiguous and not-preposition. If the categories are truly ambiguous or fuzzy, inter-observer (dis)agreement will now pertain to three categories instead of two. It is unsure whether this procedure results in higher reliability values (since the scoring procedure fits better with the ambiguous nature of the data) or if disagreement increases between observers with regards to the application of this third category. This question can be answered only in an empirical study in which these different scoring procedures are being compared.

The second prediction we make is that in the course of development inter-observer reliability increases since ambiguity decreases. The participants in the case study were observed around their second birthday. We expect a much higher agreement between observers for the observations that were made later in the trajectory. In addition to the improvement of the phonological system, older children have better defined and stable linguistic categories, which result in a decreased intrinsic ambiguity. This prediction can be tested in a developmental study on how the level of ambiguity changes over time.