

How much do children say in a day?*

KLAUS R. WAGNER

Universität Dortmund

(Received 30 April 1984)

As long ago as 1917, Schlag (1917: 218) pondered that it would be very valuable to record all the words said by a child in the course of one day. At that time, however, technology was not sufficiently developed for such a recording of a child's spontaneous speech to be feasible, and speculation was all that was available (Wagner 1974: 29-33). Not until the fifties and sixties did the emergence of tape recorders make the recording of a whole day of a child's verbal behaviour a real possibility.

The day as a unit of analysis

It is almost a rule that as technology advances, so new theoretical problems are posed. When it becomes possible to make a complete and continuous recording of the speech activity of a subject (S) then the problem arises as to what constitutes the LIMITS of communication situations in which and through which the subject moves. What factors act to constitute and alter a communication situation? A comparison of the basic components of various communication models (Wagner 1974: 103-32) provides some useful approaches for a solution to the problem; an examination of the various modes of segmentation applied to actual speech, however, also reveals the difficulties. What in one communication situation is represented as a clear break appears in another as vague and without clear divisions. As a solution to this dilemma the unit which suggests itself as an obvious unit of segmentation is the DAY AS THE PERIOD OF WAKING (*ibid.*: 34f.). Whatever doubts there may be about assumed divisions during the day, the day itself (between waking and sleeping) is surely undisputed and indisputable as the macro-unit of spontaneous speech.

The first day's recording

With these considerations in mind, I made the first day's recording of spontaneous child's speech on Monday 19 June 1972 (Wagner 1975). The subject, Teresa (9;7), wore a mini-microphone-transmitter (Telefunken Mikroport). The recording ran from 07.06 to 20.30 hours and covers a period

[*] Address for correspondence: Sprach- und Literaturwissenschaften, Journalistik und Geschichte, Universität Dortmund, Emil-Figge-Strasse 50, 4600 Dortmund 50, W. Germany.

of 13 hours and 24 minutes = 804 minutes. The transcription of the day's recording provided a corpus of 28,142 tokens and 3,825 types.

This result exceeded all expectations by a long way. Twenty-eight thousand words in one day was surely a quantity so vast as to require viewing with some reservations. Perhaps it was so much the result of individual factors as to be quite unrepresentative. That was in any case the way our thinking went at the time. It was only a second day's recording that gave us the encouragement to look again at the problem of how much a child says in a day.

The second day's recording

This was carried out by the team of Corzilius, Landskröner and Koort on Friday 20 January 1978 and analysed as described in Wagner (1974). The subject, Kai (9;6), was almost the same age as Teresa (9;7), and also the organization of the day from 07.24 to 22.00 hours (= 14 hours 36 minutes = 876 minutes) was very similar (cf. below). The transcription produced a corpus of 25,401 tokens.

What was remarkable about the second day's recording was that it reached the same area of magnitude as the first. Together they seemed to furnish reasonable expectations for the question under study. And as at that time in our preparation we were coming closer to formulating a sound hypothesis (a 10-year-old child speaks approximately 20,000-30,000 words per day), it seemed advisable as a next step to compare all the studies in the Dortmund corpus with regard to this question (Wagner 1978b: 280 ff.). It appeared likely that such a comparison would reveal within certain limitations (see further below) whether these two days' recordings had provided unique or representative results.

The Dortmund corpus

The Dortmund corpus comprises at the moment 13 part-corpora with subjects aged from 1;5 to 14;10. From this, 12 studies are drawn on here for comparison. (The Giljohann corpus, in which the language of a child aged 10;7 was recorded on six mornings at school, can unfortunately not yet be included here, as information is available only on the exact number of tokens (37,547) but not on the exact times.)

Recording and analysis procedure. In a pilot study (Wagner 1974, 1975), I described and demonstrated the methods for recording and evaluating. The subjects wear a mini-microphone-transmitter (this has always been a Telefunken-Mikroport) and are therefore not tied to a stationary microphone but free to move about as they wish. This is of immense importance for studies aiming to elicit and describe the spontaneous speech of children. Within a radius of 300 m around the recording apparatus the child can move freely, play, skip, climb trees, drive a go-cart, etc. The transcription system is that

used in the pilot study (Wagner (1974: 147-79) with certain improvements after Ehlich & Rehbein (1976) in the contextual framework):

- all the subjects' utterances verbatim, including paralinguistic;
- all interlocutor utterances in full, as far as they concern the subject, otherwise abbreviated;
- detailed information on the communicative setting (place, action, particular circumstances).

Microphone shyness vs. microphone stimulation. Both factors have their effects on the SPONTANEITY of the recording (Wagner 1974: 55-61). Whereas on the one hand microphone shyness inhibits the subject, on the other hand microphone stimulation (the Hawthorne effect) gives rise to increased speech activity, especially in children. On shorter recordings the influence of one (or both) of these factors can in certain circumstances be considerable; with long recordings (over hours), such as are used exclusively in the Dortmund corpus, both factors recede and cancel each other out.

Continuous sequence. In the studies which constitute the Dortmund corpus, great value is placed on there being an unbroken, continuous sequence, as segmentation according to types of communication situation is still a matter of considerable uncertainty and arbitrariness. The length of the recording to

TABLE 1. *Standardization of the Dortmund corpus (as of 1979) to a uniform day's length of 12 hours (720 minutes)*

Corpus No.	Age of subject	Researcher	Year	Length of sequence in minutes	Tokens	Standardization to 720 minutes
(1)	1;5	Schwarze	1975	202	3,881	13,800
(2)	1;8	Kadatz	1976	241	3,907	11,700
(3)	2;1	Wahner	1975	213	5,978	20,200
(4)	3;6	Hoffmann-Kirsch	1974	189	9,891	37,700
(5)	5;4	Brinkmann	1975	152	6,464	30,600
(6)	8;7	Häussermann	1975	193	6,630	24,700
(7)	9;2	Otto	1974	311	10,524	24,400
(8)	9;6	Corzilius/Landskröner/Koort	1979 1978 1978	869*	25,401	21,000
(9)	9;7	Wagner (1975)	1972	804*	28,142	25,200
(10)	11;8	Brunner	1974	188	9,703	37,200
(11)	12;2	Pagels/Gasse	1976	430	13,484	22,600
(12)	14;10	Vette	1974	254	8,062	22,900

* Two full-day recordings.

be analysed is limited as a rule at one end of the scale by the time limit imposed for producing undergraduate dissertations, and at the other end by the period of waking of the subject (a day). As the Dortmund corpus consists essentially of undergraduate dissertations, which for reasons of time could not involve analysis of an entire day, it seemed desirable to standardize all the part-corpora to a uniform day's recording period of 12 hours = 720 minutes. The results are presented in Table 1. The lines marked with the asterisk contain data for the two full-day recordings, and in these cases the values have not been increased but converted (reduced).

A comparison of these results (of the right-hand column of Table 1) reveals the following.

(a) The subjects aged less than two years speak considerably less than the other children (corpora 1 and 2).

(b) At 2;1, Andreas in corpus 3 already speaks 20,000 words. Here a comparison with another subject would be helpful. Ramge (1976: 17) mentions transcriptions of three days of his son Peter at ages 2;2, 2;8 and 3;2. Only the duration of each recording is indicated (8½ hours, 10 hours and 9½ hours); and the total of words spoken on these days has not yet been calculated (personal communication). However, it seems likely that the daily totals for those children who are still in the process of language acquisition will lie considerably below those whose language acquisition is completed. Corpus 3 with its 20,000 words appears here as an exception on the high side, just like corpus 4 for a child aged 3;6 with 37,700 words.

(c) Between ages 3 and 14, the figures reveal a relatively stable picture. With the exception of corpora 4 and 10, and possibly 5, a total of six recordings (Nos. 6, 7, 8, 9, 11 and 12) of children aged between 8;7 and 14;10 show figures of between 21,000 and 25,000 words per day. The reasons for corpora Nos. 4 and 10 with 37,000 tokens breaking this pattern can be illuminated by a comparison of the communication situations occurring in the various recording sequences.

Comparison of communication situations

The following section lists the communication situations in which subjects found themselves during recording, and which to a greater or lesser degree stimulated them to speak.

(1) *Schwarze corpus: Katrin* (1;5). Breakfast, playing (tap, milk lorry, dolls), helping to sort crockery, playing with bricks and dolls, nappy change, looking at guinea pigs, lunch, monologues in bed.

(2) *Kadatx corpus: Nicole* (1;8). Waking up, playing and jumping about in parents' bed, on her potty and getting dressed, breakfast and playing in her high-chair, at the kitchen window, painting, clearing the table, painting and

playing with a toy clock, playing with a big doll, playing a board game, on her potty, playing a board game, eating, getting undressed, monologues in bed.

(3) *Wahner corpus: Andreas (2;1)*. Eating a sandwich, playing (metal foil, animals, helicopter, toothbrushes, spinning top), playing with grandad and Caesar the rabbit (doll), playing with grandad and a Santa Claus doll, reciting a poem, looking at a picture book with brother and aunt (researcher), playing with a Lego tank, a candle, matches, drinking juice, playing football with a beachball.

(4) *Hoffmann-Kirsch corpus: Carsten (3;6)*. Playing (role-playing, driving a car, 'writing' = drawing), cutting up a birthday card, eating chocolate and looking at pictures with grandma, going into the cellar, playing at being a dog, going to the milkman, buying yogurt, eating yogurt, looking at and talking about pictures, having lunch, cuddling and talking to grandma, playing with cars (role-playing), cuddling and talking to his mother (researcher), crying (after being bumped), being comforted by grandma.

(5) *Brinkmann corpus: Gabi (5;4)*. Talking about her brother's birthday, breakfast, playing dominoes, eating Nutella (chocolate spread), playing dominoes again, drawing.

(6) *Häussermann corpus: Frederik (8;7)*. Waiting for the end of break; lessons: understanding things, mathematics, German, understanding things, German composition; break; lesson: braille; end of school, being driven home, arriving at home, collecting Andreas (playmate), playing with Andreas.

(7) *Otto corpus: Roman (9;2)*. Playing monopoly with Georg (younger brother), getting ready to 'go out', at the sports ground, relaxed conversation, playing with little cars, drive to the camp site, at the camp site, going home, going on with the game of monopoly with Georg, watching television (sports programme), playing with racing car set.

(8) *Corzillius/Landskröner/Koort corpus: Kai (9;6)*. Getting up, putting on the microphone transmitter, breakfast, going to school in the car, lessons (drawing, understanding things, mathematics (test), language, reading, singing) with breaks, going home, lunch, playing monopoly, driving a go-cart, playing in a Citroën 2CV, soldering, drawing, making a tassel, collecting food, watching television, memory game.

(9) *Wagner corpus: Teresa (9;7)*. Waking up, getting dressed, sewing on the microphone transmitter, breakfast, packing her bag, drive to school, before

lessons, lessons (arithmetic, language), 10 o'clock break, prizegiving, drive home, clearing things away, reading the mail, Teresa's file, picking gooseberries, playing with girl-friends (catching the cat, dressing up, clowns, ballet kidnappers), lunch, picking and cleaning gooseberries, homework, having coffee, clearing away toys, playing with Anke (coffee table, climbing a tree, playing on the grass, hopping on the patio, gold investigators, eating, ducat gold thieves), watching television, skipping, having dinner, watching television news, going to bed. (For a more detailed discussion of speech situations see Wagner (1974: 203-38).)

(10) *Brunner corpus: Markus* (11;4). Making a veteran car (toy car made by cutting out and pasting cardboard), using a microscope.

(11) *Pagels/Gasse corpus: Christiane* (12;2). Saying hello, making a crib, lunch, continuing work on the crib, playing skat, playing a word game, doing crochet, having coffee, singing Advent songs, reading aloud, conversation, watching television, drawing, having dinner, doing schoolwork.

(12) *Vette corpus: Axel* (14;10). Talking about cassette recorders, solving arithmetical problems, playing table tennis, having coffee, playing cards, recording music, playing table tennis.

A comparison of the speech situations in corpus 10 with the other corpora explains the high calculated total of words of 37,000: the language of this subject, Markus (11;4), was recorded in just two communication situations (making a veteran car with his friend Tomy and his father, using a microscope with his father) both of which were conversation-intensive. Other situations which give rise to less speech, such as watching television, playing football, etc. are not found in corpus 10. The reasons for the 37,000 words in corpus 4 are similar: here too the subject, Carsten (3;6), has in his grandmother, in whose house he is growing up, someone who is always there for him to talk to. These observations lead us to the question: how representative are the individual corpora?

Representativeness

This complex problem (cf. Schank 1973, Nikitopoulos 1974, Wagner 1974: 35 f., Bausch 1975) is discussed here under three aspects:

- (i) representativeness of the recorded SEQUENCE: does it capture a normal day? in what way is it unusual? what speech situations constitute the corpus? We are concerned here with the question of the representative filling of the period of time.
- (ii) representativeness of social LEVELS: how are the various levels of the

population represented? is it only middle-class children who speak so much or also working-class children?

(iii) representativeness of the individual IDIOLECTS: to what extent can we make inferences from 12 idiolectal corpora (the individual languages of 12 children) for the language use of various age groups?

Representativeness of the sequence. In compiling the Dortmund corpus, great emphasis was placed from the beginning on recording a continuous sequence and as far as possible a normal day's activity. For this reason all the corpora contain play situations – some even games of movement (corpus 3 [football], corpus 8 [driving a go-cart], corpus 9 [ballet, climbing a tree, hopping, skipping], corpus 12 [table tennis]). The longer the recorded sequences are, the more situations low in speech activity occur: school lessons (corpora 6, 8 and 9; also the Giljohann corpus not considered here), watching television (corpora 7, 8, 9 and 11). On the other hand, in many of the corpora the efforts of the researchers are probably not going to produce a corpus with low speech activity, e.g. corpus 10 (making a veteran car), corpus 8 (making a tassel and soldering), corpus 11 (making a crib). It could be objected that making things, even if it is something out of the ordinary daily routine introduced into the corpus sequence, is still a neutral situation with regard to speech. Depending on the subjects' own delight in speaking, they will speak more or even (much) less while modelling, depending on how involved they are in the activity. However that may be, it seems probable that the language production of subjects in the corpora is somewhat above the average.

Representativeness of social levels. As a basis for comparison, we provide information on the parents' professions.

Corpus 1: mother (researcher): qualified in child care, trainee teacher; father: parson, upper middle class.

Corpus 2: mother: saleswoman; father (researcher): trainee teacher, upper working class.

Corpus 3: mother: stopped working after the birth of her first child (subject's elder brother); father: supervisor of apprentices in an electrical workshop, is studying electrical engineering to become an electrical technician, upper working class or lower middle class.

Corpus 4: mother (researcher): trainee teacher; father: car salesman, middle class.

Corpus 5: mother: housewife; father: lawyer, upper middle class.

Corpus 6: mother (researcher): trainee teacher, entrance qualifications gained through further education, upper working class or lower middle class.

Corpus 7: mother: gymnastics teacher; father (researcher): 12 years in the armed forces (sergeant), trainee teacher, middle class.

Corpus 8: mother: landlady; father (researcher): draughtsman, invalid, died when the subject was 3 years old, upper working class.

Corpus 9: mother: teacher for eight years, then housewife; father (researcher): secondary school teacher in various school types, later university lecturer, middle class.

Corpus 10: mother (researcher): trainee teacher; father: certified engineer, architect, professor, upper middle class.

Corpus 11: mother: 10 years working in business, then housewife; father: skilled art metal worker, retrained as a teacher of art and vocational preparation at a school for mentally handicapped children, upper working class or lower middle class.

Corpus 12: mother: housewife; father: moulder in an iron-foundry, working class.

Of the 12 corpora, three (Nos. 2, 8, 12) are from working-class subjects; three further ones (Nos. 3, 6, 11) are located in the transition zone between working class and middle class; the six remaining studies (Nos. 1, 4, 5, 7, 9, 10) are from the middle class. Corpora with middle-class subjects are thus in the majority. In comparison with the working-class corpora, however, there is nothing to indicate that working-class children produce fewer tokens. (For a comparison of types, see below.) On the contrary, corpus 8 with 21,000 words and corpus 12 with almost 23,000 words lie well within the range of the middle-class corpora. In particular, a comparison of the two full-day recordings (corpus 8: subject 9; 6 working class; corpus 9: subject 9; 7 middle class) – units which have not been arbitrarily calculated – shows that the two subjects of the same age also produce approximately the same quantity of speech. There are however observable qualitative differences, as Hesse (1977) has discovered in a comparison of active use of adjectives.

Representativeness of idiolects. Representativeness is always a RELATIONAL matter. Nothing is representative in itself, but always with regard to some other entity which is elevated to become a fixed point of reference. This applies equally to the question of how representative idiolects (individual languages) are of sociolects (group languages; here age groups within language development).

It would be theoretically possible to determine the statistical representativeness of the idiolectal part-corpora which constitute the Dortmund corpus in a comparison with other empirical studies on child language. However, a calculation of this kind has not been undertaken, essentially for two reasons:

- (i) the usual magnitude of empirical corpora/samples is as a rule too small (see the discussion on 'The range of published speech corpora' (Wagner 1974: 132–46), 'Empirical studies on the vocabulary of school children' (*ibid.*: 272–80) and 'Empirical studies on child language' (*ibid.*: 68–87));
- (ii) the only corpus of a comparable magnitude (the Brunswick corpus) is

itself of low representativeness for spontaneous child language, on account of the asymmetrical situation in which recording took place (Wagner 1978b: 240; 1974: 86 f., 284-95; 1978a: 169 ff.). Here a comparative examination would be of use to discover how representative the Dortmund corpus is with regard to the Brunswick corpus, and inversely how representative the Brunswick corpus is in the light of the Dortmund corpus.

For these reasons, the question as to the representativeness of the idiolectal part-corpora of the Dortmund corpus must at this point be seen as having been raised, but not yet resolved.

Speaking time and speech situation time. No one speaks continuously. Speaking alternates with listening, and alongside this there is silence. A categorization of the various time periods could be as follows (cf. Wagner 1974: 237 ff.).

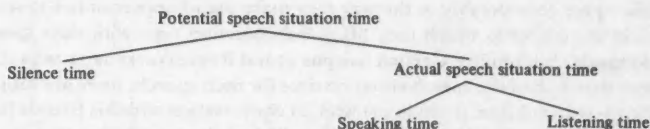
(i) *Potential speech situation time* is the duration of the complete sequence in which the subject could speak. This would not include any time at which a subject is prevented from speaking, e.g. by a plaster over the mouth.

(ii) *Silence time* is the time in which the subject could speak, but does not, and in which no other conversation takes place in which the subject participates as a listener. (On the distinction between silence time and external pauses, cf. Wagner 1974: 191 ff.)

(iii) *Actual speech situation time* when the silence time is subtracted from the potential speech situation time, what remains is the actual speech situation time, in which subjects are in situations in which speaking takes place.

(iv) *Speaking time* is the time in which the subject speaks.

(v) *Listening time* is arrived at by subtracting the speaking time from the actual speech situation time.



The speech time quotient is calculated by dividing the actual speech situation time by the speaking time. It provides a measure for establishing to what extent subjects make use of the actual speech situation for their speech, i.e. what proportion of the speech situation their speech accounts for. A value of 2 means that a subject takes up approximately one-half of the available actual speech situation time for his or her speech (value 4 = one quarter).

(vii) *Words per minute of speaking time* provides a measure for speed of delivery, for the rapidity with which a subject speaks.

Timings on the studies in the Dortmund corpus result in the values presented in Table 2. It is immediately apparent how rarely children remain silent when they have the opportunity to speak. With the exception of corpus

TABLE 2. *Timings of the studies in the Dortmund corpus (minutes)*

Corpus No.	Age of subject	Potential speech situation time	Silence time	Actual speech situation time	Speaking time	Speech time quotient	Words per minute of speaking time
(1)	1;5	202	—	202	84	2'42	46'2
(2)	1;8	241	—	241	?	?	?
(3)	2;1	213	—	213	81	2'63	74'3
(4)	3;6	189	19	170	93	1'83	106'4
(5)	5;4	152	3	148	48	3'08	134'4
(6)	8;7	193	3	190	65	2'92	102'0
(7)	9;2	311	157	154	85	1'81	123'8
(8)	9;6	876*	11	865	262	3'30	97'0
(9)	9;7	804*	60	744	191	4'11	155'5
(10)	11;4	188	—	188	58	3'24	167'3
(11)	12;2	430	?	?	90	?	149'8
(12)	14;10	254	13	241	43	5'61	187'5

* Two full-day recordings.

7, where the relatively long silence time is caused by driving to the camping site, periods of silence tend to be short or even non-existent. One possible cause of this could be the child's need to pass on information and to be in conversation. A further factor is however certainly the recording situation, in which the researcher making notes is always available to be spoken to and to be involved in conversation.

A comparison of speech time quotients reveals that the various subjects differ quite considerably in the way they make use of opportunities to speak, i.e. in the extent to which they fill speech situation time with their speech. Alongside children like Carsten (corpus 4) and Roman (corpus 7), who claim more than half of the speech situation time for their speech, there are subjects like 15-year-old Axel (corpus 12) who, in conversation with his friends (both sexes), manages to take up not even one-fifth of the speech situation time as speaking time. As if to make up for this by making the best use of his speaking time, Axel speaks the fastest of all subjects (187'5 words per minute of speaking time). After a period of practice and preparation at the beginning of language development, a speech rate of 100 words per minute is achieved by the age of just three-and-a-half (corpus 4). This increases in the course of schooling to 150-160 words per minute of speaking time.

These figures confirm that in a waking period of 12 hours children aged between 4 and 15 years take only 2-3 hours (a maximum of 4 hours) of speaking time to reel off some 20,000 words.

TABLE 3. *Lexical information in the*

Corpus No.	Age of subject	Compiler/s of lexicon	Year	General lexicon	Proper names	Pa
(1)	1;5	Boenigh-Freidank	1976	509	15	
(2)	1;8	Lexicon not yet produced	—	—	—	
(3)	2;1	Kunert	1979	596	36	
(4)	3;6	Gerdts-Pöller	1975	1,385	21	
(5)	5;4	Ligger	1977	770	27	
(6)	8;7	Freitag	1977	1,768	57	
(7)	9;2	Grosse-Ruyken	1974	1,654	86	
(8)	9;6	Lexicon not yet produced	—	—	—	
(9)	9;7	Wagner	1975	3,050	187	
(10)	11;4	Giesen-Hüls	1975	1,825	31	
(11)	12;2	Reichel (on Pagels)	1979	1,535	79	
(12)	14;10	Teuber	1974	1,289	29	

* Two full-day recordings.

n the Dortmund corpus (types)

Paralanguage	Foreign words	Surplus	Types (total)	Calculated types per day (720 minutes)
204	—	76	804	1,860
—	—	—	—	—
364	—	—	996	2,210
539	—	52	1,997	4,790
75	—	5	877	2,490
237	3	—	2,056	4,960
462	—	—	2,207	3,860
—	—	—	—	—
574	14	—	3,825*	3,520
194	—	27	2,077	5,020
203	5	—	1,822	3,580
168	57	—	1,543	3,040

HOW MUCH DO CHILDREN SAY IN A DAY?

Vocabulary

In a second analysis, a lexicon is compiled for each corpus listing words according to types and tokens. The type *ich* ('I'), for example, appears in corpus 9 with 1,331 tokens (cf. Wagner 1974: 140 ff., 309) – that is to say, Teresa (9;7) utters the word *ich* on this particular day 1,331 times. The lexica are arranged as REGISTERS OF OCCURRENCES. Each type is followed by a reference (corpus page and word number) to the positions in the corpus where the subject has used these tokens. This facilitates follow-up studies on the contexts in which particular words or word classes are used.

Under types we understand WORD FORMS and not BASIC FORMS. It is not the basic form *sprechen* ('speak') which is counted, but the word forms (for different subjects) *spreche*, *sprichst*, *spricht*, *sprechen*, *sprecht*. In addition, with *sprechen* a distinction is usually made between uses as infinitive, first person plural and third person plural. This measure also makes possible a more detailed analysis of child language use. Each lexicon for each corpus is divided into at least four sub-lexica:

(i) *general lexicon*: here all words are listed which also appear in other dictionaries (e.g. Duden, Wahrig);

(ii) *proper names*: here all personal names, place names and other names are included;

(iii) *paralanguage*: in this section all expressions are noted which occur in abundance in spontaneous speech alongside words from the general lexicon: 'öh', 'äh', 'ehm', 'hei', 'hahaha', etc. – a large and important area which up to the present has received very little attention in research; this section in the lexica is a start at sampling and collecting material;

(iv) *words with unclear meanings* this is a surplus section which covers everything which cannot be accommodated under the other categories.

This information is summarized in Table 3. In calculating the types for a day it has to be taken into account that the increase in types is exponential and not a linear function. In the course of extended speech, the rate for new types being introduced decreases steadily. (When the increase of new types reaches the value of 0, the subject's entire vocabulary has been recorded. This is to be developed with greater mathematical precision in later research.) The calculated values in the right-hand column are based on the increased quotients arrived at on a count of the first day's recording (corpus 9). Corpora 4 and 10 display again upper extreme values for types (as for tokens).

On average, children aged between 5 and 15 years appear to have a daily vocabulary of 2,500 to 3,500 different words (types). In assessing these figures, however, three points need to be considered.

(a) Along with words of a 'classic' vocabulary, paralinguistic expressions have also been counted. In the case of corpus 9, 574 paralinguistic expressions must be deducted from the total day's vocabulary of 3,825 types if the extent of vocabulary is to be made comparable with other studies.

(b) In the Dortmund corpus, word forms and not basic forms are counted (cf. above). This too needs to be taken into account in a comparison with values from other studies. A special count of corpus 9 reveals that 3,251 classic word-form types (3,825 minus 574 for paralinguistic) correspond to 2,064 basic form types (cf. Wagner 1974: 301).

(c) The distinction between ACTIVE/PRODUCTIVE and PASSIVE/RECEPTIVE VOCABULARY must be noted. All the studies in the Dortmund corpus record children's active vocabularies. Comparisons with other studies (Wagner 1974: 298 ff.) suggest that their passive vocabularies will be approximately twice as large as their active.

CONCLUSIONS

Taking into account the restrictions discussed, the following observations can be made.

Children aged between 5 and 15 years speak some 20,000 words of discourse per day (tokens). For this they take about 2-3 hours of pure speaking time.

They have an active vocabulary of some 3,000 word-form types (different words). This corresponds to an active basic form vocabulary of roughly 2,000 words.

If these values should be confirmed by other studies, this would have consequences for corpus collecting and sampling in the field of child language research. Compared with data from the part corpora constituting the Dortmund corpus, the scope of previous empirical studies in this area has been somewhat on the narrow side.

REFERENCES

- Bausch, K.-H. (1975). Zur Problematik der empirischen Basis in der Linguistik. *Zeitschrift germanistische Linguistik* 3. 123-48.
- Ehlich, K. & Rehbein, J. (1976). Halbinterpretative Arbeitstranskription (HIAT). *Linguistische Berichte* 45. 24-41.
- Hesse, G. (1977). Empirische Untersuchung zum adjektivischen Gebrauchswortschatz in drei Korpora spontaner Sprechersprache. Unpublished undergraduate dissertation, Dortmund.
- Nikitopoulos, P. (1974). Vorgriffe auf eine Thematisierung der Repräsentativität eines Corpus. *Deutsche Sprache* 1. 32-42.
- Ramge, H. (1976). *Spracherwerb und sprachliches Handeln*. Düsseldorf.
- Schank, G. (1973). Zur Korpusfrage in der Linguistik. *Deutsche Sprache* 4. 16-26.
- Schlag, J. (1917). Grundsätzliches zu einem Häufigkeitwörterbuche der Kindersprache. *Zeitschrift pädagogische Psychologie* 216-26.
- Wagner, K. R. (1974). *Die Sprechsprache des Kindes*. Teil 1: *Theorie und Analyse*. *Sprache und Lernen*, Bd. 37. Düsseldorf.
- (1975). *Die Sprechsprache des Kindes*. Teil 2: *Korpus und Lexikon*. *Sprache und Lernen*, Bd. 38. Düsseldorf.
- (1978a). *Sprechplanung Empirie, Theorie und Didaktik der Sprecherstrategien*. Frankfurt.
- (1978b). Der Erwerb von Sprecherstrategien im Kindesalter. In G. Augst (ed.), *Spracherwerb von 6 bis 16*. Düsseldorf.