Archives de Psychologie, 42, 1974, pp. 207-269.

THE FRENCH SYNTAX OF A CHILD'S NOUN PHRASES [1]

P. Suppes, R. Smith and M. Léveillé

Institute for Mathematical Studies in the Social Sciences, Stanford University

I. INTRODUCTION

This is the first of a series of articles concerned with the analysis of a young child's spoken French. In the body of the paper we give basic data on the corpus and how it was collected in Paris.

The main thrust of the work, however, is a detailed analysis of the corpus in the spirit of the work on probabilistic grammars and model-theoretic semantics begun several years ago in the Institute for Mathematical Studies in the Social Sciences at Stanford : GAMMON (1970), SUPPES (1970, 1971, 1972); and SMITH (1972).

We begin only the first stage of analysis, namely, the analysis of the grammar of noun phrases in Philippe's speech taken from the first three hours, the middle three hours and the last three hours of the corpus, which ranges from the time that Philippe was 25 months old to 39 months old. In this and subsequent articles we shall analyze the developmental aspects of Philippe's speech and shall investigate the grammar of the complete utterances, as well as the set-theoretical semantics of those utterances. However, to anticipate results reported in detail below, we did not find striking developmental trends in the grammatical structure of Philippe's noun phrases

The kind of probabilistic generative grammar applied to the corpus represents the application of a a certain set of concepts that we think are important in the detailed analysis of natural language. Some of our ideas are at variance with those commonly held by many linguists, especially the role we assign to probabilistic considerations. Consequently, even though these ideas have been set forth earlier in the publications mentioned above, it seems desirable to repeat some of our general arguments in support of the viewpoint we have adopted.

Many linguists deeply interested in generative grammars have very negative views about the kinds of frequency counts of vocabulary and

---

[1] Support for this research was provided by the United States National Science Foundation under grant NSFGJ-443X and the United States Office of Naval Research under contract N0014-67-0112-0049. This research was performed in collaboration with the Laboratoire de Psychologie Expérimentale et Comparée associé au Centre National de la Recherche Scientifique, Paris, France.

utterance length that characterize statistical linguistics. We want to urge that there is no fundamental opposition between a generative viewpoint of grammar and a probabilistic analysis of grammatical types. What we propose is, we believe, a useful marriage of the two viewpoints.

By introducing probabilistic parameters that govern the use of a given generative rule, we are able at once by standard statistical methods to introduce a goodness-of-fit criterion to discriminate between two different grammars for the same corpus. It is important that this criterion differs from the criterion used to determine which grammar accounts for a larger percentage of the utterances in the corpus. In the present paper, we consider two grammars. The second grammar accounts for a slightly higher percentage of the total utterances in the corpus; and, more importantly, its probabilistic fit is much better. It is also our feeling that the conceptual perusal of the two grammars will indicate why the second grammar is intellectually more satisfactory than the first. The requirement that the parameters attached to the generation rules match the actual frequencies with which these rules are used in derivations is to impose a criterion that requires the grammar to fit the corpus in a detailed way.

Those who have been concerned only about a competence grammar will not be deeply interested in our probabilistic results. We shall not examine in this paper the competence-performance issue, but simply say that we believe it is reasonable to think in terms of a performance grammar in dealing with a corpus of spoken speech and a probabilistic criterion is natural when dealing with a performance grammar.

We also investigate the possibility of introducing additional mechanisms for the generation of utterances, and we mirror these mechanisms in restrictions on the probabilistic parameters. In later sections we consider the mechanism that has the production rules for a given nonterminal symbol, for example the rewrite rules for determiners, be located in a push-down store. For this mechanism, a truncated geometric distribution governs the selection of which rule to use for a given generation. We recognize that this assumption is too simple, but it is interesting to see what decay in the goodness of fit results from applying this strong additional structural assumption.

Finally, we want to remark that in no sense do we consider the probabilistic account of the use of production rules an ultimate account. These probabilities themselves are overruled in particular cases by the semantics of a particular utterance. They represent the results of averaging over a number of utterances.

On the other hand it is not our belief that the introduction of semantics moves us from probabilistic to deterministic models. Probabilistic parameters are also needed at the level of semantics, for we are a very great distance from understanding the deterministic production of any nonstereotyped speech. In this article we shall consider the use of

the parameters in dealing with ambiguity — i.e., the situation where there are two distinct analyses of a given utterance. We believe that the detailed examples worked out in later sections of the paper show how the probabilistic apparatus can be used in a constructive way to help in disambiguation. In subsequent reports dealing with the Philippe corpus we shall consider appropriate probabilistic parameters and their application to semantic disambiguation.

II. COLLECTING THE CORPUS

Choice of the child. The subject for our study was chosen for two reasons. First, we wanted to ensure that the child chosen would be in close association only with native French speakers. Second, we needed to find parents who would be willing to submit to the rather demanding rule that all they said, as well as what their child said, would be recorded for one hour a week over an indefinite period of time. Thus Philippe, the only child of a couple who were in their thirties and members of a university community, was chosen as our subject.

When we visited him for the first time, Philippe was 25 months and 19 days old. He was a sociable little boy who was not shy, even with strangers. During the period of data collection he often went to the Faculty of Sciences with his father who taught there. He also visited his mother in a laboratory of psychology where she worked, and he occasionally participated in experiments in the laboratory. Usually he attended nursery school; when he did not stay there the whole day a lady in her forties stayed with him at his house. Both his mother and father talked a lot with him and provided him with a verbally and intellectually stimulating environment.

Conditions and frequency of recording sessions. During the first period, April 22 through June 24, 1971, the observer (M. LEVEILLE) visited Philippe in his house one hour a week. A group of 10 sessions was completed by June 24, 1971, before summer vacation began. After an interruption of nearly three months (83 days), when Philippe went to the country, the sessions continued at the same frequency through December 18, 1971. There was a lapse of 14 days between September 30, 1971 and October 14, 1971 due to a strike on the Métro, which paralyzed Paris. At that time 21 hours had been recorded. Then the visits became less frequent — one every fortnight. Again, between December 18 and January 6, 1972, and March 23 and May 6 a total of 63 days elapsed because Philippe was on vacation. Session 33 was the last one, because Philippe was leaving Paris for his summer vacation. The complete schedule of recording sessions is given in Table 1.

TABLE 1

Schedule of Recording Sessions

No.	Date	Age
1	April 22, 1971	25 months 19 days
2	April 29, 1971	25 26
3	May 6, 1971	26 3
4	May 13, 1971	26 10
5	May 20, 1971	26 17
6	May 29, 1971	26 26
7	June 3, 1971	27 0
8	June 10, 1971	27 7
9	June 17, 1971	27 14
10	June 24, 1971	27 21
11	September 16, 1971	30 13
12	September 23, 1971	30 20
13	September 30, 1971	30 27
14	October 14, 1971	31 11
15	October 21, 1971	31 18
16	October 28, 1971	31 25
17	November 4, 1971	32 1
18	November 11, 1971	32 8
19	November 18, 1971	32 15
20	November 25, 1971	32 22
21	December 2, 1971	32 29
22	December 18, 1971	33 15
23	January 6, 1972	34 3
24	January 20, 1972	34 17
25	February 3, 1972	35 0
26	February 10, 1972*	35 7
27	February 24, 1972	35 21
28	March 9, 1972	36 6
29	March 23, 1972	36 20
30	May 6, 1972	38 3
31	May 18, 1972	38 15
32	June 1, 1972	38 29
33	June 15, 1972	39 12

*According to the regular schedule, Session 26 should have taken place on February 17, but because Philippe was scheduled for vacation on this date, the session was recorded earlier.

For practical reasons, recording sessions always took place in the morning, generally not long after Philippe had awakened. Each session, with a few exceptions, lasted one hour. Although the recording periods were relaxed and informal, Philippe was asked not to leave the room in which the tape recorder (a UHER Recorder L 4000) was installed for any significant time. Usually the tape recorder was set up in the living room, which was at the center of the apartment. Only the microphone was moved to the kitchen during breakfast. If Philippe wanted to play in his bedroom, the tape recorder was taken there, and Philippe was asked not to go into the other rooms too often or too long.

There were few variations in the sessions, because they were held at about the same hour and under the same general circumstances. After the observer arrived, breakfast was served in the kitchen. Following breakfast, Philippe would play or look at books with the father or the mother present, or possibly both, especially in the early sessions. The observer took extensive notes on the behavior of the participants.

During the first session the mother, who had explained in advance the reason why the observer came, showed Philippe how the tape recorder worked. Philippe became accustomed rapidly to the routine and accepted it without difficulty. There was no selected topic of conversation in any of the sessions. The parents as well as the observer encouraged Philippe to speak, and consequently, asked him many questions about his school activities and what he had seen or done during the week. At times Philippe was not interested in the conversation between his parents (mostly during breakfast), but he intervened in the discussion whenever he wished.

Transcription and editing. The transcription of the recordings was made by the observer, always within a week following the session. The transcription includes utterances of both the child and the adults. In addition, comments on the perceptual and social situation based on the notes taken during the session have been included in the transcription. Additional comments by the observer that facilitate the understanding of what Philippe said have also been inserted.

The transcription was made without use of a phonetic system. The rules of written French were applied, except when a spelling different from the standard one served the purpose of revealing a clear deviation from the usual rules of spoken French. For example, when Philippe pointed to a horse drawn in a book and said *un chevau,* the *x,* which is the correct ending of *chevaux,* was omitted in the transcription in order to differentiate Philippe's special usage and to lead to easy identification of this usage in the transcribed text. As a second example, Philippe said *des monsieurs* instead of *des messieurs.* A tabulation of these nonstandard words is given in the next section.

In contrast, the words incorrectly pronounced by Philippe but clearly identified by the observer have been transcribed with the stan-

dard spelling. For instance, Philippe often said *déhors* instead of *dehors.* He also had difficulty in pronouncing correctly *cacher* and *casser,* but what he did was sufficiently clear to make his meaning understood.

The symbol &&& has been used to indicate unintelligibility of what was said by the child or by the adults present. This symbol may stand for a single word or for an entire utterance. Pauses in the conversation have been indicated by a new line, but in this as in other cases, the criterion for making the judgment was always somewhat subjective.

Because the transcription was made on a standard Model-33 teletype with ASCII characters, it was necessary to introduce certain character conventions for the transcription of French. The acute, grave, circumflex and diphthong accents were all replaced by an apostrophe. The apostrophe was replaced by the slash (/) and the cedilla by an asterisk. In the body of the present paper, however, a notation closer to the standard French accents is employed.

The coding of accents created certain minor unanticipated problems. For example, since the apostrophe was replaced by a slash, some words which were written with an apostrophe, for example, *aujourd'hui* and *quelqu'un,* appear in the vocabulary as two words. Thus, *hui* has been coded as an unclassifiable miscellaneous word in the vocabulary. In the case of *quelqu'un, quelqu* has been coded with its proper code and *un* as an article. Because the frequency of these examples is low, we did not revise the final analysis to take account of them.

In the final perusal of the dictionary, we noticed that three words were misspelled : *charriot, hibous,* and *eléphant.* The word *c** should be *c.* Moreover, the word *Oh,* which appears at the end of the vocabulary, is due to confusion between the numeral and the letter. Extensive computer reruns would have been required to correct these minor points, but these were not made, because no significant error is introduced by their retention.

As an additional remark, some words used in the corpus by the mother and father are provincial expressions and not standard French. They have been coded as standard words and will not appear as unusual words in subsequent analyses of the adult speech. This does not generate a problem for the tables of the present report because the vocabulary in this study is restricted entirely to Philippe's speech, and the vocabulary of the adults is not analyzed.

For further information, the distribution of the length of Philippe's noun phrases is given in Table 2. In Table 3 the first 100 words of highest frequency are presented.

TABLE 2

Noun-phrase length in noun-phrase corpus
(negative binomial test)

LENGTH X	FREQ X	THEOR.FREQ X	THEOR.PROB.	CHI2
1	3113	3246.90	.59	5.52
2	1747	1440.44	.26	65.24
3	368	522.32	.10	45.59
4	118	175.29	.03	18.73
5	87	56.46	.01	16.52
6	23	17.73	.00	1.57
7	5	5.47	.00	.04
8	3	1.67	.00	
9	2	.50	.00	
10	1	.15	.00	5.83

Residual Expected Frequency = .06
Chisquare sum = 159.10
Degrees of freedom = 5
parameters :
P = .72
R =1.58

TABLE 3

Rank ordering of the first 100 words of high frequency

2316	est	391	en	204	ce	123	eau
1641	le	374	tu	204	Madeleine	120	comment
1531	c'	366	fait	204	Maman	117	sont
1457	la	358	j'	199	tout	115	es
1370	&&&	348	qui	192	sais	114	cette
1351	oui	345	y	189	petite	114	mon
1164	je	334	quoi	178	mettre	111	maison
1137	non	333	ai	177	aussi	110	parce
1100	ça	326	oh	175	ils	110	train
1079	un	311	veux	174	s'	109	eh
1060	pas	308	petit	174	voilà	106	mais
964	a	307	sur	169	pourquoi	105	Philippe
871	là	297	du	166	comme	102	voir
860	de	294	avec	161	celui-là	101	ben
777	les	287	va	160	deux	98	au
771	une	277	Papa	159	plus	97	chez
653	des	272	faire	156	ah	97	se
639	elle	270	voiture	153	fais	96	bien
619	l'	262	vais	152	t'	95	chocolat
520	on	260	est-ce	148	autre	92	bateau
505	dans	260	où	148	faut	91	me
468	à	233	moi	143	monsieur	90	petits
452	et	233	que	141	camion	89	gros
445	pour	212	encore	141	dedans	89	mamie
444	qu'	209	regarde	137	si		

III. DICTIONARY I

The first step in writing a grammar and semantics of Philippe's utterances was to establish the grammatical categories for classification of terminal words. Basically, we followed classical morphological analysis and used the following eight classes of words : articles, nouns, pronouns, adjectives, verbs, adverbs, prepositions and conjunctions. Subclassifications within these broad classes are described below. Two additional classes were also used : onomatopoeias and interjections (KO), and uncodable expressions (MS), that is, words that were not understandable or sounds that could not be identified as clearly representing a word. Some words that are acceptable in French were coded as MS since, in the context of their utterance, they were not words, but rather only sounds that Philippe repeated without understanding. Examples include *cor* and *tonne.* The indistinguishable utterances classified as MS were not included in the analysis of the noun phrases. In Tables 4 and 5 we show the mispronounced and fragmented words, and occurrences of nonstandard usages in Philippe's speech.

A standard problem in the construction of dictionaries for psycholinguistic corpora is that a given word may function in different ways depending upon its context. Thus, *avoir* may function as a transitive verb, as in *J'ai de l'argent,* or as an auxiliary, as in *J'ai fait mon travail.* Words of this kind have been assigned to two or more categories, and are signaled by commas in the coding. Transitive verbs are symbolized VT, and auxiliary verbs VA, so forms of *avoir* are classified VT, VA. When this occurs, we shall say that a word has been *multiply classified.*

We have not included either all or only those multiple classifications that are in all likelihood present in Philippe's speech, but have tended to be somewhat conservative. When it is highly unlikely that Philippe would have used a certain word in more than one context, we did not multiply classify that word. As an example, the word *mets* has been coded as a verb but not as a noun, since it was our belief that the latter usage ("something related with food") was unlikely to occur.

The subclassifications of the main classes were modified from classical French grammar as represented, for example, in Grévisse (1969), although we did deviate from Grevisse in a number of respects. In the first place, his and other classical grammars involve written language rather than spoken language. Especially in the case of verbs, the inflectional patterns have been extensively simplified to correspond to spoken rather than written inflections. After our analysis had been completed, we became aware of the work by MARTINET (1958) and found that our analysis agrees to a large extent with his. The present article deals only with noun phrases; consequently we shall not expand on the coding of verbs, the principle of which is given in SUPPES, SMITH, LEVEILLE (1972). Also, we have drawn upon the work of DUBOIS (1965) and

TABLE 4

Mispronounced and Fragmented Words

Word	Interpretation	Dictionary Code
baguitte	baguette	NC2,0
banquette	"	"
beguitte	"	"
bétoleu	bétonneuse*	"
bétoleuse	"	"
blanquette	banquette	"
bouette	brouette	"
bousse	boule	"
cace	cache	VT1
chigne	cygne	NC1,0
codile	crocodile	"
drabon	dragon	"
édicaments	médicaments	"
égliaue	église	NC2,0
estatues	statues	"
fabouilles	farfouilles	VT1
festation	manifestation	NC2,0
kola	koala	NC1,0
magnéophone	magnétophone	NC1,0
mama	maman	NP2
mamaman	"	"
masou	maïs	NC1,0
migre	émigre	VI1
modeur	moteur	NC1,0
plas	pas	DV6
positoire	suppositoire	NC1,0
retourer	retourner	VN0
samedié	samedi	NC1,0
so	soleil	NC1,0
toffe	étoffe	NC2,0
verture	couverture	NC2,0
xagone	hexagone	NC1,0

*bétonneuse does not appear in "Le Petit Robert", but Philippe's parents used it instead of bétonnière.

DUBOIS-CHARLIER (1970), in particular Dubois' encoding of the definite and indefinite articles and the partitive constructions.

Because of the differences between written and spoken French, and because Grévisse does not provide a completely systematic classification, it was necessary to introduce a number of distinctions not found in his analysis. Thus, a systematic coding was initiated which distinguished whether the gender or number of an adjective or noun is undetermined, or whether the gender or number (singular or plural) is masculine or feminine. This is not to imply that Grévisse ignored the fact that adjectives may have the same form regardless of whether they are masculine or feminine, but only that we felt it necessary to introduce clear coding conventions to describe these cases.

The coding introduced for each of the classes described above is exhibited in Table 6. We turn now to detailed remarks about the nature of this classification and the reasons for it. To a large extent we have retained the standard French grammatical terminology in order to facilitate comparison with various published grammars of French.

TABLE 5

Occurrences of Nonstandard Usage

Words	Code	Session	Typical Context
allez	VD15,VI15	17	Je suis pas content que tu t'en allez
aye	VT15,VA15	30	Il faut ... un soleil chez moi pour qui aye plus de vent
ayent	VT15,VA15	27	Pour qu'ils ayent chaud
batter	VT0		(said for battre)
cheval	NC1.2	10	... et les deux cheval
cheveu	NC1.1	6	Une cheveu elle court
conducteur	NC1.0	17	Il a oublié le feu rouge le conducteur
éclant	NC1.0	2	(said for aimant)
écrise	VT5	30	Il faut que je t'écrise quelque chose
loine	DV2,LC0	26	Dans une maison très loine
monsieurs	NC1.2	12	C'est pas bien de tuer les monsieurs
morde	AQ2.0.2	33	J'ai jeté l'allumette qui était morde
mordes	AQ2.0.2	28	Les feuilles mordes
nallumettes	NC2.0	14	J'ai fait une grande maison avec des nallumettes
nanimals	NC1.1	20	Des nanimals
narbre	NC1.0	19	Il est noir ce narbre

Table 5 (following and end):

```
nautres  AQ0,0,1,PIB,AI    3      Des nautres reils
navion   NC1,0             9      Voilà le petit navion
nen      PE19,DV2,EP       27     C'est les enfants qui nen mangent
nenfant  NC0,0             21     C'est pas des petites têtes de nenfant
noiseau*
         NC1,0             3      (said for oiseau)
                                    ------
noure    NC1,0             1      Tombé le noure
oeils    NC1,2             17     C'est ça les oeils?
prendait
         VN6               26     Elle prendait toutes les sous
prende   VN5               26     Il était pas content qu'on lui prende
                                  toutes ses sous
rete     VT1               28     On rete les feuilles mordes
séqué    AQ1,0,2           19     Tu veux un petit raisin séqué?
séqués   AQ1,0,2           19     Tu veux des raisins séqués?
soye     VC15,VA15         31     Pour que la maison soye bien grande
soyent   VC15,VA15         16     ... pour elles soyent plus belles
tallé    VD0,VI0           31     ... je suis tallé a l'école
tuser*   VT0               12     (said for user)
                                    ----
zeiles   NC2,0             15     Il a pas de zeiles
zarrivent
         VT1               29     Les voilà qui zarrivent
zécoutent
         VT1               24     ... et qui zécoutent pas
zencore  DV0               26     Y en a zencore?
zoiseau  NC1,0             3      Un zoiseau sur le tonneau
zont     VT5,VA5           26     ... c'est les chats qui zont
zoutils  NC1,0             12     C'est un carré de zoutils
         * Words associated with either an action, or an object.
The interpretation is based on the notes of the observer.
```

Articles. Under this general category we have assigned the indefinite articles (*un, une*) the category IN, and the definite articles *le, la, l', les* the category DN. Even though *un* and *une* can be considered as cardinal adjectives, in the present corpus it seemed appropriate to classify them unambigously as articles.

The words *du, de, des, d'* have been coded as prepositions in the classification EP, which also includes *au* and *aux*. This follows a suggestion made in DUBOIS (1965, p. 152).

Nouns. Two main subcategories of nouns have been distinguished as in traditional grammars, namely, common and proper nouns. As already indicated, common nouns have been grammatically coded in such a way as to take into consideration the gender and the number. For example, *bouteille* and *bouteilles* have the same spoken form and both have been coded NC2.0, with the first digit indicating the gender (in this case feminine) and the second digit indicating the number (in this case undetermined). In contrast, *cheval* was coded NC1.1 and *chevaux* was coded NC1.2. Homonyms were coded 0 for both the number and the gender; for example, *moule* was coded NC0.0.

In addition, nouns followed by a hyphen and *là*, which indicates that the noun had the composed form of the demonstrative adjective as determinant, were coded with the symbol # indicating that two words were conjoined. For example. the word *aiguille-là* was coded NC2.0#AD. Thus NC2.0 is the appropriate coding for *aiguille* and AD the appropriate coding for *là*.

Proper nouns have been coded to reflect gender only, for example, *Jeanine* was coded NP2, *Papa* NP1, and *Paris* NP0.

Pronouns. Five subcategories have been distinguished : personal (PE), demonstrative (PD), interrogative and relative (PR), indefinite (PI), and possessive (PO). Personal (PE) and demonstrative pronouns (PD) have been given specific numbers, as shown in Table 6.

Interrogative and relative pronouns have been coded together because they are not distinguishable. An exception is *dont,* which is always a relative. Some indefinite pronouns (PI), which could be identified as such unequivocally, have been assigned a specific number from 1 to 7. The other forms which could also be classified as indefinite adjectives have been coded PI8. Because the possessive pronouns are formed with the definite article and the accentuated (tonique) form of the possessive adjectives, they do not appear as separate entries in the vocabulary. On the other hand, since the accentuated forms are rarely used as possessive adjectives, they have been categorized as possessive pronouns in Table 6.

Adjectives. Six main subcategories of adjectives have been distinguished : qualitative (AQ), numerical (AC for cardinal and AN for ordinal), demonstrative (AD), interrogative (AT), possessive (AO), and indefinite adjectives (AI). Relative adjectives have not been coded, since their usage is extremely rare even in written language.

Qualitative adjectives have been coded according to the gender, the number and the position (1 for anteposition, 2 for postposition, and 0 for undetermined). For example, *verte* has been coded AQ2.0.2, with the first digit indicating the gender, the second the number and the third the position. In the case of gender and number, the same three-fold classification was used as described above, namely, 0 for indeterminate, 1 for masculine, and 2 for feminine, with corresponding conventions for number.

Cardinal adjectives (such as *deux*) have been coded AC. As previously mentioned, the indefinite articles *un* and *une* could be classified as cardinal adjectives, but we classified them IN instead. Ordinal adjectives have been coded according to the gender and number; for instance, *second* has been coded AN1.0. Demonstrative, interrogative, and possessive adjectives have been coded with specific numbers, as shown in Table 6.

TABLE 6

Coding of Grammatical Categories, Dictionary I

ARTICLES

Indéfini	Défini
IN 1 un	DN 1 le, l'
IN 2 une	DN 2 la, l'
	DN 3 les

PRONOMS

Personnel

PE 1 je, j'
PE 2 me, m'
PE 3 moi
PE 4 nous
PE 5 tu, t'
PE 6 te, t'
PE 7 toi
PE 8 vous
PE 9 il, ils
PE 10 elle, elles
PE 11 lui
PE 12 eux
PE 13 se, s'
PE 14 soi
PE 15 le, l'
PE 16 la, l'
PE 17 les
PE 18 leur
PE 19 en
PE 20 y

Demonstratif

PD 1 celui
PD 2 celle, celles
PD 3 ceux
PD 4 celui-ci
PD 5 celui-là
PD 6 celle-ci, celles-ci
PD 7 celle-là, celles-là
PD 8 ceux-ci
PD 9 ceux-là
PD 10 ceci
PD 11 cela
PD 12 ça
PD 13 ç'
PD 14 ce

Interrogatif ou relatif

PR 1 qui
PR 2 que
PR 3 quoi
PR 4 qu'
PR 5 lequel
PR 6 laquelle
PR 7 auquel, auxquels, auxquelles
PR 8 duquel
PR 9 lesquels, lesquelles
PR 10 desquels, desquelles
PR 11 où
PR 12 dont (relatif)

Indéfini

PI 1 on
PI 2 quelqu'un
PI 3 quelques-uns
PI 4 quelques-unes
PI 5 personne
PI 6 rien
PI 7 chacun
PI 8 autre, etc.

Possessif

PO 1 (le, les) mien(s)
PO 2 (le, les) tien(s)
PO 3 (le, les) sien(s)
PO 4 (le, les) nôtre(s)
PO 5 (le, les) vôtre(s)
PO 6 (la, les) mienne(s)
PO 7 (la, les) tienne(s)
PO 8 (la, les) sienne(s)
PO 9 (le,la,les) leur(s)

Table 6 (following and end):

ADJECTIFS

AQ : qualificatif
AC : numéral cardinal
AN : numéral ordinal

Possessif

AO 1 mon
AO 2 ton
AO 3 son
AO 4 ma
AO 5 ta
AO 6 sa
AO 7 mes
AO 8 tes
AO 9 ses
AO 10 nos
AO 11 vos
AO 12 notre
AO 13 votre
AO 14 leur(s)

Démonstratif

AD 1 ce
AD 2 cet
AD 3 cette
AD 4 ces

Interrogatif

AT 1 quel(s)
AT 2 quelle(s)

Indéfini

AI autre(s),etc.

ADVERBES

DV 0 indéterminé
DV 1 temps
DV 2 lieu
DV 3 quantité
DV 4 manière
DV 5 affirmation
DV 6 négation
DV 7 interrogation

PREPOSITIONS

EP 0 indéterminé
EP 1 de, d'
EP 2 du, d'
EP 3 des, d'
EP 4 à
EP 5 au, aux
EP 6 de, d' (avec verbe)
EP 7 à (avec verbe)

CONJONCTIONS

CS : subordination
CC : coordination

LOCUTIONS

LC 0 indéterminé
LC 1 adverbiale
LC 2 prépositive
LC 3 interrogative

NOMS

NC : nom commun
NP : nom propre

KO : interjections et onomatopées

KS : sans signification

IV. NOUN-PHRASE GRAMMAR I

Our study of the noun phrases of Philippe's speech is based on a selection of noun phrases made from the sentences in 9 of the 33 sessions of the complete corpus. For this selection, we used Sessions 1, 2, 3, 14, 15, 16, 31, 32 and 33, which cover the beginning, the middle and the end of the corpus. In other words, each of these sets of three sessions forms a *section* of the corpus. This choice was motivated primarily by a desire to look for the changes in structure over the maximum period of time.

The noun phrases were selected from the sentences in the corpus on the basis of the intuitive judgment of M. Léveillé, who is a native French speaker. The selection was not the product of any explicit formal analysis. Nevertheless certain guidelines were followed and they are listed below.

1. Noun phrases in questions of the forms *est-ce que ?* and *est-ce qu'il ?* were not included, although other kinds of interrogatives were scanned for noun phrases.
2. Personal pronouns combined with verbs to form pronominal verbs were not included.
3. Cases of using the preposition *à,* where *de* should normally have been used (e.g., as in *je veux la montre à maman*), as well as phrases such as *sa maison à lui,* were included.
4. When the preposition *en* was combined with a substantive, such as in *la chaise en métal,* the whole expression *en métal* was excluded. On the other hand, when *en* could be replaced by *sur le* or *dans le,* as in *en vélo* or *en voiture,* it was included.
5. The preposition *pour* was analyzed in such sentences as *un bassin pour les bateaux.*
6. When adverbs such as *beaucoup* or *assez* modified a substantive, as in *beaucoup de chocolat, assez de sous,* they were included in the noun phrases.
7. When the word *là* followed a substantive, such as in *appuie sur le bouton là,* the word was considered a part of the noun phrase (*là* being analyzed as KO by grammar I).
8. When the word could serve as either a noun or another word, such as *téléphone,* the word was not included as a noun phrase if the context of usage was vague.

We expect the class of noun phrases to change somewhat in a complete grammar, and the rules of the two grammars discussed below to change in a full grammar. For example, one motivation for change will be the attempt to study differences between noun phrases that occur as

subjects in some sentences and as direct objects in others. As for the methodological justification for selecting the noun phrases by intuitive judgment, we should mention that this was only a first step in obtaining a complete grammar. The fact that noun phrases dominated Philippe's speech, as they dominate other psycholinguistic corpora of young children, indicates that the noun phrase is a reasonable place to begin.

It may be useful to expand on these methodological remarks about our method of selecting the noun phrases by individual inspection of the utterances. In principle, one would like to put the entire grammar and semantics together into one systematic whole. We think this is a worthy objective and hope to pursue it further in subsequent publications. On the other hand, given the massive character of a corpus of the size of this one, we consider it important to begin the analysis in piecemeal stages, and to do this we must make some intuitive decisions without systematic guidance from an overall model of a complete grammar or even less a complete semantics. We are certain that there could be marginal disagreements about our selection of noun phrases but we doubt that it would affect the main results of our analysis. We emphasize also that the noun phrases were selected prior to the attempt to write a grammar for them, and consequently we have in no sense tailored the selection of noun phrases to the convenience of our two grammars.

As already indicated, there are two separate and parallel analyses of the noun phrases in the corpus. In this section we describe the first grammar, the one that uses the first dictionary of grammatical categories. This grammar represents an extensive modification of those used in SUPPES (1970, 1971) and SMITH (1972) for analyzing noun phrases in English. Grammar I is of course not a grammar of English, but rather it was constructed from the modifications that would be required in the grammars for English in order to obtain an appropriate grammar for French. Grammar II, on the other hand, was written by a more direct consideration of French grammars. Detailed remarks about the goodness of fit are reserved for later discussion, after both grammars have been described and the results of the analysis presented.

Nonterminal vocabulary of Grammar I. In addition to the categories introduced in Dictionary I, the following additional nonterminal symbols were used in Grammar I. These additional symbols cannot be rewritten by lexical rules, but occur in derivation trees always at least two nodes from a terminal node. As the label of the root of all the noun-phrase trees, we adopted SN for the French *syntagme nominal.* Three additional noun-phrase notations for which we used a natural English notation, namely, NP*, NP′ and NP″, were also introduced. Note that according to Dictionary I, NP without an additional suffix was the category of proper names.

In the production rules of the grammar as shown in Table 7, we omitted all the numerical suffixes showing number, case, gender, etc.

These suffixes were included in the dictionaries, but were ignored by the grammar discussed here. In a more refined analysis it would be necessary to add these subscripts. However, it is obvious that the refinement required to do this would multiply the number of rules.While a relatively abstract level of analysis has been chosen, we feel the level selected exemplifies the structural features of Philippe's speech at about the right level of conceptual detail, insofar as noun phrases are concerned.

The notations ADJP and POSTADJ were introduced for adjectival phrases in pre-position and for adjectival phrases following the noun. The nonterminal symbol DET is the general nonterminal symbol used for determiners. Rules (7,1) to (7,6) are rewrite rules for this nonterminal symbol. The nonterminal symbol NUM designates numerical or cardinal number words (Rules (8,1) and (8,2)).

We have already mentioned that we introduced three noun-phrase symbols in addition to SN, namely, NP′, NP″ and NP*. The function of NP′ and NP″ was to provide a rewrite rule for one-word noun phrases (Rules (3,1) to (3,5) and (2,1) to (2,3)), respectively. While NP′ generates noun phrases with or without determiners, NP″ generates one word noun phrases without articles. Rewrite rules (4,1) to (4,17), that have NP* on the left, were used to construct noun phrases of a more complicated character. From a purely grammatical standpoint these nonterminal symbols were needed in order to block improper recursions. Further, they were convenient in getting a better fit from a probabilistic standpoint.

The production rules. The 46 production (or rewrite) rules of Grammar I are shown in Table 7. Perhaps the most surprising thing about the rules is their standard character. At the level of syntax, Philippe's speech is very close to standard spoken French. Not one of the 46 rules, with perhaps the exception of rule (3,2), would be regarded as inappropriate for adult spoken French, but additional rules beyond those given here would be needed in a complete grammar.

The use of the 46 production rules to produce an infinite variety of noun phrases is standard and in accordance with classical context-free grammars. An illustration of the use of the rules to produce one of the longer noun phrases in Philippe's speech is shown in the tree of Figure 1 for *IN AQ NC EP DN AQ NC* : *un petit tambour pour les petits enfants.*

There are some additional conceptual remarks that we want to make about the rules. The first group of rules constitutes the nine different rules that can be used to generate trees starting from the labeled root SN. The first three rules are standard. Rule (1,4), SN → EP NP*, a rule which permits a noun phrase to begin with a preposition, is natural not only for Philippe's speech, but also for standard spoken French. It should be recalled here that *du, de, d', des, à, au,* were all coded EP.

TABLE 7

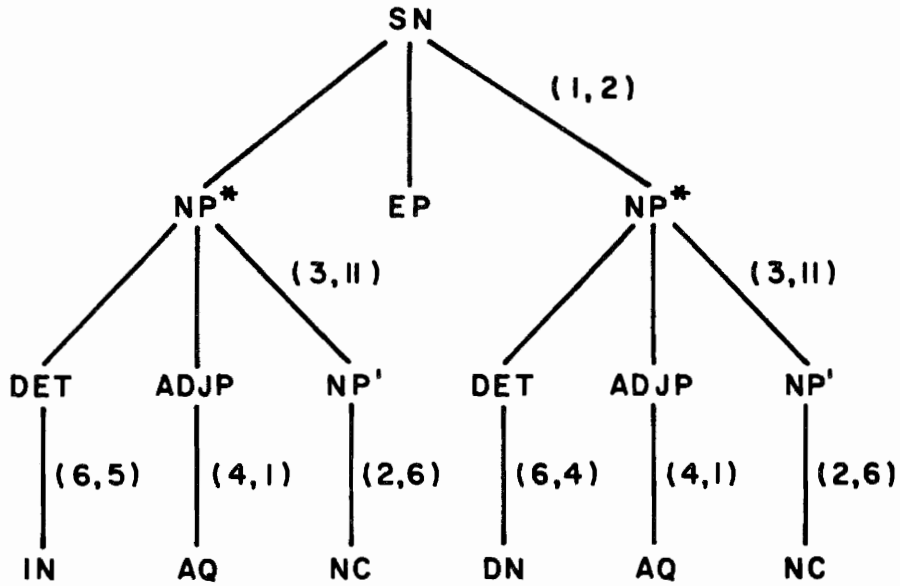
Parameters for Grammar I*
Philippe's Noun Phrases

Label	Prob.	Rule
(1,1)	.0060	SN => NP* CC NP*
(1,2)	.0250	SN => NP* EP NP*
(1,3)	.8925	SN => NP*
(1,4)	.0672	SN => EP NP*
(1,5)	.0031	SN => CC NP*
(1,6)	.0000	SN => KO NP*
(1,7)	.0052	SN => NP* KO
(1,8)	.0000	SN => LC NP*
(1,9)	.0010	SN => NP* LC
(2,1)	.5857	NP" => PE
(2,2)	.3494	NP" => PD
(2,3)	.0648	NP" => PI
(3,1)	.1388	NP' => NP
(3,2)	.7838	NP' => NC
(3,3)	.0101	NP' => NC AD
(3,4)	.0621	NP' => PR
(3,5)	.0052	NP' => PO
(4,1)	.0000	NP* => DET NUM ADJP NP' POSTADJP
(4,2)	.0000	NP* => NUM ADJP NP' POSTADJP
(4,3)	.0007	NP* => DET ADJP NP' POSTADJP
(4,4)	.0002	NP* => ADJP NP' POSTADJP
(4,5)	.0000	NP* => DET NUM NP' POSTADJP
(4,6)	.0002	NP* => NUM NP' POSTADJP
(4,7)	.0065	NP* => DET NP' POSTADJP
(4,8)	.0036	NP* => NP' POSTADJP
(4,9)	.0004	NP* => DET NUM ADJP NP'
(4,10)	.0006	NP* => NUM ADJP NP'
(4,11)	.0286	NP* => DET ADJP NP'
(4,12)	.0120	NP* => ADJP NP'
(4,13)	.0006	NP* => DET NUM NP'
(4,14)	.0037	NP* => NUM NP'
(4,15)	.2836	NP* => DET NP'
(4,16)	.1953	NP* => NP'
(4,17)	.4641	NP* => NP"
(5,1)	.9701	ADJP => AQ
(5,2)	.0299	ADJP => ADJP AQ
(6,1)	.9836	POSTADJP => AQ
(6,2)	.0164	POSTADJP => POSTADJP AQ
(7,1)	.0496	DET => AD
(7,2)	.0070	DET => AT
(7,3)	.0746	DET => AO
(7,4)	.6117	DET => DN
(7,5)	.2472	DET => IN
(7,6)	.0099	DET => AI
(8,1)	1.0000	NUM => AC
(8,2)	.0000	NUM => AN

*Correction for continuity = .5; expected < 5 are grouped together.

Figure I

TREE FOR IN AQ NC EP DN AQ NC



The instances of the use of this rule in accounting for noun phrases in the corpus are the following :

FREQ	TYPE	EXAMPLE
245	EP NC	du chocolat
60	EP DN NC	de l'eau
11	EP AQ NC	des gros animaux
11	EP NP	de maman
6	EP NC AQ	au bras droit
3	EP AI NC	de chaque côté
3	EP AT NC	de quelle couleur
2	EP IN NC	d'un cheval
2	EP PI	d'autre
1	EP AO NC	de ses mains
1	EP AQ AQ NC	des vrais petits chiens
1	EP AQ NC AQ	des petites vitres rondes
1	EP DN AQ NC	de la petite musique
1	EP PR	des quoi ?

TYPES = 14 TOKENS = 348

Rule (1,5), SN → CC NP*, introduces conjunctions at the beginning of noun phrases. This rule, like rule (1,4), is natural for adults as well as for Philippe. The instances of Rule (1,5) in the corpus are the following :

FREQ	TYPE	EXAMPLE
11	CC PD	et ça
4	CC PE	et il
1	CC DN NC	ou le tramway

Rules (1,6) and (1,7) introduce the grammatical category KO (interjections) before and after the main noun phrases, respectively. Rule (1,6) was not required for our noun-phrase fragment. The instances of Rule (1,7) are the following :

FREQ	TYPE	EXAMPLE
10	DN NC KO	le bouton là
8	PD KO	ça là
4	AD NC KO	cet accident là
3	AO NC KO	ton manteau là
1	AD AQ NC KO	ces petits bleus là (*)
1	DN PO KO	le mien là

(*) *bleu* was coded NC, AQ. The grammar did not generate AD AQ AQ, and consequently AD AQ NC was recognized.

In a similar way, Rules (1,8) and (1,9) introduce the grammatical category LC (locutions) before and after noun phrases. Rule (1,8) was not required for the noun-phrase fragment. The only instance of (1,9) was the phrase PE LC (*le voilà*), which occurred once in the corpus.

As already remarked, the rules of the second and the third group were used to generate one-word noun phrases. These noun phrases were generated by using Rules (1,3) and (4,16) and (4,17) and then rewriting NP′ and NP″ according to one of the eight rules in the second or third group.

The 17 rules of the fourth group contain the main structure of the noun phrases. They determine the way in which determiners, numerical expressions, preadjectival phrases, and postadjectival phrases can be generated. Although all 17 rules are perhaps plausible to a native French speaker, three of the rules from the fourth group were not required for the corpus. The unused rules from the fourth group were (4,1), (4,2), and (4,5). An instance of Rule (4,3) is [1]

[1] Here and subsequently we often show the frequency of a given grammatical type of noun phrase in the corpus without expressly labeling the frequency column. Thus the frequency of DN AQ NC AQ is 2 in the noun phrases *under examination*.

2 DN AQ NC AQ : *le petit filet jaune*

and an instance of Rule (4,9) is

2 DN AC AQ NC : *les deux petits pistolets*

The fifth and sixth groups of rules — (5,1), (5,2), (6,1), (6,2) — provide the standard recursions for pre- and post adjectival phrases. Except for accounting for both pre- and post positions, they are the same rules used previously for English.

The main reason for using different rules for pre- and post position adjectival phrases was to test the tendency to use a different number of adjectives in the pre-position from the postposition. After we examine the fit for the grammar in more detail, we will remark on the outcome of this test.

The seventh group of rules is for determiners and permitted rewriting the nonterminal symbol DET as any one of six nonterminal symbols standing for a grammatical category. Note that determiners were not introduced as a separate grammatical category, but were spread among the grammatical categories of Dictionary I, which is a relatively classical structure. In a sense, this was a way of having the best of both worlds. We introduced the concept of a determiner in the grammar, but kept in the dictionary of grammatical categories the more or less classical divisions.

Finally, the eighth group of rules consists of the two simple rules for rewriting the nonterminal symbol NUM for numerical expressions, which may be replaced with the grammatical categories AC and AN. Rule (8,2), which introduces AN, was not used in the corpus.

Nonprobabilistic fit of Grammar I. Probably the first empirical question to ask about a grammar for a corpus is, what percentage of the tokens and what percentage of the grammatical types does it parse ? The data are summarized in Table 8. There were 5,467 noun phrases in the nine hours analyzed. Grammar I parsed 5,191 of these noun phrases, which represents 95 percent of the tokens. There were 339 types of noun phrases in the nine hours, and the grammar parsed 222 of these, or 65.5 percent.

The entries in Table 8 are defined as follows, with the usual distinction between types and tokens. We first give the total number of types and tokens in the corpus, then the number recognized, and finally the percent recognized. The fourth line indicates how many of the types or tokens in the corpus are lexically ambiguous, that is, have at least one word assigned to more than one grammatical category. It is important to realize that lexical ambiguity, as we have defined it, results from using words in ways that are grammatically distinct; in most cases, lexical ambiguity would offer no difficulty to a speaker. It is, we believe, a test of the grammar to ask how well it is able to eliminate this superficial ambiguity.

TABLE 6

Nonprobabilistic Fit of Grammar I Noun-phrase Corpus

Description	Types	Tokens
Total	339	5467
Recognized by Grammar I	222	5191
Percent recognized	65.5%	95%
Lexically ambiguous	234	1747
Percent lexically ambiguous	69%	32%
Ambiguous recognized	161	1615
Percent of the ambiguous recognized	68.8%	92.4%
Ambiguity resolved	142	1553
Ambiguity reduced	14	35
Original ambiguity factor =	2685	
Parsed ambiguity factor =	2297	
Reduced ambiguity factor =	64	
Types after consolidation =	140	
Types after probabilistic disambiguation	124	

We then indicate how many of these lexically ambiguous types or tokens are recognized by the grammar. When the grammar recognizes a lexically ambiguous type, it is important to realize that the type is a 'shorthand' notation for several types.

For example :

(1) DN, PE NC

occurs 721 times in the original corpus. The first word, according to Dictionary I, could be either DN or PE, and hence (1) expands to

(1′) DN NC
(1″) PE NC.

In this case (1′) was recognized by Grammar I while (1″) was not; hence, we say that the lexical ambiguity was *resolved,* meaning that only one of the alternative types was recognized. Another lexically ambiguous case which occurred 13 times is :

(2) DN, PE NC, NP

It represents the types

(2.1) DN NC
(2.2) DN NP
(2.3) PE NC
(2.4) PE NP

In this case (2.1) and (2.2) were both recognized, and so we say that the lexical ambiguity was *reduced* (but not resolved). This reduction happened on 14 types (35 tokens) in the corpus. Nineteen lexically ambiguous types (62 tokens) remained that were not resolved by Grammar I. The elimination of this remaining lexical ambiguity will be discussed further after we have discussed the probabilistic fit of Grammar I.

The *ambiguity factor* is, intuitively, the number of 'extra' analyses of a noun phrase. Let S be the set of types in a corpus, for t in S, f(t) is the *frequency* of t, n(t) is the *number* of types (using simple symbols) that arise from t. Then, the *original ambiguity factor* is given by

$$(3) \qquad \sum_{t \text{ in } S} f(t)*(n(t)-1) \ .$$

We subtract 1 in order to eliminate an entry for those types that have a single classification. For example, the contribution of

9 DN, PE NC, VN

to the original ambiguity factor is

$9*(4-1)=27$,

where 9 is the frequency, and 4 is the number of simple types arising from the type in question.

The *parsed ambiguity factor* is defined in a similar fashion, except now we use S′ instead of S, where S′ is the set of types in S recognized by the grammar.

The *reduced ambiguity factor* is given by the formula

$$\sum_{t \text{ in } S'} f(t)*(n'(t)-1) \ ,$$

where S′ is the subset of S recognized by the grammar, and n′(t) is the number of simple types arising from t that are recognized by the grammar. Among the measures we consider, the reduced ambiguity factor is clearly the best measure of the lexical ambiguity permitted by the grammar. As will be seen, this number is quite small for all of the analyses presented in this work.

The resolution and reduction of lexical ambiguity alters the list of types present in the corpus. For example,

721 DN, PE NC

is resolved to

721 DN NC

(since PE NC is not recognized by Grammar I) and must be combined with

34 DN, PE NC, VT

which is also resolved to DN NC. This process is called *consolidation.* Of 222 original types recognized by Grammar I, 140 types remained after consolidation.

Probabilistic fit of Grammar I. Following the standard notion of a probabilistic grammar and the various ways of estimating the goodness of fit (SUPPES, 1970, 1971; SMITH, 1972), we used a model based on a geometric distribution for predicting the parameters of the grammars reported upon here.

Table 7 gives the parameters for the standard fit obtained by the maximum-likelihood method, and Table 9 shows the results of the goodness-of-fit test for each noun-phrase type. Types with expected frequency less than 5 were grouped together in accordance with standard statistical practice. The residual was 498.32, which indicates that 9.6 percent of the noun phrases predicted by the grammar did not occur in the corpus. There were 16 groups (low-expected types clumped together), thus reducing the degrees of freedom to 15. The chi-square was 3532 and the chi-square divided by the degrees of freedom was 235.47.

We now examine these tables, with a regard to what they tell us about Philippe's noun phrases. In Table 7 notice that the 5-rules, which generate pre-position adjectival phrases, have nearly the same parameters as the 6-rules, which generate the post position adjectival phrases. This means our hypothesis that these would be different in character is rejected. The tendency to use more than one AQ is always small, whether in pre- or postposition. Notice among the 4-rules that the rules

(4,15) $NP^* \rightarrow DET\ NP'$
(4,16) $NP^* \rightarrow NP'$
(4,17) $NP^* \rightarrow NP''$

TABLE 9

Noun-phrase Types for Grammar I

Observed	Expected	Chi-square	Count	Noun phrase type
1436	1259.5	24.6	1	PE
835	751.4	9.2	1	PD
792	630.	41.4	1	DN NC
324	125.5	312.2	1	NP
284	254.6	3.3	1	IN NC
245	53.4	683.6	1	EP NC
171	56.1	233.	1	PR
167	709.1	413.6	1	NC
158	139.4	2.4	1	PI
107	76.9	11.5	1	AQ NC
61	24.9	50.9	1	IN AQ NC
60	47.5	3.1	1	EP DN NC
55	61.6	.6	1	DN AQ NC
45	42.1	.1	1	AQ NC
37	51.0	3.6	1	AD NC
18	13.6	1.1	1	AC NC
16	14.3	.1	1	DN NC AQ
13	5.8	7.8	1	IN NC AQ
23			1	AD NC AD
23			1	IN NC EP NC
21			1	DN NC EP NP
13			1	DN NC EP NC
12			1	DN PO
GROUP 92	9.1	742.1		
11	10.2	.0	1	AI NC
11			1	CC PD
11			1	EP AQ NC
GROUP 22	5.8	42.9		
11	9.5	.1	1	EP NP
10			1	DN NC KO
10			1	NC EP NC
GROUP 20	6.7	24.3		

Table 9 (following):

8	7.2	0.1	1	AT NC
8			1	AD AQ NC
8			1	NP CC NP
GROUP				
16	5.0	21.9		
8			1	PD KO
7			1	DN NC EP DN NC
GROUP				
15	6.8	8.8		
6	12.7	3.0	1	NC AQ
5	111.5	100.8	1	DN NP
4	7.5	1.2	1	AO AQ NC
6			1	EP NC AQ
5			1	IN AQ NC EP NC
5			1	PE LC
4			1	AD NC KO
4			1	AO NC EP DN NC
4			1	CC PE
GROUP				
28	7.4	55.1		
3	13.6	7.5	1	AO NP
4			1	DN NC CC DN NC
4			1	IN NC CC IN NC
3			1	AC AQ NC
3			1	AD AQ NC AD
3			1	AI AQ NC
3			1	AO NC EP PE
3			1	AO NC KO
3			1	DN NC EP AD NC
3			1	DN NC EP IN NC
GROUP				
29	5.9	86.3		
3			1	DN NC EP PE
3			1	EP AI NC
GROUP				
6	5.6	.0		
2	9.0	4.7	1	AD NP
3			1	EP AT NC
3			1	IN AQ NC EP NC A
3			1	NC EP NP
3			1	PD CC PD
2			1	AB AB NC

Table 9 (following):

2			1	DN AC AQ NC
2			1	DN AQ NC AQ
GROUP				
18	5.5	25.7		
2	49.9	45.0	1	DN PR
2	19.2	14.5	1	EP IN NC
2	10.5	6.1	1	EP PI
2			1	DN NC CC DN NC AQ
2			1	IN AC NC
2			1	NC EP DN NC
2			1	NP AQ
GROUP				
8	5.5	.8		
2			1	NP CC NC
2			1	PD EP NP
2			1	PD EP PD
1			1	AC NC AQ
1			1	AC NC CC AC NC
1			1	AD AQ NC KO
1			1	AD NC AD CC PD
1			1	AD NC AQ
GROUP				
11	6.0	3.4		
1	6.1	3.5	1	AO PR
1			1	AQ NC CC DN NC
1			1	AQ NC EP IN NC
1			1	AQ NC EP NC
1			1	AQ NC EP NP
1			1	AT AQ NC
1			1	CC DN NC
1			1	DN AC NC
1			1	DN AQ NC EP NP
1			1	DN NC CC AO NC
1			1	DN NC EP AO NC
1			1	DN NC EP AQ NC
GROUP				
11	5.0	6.0		
1	5.8	3.2	1	EP AO NC
1			1	DN NC EP PD
1			1	DN NC EP PR
1			1	DN PO EP NC
1			1	DN PO EP NP
1			1	DN PO KO
1			1	EP AD NC AD

Table 9 (following and end):

1			1	EP AQ AQ NC
1			1	EP AQ NC AQ
1			1	EP DN AQ NC
GROUP				
9	8.0	.0		
1			1	EP PR
1			1	IN AQ AQ AQ NC
1			1	IN AQ AQ NC
1			1	IN AQ AQ NC AQ
GROUP				
4	5.0	.1		
1	9.1	6.4	1	NC AD
1			1	IN AQ NC AQ
1			1	IN AQ NC EP DN AQ NC
1			1	IN AQ NC EP DN NC
1			1	IN NC AQ CC IN NC AQ
1			1	IN NC EP AD NC
1			1	IN NC EP DN AQ NC
1			1	IN NC EP DN NC
1			1	IN NC EP IN AQ NC
1			1	IN NC EP IN NC
1			1	IN NC EP IN NC AQ
1			1	IN NC EP PR
1			1	NC AQ AQ
1			1	NC AQ EP NP
1			1	NC EP PR
1			1	NP CC PE
1			1	NP EP NP
1			1	PE CC DN NC
1			1	PE CC PE
GROUP				
18	6.6	17.8		
1			1	PI EP IN NC
GROUP				
1	.2	.4		

	498.3	498.3 RESIDUAL
5191	5191	3532.1

Degrees of freedom = 15
Chi-square/degrees of freedom = 235.473
Groups = 16

have between them 94 percent of the probability. Again this limits the expected length of the noun phrases rather sharply (see Table 2).

Turning to Table 9, one notices that the predicted values for the high-frequency types are sometimes rather different from the observed frequency. The problem arises most with the types

PE
DN NC
NP
EP NC
PR

which have far too small an expected frequency, and NC, which has far too large an expected frequency.

Notice that the 2-rules and 3-rules, which introduced NC, PE, etc., into longer phrases by replacement in the 4-rules, also performed double duty in creating, through Rules (1,3), (4,16) and (4,17), the one-word phrases that caused so much difficulty. The grammar needs a different mechanism for introducing these symbols (NC, PE, etc.) into longer noun phrases.

The residual suggests that the grammar generated some types not found in the corpus. The 1-rules and 4-rules are mainly responsible for this. Note that Rule (1,4) (SN → EP NP*) generated many types not found. Also, while

245 EP NC
11 EP AQ NC

were underpredicted by Grammar I, some other types using (1,4) were overpredicted, such as the following :

2 EP IN NC
2 EP PI

Several types using (1,2) (SN → NP* EP NP*) were underpredicted, including

23 IN NC EP NC
21 DN NC EP NP
13 DN NC EP NC
7 DN NC EP DN NC
5 IN AQ NC EP NC .

Other problems are hidden within groupings of low-frequency utterances. These include :

Observed Frequency	Expected Frequency	Chi-square	Noun-phrase type
23	.659	724.388	AD NC AD
21	.479	836.835	DN NC EP NP
8	.023	2456.477	NP CC NP
3	.002	3263.561	IN AQ NC EP NC AQ
1	.000	938.975	AC NC CC AC NC
1	.000	5189.316	IN NC AQ CC IN NC AQ

Note that DN NC EP PN, as well as IN AQ NC EP NC AQ, used Rule (1,2) as a further indication of the problems with that rule. NP CC NP and AC NC CC AC NC and IN NC AQ CC IN NC AQ used Rule (1,1) (SN → NP* CC NP*). The types using (1,1) have difficulties similar to those of (1,2), in that the more complex of them were underpredicted.

We now turn to the geometric fit for Grammar I. The method consisted of taking the rules in each class, ordering them according to the highest observation of their usage, and predicting this observed usage with a truncated geometric distribution. The mechanism assumed to generate the distribution is discussed below. The gain resulting from this simplification is a substantial decrease in the number of parameters in the model.

Table 10 shows the appropriateness of the geometric model. The rules in each class are ordered according to their observed usage. The numbers in the "Observed" column represent the number of times the rule was used in the corpus, weighed, of course, against the frequency of the types. The standard geometric distribution with a single parameter was used to predict the values given in the "Expected" column.[1] The parameter is given below each class of rules. (More precisely, the distributions are truncated geometric, since the tail of the distribution is added to the last element of the list. This practice can be troublesome on a very flat distribution, such as that for the rules in group 2, where the tail contains enough probability to inflate the last value. We have, however, found that the distributions are usually not this flat. We should remark that, when there are only two rules in a group (as in group 5), the truncated geometric simply reduces to the binomial.)

It is clear what mechanism is natural to associate with a truncated geometric distribution. We assume that the production rules for rewriting a nonterminal of a given kind are located in a push-down store. By usage and past experience, the rules are stored from the top down in approximate order of usage, that is, with the most frequent being on top, etc. When a given rule is used the speaker goes to the location, selects the first rule with average probability θ, goes to the second rule with

[1] By the maximum likelihood estimation, the parameter is the reciprocal of the mean of the distribution.

probability $(1-\theta)$; upon reaching the second rule then selects it with probability θ or with probability $(1-\theta)$ goes on to the next rule, and so forth. It is important to interpret these probabilities as being on the average. Obviously, in a given situation semantic and perceptual factors will determine often uniquely which rule will be selected, although in many cases it will still be undetermined and a probabilistic selection will be made. What is important about the assumption of such a model in the present context is that it is meant to represent averaging results of accessing without taking into account particular semantic or perceptual features of a given occurrence of use.

Table 10 gives the parameters of the grammar based on the geometric distribution for comparison with the full-parameter distribution.

Probabilistic lexical disambiguation, Grammar I. In Section III, we mentioned that the ambiguity created by multiple classifications of words in the dictionary (called lexical ambiguity) was, for the most part, resolved by Grammar I. However, 17 types, representing 62 tokens (19 types before consolidation) remained that were lexically ambiguous and not resolved by Grammar I.

As in SMITH (1972), we found that a plausible method of removing lexical ambiguity from a corpus involved the use of a probabilistic grammar. The method consisted of these steps :

1. Given an ambiguous corpus and a grammar, obtain estimates for the parameters associated with the rules of the grammar, as described above. If the lexical ambiguity remaining after resolution is small (in this case, it was less than 1 percent of the tokens), then the parameters will not be too inaccurate.
2. For each lexically ambiguous type, select the lexical reading that has the highest probability according to the parameters obtained in (1).

In Table 11 we give the 17 ambiguous types and include the resolution suggested by the above method. An examination of the actual noun phrases suggests that this method is reasonably accurate in removing spurious ambiguities.

Grammar I and the sections of the noun-phrase corpus. Our noun-phrase corpus was drawn from three parts of the 33 sessions (Section I is Sessions 1 through 3, II is Sessions 14 through 16, and Section III is Sessions 31 through 33). We analyzed each of these sessions separately, especially since we were interested in the developmental aspects of the child and his growing sophistication in the use of language. Table 12 shows the results of analyzing each of the three different time sections of the three sessions separately.

TABLE 10
Parameters for Geometric Fit of Grammar I

Label	Observed	Expected	Rule
(1,3)	4633	4408.9	SN => NP*
(1,4)	349	664.3	SN => EP NP*
(1,2)	130	100.1	SN => NP* EP NP*
(1,1)	31	15.1	SN => NP* CC NP*
(1,7)	27	2.3	SN => NP* KO
(1,5)	16	.3	SN => CC NP*
(1,9)	5	.1	SN => NP* LC
Theta =	.8493		
(2,1)	1455	1642.6	NP" => PE
(2,2)	868	556.4	NP" => PD
(2,3)	161	285.0	NP" => PI
Theta =	.6613		
(3,2)	2248	2179.6	NP' => NC
(3,1)	398	523.1	NP' => NP
(3,4)	178	125.6	NP' => PR
(3,3)	29	30.1	NP' => NC AD
(3,5)	15	9.5	NP' => PQ
Theta =	.7600		
(4,17)	2484	2797.3	NP* => NP"
(4,15)	1518	1335.3	NP* => DET NP'
(4,16)	1045	637.4	NP* => NP'
(4,11)	153	304.2	NP* => DET ADJP NP'
(4,12)	64	145.2	NP* => ADJP NP'
(4,7)	35	69.3	NP* => DET NP' POSTADJP
(4,14)	20	33.0	NP* => NUM NP'
(4,8)	19	15.8	NP* => NP' POSTADJP
(4,3)	4	7.5	NP* => DET ADJP NP' POSTADJP
(4,10)	3	3.6	NP* => NUM ADJP NP'
(4,13)	3	1.7	NP* => DET NUM NP'
(4,9)	2	.8	NP* => DET NUM ADJP NP'
(4,4)	1	.4	NP* => ADJP NP' POSTADJP
(4,6)	1	.4	NP* => NUM NP' POSTADJP
Theta =	.5227		
(5,1)	227	227.0	ADJP => AQ
(5,2)	7	7.0	ADJP => ADJP AQ

Table 10 (following and end):

Theta =	.9701		
(6,1)	60	60.0	POSTADJP => AQ
(6,2)	1	1.0	POSTADJP => POSTADJP AQ
Theta =	.9836		
(7,4)	1049	1055.9	DET => DN
(7,5)	424	405.8	DET => IN
(7,3)	128	156.0	DET => AQ
(7,1)	85	59.9	DET => AD
(7,6)	17	23.0	DET => AI
(7,2)	12	14.4	DET => AT
Theta =	.6157		
(8,1)	29	29.0	NUM => AC
Theta =	1.0000		

For example, the first time section, Sessions 1-3, contained a total of 136 types in the corpus, with 101 types recognized. Eighty-six of these types were ambiguous in the corpus, meaning that at least one word in the noun phrase had more than one grammatical classification. Of these 86, 67 were recognized and of the 67, 59 had the ambiguity resolved to a single lexical form. Of the remaining 8 types, 6 had the ambiguity reduced in a number of lexical forms. Similar remarks are to be made about the tokens, which are tabulated in the second column. The ambiguity factors for the three sections are also shown. What is important is the reduced ambiguity factor as defined above, which reflects the small amount of ambiguity in the final analysis. In other words, the reduced ambiguity factor of 14 indicates the number of tokens among the 1,512 recognized that have more than one derivation tree. For the second time period, Sessions 14-16, the reduced ambiguity factor is 17, and for the third time section, Sessions 31-33, the reduced ambiguity factor is 33.

After making the analysis shown in Table 12, we consolidated the types for each time section, and the subsequent analysis is in terms of these types. The first time section consolidated to 69 types, the second to 68, and the third to 87.

A small amount of lexical ambiguity remained in each section after resolution, and as in the corpus as a whole, we used the probabilistic method of disambiguation described above and consolidated the types

TABLE 11

Lexical disambiguation of noun phrases in Grammar I

Freq	Ambiguous type	Resolved type	Prob
16	IN NC,NP un monsieur	IN NC	.05
13	DN NC,NP le monsieur	DN NC	.12
6	NC,PI personne	NC	.14
5	DN NC AQ,LC le bras gauche	DN NC AQ	.003
3	AQ,AI NC autre chose	AQ NC	.008
3	EP NC,NP au monsieur	EP NC	.010
2	AQ,KO,NC NC,LC bon côté	AQ NC	.008
2	DN NC EP NC,NP la bouche du monsieur	DN NC EP NC	.0005
2	DN,PE NC,LC le côté	DN NC	.12
2	DN NC,NP KO le monsieur là	DN NC KO	.0007
2	EP AQ,AI NC des autres roues	EP AQ NC	.0006
1	AQ,LC AQ NC dernier petit trou	AQ AQ NC	.0002
1	DN NC CC DN NC,NP la dame et le monsieur	DN NC CC DN NC	.0001
1	DN NC EP AQ,AI NC la forme des autres personne	DN NC EP AQ NC	.00003
1	DN NC EP IN NC,NP les chaussures d'un monsieur	DN NC EP IN NC	.0002
1	NP CC NC,NP mesdames et monsieur	NP CC NC	.00002
1	PE,PO leur	PE	.24

TABLE 12
Section Comparison of Noun-Phrase Corpus

Description	Section 1 Types	Section 1 Tokens	Section 2 Types	Section 2 Tokens	Section 3 Types	Section 3 Tokens
Total	136	1601	127	1536	216	2330
Recognized	101	1512	98	1488	141	2191
% Recognized	74.26	94.44	77.17	96.88	65.28	94.03
Ambiguous	86	647	82	401	150	699
% Ambiguous	63.24	40.41	64.57	26.11	69.44	30.00
Amb. recog.	67	600	64	381	107	634
% Amb. recog.	77.91	92.74	78.05	95.01	71.33	90.70
Amb. resolved	59	588	55	364	97	601
Amb. reduced	6	10	5	8	7	17
Original Amb. factor	978		553		1154	
Parsed Amb. factor	871		512		914	
Reduced Amb. ambiguity factor	14		17		33	
Types after consolidation	69		68		87	
Types after probabilistic disambiguation	61		61		78	

after this disambiguation. After this reduction there were 61 types in the first section, 61 in the second, and 78 in the third.

There is a fairly substantial reduction in the percentage of types recognized in the third time portion. The first two are just about the same. On the other hand, the reduction in number of tokens recognized in the third time section is small, amounting to only a small percentage. The explanation for the smaller percentage of types recognized in Section 3 is probably that Section 3 is substantially larger than the first two sections (2330 tokens compared with 1601 in Section 1 and 1536 in Section 2). We have noticed in previous work that the larger the sample, the smaller the percentage of recognized types.

Some of the differences among the three sections are unexpected. Table 13 shows the most frequent types from each section, together with the observed, expected, and chi-square contribution of each type. (Types were excluded whose frequencies were low enough to require grouping under the chi-square test of the geometric model.) It is rather surprising that DN NC occurred 393 times in Section I, but only 165 and 234 times in Sections 2 and 3, respectively. In Section I, EP NC occurred 60 times and in Section 3 it occurred 92 times.

As an interesting comparison, SUPPES (1970, 1971) in previous work on ADAM 1, found that in the earliest stages of development of an English-speaking child, the noun by itself is the most frequently occurring noun phrase. Also, in the study of the ERICA corpus (SMITH, 1972), the preponderance of nouns occurring alone decreases as the child nears three years of age. Quite plausibly, for a French-speaking child, the noun does not occur by itself (even at first), but rather with an article, which is natural in French for obvious reasons. Therefore, we see in Philippe's speech the same tendency that we have seen before in English corpora, but with the straightforward addition of the required article.

More general remarks of a developmental nature are reserved until the end of the section on Grammar II.

V. DICTIONARY II

Our second analysis, built directly on the first analysis, is aimed at an approach less based on the earlier work with English. We especially use the ideas found in DUBOIS (1970) to describe the construction of the determiner phrase that modifies a noun. These rules of the grammar are examined in more detail in Section VI.

To accommodate Dubois' ideas, Dictionary I required some modifications. In passing to Dictionary II, we have changed the grammatical categories as little as possible, in order to maintain a meaningful comparison. This means that some of our abbreviations, etc., still make Dictionary II look very much like Dictionary I. We have, however, at critical points introduced changes.

TABLE 13
Section Comparison
High-frequency Noun Phrases

Observed	Expected	Chisquare	Count	Noun phrase type
		Section I (Sessions 1-3)		
393	307.5	23.5	1	DN NC
212	185.5	3.6	1	PE
191	169.5	2.6	1	PD
177	69.3	165.8	1	NP
97	275.4	114.9	1	NC
82	68.2	2.6	1	IN NC
60	22.6	60.4	1	EP NC
39	30.8	1.9	1	AQ NC
38	33.7	.4	1	PI
35	25.2	3.4	1	EP DN NC
22	13.4	4.9	1	AQ NC
22	19.7	.2	1	DN AQ NC
		Section II (Sessions 14-16)		
427	372.4	7.9	1	PE
302	277.3	2.1	1	PD
165	124.2	13.1	1	DN NC
93	17.9	311.6	1	EP NC
85	36.3	64.2	1	NP
62	52.2	1.6	1	IN NC
60	22.8	58.8	1	PR
47	41.0	.7	1	PI
39	26.7	5.2	1	AQ NC
33	201.9	140.5	1	NC
21	21.5	.0	1	AD NC
16	6.8	11.2	1	IN AQ NC
15	16.2	.0	1	DN AQ NC
		Section III (Sessions 31-33)		
797	708.6	10.9	1	PE
342	306.1	4.1	1	PD
234	198.3	6.3	1	DN NC
140	133.4	.3	1	IN NC
105	30.3	181.6	1	PR
92	14.2	422.1	1	EP NC
73	64.9	.9	1	PI
62	21.6	73.6	1	NP
46	36.0	2.5	1	AQ NC
37	228.2	159.3	1	NC
30	15.0	13.9	1	IN AQ NC
18	22.3	.7	1	DN AQ NC
14	26.4	5.4	1	AD NC

The comparison of the grammatical categories of Dictionary I and Dictionary II is shown in Table 14.

TABLE 14
Correspondence between Dictionary I and II.

Dic. I	Dic. II	Dic. I	Dic. II
IN	IN	AI	DT
DN	DN		QA
PE	PE		NU
PD	PD		
PR	PR	NC	NC
PI1-7	PI1-7	NP	NP
PI8	DT	DV	DV
	QA		QR
	NU		QA
PO	PO	EP	EP
AQ	AQ	CS	CS
AC	AC	CC	CC
AN	AQ	KO	KO
AD	AD	LC	LC
AT	AT	MS	MS
AO	AO		

First, to facilitate the rewrite rules of Grammar II described in the next section, we broke the category of PI8 of Dictionary I into three categories, *denotative* (DT), the category that Dubois calls *numerical* (NU) and Dubois' category of *absolute quantitative* (QA). Associated with this last category is also Dubois' category of *relative quantitative* (QR), which is a subdivision of DV in Dictionary I. Note that in the categories QA and QR we classified words which were classified as indefinite pronouns or possibly as adjectives or adverbs in Dictionary I. In one point we did not follow Dubois completely, for we still entered many of the words he classifies as QR also as adverbs in Dictionary II. For example, *beaucoup* was coded both as QR and DV0. Also, to anticipate Grammar II, in the rewrite rules using QR the category of lexical items in the category QR are ordinarily followed by *de*.

As might be expected, from the recoding of the category *pronom indéfini* (PI8), there is a corresponding recoding from Dictionary I to Dictionary II of the category of *adjectifs indéfinis* (AI) into the same three categories DT, QA and NU.

The third major change is the classification of adverbs (DV). In Dictionary II this classification is broken up into the adverb category, the absolute quantitative category and the relative quantitative category. We have already remarked about the two latter categories in Dictionary II, and consequently, nothing more is said here.

However, because the number of lexical items in these new categories is small, and because there is a lack of systemization of all the words to be considered by Dubois, we list the complete category for DT, NU, QA and QR. The following words were placed in the denotative category DT : *autre(s)*, *là* (when following a noun), *même(s)*. The numerical category NU contained the words : *chaque, nulle, quelque, plusieurs.* The absolute quantitative category contained the words : *tout, toute(s), tous.* Finally, the relative quantitative category contained the words : *assez, beaucoup, peu, plus, trop.*

The coding of *là* was DT when it followed a noun and seemed to have the interpretation of the composed form of the demonstrative adjective in such sentences as *Appuie sur le bouton là.* The ordinal numbers have been coded as qualificative adjectives, that is, category AQ, following DUBOIS (1970, p. 54).

Some of the decisions made in establishing the categories of Dictionary II will be more evident when we turn to the production rules of Grammar II.

VI. NOUN-PHRASE GRAMMAR II

Grammar II, which is derived from DUBOIS (1970), is closer than our Grammar I to the current literature on generative grammars of the French language. It is already clear from discussing the construction of Dictionary II how the grammatical categories have been changed. In order to facilitate the comparison of the two grammars, the main subheadings of this section are precisely the same as those of the section dealing with Grammar I.

Nonterminal vocabulary of Grammar II. In addition to the grammatical categories introduced in Dictionary II, we have used the following additional nonterminal symbols in Grammar II. These additional symbols cannot be rewritten by lexical rules, but they occur in derivation trees always at least two nodes from a terminal node. As in Grammar I, the label of the root of all the noun phrases is SN (*syntagme nominal*). Following Dubois we also introduced GN and, for more specialized purposes of the final probabilistic analysis, GN′. The nonterminal GN is an abbreviation for *groupe nominal.* We also have NP* playing a similar role to NP* in Grammar I.

Additional nonterminals taken from Dubois are PREART for prearticles, POSTART for post-articles, ART for articles, and NUM for numerals. As in the case of Grammar I, we also used ADJP for adjective

phrases in pre-position and POSTADJP for adjective phrases in post-position. Finally, we used N for nouns with N being rewritten as either a common noun or a proper noun, that is, as a NC or NP.

By again omitting the subscripts introduced in the original dictionary coding, we achieved the same level of abstraction as in Grammar I.

The production rules. The 56 production rules of Grammar II are shown in Table 15. As in the case of the rules of Grammar I, the production rules have a surprisingly standard character. The prototypes of a good many of them are to be found in DUBOIS (1970).

In our formalization of the notions of Dubois, we first used a grammar somewhat closer to that of Dubois; we discuss below the reasons for the main modifications.

The basic structure of the production rules reflects Dubois' structural representation of the nominal group (GN) as the determiner plus a noun, where the determiner itself is rewritten as shown in rules (4,1) to (4,11). Thus, the structure of the determining phrase and its importance to the noun phrase is a critical part of the Dubois treatment. We have taken over Dubois' distinctions of pre-article, demonstrative articles and post-articles, as mentioned above. As the production rules for pre-articles we have also used his division into absolute and relative quantities as reflected in rules (8,1) and (8,2). Unlike Dubois, we have also coded as adverbs the words coded as QR. We adapted his rewriting rules for the post-articles, and this is reflected in rules (6,1) (6,2) and (6,3). Dubois included as one of the possible rewrite rules for post-articles the introduction of the word *tel,* but we omitted this rule because *tel* does not appear in the Philippe corpus.

Examination of Table 15 shows that we have a larger number of groups of rules in Grammar II than in Grammar I; in particular there are 12 groups of rules in Grammar II compared to 8 in Grammar I. In the first group we have 11 rules compared to 9 in the first group of Grammar I. These are the rules that govern the first step from the labeled root of the tree SN.

The two rules in the second group, starting from GN, generate either symbol GN′ or the single terminal NC. The reason for introducing these rules was to generate the terminal NC (common nouns) at a high level in the grammar in order to obtain a better probabilistic fit. Comparing this feature of Grammar II with Grammar I, we see that most of our modifications have mainly changed the level of introduction of certain symbols, and not the general nature of the grammar or to any degree the noun phrases recognized.

Rule (1,10) requires some separate remarks. The purpose of this rule is to generate, at a high level in the grammar, phrases like *d'argent.* In these phrases, the preposition EP is in fact the partitive construction, as is indicated by the common English translation *some money.* Our

TABLE 15

Parameters for Grammar II

Philippe's Noun Phrases*

Label	Prob.	Rule
(1,1)	.8616	SN => GN!
(1,2)	.0255	SN => GN EP GN
(1,3)	.0059	SN => GN CC GN
(1,4)	.0320	SN => NC
(1,5)	.0031	SN => CC GN
(1,6)	.0000	SN => KO GN
(1,7)	.0052	SN => GN KO
(1,8)	.0000	SN => LC GN
(1,9)	.0002	SN => GN LC
(1,10)	.0470	SN => EP NC
(1,11)	.0196	SN => EP GN!
(2,1)	.8065	GN => GN!
(2,2)	.1935	GN => NC
(3,1)	.2963	GN! => PE
(3,2)	.1773	GN! => PD
(3,3)	.0357	GN! => PR
(3,4)	.0310	GN! => PI
(3,5)	.0161	GN! => NP*
(3,6)	.0031	GN! => DN PO
(3,7)	.0351	GN! => D NP*
(3,8)	.0055	GN! => ART POSTART
(3,9)	.0794	GN! => NP
(3,10)	.3133	GN! => D NC
(3,11)	.0051	GN! => D NC AD
(3,12)	.0020	GN! => D NP
(4,1)	.0000	D => PREART AD POSTART
(4,2)	.0000	D => PREART ART POSTART
(4,3)	.0000	D => AD POSTART
(4,4)	.0264	D => POSTART
(4,5)	.0000	D => PREART AD
(4,6)	.0000	D => PREART POSTART
(4,7)	.0178	D => ART POSTART
(4,8)	.0075	D => PREART ART
(4,9)	.0471	D => AD
(4,10)	.8978	D => ART
	.0034	D => PREART

Table 15 (following and end):

(5,1)	.0802	ART => AQ
(5,2)	.0073	ART => AT
(5,3)	.2656	ART => IN
(5,4)	.6469	ART => DN
(6,1)	.4712	POSTART => NUM
(6,2)	.0000	POSTART => NUM DT
(6,3)	.5288	POSTART => DT
(7,1)	.6531	NUM => AC
(7,2)	.3469	NUM => NU
(8,1)	.6842	PREART => QA
(8,2)	.3158	PREART => QR EP
(9,1)	.0199	NP* => ADJP N POSTADJP
(9,2)	.7610	NP* => ADJP N
(9,3)	.2191	NP* => N POSTADJP
(10,1)	.9751	ADJP => AQ
(10,2)	.0249	ADJP => ADJP AQ
(11,1)	.9836	POSTADJP => AQ
(11,2)	.0164	POSTADJP => POSTADJP AQ
(12,1)	.9920	N => NC
(12,2)	.0080	N => NP

*Correction for continuity = .5; expected < 5 are grouped together.

original "Dubois" grammar did not allow for this construction, and our first effort to recognize these phrases was to add as a rule to group 3.

$$GN' \rightarrow EP\ NP^* .$$

The problem with this rule, from a probabilistic point of view, is that it does not allow for the rather common occurrence of EP NC (which occurred 245 times in the consolidated version of Philippe's speech under Grammar II), as opposed to phrases like EP NC AQ, which occurred only 6 times. Therefore, we changed the grammar to allow for the separate generation of EP NC by means of (1,10), and took care to see that no ambiguities were introduced.

The third group of rules, (3,1) to (3,12), provide the rewrite rules for the nominal group corresponding fairly closely to suggestions in

Dubois. There are however several differences. DUBOIS (1970, p. 35) discusses the view that the GN structure has these forms :

GN → D N
GN → NP
GN → PRONOUN

and proposes as the only rule GN → DN. The first change that we made was to add rules (3,1) to (3,4) to accommodate the several kinds of pronouns that we have defined. This change is hardly more than terminology. A deeper change, but one that is consistent with Dubois' treatment of adjectives (see pp. 126-131), is to allow adjective phrases to be introduced after the determiner element, using rules (9,1) through (9,3). Thus, we have added rules (3,5) and (3,7) to the rules in group 3.

One additional rule that deserves mention is rule (3,6), which is also a departure from Dubois. This rule was introduced to remedy a rather bad probabilistic discrepancy that had occurred in Grammar I. Noun phrases of the type DN PO can be generated in Grammar I by the use of rules (4,15), (3,5) and (7,4). However, as Table 7 shows, the theoretical frequency of this type of utterance was negligible, while the type has an actual frequency of 12 in our selected nine hours of the corpus. The much better fit of the new rule introduced in Grammar II is seen from the data in Table 17 below. The predicted frequency is now 13,769, which matches closely the observed frequency of 12. Also, rule (3,8) has been introduced to generate noun-phrases of the type DN DT.

The fourth group of rules, the rewrite rules for determiners, constitute the most important contribution of Dubois' analysis to Grammar II. Here we have followed closely his suggestions, because we think they provide an excellent framework for handling the various parts of speech that precede nouns in noun phrases. We can summarize these rules by saying that they generate the 11 grammatically sensible possibilities out of the 16 possibilities obtained by either using or not using each of the categories PREART, AD, ART, POSTART.

The fifth group of rules are the rules for the post-articles to be rewritten as "numerical" terms, which can lead to a single word in the NUM category or NUM followed by a denotative or a denotative alone. The seventh group of rules are the rewrite rules for splitting the numerical category NUM into the grammatical categories AC and NU, which are discussed above in the section on Dictionary II. The pre-article rewrite rules form the eighth group and provide for rewriting pre-articles as either absolute or relative quantities, with of course the relative quantities followed by a preposition. The ninth, tenth, eleventh and twelfth groups of rules are close to rules occurring in Grammar I and do not require additional comment.

Nonprobabilistic fit of Grammar II. The data are summarized in Table 16, whose entries parallel those of Table 8. Grammar II does better

than Grammar I in terms of the total number of noun phrases recognized. Grammar II parses 5,217 tokens in contrast to the 5,191 for Grammar I. The number of types that are parsed is about the same — 226 for Grammar II, 222 for Grammar I. The reduced ambiguity factor is also similar: 62 for Grammar II and 64 for Grammar I.

TABLE 16
Nonprobabilistic Fit of Grammar II Noun-phrase Corpus

Description	Types	Tokens
Total	339	5467
Recognized by Grammar II	226	5217
Percent recognized	66.7%	95.4%
Lexically ambiguous	225	1723
Percent lexically ambiguous	66.4%	31.5%
Ambiguous recognized	158	1604
Percent ambiguous recognized	70.2%	93%
Ambiguity resolved	142	1544
Ambiguity reduced	11	33
Original ambiguity factor	2516	
Parsed ambiguity factor	2233	
Reduced ambiguity factor	62	
Types after consolidation	147	
Types after probabilistic disambiguation	132	

After initial parsing, the 226 recognized types consolidated to 147 types. This compares to a somewhat lower figure of 140 types for Grammar I.

It might be thought from this discussion that Grammar II is not significantly better than Grammar I. However, under the much tighter and stricter probabilistic criterion that we next discuss Grammar II generates a language that has a much better fit to the corpus of noun phrases than does that generated by Grammar I.

Probabilistic fit of Grammar II. Table 15 gives the parameters for the standard fit of Grammar II, obtained by the maximum-likelihood method, and Table 17 shows the results of the goodness-of-fit test for each noun-phrase type. Once again, types with expected frequency less than 5 were grouped together, and we used a continuity correction of .5.

TABLE 17

Noun-Phrase Types for Grammar II

Observed	Expected	Chisquare	Count	Noun-phrase type
1436	1331.9	8.1	1	PE
835	796.7	1.8	1	PD
792	817.7	.8	1	DN NC
324	357.1	3.0	1	NP
284	335.8	7.8	1	IN NC
245	245.0	.0	1	EP NC
171	160.6	.6	1	PR
167	167.0	.0	1	NC
152	139.5	1.0	1	PI
107	101.3	.3	1	AQ NC
60	18.6	90.3	1	EP DN NC
52	27.7	20.4	1	IN AQ NC
43	53.4	1.8	1	AQ NC
40	67.5	10.8	1	DN AQ NC
37	66.3	12.5	1	AD NC
18	11.4	3.2	1	AC NC
16	19.6	.5	1	DN NC AQ
15	8.6	4.1	1	DN DT NC
23			1	AD NC AD
23			1	IN NC EP NC
21			1	DN NC EP NP
13			1	DN NC EP NC
GROUP				
80	7.7	674.0		
13	8.0	2.5	1	IN NC AQ
12	8.5	1.1	1	DN DT
12	13.8	.1	1	DN PO
11	8.1	.7	1	EP NP
11	6.1	3.2	1	NU NC
11			1	CC PD
10			1	DN NC KO
GROUP				
21	6.2	32.5		
10			1	IN DT
10			1	NC EP NC
GROUP				
20	8.5	14.4		

Table 17 (following):

8	5.5	.8	1	AD AQ NC
8	9.3	.1	1	AT NC
9			1	EP AQ NC
9			1	IN DT NC
8			1	NP CC NP
8			1	PD KO
GROUP				
34	8.7	70.4		
7			1	DN NC EP DN NC
7			1	QA DN NC
GROUP				
14	7.5	4.8		
6	15.5	5.2	1	NC AQ
5	5.3	.0	1	DN NP
4	8.4	1.8	1	AO AQ NC
6			1	EP NC AQ
5			1	IN AQ NC EP NC
5			1	QR EP NC
4			1	AD NC KO
4			1	AO NC EP DN NC
4			1	CC PE
GROUP				
28	6.5	67.7		
4			1	DN NC CC DN NC
4			1	IN NC CC IN NC
3			1	AC AQ NC
3			1	AO NC EP PE
3			1	AO NC KO
3			1	AO NP
3			1	DN AC
GROUP				
23	8.4	23.8		
3			1	DN NC EP AD NC
3			1	DN NC EP IN NC
3			1	DN NC EP PE
GROUP				
9	6.1	1.0		
3			1	EP AT NC
3			1	EP NU NC
3			1	IN AQ NC EP NC AQ
3			1	NC EP NP
3			1	NU AQ NC
3			1	PD CC PD
3			1	QA AQ NC

Table 17 (following):

3			1	QA DN NC EP NP
2			1	AD NP
2			1	AQ AQ NC
GROUP				
28	5.5	88.5		
2	19.7	15.0	1	DT NC
2	7.6	3.4	1	EP IN NC
2			1	DN AC AQ NC
2			1	DN AQ NC AQ
2			1	DN NC CC DN NC AQ
2			1	EP DT NC
2			1	IN AC NC
2			1	NC EP DN NC
GROUP				
12	8.4	1.1		
2			1	NP AQ
2			1	NP CC NC
2			1	PD EP NP
2			1	PD EP PD
1			1	AC NC AQ
1			1	AC NC CC AC NC
1			1	AD AQ NC KO
1			1	AD NC AD CC PD
1			1	AD NC AQ
GROUP				
13	6.3	6.0		
1			1	AQ DT
1			1	AQ NC CC DN NC
1			1	AQ NC EP IN NC
1			1	AQ NC EP NC
1			1	AQ NC EP NP
1			1	AT AQ NC
1			1	CC DN NC
1			1	DN AC NC
GROUP				
8	9.6	.1		
1			1	DN AQ NC EP NP
1			1	DN NC CC AQ NC
1			1	DN NC EP AQ NC
1			1	DN NC EP DT NC
1			1	DN NC EP PD
1			1	DN NC EP PR
1			1	DN PO EP NC
1			1	DN PO EP NP
1			1	DN PO KO
1			1	EP AD NC AD

Table 17 (following and end):

```
        1                                       1      EP AO NC
GROUP
       11          6.4         2.6
--------------------------------------------------------------------------------
        1                                       1      EP AQ AQ NC
        1                                       1      EP AQ NC AQ
        1                                       1      EP DN AQ NC
        1                                       1      EP PR
GROUP
        4          5.2          .1
--------------------------------------------------------------------------------
        1                                       1      IN AQ AQ AQ NC
        1                                       1      IN AQ NC AQ
        1                                       1      IN AQ NC EP DN AQ NC
        1                                       1      IN AQ NC EP DN NC
        1                                       1      IN DT AQ NC
        1                                       1      IN DT AQ NC AQ
        1                                       1      IN DT EP NC
        1                                       1      IN NC AQ CC IN NC AQ
        1                                       1      IN NC EP AD NC
        1                                       1      IN NC EP DN AQ NC
        1                                       1      IN NC EP DN NC
        1                                       1      IN NC EP IN AQ NC
        1                                       1      IN NC EP IN NC
        1                                       1      IN NC EP IN NC AQ
        1                                       1      IN NC EP PR
        1                                       1      NC AQ AQ
        1                                       1      NC AQ EP NP
        1                                       1      NC EP PR
        1                                       1      NP CC PE
        1                                       1      NP EP NP
GROUP
       20          5.5         37.8
--------------------------------------------------------------------------------
        1                                       1      PE CC DN NC
        1                                       1      PE CC PE
        1                                       1      PE LC
        1                                       1      QR EP AQ NC
GROUP
        4          3.2          .0
--------------------------------------------------------------------------------

              265.8       265.8                       Residual

  5217        5217        1491.3                      Totals

Degrees of freedom =  9

Chi-square/degrees of freedom =  165.7

Groups =   16

--------------------------------------------------------------------------------
```

We may compare Table 17 with Table 9 to substantiate our claim that Grammar II has a better fit than Grammar I. Notice first that the residual of Grammar II is 265, whereas that of Grammar I is 498. This means that 10 percent of the probability in Grammar I was distributed among noun-phrase types that did not occur in Philippe, while in Grammar II only about 5 percent of the probability is so distributed. This is not a measure of the adequacy of a grammar, but instead of the tightness of the grammar relative to the corpus under consideration. Grammar I, based on our previous experience with English, is too broad to provide a sharp characterization of the speech of Philippe.

The goodness-of-fit of Grammar II is even more encouraging. With 9 degrees of freedom, the chi-square sum over the noun-phrase types is 1491 for Grammar II, compared to 3532 with 15 degrees of freedom for Grammar I. On comparing the high-frequency noun phrases for the two grammars we quickly see that the predictions of Grammar II are much more accurate than those of Grammar I. The high-frequency noun-phrase types (which happen to be the same in both dictionaries) are shown in Table 18.

TABLE 18

Comparison of Grammars I and II

Philippe Noun phrases

Observed	Chi-square contributions Grammar I	Grammar II	Noun-phrase
1436	24.5	8	PE
835	9.1	1.7	PD
792	41.3	.7	DN NC
324	312.1	2.9	NP
284	3.2	7.8	IN NC
245	683.6	.001	EP NC
171	232.9	.6	PR
167	413.6	.001	NC
Totals (in the above sample)			
4254	1720.3	21.702	

There are other types not so well handled by Grammar II, but among the high-frequency types shown in Table 18, which account for 4254 noun phrases out of 5467 in the corpus, it is clear that Grammar II is the better fit.

We now turn to the geometric fit of Grammar II. Table 19 gives the parameters of that fit. (This table corresponds to Table 10 of Section IV.)

When we use geometric distributions for each group of production rules, Grammar II again has a better fit than Grammar I in relation to the highest frequency noun-phrase types. With geometric parameters, Grammar II has a chi-square sum of 4956 (24 degrees of freedom) while Grammar I has a chi-square sum of 6769 (35 degrees of freedom). Perhaps more significantly, Grammar II has a residual of 105 compared to a residual of 695 for Grammar I.

In all of the tables we have indicated not only the chi-square for the fit of the language generated by the grammar to the actual corpus, but we have also indicated the magnitude of chi-square divided by the number of net degrees of freedom. We have used this method as a rough and ready way of comparing the fit of different grammars with different numbers of degrees of freedom. For example, in the case of the geometric distribution we have 24 degrees of freedom rather than 9 in the case of Grammar II with the regular distribution. However, when we look at the ratio of chi-square divided by degrees of freedom we see, without even consulting a significance table, that in spite of the greater degrees of freedom the fit of the general Grammar II is better than the geometrically constrained grammar (166 compared to 207). We emphasize that we are not actually attaching a statistical significance to this ratio of chi-square to degrees of freedom. It is introduced for the purpose of easy, descriptive comparison.

Probabilistic lexical disambiguation, Grammar II. As in the case of Grammar I we used probabilistic methods to resolve the lexical ambiguity remaining in the corpus. There were 15 types (60 tokens) that were recognized with lexical ambiguity by Grammar II. In Table 20 we give the results of the probabilistic disambiguation (Table 20 is similar to Table 11 in Section IV).

Grammar II and the sections of the noun-phrase corpus. Table 21 shows the comparison of the noun phrases in the three sections of the noun phrases that were used in this study. (Section I is Sessions 1 through 3, II is 14 through 16, and III is 31 through 33.) We have applied the push-down store model with a geometric parameter for each group of rules to each of the sections. Our original intention was to look at the changes in the geometric parameters over time as evidence of development. Because the corpus itself does not show striking differences of a systematic kind over the three sections, we have not found any striking developmental trends at the more abstract level of the geometric parameters.

TABLE 19
Parameters for Geometric Fit of Grammar II

	Observed	Expected	Rule
(1,1)	4495	3865.9	SN => GN'
(1,10)	245	1001.2	SN => EP NC
(1,4)	167	259.3	SN => NC
(1,2)	153	67.2	SN => GN EP GN
(1,11)	102	17.4	SN => EP GN'
(1,3)	31	4.5	SN => GN CC GN
(1,7)	27	1.2	SN => GN KO
(1,5)	16	.3	SN => CC GN
(1,9)	1	.1	SN => GN LC
Theta =	.7410		
(2,1)	300	300.0	GN => GN'
(2,2)	72	72.0	GN => NC
Theta =	.8065		
(3,10)	1534	1845.6	GN' => D NC
(3,1)	1451	1150.0	GN' => PE
(3,2)	868	716.6	GN' => PD
(3,9)	389	446.6	GN' => NP
(3,3)	175	278.2	GN' => PR
(3,7)	172	173.4	GN' => D NP*
(3,4)	152	108.0	GN' => PI
(3,5)	79	67.3	GN' => NP*
(3,8)	27	41.9	GN' => ART POSTART
(3,11)	25	26.1	GN' => D NC AD
(3,6)	15	16.3	GN' => DN PO
(3,12)	10	26.9	GN' => D NP
Theta =	.3769		
(4,10)	1563	1449.4	D => ART
(4,9)	82	242.7	D => AD
(4,4)	46	40.6	D => POSTART
(4,7)	31	6.8	D => ART POSTART
(4,8)	13	1.1	D => PREART ART
(4,11)	6	.2	D => PREART
Theta =	.8325		
(5,4)	1057	1125.9	ART => DN
(5,3)	434	350.1	ART => IN

Table 19 (following and end):

```
(5,1)      131       108.9          ART => AO
(5,2)       12        49.1          ART => AT

Theta =       .6890

(6,3)       55        55.0          POSTART => DT
(6,1)       49        49.0          POSTART => NUM

Theta =       .5288

(7,1)       32        32.0          NUM => AC
(7,2)       17        17.0          NUM => NU

Theta =       .6531

(8,1)       13        13.0          PREART => QA
(8,2)        6         6.0          PREART => QR EP

Theta =       .6842

(9,2)      191       198.5          NP* => ADJP N
(9,3)       55        41.5          NP* => N POSTADJP
(9,1)        5        11.0          NP* => ADJP N POSTADJP

Theta =       .7910

(10,1)     196       196.0          ADJP => AQ
(10,2)       5         5.0          ADJP => ADJP AQ

Theta =       .9751

(11,1)      60        60.0          POSTADJP => AQ
(11,2)       1         1.0          POSTADJP => POSTADJP AQ

Theta =       .9836

(12,1)     249       249.0          N => NC
(12,2)       2         2.0          N => NP

Theta =       .9920
```

TABLE 20

Lexical Disambiguation of Noun Phrases in Grammar II

Freq	Ambiguous type	Resolved type	Prob
16	IN NC,NP un monsieur	IN NC	.064
13	DN NC,NP le monsieur	DN NC	.157
6	NC,PI personne	NC	.032
5	DN NC AQ,LC le bras gauche	DN NC AQ	.004
4	DN,PE DT,LC le même	DN DT	.002
3	EP NC,NP au monsieur	EP NC	.047
2	AQ,KO,NC NC,LC bon côté	AQ NC	.010
2	DN NC EP NC,NP la bouche du monsieur	DN NC EP NC	.0007
2	DN NC,NP KO le monsieur là	DN NC KO	.0008
2	DN,PE NC,LC le côté	DN NC	.16
1	AQ,LC AQ NC dernier petit trou	AQ AQ NC	.0003
1	DN NC CC DN NC,NP la dame et le monsieur	DN NC CC DN NC	.0001
1	DN NC EP IN NC,NP les chaussures d'un monsieur	DN NC EP IN NC	.0002
1	NP CC NC,NP mesdames et monsieur	NP CC NC	.00007
1	NU,AQ NC nulle part	AQ NC	.010

TABLE 21

Section Comparison of Noun-Phrase Corpus

Description	Section 1 Types	Section 1 Tokens	Section 2 Types	Section 2 Tokens	Section 3 Types	Section 3 Tokens
Total	136	1601	127	1536	215	2330
Recognized	104	1524	100	1491	144	2202
% Recognized	76.5	95.1	78.7	97.1	67.0	94.5
Ambiguous	83	641	79	394	142	688
% Ambiguous	61.0	40.0	62.2	25.7	66.0	29.5
Amb. recog.	66	602	61	372	105	630
% Amb. Recog.	79.5	93.9	77.2	94.4	73.9	91.6
Amb. resolved	59	589	54	357	96	598
Amb. reduced	5	11	3	6	6	16
Original Amb. factor	914		514		1088	
Parsed Amb. factor	852		476		905	
Reduced Amb. factor	15		15		32	
Types after consolidation	75		73		93	
Types after probabilistic disambiguation	68		67		84	

The parameters of the geometric distribution for the twelve groups of rules are shown in Table 22. The one thing to note is that the changes in the parameters across the three sections are not monotonic. We might, for example, expect for the more complex groups of rules that the geometric parameter would decrease in size because the variety of use of rules of a given type would increase with time, or put more exactly, the frequency of use would be more diverse. This is not the case for a number of groups of rules, as may be seen from a perusal of Table 22.

TABLE 22

Section Comparison of Geometric Parameters

Label	Observed	Expected	Rule
		Section I (Sessions 1-3)	
(1,1)	1247	1056.1	SN => GN'
(1,4)	97	324.2	SN => NC
(1,10)	60	99.6	SN => EP NC
(1,2)	51	30.6	SN => GN EP GN
(1,11)	47	9.4	SN => EP GN'
(1,3)	17	2.9	SN => GN CC GN
(1,7)	5	1.3	SN => GN KO
Theta =	.6930		
(2,1)	109	109.0	GN => GN'
(2,2)	32	32.0	GN => NC
Theta =	.7730		
(3,10)	595	548.5	GN' => D NC
(3,9)	221	334.1	GN' => NP
(3,1)	217	203.5	GN' => PE
(3,2)	201	123.9	GN' => PD
(3,5)	55	75.5	GN' => NP*
(3,7)	45	46.0	GN' => D NP*
(3,4)	37	28.0	GN' => PI
(3,8)	13	17.1	GN' => ART POSTART
(3,6)	9	10.4	GN' => DN PO
(3,3)	7	6.3	GN' => PR
(3,12)	3	9.9	GN' => D NP
Theta =	.3910		
(4,10)	619	592.1	D => ART
(4,7)	11	46.9	D => ART POSTART
(4,9)	4	3.7	D => AD
(4,4)	3	.3	D => POSTART
(4,8)	3	.0	D => PREART ART
(4,11)	3	.0	D => PREART
Theta =	.9209		
(5,4)	508	514.2	ART => DN
(5,3)	116	104.9	ART => IN

Table 22 (following):

(5,1)	22	26.9	ART => AQ
Theta =	.7959		
(6,3)	23	23.0	POSTART => DT
(6,1)	4	4.0	POSTART => NUM
Theta =	.8519		
(7,1)	4	4.0	NUM => AC
Theta =	1.0000		
(8,1)	3	3.0	PREART => QA
(8,2)	3	3.0	PREART => QR EP
Theta =	.5000		
(9,2)	79	81.8	NP* => ADJP N
(9,3)	20	14.9	NP* => N POSTADJP
(9,1)	1	3.3	NP* => ADJP N POSTADJP
Theta =	.8182		
(10,1)	80	80.0	ADJP => AQ
(10,2)	2	2.0	ADJP => ADJP AQ
Theta =	.9756		
(11,1)	21	21.0	POSTADJP => AQ
(11,2)	1	1.0	POSTADJP => POSTADJP AQ
Theta =	.9545		
(12,1)	100	100.0	N => NC
Theta =	1.0000		

Section II (Sessions 14-16)

(1,1)	1295	1128.7	SN => GN'
(1,10)	93	274.3	SN => EP NC
(1,4)	30	66.6	SN => NC
(1,11)	24	16.2	SN => EP GN'
(1,2)	20	3.9	SN => GN EP GN
(1,7)	11	1.0	SN => GN KO
(1,5)	10	.2	SN => CC GN
(1,3)	8	.1	SN => GN CC GN
Theta =	.7570		

Table 22 (following):

(2,1)	62	62.0	GN => GN'
(2,2)	15	15.0	GN => NC
Theta =	.8052		
(3,1)	427	534.1	GN' => PE
(3,10)	364	327.5	GN' => D NC
(3,2)	318	200.9	GN' => PD
(3,9)	98	123.2	GN' => NP
(3,3)	62	75.5	GN' => PR
(3,4)	50	46.3	GN' => PI
(3,7)	48	28.4	GN' => D NP*
(3,5)	8	17.4	GN' => NP*
(3,11)	3	10.7	GN' => D NC AD
(3,8)	2	6.6	GN' => ART POSTART
(3,6)	1	10.4	GN' => DN PO
Theta =	.3868		
(4,10)	343	320.4	D => ART
(4,9)	37	73.1	D => AD
(4,4)	22	16.7	D => POSTART
(4,7)	11	3.8	D => ART POSTART
(4,11)	2	1.1	D => PREART
Theta =	.7720		
(5,4)	213	226.1	ART => DN
(5,3)	91	82.5	ART => IN
(5,1)	46	30.1	ART => AD
(5,2)	6	17.3	ART => AT
Theta =	.6352		
(6,1)	23	23.0	POSTART => NUM
(6,3)	12	12.0	POSTART => DT
Theta =	.6571		
(7,1)	16	16.0	NUM => AC
(7,2)	7	7.0	NUM => NU
Theta =	.6957		
(8,2)	2	2.0	PREART => QR EP
Theta =	1.0000		
(9,2)	45	46.0	NP* => ADJP N
(9,3)	10	8.2	NP* => N POSTADJP
(9,1)	1	1.8	NP* => ADJP N POSTADJP

Table 22 (following):

```
Theta =       .8209

(10,1)      46        46.0          ADJP => AQ

Theta =      1.0000

(11,1)      11        11.0          POSTADJP => AQ

Theta =      1.0000

(12,1)      54        55.0          N => NC
(12,2)       2         2.0          N => NP

Theta =       .9643

                          Section III (Sessions 31-33)

(1,1)     1959      1740.9          SN => GN'
(1,10)      92       364.6          SN => EP NC
(1,2)       62        76.3          SN => GN EP GN
(1,4)       34        16.0          SN => NC
(1,11)      31         3.3          SN => EP GN'
(1,7)       11          .7          SN => GN KO
(1,3)        6          .1          SN => GN CC GN
(1,5)        6          .0          SN => CC GN
(1,9)        1          .0          SN => GN LC

Theta =       .7906

(2,1)      129       129.0          GN => GN'
(2,2)       25        25.0          GN => NC

Theta =       .8377

(3,1)      807       843.4          GN' => PE
(3,10)     575       507.7          GN' => D NC
(3,2)      349       305.6          GN' => PD
(3,3)      106       184.0          GN' => PR
(3,7)       79       110.7          GN' => D NP*
(3,4)       71        66.7          GN' => PI
(3,9)       70        40.1          GN' => NP
(3,11)      22        24.2          GN' => D NC AD
(3,5)       16        14.5          GN' => NP*
(3,8)       12         8.6          GN' => ART POSTART
(3,12)       7         5.3          GN' => D NP
(3,6)        5         8.0          GN' => DN PO

Theta =       .3980
```

Table 22 (following and end):

(4,10)	601	557,2	D => ART
(4,9)	41	102,6	D => AD
(4,4)	21	18,9	D => POSTART
(4,8)	10	3,5	D => PREART ART
(4,7)	9	,6	D => ART POSTART
(4,11)	1	,1	D => PREART
Theta =	,8158		
(5,4)	336	396,8	ART => DN
(5,3)	227	147,7	ART => IN
(5,1)	63	54,9	ART => AO
(5,2)	6	32,6	ART => AT
Theta =	,6279		
(6,1)	22	22,0	POSTART => NUM
(6,3)	20	20,0	POSTART => DT
Theta =	,5238		
(7,1)	12	12,0	NUM => AC
(7,2)	10	10,0	NUM => NU
Theta =	,5455		
(8,1)	10	10,0	PREART => QA
(8,2)	1	1,0	PREART => QR EP
Theta =	,9091		
(9,2)	67	71,1	NP* => ADJP N
(9,3)	25	17,9	NP* => N POSTADJP
(9,1)	3	6,0	NP* => ADJP N POSTADJP
Theta =	,7480		
(10,1)	70	70,0	ADJP => AQ
(10,2)	3	3,0	ADJP => ADJP AQ
Theta =	,9589		
(11,1)	28	28,0	POSTADJP => AQ
Theta =	1,0000		
(12,1)	95	95,0	N => NC
Theta =	1,0000		

TABLE 23

Noun-phrase Types Not Recognized by The Grammars

(Frequency of at least 2)

Not parsed by Grammar I		Not parsed by Grammar II	
Freq.	String	Freq.	String
14	ap aq	14	ap aq
13	in	13	in
12	in aq	12	in aq
11	dn,pa aq	11	dn,pa aq
10	ap nc ap nc	10	ap nc ap nc
10	in aq,pi,ai		
9	in dv	9	in qr,dv
8	aq ap nc	8	aq ap nc
8	dn,pa aq cc dn,pa aq	8	dn,pa aq cc dn,pa aq
7	ac	7	ac
7	ap nc cc ap nc	7	ap nc cc ap nc
7	dn,pa aq,pi,ai		
6	dv nc		
5	in ms	5	in ms
5	np np	5	np np
5	aq,pi,ai,dv pd	5	qa,aq,dv pd
4	in aq,pi,ai,dv aq dv	4	in qa,aq,dv aq,qr dv
4	nc nc	4	nc nc
		4	qa,aq,dv
		4	qr,dv nc
3	aq,pi,ai ao nc		
3	aq,pi,ai dn,pa nc		
3	aq,pi,ai dn,pa nc ap nc		
3	aq,pi,ai,dv dn,pa nc		
3	dn,pa ac		
3	dv ap nc		
3	ap nc ap,dv ap nc	3	ap nc ap,dv ap nc
3	in aq dv	3	in aq qr,dv
3	in ap pi,ai nc,le	3	in ap nu nc,le
2	ac aq	2	ac aq
		2	ad aq nc ad
2	aq,pi,ai dn,pa ac		
2	aq ap dn,pa nc	2	aq ap dn,pa nc
2	cc in ms	2	cc in ms
2	dn,pa nc ap pa,dv,ap	2	dn,pa nc ap pa,dv,ap

Table 23 (following and end):

	nc,dv,ep		nc,dv,ep
		2	dn,pe pr
2	dv ep nc,ep		
2	dv ic pr	2	dv ic pr
		2	dv nc
2	ep aq nc ep nc	2	ep aq nc ep nc
		2	ep dt
2	ep nc ep np	2	ep nc ep np
2	in aq cc in aq	2	in aq cc in aq
2	in aq nc ep,dv ep nc	2	in aq nc ep,dv ep nc
2	in aq,dv ep nc	2	in aq qr,dv ep nc
2	in ep nc	2	in ep nc
2	nc,np np	2	nc,np np
		2	qa,aq
		2	qa,aq dn,pe ac

Table 23 lists the noun-phrase types that were not recognized by the two grammars and that had a frequency of at least two in the corpus of the three sections. We shall not discuss in any detail methods for incorporating these types into the two grammars by appropriate extensions. The main reason for including Table 23 is to give the reader a sense of the main types that occurred in the corpus but are not parsed by the grammars.

ABSTRACT

This is the first of a series of reports concerned with the analysis of a young child's spoken French, which was collected in Paris in 1971 and 1972. The corpus covers the period when the child was 25 months old to 39 months old. The present article gives a formal generative grammar for the noun phrases taken from the first three hours, the middle three hours and the last three hours of the corpus. The analysis deals not only with the proportion of the noun phrases in the corpus parsed by the grammar, but also with the probabilistic fit of the grammar, given an assignment of probability parameters to the production rules of the grammar. The probabilistic analysis is also applied to problems of syntactic ambiguity.

RESUME

Cet article est le premier d'une série de travaux consacrés à l'étude du langage d'un enfant de langue française. Le recueil du corpus, entrepris lorsque l'enfant était âgé de 25 mois, s'est achevé lorsque l'enfant avait 39 mois. Cette étude présente une grammaire générative du syntagme nominal qui a été testée sur les trois premières. les trois dernières et les trois sessions d'enregistrement du milieu. L'analyse rend non seulement compte du nombre de syntagmes nominaux analysés, mais encore propose une méthode d'évaluation de la grammaire basée sur les paramètres probabilistes attachés aux règles de production. L'analyse en termes de probabilités est également appliquée à la résolution de l'ambiguïté syntaxique.

ZUSAMMENFASSUNG

Dies ist der erste Bericht aus einer Serie, die sich mit der Analyse des gesprochenen Französisch eines Kleinkindes beschäftigt. Das Material wurde 1971 und 1972 in Paris aufgenommen. Der vorliegende Artikel liefert eine Analyse der Nomenphrase aus den ersten, den mittleren und den letzten drei Stunden. Diese Stichprobe umfasst den Zeitraum vom 25, bis 39, Lebensmonat des Kindes.

Die Analyse untersucht nicht nur den relativen Anteil der Nomenphrase am Gesamtcorpus gemäss der Grammatik, sondern such die probabilistische Angemessenheit der Grammatik, wobei den Produktionsregeln der Grammatik Wahrscheinlichkeitsparameter zugeordnet sind.

Die Wahrscheinlichkeitsanalyse wurde auch auf Probleme syntaktischer Mehrdeutigkeit angewandt.

REFERENCES

DUBOIS, J. : *Grammaire structurale du français : nom et pronom*. Paris : Larousse, 1965.

DUBOIS, J. et DUBOIS-CHARLIER, F. : *Eléments de linguistique française : syntaxe*. Paris : Larousse, 1970.

GAMMON, E.M. : A syntactical analysis of some first-grade readers. Technical Report No. 155, June 22, 1970, Stanford University, Institute for Mathematical Studies in the Social Sciences. Published in K.J. Hintikka, J. Moravcsik and P. Suppes (Eds.), *Approaches to natural language*. Dordrecht, Holland : Reidel, 1973.

GREVISSE, M. : *Le bon usage. Grammaire française avec des remarques sur la langue française d'aujourd'hui.* Gemblouz, France : Duculot, 1969.

MARTINET, M.A. : De l'économie des formes du verbe en français parlé. In A.G. Hatcher and K.L. Selid (Eds.), *Studia philologica et litteraria in honorem L. Spitzer.* Bern, Switzerland : Francke, 1958.

ROBERT, P. : *Le petit Robert.* Paris : Société du nouveau Littré, 1967.

SMITH, R.L., Jr. : The syntax and semantics of ERICA. Technical Report No. 185, June 14, 1972, Stanford University, Institute for Mathematical Studies in the Social Sciences.

SUPPES, P. : Probabilistic grammars for natural languages. Technical Report No. 154, May 15, 1970, Stanford University, Institute for Mathematical Studies in the Social Sciences. Published in *Synthèse,* 1970, *22,* 95-116.

SUPPES, P. : Semantics of context-free fragments of natural languages. Technical Report No. 171, March 30, 1971, Stanford University, Institute for Mathematical Studies in the Social Sciences. Published in K.J. Hintikka, J. Moravcsik and P. Suppes (Eds.), *Approaches to natural language.* Dordrecht, Holland : Reidel, 1973.

Manuscript received July 1963.