Linguistic and Computational Problems for the Creation of an Italian Children's Corpus of Spoken Language

Laura Pecchia, Giuseppe Cappelli, Elisabetta Guazzini
Istituto di Linguistica Computazionale, CNR
Via Moruzzi, 1 - 56124 Pisa, Italy
laura.pecchia@ilc.cnr.it, beppe.cappelli@ilc.cnr.it, elisabetta.guazzini@ilc.cnr.it

Abstract

In this paper we describe the criteria adopted for the creation of a corpus of spoken language produced by children of six to eleven years of age in different communicative situations, the methodology used for the collection of data, the transcription, coding and lemmatization phases. We also give some quantitative descriptions about nouns, verbs and adjectives present in the corpus. Qualitative analyses on the adjectives are underway.

This work is to be included among the activities carried out within the framework of the "Corpus di Linguaggio Infantile" (C.L.I.), a special project of the Italian National Research Council (CNR).

1. Introduction

The development and rapid diffusion of computer technology have encouraged and improved considerably the growth of a great many disciplines, among which the study of language. In linguistics, a particularly important role has been played by the availability of large textual corpora to be used as important research resources.

In the last years the need for spoken corpora, as part of the general framework of language resources, has been worldwide recognized considering the advantages for both linguistic and more application-oriented research.

The collections of spoken language in Italy are scarce and are often restricted to adult language (De Mauro T. et al., 1993).

As for children's language, there are collections of different size concerning communicative situations which are specifically designed to study particular linguistic phenomena or collections of pathological language (IRCCS, Stella Maris, Pisa).

Standards for the creation of corpora already exist for children's written language (Marconi L. et al., 1994) while there are not, as far as we know, standards for children's spoken language. The creation of a spoken corpus represents a precious and particularly interesting tool which allows the study of the process of language acquisition and witnesses the lexical development of children in this age range.

2. The Corpus

This work describes one of the activities performed within the framework of the "Corpus di Linguaggio Infantile" (C.L.I.), a special project of the Italian National Research Council aimed at the creation of an Italian child oral language corpus produced by children between six and eleven years of age in different communicative situations.

The size of the corpus was fixed in the number of 500,000 occurrences which seem to ensure (Greenbaum S. & Svartvik J., 1990) the significance of the different analyses the researcher will perform.

We chose the public school as reference point; the subjects examined were subdivided into five age groups and the children were recorded separately class by class. We tried to have the same number of males and females, and to represent both urban and suburban, full-time and part-time schools.

In order to provide appropriate linguistic variety the children were assigned different tasks chosen according to a careful study of the syllabus and programs of the Ministry of Education for this age range. It was important to choose activities which reflected as much as possible the various interests and 'experiential fields' of the children at the different ages. There was a strong collaboration and interaction with the teachers, with whom we discussed and identified certain activities already included in the curriculum of each class.

The activities proposed were considered according to a variety of oral language usages: to exchange information in conversation, to share stories, to retell familiar stories or fables, to share television programs or movies, to express needs and feelings, to evince comments or questions, etc..

Conversation, narration and description offered the opportunity to examine the productive language skills, such as vocabulary, syntax and semantics as well as to maintain a topic or story structure, providing the beginning, middle and ending of a story; relating story characters; ordering of events, etc..

The corpus is subdivided into the following sectors:

1) *Story Telling* (Sector U[1]): the activities proposed in this sector include stories invented or prompted by picture books (UA); account of classical or familiar stories, books or comic-strips (UB); account of animated cartoons or films (UC).

2) *Narration of past events* (Sector D[1]): this sector includes the narration of different types of children's personal experiences such as excursions, parties, anniversaries, visits to exhibitions, museums, etc. (D).

3) *Descriptions* (Sector T[1]): the descriptions concern outdoor environments such as garden, district, beach, etc. (TA); indoor environments, for example, bedroom, school, gym, etc. (TB); people or animals: relatives, friends, television characters, pets, etc. (TC).

[1] Sectors U, D, T, Q and C stand respectively for Uno (one), Due (two), Tre (three), Quattro (four) and Cinque (five).

4) *Explanations and hypotheses* (Sector Q[1]): this sector comprises activities in which the children were requested to provide explanations regarding several topics, as for example the functioning of objects like washing machines, telephones, kites, etc. or the description of a recipe, a cake, pizza, spaghetti, etc. (QA); to make hypotheses on natural phenomena such as lightning, wind, rain, etc. (QB); explanation of the rules of a game, for example, Monopoli, basketball, football, etc. (QC).
5) *Conversations on different topics* (Sector C[1]): this sector includes conversations on desires and projects, like becoming a millionaire, going on a journey, receiving a present, etc. (CA); feelings like fear, happiness, sadness, etc. (CB); preferences about food, clothes, seasons, etc. (CC).

It is important to underline that the children's productions were not assessments of the school activities nor were they prepared in advance but were personal elaborations of the different topics.
The child was generally allowed to choose, according to his preferences, a particular activity which could be story telling, description or conversation.
The children's entire linguistic production was recorded without time limits and all the recordings were transcribed and coded. The productions which were too long were cut to mantain an average length with other children's productions within the considered sector; for instance in Sector U we seeked for the beginning, the development and the end of the story and the parts which were not strictly relevant to the story and its development were not considered.

3. Data Collection

A fundamental role for the collection of the data was played by the school, chosen as reference point both because it was the only environment in which it was easy to collect large amounts of data, and because it is a language intensive ambience where children spend most of their time engaging in different types of language-learning activities and in a variety of speaking situations.

The data were collected in the town of Pisa and province, with extremely useful results which led us to vary some original planning hypotheses and to adjust some procedures along the way. Eleven schools[2] participated in the project, for which the production of 834 children was transcribed for a total of about 70 hours of recording and for a total of about 160,000 occurrences.

3.1. Methodology for the collection

Interaction established with the children was basic to language collection whose main goal it was to obtain the most spontaneous language. In order for the productions to be as natural and spontaneous as possible, it was decided to spend at least one hour in each classroom before the beginning of the activity, thus allowing the observer to familiarize with the children, to observe them and try to understand the type of approach to be adopted during the recordings.

A protocol was prepared for each activity in order to make collection of the data homogeneous and independent of the person recording the data.
The protocol contains: the aim of the activity; the teacher's role; the material needed; suggestions for the presentation of the activity and finally indications are given for a correct performance of the activity itself.
This protocol was given to each teacher together with a number of forms corresponding to the number of children in each class. Each form contains on the top the name and code of the chosen activity and is divided into three parts; in the first part the observer is to write the name of the school, class, section, date of recording. The second part contains information about each child (sex, age, family, etc.) and the number occupied by the child in the class-register. We intentionally avoided to use the family name in order to ensure total privacy of the children and of their productions. Finally, the observer had to compile the last part of the form, containing specific features, as for example the child's behaviour – whether calm, anxious, etc. - during the activity, as well as particular words or expressions which were difficult to understand, external interruptions, etc..
All the productions were appropriately recorded using a digital audio-tape recorder assuring good quality recordings and therefore reliability of the transcriptions.
The children participated in the activities with willingness and great enthusiasm; they realized that the activities were somehow different from the regular routine and after the first task they often asked to perform another one. However, each child was involved in a maximum of two activities belonging to the different sectors in order to obtain as much variety of production as possible.
We wish to underline that extremely important and necessary for the collection of the data was the collaboration with the teachers, who participated in the different planning phases of some of the activities of the project and in many cases helped us with the recording of the data.

4. Transcription and Coding

Speech transcription is a difficult, lengthy and economically heavy task, by no means insignificant. One is faced with the problem of what and how to transcribe the various types of oral production such as intonation, pause, correction, repetition, turn-taking, overlaps, etc..
The Child Language Data Exchange System (CHILDES) project, proposed da B.MacWhinney and C.Snow (1985), was originated by the strong need in the world of child language research to dispose of large quantities of uniformly transcribed data and common computational tools aimed at checking linguistic hypotheses and comparing the results obtained with a vast numbers of researchers.
The main goals of the project consisted in the creation of a simple, flexible and complete system, able to create a standard in the coding of data transcriptions concerning child language; the implementation of a package of easy-

[1] Sectors U, D, T, Q and C stand respectively for Uno (one), Due (two), Tre (three), Quattro (four) and Cinque (five).
[2] Alighieri, Collodi, De Sanctis, Filzi, Frati Bigi, Genovesi, Gereschi, Moretti, Oberdan, Pascoli, Rosati.

to-use programs, promoting the automatic analysis of this data; the availability of an environment encouraging the exchange of data and standardized transcriptions among the world researchers.

We decided to transcribe and encode the data so far collected in CHAT (Codes for the Human Analysis Transcripts), which is the transcription system envisaged by CHILDES. This system uses a set of conventions which symbolize spoken and not spoken messages within the context of communicative interactions.

The choice of CHAT was made essentially taking into account: clarity (each symbol used for coding has a clear and definable real world referent); systematicity (codes, words and symbols are used consistently across the transcripts); readability (a variety of CHAT options is available in order to allow the users to maximize the readability of a transcript); reliability, flexibility and extensibility (the system is largely used and therefore very well tested; the user can introduce and integrate any particular code which is not present in the system).

CHAT is provided with three main parts: headers, main lines and dependent tiers.

The headers provide the reader with important information (participants, subject examined, date of birth, sex, etc.) concerning transcription and are generally made according to the researcher's requirements. For our purposes we add to the header lines: family, school and class so as to be able to perform specific analyses; we also introduced transcriber and coder in order to have a double check on the data.

The main lines contain the transcription of the participants' production.

Dependent tiers, placed below the main line, contain codes, comments and all the information relevant to the dialogue which can be useful to the researcher.

In the transcription phase it was particularly important to highlight all the 'special' forms produced by the child. These were evidenced by the symbol "@ " used together with one or two letters at the end of a word, to distinguish and categorize that particular word form. This was very useful because some word forms expressed by the children were not mistakes, but they were completely different from standard and therefore needed to be indicated with special markers.

Some of the most meaningful markers used for the transcription and coding of the corpus follows:

@d dialect form	ghiozzo@d	(for rude)
@f family form	bombo@f	(for drinking)
@fp filled pause	e@fp	(for and)
@i interjection	hum@i	(for hum)
@n neologism	appendigiubbotti@n	(for clothes-tree)
@o onomatopoeia	baubau@o	(for bow-wow)
@s 2nd language form	yes@s	

Each oral production included in C.L.I. is a file distinguished by eight characters: the first two represent the code of the school, the third and fourth refer to the number of the child in the school register, the fifth and sixth regard the type of activity and the last two the class number.

4.1. Grammatical tagging

Lemmatization consists in manual or automatic processing through which each form is reconducted to its relevant lemma, at the same time providing information on part of speech.

The Child Language Analysis (CLAN) programs of the CHILDES project are used for analysis of the data encoded in CHAT. These tools make it possible to perform frequency counts, context extractions, median length utterance calculations, search through Boolean operators, etc.. The use in succession and/or combination of the different CLAN programs allows various types of detailed analyses (lexical, morphological, syntactic). However, these analyses are possible only if the corpus has been lemmatized. This is a rather burdensome operation for any inflected language like Italian. Unfortunately, the use of automatic analyzers for the disambiguation of homographies, largely present in Italian, has not yet provided satisfactory results for the analysis of vast quantities of variable data like those of spoken language.

4.2. Method of analysis

We decided to perform lemmatization using AyDA[3] (Analizador y Desambiguator Automatico), a system for the linguistic tagging of data which disambiguates functional homographs automatically.

AyDA uses two dictionaries: a reference dictionary of around 80,000 forms obtained from the Frequency Dictionary of Italian Words by Juilland A. and Traversa V., and a children's dictionary list containing the forms produced by the children interviewed during the different activities. This dictionary has been increased since the beginning of the analysis and now amounts to about 7,000 forms.

The main modules of AyDA are: MORPH and MDS.

MORPH performs linguistic tagging and is able to process data coded in CHAT and the obtained output can be processed by CLAN programs. The application of the MORPH module produces, below the main line containing the children's production, a secondary line reporting the linguistic tagging(s) for each word. The program is able to identify and treate the "non-standard" forms (erroneously pronounced words, children's typical words, dialect forms, etc.), reconducting them to the standard form.

MDS performs the disambiguation of functional homographs; it relies on a transition matrix (which is a repository of syntactic structures collected during the

[3] The system AyDA was implemented in collaboration with the Instituto de la Lengua Espanola del Consejo Superior de la Investigaciones Cientifica (CSIC). The joint research project "Analisi Morfologica del Linguaggio Infantile: Creazione di un Corpus bilingue Italiano-Spagnolo", falls within the framework of the scientific agreement between CNR and CSIC. The researchers collaborating to this project are Cappelli G. e Saba A. for CNR, Albalà Hernandez M.J., Marrero V. and Torres I.M. for CSIC.

lemmatization performed on the texts) and on some specific rules. Each structure has a number which indicates the most likely disambiguation. MDS chooses a morphological classification, not necessarily the right one, using a statistical method.

The MDS output has been checked manually and where necessary completed.

As for lemmatization, in order to facilitate and speed up the following phase of manual checking, MORPH uses the children's reference list first then it consults the reference dictionary when it is unable to identify a word.

The use of a children's list makes it possible to exclude from lemmatization particular forms which are difficult to find in child's language (technical words, archaic forms, etc.). At the end of this work a children's frequency list will be available thus reducing as much as possible the use of the Italian form list.

5. Lemmatization

A tag was associated to each word form according to its part of speech and the main parts of the sentence were considered. Some special tags for further distinction inside the grammatical classes or otherwise particular lexical sets were added for more detailed future studies.

We have considered: nouns, verbs, adjectives, adverbs, conjunctions, articles, prepositions, pronouns, enclitic forms, dialectal forms, neologisms, interjections.

So far we have lemmatized 405 files out of a total of 835. In particular 3,200 words were chosen for each sector and for the five classes for a total of about 85,000 words.

Lexicographical reference for our work was the Vocabolario della Lingua Italiana edited by Zingarelli N. (1995).

6. Some quantitative descriptions

The following pages contain some tables with the figures relevant to the verbs, nouns and adjectives present in the corpus.

These data refer to the quantitative analyses carried out on the 85,000 lemmatized words for which more precise and detailed studies will be conducted in the future. The data which appear were calculated according to the various activities proposed in the five classes; the total figures for each sector are also reported.

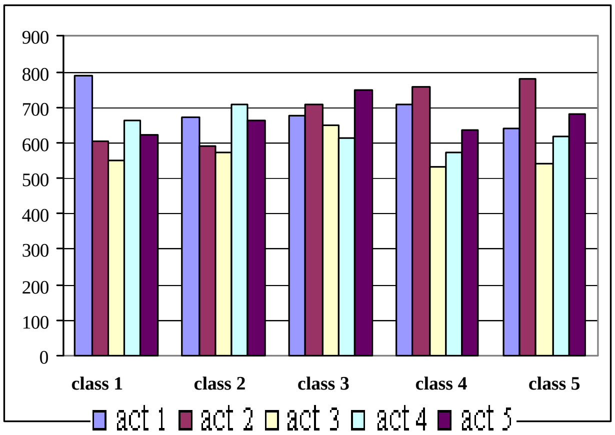

Figure 1 verbs

	class 1	class 2	class 3	class 4	Class 5
1a	253	241	233	223	240
1b	258	201	217	224	195
1c	277	231	227	260	204
	788	673	677	707	639
2	603	590	706	756	782
3a	172	177	206	194	178
3b	177	176	209	186	149
3c	201	221	233	151	215
	550	574	648	531	542
4a	222	220	205	226	186
4b	214	225	185	124	197
4c	228	264	222	224	234
	664	709	612	574	617
5a	208	234	259	206	234
5b	241	219	242	240	216
5c	173	209	246	188	230
	622	662	747	634	680

Table 1 verbs

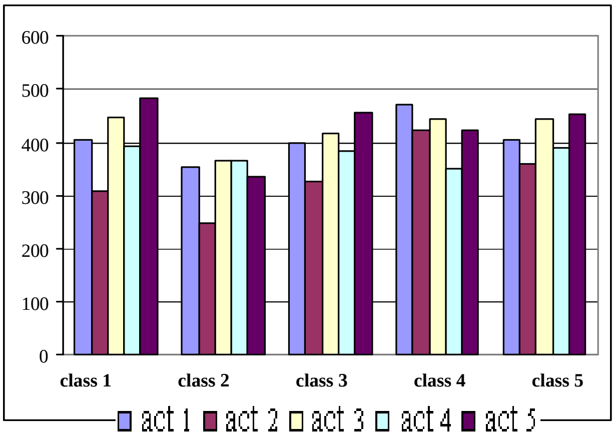

Figure 2 adjectives

	class 1	class 2	class 3	class 4	class 5
1a	121	109	126	178	154
1b	159	134	137	152	127
1c	124	112	135	141	123
	404	355	398	471	404
2	308	248	327	423	358
3a	159	150	112	150	162
3b	94	97	156	158	117
3c	193	120	148	135	165
	446	367	416	443	444
4a	112	110	129	149	129
4b	154	133	137	88	151
4c	127	124	117	113	110
	393	367	383	350	390
5a	143	104	160	188	160
5b	192	119	156	121	176
5c	148	111	139	114	118
	483	334	455	423	454

Table 2 adjectives

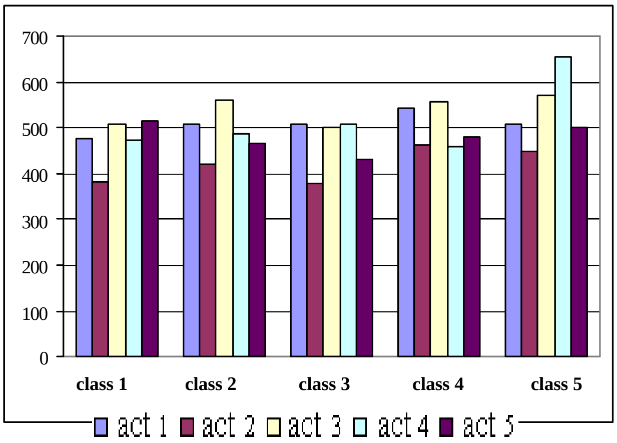

Figure 3 nouns

	class 1	class 2	class 3	class 4	Class 5
1a	166	200	184	208	151
1b	164	182	168	169	169
1c	147	128	158	167	187
	477	510	510	544	507
2	381	420	380	462	449
3a	191	219	190	215	175
3b	175	194	181	209	223
3c	141	149	132	133	172
	507	562	503	557	570
4a	161	175	154	181	232
4b	173	159	202	99	237
4c	141	152	151	181	187
	475	486	507	461	656
5a	175	144	145	157	160
5b	160	155	137	156	182
5c	182	166	150	166	158
	517	465	432	479	500

Table 3 nouns

Tables 1, 2 and 3 show the subdivision of verbs, adjectives and nouns in the various activities of the five classes. Table 2 displays the data regarding all types of adjectives (qualificative, demonstrative, relative, etc.).

The captions placed before each table help the reader understand the trend of the various activities more clearly.

Table 4 shows the data referred to another type of analysis concerning the different activities carried out in the five classes: this is the type-token ratio that is the measure of the lexical diversity. It is computed dividing the number of the different words used by a speaker by the total number of words the speaker produces in the speech sample.

What follows are the legends relative to Fig.4:

the ✍ refers to the first class;
the ✍ refers to the second class;
the ↴ refers to the third class;
the ? refers to the fourth class;
the ✍ refers to the fifth class.

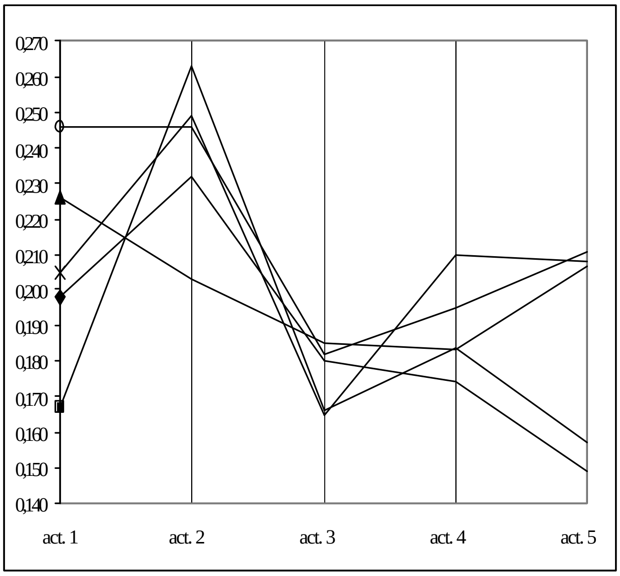

Figure 4 type/tokens

	class 1	*class 2*	*class 3*	*class 4*	*class 5*
sector 1	0,198	0,167	0,226	0,205	0,246
sector 2	0,232	0,263	0,203	0,249	0,246
sector 3	0,180	0,166	0,185	0,165	0,182
sector 4	0,174	0,184	0,183	0,210	0,195
sector 5	0,149	0,157	0,207	0,208	0,211

Table 4 type/tokens

7. Conclusions

Morphological-grammatical analyses will also be performed with particular regard to the lexical aspects. It will be interesting to discover the incidence of the different morphological categories, the use of verbal tenses, auxiliaries, etc., according to age, sex, and type of language.

From a lexical point of view, the information obtained will be useful to observe the moment in which new words are introduced, to evaluate the children's lexical competence, to follow their development and check the children's knowledge of words by automatic extraction of the contexts which make it possible to identify the familiar polysemies and to assess their correct use.

It will be interesting to create specific lists relevant to particular sectors of the lexicon, as for example animals, clothes, parts of the body, tools, etc. and it will be amusing to see the children’s naif beliefs concerning physical phenomena, functioning of tools, etc..

Further analyses of the data will allow identification of the features and developmental stages of language acquisition, while more detailed studies performed on the subsets of the corpus will evidence particular aspects or phenomena of the different age groups and/or specific types of language (narration, descriptions, conversations).

These analyses are only a few examples of the many possibilities offered by an encoded and lemmatized corpus of this type.

REFERENCES

Aarts, J. & Mejis, W. (Eds.). (1984). Corpus Linguistics, Recent Development in the Use of Computer Corpora in English Language Research, Amsterdam.

Aarts, J. & de Haan, P. & Oostdijk, N. (Eds.). (1993). English Language Corpora: Design, Analyses and Exploitation. Amsterdam - Atlanta GA, Rodopi.

Atwell, E.S. (Ed.). (1983). Corpus-based Computational Linguistics: Selection of Recent Papers. Oxford, Basil Blackwel.

Camaioni L. (Ed.). (1993). Manuale di psicologia dello sviluppo. Bologna, Il Mulino.

Camaioni L. (Ed.). (1978). Sviluppo del Linguaggio e Interazione Sociale. Bologna, Il Mulino.

Cook-Gumperz, J. & Corsaro, A. W. & Streeck J. (Eds.). (1986). Children's Worlds and Children's Language, Berlin, Mouton de Gruyter.

Bazzanella C. (1994). Le facce del parlare. Firenze, La Nuova Italia.

Bortolini, U. & Tagliavini, C. & Zampolli A. (1972). Lessico di Frequenza della Lingua Italiana Contemporanea. Milano, Garzanti.

Bungarten T. (1978). The Role Organization of a Corpus in Literary and Linguistic Research in General. In ALLC Bulletin. 6 1: 8-9.

De Mauro T. (Ed.). (1994). Come Parlano gli Italiani, Scandicci (FI), La Nuova Italia.

De Mauro T. & Mancini F. & Vedovelli M. & Voghera M. (1993). Lessico di Frequenza dell'Italiano Parlato. Milano, Etaslibri.

Fletcher P. & Garman M. (1986). Language Acquisition. Studies in the First Language Development, Cambridge, Cambridge University Press.

Greenbaum S. & Svartvikk J. (1990). The London-Lund Corpus of Spoken English. In J. Svartvik (Ed.), 11--63.

Harris M. & Coltheart M. (Eds.). (1986). Language Processing in Children and Adults. London, Routledge & Kegan Paul.

Juilland A. & Traversa V. (1973). Frequency Dictionary of Italian Words. The Hague.

Leech G. (1991). The State of the Art in Corpus Linguistics. In K. Aijmer & B. Altenberg (Eds.), English Corpus Linguistics (pp 8--29). New York, Longman.

MacWhinney B. (1995). The CHILDES Project, Tools for Analyzing Talk. Hillsdale, NJ: Lawrence Erlbaum Associates .

MacWhinney B. & Snow C. (1985). The Child Language Data Exchange System. Journal of Child Language, 12, 271--296.

Marconi L. & Ott M. & Pesenti E. & Ratti D. & Tavella M. (1993). Lessico Elementare. Bologna, Zanichelli.

Moore T.E. (Ed.). (1973). Cognitive Development and the Acquisition of Language. New York, Academic Press.

Oostdijk N. & de Haan P. (Eds.). (1994). Corpus-Based Research into Language. Amsterdam - Atlanta, GA.

Sinclair J.M. (Ed.). (1987). Looking up. An Account of the COBUILD Project in Lexical Computing. London, Collins .

Sornicola R. (1981). Sul Parlato, Bologna , Il Mulino.

Vygotskij L.S. (1990). Pensiero e Linguaggio. Bari, Laterza.

Wanner E. & Gleitman L.R. (Eds.). (1989). Language Acquisition. Cambridge, Cambridge University Press.