French Language Studies **22** (2012), 7–12, © Cambridge University Press
doi:10.1017/S095926951100055X

The Paris Corpus

ALIYAH MORGENSTERN[a] and CHRISTOPHE PARISSE[b]

[a] *Université Sorbonne Nouvelle – Paris 3*
[b] *MoDyCo, INSERM, CNRS, Université Paris Ouest Nanterre La Défense*

PRESENTATION OF THE CORPUS

The *Paris corpus*[1] was financed by the Agence Nationale de la Recherche, in the context of two research programmes entitled 'Acquisition du Langage et Grammaticalisation' (2005–2008, http://anr-leonard.ens-lsh.fr/) and 'Communication Langagière chez le Jeune Enfant' (CoLaJE, 2009–2012, http://colaje.risc.cnrs.fr). The aim of the two programmes was to collect new French data and add five new longitudinal corpora to the international database of the CHILDES project (http://childes.psy.cmu.edu/, MacWhinney, 2000), improve researchers' transcription and coding systems to enable them to study the emergence and development of grammatical patterns used by children between age one and seven, and compare child and adult speech. The programmes brought together specialists from various fields of language acquisition in order to study language development in the same longitudinal corpus from a multimodal and interdisciplinary perspective. The analyses aimed to find regularities in acquisition for each child and across the children.

For this Special Issue of the *Journal of French Language Studies*, all the authors were given the video recordings and transcriptions of the same four longitudinal corpora. The researchers chose to analyse either one or several children within the same data set according to their own field of competence.

THE CHILDREN

We focused the analyses for this Special Issue on four children from the *Paris Corpus*, two girls and two boys. ***Madeleine*** was filmed by Martine Sekali from age 0;10, ***Théophile*** by Aliyah Morgenstern from age 0;07, ***Antoine*** by Christophe Parisse from age 0;01 and ***Anaé*** by Aliyah Morgenstern and Marie Leroy from age 1;00. The four children are still being filmed and will be until age 7;00. The analyses in this volume are conducted up to 4;00. All the children live in Paris or in surrounding suburbs. They have middle-class college-educated parents, and were filmed at home about once a month for an hour in daily life situations (playing, taking a bath, having dinner). Madeleine has an older sister, and a brother was born during the course of the recordings. Théophile is a first-born child, and in

[1] The recordings and transcriptions can be downloaded from
- http://childes.psy.cmu.edu/data/Romance/French/, or
- http://colaje.risc.cnrs.fr/index.php/corpus/corpus-colaje.

the course of the recordings a brother and then a sister were born. Antoine is a first-born child and now has a younger brother. Anaé has two older brothers. The parents all worked throughout the data collection period; they used various forms of childcare when the children were young, and put the children in kindergarten when they were around three years of age.

THE TRANSCRIPTIONS

The recordings for ages 1;00 to 3;03 have already been transcribed and transcriptions up to age 5;00 are currently in progress. The new data will be given to CHILDES[2] at the end of the CoLaJE research program. The transcriptions were done in CHAT format,[3] thus enabling the use of CLAN software tools for analysing and searching the data (Mean Length of Utterance; word frequency; number of word types and word tokens; morphological categorisation; word and expression search).

The CHAT format enables transcribers to integrate various fields of information. The main symbols used in the examples given in the studies presented in this Special Issue are the following:

- @ followed by general comments on the situation of the session;
- * followed by three capitals to refer to the speaker (example CHI for the target child, MOT for the mother, FAT for the father, BRO for the brother, FRI for a friend of the family or the three first letters of their name, OBS for the observer, i.e. usually the person who is filming);
- % followed by a three letter code for secondary tiers (pho for phonetic transcription, act for action, gaz for gaze, sit for situation). Transcribers can use as many existing secondary tiers as they need from the user guide or create more codes.
- yy is used when a syllable or word cannot be identified but can be transcribed phonetically (it is then transcribed in the %pho line). yyy is used when the meaning of a longer string is not recognised by the transcriber.

Example:

- @Situation: CHI and MOT are seated at the table.
- *MOT: j'ai fait rouge. [I used red]
- %act: MOT draws.
- *MOT: c'est fini. [it's finished]
- %act: MOT takes her hand away from the paper. CHI violently shakes his head.
- *CHI: yy donne. [give]
- %pho: ma don
- %act: CHI tries to take the pen away from mother's hands.
- %gaz: CHI towards MOT

[2] See MacWhinney 2000.
[3] See Morgenstern and Parisse 2007, and Parisse and Morgenstern 2009 for justification and details on the choice of the transcription system.

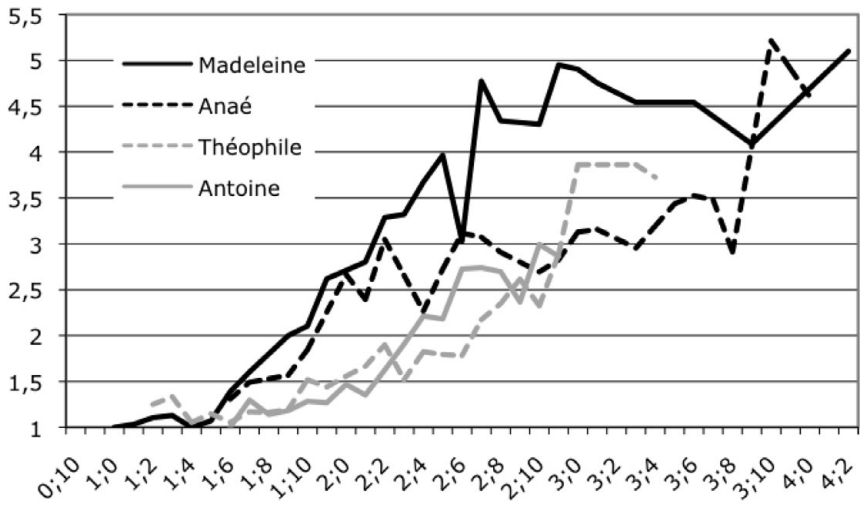

Figure 1. *Mean Length of Utterance per hour of recording according to age.*

INDIVIDUAL DIFFERENCES

Despite the fact that the four children come from middle-class families and all spend about the same amount of time with their parents and siblings after work, during weekends and vacations, their language development is quite different in many ways. The two girls are more precocious than the two boys. ***Madeleine***'s language development has been extremely fast: her phonological system was almost complete at 2;04. Her grammatical development, the increase in her vocabulary and the complexification of her utterances are extremely quick. Her logic and argumentation are quite advanced for her age. Her mother has treated her as a fully-fledged co-speaker from very early on. Her data has been studied extensively by researchers from the two projects[4] and various linguistic markers could be analysed in detail as early as the first transcription set (from 1;00 to 3;03). ***Anaé***'s language development is more varied: it has also been quite fast, but she often makes nonstandard productions such as gender mistakes (*un fleur* for *une fleur*/ a flower) and morphological creations (*elles sontaient* for *elles étaient* / they were) that provide clues about how she processes and analyses the input (Leroy, 2010). The mother and child have a very engaging relationship, with a lot of complicity and humour. ***Théophile***'s household is quite a fun place to be raised in, full of music and laughter. His language development was slow at first, but at 5;00, he has now become a talkative little boy and full of humour. ***Antoine*** is a cautious little boy,

[4] See for example Morgenstern (2006; 2009); Morgenstern and Sekali, 2009; Leroy et al., 2009; Mathiot et al., 2009.

https://doi.org/10.1017/S095926951100055X Published online by Cambridge University Press

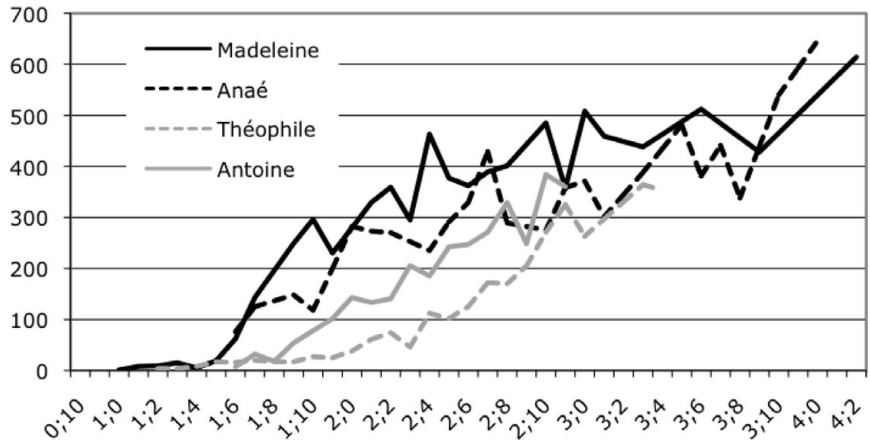

Figure 2. *Number of word types per hour of recording according to age.*

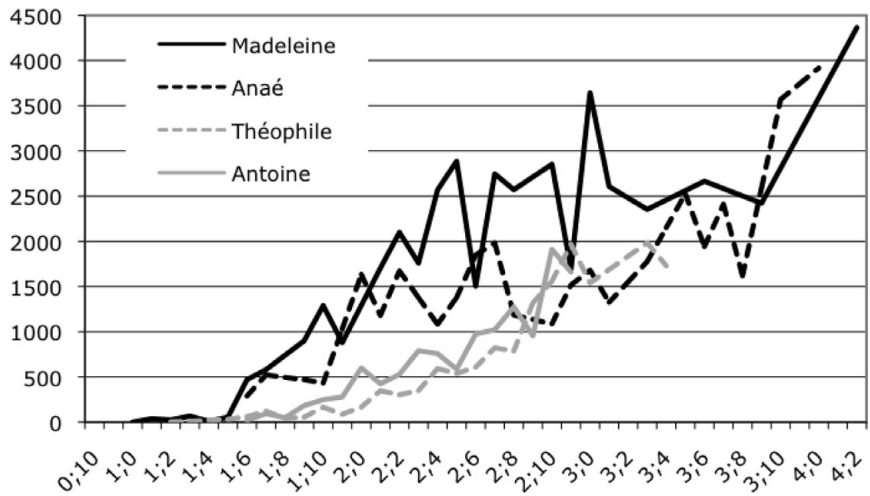

Figure 3. *Number of words per hour of recording according to age.*

whose productions are quite infrequent but varied and with few deviations from the target. He is very attentive to his interlocutors and his environment, and reacts with a lot of sensitivity and humour to his family's input.

Figures 1 to 4 show various objective measures used to compare the four children's language development, many of which are used in the analyses presented in this special issue: Mean Length of Utterance, number of word types, number of word tokens, and number of utterances according to age.

https://doi.org/10.1017/S095926951100055X Published online by Cambridge University Press

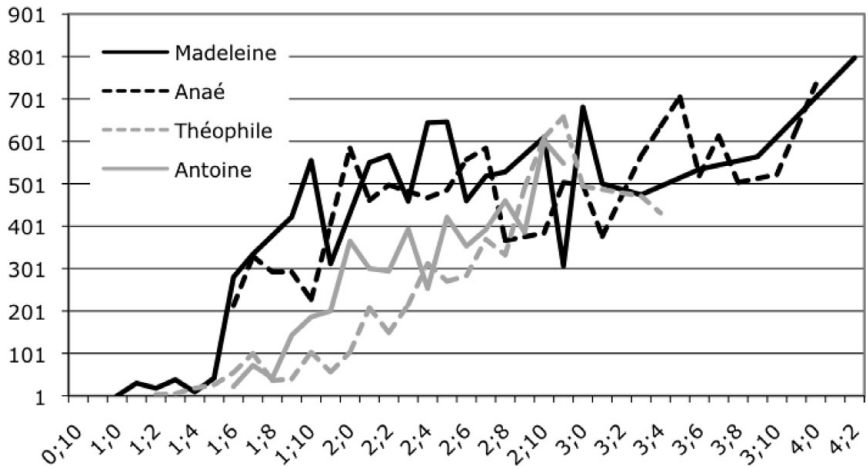

Figure 4. *Number of utterances per hour of recording according to age.*

The four measures shown in the figures reflect a large degree of variation across the different recording situations. Madeleine's language productions (black) are richer overall than the other children's, though Anaé (dotted black) is more talkative during certain recordings. Théophile's productions (dotted grey) are the least rich and numerous, but towards the end of the recordings, he seems to become more talkative (more utterances than Madeleine in some recordings, MLU higher than Anaé's at some points). Antoine (grey) is a very steady and measured speaker, but he seems to catch up with Anaé, at least from recordings around the age of 3;00 onwards. It will be very interesting to compare the four children between 3;00 and 5;00 to see if the rate of acquisition between 1;00 and 3;00 has any impact on the quantity and quality of their later productions.

These measures provide a general overview of the various children's language development. The quantitative and qualitative analyses conducted in this volume on various aspects of their grammatical development explore important specific features of their individual pathways towards full mastery of their target language.

Addresses for correspondence:
Aliyah Morgenstern
Université Sorbonne Nouvelle - Paris 3
5 rue de l'Ecole de médecine
75006 Paris
France
e-mail: Aliyah.Morgenstern@univ-paris3.fr

Christophe Parisse
Modyco, Bat A, Université Paris Ouest Nanterre La Défense
200 Avenue de la République
92001 Nanterre cedex
France
e-mail: cparisse@u-paris10.fr

REFERENCES

Leroy-Collombel, M. (2010). Eveil de la conscience grammaticale chez un enfant français entre 18 mois et 3 ans. In: F. Neveu, V. Muni Toke, J. Durand, T. Klingler, L. Mondada and S. Prévost (eds), *Congrès Mondial de Linguistique Française - CMLF 2010*. Paris, 2010, Institut de Linguistique Française.

Leroy, M., Mathiot, E. and Morgenstern, A. (2009). Pointing gestures and demonstrative words: deixis between the ages of one and three. In: J. Zlatev, M. Johansson Falck, C. Lundmark and M. Andrén (eds), *Studies in Language and Cognition*, Cambridge Scholars Publishing, pp. 386–404.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*, 3rd Edition. Vol. 2: *The Database*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mathiot, E., Leroy, M., Limousin, F. and Morgenstern, A. (2009). Premiers pointages chez l'enfant entendant et l'enfant sourd-signeur : deux suivis longitudinaux entre 7 mois et 1 an 7 mois. In: S. Benazzo (ed.), *Au croisement de différents types d'acquisition : pourquoi et comment comparer. AILE-LIA* 1: 141–168.

Morgenstern, A. with the collaboration of Benazzo, S., Leroy, M., Mathiot, E., Parisse, C., Salazar Orvig, A., Sekali, M. (2009). *L'enfant dans la langue*. Paris: Presses de la Sorbonne Nouvelle.

Morgenstern, A. and Parisse, C. (2007). Codage et interprétation du langage spontané d'enfants de 1 à 3 ans. *"Interprétation, contextes, codage"*, *Corpus* 6: 55–78.

Morgenstern, A. and Sekali, M. (2009). What can child language tell us about prepositions? A contrastive corpus-based study of cognitive and social-pragmatic factors. In: J. Zlatev, M. Johansson Falck, C. Lundmark and M. Andrén (eds), *Studies in Language and Cognition*, Cambridge: Cambridge Scholars Publishing, pp. 261–275.

Parisse, C. and Morgenstern, A. (2010). Transcrire et analyser les corpus d'enfant. In: E. Veneziano, A. Salazar Orvig and J. Bernicot (eds), *Acquisition du langage et interaction*. Paris: L'Harmattan, pp. 201–222.