THE INFLUENCE OF FREQUENCY ON WORD-INITIAL OBSTRUENT ACQUISITION IN HEXAGONAL FRENCH

Julia Monnin[1,2], Hélène Lœvenbruck[1] and Mary E. Beckman[3]

[1]EA Transcultures, Université de la Nouvelle-Calédonie, Nouméa, France; [2]ICP, Speech and Cognition Department, GIPSA-lab, Grenoble, France; [3]Ohio State University, Columbus, OH, USA
monninjulia@yahoo.fr; helene.loevenbruck@gipsa-lab.inpg.fr; mbeckman@ling.ohio-state.edu

ABSTRACT

This study extends a cross-linguistic collaboration on phonological development, which aims to compare production of word-initial obstruents across sets of languages which have comparable consonants that differ in overall frequency or in the frequency with which they occur in analogous sound sequences. By comparing across languages, the influence of language-specific distributional patterns on consonant mastery can be disentangled from the effects of more general phonetic constraints on development. In preparation for extending the comparison to Hexagonal French, we counted type frequencies in French databases and did a preliminary experiment with French-acquiring two-year-old children.

Keywords: Phonological development, obstruents, French, universal, language-specific.

1. INTRODUCTION

Two different types of trend can be observed during phonological development. First, some "phonetically difficult" sounds and sound sequences are acquired much later than the sounds observed universally in canonical babble, whatever the child's ambient language. For example, sibilant fricatives and affricates are mastered later than stops [1, 2], and dorsal stops are mastered later when the following vowel is /i/ as compared to /u/ [3]. Second, some of these "phonetically difficult" sounds and sound sequences are acquired relatively sooner in some languages than in others. For example, /v/ seems to be mastered earlier in Swedish, Estonian, and Bulgarian than it is in English [4]. Ingram [4] suggests this is because of the low type frequency of /v/ in English compared to the other cited languages. It seems reasonable to hypothesize that at least some language-specific differences in phonological development might be related to differences in phoneme and phoneme sequence frequency across languages in this way.

This study is framed within the Paidologos Project (http://www.ling.osu.edu/~edwards/), a cross-linguistic investigation of phonological acquisition. The aim is to compare the production of word-initial obstruents across many languages which contain sound sequences that differ in frequency. Children aged 2 to 5 were tested in the languages already studied (Cantonese, English, Greek, Japanese). The present study aims at extending the project to Hexagonal French. We first report type frequency data obtained on French and second the results of a preliminary experiment with French-acquiring 2 year old children.

2. FREQUENCY DATA ON WRITTEN AND ORAL FRENCH

Some studies suggest that there may be small but significant differences in some phoneme or phoneme-sequence frequencies between some registers of adult- directed speech and some styles of child-directed speech, in some languages [5]. Typical input to children may therefore differ from adult lexicons. Child-directed speech data on French is limited, however. We have therefore chosen to examine 3 different corpora: a large written corpus of adult-directed speech, a smaller corpus of oral adult-directed speech and a pilot recording of oral child-directed speech.

2.1. Corpora

The first corpus is the adult lexicon LEXIQUE 2 (http://www.lexique.org) which was obtained from 3200 written texts in French. LEXIQUE 2 contains 31 million items, from which a list of 130 000 orthographically distinct items have been derived. Phonetic transcriptions are provided for all items, as well as word class.

The second corpus is the adult lexicon LEXIQUE 3 which was obtained from written movie subtitles in French and therefore corresponds to spoken dialogues. LEXIQUE 3

contains 14.7 million items. Phonetic transcriptions and word class are available.

The last corpus consists of recordings of adults (more often the mother) talking to children. The first set of recordings was made by the first author, for 5 children aged 22-26 months, at each child's home. Eight recording sessions are available for a total of 4 hours 45 minutes. They include 9620 content word tokens.

The recorded data have been transcribed by a trained phonetician (first author) using CHILDES criteria (http://childes.psy.cmu.edu/). Adult utterances that were directed to another adult and not to the child have not been transcribed.

It should be noted that these recordings took place in Nouméa (New-Caledonia) where it can be suspected that a different dialect from Hexagonal French is spoken. However, the recordings took place in educated middle-class families whose dialects are close (if not identical) to Hexagonal French. The second author of the manuscript, who is from Hexagonal France, did not notice any particular accent in the data.

A second set of recordings is considered, which corresponds to similar data (child's age, adult participant). These recordings were extracted from the York Corpus of Child French [6]. Two recording sessions of respectively 2000 and 3000 tokens are considered. The child was between 22 and 23 months old. The adult participants were the father, mother and experimenter. Only orthographic transcriptions were available. An automatic phonetizer was used to obtain phonetic transcriptions [7], which were checked and corrected when necessary. The data from these 2 sets were grouped into what constitutes our Child-Directed Speech (CDS) database.

2.2. Corpus analyses

Function words are often reduced in French, therefore frequency was counted on content words or strong function words only. Determiners, unstressed prepositions (e.g. "de" (of)) were excluded. The remaining words include nouns, verbs, adjectives, adverbs, wh-question and relative clause words, stressed prepositions, strong and interrogative pronouns; exclamations, onomatopeio, referred to as "content words" hereafter.

Sandhi phenomena (such as liaison and enchainement) are frequent in French. The word "ours" (/urs/, bear) for instance may be differently resyllabified: un ours /ɛ̃.nurs/ (a bear), des ours /de.zurs/ (some bears); l'ours /lurs/ (the bear). Recent works suggest that these variants of the same word are all part of the input the child receives and memorizes [8]. We have therefore considered resyllabified versions of lexically vowel-initial words as CV-initial words.

Before choosing which CV sequences to use in comparing French to other languages in the Paidologos Project, we decided to examine all possible initial CV sequences. The 16 consonant categories (symbolized hereafter in WorldBet) are:

/p, b, t, d, k, g, f, v, s, z, S (=ʃ) , Z (=ʒ), l, r, m, n/.

Open and closed vowels that are frequently substituted for each other were grouped and nasal vowels are grouped with their oral counterparts, to make the following 7 vowel categories:

A: /a/, /ɑ/ and /ɑ̃/ O: /o/, /ɔ/ and /ɔ̃/

8 (oe): /ø/ and /œ/ E: /e/, /ɛ/, /ɛ̃/ and /œ̃/

u: /u/; i: /i/; y: /y/

To calculate type frequency for each CV sequence in each of the three databases, we used an awk script to count the number of words which contained the sequence in word-initial position and we divided this number by the total number of "content word" types in the database (128 891 in LEXIQUE 2, 137 385 in LEXIQUE 3 and 1 108 in CDS). Using this ratio allows us to correct for the different sizes of the three sets of databases.

Figs. 1 and 2 show the log frequencies of each consonant (pooling over all following vowels) and of each vowel (pooling over all preceding consonants) in the CV-initial content words in each of the three corpora. The three different bar patterns are for the three different corpora.

Looking first at the results for LEXIQUE 2 (the left-most bars in each set), the results show that the six highest-frequency initial consonants are /r/, /d/, /k/, /s/, /p/, /m/. The lowest frequencies are observed for /z/, /n/, /Z/, /S/ and /g/. It can be noted that, word-initially, /k/ is much more frequent than /t/ or /p/, and /p/ is more frequent than /t/. The most frequent vowel contexts are /E/ and then /A/. The least frequent contexts are /u/ and /y/.

Turning now to the frequencies obtained for LEXIQUE 3 (the middle bars in Figs. 1 and 2), the six most frequent consonants are the same, in a slightly different order: /d/, /k/, /s/, /r/, /m/, /p/. The five least frequent consonants also are the same: /z/, /Z/, /n/, /S/, /g/. Again, /k/ is more frequent than both /t/ and /p/, and /p/ remains more frequent than /t/. The most frequent vowel contexts are /E/ then /A/. The least frequent are /8/, /y/ and /u/.

Figure 1: Log frequencies of consonants in CV-initial content words when all following vowel contexts are combined, in each of 3 corpora.

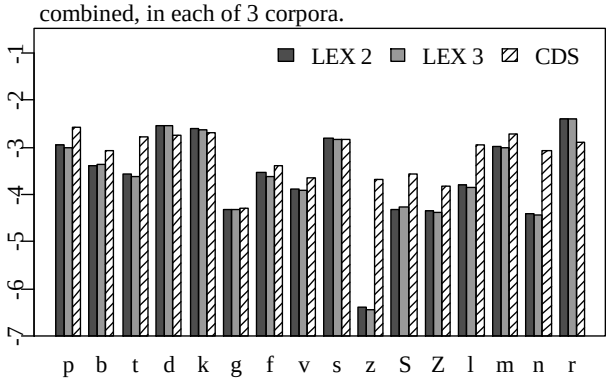
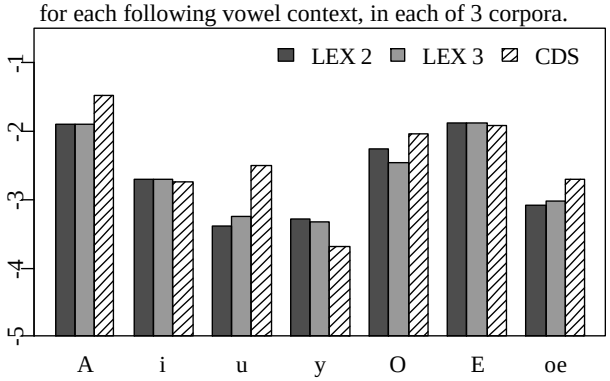


Figure 2: Log frequencies of the first vowels in CV-initial content words when all initial consonants are combined for each following vowel context, in each of 3 corpora.



Turning finally to the CDS database (the solid grey, right-most bars in each set), we see that the most frequent consonants are /p/, /k/, /m/, /d/, /t/, /s/, /r/. Again, /k/ is more frequent than /t/ but /p/ is more frequent than both /k/ and /t/. The five least frequent consonant contexts are slightly different from the adult databases: /g/, /Z/, /z/, /v/, /S/. The fact that /z/ is more frequent in the CDS database is probably due to the fact that resyllabification with liaison was taken into account in this oral corpus. The most frequent vowel contexts are /A/ then /E/. The least frequent context is /y/.

All our databases display similar trends for CV-initial words, namely that /r/, /s/, /d/, /p/, /m/ and /k/ are very frequent. One differences is that /t/ is less frequent in the adult databases but is quite frequent in the CDS. In all three databases, /g/, /Z/, /z/, /S/ are very infrequent; /y/ is a very infrequent vowel category in all databases whereas /E/ and /A/ are the most frequent (note that these two categories represent in fact more than one vowel).

The relative frequencies of /t/ and /k/ in CV-initial position are quite different according to the language. In Japanese, for instance, /k/ occurs far more frequently than /t/ both in the adult lexicon and in CDS [9]. In English, however, although /k/ is somewhat more frequent than /t/ in the adult lexicon, /t/ occurs more often than /k/ in CDS [9].

In all three of our French databases, /k/ is always more frequent than /t/. In the adult databases, /k/ is also more frequent than /p/, whereas in CDS /p/ is more frequent than /k/.

3. PILOT CHILDREN STUDY

3.1. Methods

In this pilot study, a reduced set of CV-sequences was examined, in order to draw comparisons with data for the other languages studied in the Paidologos Project. The consonants /t/, /s/, and /k/ were analyzed in the context of /a/, /i/, /y/, /u/. Complex consonants /tw/ and /kw/ were examined before /a/ and /i/. A list of 16 words containing all the CV combinations in initial word position was created. The words were chosen to be as familiar as possible. They contained one to two syllables.

Six children aged 20 to 24 months were recorded (one girl, five boys). They all lived in the Grenoble region (Hexagonal France), were all monolingual, with monolingual parents and had no known hearing deficits. The testing took place at the child's home, with the mother present. Each trial consisted of a picture and the associated sound pronounced by the experimenter. The pictures were presented over a laptop. The children were instructed to repeat each word exactly as they heard it. They wore a tie-clip microphone. When possible, the experiment was repeated a few times to gather several repetitions of the same word. The children's responses were recorded directly onto the computer using CoolEdit. To be able to track articulatory movements if necessary, video recordings were simultaneously made. The data from two children had to be discarded (one child refused to take part in the experiment, and the other one was still in the babbling-only stage).

A native speaker who is a trained phonetician listened to the responses and examined the acoustic waveforms. The target consonants were coded as *correct* or *incorrect*. Substitutions were noted. A second trained phonetician retranscribed part of the data. Inter-transcriber reliability was high.

3.2. Results

The mean accuracies for the 4 children are presented in Fig. 3. For all vowel contexts except /A/, /k/ and /s/ are less error-prone than /t/.

Figure 3: % correct for /s/, /t/ and /k/ in 4 vowel contexts

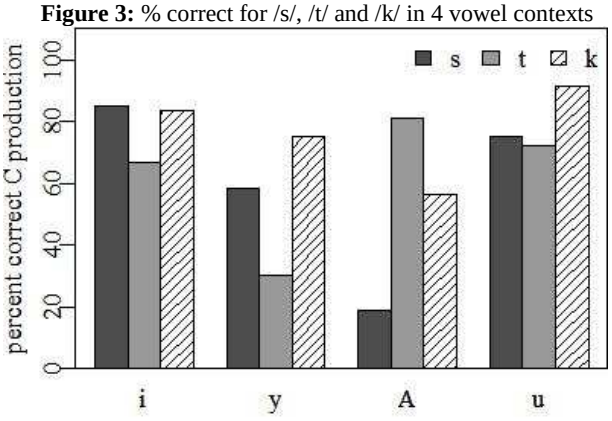


Figure 4: Log frequencies of CV-initial content words beginning with several CV sequences, in each corpus

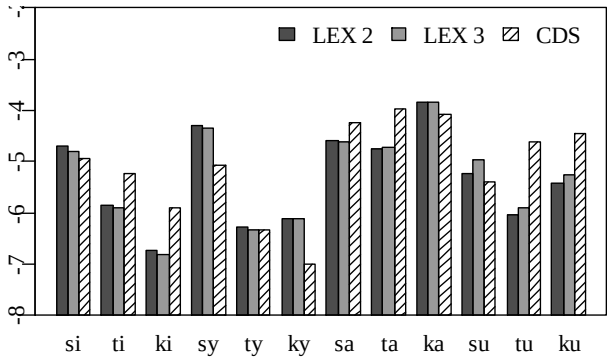
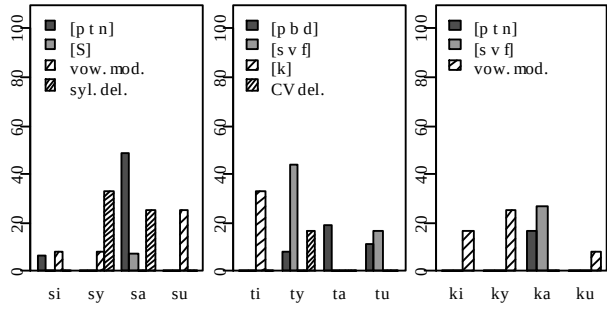


Figure 5: Error pattern for /s, t, k/



A comparison between CV frequencies (Fig. 4) and the children's accuracy rates on repetition (Fig. 3) shows that consonants in less frequent sequences tend to be less accurate. For example, /sa/ is less frequent in the input than /ta/, and children produce /s/ less accurately than /t/ in the context of /a/. Also, /ku/ is more frequent than /tu/ or /su/, and /k/ is more accurate than /t/ or /s/ in this context. The same pattern is observed for /si/.

Error patterns for each of the consonants are presented in Fig. 5. The figure shows that /k/ is never replaced by /t/, an error often made by English-acquiring children. On the contrary, /t/ can be replaced by /k/ and by fricatives.

4. DISCUSSION AND CONCLUSION

In this preliminary study of consonant accuracy in word-initial CV sequences produced by Hexagonal French acquiring children, /k/ and /s/ are shown to be accurately produced except in the vowel context /A/ where /t/ happens to be better produced. /k/ is shown to be less error-prone (in /A/, /i/, /y/, /u/ contexts) than /t/ and /s/. In addition, /k/ is never replaced by /t/ whereas /t/ can be replaced by /k/.

In all three of our databases, /s/ and /k/ are more frequent than /t/, and this probably explains the relatively high accuracy of the children's productions of these two consonants. The high accuracy rates are notable especially because /s/ is often considered to be "articulatorily difficult" and /k/ is a back consonant. These accuracy rates and error patterns differ from English (where /k/ is less frequent than /t/ in CDS) and are similar to Japanese and Greek, for which /k/ has been shown to be more frequent than /t/ in CDS. Although more data need to be collected, the present results seem to highlight the influence of the ambient language on phonological development.

5. REFERENCES

[1] Smit, A. B., Hand, L., Freilinger, J. J., Bernthal , J. E. & Bird, A. (1990). The Iowa articulation norms project and its Nebraska replication. *J. of Speech and Hearing Disorders*, 55, 779-798

[2] Hua, Z. & Dodd, B. (2000). The phonological acquisition of Putonghua (Modern Standard Chinese). *J. of Child Language*, 27, 3-42.

[3] Davis, B.L., MacNeilage, P.F. & Matyear, C. Acquisition of serial complexity in speech production: A comparison of phonetic and phonological approaches to first word production. *Phonetica*, 2002, 59, 75-107.

[4] Ingram, D. (1988). The acquisition of word-initial [v]. *Language and Speech*, 31, 77-85.

[5] Tserdanelis, G. (2005). *The role of segmental sandhi in the parsing of speech.* Unpublished doctoral dissertation, Ohio State University, Columbus.

[6] Plunkett, B. (2002). Null Subjects in child French interrogatives: A view from the York Corpus. In Claus D. Pusch, & Wolfgang Raible (Eds), *Romance corpus linguistics:Corpora and spoken language*, 441-452.

[7] Bailly, G. and M. Alissali (1992). COMPOST: a server for multilingual text-to-speech system. *Traitement du Signal* 9(4), 359-366.

[8] Chevrot, J.-P., Dugua, C. & Fayol, M. (2005). Liaison et formation des mots en français : un scénario développemental, *Langages*, 158, 38-52.

[9] Beckman, M. E., Yoneyama, K., & Edwards, J. (2003). Language-specific and language universal aspects of lingual obstruent productions in Japanese-acquiring children. *J. of the Phonetic Society of Japan*, 7, 18-28.