

## Phonological description of the Greek-speaking Children Corpus (GSCC)

Zoe Gavriilidou  
Democritus University of Thrace

Penelope Kambaki-Vougioukli  
Democritus University of Thrace

### 1 Introduction

Previous research in phonological acquisition has attempted to determine by what age a child should have mastered every phoneme. Groups of children were studied at succeeding ages and the number and percentage of children who had mastered a particular phoneme at each age level was determined. These results were used to create tables with recommended age standards in the acquisition of each phoneme.

Researchers agree on the general patterns and error patterns that occur in phoneme acquisition in English (Irwin and Wong 1983, Stoel-Gammon and Dunn 1985, Grunwell 1987) and several other languages such as Dutch, French or Spanish. For example, English-acquiring children generally produce labial and alveolar stops and nasals quite early, while production of most fricatives and affricates is acquired later (Prather *et al.* 1975, Smit, Hand, Freilinger, Bernthal and Bird, 1990, Kent 1992). In general, stops, nasals and glides tend to be mastered earlier than liquids, fricatives and affricates. According to Jacobson (1941) this is due to the distribution of these sounds among the world's languages. However, there is not much agreement on the age of acquisition of phonemes. For example, Prather *et al.* (1975) found that the English phonemes /z ʒ/ were acquired around the age of 4;0–4;6 by 75% of the subjects, while Stoel-Gammon and Dunn (1985) reported that these sounds are not mastered before 7;0 by all subjects.

A series of language specific issues, such as the influence of the frequency of a phoneme or neighbouring phonemes on a target-phoneme production, the role of syllable structure or accent in the acquisition of a phoneme and finally word length, have also attracted the scientific interest. For example, it was found that not only the frequency of the target phoneme, but also its presence in a high frequency CV, VC and CC sequence affect the time of mastery of consonants (Ingram 1988, Yoneyama *et al.* 2003, Munson *et al.* 2005). Arvaniti (2000) and Kehoe and Stoel-Gammon (1997) among others claimed that consonant accuracy is related with accented syllables. MacNeilage and Davis (1990) consider that CV syllables affect phoneme acquisition, because they do not exert motoric constraints. Finally, Edwards and Beckman (2008) reported a significant effect of word length to the accuracy of target phonemes.

A number of researchers have also focused on the phonological acquisition of Greek speakers (Arvaniti 2000, Kappa 2000, 2002, Nicolaidis 1994, 2002, 2003, Nicolaidis *et al.* 2004, Papadopoulou 2000, Tzakosta 2001, 2003, 2004, Sanoudaki 2009). In these studies, stress, consonant clusters and phoneme acquisition in various positions were thoroughly investigated.

In the past, diary studies—among others—were used to investigate phonological acquisition. Nowadays, research can rely on electronic linguistic resources accessible to the scientific community. By electronic linguistic resources we refer to electronic dictionaries, grammars, terminology data bases and corpora, oral or written. According to Sinclair (2004) *written corpora* are “collections of pieces of language text in electronic form, selected based on external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research”. On the other hand an *oral corpus* is a database of speech audio files and text transcriptions in a format that can be used to create acoustic models or provide data for linguistic research (Llisteri 1996). Oral corpora are divided in two further categories: (a) the read oral corpus, which include broadcast news, lists of words etc. and (b) the spontaneous oral corpus, which include dialogs, narratives etc. Each of them has its advantages and disadvantages. For instance, according to Edwards and Beckman (2008), using single word elicitation via picture-naming or word repetition tasks has the advantage of allowing the researcher to control the phonetic context and to know what the target is, while in natural speech samples a child may not produce all of the sounds of interest. However, natural speech samples “have the strong advantage of being ecologically valid, since children need to be able to talk intelligibly in connected speech about topics of interest to them and their conversational partners” (Edwards and Beckman 2008: 941).

It has to be mentioned that the increasing interest on oral corpora is the result of the influence, on the one hand, of experimental phonetics and on the other, of speech technologies and corpus linguistics (Llisteri 1996). The GSCC is a spontaneous oral corpus which includes children’ speech and was created at the Democritus University of Thrace.

## **2 Purpose and rationale**

The purpose of the present study is to analyze all cases included in the GSCC and try to investigate:

- (a) The percentage of children with or without phonological errors.
- (b) The types of phonological errors encountered in the GSCC.
- (c) The mispronounced phonemes.
- (d) The couples of phoneme substitutions (correct phoneme vs wrong phoneme which substituted a correct one).
- (e) The type of syllable in cases of substitutions (CCV, CVC, CV, V, VC).
- (f) The place of accent in mispronounced words.

### 3 GCSS characteristics

The GSCC includes interviews produced by sixty one children aged from three to six years old<sup>1</sup>. More precisely, 18 of the 61 interviews included belong to children aged from 3 to 4 years, 23 interviews belong to children aged from 4 to 5 years and 20 interviews belong to children aged from 5 to 6 years.

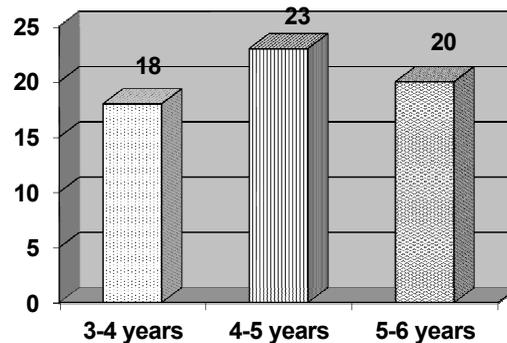


Figure 1. Age distribution

All of the children are Greek native speakers. The collection includes samples of children speaking standard Greek (N=47) or the Cypriot dialect (N=13). The geographical distribution of the sample includes the following cities: *Athens, Alexandroupolis, Drama, Edessa, Thessalonica, Kavala, Katerini, Kozani, Corinthus, Livadia, Viotia, Naousa, Orestiada, Trikala, Florina, Lemesos, Nicosia* and *Pafos*.

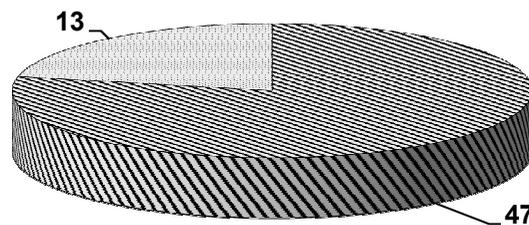


Figure 2. Geographical distribution

---

<sup>1</sup> In one case two children participated in the same interview, so we have sixty interviews and sixty one children overall.

36 of the children who participated are female and 25 are male.

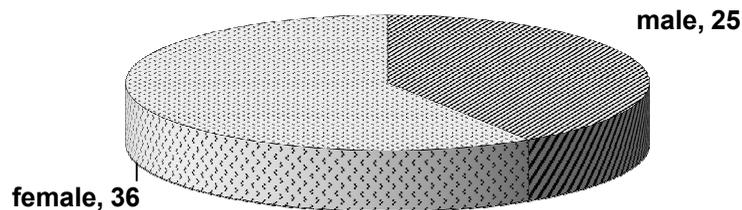


Figure 3. Gender distribution

The size of the corpus comes up to 151.380 words and the audio recordings are approximately calculated in 45 hours of speech.

### **3.1 GCSS compilation procedure**

The data of the GSCC had been collected by students of the Department of Education Sciences in Pre-School Age of the Democritus University of Thrace. Most of the interviews were held in surroundings familiar to the children and in all cases the interviewee (child) already knew the interviewer (student). In some cases, the parents of children or other relatives were present during the recording. The children knew that their voices were being recorded. The interviewer, in order to collect spontaneous production in a naturalistic setting (McEnery and Wilson 1996), asked the interviewee questions to elicit a dialogue and let children speak freely about what they wanted. Then the children were asked to tell a story or a fairy tale and finally used flash cards or pictures in order to elicit descriptions. Even though this method is sometimes criticized, if used to collect data only for phonological analysis (Edwards and Beckman 2008), it was chosen, because we thought it would allow children to talk in connected speech about topics of their interest, thus providing data for children's vocabulary, as well as semantic and/or pragmatic acquisition.

After the completion of this process the interviews were de-recorded and the oral texts were transcribed. Finally, the texts were transcribed in the International Phonetic Alphabet (IPA) phonemically. The process of collection began in 2004 and was completed in 2005.

### **3.2 Qualitative Characteristics of the GSCC**

The GSCC includes the basic vocabulary of children as well as stereotyped words and expressions that are usually used by children aged 3 to 6. Geographical variation of language is also included.

Each interview of the GSCC provides information about the language-level of the specific child and the level of communication that each child develops with adults. In general, the GSCC as a whole provides to researchers working in the area

of language development the possibility of studying the levels of language acquisition between the ages from 3 to 6 years.

## 4 Phonetic and phonological description of the GSCC

### 4.1 Data coding and analysis

A data base including 10 variables was created for the data processing. In these variables we coded (i) the subjects, (ii) their geographical origin, (iii) their gender, (iv) age (in months) and (v) age group (3–4, 4–5 and 5–6), (vi) the total of phonological mistakes produced by each subject, (vii) the distorted words and phonemes, (viii) the syllable type where each mis-produced phoneme was found, (ix) the place of accent in cases of mis-produced phonemes and (x) the kind of phonological error (substitution, omission, assimilation) or correct production for each phoneme.

Descriptive statistics, such as frequencies, were used to identify the phonological profile of the subjects. The sample characteristics did not allow gender and age comparisons.

### 4.2 Results

Descriptive statistical analysis of the GSCC revealed that 61,1% of the sample of children aged three to six presented at least one phonological error while only 38,9% had no phonological errors at all. The mean age of subjects who committed at least one mistake was 3;8 years old, while the mean age of children with no mistakes was 4;3. The most frequent phonological errors were substitutions (73,8%), omissions (19,6 %) and assimilations (3,5%) or metathesis (3%).

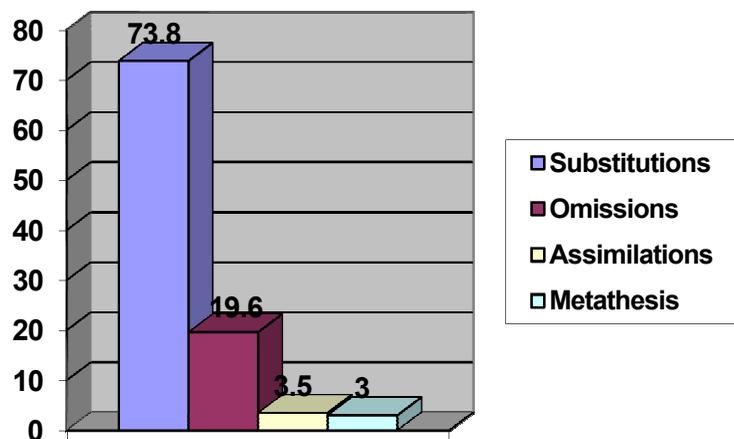
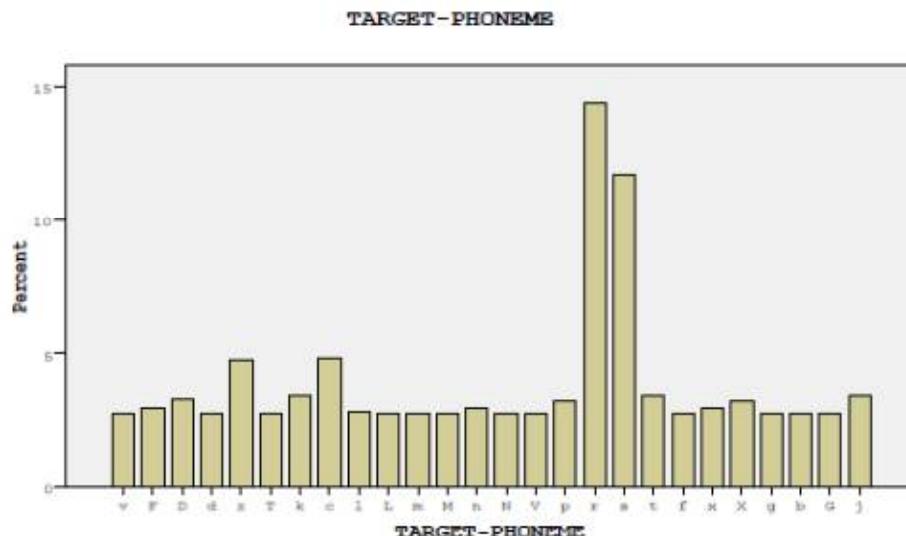


Figure 4. Error distribution

The group age of 3–4 years old made more omissions than substitutions and assimilations, while the age group 4–5 made more metathesis than substitutions and omissions. Finally, the age group 5–6 years old committed no assimilations, or metathesis at all but mainly substitutions and fewer omissions.

The most frequent distorted phonemes were the voiced liquid /r/ (14,1%) and voiceless sibilant fricative /s/ (11,4%). The palatal voiceless plosive /c/ (4,7) and the voiced fricative alveolar /z/ (4,6) followed. More precisely /s/ was replaced by /θ/ in 8,2%, /l/ replaced /r/ in 7,9% and /z/ replaced /z/ in 1,3%. 7,4% of phonemes were omitted with no substitution.



Most of the errors occurred in non accented syllables (65,3%), while 34,7% occurred in accented ones. As far as the syllabic structure was concerned, most of the phonological errors happened in CCV syllables (38%) and concerned cluster reduction or substitution, followed by the CVC syllable (35%), the CV syllable (25%), the V syllable (3,7%) and the VC syllable (2,3%).

## 5 Discussion

The purpose of the present study was to analyze the GSCC in order to determine the percentage of children with or without phonological errors. It was found that 61,1% of children aged three to six years old presented at least one phonological error with the mean of age of children with no mistakes being 4;3. This finding agrees with the results of Mann and Hodson (1994) who found that children speaking Spanish, a quite similar language to Greek at the phonological level, acquired the majority of Spanish phonemes at the age of four. The same age of phoneme acquisition is also reported for German (Foxx and Dodd 1999), while for other languages the mean age

of acquisition is 3;0 for Turkish (Topbas 1997), 5;0 for Portuguese (Yavas and Lamprecht 1988), 4;0 for English according to Prather *et al.* (1975) and 7;0 according to Stoel-Gammon and Dunn (1985).

The main error patterns cited in literature (Ingram 1981, Grunwell 1987, Dean *et al.* 1990) for other languages were also found for Greek. Omissions were committed mainly by the age group of 3–4 years old. This was also found for English (Zhu Hua 2002). Children of that age may prefer omissions, because omissions facilitate word pronunciation from a motoric-articulatory point of view. Syllable or phoneme metathesis is found to be the most frequent error in the age group 4–5. This type of error was also frequent in Italian-speaking children of that age (Bortolini and Leonard 1991) and could be explained as “an outcome at children’s attempts at words with challenging complexity” (Zhu Hua 2002: 11). It has to be noted that in the products of metathesis children demonstrated a strong preference for front consonants, which are the first in the order of phoneme acquisition, when they are word initial. Finally, 5–6 years old children committed more often substitutions with more persistent being the substitution of /s/→/θ/.

It was also found that the most problematic sounds were /r/ and /s/ followed /z/, /z/, /ç/ and /c/. This result shows that consonant acquisition in Greek is similar to other languages (Irwin and Wong 1983, Stoel-Gammon and Dunn 1985, Grunwell 1987), where affricates follow stop acquisition. The liquid /r/ and the fricative /s/ present the highest scores in errors. Once more, this finding agrees with consonant acquisition in Spanish, where the phonemes last mastered are /r s l/ (Mann and Hodson 1994).

Most of the errors were found in non accented syllables. This is probably due to the fact that stressed syllables in Greek are typically more clearly articulated than unstressed syllables, with higher amplitudes and longer durations (Arvaniti 2000).

Finally most of the phonological errors concerned CCV syllables, especially consonant cluster reductions. Recent research on phonological acquisition, based on Jakobson’s notion that the first structures to be acquired would be those that are unmarked, assumed that children will acquire CV structures before acquiring other types of structures. This could explain the high percentage of errors in CCV and CVC syllables.

## **6 Perspectives**

The GSCC is one of the very few oral corpora existing for the Greek language, offered freely for research on child language acquisition, sociolinguistics, psycholinguistics, etc. In the next phase of the GSCC project, an increase in the number of interviews and consequently the total of speaking hours is scheduled, in order to create tagged corpora, to improve the corpus’s geographical representativeness by adding data concerning Greek dialects and finally providing data of children speaking Greek as a second or foreign language. This perspective will permit research on bilingual language development as well.

More advanced statistical analyses of the GSCC cases will provide gender and age comparisons for each specific error-type, as well as new data about phoneme acquisition in comparison with the frequency of words or syllable types in which they are found.

## References

- Arvaniti, A. 2000. The acoustics of stress in Modern Greek. *Journal of Greek Linguistics*, 1.9–39.
- Aston, G. 1997. Small and Large Corpora in Language Learning. In *Proceedings of the 1<sup>st</sup> International Conference on Practical Applications in Language Corpora*, ed. by B. Lewandowska-Tomaszczyk and J. Melia, 51– 61. Lodz: Lodz University Press.
- Bortolini, U. and L. B. Leonard. 1991. The speech of phonologically disordered children acquiring Italian. *Clinical Linguistics & Phonetics* 5.1–12.
- Dean, E. Howell, J. Hill, and D. Waters. 1990. *Metaphon resource pack*. Windsor, Berks: NFER Nelson.
- Edwards, J. and M. E. Beckman. 2008. Methodological questions in studying phonological acquisition. *Clinical Linguistics and Phonetics*, 22(12).939–958.
- Foxx, A. V. and B. J. Dodd. 1999. Der erwerb des phonologischen systems in der Deutschen sprache. *Sprache Stimme Gehor* 23.183–191.
- Garnham A., Shillock R., Brown G., Mill A and A. Cutler. 1981. Slips of the tongue in the London-Lund corpus of spontaneous conversation. *Linguistics* 19.805–817.
- Grunwell, P. 1987. *Clinical phonology*. Baltimore: Williams and Wilkins.
- Ingram, D. 1981. *Procedures for the phonological analysis of children's language*. Baltimore, Md.: University Park Press.
- Ingram, D. 1988. The acquisition of word initial [v]. *Language and Speech* 31. 77–85.
- Irwin J. V. and S. P. Wong (eds). 1983. *Phonological Development in Children 18 to 72 Months*. Carbondale: Southern Illinois University Press.
- Jacobson, R. 1941. *Child language, aphasia, and phonological universals*. The Hague: Mouton.
- Kappa, I. 2000. On the Acquisition of Fricatives. In *Proceedings of the 13th International Symposium on Theoretical and Applied Linguistics. Festschrift for Prof. Athanasios Kakuriotis* , ed. by Katerina Nicolaidis and Marina Mattheoudaki, 197–205.
- Kehoe, M. and C. Stoel-Gammon. 1997. Truncation patterns in English-speaking children's word productions. *Journal of Speech, Language, and Hearing Research*, 40.526–541.

- Kent, R. D. 1992. The biology of phonological development. In *Phonological development: Models, research, implications*, ed. by C. A. Ferguson, L. Menn, and C. Stoel-Gammon, Parkton, MD: York Press.
- Llisteri J. 1996. Els corpus linguistics orals. In *Corpus corpora.*, ed. by Payrato, Boix Llorat, Lorente. Actes del 1er y 2<sup>nd</sup> colloquis linguistics de la Universitat de Barcelona. Promociones y Publicaciones Universitarias. Barcelona, 27–70.
- MacNeilage, P. and B. Davis. 1990. Acquisition of speech production: Frames, then content. In *Attention and Performance XIII: Motor Representation and Control*, ed. by M. Jeannerod. Hillsdale, NJ: Lawrence Erlbaum.
- Mann D. M. and B. Hodson. 1994. Spanish-speaking children's phonologies: Assessment and remediation of disorders. *Seminars in Speech and Language* 15(2).137–148.
- Munson, B., B. A. Kurtz and J. Windsor. 2005. The influence of vocabulary size, phonotactic probability, and wordlikeness on nonword repetitions of children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research*, 48.1033–1047.
- McEnery T. and A. Wilson. 1996. *Corpus Linguistics*. University Press: Edinburgh.
- Nicolaidis, K. 1994. Aspects of lingual articulation in Greek: an electropalatographic study. In *Themes in Greek Linguistics*, ed. by I. Philippaki-Warbuton, K. Nicolaidis and M. Sifianou. 225–232. Amsterdam: John Benjamins.
- Nicolaidis, K. 2002. Durational variability in Vowel-Consonant-Vowel sequences in Greek: the influence of phonetic identity, context and speaker. *Proceedings of the 14th International Symposium on Theoretical and Applied Linguistics*. 280–294. Thessaloniki, Greece.
- Nicolaidis, K. 2003. Acoustic variability of vowels in Greek spontaneous speech. *Proceedings of the 15th International congress of Phonetic Sciences*. Barcelona. 3221–3224.
- Nicolaidis, K. Edwards, J. Beckman, M. and G. Tserdanelis. 2004. Acquisition of lingual obstruents in Greek. In *Proceedings of the 6th International Conference of Greek Linguistics*, Rethymno, Crete.
- Papadopoulou, K. 2000. *Phonological acquisition of Modern Greek*. Unpublished BSc Honours dissertation. University of Newcastle upon Tyne.
- Prather, E., Hedrick, D. and C. Kern. 1975. Articulation development in children aged two to four years. *Journal of Speech and Hearing Disorders* 40.179–191.
- Sanoudaki, I. 2009. Strength relations and first language acquisition. In *Strength relations in Phonology*, ed. by K. Nasukawa and P. Backley. Mouton de Gruyter, 149–183.
- Schiel, F. 2003. *The Validation of Speech Corpora*. Bavarian Archive for Speech Signals. Munich.
- Sinclair, J. 2004. Intuition and annotation - the discussion continues. In *Advances in Corpus Linguistics*, ed. by K. Aijmer and B. Altenberg. Rodopi, Amsterdam, 39–59.

- Smit, A., Hand, L., Freilinger, J. Bernthal, J. and A. Bird. 1990. The Iowa Articulation Norms Project and its Nebraska Replication. *Journal of Speech and Hearing Disorders* 55.779–798
- Stephany U. (1997). The Acquisition of Greek. In *The Crosslinguistic Study of Language Acquisition*, ed. by Dan I. Slobin, Mahwah, NJ/London: Lawrence Erlbaum Ass, Vol. 4, 183–333.
- Stoel-Gammon, C. and Dunn 1985. *Normal and Disordered Phonology in Children*. Austin, TX: Pro-Ed.
- Tzakosta, M. 2001. The Realization of Consonant Clusters in Child Language [in Greek]. *Studies in Greek Linguistics* 21.736–746.
- Tzakosta, M. 2003. Four of a kind: stress patterns in child language. In *Proc. 23<sup>rd</sup> Annual Meeting of Greek linguistics*, Faculty of Philosophy, Aristotle University of Thessaloniki, Thessaloniki, 467–478.
- Tzakosta, M. 2004. *Multiple Parallel Grammars in the Acquisition of Stress in Greek LI*. Ph.D. dissertation, Leiden University. (=LOT Dissertation series 93].
- Yavas, M. and R. Lamprecht. 1988. Processes and intelligibility in disordered phonology. *Clinical Linguistics and Phonetics* 2.329–345.
- Yoneyama, K., M. Beckman and J. Edwards. 2003. Phoneme frequencies and acquisition of lingual stops in Japanese. Submitted to *Language and Speech*.
- Topbas, S. 1997. Phonological acquisition of Turkish children: implications for phonological disorders. *European Journal of Disorders of Communication* 32.377–396.
- Tribble C. 1997. Improving corpora for ELT: quick and dirty ways of developing corpora for language teaching. In *Practical Applications in Language Corpora*, ed. by B. Lewandowska-Tomaszczyk and P. Melia. Lodz: Lodz University Press, 107–117.
- Zhu Hua. 2002. *Phonological development in specific contexts*. Clevedon: Multilingual matters.