

# Automated Text Segmentation of Russian Child-Directed Speech: A Statistical Approach

Natalya Muzinich  
Indiana University  
panteley@indiana.edu

## ABSTRACT

This paper describes how distinctive features that classify speech sounds emerge from statistical analysis of Russian child-directed speech. The analysis is based on transcriptional representation of individual speech sounds. From the analysis of bigram distribution major natural classes such as consonants and vowels further subdivided into non-palatalized versus palatalized consonants and front versus non-front vowels can be computed. The results in the form of a probabilistic FSA exhibit strong associations between the uncovered subclasses of consonants and vowels that are supported by traditional linguistic analysis.

## Categories and Subject Descriptors

I.2.7. [Artificial Intelligence] Natural Language Processing. Language parsing. I.5.3. [Pattern Recognition] Clustering. I.5.4. [Pattern Recognition] Applications. Text processing.

## General Terms

Algorithms, design, measurement, theory.

## Keywords

Principal component analysis. Singular value decomposition. Ward's method.

## 1. INTRODUCTION

How very young infants learn to segment a continuous stream of speech into meaningful units such as phrases, words and morphemes, is an important problem in language acquisition that currently lacks a definitive account. Because infants as young as several months of age are sensitive to statistical properties of sound pattern distribution in the input [11], this sensitivity has been interpreted as evidence for their use in guiding language acquisition. In addition, it has been proposed that in learning to segment a stream of child-directed speech infants rely on multiple cues to which they show sensitivity [6] The extent of reliance on each of these cues is not addressed because it is unclear how it can be measured. This paper describes a statistical approach that derives a hierarchical structure of speech sounds from the textual representation of child-directed speech based on the distributional properties of co-occurrences of individual symbols.

The resulting groupings of symbols closely correspond to the natural classes of speech sounds that are products of traditional linguistic analysis. The unifying feature within each class can be interpreted as one of the cues such as stress to which the infants show sensitivity.

Although textual representation of speech is an idealization which transforms a raw acoustic signal into discrete symbols, the psychological reality of perceiving speech as a linear sequence of discrete units is rarely disputed. Furthermore, the approach taken

here does not imply that individual speech sounds are the units of infants' perception. The results exhibit strong coupling between different sound classes supporting grouping of individual sounds into syllables.

## 2. DATA

Textual representation of child-directed speech analyzed in present work comes from the CHILDES corpus. The author transcribed samples of Russian child-directed speech using Latin alphabet and adapting standard transcriptional symbols to maintain ASCII single character representation for individual sounds. 0 and 1 mark the beginning and the end, respectively, of an adult's turn in their conversation with the child, while all other word boundaries were removed. The input data file consists of 40749 characters. 49 distinct transcriptional symbols were employed.

## 3. METHOD

A 49-by-49 matrix with the rows representing the preceding speech sounds and the columns - the following speech sounds was constructed. The initial data in the matrix was raw bigram count in the input file. The standard technique of  $\log+1$  transform was performed to eliminate zero counts and scale down individual sound frequencies. By-column normalization converted the logarithms of counts to their Z-scores:  $Z=(H_c - \text{Mean}H_c)/\text{Stddev}H_c$ , where for a given following speech sound H,  $H_c$  designates its count in the input data,  $\text{Mean}H_c$  and  $\text{Stddev}H_c$  are the mean and the standard deviation, respectively, of the count of this sound in the data. High positive Z-scores in a column representing the sound  $a$  indicate the sounds that precede  $a$  unusually frequently. Low negative Z-scores designate the sounds that markedly rarely follow  $a$ . Similarity factors were derived from the normalized matrix via singular value decomposition, or SVD, described in the next section.

### 3.1. Vector Space and Latent Structure Modeling

Vector representation suggests high-dimensional space for modeling of the concept of similarity as either distance, such as Euclidean, or direction such as cosine. The current project adheres to the view under which the count correlations signify mutual dependency of both preceding and following sounds on a set of orthogonal factors. These latent variables constitute a coordinate system in a single high-dimensional space for both the preceding and the following sounds. Each dimension makes a largest possible new dissection in the cloud of data points capturing the maximum of the variance in the remaining unaccounted for so far data. If the data are structured, a small subset of earlier dimensions explains most of the variance in the data, while the remaining ones contribute very little to the data distribution. The method, known as the principal component analysis, or PCA, performs a significant dimensionality reduction which retains those factors that impose structure onto the data and discards others that account for the





*guage Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall. 2000

- [6] Jusczyk, Peter. *The Discovery of Spoken Language*. A Bradford Book, MIT Press, Cambridge, Massachusetts, London, England. 1997.
- [7] Mccallum, Andrew, Dayne Freitag, Fernando Pereira. *Maximum Entropy Markov Models for Information Extraction and Segmentation*. Proc. 17th International Conf. on Machine Learning. 2000.
- [8] Manning, Chris and Hinrich Schütze.. *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA. 1999
- [9] Paolillo, John *Analyzing Linguistic Variation: Statistical Models and Methods*. CSLI Publications. 2002
- [10] Doug Beeferman, Adam Berger, John Lafferty. *Statistical Models for Text Segmentation*. Machine Learning 34(1-3), 117-210. 1999.
- [11] Saffran, Aslin and Newport. *Statistical cues in language acquisition: Word segmentation by infants*. COGSCI-96, 376-380, 1996.
- [12] Ratnaparnakhi.. *A Linear Observed Time Statistical Parser Based on Maximum Entropy Models*. Proceedings of the Second Conference on Empirical Methods in Natural Language Processing. UPenn, 133-142, ACL. 1997.