Proceedings of the International Multiconference on Computer Science and Information Technology pp. 359–362
ISBN 978-83-60810-14-9
ISSN 1896-7094

An application supporting language analysis within the framework of the phonetic grammar

Krzysztof Dyczkowski
Adam Mickiewicz University
Faculty of Mathematics and Computer Science,
Umultowska 87, 61-614 Poznań, Poland
Email: chris@amu.edu.pl

Norbert Kordek, Paweł Nowakowski, Krzysztof Stroński
Adam Mickiewicz University,
Institute of Linguistics,
al. Niepodległości 4, 61-874 Poznań, Poland
Email: {norbert, gpn, stroniu}@amu.edu.pl

***Abstract*—The aim of the paper is to present an application supporting language analysis within the framework of the phonetic grammar. The notion of the phonetic grammar has been concisely introduced and the basic potential of the application and the algorithms employed in it are briefly discussed. The application is to enable a uniform description and a comparative analysis of many languages. At the first stage the languages taken into consideration are Polish, Chinese and Hindi.**

I. The phonetic grammar

A. Introduction

PHONETICS is a field of linguistics which is concerned with the articulatory, acoustic, auditory and distributive properties of phones. Phonetics of a given language is also sometimes understood as a set of phones relevant to a given language (e.g. the phonetics of Polish language). The phone is a set of all hic et nunc pronounced homophonous speech sounds. The speech sounds are of temporal character and their number is actually infinite. To reduce the number of elements belonging to hic and nunc pronounced speech sounds we classify them into sets of phones e.g. the set of all homophonous temporal realizations of the speech sounds p1, p2, p3, p4, ... is considered to be the phone [p]. All phones are described in terms of the articulatory features. E.g. the relevant features of [p] are: voiceless, oral, hard, plosive, labial etc. Assigning an exhaustive feature set to a given phone is equal with the definition of the phone.

Phonetic grammar is understood as a set of relations between articulatory features and articulatory dimensions (sets of homogenous articulatory features). The concept is based on the theory introduced in the works of Jerzy Bańczerowski ([1], [2]).

The aims of the present project are as follows:

- elaboratation of a uniform (for each language) set of articulatory features,
- elaboratation of a mathematical model of the relations between languages,
- preparation of computer tools for the processing of the collected data, i.e. to elaborate a model of the linguistic data collecting,
- implementation of algorithms of the data processing,
- preparation of algorithms of comparative analysis of languages using the mathematical model,
- comparative analysis of the phonetic grammars of Polish, Chinese and Hindi.

To enable the comparative analyses of languages we introduce a uniform description of the phones of a given language as a set of articulatory features belonging to one of the following articulatory dimensions:

- the mechanism of the air flow origin,
- the direction of the air flow,
- the state of glottis,
- the way of air flow
- the place of articulation,
- the articulator,
- the degree of supraglottal aperture,
- the vertical position of the tongue—the horizontal position of the tongue,
- the degree of labialization,
- the degree of delabialization,
- the duration of articulation
- the degree of supra- and subglottal tension,
- the frequency of articulatory approximation.

B. Phones as objects in n-dimensional space

The original method introduced the notion of the articulatory distance between phones (see [2]). The distance is interpreted as a number of differential features (features by which given phones differ from others). It is thus reducible to the well known Hamming distance.

Our team has proposed to introduce numerical interpretation of the articulatory dimensions. Let G be a set of all phones within which the subsets G_l of the phones belonging to a given language can be specified (where l is an index of a given language) and let $W = \{W_1, W_2, \ldots, W_n\}$ be a set of articulatory dimensions, where n is a number of articulatory dimensions. Thus each phone g from the set G is specified by the vector in n-dimensional metric space $\mathbb{R}^n$. Each articulatory feature is uniquely specified by one numerical value from the interval $[0, k]$, where k is a maximal number of features in a dimension. Ascribing a proper numerical value to the feature mirrors the natural order of the features in a given dimension.

The research project is supported by Ministry of Science and Higher Education grant N N104 327434.

Thus each phone $g = (c_1, c_2, \ldots, c_n)$ where c_i belongs to the set of features of a given dimension W_i.

The notion of the phone as a point in n-dimensional space enables application of well known measures of distances. For example for the pair of phones $a, b \in G$ we can specify the following measures of distances:

- The Minkowski distance for $m \geq 1$:

$$Dist_M(a,b) = \left(\sum_{i=1}^{n} |a_i - b_i|^m \right)^{1/m}$$

- The Manhattan distance:

$$Dist_H(a,b) = \sum_{i=1}^{n} |a_i - b_i|$$

being a particular instance of the Minkowski distance for $m = 1$,

- The Euclidean distance:

$$Dist_E(a,b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

being a particular instance of the Minkowski distance $m = 2$.

The distances defined in this manner will enable us to build similarity measures between phones and between phonetic systems of given languages. We assume that sound more distant from each other in the sense of the appropriate metrics are less similar to each other. And this seems to be in accordance with the intuition.

II. Computer application

A. The tool for collecting phone inventories

The first essential element in the system has been to build a database and a relevant interface enabling data insertion using the standardized International Phonetic Alphabet (IPA).

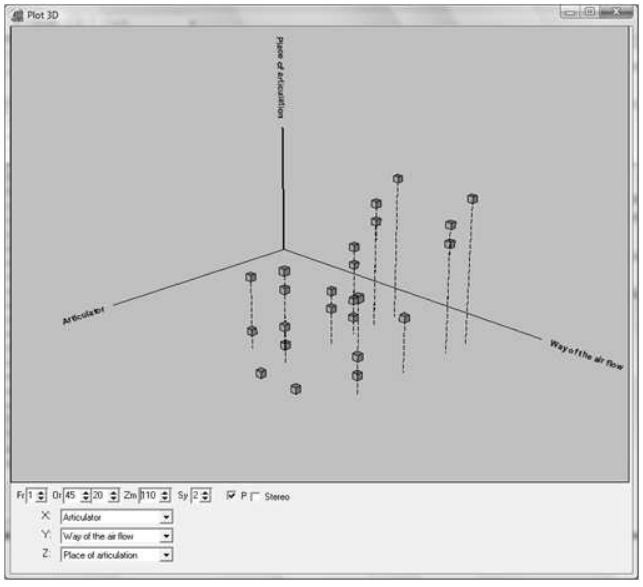

Fig. 1. Phones in selected 3 dimensions

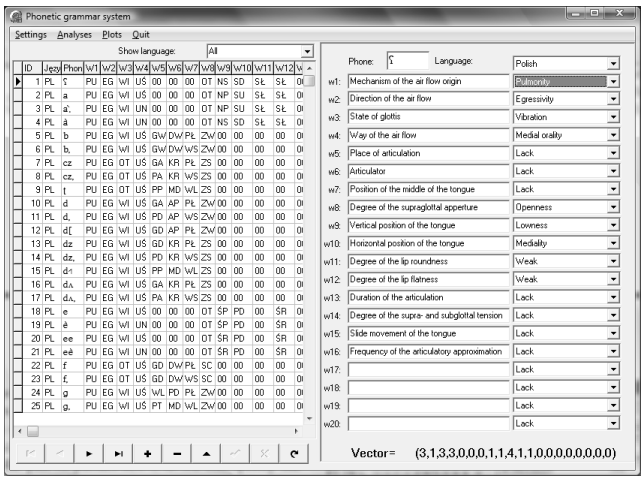

Fig. 2. Inventory of phones

The application enables:

- defining dimensions and features occurring in them,
- ascribing proper numerical values to the dimensions,
- defining the number of languages,
- introducing a repertoire of phones of a given language and the description of the phones in terms of the relevant set of articulatory features.

B. Basic analyses

The application is to generate the data concerning detailed levels of analysis in the phonetic grammar of each of the analyzed languages:

- the phone articulemization[1],
- the combining of the articulatory features,
- the articulatory opposition and similarity of phones,
- differential and identifying articulatory dimensions,
- the articulatory distance and proximity.

The computer application will automatically generate:

- the articulatory distance of any two phones in a given language,
- the articulatory similarity of any two phones in a given language,
- the articulatory features of a given phone,
- the articulatory category of a given articulatory feature,
- the dimensions in which given phones differ,
- the dimensions in which given phones are identical,
- the set of phones which have a specified articulatory distance,
- the set of phones which have specified articulatory features,
- the combination of a given set of articulatory features,
- the average articulatory distance between phones,
- the most numerous articulatory category specified by a given number of features,

[1]The operation of articulemization consists in ascribing the articulatory characteristics to a given phone.

- the least numerous articulatory category specified by a given number of features,
- the set of relevant features discerning at least one pair of sounds,
- the number of pairs of phones being discerned by particular features,
- the number of pairs of phones being discerned by particular sets of features,
- the most frequently combined articulatory features in a given articulatory distance.

C. Applied data-mining algorithms

The analyses presented in the last section are the basis of language analysis. They apply rudimentary statistical and combinatorics methods. In the present section we are going to explore methods from the data-mining domain which will allow to discover automatically new interdependencies between phones. It will in turn enable to show certain relations between languages which have been so far unnoticed. All algorithms applied here use measures of distances as measure of similarity between phones. These algorithms can be used for phones from one ore more languages.

1) **K-means algorithm** *([9],[12]):* The first of the algorithms requires an input of expected number of phone clusters. It allows to divide the phone inventory into particular number of disjunctive classes. For example put $k := 2$ results in the division of the set of phones into vowels and consonants.

The algorithm is composed of the following steps:

1. Place K points into the space represented by the phones that are being clustered. These points represent initial group centroids.
2. Assign each phone to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

2) **The connected subgraphs algorithm** *([5],[7]):* This algorithm does not require an input of the number of clusters. It finds them itself on the basis of regularities in the data.

The algorithm operates on the basis of the matrix of distance D. It is a symmetric matrix $n \times n$ in which on the intersection of the columns and verses one receives the distance between the proper pair of phones and on the diagonal zero.

The algorithm operates in the following steps:

1. The distance matrix D is calculated using the fixed distance.
2. Below the fixed threshold α all values in the matrix D are zeroed. Finding the threshold is the basic element of the algorithm. In the simplest case it can be fixed as an average distance in the phone inventory reduced by the standard deviation of the average distance. The choice of the proper threshold mirrors our understanding how big the distance between phones must be to consider them too distant to be the members of the same group.

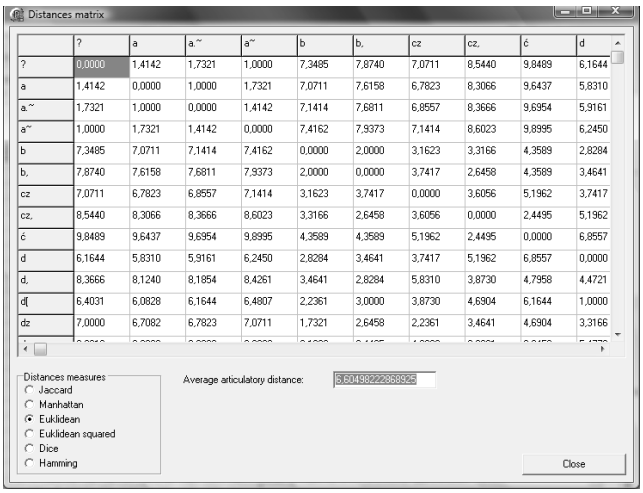

Fig. 3. The distances matrix for the Euclidean metric

3. Such a matrix is treated as a directed weighted graph in which non-zero values will mark weighted edges between the phones—the vertices. (also called threshold graph)
4. The algorithm Depth-first search (DFS) is applied. The algorithm results in finding connected subgraphs. The subgraphs are wanted clusters.

3) **Agglomerative hierarchical clustering and dendrograms** *(see [10],[11]):* An agglomerative hierarchical clustering procedure produces a series of partitions of the data, $P_k, P_{k-1}, \ldots, P_1$. The first P_k consists of k single phone clusters, the last P_1, consists of a single group containing all k phones.

At each particular stage the method joins together the two clusters which are closest together (most similar). At the first stage, this amounts to joining together the two objects that are closest together, since at the initial stage each cluster has one phone.

There are some different methods that use different ways of defining distance (or similarity) between clusters.

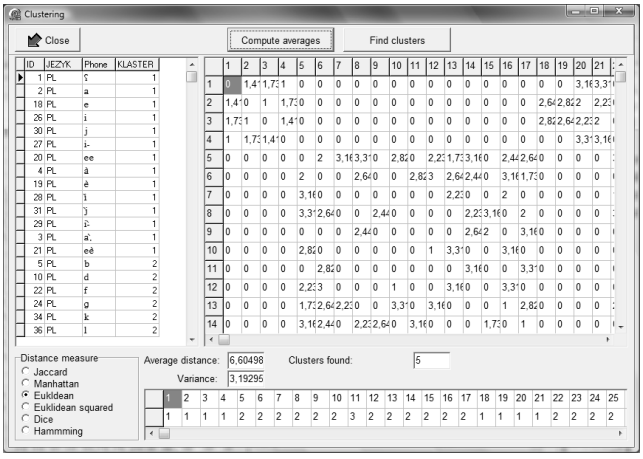

Fig. 4. The effect of the operation of the connected subgraphs algorithm

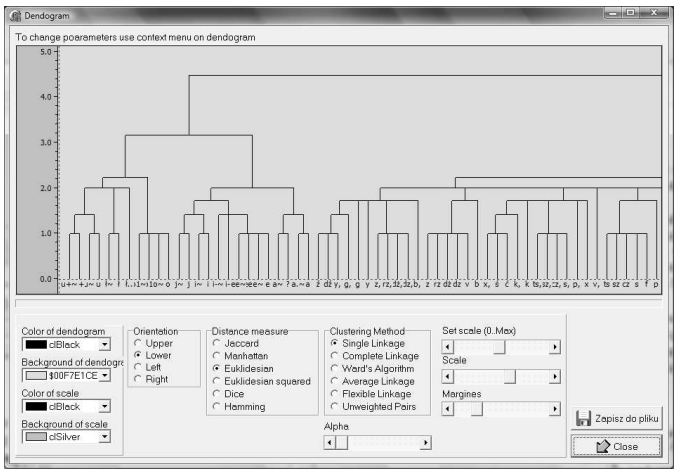

Fig. 5. The example of the dendrogram

1. **Single linkage clustering**: It is one of the simplest agglomerative hierarchical clustering methods. It is also known as the nearest neighbor technique. The defining feature of the method is that distance between groups is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each group are considered.
2. **Complete linkage clustering**: It is also called farthest neighbor, clustering method is the opposite of single linkage. Distance between groups is now defined as the distance between the most distant pair of objects, one from each group.
3. **Average linkage clustering**: Here the distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group.
4. **Average group linkage**: With this method, groups once formed are represented by their mean values for each variable, that is, their mean vector, and inter-group distance is now defined in terms of distance between two such mean vectors.

The result of the operation of the algorithm is presented in the dendrogram. It is a special type of the dendric structure which assures an easy way of presenting the results of the hierarchical grouping.

Cutting the dendrogram at the selected level we can receive a proper division into a particular number of the groups of phones.

III. Summary

The paper has presented the first stage of the realization of the more complex project which is meant to apply computer methods for the purpose of linguistic analyses.

At the next stages, besides the interpretation of the results we are going to apply the methods of fuzzy sets (mainly the notion of the linguistic variable) for the description of the repertoires of phones. The methodology of the linguistic summarization as a tool of the data analysis also seems to be very promising.

The results can be further used in different linguistic disciplines (also applied linguistics), especially in teaching foreign languages, speech analysis and in basic research on natural languages (in theory of linguistics and literary phonostylistics, comparative linguistics, typology).

References

[1] J. Bańczerowski, “Phonetic Relations in the Perspective of Phonetic Dimensions“, *In: Pieper U., Stickel G. (eds.) Studia Linguistica Diachronica et Synchronica*. Berlin, 1985
[2] J. Bańczerowski, J. Pogonowski, T. Zgółka, “Wstęp do językoznawstwa.“ UAM, Poznań.
[3] T. Benni, “Fonetyka opisowa języka polskiego“, Ossolineum, Wrocław, 1964.
[4] C.K. Bhatia, “Consonant sequences in Standard Hindi“, Indian Linguistics, 1964, 25.206-12.
[5] U. Brandes, M. Gaertler, and D. Wagner. Experiments on graph clustering algorithms. *Lecture Notes in Computer Science, Di Battista and U. Zwick (Eds.)* :568–579, 2003.
[6] Yuen-ren. Chao, “A Grammar of Spoken Chinese“, University of California Press, Berkeley L’ Los Angeles L’ London, 1968.
[7] S. van Dongen, “Graph Clustering by Flow Simulation“, PhD thesis, University of Utrecht, 2000.
[8] L. Dukiewicz, T. Sawicka, “Fonetyka i fonologia“, W: *Urbańczyk S. (red.) Gramatyka współczesnego języka polskiego*, IJP PAN, Kraków, 1995.
[9] J. Han, M. Kamber, “Data Mining: Concepts and Techniques“, Morgan Kaufman, 2000.
[10] J. Hand, H. Mannila, P. Smyth, “Pricinciples of Data Mining“, MIT Press, 2001.
[11] A. Jain and R. Dubes,“Algorithms for Clustering Data“,Prentice-Hall, 1988.
[12] D.T. Larose,“Discovering Knowledge in Data: An Introduction to Data Mining“, Wiley, 2005.
[13] M. Ohala, “Aspects of Hindi Phonology“, Motilal Banarisidas, Delhi, 1983.
[14] M. Steffen-Batóg, “Studies in Phonetic Algorithms“, Sorus, Poznań, 1997.
[15] M. Steffen-Batóg, T. Batóg, “A Distance Function in Phonetics“, Lingua Posnaniensis 23, 47 L’ 58, 1980.
[16] B. Wierzchowska B. “Opis fonetyczny języka polskiego“, PWN, Warszawa, 1967.
[17] Qin. Zhong, “On Chinese Phonetics“, Commercial Press, Beijing, 1980.