

Computational tools in the analysis of phonetic grammar[#]

Krzysztof Dyczkowski,* Norbert Kordek,**
Paweł Nowakowski,** and Krzysztof Stroński**

*Adam Mickiewicz University, Faculty of Mathematics and Computer Science,
Poznań, Poland

**Adam Mickiewicz University, Institute of Linguistics, Poznań, Poland
chris@amu.edu.pl
{norbert, gpn, stroniu}@amu.edu.pl

ABSTRACT

The aim of this paper is to present an application supporting language analysis within the framework of phonetic grammar. The notion of phonetic grammar is concisely introduced and the basic potential of the application and the algorithms employed in it are briefly discussed. The application is intended to enable a uniform description and comparative analysis of many languages. In the initial stage the languages considered are Polish, Chinese and Hindi.

1. Phonetic grammar

1.1. Introduction

The idea of using mathematical, logical and computational tools to analyze natural language is, of course, neither new nor revealing. Even in the field of phonetic and phonological research alone, there have been a considerable number of proposals in a similar vein to the one presented here. In this context one could mention Batóg [6], Steffen-Batóg & Batóg [16], Steffen-Batóg [15], Meyer-Eppler [14] – to name only a few. The limitations of this paper do not allow us to elaborate more on those proposals and how they differ from with our framework.

The notion of grammar may be understood in many ways. Probably in the most abstract sense it can be conceived as a calculus of a particular subsystem of a language. In the traditional sense of this term, “phonetic grammar” may sound odd as “grammar” is mainly associated with such subsystems as morphology, semantics and syntax, more loosely with phonology and hardly ever with phonetics. This may be a consequence of the anatomical, acoustical or auditory character of basic phonetic studies, which indeed are not grammatical in nature.

We follow the theoretical proposals of Bańcerowski [1–4], Bańcerowski *et al.* [5] in which the results of basic phonetic research are used as data in a framework that has all the traits of a grammatical theory, i.e. it operates on mutually related linguistic units. The most relevant property of this proposal is that dependencies and relations

[#]The research project is supported by Ministry of Science and Higher Education grant NN104327434.

between phonetic units are described in the same way, as far as methodology is concerned, as those between morphemes in morphological theory, between words in syntactic theories etc.

The foundations of the theoretical framework that we are adapting for the project are extensively introduced in Bańcerowski [1] (and more concisely in Bańcerowski *et al.* [5]). Here we will restrict ourselves to introducing only those theoretical aspects of phonetic grammar which are necessary to give the reader a clear understanding of the project.

Articulatory phonetics can be treated extensionally and thus it can be understood as a set of phones (cf Batóg [6]: 32), each having an articulatory description of some sort. Phonetic grammar is no different in this respect, but it extends beyond just the inventory of phones, treating them as input data for analytical operations that will be briefly introduced in this paper.

The relevance of the phonetic inventory must be a property shared by all phonetic studies. Our approach in this aspect is special in that it takes into account only an anatomical description of the process of articulation. It is probably impossible to completely eliminate the influence of the phonological properties of a given language on the shape of the articulatorily driven phone inventory, but nevertheless it is one of the theoretical assumptions of phonetic grammar that it should be kept separate from phonological issues. This assumption presents a problem when it comes to making use of existing inventories of the phones of particular languages. It is common practice that drawing up a phone inventory involves minimizing the number of elements in the set. Minimizing usually means taking into account phonological properties of the system. This kind of procedure is directly contrary to what we assume in our study. The number of elements in the inventory is not important; we try to take into account all articulatorily distinct phones that are found in a particular language. This includes, for example, all distributional allophones of what is regarded as one element of the inventory from the phonological point of view.

At the foundation of the theory of phonetic grammar lie several primitive terms, among which the following may be listed:

- speech sound (phone), *hic et nunc* – a physical entity produced at a certain time,
- the set of articulatory features,
- the relation of homophony,
- the relation of homogeneity.

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 infinite. To reduce the number of elements of the set of *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].

The process of building an inventory within the framework of the phonetic grammar requires the assignment of articulatory features to each phone. The operation of assigning articulatory features to a phone is called *articulemization*. An articulatory feature is a single characteristic of phones in some articulatory aspect. For example, the relevant features of [p] are: voiceless, oral, hard, plosive, labial etc. Assigning an exhaustive feature set to a given phone is equivalent to definition of the phone. Thus the set of all features can be identified with the set of all phones in a given languages (cf Batóg [6]: 32; Bańcerowski *et al.* [5]: 148).

The set of all features is categorized into subsets by the relation of articulatory homogeneity, which holds between features that are characteristics in the same articulatory aspects, in other words features that can be compared. Such subsets are called articulatory dimensions – a dimension is a set of homogeneous features and will be denoted as D_i . By D we will denote the set of all articulatory dimensions. Below we give a tentative list of dimensions that is sufficient for an exhaustive description of the phone repertoires of Chinese, Hindi and Polish:

- D_1 the mechanism of the airflow origin,
- D_2 the direction of the airflow,
- D_3 the state of the glottis,
- D_4 the path of the airflow,
- D_5 the place of articulation,
- D_6 the articulator,
- D_7 **the position of the middle of the tongue,**
- D_8 the degree of supraglottal aperture,
- D_9 the vertical position of the tongue,
- D_{10} the horizontal position of the tongue,
- D_{11} the degree of labialization,
- D_{12} the degree of delabialization,
- D_{13} the duration of articulation,
- D_{14} the degree of supra- and subglottal tension,
- D_{15} **the slide movement,**
- D_{16} the frequency of articulatory approximation,
- D_{17} **the degree of glottal aperture,**
- D_{18} **the strength of approximation.**

Bold type marks the dimensions that have been added to the original proposal in Bańczerowski [4]. One of them, namely *the degree of supra- and subglottal tension*, although irrelevant for the description of the above-mentioned languages, has not been removed from the list since it can be utilized later for the comparison of other phonetic systems. There are two important properties of articulemization to be mentioned at this point:

- since in each dimension there is a “none” feature, every phone is assigned a feature in each dimension, even if it is articulatorily unspecified in a given respect,¹
- not only relevant features² are assigned to each phone; the number of features of each phone in all languages is the same and is equal to the total number of dimensions.

1.2. Phones as objects in n -dimensional space

Another important change to the original framework is the introduction of numerical interpretation of features in each dimension D_i . Let G be the set of all phones in the inventory and G_l be the set of all the phones belonging to a given language (where l is the index of that language). 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 D_i . Due to space limitations we will not give an extensive description of each

¹For example all vowels in the dimension “place of articulation”.

²I.e. the unique set of features that distinguish a given phone from any other.

Table 1. Features in the Dimension D_5 : Place of articulation

n	feature
0	None
1	Upperalabiality
2	Upperdentality
3	Postdentality
4	Upperalveolarity
5	Postalveolarity
6	Praepalatality
7	Postpalatality
8	Velarity

dimension – we will restrict ourselves only to one example – D_5 : Place of articulation (see Table 1).

In this way each phone g is specified by a vector $g = (f_1, f_2, \dots, f_n)$ where f_i is a feature from the set of articulatory features of a given dimension D_i . As an example, in the above exemplary dimension the phone [p]³ is assigned the value 1, which denotes upperlability. The whole vector is coded by repeating the procedure in each dimension. As a result each phone becomes an object in n -dimensional space represented by a vector. The numerical values of features in every dimension are not random – for instance in the example dimension the increasing numbers indicate the actual order of relevant anatomic parts of the oral cavity from the furthest front to the furthest back. Thus each phone g from the set G is specified by a vector in n -dimensional metric space \mathbb{R}^n .⁴

The phonetic grammar may now be introduced as a set of relations between phones (objects in n -dimensional space) and articulatory dimensions (cf Fig. 1). Bańcerowski's original framework introduced the notion of articulatory distance equivalent to Hamming distance ($Dist_H$); this is defined for $a, b \in G$ as

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

where

$$H(a_i, b_i) = \begin{cases} 1 & \text{if } a_i \neq b_i, \\ 0 & \text{otherwise.} \end{cases}$$

It is measured by the number of differential features (features in which given phone differs from the other). For example the Hamming distance between the phones [p] and [b] is equal to 1, because these phones are differentiated by one feature (or differ in one dimension).

The vector interpretation that we propose makes it possible to apply other well-known distance measures, such as Minkowski distance, Manhattan distance, Euclidean distance,

³A particular phone will be given in brackets.

⁴Where n is simply the number of articulatory dimensions, which is 18 at the present stage.

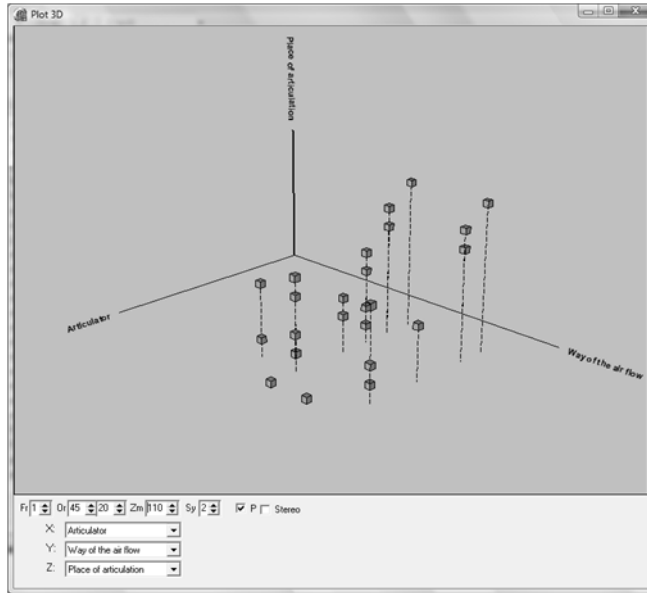


Figure 1. Phones in three selected dimensions.

etc., which can be employed to measure more precisely similarities between the phones and phonetic systems of different languages.

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:

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

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 $m = 2$.

To exemplify the difference in precision, let us consider the Hamming ($Dist_H$) and Euclidean ($Dist_E$) metrics applied to the phones [p] and [k] in only one dimension D_5 – “Place of articulation”:

- $Dist_H = 1$ – this simply means that the two phones differ in this dimension; $Dist_H$ is not able to indicate any more subtle differences,

- $Dist_E = 7$ – this takes into account the vector value of both phones in the dimension, which is 8 (upperlabiality) for [k] and 1 (velarity) for [p].

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

2. Computer application

Measuring the similarities between phones and articulatory systems of different languages is only one of many analytic possibilities that our framework allows. Regardless of the kind of analysis that is to be conducted within the phonetic grammar, the best results are going to be achieved by means of computer aided computational methods. The mathematical model of articulation makes this approach even more plausible. Another important part of our phonetic project is to design a computer application to facilitate the research within this framework.

2.1 Tool for collecting phone inventories

The first essential element in the system has been to build a database and a relevant interface that would enable data insertion using the standardized International Phonetic Alphabet (IPA) (Fig. 2).

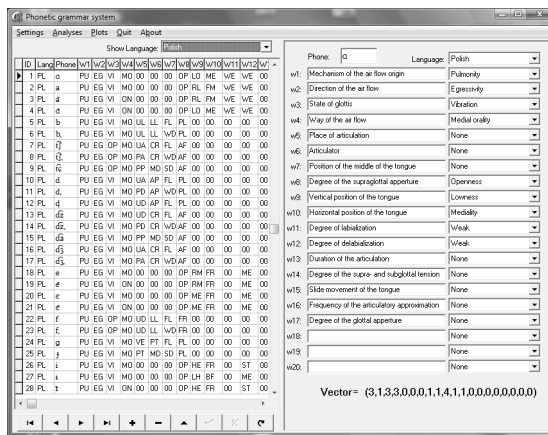


Figure 2. Repertoire of phones.

The application enables:

- definition of dimensions and features occurring in them,
- assignment of appropriate numerical values to the dimensions,
- definition of the number of languages,
- entry of a repertoire of phones of a given language and description of the phones in terms of the relevant set of articulatory features.

2.2. Basic analyses

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

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

Below there is a list of results of our basic analytical algorithms which have already been implemented in the software. Thus the computer application is able to automatically generate:

- the articulatory distance 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,
- the least numerous articulatory category specified by a given number of features,
- the set of relevant features discerning at least one pair of sounds from each other,
- 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.

2.3. Applied data-mining algorithms

The analyses presented in the last section are the basis of language analysis. They apply rudimentary statistical and combinatorial methods. In the present section we will explore methods from the field of data-mining which will allow us to discover automatically new interdependencies between phones. This will in turn enable us to identify certain relations between languages which have been so far unnoticed. All the algorithms applied here use measures of distance as measures of similarity between phones.

2.3.1. *K-means algorithm*⁵

The first of the algorithms requires as input the expected number of phone clusters. This makes it possible to divide the phone inventory into a particular number of disjoint classes. For example, putting $k := 2$ results in a division of the set of phones into vowels and consonants.

The algorithm consist 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.

⁵cf Han & Kamber [10], Larose [13].

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.

2.3.2. *The connected subgraphs algorithm*⁶

This algorithm does not require the number of clusters as input. It finds them itself on the basis of regularities in the data.

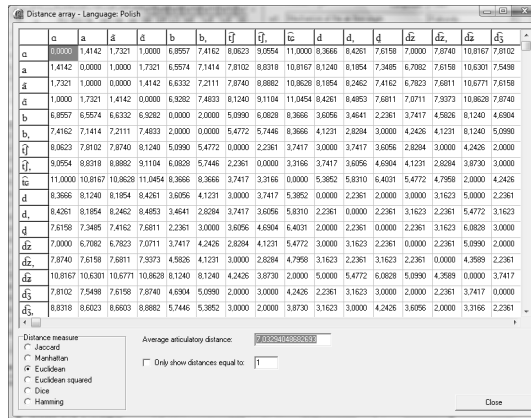


Figure 3. Distances matrix for the Euclidean metric.

The algorithm operates on the basis of the matrix of distance DM . It is a symmetric matrix $N \times N$ (where N is the number of phones taken into account) in which on the intersection of the columns and rows one obtains the distance between the proper pair of phones, and the diagonal consists of zeros.

The algorithm consists of the following steps:

1. The distance matrix DM is calculated using the fixed distance.
2. Below the fixed threshold α all values in the matrix DM are zeroed. Finding the threshold is the basic element of the algorithm. In the simplest case it can be fixed as the average distance in the phone inventory reduced by the standard deviation of the average distance. The choice of the proper threshold mirrors our understanding of how large the distance between phones must be to consider them too distant to be members of the same group.
3. Such a matrix is treated as a directed weighted graph in which non-zero values mark weighted edges between phones – the vertices (also called a threshold graph).
4. The algorithm Depth-First-Search (DFS) is applied. The algorithm produces connected subgraphs.

The subgraphs form the clusters (Fig. 4).

⁶cf Ulrik et al. [17].

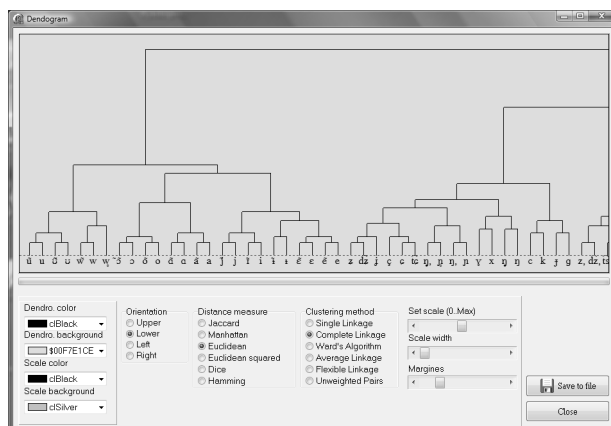


Figure 5. Example of a dendrogram.

By cutting the dendrogram at a selected level we can obtain a proper division into a particular number of groups of phones.

3. Summary

The paper presents the first stage of the realization of a more complex project which is intended to apply computational methods for the purpose of linguistic analyses.

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

The results can be further applied in various linguistic disciplines (including 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 and typology).

BIBLIOGRAPHY

- [1] Bańczerowski, Jerzy. 1985. "Phonetic Relations in the Perspective of Phonetic Dimensions". In: Pieper U., Stickel G. (eds.) *Studia Linguistica Diachronica et Synchronica*.
- [2] Bańczerowski, Jerzy. 1987. "Towards a dynamic approach to phonological space". *Studia Phonetica Posnaniensia*, vol. 1, 5–30.
- [3] Bańczerowski, Jerzy. 1990. "Undular aspect of phonological space-time". *Studia Phonetica Posnaniensia*, vol. 2, 13–42.
- [4] Bańczerowski, Jerzy. 1992. "Formal properties of neostructural phonology". *Studia Phonetica Posnaniensia*, vol. 3, 5–28.
- [5] Bańczerowski, Jerzy, Pogonowski, Jerzy, & Zgółka, Tadeusz. 1982. *Wstęp do językoznawstwa*. Poznań: Wydawnictwo UAM.

- [6] Batóg, Tadeusz. 1967. *The Axiomatic Method In Phonology*. London: Routledge and Kegan Paul Ltd.
- [7] Benni, Tytus. 1964. *Fonetyka opisowa języka polskiego*. Wrocław: Ossolineum.
- [8] Wierzchowska, Bożena. 1967. *Opis fonetyczny języka polskiego*. Warszawa: PWN.
- [9] Dukiewicz, Leokadia, & Sawicka, Irena. 1995. "Fonetyka i fonologia". In: Urbańczyk Stanisław (ed.) *Gramatyka współczesnego języka polskiego*. Kraków: IJP PAN
- [10] Han, Jiawei, & Kamber, Micheline. 2000. *Data Mining: Concepts and Techniques*. Morgan Kaufman.
- [11] Hand, David, Heikki, Mannila, & Padhraic, Smyth. 2001. *Principles of Data Mining*. MIT Press.
- [12] Jain, Anil, & Dubes, Richard. 1988. *Algorithms for Clustering Data*. Prentice-Hall.
- [13] Larose, Daniel T. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley.
- [14] Meyer-Eppler, Werner. 1959. *Grundlagen und Anwendungen der Informationstheorie*. Berlin: Springer-Verlag.
- [15] Steffen-Batóg, Maria. 1997. *Studies in Phonetic Algorithms*. Poznań: Sorus.
- [16] Steffen-Batóg, Maria, & Batóg, Tadeusz. 1980. "A Distance Function in Phonetics". *Lingua Posnaniensis*, 23, 47–58.
- [17] Ulrik, Brandes, Gaertler, Marco, & Wagner, Dorothea. 2003. "Experiments on graph clustering algorithms". *Lecture Notes in Computer Science*, Di Battista and U. Zwick (Eds.), 568–579.

