Displaying Phonological Diachronic Changes through a Database Application

Marta Manfioletti and Mattia Nicchio

Department of Linguistics and Performing Arts, University of Padua
Via Beato Pellegrino, 1 – 35137 Padua, Italy
{marta.manfioletti,mattia.nicchio}@gmail.com

Abstract. This paper presents a project which aims to provide a new digital instrument for linguistic research. This new tool will be able to show the historical evolution of a language into one or more daughter languages, and it will allow users to perform a comparative and typological analysis of diachronic processes. The originality of this project is given by two factors: first, its developers are linguists with notions in computer science, which prevents any communication issue between different teams of experts; second, the data feeding database, though derived from well known corpora, have been processed in a specialist way to display the evolution of words from a mother language to the daughter languages. The instrument will account for all the diachronic phonological rules which occur during the word change.

1 Introduction

Computer science has broadened the horizons of many scientific and humanistic disciplines by providing a huge number of new instruments to the experts of all fields of research. The main issue about the development of those instruments is its interdisciplinary nature. To obtain good results a good level of collaboration and communication is needed between the computer scientists developing the digital instrument and the experts of the discipline the instrument is about. The more the experts of both fields mutually share their specific knowledge, the more accurate and detailed the structure will be.

The best solution, however, would be to have a single team with all the necessary competences: this means that the digital instrument would be wholly conceived by the same people, and thus there would be no discrepancies or lack of detail that would require further adjustments or even compromise the functions of the instrument.

The division of the work through the areas of expertise can often lead to incomplete or approximate results; the best way to avoid this problem would be for the theoretical structure to be directly conceived according to the structure of the database.

This is precisely what is new about this work: the approach to the matter was twofold right from the outset, and allowed the development of a database which performs its tasks while fully accounting for a complex phonological reality. The

M. Agosti et al. (Eds.): IRCDL 2011, CCIS 249, pp. 49–60, 2011.

effectiveness of this approach becomes even clearer when examining the results: the outcome of the work is not simply the digitalization of a group of corpora, but a completely new research instrument.

In this paper we describe the first step of a broader project aiming to create a digital library which gathers diachronic phonological data to potentially account for the historical linguistic changes in every human language. The first phase of the work deals with the design of the database which represents the backbone of the entire digital library. The aims of this first stage will be explained in the next sections, but the most important achievements are listed below:

- To automatize the preliminary stages of diachronic phonological analyses and create new knowledge through an original form of data processing.
- To allow immediate comparison of data in order to give a broader perspective to comparative and typological research.
- To gather information from historical grammars and organize it in a single instrument, thus allowing the comparison of data that have always been isolated to date.
- To be able to represent the diachronic phenomena present in every human language.
- To account for phonological reality in a new way: for the first time it can be represented by means of the entity-relationship (ER) model.
- To explain existing data using new phonological theories [1].

This paper is organized as follows: Section 2 contains the motivations and the purposes of the project; Section 3 introduces some basic linguistic concepts; Section 4 presents the dual nature of the work, explaining the linguistic aspect and accounting for some choices that have been made about the database; Section 5 contains the conclusions and some thoughts about the possible futire developments of this work.

2 Motivations

This section will deal with the preliminary choices that shaped the entire project, and its purpose will be to show how even a simple database structure can give another discipline a new approach to the studied object, thus producing new knowledge.

Databases can be extremely useful in a wide range of activities, since they allow users to organize information and to store it so that it will be easy to retrieve and modify later on. However, some distinctions must be made.

Consider the case in which a bank wants to organize all the data about every bank account. The ER model will provide the conceptual schema, then a logic schema will be used to design the database structure, which will presumably be quite simple. Eventually, data and metadata will be fed into the database, and the digital tool will be ready to step in and take the place of all the paperwork. This kind of procedure is perfectly ordinary and does not bring any innovation, neither to the banking sector nor to computer science. The only (and yet, not

minor) improvement as opposed to the previous system is the rationalization of information which was already there at the time of implementing of the database.

In the present project, things are different. The contribution given by computer science to linguistics goes much further, since the database will make it possible to create new knowledge. Of course, it is not simply supposed to do such a thing. There are plenty of possible applications of databases to linguistics that do not involve a different perspective on the discipline, but simply allow swifter access to information by digitalizing preexisting corpora. This database, though, aims to give linguists a new way of approaching phonological change and comparing languages and their changing patterns.

First of all, the digital tool will be able to store data which has never been gathered in the same place before, and to allow immediate comparison between historical grammars and etymological dictionaries by representing their contents through a common standard. This aspect will not consist of a mere digitalization, since the uniformity of data is yet to be achieved: in fact, human work will be required to extract and elaborate data to make them fit the database structure.

There is, however, another main innovation besides the new representation of linguistic data: typological comparison. Once the database is fed with data representing many different languages, it will be possible to compare the evolution of different linguistic varieties in order to make generalizations about the way linguistic changes occur. This approach would not be possible without the digital instrument presented in this paper.

Furthermore, the database will help any linguist perform preliminary research before the analysis of a set of phonological changes.

Several linguistic databases with different functions already exist or are being created; some of them are listed and described below:

- The *ASIt* project[1] [2], which is being carried out by the Department of Information Engineering and the Department of Linguistics and Performing Arts of the University of Padua. The ASIt project consists of a syntactic database where the aim is to gather variants of grammatical structures within a sample of 200 Italian dialects.
- The *Multimedia Atlas of Venetian Dialects* (AMDV[2] is being carried out by the Department of Linguistics and Performing Arts of the University of Padua with the collaboration of the ISTC of Rome and the Department of Historical Studies of Ca Foscari University of Venice. The aim of this project is to digitalize the *Sprach- und Sachatlas Italiens und der Südschweiz* (AIS) by K. Jaberg and J. Jud which represents one of the most important corpora on Italian dialects. The team behind this project has developed a new software (NavigAIS) which enables browsing of a digitalized version of the AIS, thus allowing research based on geographical or lexical criteria.
- The *Dictionnaire Étymologique Roman* (DÉRom) [3] is a European project that involves several European scholars and especially the Universities of

[1] `http://asit.maldura.unipd.it/`
[2] `http://www3.pd.istc.cnr.it/navigais`

Nancy and Saarbrücken. This project goal is to digitalize the *Romanisches Etymologisches Wörterbuch* by W. Meyer-Lübke.
- A specific phonological database is the *UCLA Phonological Segment Inventory Database* (UPSID). This database is a statistical survey of the phoneme inventories in 451 human languages; it was created by Ian Maddieson in 1984 [4].

3 Notes on Phonology

Phonology is the linguistic field that studies the sounds of human languages, and it specifically deals with the sounds which have a distinctive relevance. A particular research field of phonology studies the diachronic processes responsible for phonetic mutation; in fact, languages change continuously and they evolve as time goes by.

One of the aims of phonology is to find and motivate the causes that brought a word of a parent language to develop into different shapes in several daughter languages. This particular approach to phonological change is adopted in this work, and it belongs to what is known as diachronic or historical phonology.

This work aims to account for two specific aspects of phonology: the chronological order of phonetic and phonological rules and the feature structure of sounds[3].

In phonology the ordering of rules represents a main issue, since different rule orders can result in different words. In fact the application of a certain rule can affect the following phonological processes by creating or deleting contexts of application. The rule ordering can be either synchronic or diachronic: since this project deals with historical phonology the speculation involves only the diachronic aspect.

The second crucial theoretical aspect that this work aims to underline is the feature structure: sounds are not the minimum unit of analysis, they can be separated into smaller units known as features. Features are the basic phonological units, they are categorized according to the natural class of the segments they describe and they are grouped in force as of their articulatory properties (see Figure 1).

The result that is expected to be obtained with this project is a digital library which is able to provide information about the entire process that leads to a word assuming different shapes in different languages from a common origin. The kind of data that will be provided by this application is not available in other databases or books since it represents the result of original linguistic research and processing of data. The innovative aspects of this work are both the design of the database structure and the aim of the instrument.

As a matter of fact, the main innovation of this project consists of the new way data are processed, which allows an immediate overview of the diachronic change of specific languages and to account for the rule order and, at the same

[3] For Autosegmental Theory see Halle M., Vaux B., Wolfe, A. (2000); for rule ordering see Kenstowicz, 1994 (pp. 89-135) [5] and Odden, 2005 (pp. 225-300) [6].

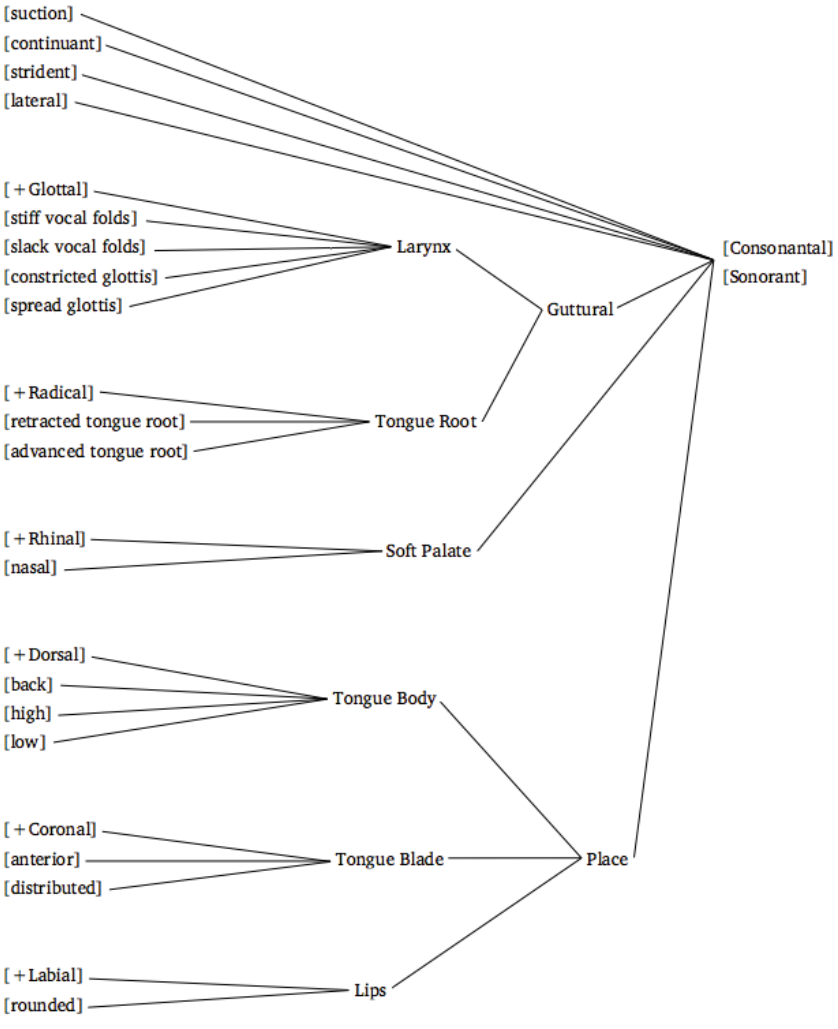


Fig. 1. Feature geometry according to Halle, Vaux and Wolfe (2000)

time, the dual nature of the rules (specific and general representations). It is important to notice that the linguistic data are processed and treated with a specialist approach and for this reason this instrument will be suitable for users who have a specific phonological knowledge.

Although this version of the database is filled with data regarding a specific romance reality (in particular this database treats the evolution of words from Latin to Italian and some northeastern Italian dialects), the database structure is designed to be filled with data from every human language.

4 Approach

4.1 Linguistic Approach

The aim of the instrument is to account for the phonological changes that occurred during the evolution of a parent language into one or more daughter languages. Theoretically, the structure of the database would allow it to represent phonological diachronic processes taken from any human language (except for tonal processes, which would require a slightly more complex structure), but of course a sample of linguistic varieties was required to test the efficiency of the instrument.

The chosen sample consists of standard Italian and dialectal varieties from Veneto and Friuli. Word stems are taken from different corpora and etymologic dictionaries:

- *Romanisches Etymologisches Wörterbuch* by Wilhelm Meyer-Lübke;
- *Sprach- und Sachatlas Italiens und der Südschweiz* (AIS) by Karl Jaberg and Jakob Jud;
- *Atlante Linguistico Italiano* by Ugo Pellis;
- *Lessico Etimologico Italiano* (LEI) by Max Pfister;
- *Dizionario Etimologico Storico Friulano* (DESF).

The first step in representing diachronic changes is to make it possible to identify the Latin word from which each present word derives. If the processing of the data stopped here, though, we would have nothing more than a digitalized corpus, which would certainly be useful, but not innovative; and that is why this database aims to explain how phonological processes take place, step by step.

The theoretical structure of the project is based on two assumptions: every human language is made up of sounds and every language changes over time.

Every human language is made up of sounds which are called phones. It is important to clarify that there is a limited set of executable phones, and every language selects its phonetic inventory from that common set (see Figure 2). Furthermore, every language also selects a set of phonemes as a subset of the phone set. Phonemes are distinctive mental units with which the human mind composes words. Phones are the surface realizations of underlying phonemes.

If two segments, occurring in the same context, cannot be swapped without changing the meaning of that word, those two segments have a distinctive value, and they are phonemes of that language.

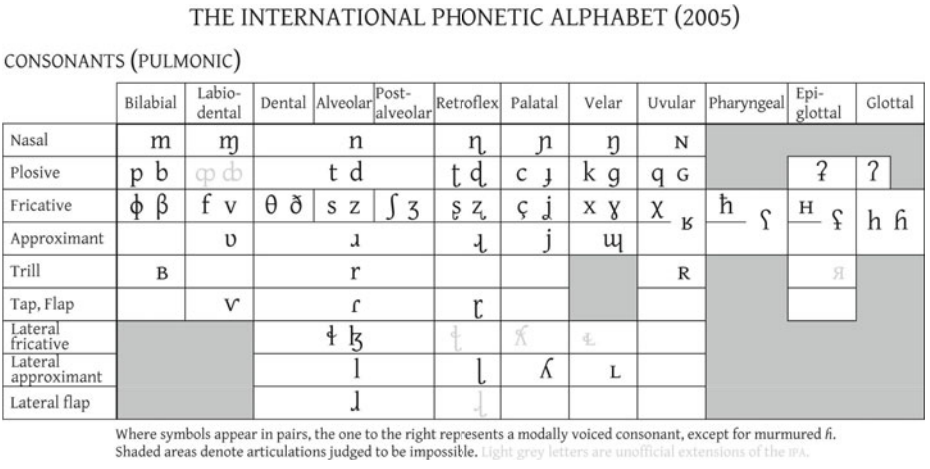


Fig. 2. Every human executable consonant according to the International Phonetic Association

As we can see in Table 1, [k] and [p] are Italian phonemes since it is impossible to swap the sounds [k] and [p] without changing the word meaning.

Table 1.

[kane] ∼ [pane]

It is important to underline that a sound can be a phoneme in one language and a simple phone in another, as we can see in Table 2. As a matter of fact [n] represents a phoneme both in Italian and English, whereas [ŋ] is an English phoneme but an Italian phone.

Table 2.

Italian	English
[mano] = mano	[θm] = thin
[aŋke] = anche	[θɪŋk] = think

Phonological change is determined by rules, in fact words change according to those rules, which are always applied in a given context. When its application context occurs, a rule always applies, and consequently phonological change is regular. The number of possible sounds is limited, and so is the number of natural phonological rules: potentially, all human languages are subject to the same rules.

Table 3.

Intervocalic context	Not intervocalic context
AURIC(U)LA > [orek:ja]	CLAVE(M) > [kjave]
SPEC(U)LU(M) > [spɛk:jo]	CLAUSU(M) > [kjuzo]
MAC(U)LA(M) > [mak:ja]	MAC(U)LA(M) > [mak:ja]

To prove that a rule is always applied in a given context we can consider the Italian evolution of the Latin consonant cluster [kl].

As we can see, the Latin consonantal cluster [kl] becomes Italian [k:j] if it occurs in intervocalic context, and [kj] anywhere else (in the beginning of the word or after another consonant). The two different contexts give slightly different outputs of the cluster. It is also essential to consider the order in which phonological phenomena occur, since the application of a rule can create or delete the context for the application of another rule; thus, some processes can only occur after the creation of their optimal context has been performed by a previous process. A practical example is provided by the case of the Latin word GENŬCULUM, which becomes *ginocchio* in Italian. The actual form of the Italian word is given by the specific order in which phonological rules have occurred over time: without the deletion of the second U, the consonants [k] and [l] would not have formed an intervocalic cluster, and thus would not have formed the Italian sequence [k:j] (see Table 4).

Table 4.

GENŬCULUM > ginocchio
[genukulum] > [dʒinɔkkjo]

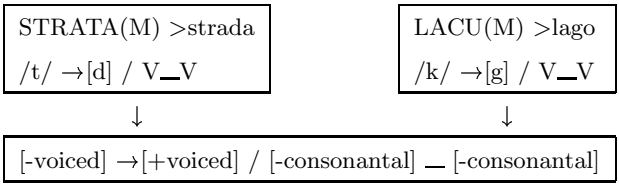
/g/ → [dʒ]
/e/ → [i]
/u/ → [ɔ]
/u/ → Ø
/kl/ → [k:j]
/u/ → [o]
/m/ → Ø

The structure of the database must also account for a deeper level of analysis. Every phoneme is defined by means of a set of distinctive features. Features are the smallest distinctive phonological unit and they are hierarchically organized following an articulatory criterion, as we have shown in Figure 1.

Feature representation of rules provides a wider generalization of the rules themselves: when a rule is expressed using distinctive features it can account for several specific rules (see Table 5).

Two word changes are examined in the provided example. In both the input is an unvoiced segment, which eventually becomes voiced. The two inputs /t/ and

Table 5.



/k/ are different, but they share fundamental articulatory characteristics: they both are stop consonants, and they both are unvoiced. Furthermore, they occur in the same context, i.e. intervocalically. The rule which is expressed through features describes both these cases, and thus it accounts for both of the specific rules.

Many specific phenomena can be explained with the same feature representations, which is why the database must provide a correct feature formalization of phonological rules.

In conclusion, the linguistic aspects this database must account for are:

- the chronological rule order;
- feature representation of segments and rules.

4.2 Database Development

The design of the conceptual schema (see Figure 3) required meticulous attention to every phonological aspect and for this reason the first phase of the design consisted of identifying and isolating the main entities, which are:

- Root: the Latin word from which the Romance words developed into different shapes;
- Word: the outcome that results from the phonological processes;
- Rule: the formalization of the phonological change in a linguistic code; it is always composed of a phenomenon which occurs in a context;
- Phenomenon: the effective linguistic change;
- Context: the linguistic environment in which a phenomenon takes place.

Starting from this general schema, a deeper structure was designed considering the two aspects of every phonological change: the concrete and the theoretical. For this reason the conceptual schema shows two separated but specular areas: the first area, shown in the upper part of Fig. 2 (composed of Context, Phenomenon and Rule), aims to represent the concrete phonological reality displaying a specific rule closely related to a particular word; instead the second area (composed of all the entities shown beneath the first area) groups a set of concrete rules into a general and abstract rule. This conceptual bipartition allows two levels of representation which are both essential in the phonological analysis.

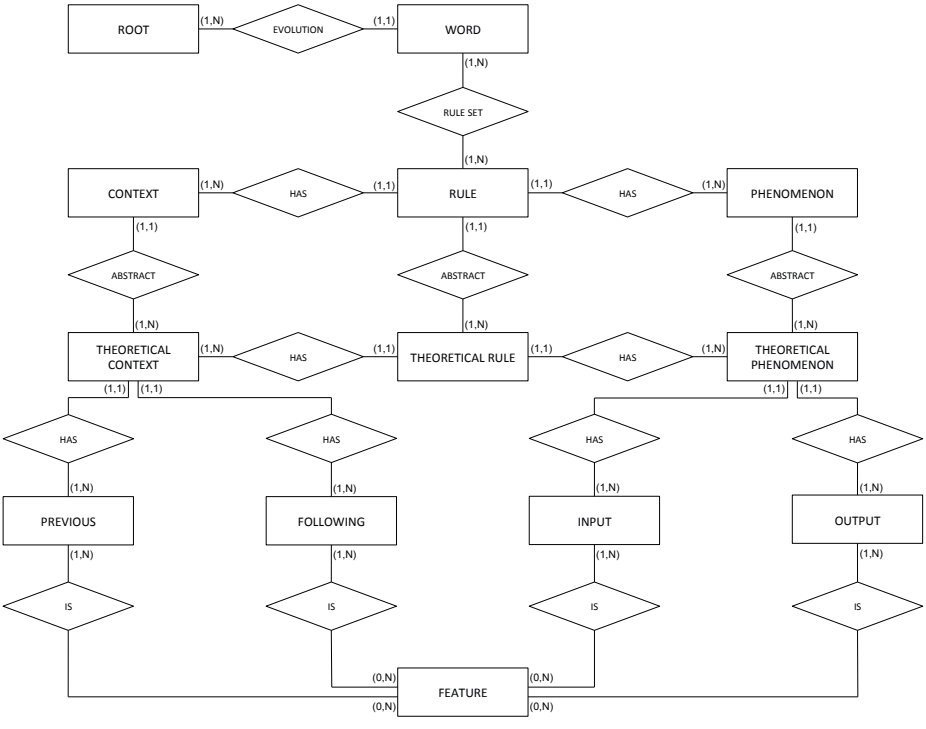


Fig. 3. Conceptual schema of the database

5 From the Database Application to a Digital Library

This project was conceived as a dissertation for a Master degree, but its development could go well beyond the planned shape it will have once completed. In other words, the completion of this project could be set much further along: although the expected outcome of this work will be a useful instrument for analyzing the language changes that occurred during the transition from Latin to the northeastern Italian dialects and Italian itself, the possible applications of this database are much broader.

The main possible implementation consists of a software application whose aim would be to allow any user to use the database with ease. As mentioned before, this instrument is conceived to be used mainly by linguistics experts, since it would be neither useful nor interesting to those who are not performing specialist research in this field. The first task the software will have to accomplish will be to make it easy for linguists to browse the content of the database and to make specific interrogations about the data.

A further development could involve a set of maps, on which the software should be able to draw isoglosses according to the queries performed by the users. Such an application would allow users to witness the spread of linguistic changes throughout a geographical area.

Lastly, a final version of this work should be made available on the web, in order to achieve the most far-reaching aim of the whole project: since the theoretical and practical model of this database is meant to suit all human languages, data insertion should be made possible for every scholar interested in sharing his knowledge. Such a collaboration would grant the digital library a huge amount of detailed information about an extremely wide range of languages from all over the world, thus multiplying and enhancing the opportunities to perform comparative linguistic studies and allowing analysis of the overall distribution of the phonological processes in a new and broader perspective. This objective may seem somewhat ambitious, and it would surely take some time before such a sharing process could begin, but once the guidelines for data insertion were set, the foundation for a common work would immediately be ready. Of course, these predictions need not be taken to their maximum extent: this kind of collaboration can be made on a smaller scale. At any rate, it would still be the starting point of a virtuous circle that would this instrument to reach many universities and, in doing so, help research in this field.

In this paper we have shown the first stage of the project, i.e. the database design. The implementations discussed above would enhance the functionality of the digital tool, and would make it an actual digital library with a broad collection of linguistic documents. Such a digital library, as we explained in this work, would not only allow users to easily witness the variety of languages and linguistic changes, but it would also help producing new knowledge by providing linguists new research methods.

Acknowledgments. This work is part of our MA dissertation projects which will be defended in the near future at the University of Padua.

We would like to thank prof. Maristella Agosti (Department of Information Engineering, University of Padua) and prof. Laura Vanelli (Department of Linguistics and Performing Arts, University of Padua), who are both supervising the projects.

References

1. Halle, M., Vaux, B., Wolfe, A.: On Feature Spreading and the Representation of Place of Articulation. Linguistic Inquiry 31, 387–444 (2000)
2. Agosti, M., Benincà, P., Di Nunzio, G.M., Miotto, R., Pescarini, D.: A Digital Library Effort to Support the Building of Grammatical Resources for Italian Dialects. In: Agosti, M., Esposito, F., Thanos, C. (eds.) IRCDL 2010. CCIS, vol. 91, pp. 89–100. Springer, Heidelberg (2010)

3. Buchi, E., Schweickard, W.: Le Dictionnaire Étymologique Roman (DÉRom): en Guise de Faire-part de Naissance. Lexicographica. International Annual for Lexicography 24, 351–357 (2008)
4. Maddieson, I.: Patterns of Sounds. Cambridge Studies in Speech Science and Communication. Cambridge University Press, Cambridge (1984)
5. Kenstowicz, M.: Phonology in Generative Grammar. Blackwell, Oxford (1994)
6. Odden, D.: Introducing Phonology. Cambridge University Press, Cambridge (2005)