AN INFRASTRUCTURE FOR STORING AND PROCESSING DIALECT DATA

Petya Osenova and Kiril Simov
The Sofia University and Linguistic Modeling Laboratory

The paper titled An Infrastructure for Storing and Processing Dialect Data discusses both – linguistic and software aspects when building databases with Bulgarian dialect data and navigating in it. The authors first focus on combining user and machine demands in adequate encoding of the data. Then, some basic techniques within CLaRK system are introduced, which facilitate various types of data processing depending on the task.

1. Introduction

In the field of dialectology the computerized storage of data is becoming more and more crucial for testing different linguistic hypotheses and for more objective generalizations in research. However, there are various problems to face before the achievement of an efficient infrastructure.

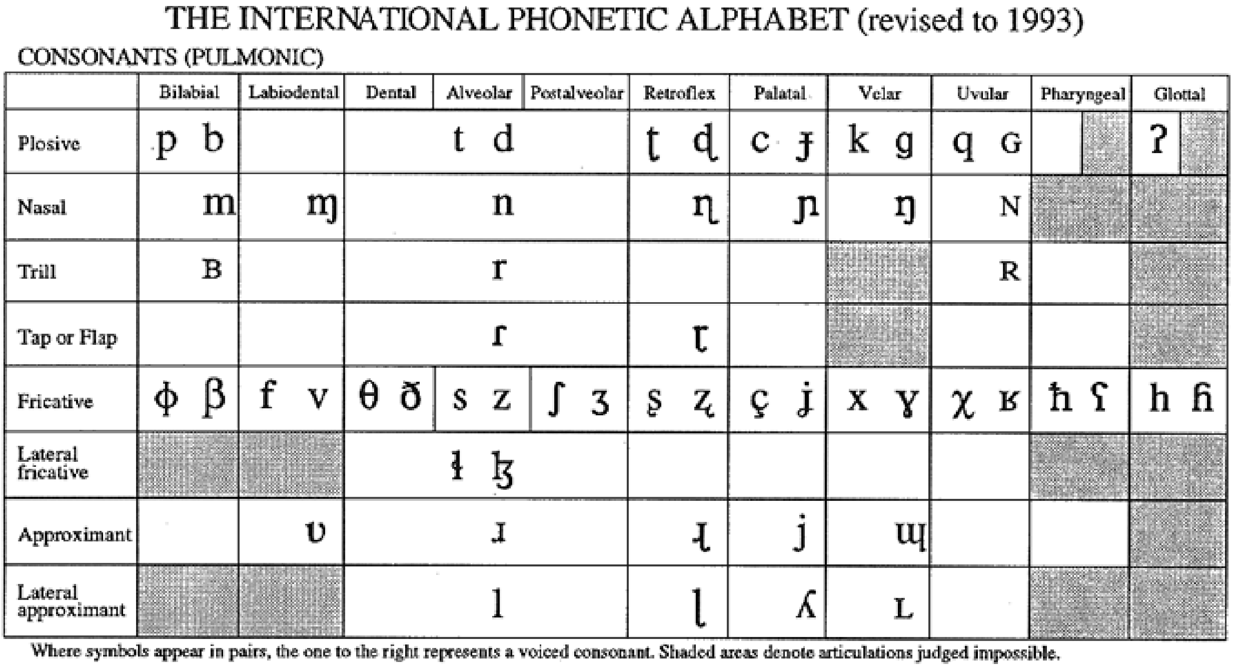
First of all, the dialect data is phonetically richer than the phonetic system of the standard language. This fact poses two problems - adequate linguistic and computer-friendly encoding. The first one is related to: a better coverage of all phonetic varieties and a maximally correct mapping to language presentation. The second is concerned with an appropriate re-encoding for easy processing by computer programs. Then, the question comes how to structure the data in a database which is easy to explore for various purposes. In this presentation we show an XML-based model for storing dialect data in CLaRK system. The data is stored in UNICODE, which makes it easily transferable. We can have several versions of the data depending on the task.

From content point of view the data can include different layers of linguistic information. For example, the concepts, the standard pronunciation, the base form, the recorded word form, the different pronunciations per site of the word form.

The paper is structured as follows: In the next section the problems of data storage are discussed. Section 3 presents in brief the CLaRK System and some of its tools that can be extensively used in data maintenance and processing. The last section concludes the paper.

2. Data storage

The dialect data should be stored with respect to two main parameters: linguistic adequacy, and software efficiency and portability. The first parameter allows the user to decide on the level of detailness according to the task. For example, the phonetician can choose whether to reflect three-level of consonant palatalness in the dialects (*palatal–semipalatal–nonpalatal*), or to underspecify and encode only the dichotomy *palatal-nonpalatal.* Then (s)he can decide what linguistic and non-linguistic information to add – lemma, wordform, sense, site, example etc. Additionally, a mapping to some International Standard is required (for example, the specifications of the International Phonetic Association (IPA 2003)) – see Figure 1. This mapping is not a trivial process, because: (1) the standard covers mostly literary languages and (2) there is no always one-to-one interpretation of the sounds in Bulgarian dialects to the IPA. Therefore, the important thing of the mapping is it to be consistent.



THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ	▨	ʔ ▨
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ	▨	▨
Trill	ʙ			r				▨	ʀ		▨
Tap or Flap				ɾ		ɽ		▨			▨
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative	▨	▨		ɬ ɮ						▨	▨
Approximant		ʋ		ɹ		ɻ	j	ɰ			▨
Lateral approximant	▨	▨		l		ɭ	ʎ	ʟ		▨	▨

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Figure 1

The second parameter, as it was said above, concerns software efficiency and portability of the data. Additionally to the adequate encoding of linguistic data, we would like to process the data easily in automatic or semi-automatic way. Therefore, we have to rely on a simpler encoding than IPA uses. This step is necessary in order to facilitate the work of both – the human and the computer. For that reason many people use X-SAMPA encoding (see Wells), which, on the one hand, is based on the combination of only ASCII symbols, and on the other hand, is convertible to the standard of IPA – see below Figure 2 for an illustrative encoding of nasals in X-SAMPA.

Nasals

bilabial nasal	**m**
labiodental nasal	**F**
dental/alveolar nasal	**n**
retroflex nasal	**n’**
palatal nasal	**J**
velar nasal	**N**
uvular nasal	**N**

Figure 2

In order to see the difference between the two encodings more clearly, we also enclose the following small table in Figure 3.

IPA (Unicode)	X-SAMPA (ASCII)
æ	{
ɛ	E

Figure 3

In the table above we can see that in IPA encoding the symbols are more linguistically motivated, while in X-SAMPA conversion they are computer friendly and hence are easily tractable. Recall that these two encodings are mapped to each other and therefore, there is no problem to change one into the other, and vice versa, when necessary.

3. Data processing in CLaRK

CLaRK is an XML-based software system for corpora development. It incorporates several technologies: XML technology; Unicode; Regular Grammars and Constraints over XML Documents (see more about the facilities of the system at: http://www.bultreebank.org/clark/index.html and in (Simov et. al. 2004)). Let us show some practical processing steps that can be performed within the system.

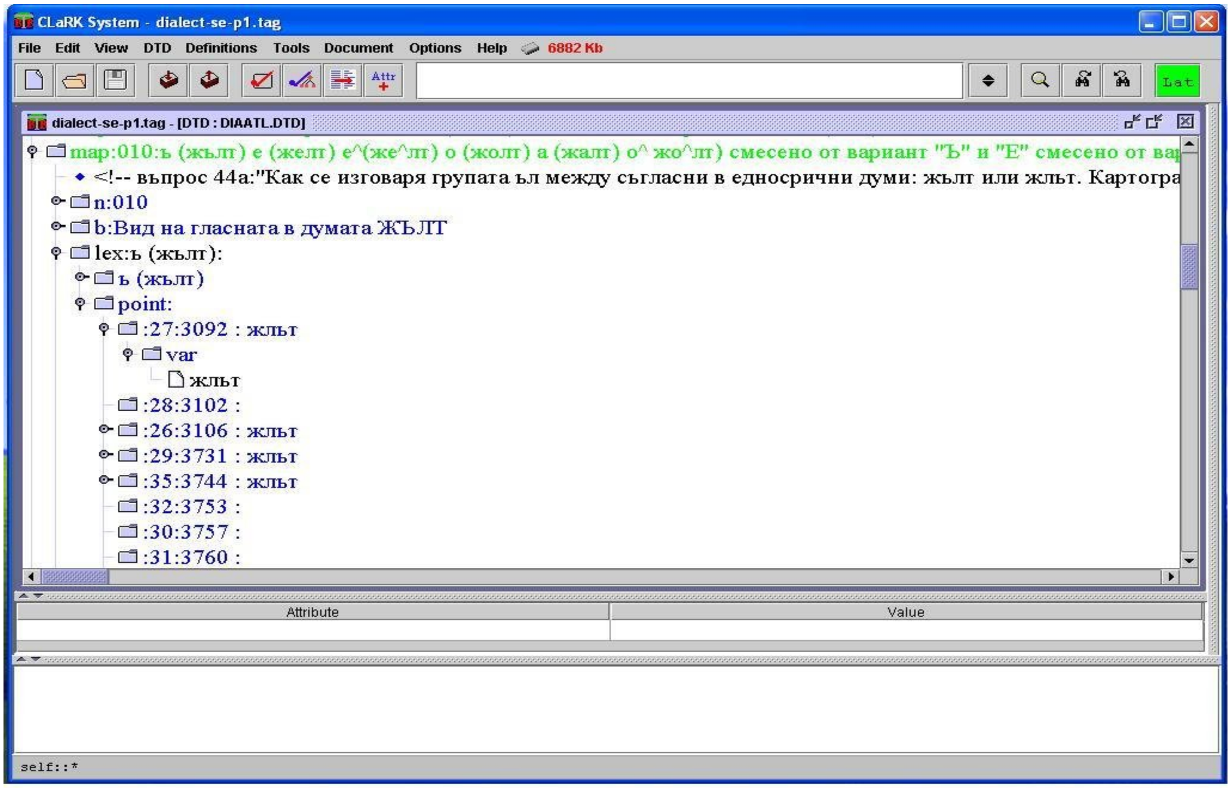
3.1. Structuring the data in XML format

To structure the data in a database, which is easy to explore for various purposes, we rely on XML. From content point of view the data can include the concept, the standard pronunciation, the base form, the recorded word form, the different pronunciations per site of the word form. The richer the data, the better, because we can remove anytime the unnecessary information or transform, add, derive it in different formats. Let us take an example. The targeted format for further processing of the data is as follows:

```
# zhylt
* 1
: Standard
- Z@lt
: Brezovo
- Z@lt
: Gradina
- Z@lt
: Seltsi
- Z@lt
```

Here we have the adjective `zhylt’ (yellow) in X-SAMPA encoding, as it is pronounced in some Bulgarian sites. The model is as follows: first, the site is given, and then the pronunciation of the word. For clarity, we included the Standard pronunciation as well. First, we have encoded these pronunciations, as it is shown in Figure 4[1]. There we can see the sites encoded as numbers together with the pronunciations of the word. Note that more information than that is present in the XML representation. For example, the content of the question on this linguistic phenomenon, the number of the map in the dialect atlas, the variants of the word in each site. As the next step is conversion to X-SAMPA, we will discuss it in the next subsection. Below, in Figure 5, we will show only the result from CLaRK's Constraint Manager, Insert and Remove operations.

[1] The illustrative data is from (Stoykov and Bernshteyn 1964).

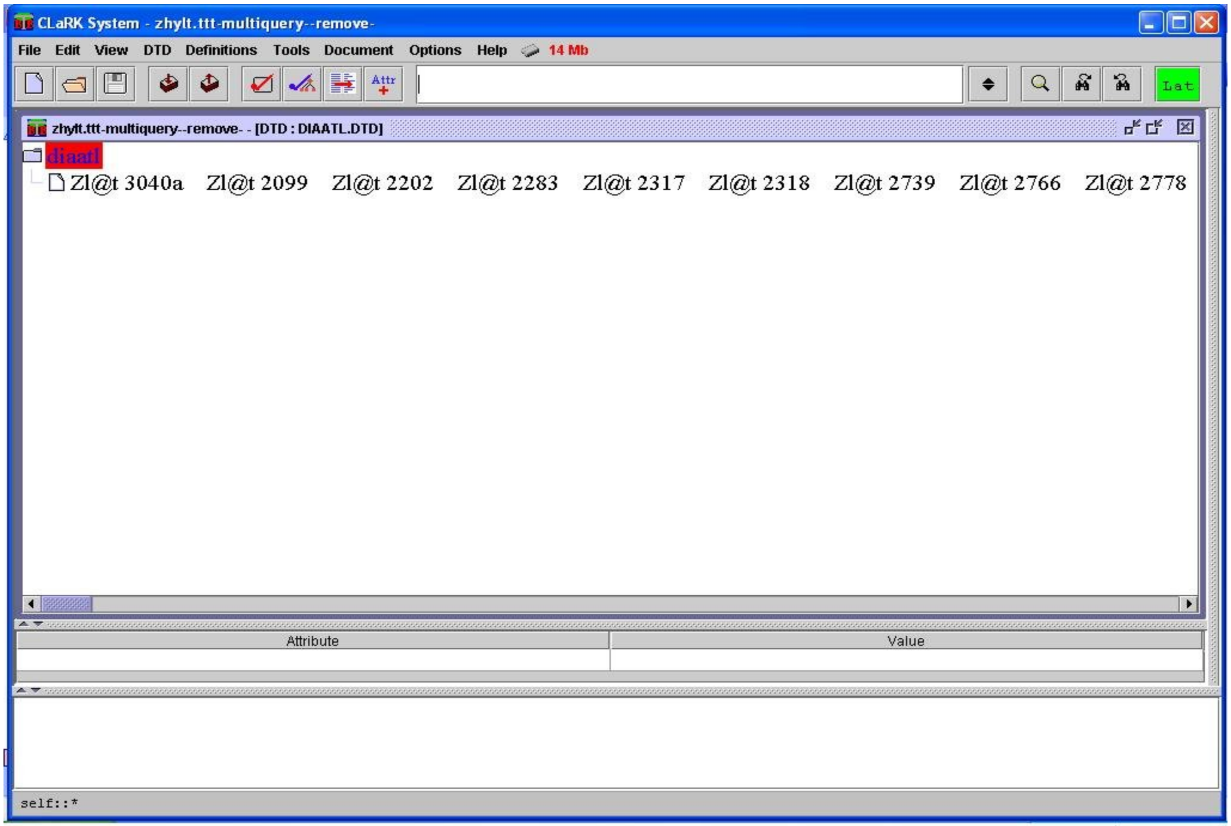

CLaRK System - dialect-se-p1.tag
File Edit View DTD Definitions Tools Document Options Help 6882 Kb
Attr
Lat
dialect-se-p1.tag - [DTD : DIAATL.DTD]
Attribute
Value

```
map:010:ъ (жълт) е (желт) е^(же^лт) о (жолт) а (жалт) о^ жо^лт) смесено от вариант "Ъ" и "Е" смесено от ва
  <!-- въпрос 44а:"Как се изговаря групата ъл между съгласни в едносрични думи: жълт или жлът. Картогра
  n:010
  b:Вид на гласната в думата ЖЪЛТ
  lex:ъ (жълт):
    ъ (жълт)
    point:
      :27:3092 : жлът
        var
          жлът
      :28:3102 :
      :26:3106 : жлът
      :29:3731 : жлът
      :35:3744 : жлът
      :32:3753 :
      :30:3757 :
      :31:3760 :
```

```
self::*
```

Figure 4

In order to structure the data in this way, we need a DTD (Data Type Declaration), which to constrain the appropriate elements in XML. In the picture above we can see that there are several elements: **map, n** (number of the map in the atlas), **b** (the short description of the phenomenon), **lex** (the lexeme), **point** (the site, or in other words, the village), **va**r (the variation of the lexeme as pronounced in the site in question).


CLaRK System - zhylt.ttt-multiquery--remove-
File Edit View DTD Definitions Tools Document Options Help 14 Mb
Attr
Lat
zhylt.ttt-multiquery--remove- - [DTD : DIAATL.DTD]
Attribute
Value

diaatl

Zl@t 3040a Zl@t 2099 Zl@t 2202 Zl@t 2283 Zl@t 2317 Zl@t 2318 Zl@t 2739 Zl@t 2766 Zl@t 2778

```
self::*
```

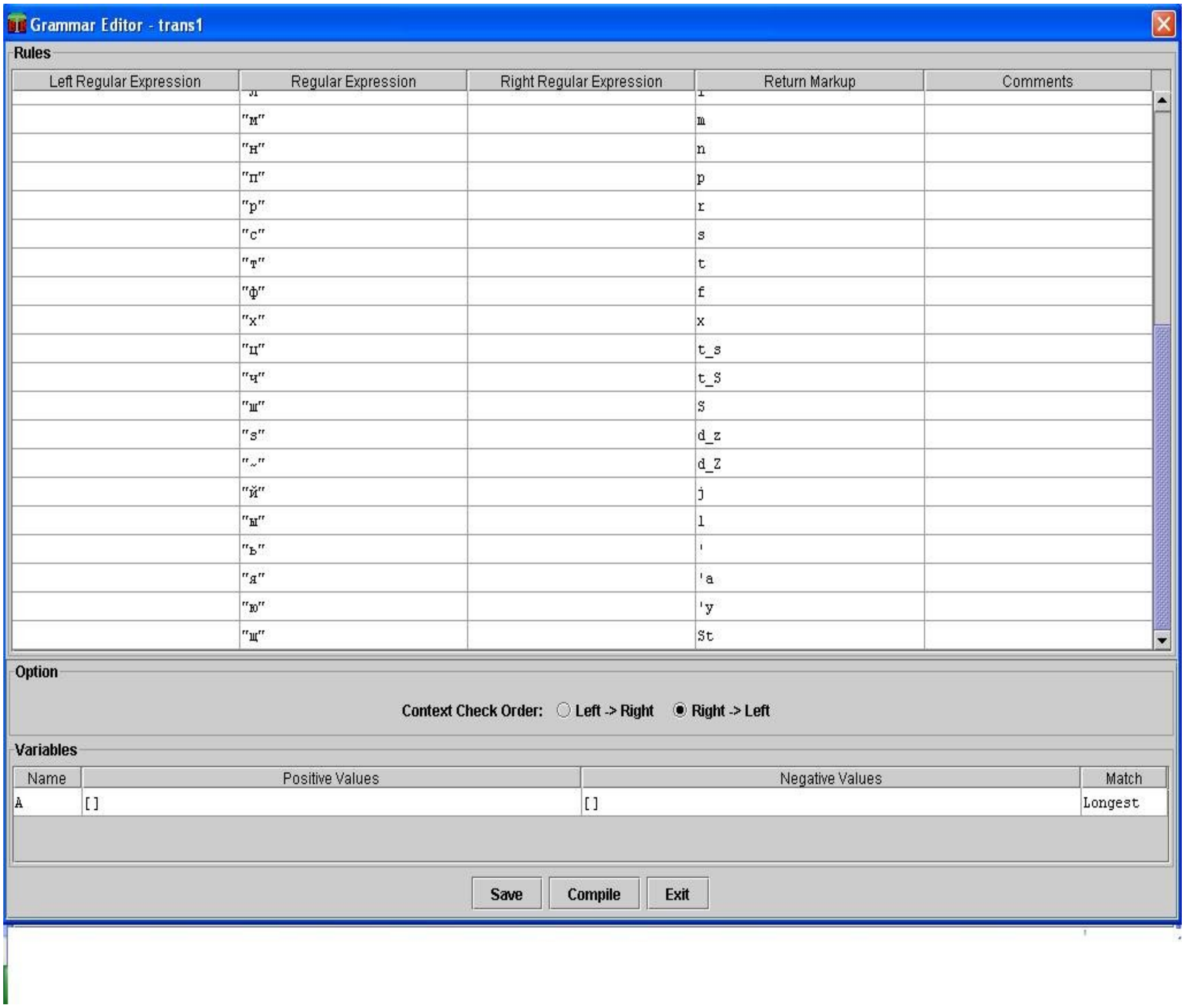
Figure 5

3.2. Conversion of the data from some ‘working digitization form’ into X-SAMPA format

The conversion into X-SAMPA can be performed by a set of grammars in a cascaded order. First the grammar with the longest match is executed. The regular expression rules are of the following kind:

```
<RE>"w"</RE> <RM>B</RM>
```

RE means regular expression and RM stands for Return mark-up. In this way every instance of a letter in the dialect transcriptions is converted to the corresponding X-SAMPA symbol. In the above rule this is illustrated by the fricative bilabial sound "w", whose corresponding X-SAMPA symbol is B. For example, see Figure 6 below.



Left Regular Expression	Regular Expression	Right Regular Expression	Return Markup	Comments
	"л"		l	
	"м"		m	
	"н"		n	
	"п"		p	
	"р"		r	
	"с"		s	
	"т"		t	
	"ф"		f	
	"х"		x	
	"ц"		t_s	
	"ч"		t_S	
	"ш"		S	
	"ѕ"		d_z	
	"~"		d_Z	
	"й"		j	
	"ы"		l	
	"ь"		'	
	"я"		'a	
	"ю"		'y	
	"щ"		St	

Context Check Order: ○ Left -> Right ● Right -> Left

Name	Positive Values	Negative Values	Match
A	[]	[]	Longest

Figure 6

Each symbol in the column “Regular expression” receives its counterpart in X-SAMPA, which is in the column “Return Markup” in the Grammar Manager.

4. Conclusion

In this paper we discussed the need of a rich dialect database, from which to take the required information when necessary. In our opinion, the data should be conformant to some international phonetic standard like IPA. At the same time, this standard should be convertible to another, simple enough format, which to ensure a good processing of the data without losing or damaging the linguistic adequacy of the encoded information.

Bibliography

IPA 2003: Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge University Press.
Simov et. al. 2004: Simov K., Simov A., Ganev H., Ivanova K., Grigorov I. The CLaRK System: XML-based Corpora Development System for Rapid Prototyping}. In: Proceedings of LREC 2004, Lisbon, Portugal, pp. 235—238
Stoykov and Bernshteyn 1964: Stoykov St. and Bernshteyn S. B. Atlas of Bulgarian Dialects: Southeastern Bulgaria, Publishing House of the Bulgarian Academy of Sciences, 1964, volume I, Sofia, Bulgaria. (In Bulgarian)
Wells: Wells John. Computer-coding the IPA: a proposed extension of SAMPA. Available at: http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm