View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Bulgarian Digital Mathematics Library at IMI-BAS

Serdica J. Computing **1** (2007), 73–86

Serdica
Journal of Computing

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

APPLYING A NORMALIZED COMPRESSION METRIC TO THE MEASUREMENT OF DIALECT DISTANCE

Kiril Simov, Petya Osenova

Abstract. The paper discusses the application of a similarity metric based on compression to the measurement of the distance among Bulgarian dialects. The similarity metric is defined on the basis of the notion of Kolmogorov complexity of a file (or binary string). The application of Kolmogorov complexity in practice is not possible because its calculation over a file is an undecidable problem. Thus, the actual similarity metric is based on a real life compressor which only approximates the Kolmogorov complexity. To use the metric for distance measurement of Bulgarian dialects we first represent the dialectological data in such a way that the metric is applicable. We propose two such representations which are compared to a baseline distance between dialects. Then we conclude the paper with an outline of our future work.

1. Introduction. The measurement of the distance between dialects has many applications in the area of dialectology and recently in the area of language contacts. The latter will be of a great importance with respect to the development of a new regional communication which earlier was interrupted by political restrictions such as state borders. We can consider the measurement of

ACM Computing Classification System (1998): J.5.
Key words: Kolmogorov Complexity, compression metric, dialect distance, language contacts

dialect distance in two respects: (1) how much effort a speaker of one dialect has to invest in order to understand a speaker of another dialect; and (2) what common information is shared by the speakers of the corresponding dialects. In this paper we consider the second option by applying an information-based method for distance measurement.

The paper discusses the application of a similarity metric based on compression to the measurement of the distance among Bulgarian dialects. A very appealing property of the compression metric is the fact that it is not necessary to define which features of the data are used for the measurement. Ideally, the method has to take into account all features of the data. In practice, the method gives good results only for large data sets, because the real compressors (like WinZip, WinRar, 7-zip, etc) are only approximations of the ideal ones. Thus, one problem with the application of the method to dialect data and, accordingly, language contact phenomena, is that there is not enough natural data available for the method.

To overcome this problem we defined two new methods for the creation of data sets with dialectological data. The data sets constructed by these two methods are large enough to allow the application of the method based on a compression metric. Here we assume that the information about the dialects is represented in digital form. Thus, in fact we measure the distances between the files for the corresponding dialects.

The next step in our investigation will be to check the applicability of the method to data on language contacts. The direct comparison by word-based methods seems to be applicable only to languages from the same language family like Bulgarian and Serbian, or German and Dutch, because they share a lot of words with the same root. In such cases one can expect that the word itself provides good context for comparison. But in the case of languages from different language families these methods do not seem to be applicable because of the great lexical diversity. Additionally, it is not always clear where the boundary is between measuring two dialects and measuring two languages. Also, it is not always easy to control the data in case of having only phonetic or only lexical variants. As a consequence, the data becomes noisier for interpretation. In order to overcome this problem more elaborate methods for generation of data sets are discussed below.

The structure of the paper is as follows: the next section presents the similarity metric based on compression; then we present the data used in our experiments; in the following section some preliminary experiments are discussed and the two ways of generating dialectological data are presented; the last section concludes the paper and outlines some future work.

2. Similarity metric based on compression. Ideally, compression has to delete all the redundancy from the file. Then the compressed file will contain the really important information for the original file. Intuitively, if we have two files that share some information, then the compression of the concatenation of the two files will be smaller if the two files share more information and it will be bigger if the files do not have much information in common. This intuition is captured by the metric presented here. The metric is defined on the basis of the Kolmogorov complexity of a file, which is the ultimate compressed form of the file. A metric defined in this way is information-based inasmuch as the Kolmogorov complexity was proposed by Kolmogorov as a measure of the information in a file (or more exactly in a binary string). A presentation of his work on information theory can be found in [2] and [3]. He defined the quantity of information in a binary string to be the size in bits of the shortest program which can reproduce the string. Thus, the Kolmogorov complexity could be considered as the ultimate compression of the string.

The metric described here is reported in [1]. They consider as their goal the definition of a Non-Feature Similarity which is a single similarity metric for all features represented in a file. The advantage of the non-feature similarity is that it is not necessary for the features on which the similarity is defined to be determined in advance, they are discovered by the similarity itself. This advantage might also be a problem because we cannot be sure which features represented in the two files are reasons for the similarity or dissimilarity between them.

Similarity metrics are defined as a distance function $d(.,.)$ such that: $d(a,b) = 0$ iff $a = b$; $d(a,b) = d(b,a)$ (symmetry); $d(a,b) \leq d(a,c) + d(c,b)$ (triangle inequality). Additionally, the density condition should be met: for each object there are objects at different distances from it; and the normalization condition: the distance between two objects depends on the size of the objects. Distances are in the interval $[0,1]$.

As we mentioned above, the metric defined here, is based on the notion of Kolmogorov complexity. Let x be a file, then $k(x)$ (Kolmogorov complexity of x) is the length in bits of the ultimately compressed version of the file x. In order to define a metric the notion of conditional Kolmogorov complexity is used: $k(x|y)$ is the length of the ultimately compressed version of y if the compressed version of x is available. That is, the compress version of x is used to compress y. As an approximation of $k(x|y)$ $k(x,y) = k(xy)$ is used – the Kolmogorov complexity of the concatenation of the two files. $k(x,y)$ is almost a similarity metric:

$$\begin{aligned} k(x,x) &= k(xx) \approx k(x); \\ k(x,y) &= k(y,x); \\ k(x,y) &\leq k(x,z) + k(z,y). \end{aligned}$$

It is necessary to normalize it in order to turn it into a similarity metric. We can do this by the following sequence of equivalent inequalities:

$$\min(k(x), k(y)) \leq k(x, y) \leq k(x) + k(y);$$
$$0 \leq k(x, y) - \min(k(x), k(y)) \leq k(x) + k(y) - \min(k(x), k(y));$$
$$0 \leq k(x, y) - \min(k(x), k(y)) \leq \max(k(x), k(y));$$
$$0 \leq (k(x, y) - \min(k(x), k(y)))/ \max(k(x), k(y)) \leq 1.$$

Therefore, we define the Kolmogorov similarity metric $ksm(x, y)$ by the following formula:

$$ksm(x, y) = (k(x, y) - \min(k(x), k(y)))/ \max(k(x), k(y)).$$

This metric uses all the information represented in the two files x and y in order to determine the distance between them. One unfortunate fact is that the Kolmogorov complexity is undecidable problem. Thus, the above metric cannot be used in practice. It can be only approximated by a similar metric which uses a real life compressor c. Therefore, [1] defined a normalized compression distance $ncd(.,.)$ which is an approximation of the Kolmogorov metric. It is defined by the following formula:

$$ncd(x, y) = (c(x, y) - \min(c(x), c(y)))/ \max(c(x), c(y))$$

where $c(x)$ is the size of the compressed file x. Of course, the properties of $ncd(.,.)$ depend crucially on the properties of the compressor c. In order to determine the properties of $ncd(.,.)$ the authors defined the notion of a normal compressor. The compressor c is normal if it satisfies (asymptotically to the length of the files):

1. Stream-basedness: compress first x, then y;
2. Idempotency: $c(xx) = c(x)$;
3. Symmetry: $c(xy) = c(yx)$;
4. Distributivity: $c(xy) + c(z) \leq c(xz) + c(yz)$.

If c is normal, then $ncd(.,.)$ is a similarity metric. The stream-basedness of the compressor means that it produces first the compressed version of the file x and then the compressed version of y using the compressed version of x, i.e. the compressor uses the information in x in order to compress y. The idempotency ensures that the compressor is capable of recognizing the repeating structures in the file. The symmetry is required to guarantee that the compressor uses the common information of x and y when it is compressing the concatenation of both files. The distributivity ensures that $ncd(.,.)$ will meet the triangle inequality. Based on a normal compressor $ncd(.,.)$ defines a non-feature similarity metric

because it uses (almost) all information presented in the files in order to compare them.

The normality conditions on the compressor c are used in order to select the best real life compressor for our experiments. The best approximation to a normal compressor we received for the 7-zip and rar compressing programs. In the next part of the paper we proceed in the following way. First, we describe the dialectological data that we used in our experiments, then we report on the results from two different generations of dialectological data.

3. The Bulgarian dialectological data[1]**.** The data was digitized from the four volumes of Bulgarian dialect atlases which cover the entire country area: Volume I – Southeastern Bulgaria [11], Volume II – Northeastern Bulgaria [10], Volume III – Southwestern Bulgaria [12] and Volume IV – Northwestern Bulgaria [13]. Unlike similar atlases for other languages, the data is gathered only from villages with exclusively Bulgarian populations regardless of geography. This means that the sites are not distributed uniformly within Bulgaria. For example, because most of the original Bulgarian sites with uniform dialects are in the mountains, there are more mountainous sites than non-mountainous ones. We used 490 dialect sites within Bulgaria, and we included the Standard pronunciation as well.[2] The sites were selected with respect to two main criteria: good coverage of the entire area presented in the atlas, and a representative number of varieties and subvarieties. There are 1682 sites altogether. This means that roughly one third of the sites is presented in our data. Then, we used all the 36 words which are common for all the sites. The list of these words is as follows:[3]

бъчва	ˈbətʃva	'barrel'	зълва	ˈzəlva	'sister-in-law'
дошъл	doˈʃəl	'has come-he'	жълт	ʒəlt	'yellow'
зъб	zəb	'tooth'	събота	ˈsəbota	'Saturday'
къща	ˈkəʃta	'house'	бяла	ˈbʲala	'white'-fem
бели	ˈbɛli	'white'-pl	язди	ˈjazdi	'ride'-3per
неделя	nɛˈdɛlʲja	'Sunday'	млекар	mlɛˈkar	'milkman'
грешка	ˈgrɛʃka	'mistake'	венчило	vɛnˈtʃilo	'married life'
ключ	klʲutʃ	'key'	чаша	ˈtʃaʃa	'glass; cup'

[1]This section is based on [6].

[2]The standard pronunciations are in accordance with [8].

[3]First, the Cyrillic presentations of the words are given, then their phonetic correspondences in IPA (International Phonetic Alphabet) and a translation into English.

път	pət	‘road’	жаби	ˈʒabi	‘frogs’
нощви	ˈnoʃtvi	‘hutch’	поляна	poˈlʲana	‘glade’
овче	ˈovtʃe	‘sheep’s’	тънко	ˈtənko	‘thin-neut’
гуляй	guˈlʲaj	‘feast’	овчар	ovˈtʃar	‘shepherd’
кон	kon	‘horse’	сън	sən	‘dream’
отишъл	otiˈʃəl	‘has gone-he’	вътре	ˈvətrɛ	‘inside’
тенджера	ˈtɛndʒɛra	‘pot’	джоб	dʒob	‘pocket’
няма	ˈnʲama	‘there is no’	череша	tʃɛˈrɛʃa	‘cherry’
гръб	grəb	‘back’	живя	ʒiˈvʲa	‘lived’
сол	sol	‘salt’	ден	dɛn	‘day’

These 36 words highlight important features for Bulgarian dialect classification. The expectations are that any good measurement of dialect distance given these words as a base for the comparison of Bulgarian dialects will give a satisfactory result with respect to the expert division of these dialects. In the rest of the paper we present methods that show the applicability of the normalized compression metric to the data.

4. Experiments on Bulgarian dialectological data. We have divided our experiments in two phases. During the first phase we studied the properties of some of the common real life compressors in order to gain an idea of which one satisfies the normality conditions and how we can represent the dialectological data in order to have good results. Thus, before our experiments with the real dialectological data we made several preliminary experiments with different kinds of text from the BulTreeBank corpus ([9]). The best results were obtained with the 7-zip and rar compressors. For that reason these compressor were further used in our work. Here are some of our findings:

- *Good results can be obtained only for large data sets.* Even the Kolmogorov metric (ideal compressor) depends on the size of the files. Thus, in order to have a good measurement we need longer representative files;
- *Each feature in the data set is a basis for a comparison.* Here we need to balance between the non-feature basedness of the method and minimization of the impact of the unimportant features represented in the files;

- *Most compressors are byte-based, thus some intra-byte features cannot be captured well.* Generally this observation means that features that are defined within one byte could have no impact on the result of the metric. Thus, the important features have to be encoded in several bytes. For example, the fact that the phonemes b and p have many common features will not be captured if they are represented simply as b and p, but they need a more elaborate representation like p' and p;

- *Systematic repetition (even on large scale) in the data is captured by the compressors.* A normal compressor very easily captures the repetition of strings. (Idempotency);

- *Some insignificant reordering of the data does not play a role for the size of the compression.* Changes in the order within small context do not cause changes in the size of the compression.

The conclusion from these findings is that in order to use this metric we needed large dialectological, naturally created data sets. Unfortunately, such data sets for dialects are missing. Thus, one option would be trying to create such data by simulating 'naturalness'. Consequently, we decided to generate dialectological 'texts' for each Bulgarian dialect. The methods for generation of these 'texts' have to reflect the above observations about the real life compressor. The texts need to encode the features in an explicit manner (something already done in the representation of the words in IPA), the order of the words in the text needs to not be easily predictable in order for the compressed files to be bigger, the unimportant features have to be presented in the texts for the different dialects in the same way in order to minimize their impact on the measurement. As a main unimportant feature we determined the order of the words in the generated texts, thus, we applied the same order of the words for each dialect.

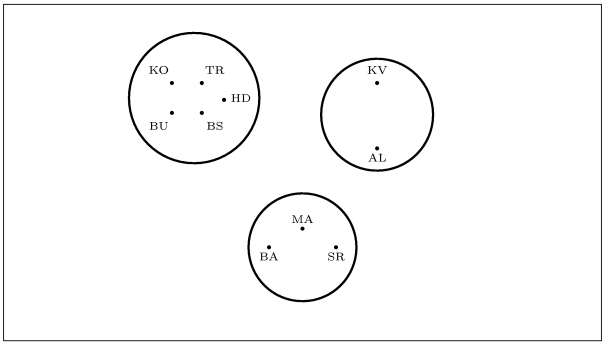
Recall that we relied on the 36 words used in the experiments performed by Petya Osenova, Wilbert Heeringa and John Nerbonne, at Groningen University. We used this set as an initial base for further generation. We selected ten villages which were grouped in three clusters by the methods developed in Groningen:

[Alfatar, Kulina-voda]

[Babek, Malomir, Srem]

[Butovo, Bylgarsko-Slivovo, Hadjidimitrovo, Kozlovets, Tsarevets]

The clusters are depicted graphically in the following picture. The names of the villages are abbreviated: AL (Alfatar), BA (Babek), BU (Butovo), BS



(Bylgarsko-Slivovo), HD (Hadjidimitrovo), KO (Kozlovets), KV (Kulina-voda), MA (Malomir), SR (Srem), TR (Tsarevets).

These clusters will be used as a baseline in order to check whether our methods for generation of dialectological texts and the normalized compression distance give similar results. We generated dialectological 'texts' in two ways, described in the following subsections.

4.1. Corpus-based text generation. Ideally we need naturally occurring texts for each dialect, but as we have mentioned, such do not exist and it is very hard to collect them, because there have to be special records. Unfortunately, there are not enough published texts in various dialects. In order to generate a 'text' which is as closer as possible to a natural text we decided to use a corpus of texts in the standard Bulgarian language which is available. Thus, we have performed the following steps:

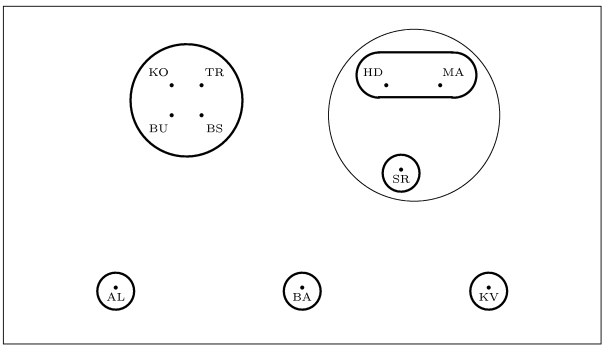
- From a corpus of about 55 million words we deleted all word forms except for the 36 from the list;
- Then we concatenated all the remaining word forms in one document;
- For each dialect we substituted the normal word forms with the corresponding dialect word forms;
- Then we applied the $ncd(.,.)$ metric to these texts.

The repetition and the order of the dialect word forms follow the repetition and the order of the standard word forms in the original corpus. Thus, they are not easily predictable and the result is close to a natural text. Therefore, the generated text is a good basis for comparison of the dialects. The following is the distance matrix calculated on the basis of these texts:

v/v	AL	BA	BU	BS	HD	KO	KV	MA	SR	TR
AL	0	0.9583	0.9673	0.9675	0.9626	0.9675	0.9915	0.9583	0.9677	0.9675
BA	0.9583	0	0.9894	0.9896	0.9875	0.9896	0.9928	0.9848	0.9839	0.9896
BU	0.9673	0.9894	0	0.0366	0.6214	0.0365	0.9735	0.6634	0.5072	0.0365
BS	0.9675	0.9896	0.0366	0	0.6245	0.0023	0.9738	0.6624	0.6598	0.0023
HD	0.9626	0.9875	0.6214	0.6245	0	0.6249	0.9699	0.466	0.7584	0.6249
KO	0.9675	0.9896	0.0365	0.0023	0.6249	0	0.9738	0.6623	0.5067	0.0022
KV	0.9915	0.9928	0.9735	0.9738	0.9699	0.9738	0	0.9749	0.9791	0.9729
MA	0.9583	0.9848	0.6634	0.6624	0.466	0.6624	0.9749	0	0.7057	0.6605
SR	0.9677	0.9839	0.5072	0.6598	0.7584	0.5067	0.9791	0.7057	0	0.5202
TR	0.9675	0.9896	0.0365	0.0023	0.6249	0.0022	0.9729	0.6605	0.5202	0

Here the names of the villages are abbreviated as follows: AL (Alfatar), BA (Babek), BU (Butovo), BS (Bylgarsko-Slivovo), HD (Hadjidimitrovo), KO (Kozlovets), KV (Kulina-voda), MA (Malomir), SR (Srem), TR (Tsarevets). The cell in row i and column j contains the distance between the village for row i and the village for column j, where $1 \leq i \leq 10$ and $1 \leq j \leq 10$. Row 0 and column 0 contain the names of the villages.

Here are the clusters we received according to these texts:



Here we can see two non-singleton clusters [Butovo, Bylgarsko-Slivovo, Kozlovets, Tsarevets] and [Hadjidimitrovo, Malomir]. The other villages form singleton clusters. [Srem] is close to the second cluster [Hadjidimitrovo, Malomir].

The experiment shows that the really close dialects are also grouped together. The divergence in the other clusters is because of the role of the word frequency in the generated texts. More frequent word forms play a bigger role. For example, the word form *няма* ('there is no') appears 106246 times vs. *млекар*

('milkman') – 5 times from 230100 word forms. Thus, the features represented by *няма* influence the result to a greater degree than the features represented by the word form *млекар*. This experiment posed an important question: *What is the role of the frequency of a given phenomenon in the dialect distance measurement?*[4] The adequate impact of the frequency factor seems to depend on two conditions: (1) availability of a bigger set of dialectological word forms; (2) good reflection of the natural distribution of the dialect features.

4.2. Permutation-based text generation. The second method relies on the idea of generating non-predictable text chunks. In order to do this we performed the following steps:

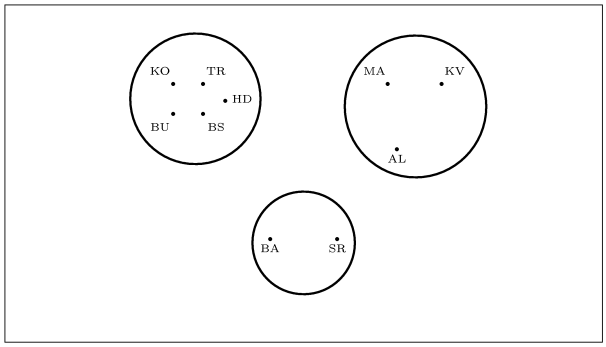
- All 36 words were manually segmented into meaningful segments. In our case the segments correspond to the phonemes in the representation of the words;
- Then for each site we made all permutations for each word, and concatenated them.

The permutation of the segments in a word representation ensures good (non-predictable) non-repetitiveness of the segments in the data and sufficiently large files (a high number of possible permutations). Thus, the generated dialectological data sets are good for the application of the compression metrics. Here is the distance matrix for this experiment:

v/v	AL	BA	BU	BS	HD	KO	KV	MA	SR	TR
AL	0	0.7149	0.5077	0.4832	0.6557	0.5319	0.5701	0.4321	0.6992	0.4793
BA	0.7149	0	0.6588	0.6327	0.573	0.7067	0.5513	0.5111	0.2886	0.6384
BU	0.5077	0.6588	0	0.078	0.3616	0.1485	0.7231	0.633	0.717	0.079
BS	0.4832	0.6327	0.0783	0	0.3152	0.0999	0.7838	0.6615	0.7534	0.014
HD	0.6557	0.573	0.3616	0.3152	0	0.3606	0.7149	0.6684	0.6379	0.2591
KO	0.5319	0.7067	0.1485	0.0999	0.3606	0	0.7515	0.746	0.7449	0.0587
KV	0.5701	0.5513	0.7231	0.7838	0.7149	0.7515	0	0.4227	0.5884	0.6791
MA	0.4321	0.5111	0.633	0.6615	0.6684	0.746	0.4227	0	0.5783	0.6192
SR	0.6992	0.2886	0.717	0.7534	0.6379	0.7449	0.5884	0.5783	0	0.6436
TR	0.4793	0.6384	0.079	0.014	0.2591	0.0587	0.6791	0.6192	0.6436	0

[4]The frequency of phonemes in dialect distance measuring is included as a parameter in methods like *Corpus frequency method* — see [4]. The problem arising in our work is how the corpus is compiled.

As above, the names of the villages are abbreviated as follows: AL (Alfatar), BA (Babek), BU (Butovo), BS (Bylgarsko-Slivovo), HD (Hadjidimitrovo), KO (Kozlovets), KV (Kulina-voda), MA (Malomir), SR (Srem), TR (Tsarevets). The clusters are represented graphically in the following picture:



There are three distinctive clusters here: [Hadjidimitrovo, Butovo, Bylgarsko-Slivovo, Kozlovets, Tsarevets], [Babek, Srem, Malomir] and [Kulina-voda, Alfatar]. It is obvious that this clustering is closer to the initial one. It kept the first cluster as it is and almost preserved the other two clusters. Hence, this way of generating texts gives results closer to the results based on Levenshtein distance ([5]) used in the experiments by Osenova, Heeringa and Nerbonne. One possible explanation for the better results might be that this method treats all phenomena equally with respect to the frequency in large corpora. Thus, the segments that have zero Levenshtein distance are the same in two dialect representations and they are a basis for better compression when the dialects are closer to each other.

5. Conclusions and future work. Our experiments proved that compression methods are feasible with generated data sets. Different ways of generation of such texts give us different measurements of the distance of dialects. Thus there is room for experiments, such as: which way of generation of text is more reliable with respect to linguistic intuition about the dialect distance measurement.

Another future task is to compare the compression method to other methods for dialect distance measurements and language contact measurement and also to methods for document similarity measurement. Our expectations are that word-based methods for dialect distance measurement will be close to the permutation-based method, because the frequency is not taken into account.

The general methods for document similarity such as cosine between the vector of words in the documents, used in Information Retrieval, or Latent Semantic Indexing will not work in our case because they compare word forms, which in the case of dialectological data (and between languages) will have different representations. Thus, methods that do not take into account the intraword structure are not good for our purposes.

As very important direction of future work we have recognized the problem of characterizing language contacts. In the following section we present some initial ideas of how our methods could be applied to this problem.

5.1. Application to language contacts. There are several ways in which language contact measurement can be defined. On the phonological level we would like to consider the closeness of the phonetic systems of the corresponding languages[5]. Here we can proceed on the basis of dialects in each language and check whether dialects of both languages which are spoken in geographically close sites demonstrate greater similarity in phonetic systems that dialects that are spoken in distant sites. Another way is just to check the similarity between standard languages.

In both cases the language data have to explicate the most prominent characteristics of the phonological systems of the two languages (or dialects). Because the measurement of similarity between languages from different language families is a harder and thus more interesting problem we will concentrate on it in future. In general, we cannot expect that similar phonetic features are explicated in the same lexemes (in the sense of translations between the lexemes in both languages), we need to segment the word forms in smaller segments.

Based on the above considerations we plan to do the following experiments:

- **Standard Languages Similarity Measurement.** Because we measure the similarity of the phonetic system of the two languages we can start with two corpora comparable in size. First step of preparing the data is to encode the word forms in phonetic alphabet. We cannot expect that we will be able to encode the interword forms dependencies, thus, we will concentrate on the intraword phonetic features. Then we can apply the method over the phonetically encoded corpora.

  As second option is to extract a list of word forms from both corpora and to apply the permutation method to each of the lists and compare the resulting data.

[5] An attempt in this direction was already made in [7]

- **Dialect to Standard Language Similarity Measurement.** In this experiments we will measure the distance between a Bulgarian dialect and the standard language of some of the neighboring countries. In this case we will use generation of a data set for Bulgarian dialects as it was described above. For the standard language we can use some of the methods described in the above point.

- **Dialect to Dialect Similarity Measurement.** In this case we can generate data sets as in the cases of the current experiments. The main difference will be in the case of corpus-based generation when we will use different corpora for the different languages. In this case we assume that the initial word forms are phonetically represented. This experiment will depend on the available data for the dialects in the neighboring countries.

6. Acknowledgments. The work presented in the paper is supported by a grant from the Volkswagen Foundation awarded jointly to the University of Groningen, the University of Tübingen, the University of Sofia, and the Institute for Parallel Processing, Bulgarian Academy of Sciences.

REFERENCES

[1] Cilibrasi R., P. Vitanyi. Clustering by Compression. *IEEE Trans. Information Theory* **51**, No. 4 (2005), 1523–1545.

[2] Grunwald P., P. Vitanyi. Kolmogorov complexity and information theory. With an interpretation in terms of questions and answers. *Journal of Logic, Language, and Information* **12**, No. 4 (2003), 497–529.

[3] Grunwald P., P. Vitanyi. Shannon Information and Kolmogorov complexity. *IEEE Trans. Information Theory* (submitted).

[4] Heeringa W. Measuring Dialect Pronunciation Differences using Levenshtein Distance. PhD thesis. University of Groningen. Groningen, The Netherlands, 2004.
`http://www.let.rug.nl/~heeringa/dialectology/thesis/`

[5] Levenshtein V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR* **163**, No. 4 (1965), 845–848 (in Russian).

[6] Osenova P., W. Heeringa, J. Nerbonne. A Quantitative Analysis of Bulgarian Dialect Pronunciation. In: Zeitschrift für Slavische Philologie. Tübingen, Germany, in press.

[7] Osenova P., W. Heeringa, J. Nerbonne. Using Pronunciation Difference to Measure Language Contact Effects. In: Proceedings of the First Conference on Language Contact in Times of Globalization (LCTG). University of Groningen. Groningen, The Netherlands, in press.

[8] Popov D., K. Simov, S. Vidinska. A Dictionary of Writing, Pronunciation and Punctuation of the Bulgarian Language. Atlantis LK, Sofia, Bulgaria, 1998 (in Bulgarian)

[9] Simov K., P. Osenova, S. Kolkovska, E. Balabanova, D. Doikoff. A Language Resources Infrastructure for Bulgarian. In: Proceedings of LREC, 2004, Lisbon, Portugal, 1685–1688.

[10] Stoykov S. Atlas of Bulgarian Dialects: Northeastern Bulgaria. Publishing House of the Bulgarian Academy of Sciences, 1966, volume II, Sofia, Bulgaria (in Bulgarian).

[11] Stoykov S., S. Bernshteyn. Atlas of Bulgarian Dialects: Southeastern Bulgaria. Publishing House of the Bulgarian Academy of Sciences, 1964, volume I, Sofia, Bulgaria (in Bulgarian).

[12] Stoykov S., K. Mirchev, I. Kochev, M. Mladenov. Atlas of Bulgarian Dialects: Southwestern Bulgaria. Publishing House of the Bulgarian Academy of Sciences, 1975, volume III, Sofia, Bulgaria (in Bulgarian).

[13] Stoykov S., I. Kochev, M. Mladenov. Atlas of Bulgarian Dialects: Northwestern Bulgaria. Publishing House of the Bulgarian Academy of Sciences, 1981, volume IV, Sofia, Bulgaria (in Bulgarian).

Kiril Simov, Petya Osenova
Linguistic Modelling Laboratory
Institute for Parallel Processing
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 25A
1113 Sofia, Bulgaria
e-mail: kivs@bultreebank.org
petya@bultreebank.org

Received November 7, 2006
Final Accepted February 23, 2007