Detecting Contact Effects in Pronunciation

Wilbert Heeringa

Variationist Linguistics, Meertens Institute, Amsterdam

wilbert.heeringa@meertens.knaw.nl

John Nerbonne

Humanities Computing, University of Groningen

j.nerbonne@rug.nl

Petya Osenova

Linguistic Modelling Department, IPOI, Bulgarian Academy of Sciences, Sofia

petya@bultreebank.org

J.Nerbonne

Information Science

Rijksuniversiteit Groningen

P.O. Box 716

NL 9700 AS Groningen

The Netherlands

Wilbert Heeringa

Variationist Linguistics

Meertens Institute

Postbus 94264

1090 GG Amsterdam

The Netherlands

Petya Osenova

Linguistic Modelling Department

Institute for Parallel Processing of Information

Bulgarian Academy of Sciences

Acad. G.Bonchev St. 25A

1113 Sofia, Bulgaria

Abstract

We investigate language contact effects between Bulgarian dialects on the one hand, and the languages of the countries bordering Bulgaria on the other. The Bulgarian data comes from Stojkov's Bulgarian Dialect Atlases. We investigate three techniques to detect contact effects in pronunciation, the phone frequency method and the feature frequency method, both of which are insensitive to the order of phonological segments within words, and also Levenshtein distance, a word-based method which is order-sensitive. We also examine pronunciation effects under the hypothesis that pronunciation influences should be strongest as one approaches the border of a country which speaks the putatively influential language. The study aims to contribute to the development of more exact tools for studying language contact.

Detecting Contact Effects in Pronunciation

Wilbert Heeringa, John Nerbonne, Petya Osenova

Meertens Institute, Amsterdam / University of Groningen / IPOI, Sofia

1. Introduction[1]

Although computational techniques have recently enabled large scale investigations of language varieties (Nerbonne & Kretzschmar 2006, and references there), little computational attention has to-date been paid to techniques for assaying language contact effects. Heeringa, Nerbonne, Niebaum, Nieuweboer & Kleiweg (2000) studied Dutch-German contact in and around the German county Bentheim. They found that dialects at the Dutch side of the border have become more Dutch while the German dialects have become more German. Measurements were made with the use of Levenshtein distance, which measures pronunciation differences between pairs of words, preferably pairs of cognates.

[1] The authors would like to thank Kiril Simov for help in the digitizing the data; Luchia Antonova for comments on the IPA conversion, for the selection of the Bulgarian sites and for general recommendations on Bulgarian dialectology; Christine Siedle for her help with geographical coordinates and the maps; and Peter Kleiweg for software and his quick reactions on software questions. We also owe thanks to the audience at *Language Contact in Times of Globalization* for very useful discussion and suggestions on a preliminary version of this paper. We especially thank Peter Houtzagers and Muriel Norde for their valuable remarks. This work is funded by NWO, Project Number 048.021.2003.009, P.I. J.Nerbonne, Groningen, and also a grant from the Volkswagenstiftung "Measuring Linguistic Unity and Diversity in Europe", P.I. E. Hinrichs, Tübingen.

For centuries Bulgaria has been in intensive contact with its neighboring countries. This contact includes relations not only in the areas of politics and economics, but also among languages. In this paper we compare dialects throughout Bulgaria to the five standard languages on its borders, viz., Macedonian, Serbian, Romanian, Greek and Turkish. We use a design intended to capture areal effects in language contact (Kurath 1972). We hypothesize that the areal spread of linguistic features should result in gradients of increasing similarity between the various dialects and each of the putative sources of contact effects. For example, in the case of Romanian, this predicts that varieties closest to the Romanian border will be most similar to Romanian, and those furthest away most dissimilar. The thesis that pronunciation should be subject to mixing effects stands in contrast to the general position of Balkanologists, who regard pronunciation as little affected by widespread contact (Birnbaum 1965). Our study uses a simple model of geography (effectively, just linear distance) and studies whether phonological similarity is related to it.

Naturally we need to operationalize the notion 'phonologically similar' in order to do this. We cannot rely exclusively on the human observations to adjudge phonological similarity since we need a method that can be applied to large amounts of material automatically, i.e. a computational technique. The Bentheim study used Levenshtein distance, a technique which aligns corresponding segments of the words to be compared, and sums the differences between the segments. But Levenshtein distance is sensitive to the order of segments in words, and insensitive to differences in segments that do not correspond. If we consider the example of the spread of uvular /r/ in the languages of Europe (Chambers & Trudgill 1998 [1980], § 11.4), it is clear that we notice changes even when they do not involve corresponding words. The uvular /r/ is present e.g. in the German word [ʁaux, ʁaux] 'smoke', even though it is completely absent in the nearest French equivalent *fume* /fym/ (and even though French is the source of the uvular /r/,

as scholars agree). We are therefore cautious about applying Levenshtein distance to materials from very different languages.

We therefore also consider two other corpus-based techniques where the difference between two dialects is equal to the sum of phones or, alternatively, features frequency differences of the respective corpora. The phone frequency method (PFM) was introduced by Hoppenbrouwers & Hoppenbrouwers (2001) and the feature frequency method (FFM) was firstly introduced by Hoppenbrouwers & Hoppenbrouwers (1988), but described in its most mature form in Hoppenbrouwers & Hoppenbrouwers (2001). In a nutshell, PFM compares two languages or language varieties by counting how many tokens there are of each phoneme in comparable corpora. FFM is a step more abstract, counting how many tokens there are of segments with specific values for given phonological features. Both of them seem poised to detect interlanguage effects that might escape Levenshtein distance.

The structure of the paper is as follows: the next section provides some background on Bulgarian dialectology. Section 3 focuses on the data source and the preparation of the data. In Section 4 the dialect distance metrics are explained and the procedure to measure the geographical course of the influence of surrounding languages to the Bulgarian dialect continuum. Section 5 discusses the results, and postulates that one enigmatic aspect of the present analysis has its roots in earlier patterns of settlement in Bulgaria. Section 6 sketches conclusions and prospects for further work along these lines.

2. Background Bulgarian Dialectology

Since we shall test a hypothesis about language contact by examining whether Bulgarian dialects become more and more similar to contact languages as one approaches the borders, we review the basic facts of Bulgarian dialectology here, focusing on pronunciation. It will be important later to

conclude that the measurements we are making do not contradict what is known about Bulgarian dialects. Our presentation of this background follows Stojkov (2002).

There is a major east/west division following the pronunciation of the old Bulgarian vowel 'yat' (in Bulgarian: 'ят'). In western Bulgarian dialects 'yat' has only the reflection /e/, e.g. *bel* 'white' - *beli* 'white-pl', while 'yat' in eastern dialects shows both reflections, /e/ and /ja/, e.g. *bʲal* 'white'- *beli* 'white-pl'. This single characteristic does not by itself distinguish the dialects consistently, but it remains quite important.

The various historical developments of the old Bulgarian 'big nosovka' (in Bulgarian: '*голяма носовка'*), a nasal vowel, divide Bulgarian dialects into five groups: ə-dialects (Northeastern and Northwestern Bulgaria and the eastern part of Southeastern Bulgaria); a-dialects (Western Bulgaria and the eastern dialect of Pirdop); ɒ-dialects (the Rodopi mountain); æ-dialects (the Teteven region and two villages in Eastern Bulgaria, Kazichino and Golitsa); and u-dialects (Western Bulgarian areas near the Bulgarian-Serbian border). This classification is admirably simple but also encounters numerous exceptions.

Morphological and lexical research shows Bulgaria to be divided into a central part (Northeastern and Central Bulgaria) and a peripheral part (Northwestern, Southwestern and Southeastern Bulgaria; Stojkov, 2002, p. 93).

Because of the instability and conflicting nature of various linguistic criteria Stojkov (2002) suggests a classification of Bulgarian dialects which respects geographical continuity, as well. In his standard work he distinguishes six, rather than five areas, concluding first that Bulgarian dialects are not separated categorically, but rather form a continuum. Second, there is a central (typical) area as well as peripheral (transition) areas among Bulgarian dialects. Third, Stojkov agrees with traditional scholarship that the most striking distinction of Bulgarian dialects is between East and West along the 'yat' border. In Figure 1 the six most significant geographical groups of Bulgarian dialects are shown as presented in Stojkov (2002, p. 416).

Figure 1 comes here

The vertical lines represent Moesian dialects; the horizontal lines represent Balkan dialects, the broken slanting lines - Southwestern dialects, the crosses - Northwestern dialects. The thick broken line represents the 'yat' borderline that divides the dialects into two major groups: Western and Eastern. The nearly horizontal slanting lines on the left side show transitional zones, and the steeply slanting lines at the bottom of the map represent the Rupskian (Rodopian) dialects.

For the purposes of this paper it is important to note that there are dialect divisions which indeed correspond to the various "peripheral areas". Naturally, this does not mean that these areas are therefore more similar to the languages spoken on the other side of the border, this is something we shall test. As we shall show below, the areas of similarity are in any case more diffuse than the division into areas suggests (see Section 4, Figures 3, 4 and 5).

3. The Data

We consider in turn the sources of our data and the selection we made, its preparation, and its conversion to digital form.

3.1. Sources

The data was digitized from the four volumes of Bulgarian dialect atlases which cover the entire country. These volumes are described in Stojkov (2002) and also Osenova, Heeringa and Nerbonne (2007), and we shall repeat only the most important information for our purposes here.

The atlases were compiled over a period of thirty years by various fieldworkers, who transcribed consistently into a broad phonetic transcription. Fieldworkers did not rely on single informants, but instead used several, and attempted to elicit material indirectly in extensive interviews, rather than via direct questions.

We extracted words from these atlases which we then compared in pronunciation. Our method (described below) relies on transcriptions of entire words, which we took from the atlases as best we could. Where we needed (infrequently) to extrapolate, we always did this conservatively, i.e. using no additional phonetic detail.

The sites sampled in the atlases were all exclusively ethnic Bulgarian populations regardless of geography. We speculate that the atlas designers chose only such sites because they were interested in the historical roots of Bulgarian. Whatever the reason, the selection is clearly suboptimal for the purpose of gauging contact effects; indeed, it seems better designed to hide contact effects rather than document them. However, instead of giving up in recognition of this problem, we choose to forge ahead, reasoning that long-standing effects of the sort we are interested in should not occur only in ethnically heterogeneous settlements. Further, we suspect the effects of restricting attention to ethnically homogeneous towns and villages should not confound the study, since it affects all areas in roughly the same way. But it remains the case that the sites sampled in the atlas certainly under-represent the degree of contact influence in the country.

3.2. Sites

In Stojkov's Bulgarian Dialect Atlases data from 1682 sites is available. We use a subset of 488 sites which were selected with respect to two main criteria: maximally complete coverage of the area covered by the atlas, and a representative number of varieties and sub-varieties. We would have preferred using sites selected randomly from a regular grid throughout Bulgaria, but there

were no collection sites in large stretches of the country, which explains the patchy impression of the map. The distribution of the 488 sites is shown in Figure 2.

When studying the influence of a particular language on the Bulgarian dialects, and especially the course of the influence in the Bulgarian dialect landscape, we need to measure the shortest geographic distances to the border of the country in which that language is spoken. We measured these distances manually using a paper map and a ruler. Because this turned out to be time-consuming, we restricted the analysis to a subset of 50 representative varieties (from the original 488), which were scattered as regularly as possible. The 50 sites are represented by circles in Figure 2. For clarity, we should note that in this paper we use the 488 sites for calculating and visualizing dialect distances compared to the standard languages (see Sections 5.1 and 5.2 and Figures 3, 4 and 5), and we use the selection of 50 sites for the regression analysis (see Section 5.3) aimed at detecting contact effects.

Figure 2 comes here

3.3. Words and Conversion

We digitized a set of 54 words, which turned out not to be instantiated at every site, but which includes a subset of 36 words that were instantiated in all the atlas volumes. This differentiation of two sets arose because, as noted above, the lexical material differs across the four atlases.

The digitization step involved transliterating from a Bulgarian system of phonetic transcription into IPA, which was processed in its computerized form, X-SAMPA. We include two tables in an appendix to show how we interpreted the Bulgarian phonetic transcription system in terms of equivalents in the *International Phonetic Alphabet* (IPA 2003).

Table 1 in the Appendix provides the list of 36 words that were common to all of the 488 sites selected from the atlases. The phonetic transcriptions of the standard Bulgarian, Macedonian, Serbian, Romanian, Greek and Turkish pronunciation are given. The transcriptions are the same as used for the experiments in this paper. Osenova, Heeringa and Nerbonne (2007) discuss the word sample and its properties in more detail.

The words in Table 1 represent many of the most important phonetic features of Bulgarian varieties. They reflect the following phenomena:

1. the reflections of 'yat' in different phonetic contexts (stressed and unstressed, word-finally, after fricatives, etc.): [bʲala, 'bɛli, 'greʃka, mlɛ'kar, 'vətrɛ, vɛn'ʧilo]
2. the reflections of the etymological 'ja': ['jazdi] or ['jɛzdi], [po'lʲana] or [po'lʲɛna], [gu'lʲaj] or [gu'lɛj], [dɛn] or [dɛnʲ]
3. nonpalatal-semipalatal-palatal distinction word-finally: [sol] or [solʲ] or [sol'], [pət] or [pətʲ] or [pət'], [kon] or [konʲ] or [kon'], [dɛn] or [dɛnʲ] or [dɛn'][2]
4. the realizations of 'schwa' under stress: ['bəʧva] or ['boʧva] or ['baʧva] etc. The same for other words: ['zəlva, sən, 'tənko, oti'ʃəl, do'ʃəl]
5. the realizations of the nasal vowel: [zəb] or [zob] or [zab] etc. Similarly for other words: ['kəʃta, 'səbota]

[2] The IPA system (revised to 2005) provides a diacritic for palatalized segments, but does not distinguish between semipalatalized and palatized segments. In the Bulgarian atlas, however, this distinction is made. Here we add a superscript j to semipalatalized segments (e.g. [tʲ]) and a ' to palatalized segments (e.g. [t']). When processing the data, we do not yet process the semipalatalized diacritic, but ignore it.

6. the metatheses ‘əl-lə’ and ‘ər-rə’: [ʒəlt] or [ʒlət], [gərb] or [grəb]

7. the realizations of various vowels in different contexts: [‘ovʧɛ] or [‘ovʧo], [klʲuʧ] or [kliʧ]

8. the reduction of the open vowels in unstressed position: [mlɛ’kar] or [mli’kar], [vɛn’ʧilo] or [vin’ʧilo], etc.

3.4. Contact Language Material

To compare the pronunciations in the contact languages, we used the most frequent lexicalization of the concepts used in the word list above. We sought these for each of the four contact languages examined: Macedonian, Serbian, Romanian, Greek and Turkish. Macedonian and Serbian are closely related South Slavic languages, while Romanian belongs to the Romance language family, Greek to Greek and Turkish to Turkic. This appears to be unfortunate from the point of view of language contact studies, as it will be impossible to separate genealogical influence (stemming from the common historical source of the Slavic languages) from contact influence in the case of Macedonian and Serbian, but it is quite fortunate in that we can use the well-known proximity of Bulgarian to Macedonian and Serbian as a test of how well the different candidate techniques are working.

The set of 36 words which we used for the comparison comprises almost exclusively words of Slavic origin. Only two loanwords are present: 'pocket' and 'pot', both from Turkish. The nearest equivalents in Macedonian and Serbian were obtained from Bulgarian experts on these languages on the basis of the Bulgarian words. The nearest equivalents in Romanian, Greek and Turkish were obtained by asking native speakers of these languages for the nearest equivalent, using English translations as a basis for comparison.

It was naturally difficult at times to settle on a single closest word for a given concept. For example, Turkish has two words for the concept 'mistake'; Romanian two words for 'feast';

and Serbian two words for 'cup, glass' (as does English). In all these cases, both words were used and the differences averaged. In cases of morphosyntactic asymmetry, in which single lexical items in Bulgarian were closest to multiword lexical items (e.g. 'ride' in Turkish) we encode the sequence of words and used that as a basis of comparison.

4. Methods

4.1. Measuring Linguistic Distances

4.1.1. Phone frequency method

Hoppenbrouwers and Hoppenbrouwers (2001, p. 1) describe an experiment in which languages were compared on the basis of phonetic texts from *The Principles of the International Phonetic Association* (IPA 1949). In this IPA pamphlet the fable 'The North Wind and the Sun' is produced in 51 languages and rendered in phonetic transcription. For each text the frequencies of phones are determined. Since not all samples have the same size, relative frequencies are used. The distance between two languages is equal to the sum of the absolute values of the differences between the corresponding (relative) phone frequencies.

Even though the difference between palatalized and non-palatalized consonants is represented only through the use of the diacritics (see examples, Appendix Table 1), palatal and nonpalatal consonants were regarded as distinct when we counted the phones. Applying this method to our material (488 dialects, standard Bulgarian, Macedonian, Serbian, Romanian and Turkish) 69 different phones were found.

4.1.2. Feature frequency method

The phone frequency method introduced above does not take into account that for example the [i] and [I] are more similar to each other than the [i] and [ɑ]. Therefore Hoppenbrouwers and Hoppenbrouwers (1988) developed the feature frequency method. Using this method, each phone is described by a range of binary features. For example the feature *round* is set to 1 when a vowel is rounded (e.g. the [y]) and set to 0 if a vowel is not rounded (e.g. [i]). The feature *voiced* is set to 1 when a consonant is voiced (e.g. [v]) and set to 0 if the consonant is not voiced (e.g. [f]). If we have a corpus of 36 phonetic transcriptions per variety, for each feature we count the number of segments for which that feature is marked positively. We count the number of rounded sounds, the number of voiced sounds, etc. The frequencies are divided by the total number of phones in the corpus to obtain relative frequencies. We calculate the distance between two languages as the sum of the differences between the corresponding feature frequencies.[3]

When defining phonetic segments in terms of features, one has to choose the right features. Hoppenbrouwers and Hoppenbrouwers (2001) used a modified version of *The Sound Pattern of English* (SPE) (Chomsky & Halle 1968). We used the system of Almeida & Braun (1986), since this system is directly based on the well-known IPA system. When using this system, we separate vowels and consonants. It means that vowel feature counts are divided by the number of vowels in the corpus, and consonant feature counts are divided by the number of consonants in the corpus. The vowel features are listed in Table 2 and the consonant features are listed in Table 3.

Table 2 comes here

Table 3 comes here

[3] Hoppenbrouwers and Hoppenbrouwers (1988) give several alternatives for calculating the difference of two feature frequency histograms. It is beyond the scope of this article to discuss them fully.

We converted the IPA-inspired Almeida-Braun system to a binary system (like SPE). The binary system chosen is designed to avoid the obscuring effects of multivalued systems in which contrasting differences may be neutralized. We illustrate the danger with a small example. Assume that one variety has one front vowel and one back vowel. The mean value will be equal to (1+3)/2 = 2. Another variety with two central vowels would have a value of (2+2)/2 = 2. In this way it looks if the two varieties do not differ with respect to the feature *advancement*. This problem is solved by converting the multivalued feature into a vector of binary features if we use a somewhat verbose format. In general a feature with *n* values is always converted to a vector of *n-1* binary values. We illustrate this with the feature *advancement* which will be represented by three binary features:

	advance 1	advance 2	advance 3
front	1	0	0
central	1	1	0
back	1	1	1

We also need to pay special attention to affricates. When we find for example a [ts], we use the average values of the binary feature representations of the [t] and the [s].

As mentioned above, many consonants have palatalized counterparts. We represented e.g. [k^j] by averaging the place of articulation of the [k] with palatal. Averaging is again done on the basis of the binary representations.

4.1.3. Levenshtein distance

Using the Levenshtein distance, two varieties are compared by comparing the pronunciation of words in the first variety with the pronunciation of the same words in the second. We determine how one pronunciation might be transformed into the other by inserting, deleting or substituting sounds. Costs are assigned to these three operations. In the simplest form of the algorithm, all operations have the same cost, e.g., 1. We illustrate this with an example of two varieties of a word pronunciation in northwestern dialects.

Changing one pronunciation into the other can be done as follows (ignoring suprasegmentals and diacritics):

tsrɛʃna	substitute ts by ʧ	1
ʧrɛʃna	insert ɛ	1
ʧɛrɛʃna	delete n	1
ʧɛrɛʃa		
		3

In fact many sequence operations map [tsrɛʃna] to [ʧɛrɛʃa]. The power of the Levenshtein algorithm is that it always finds the cost of the cheapest mapping. Levenshtein distance is then the distance assigned by the Levenshtein algorithm, the cost of the least expensive means of mapping one string to another.

To deal with syllabicity, the Levenshtein algorithm is adapted so that only vowels may match with vowels, and consonants with consonants, with several special exceptions: [j] and [w] may match with vowels, [i] and [u] with consonants, and central vowels (in our research only the schwa) with sonorants. So the [i], [u], [j] and [w] align with anything, but otherwise vowels align

with vowels and consonants with consonants. In this way unlikely matches (e.g., a [p] with an [a]) are prevented. In our example we thus have the following alignment:

ts	0	r	ɛ	ʃ	n	a
ʧ	ɛ	r	ɛ	ʃ	0	a
1	1				1	

In earlier work we divided the sum of the operation costs by the length of the alignment. This normalizes scores so that longer words do not count more heavily than shorter ones, reflecting the status of words as linguistic units. However, Heeringa, Kleiweg, Gooskens & Nerbonne (2006) showed that results based on raw Levenshtein distances approximate dialect differences as perceived by the dialect speakers better than results based on normalized Levenshtein distances. Therefore we do not normalize the Levenshtein distances in this paper.

Here we use Levenshtein as demonstrated in the examples above, i.e. with binary operation costs. One might expect the use of gradual costs to be more obvious, but in a validation study Heeringa (2004) showed that, generally speaking, the use of binary costs outperforms the use of gradual costs.

Again we need to pay some special attention to affricates and palatalized consonants. Affricates as processed as sequences of two consonants. For example the [ʦ] is processed as a [t] followed by an [s]. Following our procedure for the phone frequency method, we considered a palatal sound and its non-palatal counterpart as fully different. For example the [k] and the [k^j] are considered as different as the [k] and [v].

The distance between two varieties is calculated as the average of the 36 Levenshtein distances which correspond with 36 word pairs.

4.2. Design

Trubetzkoy (1930) suggested that superficial similarity in pronunciation–in the absence of regular sound correspondence–should constitute evidence of a *Sprachbund*, using Bulgarian's relation to the other Balkan languages as an example.[4]

If we add to this the conjecture that such groups originate in language contact, and that such contact is most intense near borders, then we should expect to see that pronunciation similarity is most extreme near borders, a hypothesis which we can test readily using a regression analysis, once we have settled on a suitable measure of pronunciation similarity.

We therefore measure, for each of the varieties in the subset of 50 sites, taken from the sample of 488 sites (see Section 3.2) the distance to the nearest border for each of the four contact

---

[4] From Trubetzkoy's (1930) brief note:

> Gruppen, bestehend aus Sprachen, die [...] manchmal auch äussere Ähnlichkeit im Bestande der Lautsysteme, - dabei aber keine systematischen Lautentsprechungen, keine Übereinstimmungen in der lautlichen Gestalt der morphologischen Elemente und keine gemeinsamen Elementarwörter besitzen, - solche Sprachgruppen nennen wir Sprachbünde. So gehört z.B. das Bulgarische einerseits zur slawischen Sprachfamilie [...] andererseits zum balkanischen Sprachbund [...].

Translated in English:

> Groups consisting of languages which […] often have external similarities in the inventories of sound systems, but no systematic sound correspondences, no similarities in the sound shape of morphological elements and no shared elementary words, such linguistic groups we call Sprachbünde. For example Bulgarian belongs to the Slavic language family on the one hand […] and to the Balkan Sprachbund on the other hand […].

languages. We hypothesize that the distance to the border will correlate positively with the pronunciation distance as measured by PFM, the FFM and Levenshtein distance.

5. Results

We first examine the overall measurements in order to determine which of the measurement techniques appears to be successful in detecting linguistic affinity. We then turn to the correlation with geography.

5.1. Distances to standard languages

We examine the overall measurements in two respects to see whether they were sensitive to the sort of linguistic similarity we wish to detect. First, we examined what Nerbonne & Kleiweg (2007) call *local incoherence* to see how well the measurement was detecting a signal of geographic coherence. Levenshtein distance was far and away the best technique in this respect. We applied Levenshtein distance irrespective of whether words of a comparison pair are cognates or not.

Second we checked whether the consensus view, i.e. that Macedonian is most similar to Bulgarian, followed closely by Serbian, is in fact reflected by all of the measurement techniques. In order not to be confused by the similarity of some varieties, even in the face of substantial overall differences, we examine not only the average degree of similarity, but also the degree of similarity of the most similar varieties (the first quartile of measurements).

The table shows the mean and the standard deviation of the distances to each standard language (in the two columns on the right), while the "first quartile" columns show mean and standard deviations for the closest quarter of the dialects (per language). The Levenshtein distances are averaged over the number of dialects, and over the number of words per dialect.

Table 4 comes here

To do this, we calculated the average linguistic distance between each of the reference languages and all of the 488 Bulgarian dialects for which we had data. The same descriptive statistics were calculated while restricting attention to the top 25% of most similar varieties. The results in Table 4 show the mean and the standard deviation for each standard language using the Levenshtein distance, which have been averaged over the number of dialects, and over the number of words per dialect. Levenshtein distances conform to the expectation that Macedonian is closest, followed by Serbian. Both have relatively small standard deviations. Romanian is more distant, followed by Turkish and Greek, and the distances to these standard languages have relatively high standard deviations.

It turns out that the order-insensitive methods, PFM and FFM, are only marginally less successful in detecting linguistic similarity. However, when attention is restricted to the most similar quartile, PFM and FFM agree with Levenshtein in showing that Macedonian is closest, followed by Serbian. FFM differed from the other two when the entire set of Bulgarian varieties was examined, where it led to results in which Serbian was most similar. As we noted above, the analysis of the most similar varieties is probably the better pole of comparison when examining these results.

So it turns out that PFM and FFM, which we suspected would be more suitable for the comparison of (strongly) unrelated varieties, are not clearly better. On the other hand, we do not conclude that they are clearly worse either, only marginally so. In particular, all three methods result in analyses of the first quartile of data in which the consensus view of experts is respected.

5.2. Geographic gradient of contact

We turn then to our second topic, the degree to which we can detect a gradient of similarity approaching the borders of other languages areas. We shall continue to examine alternative measurement techniques since we do not regard any as clearly superior, even if Levenshtein distance seems (marginally) preferable to the alternatives.

Figure 3 displays Levenshtein distances of 488 Bulgarian varieties compared to Macedonian and Serbian, in Figure 4 the same varieties are compared to Romanian and Greek, and in Figure 5 they are compared to Turkish. In Figure 2 the varieties are represented by dots, which represent locations. In Figures 3, 4 and 5 not only the dots are colored, but the areas surrounding the dots as well, in order to get clearer pictures. In general (nearly) the same dialect is spoken in the direct neighborhood of a location, although there may be exceptions, especially as regards (large) cities.

The Macedonian and Serbian map clearly show a gradient of similarity toward the border. But we note again here that the gradient of similarity may not indicate language contact effects at all, but rather the pronunciational residue of a continuum in the South Slavic languages. The Romanian map, the Greek map and the Turkish map do not suggest a strong gradient of similarity toward the relevant borders, but we shall examine the gradient numerically, as well. It is striking that the Greek map shows a gradient of similarity toward the Macedonian border. Bulgarian dialects which are relatively close to Greek, are close to Macedonian as well.

Figure 3 comes here

Figure 4 comes here

Figure 5 comes here

5.3. Correlation between geographic and linguistic distances

For a subset of 50 Bulgarian dialects we measured the geographic distances to the (closest) borders of Macedonia, Serbia, Romania, Greece and Turkey. We calculated the correlations between these geographic distances and the linguistic distances to the corresponding standard languages of these countries. The results are given in Table 5. Linguistic distances were calculated using the PFM, the FFM and Levenshtein distance.

Table 5 comes here

We first note that the results agree to some extent. All of the techniques detect clines of increasing similarity approaching the borders of Macedonia, Serbia and Romania, and none of them see any such (positive) gradient when approaching the Greek or Turkish border.

In fact, PFM and Levenshtein actually detect significant *negative* correlations between linguistic and geographic distances involving Greek or Turkish on the one hand and the Bulgarian varieties on the other. FFM measures a nonsignificant correlation, but again a correlation in the direction opposite from the one predicted.

The relatively strong correlation with Romanian when using the PFM and the FFM is all the more remarkable given the large consensus among Balkanists that pronunciation plays a subordinate role in the *Sprachbund* (Birnbaum 1965). Romanian is, of course, a Romance language, and it is surprising to see that its phonological properties are increasingly shared as one proceeds toward its borders, given the usual tenet that Balkan language contact does not involve phonology, at least not primarily. Perhaps Asenova (1989), is correct in identifying the similar vowel systems of the Balkan languages as a unifying feature (but we note that Asenova does not regard Turkish as participating in the *Sprachbund*, an issue outside the scope of this paper and

one we have attempted to avoid taking a stand on). Investigating the linguistic basis of the Romanian gradient will have to await a next paper.

We return to the cases of Greek and Turkish. For Greek, both the PFM and the FFM measurements result in significant negative correlations. For Turkish the FFM measurements result in a significant negative correlation. The Bulgarian varieties we collected and analyzed become less similar to Greek/Turkish as one approaches the border. Counseled by caution, we emphasize that techniques we are applying are novel in this area so that we cannot rule out problems in the measurement techniques. But simple error is unlikely to result in statistical significance.

A more interesting conjecture for Turkish is that the explanation lies in the more complicated relation between Turkish contact and Bulgarian. After all, Bulgarian was a part of the Ottoman empire from 1393 on for nearly five centuries. Hence, the sites with substantial Turkish populations are not only located near the Turkish border, but practically all over the country. For example, there are compact Turkish populations in the Northeast (Shumen, Targovishte, Razgrad, Silistra), in the south central part of the country (Plovdiv), and in southern parts (Kardzali, Smolyan). We would be interested in following up this conjecture with a study involving such demographics (if the relevant quantitative information is available). Linguistically we found that palatalization of [b], [t], [d], [v], [n] and [r] is most frequently found in the eastern Bulgarian dialects. However, none of these palatal sounds occur in Turkish, which makes the geographically more distant western Bulgarian varieties linguistically closer to Turkish than the eastern ones.

In the southern part of Bulgaria a large Greek population lived, especially in the area which was known as Eastern Rumelia in the period 1878-1885 when it was an autonomous province in the Ottoman empire. This population was largely exchanged in the aftermath of the Balkan wars and the second world war. Today, several thousand Bulgarians of Greek descent still inhabit the region, especially the *Sarakatsani*, transhumant shepherds. Actually we may expect a

positive correlation from this, but probably the Bulgarians want to distinguish themselves from the Greek by contrasting their dialect pronunciation to the Greek pronunciation.

Although the results in Table 5 agree to some extent, the correlation measures do not agree with each other well, in particular the phone frequency and feature frequency methods as applied to Turkish and the feature frequency method and Levenshtein distance as applied to Serbian. This is important with respect to the methodological goal of developing techniques which detect contact effects. The failure of the techniques to agree indicates that they are not *all* functioning as wished.

6. Conclusions and prospects

Bulgaria and the Balkans are most famous linguistically for the extensive language contact which has developed there (Trubetzkoy 1930), and it is fascinating to apply quantitative techniques developed for dialectology in order to explore and analyze language contact.

In this paper we applied a measurement of pronunciation differences to a large database of Bulgarian.

We see the future work in several directions. First, we would like to examine different dialect data, and in particular data collected from sites that were not selected for being purely Bulgarian. Second, it would be important to identify the regular aspects of the distinctions at the base of the analysis here, i.e. the linguistic basis of the aggregate analysis, and, in fact, we have initiated that work in collaboration with a Ph.D. student. Third, it would be interesting to include lexical variation in a parallel analysis, and to examine the degree to which lexical differences correlate with differences in pronunciations. We hasten to add that a great deal more material would be needed in order to obtain reliable lexical measurements.

Appendix

Bulgarian cyrillic written form	**Bulgarian pronun-ciation**	**Macedonian pronun-ciation**	**Serbian pronun-ciation**	**Romanian pronun-ciation**	**Greek pronun-ciation**	**Turkish pronun-ciation**	**English translation**
бъчва	‘bətʃva	‘boʧva	‘batʃva	bu’toj	vi’tion	‘fɯʃɯ	barrel
зълва	‘zəlva	‘zolva	‘zaova	kum’natə	anða’ðɛlfi	‘□œɾymʤɛ	sister-in-law
дошъл	do’ʃɛl	do’ʃol	do’ʃao	vɛ’nit	‘irΘɛ	‘□ɛlʲmiʃ	has come-he
жълт	ʒəlt	ʒolt	ʒut	‘galbɛn	‘kitrinos	sa’ɾɯ	yellow
зъб	zəp	zap	zup	‘dintɛ	‘ðodi	diʃ	tooth
събота	‘səbota	sa’bota	‘subota	‘sɨləbtə	‘savaton	ʤumaɾtɛsi	Saturday
къща	‘kəʃta	‘kukʲa	‘kuʧʲa	‘kasə	‘spiti	ɛf	house
бяла	‘bʲala	‘bɛla	‘bɛla	‘albə	aspri’a	ak	white-fem
бели	‘beli	‘bɛli	‘bɛli	‘albɛ	as’pri	ak	white-pl
язди	‘jazdi	‘yazdi	‘jaʃɛ	kələrɛʃtɛ	piɟɛ’ni	a’tabinojoʃ	ride-3per
неделя	nɛ’dɛlʲa	‘nɛdɛla	nɛ’dɛlja	du’minikə	kiria’ki ɛvðo’maða	‘pazaɾ	Sunday
млекар	mlɛ’kar	‘mlɛkar	‘mlɛkar	ləp’tar	ɣalakto’polis	‘syʧy	milkman
грешка	‘grɛʃka	‘grɛʃka	‘grɛʃka	grɛ’ʃʲala	‘laΘos	‘hata ‘janɯʃlɯk	mistake
венчило	vɛn’ʧilo	vɛn’ʧilo	vɛn’ʧanjɛ	kunu’niɛ	nifi’kos	ni’kʲah	married life
ключ	klʲuʧ	kluʧ	kljuʧ	‘kɛjɛ	‘kliði	‘anahtaɾ	key
чаша	‘ʧaʃa	‘ʧaʃa	‘ʃolja ‘ʧaʃa	pa’xar	po’tiri	‘baɾdak	glass; cup
път	pət	pat	put	drum	‘ðromos	jol	road
жаби	‘ʒabi	‘ʒabi	‘ʒabi	‘broaʃtɛ	‘vatraçi	kuɾba’ɣałaɾ	frogs
нощви	‘noʃtvi	‘nokʲvi	‘natʃjvɛ	‘kuʃkə	zumo’tiri	ɛk’mɛktɛknɛsi	hutch
поляна	po’lʲana	‘poljana	‘proplanak	po’janə	ksɛf’to	‘ʧimɛn a’lan	glade
овче	‘oftʃɛ	‘oftʃo	‘oftʃji ‘oftʃɛtina	dɛ’oajɛ	‘provios	‘kojundan	sheep’s
тънко	‘tənko	‘tanko	‘tanko	sub’tsirɛ	lɛp’tos psi’los	‘inʤɛ	narrow-neut
гуляй	gu’lʲaj	‘pijanka	‘pijanka	pɛtrɛ’ʧɛrɛ kɛf	krɛ’pali ‘orɟia	ʃo’lɛn	feast
овчар	of’ʧar	‘ovʧar ‘ʧoban	‘ʧoban	ʧo’ban	vos’kos tso’panis	‘ʧoban	shepherd
кон	kon	konʲ	konʲ	kal	‘aloɣos ‘ipos	at	horse
сън	sən	son	sanʲ	vis	‘oniron ‘ipnos	ɾy’ja	dream

отишъл	oti'ʃɛl	‘otiʃol	oti'ʃao	plɛ'kat	pi'ɟɛ	‘□itmiʃ	has gone-he
вътре	‘vətrɛ	‘vnatrɛ	‘unutra	ɨnə'unutru	‘mɛsa	i'ʧardɛ	inside
тенджера	‘tɛnʤɛra	‘tɛnʤɛrɛ	‘lonats ‘ʃɛrpa	‘kratitsə	tɛnʤɛ'rɛs ‘tɛnʤɛrɛis	‘tɛndʒɛrɛ	pot
джоб	ʤop	ʤɛp	ʤɛp	buzu'nar	‘tsɛpi	ʤɛp	pocket
няма	‘nʲama	‘nɛma	‘nɛma	nu'ɛstə	ðɛni'parçi	jok	there is no
череша	ʧɛ'rɛʃa	‘ʧɛrɛʃna	‘ʧrɛʃnja	ʧi'rɛʃ ʧi'rɛaʃə	kɛra'sɛa kɛ'rasioni	‘kɪraz	cherry
гръб	grəp	grp	‘lɛʤʲa	‘spatɛ	‘plati ‘raxi	sɯrt	back
живя	ʒi'vʲa	ʒi'vɛ	‘ʒivio	trə'it	ɛsizɛ	‘jaʃadi	lived-he/she/it/you
сол	sol	sol	so	‘sarɛ	a'lati	tuz	salt
ден	dɛn	dɛn	dan	Zi	mɛra	□jun	day

Table 1: The thirty-six Bulgarian words which formed the base of the study in phonemic transcription. The transcriptions of the Macedonian, Serbian, Romanian, Greek and Turkish equivalents are given as well. All 488 sites used in this study included phonetic transcriptions of these thirty-six words.

References

Almeida, A. & Braun, A. 1986. "Richtig" und "Falsch" in phonetischer Transkription; Vorschläge zum Vergleich von Transkriptionen mit Beispielen aus Deutschen Dialekten'. *Zeitschrift für Dialektologie und Linguistik* LIII(2), 158-172.

Asenova, P. 1989. Балканско езикознание. *Balkan Linguistics* (in Bulgarian), Faber, Veliko Tarnovo.

Birnbaum, H. 1965. Balkanslavisch und Südslavisch. Zur Reichweite der Balkanismen im Südslavischen Sprachraum. *Zeitschrift für Balkanologie* 3, 12-63.

Chambers, J. K. & Trudgill, P. 1998, [[1]1980]. *Dialectology.* Cambridge: Cambridge University Press.

Chomsky, N. A. & Halle, M. 1968. *The sound pattern of English.* New York: Harper & Row.

Heeringa, W. 2004. Measuring dialect pronunciation differences using Levenshtein distance, Ph.D. thesis, University of Groningen, Groningen. URL: http://www.let.rug.nl/~heeringa/dialectology/thesis.

Heeringa, Wilbert, John Nerbonne, Hermann Niebaum, Rogier Nieuweboer & Peter Kleiweg. 2000. Dutch-German contact in and around Bentheim. In *Languages in Contact. Studies in Slavic and General Linguistics*, D. Gilbers, J. Nerbonne & J. Schaeken (eds). Vol. 28, 145-145, Amsterdam and Atlanta GA: Rodopi.

Heeringa, W., Kleiweg, P., Gooskens, Ch. & Nerbonne, J. 2006. Evaluation of string distance algorithms for dialectology. In *Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney, July, 2006,* J. Nerbonne & E. Hinrichs (eds.), 51-62.

Hoppenbrouwers, C. & Hoppenbrouwers, G. 1988. De featurefrequentiemethode en de classificatie van Nederlandse dialecten. *TABU: Bulletin voor taalwetenschap* 18(2), 51-92.

Hoppenbrouwers, C. & Hoppenbrouwers, G. 2001. *De indeling van de Nederlandse streektalen. Dialecten van 156 steden en dorpen geklasseerd volgens de FFM.* Assen: Koninklijke Van Gorcum B.V.

IPA. 1949. *The principles of the International Phonetic Association: being a description of the International Phonetic Alphabet and the manner of using it, illustrated by Texts in 51 Languages*. London: International Phonetic Association.

IPA. 2003. *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet.* Cambridge: Cambridge University Press.

Kurath, H. 1972. *Studies in areal lingusitics.* Bloomington/London: Indiana University Press.

Nerbonne, J. & Kleiweg, P. 2007. Toward a dialectological yardstick, *Quantitative Linguistics* 14(2), 148-167.

Nerbonne, J. & Kretzschmar, W. (eds.). 2006. *Progress in dialectometry*. Special issue of *Literary and Linguistic Computing*, 21(4).

Osenova, P., Heeringa, W. & Nerbonne, J. 2007. A quantitative analysis of Bulgarian dialect pronunciation. *Zeitschrift für Slavische Philologie*, accepted for publication.

Stojkov, S. 2002. Българска диалектология, Bulgarian Academy of Science, Sofia, fourth edition.

Trubetzkoy, N. S. 1930. Proposition 16. Über den Sprachbund. *Actes du premier congrès international de linguistes à la Haye du 10.-15.1928*, Vol. 1, 17-18. Leiden: A. W. Sijthoff's.

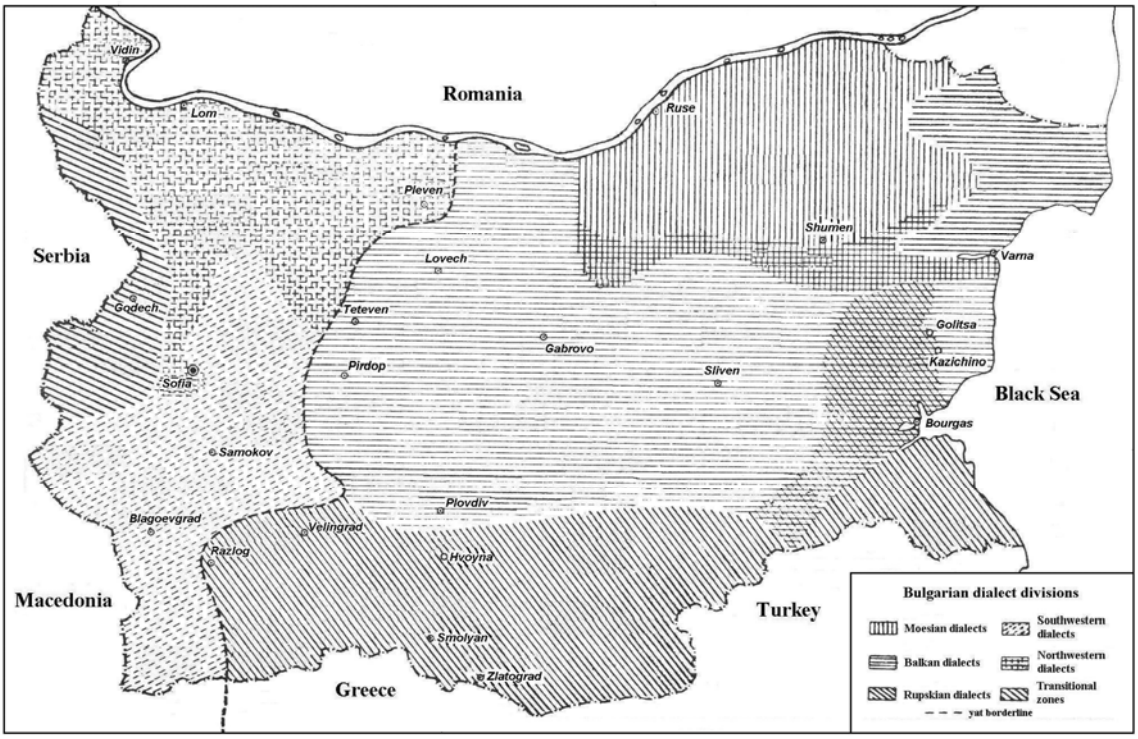

Figure 1: The map of Bulgarian dialect divisions as presented in Stojkov (2002, p. 416).

Figure 2: The distribution of the 488 Bulgarian sites selected. We use a subset of 50 dialects to study the relationship between geography and language contact with languages of bordering countries. These sites are represented by circles. The boundary lines indicate administrative divisions, not dialect areas.[5]

[5] Here the older administrative division is presented (valid up to 1997). We prefer this representation, because the areas are few, and thus easily detectable. The new division includes 28 regions. The interested reader is referred to: http://bg.wikipedia.org/wiki/Административно_деление_на_България .

Feature	Value	Meaning
advancement	1	front
	2	central
	3	Back
height	1	close
	2	near-close
	3	close-mid
	4	central
	5	open-mid
	6	near-open
	7	open
roundedness	0	no
	1	Yes

Table 2: The vowel features of Almeida & Braun and their possible values.

Feature	Value	Meaning
place	1	bilabial
	2	labiodental
	3	dental
	4	alveolar
	5	postalveolar
	6	retroflex
	7	palatal
	8	velar
	9	uvular
	10	pharyngeal
	11	Glottal
manner	1	plosive
	2	nasal
	3	Trill
	4	tap or flap
	5	fricative
	6	lateral fricative
	7	approximant
	8	lateral approximant
voice	0	no
	1	Yes

Table 3: The consonant features of Almeida & Braun represented in a binary system.

	first quartile		all distances	
	mean	sd	mean	sd
Macedonian	1.1	1.3	1.3	1.7
Serbian	2.6	7.0	2.5	6.1
Romanian	4.9	23.7	4.8	23.5
Greek	5.3	28.2	5.5	29.8
Turkish	5.4	29.3	5.4	29.6

Table 4: The Levenshtein distances between all of the 488 Bulgarian dialects and each of the putative sources of contact influence.

Figure 3: Average Levenshtein distances of 488 Bulgarian dialects compared to Macedonian (top) and Serbian (bottom). Dialects are represented by polygons. Lighter polygons represent closer dialects, and darker ones more distant dialects. Notice the clear gradient in similarity with respect to the western (Serbian) border.

Figure 4: Average Levenshtein distances of 488 Bulgarian dialects compared to Romanian (top) and Greek (bottom). Dialects are represented by polygons. Lighter polygons represent closer dialects, and darker ones more distant dialects. We see little visual reflection of the gradients hypothesized with respect to the Romanian and Greek borders.

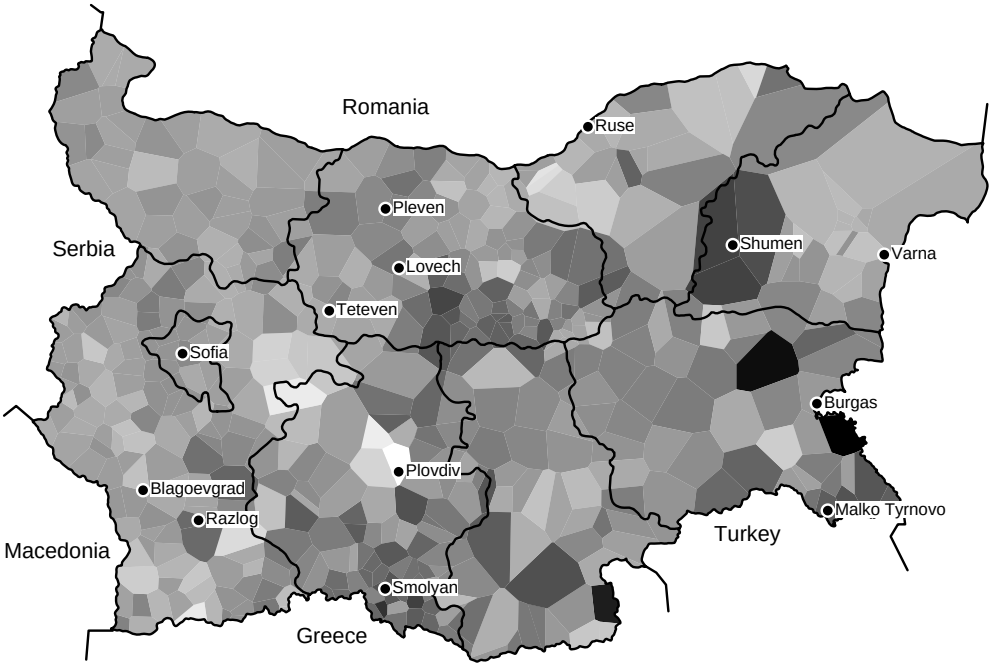

Figure 5: Average Levenshtein distances of 488 Bulgarian dialects compared to Turkish. Dialects are represented by polygons. Lighter polygons represent closer dialects, and darker ones more distant dialects. Again we see little reflection of a gradient with respect to the Turkish border.

	phone frequency method	feature frequency method	Levenshtein distance
Macedonian	0.52 ***	0.41 **	0.65 ***
Serbian	0.59 ***	0.24	0.80 ***
Romanian	0.53 ***	0.52 ***	0.34 *
Greek	-0.24	-0.56	-0.11
Turkish	-0.49 ***	-0.04	-0.21

Table 5: For each standard language linguistic distances to 50 Bulgarian dialects are calculated. Geographic distances are measured between the border of the corresponding country and the 50 dialects. The table shows the correlations between the linguistic distances and the geographic distances. We measured linguistic distances with the PFM, the FFM and Levenshtein distance. Correlations with $p<0.05$ are marked with *, those with $p<0.01$ are marked with ** and those with $p<0.001$ with ***.