Measuring linguistic unity and diversity in Europe

Erhard Hinrichs[1] **Dale Gerdemann**[1]
John Nerbonne[2]

[1] Eberhard-Karls Universität Tübingen
Seminar für Sprachwissenschaft, Abt. Computerlinguistik
Wilhelmstr. 19, 72074 Tübingen, Germany

[2] Rijksuniversiteit Groningen
Alfa-informatica, Fac. der Letteren
Oude Kijk in 't Jatstr. 26, Postbus 716
9700 AS Groningen, The Netherlands

in cooperation with: **Luchia Antonova-Vassileva, Petya Osenova, and Kiril Simov**
Bulgarian Academy of Sciences, Sofia, Bulgaria

1 Introduction

The use of language is a key factor in establishing and maintaining cultural identity. A common language provides the basis for passing on cultural heritage and for enabling social interaction and identification. Due to the intricate interplay of language and culture, cultural diversity is often reinforced by (real or perceived) differences among the language varieties of groups, while group identity can be solidified by the (real or perceived) unity of the group's language, even despite some internal diversity.

Linguistic diversity has important political ramifications, in particular for 21st-century Europe. A central example is minority language rights, which are guaranteed by the European Union, obliging member states to evaluate their own internal linguistic diversity in order to respect the conditions of the relevant European Charter.

For these reasons, the study of linguistic unity and diversity must be recognized as a central component of the study of cultural unity and diversity and their political consequences. The concepts of unity and diversity are however not clear-cut; rather they represent contrary directions on a continuum, and their application is conditioned by cultural and political context. A cultural or linguistic group may be treated, or treat itself, as a coherent unit, with the justification of its external difference to others, despite some degree of internal diversity within the group. This creates a conceptual conflict, raising the question of how much diversity can exist within a putative linguistic/cultural unit. For this reason, it is desirable when possible to go beyond qualitative assessments and seek methods to quantify the degree of unity vs. diversity in a given situation.

The project proposed here aims to take well-established quantitative techniques for measuring language diversity and apply them to the language varieties of Bulgaria and the surrounding territory. These methods have already been used successfully for the study of language diversity and unity in Western Europe, but the cultural and linguistic context of Bulgaria within the larger Balkan language area presents new challenges not encountered in the Western context.

Bulgarian in its Balkan context is an optimal domain of application for this research for the following reasons. First of all, Bulgarian is a central participant in the so-called Balkan *Sprachbund*, in which diverse languages share a large number of common linguistic features due to centuries of cultural contact. Furthermore, the West Germanic languages to which our techniques have been applied previously, and the Balkan Slavic languages (Bulgarian and its neighbors Macedonian and Serbian) to which we aim to extend, together form two classic examples of *dialect continua*, i.e. situations of gradual increase in dialect diversity over space. These two areas are thus perfect arenas for the application of quantitative techniques of linguistic distance measurement and of comparison between dialect continua which occur in different cultural contexts.

This three-year project will rely on the collaboration of three teams with expertise in linguistics and linguistic computation. A Bulgarian team will supply machine-readable data on dialects of Bulgaria and Macedonia; a Dutch team will use computational dialectometry to quantify language and dialect diversity and unity; and a German team will apply methods from machine learning and mathematical Information Theory to quantify the impact of language diversity on communication and mutual predictability between diverse language variants. The Western European teams involved have experience studying the measurement of linguistic diversity in Western Europe; thanks to sustained efforts of data collection by the Institute for Bulgarian Language at the Bulgarian Academy of Sciences, extensive Bulgarian dialect data are now available, some of them already in electronic form. This provides a unique opportunity to extend the quantitative study of language diversity to a Southeastern European context. The goals of this project could not be achieved without the sustained collaboration among the three partner institutions for the entire duration of the project.

Besides contributing to the study of Balkan linguistic unity and diversity, this work will have an important comparative dimension: we believe that investigating parallels and points of divergence between the previous western case study and this Southeastern European counterpart will yield more widely applicable insights into the powers and limitations of the quantitative study of language diversity.

1.1 Political dimensions of intra-national language diversity

The *European Charter for Regional or Minority Languages* (Strasbourg, 5.XI.1992) obliges European states to promote the use of regional or minority languages.[1] According to this Charter, speakers of a language variety[2] may apply for the status of regional or minority language to the European Community. In Part 1 Article 1, regional or minority languages are defined for the purpose of the charter as languages that are:

1. traditionally used within a given territory of a State by nationals of that State who form a group numerically smaller than the rest of the State's population; and

2. different from the official language(s) of that State.

Dialects of the official language(s) of the state, and the languages of migrants are excluded. Regional languages are those defined by the region within which they are spoken (e.g., Low German or *Plattdeutsch*), and minority languages are spoken by recognized ethnic minorities (e.g., the Sorbians in the eastern part of Germany or the Danes in the North).

[1] The charter can be found at: `http://conventions.coe.int/Treaty/EN/Treaties/Html/148.htm`.

[2] So as not to prejudge issues of language boundaries and dialect status, we use the neutral, and necessarily somewhat vague, term "variety" to refer to the speech of a community or group of communities.

It is instructive to examine the effects of this policy in one of the community states. The charter distinguishes two classes of regional or minority languages and two levels of state responsibility toward them. Both levels have played a role in the recognition of regional or minority languages in the Netherlands, where different groups of varieties are recognized as minority or regional languages. In 1995 Low Saxon was recognized as a regional language according to Part II of the charter, followed in 1997 by Limburg. This means that administrative divisions (e.g., subdivision into states, provinces, districts or other official territories) may not constitute an obstacle to the promotion of the regional language. The government is further obliged to facilitate and encourage the use of the regional or minority language in speech and writing, in public and private life. Means for teaching and studying the language should be available.

In 1996 Frisian was recognized accordingly as a minority language under Part III of the charter. Part III requirements include those of Part II. In addition, a language with this status is a compulsory subject in primary schools, and it may be used in courtrooms, and in contact with administrative authorities, local and regional authorities and in public services. It is clear that both statuses carry with them serious political and administrative consequences, and therefore that there should be clear criteria about when a group of varieties deserves the status of regional or minority language.

If European policy depends on a distinction between language and dialect, then it is desirable to have an objective and scientifically verifiable method to distinguish the two. It is similarly desirable to find objective ways of determining when the status of "regional or minority language" should be considered.

It would be naïve to expect a mathematical/computational approach to language diversity to provide an all-encompassing solution to this problem. In general, linguists are loath to make claims indicating when two linguistic varieties should be regarded as dialects of a single language and when they should be regarded as separate languages. Some attempts at definition result in counterintuitive consequences. For example, it is sometimes claimed that two varieties should be regarded as dialects of the same language if, and only if they are mutually intelligible. Where dialect continua exist, however, neighboring varieties are generally mutually intelligible and more distant varieties are not. There are then no neat lines dividing mutually intelligible subsets of varieties from one another. Max Weinreich famously quipped that "a language is a dialect with an army and a navy" (*"a shprakh iz a diyalekt mit an armey un a flot"*, in Weinreich 1945).[3]

The criterion of "mutual intelligibility" has itself been challenged as too vague. The famous Harvard Germanist, Einar Haugen, noted that Scandinavians of different native languages often converse, each using his own language, resulting in what Haugen dubbed "semi-communication" (Haugen, 1966). Clearly, mutual intelligibility is not ideal as a categorical criterion of language identity; instead varieties can be related by gradient degrees of intelligibility which do not result in clear-cut linguistic boundaries.

Despite this conceptual hurdle, mathematical and computational techniques can make an important contribution to language status debates. The goal of the project proposed here is to employ measurements of linguistic distance (including their relation to intelligibility) in an attempt to provide a quantitative basis for addressing issues concerning state languages, minority and regional languages, and dialects in Europe. The work proceeds from a measurement of linguistic distance which has been validated in extensive dialectological work and which more recently has been applied to questions of language policy. The work crucially involves two western teams with a history of technical collaboration and also a team from a candidate state with access to novel data involving various minority languages, a range of varieties, and a

[3]Wales Browne has more recently updated Weinreich's criteria, perhaps only half-jokingly, to include "an airline…, a seat in the United Nations, and a soccer team with the national colors" (Browne, 2002).

notoriously complex case of linguistic "mixing."

1.2 Context: Linguistic unity and diversity in the Balkans

Language variation concerns both dialectal differences of the same language as well as the distance of individual dialects to a commonly accepted standard variety of a language. While dialectal variation is due to a complex interaction of geographical and sociological factors within a speech community, language contact is generally regarded as the crucial external source for language variation and language change. Therefore, situations of extended language contact, particularly of typologically and genealogically diverse languages, provide the best, albeit also the most complex, environments for case studies of language diversity and unity (cf. Labov (1972); Chambers and Trudgill (1980); Hock (1991); Wolfram and Schilling-Estes (1998)).

In situations of extended language contact, issues of language policy such as the definition and promotion of a so-called standard language as well as the attitude towards non-standard dialects typically arise. The most well-known case of extensive mutual influence is the so-called *Balkan Sprachbund* (Trubetzkoy, 1930). It involves five distinct subbranches of the Indo-European language family: Slavic (Bulgarian, Macedonian, Serbian), Albanian, Hellenic (Greek), Romance (Romanian), and Indic (Romany) and is characterized by common grammatical features that are cross-fertilized among distinct language families due to geographical proximity (Asenova, 1989; Gilbers et al., 2000) rather than due to membership in the same language family. In addition, due to centuries of occupation, there is extensive lexical borrowing from Turkish.[4]

Due to its central geographical location within the Balkan region, Bulgarian exhibits to a high degree the linguistic innovation of the Balkan Sprachbund. Trubetzkoy (1930), who introduced the concept of the Sprachbund, described Bulgarian as the prime example of the effect of areal influence of unrelated or distantly related language on the form of a language. A large-scale data collection effort of Bulgarian dialects at the Bulgarian Academy of Sciences, whose beginnings date back to the 1950s, has produced an invaluable basis for a systematic study of language unity and diversity in a Sprachbund situation.

Besides the complexities of the relation between Balkan Slavic languages and their non-Slavic neighbors, the relationships between the Slavic varieties spoken by the majorities in Bulgaria and the former Yugoslavia, and by language minorities in neighboring countries, present an extremely complex picture. On the one hand, their diversity is sufficiently great to exclude understanding between distant varieties. Yet at the same time, a number of features unite all these varieties, justifying their universal recognition as South Slavic language family separate from both East Slavic (e.g. Russian, Ukrainian) and West Slavic (e.g. Czech, Polish).

Furthermore, the official language boundaries and official status of many varieties remain points of contention and are bound up in complex ways with historical and political factors. Before the 19th century, the primary identity marker distinguishing cultural groups in the Balkans was religion (Friedman, 2003). However, in this role, religion was replaced by language in the 19th century, as new nation-concepts arose as a challenge to the hegemony of the external power bases with their religious associations (Islam with Turkey, Orthodox Christianity with Greece). The Slavic nations sought to legitimize their separation from the Greeks and Turks by the creation of standard national languages based largely on spoken varieties (rather than on the previous historical model provided by the archaic Old Church Slavonic). This provided a pressure in the direction of increased unity, as actual diversity was glossed over to empha-

[4]It is interesting to note in this context, as Friedman (2003) does, that in political discourse the term "Balkanization" has come to emphasize diversity and fragmentation, while its linguistic sense is exactly the opposite: it calls up the shared features that unite the Balkan languages despite their diverse origins.

size the common element of the Slavic language community as against the Turkish or Greek "Other". However, subsequent schismatic forces emphasized the diversity of the spoken varieties underlying these newly-created standards to the end of establishing further distinctions, notably in the efforts to separate a Croatian standard language from the Serbian one and to separate a Macedonian standard language from the Bulgarian one. In Friedman's terms, "[i]t is the politicization of linguistic drift (the natural tendency of languages to differentiate over time) that has had particularly dramatic effects in Southeastern Europe, where language has become a vehicle of conflicting *centripetal* [i.e. unifying] and *centrifugal* [i.e. diversifying] forces" (p. 26; emphasis added).

Most important for our research in its relationship to Bulgarian is the status of Macedonian. The term "Macedonia" is traditionally applied to a region which after World War I was divided between Serbia, Bulgaria, and Greece (as well as referring to a district of Greece). The Yugoslav portion of this territory was granted partial autonomy in 1943, leading to official recognition by Yugoslavia of a Macedonian language in 1944 and a codification of the orthography in 1945, based on a Serbian, rather than Bulgarian, model (Schaller, 1975). However, disagreement persists over this variety's status as a "language". Official recognition has in particular been withheld by Greece and Bulgaria, the latter categorizing it as a "regional norm" of Bulgarian (Friedman, 1993). Moreover, varieties recognized as Macedonian (according to the official position in the former Yugoslavia and current Macedonia) are spoken outside the political borders of Macedonia, in Bulgaria and Greece inter alia.

The linguistic facts are similarly intricate. Bulgarian and Macedonian language varieties form a subgroup of comparative unity within the range of South Slavic diversity. This grouping is supported by shared innovative features in these two regions. Some are phonological, such as the loss of the vowel-length distinctions (which persist in Serbian/Croatian). Syntactic and morphological features also unite the two (e.g. strongly analytic sentence structure with the absence of the noun cases found in Serbian and other Slavic languages).

The picture of Macedonian-Bulgarian linguistic unity is however disrupted by several dimensions of diversity, which are sometimes cited as linguistic justification for the recognition of Macedonian as a separate language. Unlike standard Bulgarian, Macedonian varieties exhibit such distinctive characteristics as: fixed antepenultimate stress (as opposed to Bulgarian's free, movable stress); a three-way neutral/proximal/distal distinction in the definite article; and certain historical phonological changes which have led to persistent pronunciation differences, e.g. the development of the Proto-Slavic *jat′* vowel into *e* (as opposed to the default standard Bulgarian outcome *ja*); and the development of the Proto-Slavic consonant sequences *tj* and *dj* vowel as *k′* and *g′*, resp. (as opposed to the standard Bulgarian outcomes *št* and *žd*). Thus the picture of Bulgarian-Macedonian linguistic unity must be strongly qualified to take these differences into account. But even this qualification must be further qualified by the observation that many of these characteristically Macedonian features are shared with varieties spoken within Bulgaria (such as the developments *jat′* $>$ *ja* and *tj*/*dj* $>$ *k′*/*g′* in Western Bulgarian, and the existence of a tripartite definite article in Bulgarian Rhodope varieties).

Furthermore, even the separation of the Bulgarian-Macedonian unit from Serbo-Croatian is far from categorical, due to the existence of intermediate dialects, such as that of the Timok region, which has Serbian characteristics alongside a postposed definite article and analytic nominal inflection.

As the preceding paragraphs demonstrate, the Balkan Slavic linguistic situation exhibits a pronounced tension between diversity and unity, both on political and linguistic levels. We believe that it is not possible to fully do justice to the complexity of the linguistic aspects of this situation without the use of quantitative methods of dialect and language comparison.

1.3 The quantitative study of linguistic diversity

Quantitative techniques have several advantages in the study of linguistic unity and diversity. They are better equipped to describe the gradient nature of diversity in a continuum situation, and similarly to describe the gradient nature of the concepts of (linguistic) unity and diversity themselves, as discussed above. Qualitative characterizations of the distribution of linguistic features run the risk of oversimplification, with potentially dangerous consequences. For example, Friedman describes how at the end of the 19th century, "...linguists were putting their knowledge at the service of politicians by choosing one or another isogloss [linguistic feature boundary] as the definitive justification for the ethnic identity—and therefore the nationality—of the Slavic speakers on Macedonian and adjacent territory" (Friedman, 2003, p. 20). To the degree that quantitative techniques can simultaneously encompass a variety of features, they can, when applied with care, avoid positing such arbitrary and artificial categorical distinctions which the empirical facts do not unambiguously support. Of course, quantitative methods are also not immune from potential abuse, but insofar as the measurements made rely on extensive empirical data, it is more difficult to (consciously or subconsciously) finesse details to support some ideal categorization.

Similarly, the subjective judgments of speakers of varieties in a dialect continuum about dialect distance often show a disconnect with the purely linguistic details. For instance, Herson Finn (1996) investigates the attitudes of Macedonians about which varieties are closest to their own;[5] her results clearly demonstrate the degree to which nonlinguistic factors influence such judgments (for instance, Croatian is widely perceived by Macedonians as closer to Macedonian than Bulgarian is, despite all linguistic evidence to the contrary). Although these attitudes are themselves an important component of the Balkan sociolinguistic situation (and amenable to quantitative investigation in their own right; witness Herson Finn's data analysis), they show the unreliability of subjective qualitative judgments as an indicator of dialect distance in a purely linguistic sense.

For these reason, the study of Balkan Slavic linguistic diversity in general, and in particular the study and the relative influence of Macedonian, Serbo-Croatian, and other standard language varieties on neighboring Bulgarian varieties, stands to be informed by quantitative methods for measuring linguistic diversity. Whereas this field of study has so far relied solely on qualitative criteria, our research, by applying more fine-grained and state-of-the-art mathematical/computational techniques, can provide a richer quantitative characterization of the linguistic distance of non-standard Bulgarian dialects from the various standard South Slavic languages, allowing in particular for a more informed evaluation of the status of non-official varieties.

2 Previous Research and State of the Art

As discussed in section 1.1 above, the *European Charter for Regional or Minority Languages* relies on a criterion of external diversity, that the candidate variety be "different" from the official language. The charter does not describe how one might determine whether a candidate regional or minority language is different from a standard language, and certainly not how the difference might be operationalized (or measured) in an objective way. In this section we review, special emphasis on our own contributions to the state of the art, research on the measurement of linguistic distance which can shed light on this issue, with the goal of demonstrating the

[5] Besides questionnaires about language attitudes, Herson Finn uses the ingenious method of presenting subjects with a text which, by means of gradual transitions through sections of mixed vocabulary and grammar, shifts from Serbian, through Macedonian, and finally to Bulgarian; subject are asked to mark off where "your language" begins and ends.

feasibility of measuring the degree to which a candidate minority or regional language should qualify as "different".

2.1 Applying linguistic distance metrics to Dutch

Our research has focused on the application of linguistic distance metrics to Dutch and German and their implications for the special status of minority varieties (Nerbonne et al., 1999*a*; Heeringa and Nerbonne, 2001; Heeringa, 2004; Nerbonne and Siedle, 2003). To find varieties in the Dutch dialect area which might be candidates for special status, we used a sample from the well-studied *Reeks Nederlandse Dialectatlassen* (RND), a series of Dutch dialect atlasses which were edited by Blancquaert and Peé in the period 1925–1982. From the nearly 2 000 RND sites in the RND, we selected 360 evenly distributed over the Dutch language area. Each dialect transcription in the atlas records 139 sentences from which we examined the same 125 words, which may be regarded more or less as a random sample (of the RND material).

To be able to compare the varieties with respect to Standard Dutch in order to evaluate their potential as candidates for the status of regional or minority languages, we added a transcription of Standard Dutch. We tried to guaranteee consistency with existing RND transcriptions by using the instructions Blancquaert gave to his field worker-transcribers (1939). We took the liberty of deviating from Blancquaert's guidelines only to conform to one completely uncontroversial aspect of contemporary pronunciation, the loss of the *n* in the weak final syllable *-en*.

We also added Standard German, based on the *Wörterbuch der deutschen Aussprache* (Krech and Stötzer, 1969, pp. 88–90). We included German in order to measure the degree to which the Dutch dialects along the German border are distinct not only from the Dutch standard language, but also from the German standard.

We first sought candidates for the status of regional language by determining the prominent groups in the collection of 360 varieties. We determined the difference in pronunciation between all of the pairs of the 360 varieties using LEVENSHTEIN DISTANCES (see section 4.2 below for an explanation of Levenshtein distance). This yields a distance matrix of 360×360 dialects which we can analyze in various ways (especially cluster analysis and multi-dimensional scaling). (Because cluster analysis is normally an exploratory technique, we also subjected our clusters to revalidation analysis.) Cluster analysis identifies *clusters* of varieties, i.e., groups of similar varieties and therefore presumably strong candidates for the status of minority or regional language.

In Figure 1 the 13 most significant Dutch groups—as derived by clustering—are shown. For each of these groups we further calculate an INDEX OF LINGUISTIC INDEPENDENCE. This index is calculated in two steps, (i) external diversity and (ii) internal unity, which we now explain.

2.2 External diversity: Distances to official language

We measure external diversity by examining the distance between a candidate regional or minority language and the standard language or languages it might be identified with. Figure 2 shows each group's average distance to Standard Dutch and Standard German. Each group is on average closer to Standard Dutch than to to Standard German. Since there are no other candidates to function as reference languages (other than Dutch or German), it is reasonable to view all groups as Dutch varieties.

In the graph we see that Frisian is furthest from Standard Dutch. This is in accordance with general opinion, and also in accordance with Frisian's unique primary status as a regional or

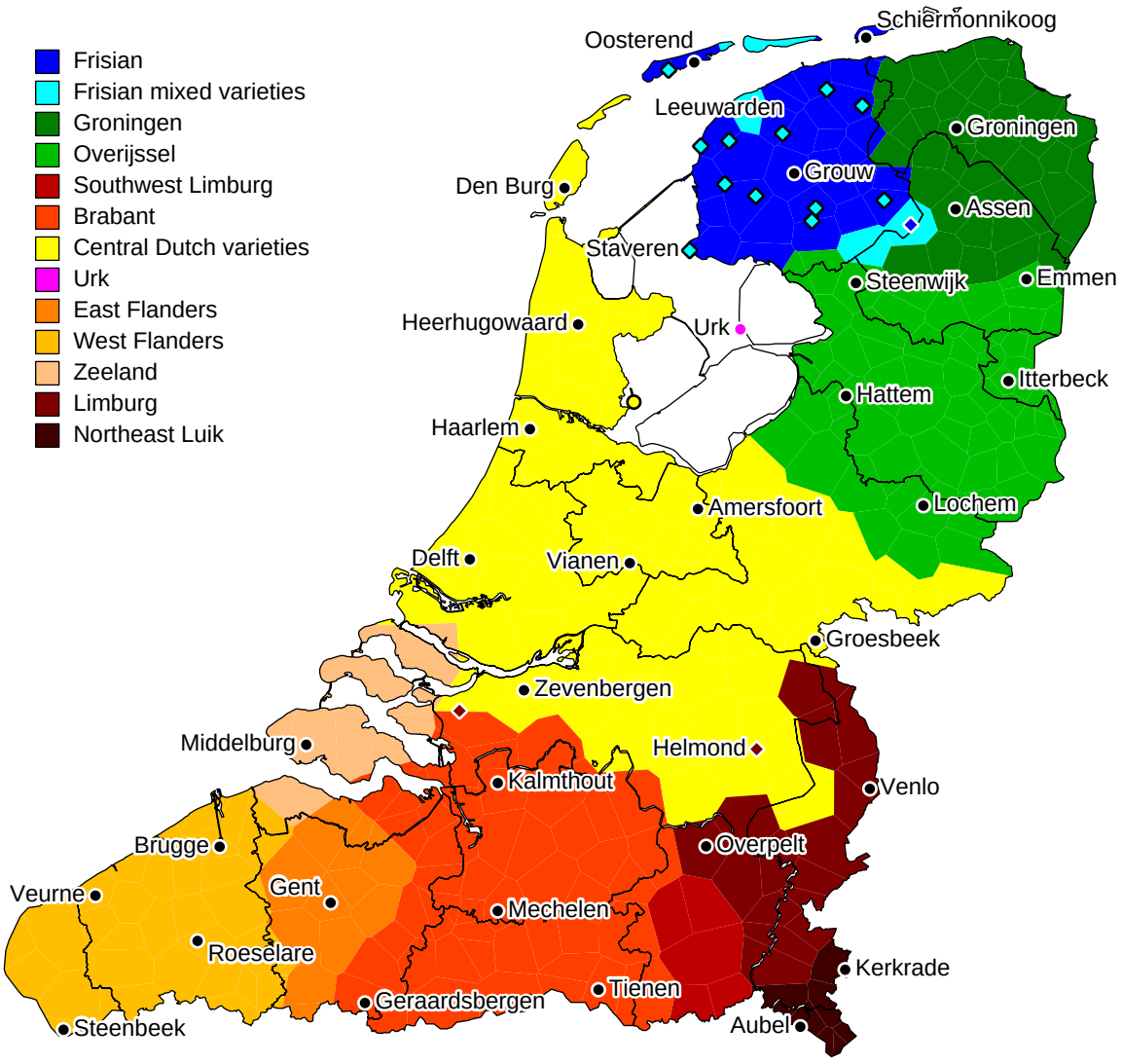

Figure 1: The 13 most significant groups among more than 350 language varieties spoken within the Dutch-speaking area. Diamonds represent language islands. Note that varieties from Frisian (blue and light blue), as well as from Low Saxon (Groningen and Overijssel, both in dark green) and Limburg (the red-brown areas in the Southeast) are included. Frisian has been recognized as top-status regional or minority language, and Low Saxon and Limburg have been awarded secondary status. The scattered light blue "mixed Frisian" varieties constitute a well-known distributed variety also known as "town Frisian".

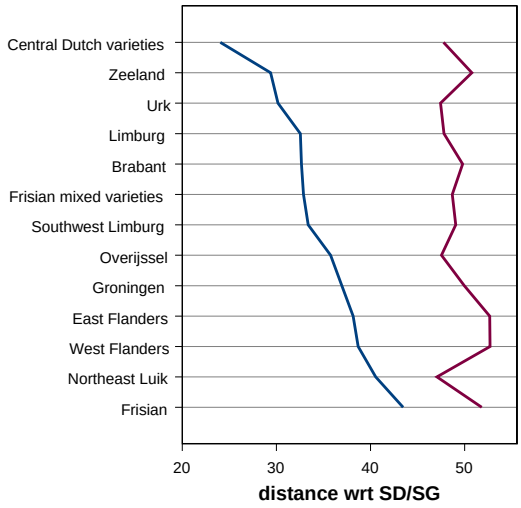

Figure 2: For each of the 13 most significant groups in the clustering the average distance with respect to Standard Dutch (SD, blue, left) and Standard German (SG, red, right) is calculated. The distance between Standard Dutch (SD) and Standard German (SG) is 43.4%.

minority language. Frisian has a distance of 43.5%, while the distance between Standard Dutch and Standard German is 43.4%. For Northeast Luik we also found a large distance: 40.6%. Luik's distance to Standard German (47.1%) is even larger, but of all groups it is closest to Standard German. Other relatively large average distances from Standard Dutch are found for West Flanders (38.7%) and East Flanders (38.2%).

2.3 Internal unity: Group coherence

Although the *European Charter* mentions only a criterion of external diversity for granting special status on a group of varieties, we would find it implausible to award a special status to a group of varieties which showed little linguistic coherence among themselves. In an extreme case any collection of remote varieties might qualify as sufficiently different from a standard language, but they should not therefore count as a regional or minority language themselves unless they were minimally mutually intelligible. We therefore suggest that one needs to measure the internal unity of candidate minority or regional languages as well as their difference from standard languages. To measure the internal unity of a group of varieties, we use INTERIOR DISTANCE, i.e. average distance among the varieties within each candidate group. If this is not sufficiently small, we suggest that the set of varieties under consideration for the status 'minority or regional language' is internally too diverse for recognition. In Figure 3 the interior distance for each group is shown.[6]

We found the lowest interior distance for Frisian: 15.3%, indicating that the Frisian area is quite homogeneous. Frisian is followed by the two Lower Saxon groups: Groningen (19.4%) and Overijssel (19.6%). Although the Frisian mixed varieties are not found in a single area, but rather as scattered "dialect islands" for the greater part (see Figure 1), the mean interior

[6]Since one "group" is based on a single variety (Urk), its interior distance is zero. In practice such occurrences would presumably not arise: we would probably refuse to consider the linguistic independence of a putative group without several measurements.

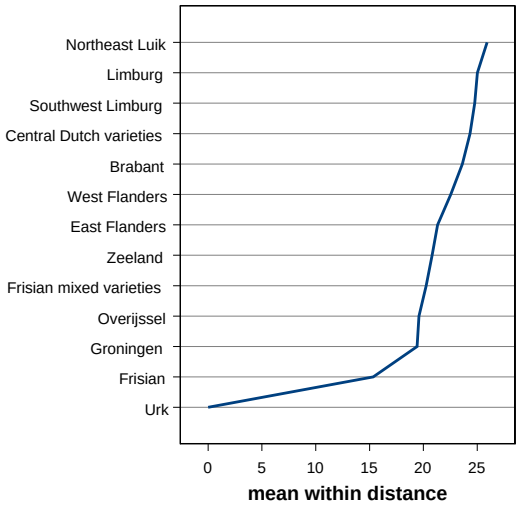

Figure 3: For each of the 13 candidate groups the *mean interior distance*, i.e., the mean distance between each pair among the varieties in the group is calculated.

distance for this distributed variety is still 20.3%. Most heterogeneous are the three Limburg groups: Southwest Limburg (24.8%), Limburg (25.0%) and Northeast Luik (25.9%).

2.4 Index of Linguistic Independence

Given our measurements of external diversity and interior unity, we can combine them into index of linguistic independence. The index of linguistic independence may be calculated very simply as the average distance of a group of varieties with respect to the closest standard language minus the average interior distance within that group. In Figure 4 for each of the 13 groups in Figure 1 the index of linguistic independence is given (naturally in the course of the proposed project we would examine variation on this definition).

We clearly find Frisian as the most convincing candidate.[7] This is not surprising since is has the highest average distance with respect to Standard Dutch *and* the lowest group incoherence (the smallest interior distance). The index of independence is equal to 28.1%. This indicates that Frisian's recognition according to section III of the *European Charter for Regional or Minority Languages* is strongly supported by the linguistic evidence.

We find Groningen as the second candidate (17.6%) and Overijssel as the fourth (16.2%). Groningen and Overijssel together form the Lower Saxon group which is recognized under section II of the *European Charter for Regional or Minority Languages*.

If Low Saxon is recognized, than East Flanders (16.3%) and West Flanders (16.1%) are likewise excellent candidates. The recognition of Limburg has little justification (7.5%) because the area is too heterogeneous. It is not surprising that the Central Dutch varieties have the lowest language index (-0.3%). They are neither particularly different from Standard Dutch nor particularly coherent among themselves.

We are naturally aware of the many slight alternatives which might be proposed (various weightings of the external distances vs. the interior homogeneity), as well as alternative per-

[7] Ignoring Urk—see footnote 6.

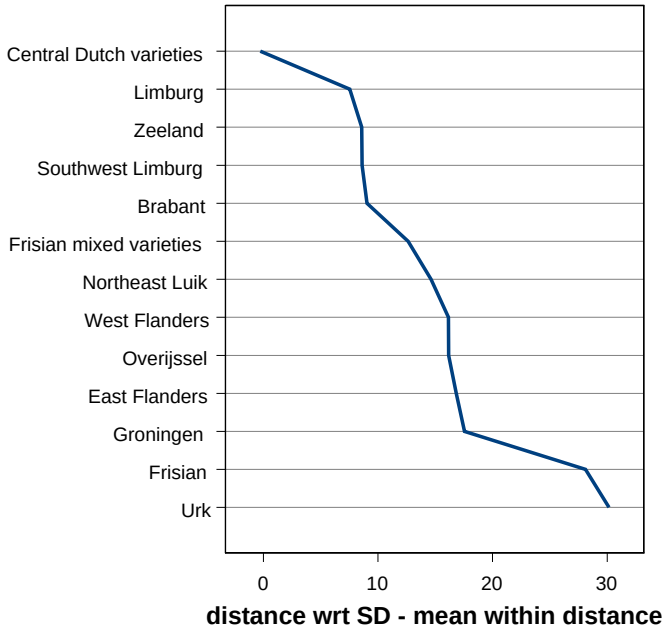

Figure 4: For each of the 13 most significant groups in the clustering *distance with respect to Standard Dutch − mean interior distance* is calculated.

spectives (see Section 4.3 below). However, the point of these calculations has been to demonstrate the feasibility of the research program which is proposed here.

3 Project Goals, Hypotheses, and Expected Results

The goal of the project proposed here is to extend our work on linguistic distance measurement from its previous applications in Western Europe, described in section 2, to the more complex language situation in the Balkans. Specifically, we will calculate and evaluate metrics of linguistic independence using the Bulgarian dialect data described in section 4 below. We will measure the linguistic distance between non-standard varieties of Bulgarian, between these varieties and standard Bulgarian, and furthermore between these varieties and standard Macedonian. Besides their pairwise distance, we will also use interior distance in the technical sense described in section 2 as a quantitative measure of the linguistic unity vs. diversity of groups of varieties in order to detect internally cohesive units. At the same time, we will apply computational techniques described in section 4.3 below for the automatic discovery of regular sound correspondences between varieties and use these correspondences to quantify mutual predictability between varieties as an additional metric of distance.

As discussed in section 1.1 above, the wider goal of this work is to provide a quantitative operationalization of the linguistic condition of the European Charter's definition of a minority or regional language, i.e. that such a variety be "different from the official language(s)" of the state. We do not claim that our method will always be able to provide a categorical yes/no answer to questions of language status; however, ours represents the first attempt to bring quantitative scientific rigor to this politically charged issue. We also believe that our measurement of

homogeneity will inform the issue with a subcriterion of internal unity (internal cohesiveness of a group of varieties) to complement the official criterion of external diversity (difference from the official language).

A related subgoal of this research is to critically examine our quantitative methods' applicability to language situations differing in essential ways from the Western European arena in which they have thus far been tested. The study of language diversity in Southeastern Europe, and in Bulgaria in particular, forces our methods to confront issues of language contact and areal diffusion to a much greater extent than has been necessary in Western Europe. Consequently, we anticipate that these factors will manifest themselves in the measures of linguistic distance to be described in more detail in section 4. In particular, we hypothesize that, due to loan words, there will be a higher degree of irregularity among isoglosses than is witnessed in the, by comparison, rather homogeneous West Germanic dialect situation.

Another hypothesis that we will seek to verify concerns the status of Macedonian vis-à-vis standard Bulgarian. Here we anticipate that our measures of linguistic distance will demonstrate that Macedonian shows significant differences from other varieties, whose status as Bulgarian dialects is undisputed, in terms of external diversity and internal homogeneity.

A third hypothesis concerns the rate of change and innovation among Bulgarian dialects due to recent political developments. The mobility of people within a language area, and forces which promote interchange, most notably trade, jointly contribute to a well-known leveling—or convergence—of language varieties within a language area. We therefore hypothesize that recent political developments that enable a much more extensive interaction of Bulgarians with their neighbors should likewise improve preconditions for more extensive contact-induced changes. As a consequence, we expect to find an accelerated rate of change in the data compared to the data collected for previous decades.

4 Methods

This section describes the Bulgarian dialect data which will be analyzed (section 4.1 and the techniques to be used for their analysis (sections 4.2 and 4.3).

4.1 The Bulgarian Data

The Bulgarian dialect data cover all main regions of Bulgaria and also[8] the Tetovo, Skopie, Kumanovo, Kratovo, Deber, Kichevo, Veles, Stip, Radovish, Struga, Ohrid, Resen, Bitolya, Prilep, Negotino, Strumica, Gevgeli regions of Macedonia, the Zajchar, Balevac, Knyazhevec, Aleksinac, Pirot, Bela Palanka, Nish, Leskovac, Vranya regions of Eastern Serbia, the Korcha region of Albania, the Kostur, Lerin, Kajlyar, Voden, Enidzhevardar, Kukush, Solun (Tesaloniki), Seres, Valoviste, Drama, Kavala, Ksanti, Gyumyurdzhina, Dedeagach, Soflu, Dimotika regions of Northern Greece, the Odrin, Uzunkyopryu, Lyuleburgaz, Viza, Chorlu, Chataldzha regions of Turkey, and the Kalafat, Krajova, Karakal, Slatina, Turnu Mygurele, Bukurest, Kylyrash, Tulcha regions of Romania. The data were collected from 2 200 villages on the basis of 49 phonetic, 58 morphological questions and 100 lexical questions. The dialect data collection consists of 2 million lexical records that have been collected over a span of 45 years dating back to 1956.

These data contain a wide range of varieties, including dialects which, as discussed in section 1.2, show features found in the standard languages of Macedonia and Serbia but not in

[8] All place-names are listed here in their Bulgarian form, in accordance with the data of the atlas.

standard Bulgarian. They allow us therefore to compare the dialects in southwest part of Bulgaria and southeast part of Macedonia with each other and with the literary languages of Bulgaria and Macedonia; likewise we can compare the dialects of the northwest part of Bulgaria, the east part of Serbia, and the north part of Macedonia with each other and also with the three literary languages (Bulgarian, Macedonian and Serbian).

The data have been published in book form as the Bulgarian Dialect Atlas, consisting presently of 6 volumes (Stoykov and Bernshteyn, 1964; Stoykov, 1966; Stoykov et al., 1975, 1981; Mirchev, 1972; Kochev et al., 2001) with two additional volumes in preparation (Tsanov, in preparation; *Atlas of Bulgarian Dialects: Morphology*, in preparation).

Figures 5 and 6 show two representative samples of different dialect maps for Bulgarian isoglosses that have been compiled from the dialect data.

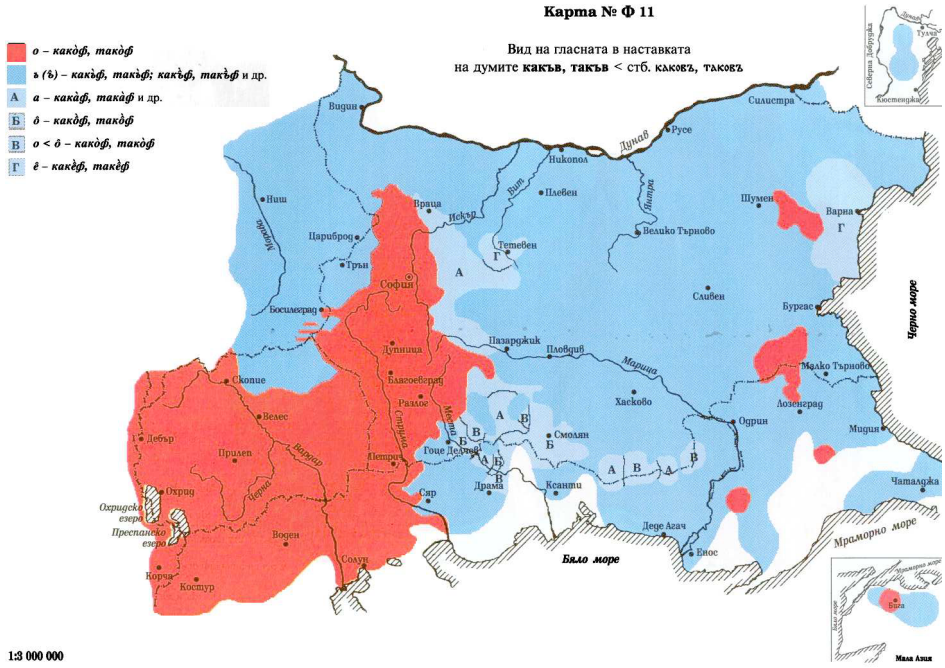

Figure 5: Description of the vowels in the suffix of the words ‘kakav’, ‘takav’ (‘what, such’)

Approximately 20% of the data are also available in digital form, stored as XML records. One of the first tasks of the present project will be to extend the digitization of the data by automatic methods using Optical Character Recognition software and by manual postediting and correction of these automatically acquired records.

As examples, we give in figure 7 two lexical records from the archive for the same literary word `de"lva` (in English ‘pot’). The entries demonstrate a difference in the phonetic variants of this word in the two villages. This difference concerns the phenomenon of ‘non-reduction vs. reduction’ of the final vowel ‘a’ in an post-tonic position. Note that the transliteration is performed in accordance with SAMPA specifications.

Each lexical record consists of the following elements:

- normalized variant (form of lexeme in Standard Bulgarian): XML element `<nw>`
- phonetic transcription of variant: XML element `<phV>`

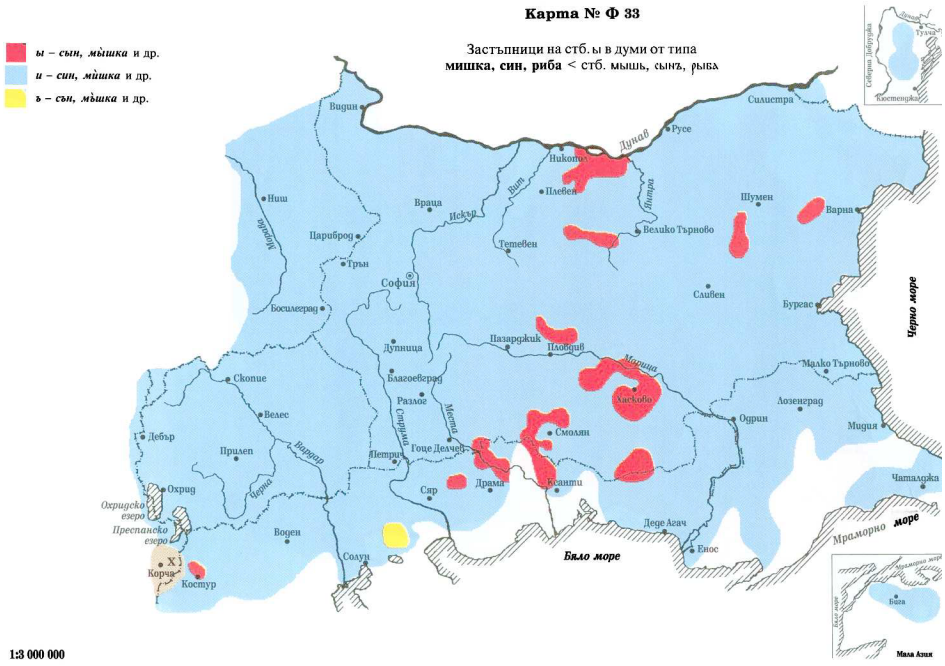

Figure 6: Description of the old Bulgarian vowel ‘y’ in the words from the type ‘mishka’, ‘sin’, ‘riba’ (‘mouse, son, fish’)

- morpho-syntactic features (part of speech, gender, etc.: XML element `<morph>`
- definition: XML element `<sense>`
- semantic category (a semantic classification of the lexeme with respect to a set of predefined semantic classes, such as *food*, *tools*, *clothes* etc.): XML element `<semGr>`
- example of usage (transcription of actual recording in real situation): XML element `<example>`)
- village/town where recorded: XML element `<vil>`
- signature (who recorded the example or other sources of information such as books, publications, dissertations): XML element `<sign>`
- fixed phrases in which the lexeme is used in the corresponding village (no examples in figure 7): XML element `<combW>`

The form of the entries facilitates studies of language contact and areal diffusion by also including information about the origin of each lexeme—whether it is from a Slavonic language, Greek, Turkish, or Romanian. This is useful when the Bulgarian data is compared with the other Balkan languages.

Our data on the Macedonian standard language is drawn from *Rechnik na makedonskiot jazik* (1961–1966).

```
<entry>                                  <entry>
  <nw>de"lva</nw>                          <nw>de"lva</nw>
  <phV>d'e"lva</phV>                       <phV>d'e"lv@</phV>
  <morph>fem</morph>                       <morph>fem</morph>
  <sense>Delva.</sense>                    <sense>Delva.</sense>
  <semGr>S@dove.</semGr>                   <semGr>S@dove.</semGr>
  <example>V@v d'e"lva pr@"st'ena,         <example>D'e"lvi.</example>
           swar'e"nu ka"ktu
           z@ kva"s'en'e.</example>
  <vil>Orizovo, Slivensko</vil>            <vil>KotSovo, Preslavsko</vil>
  <sign>LA</sign>                          <sign>LA</sign>
</entry>                                 </entry>
```

Figure 7: Two lexical records for the literary word de"lva *pot*

4.2 Measuring linguistic independence: Levenshtein distance

Levenshtein distance[9] was introduced in Levenshtein (1965) as a measurement of the distance between two strings. It is defined in terms of string editing option (insertion, deletion, and substitution of symbols) and gives the least-cost sequence of edit operations necessary to change one string into the other. The cost of a sequence is the sum of the costs of the individual edit operations, which can be set equally or weighted unequally, depending on the application.

Nerbonne et al. (1999*b*) demonstrate that an appropriately defined Levenshtein distance algorithm provides a reliable measure of pronunciation differences, which can be used on language samples to characterize aggregate dialect differences.

Proceeding from this notion of linguistic distance, we can examine candidate minority or regional languages as instantiated in a set of varieties. As discussed in section 2 above, we measure on the one hand the external diversity of a variety in terms of its average distance from the official language (d_o), and on the other hand the internal unity of the group in terms of its interior distance (d_i).

Our attempt to operationalize the determination of state will consist in the examination of an *index of linguistic independence*. We shall calculate this in what follows as $d_o - d_i$. In the course of the project we will wish to examine alternative measures, e.g., $d_o/(\sigma_o^2 + \sigma_c^2)$, where σ_c^2 is a measure of variance with respect to a "center" of the candidate regional or minority language and σ_o^2 the variance in the official language. This follows standard steps in data analysis (Schalkoff, 1992, pp.90-94).

As earlier sections document we believe Levenshtein distance to be an appropriate measure of linguistic distance. Our choice is confirmed by validation work (see Heeringa 2004, pp. 165–198), and a first attempt to deploy the Levenshtein distance metric in order to assess the linguistic independence of candidate regional or minority languages can be found in Gooskens and Heeringa (2004, pp. 69–77).

4.3 Regular sound correspondences and mutual intelligibility

The measurement of linguistic difference based on Levenshtein distance is a purely aggregate technique: it compares word-pairs in isolation and averages over the distances of a set of such pairs. However, as has been known since pioneering work in the 19th century by Osthoff, Brugmann, Paul, and other, related varieties tend to show patterns of *regular sound correspon-*

[9] This distance is also known under the name of *edit distance*.

dences. We will therefore complement the Levenshtein distance metric with a technique whose goal it is to automatically discover such correspondences.

As an example of a regular sound correspondence, consider the fact that an English word with an initial *p* often has a German cognate with an initial *pf* (plant-Pflanze, path-Pfad, penny-Pfennig, etc.). This correspondence is the result of a regular development of German initial *pf* from Proto-Germanic *p*, which is still retained in English. The fact that this correspondence is regular does not imply that it has no exceptions. Many exist, for instance as the result of borrowing from a common third language (e.g. English *pause* vs. German *Pause*).

Traditionally, regular sound correspondences are discovered by intensive labor on the part of trained linguists. However, the task of automatically detecting such correspondences in lists of word pairs has recently been studied by Kondrak (2002) inter alia. His method adapts techniques from statistical Machine Translation (Brown et al., 1990) designed to learn translation lexicons from bilingual texts: by seeing enough examples, the method learns that German *pf* is a common "translation" of English *p*.

The advantages of automatic correspondence detection are threefold. (1) Regular correspondences are interesting in and of themselves and are an essential part of historical linguistics; automatic techniques can be more efficient than human effort. (2) Correspondences provide the basis for state-of-the-art techniques of cognate detection. In the previous research on Western European languages described in section 2 above, determination of which pairs of words are in fact cognates is carried out by hand. However, these pairs can be extracted automatically from word-lists in tandem with the process of correspondence detection, using the technique of Kondrak (2002). (3) Regular sound correspondences play a role in making communication possible between related varieties. By automatically detecting them, we can provide a basis for a metric of the *mutual predictability* of two varieties, based on the mathematics of Information Theory.

Therefore, we will supplement the aggregate distance metric research described in section 4.2 with the automatic detection of regular sound correspondences. Corresponding to the three advantages described in the previous paragraph, we will put these to three uses. First, we will collect lists of correspondences to supplement linguists' previous dialectological and historical research on Bulgarian. Secondly, we will use the discovered correspondences as a basis for cognate detection to provide the cognate pairs for the Levenshtein distance metric, eliminating the subjective step in the initial collection of word-pairs.

Thirdly, we will use the discovered correspondences as a basis for a metric of the mutual predictability of related dilects. This metric will supplement the Levenshtein-distance metric with a sensitivity to predictable patterns of correspondence and will provide a measurement that better reflects the properties of related varieties that facilitate or impede communication. In the next subsection we describe how to exploit regular correspondences to derive such a metric.

4.3.1 Mutual predictability and Information Theory

Mutual intelligibility between related varieties is enhanced by lexical and grammatical similarities. In the phonological domain, which is our focus, intelligibility can be possible despite differences because of regular correspondences between varieties: a speaker of variety X can comprehend a word w_1 from a related variety Y insofar as s/he can *predict* which word w_2 of X is the most likely equivalent for w_1. When w_1 and w_2 are cognate, prediction can be based on sound correspondences.

Of course, the degree to which any given speaker of X can do such prediction Y is heavily dependent on factors such as his/her experience with Y, attitude towards Y, etc. For dialectometry it is therefore desirable to abstract away from these differences. To this end, Cheng

(1996), in his research measuring on Chinese dialect diversity, introduces the notion of "systemic mutual intelligibility" which is meant to be dependent only on the phonological mutual predictability of words of two varieties, absent any background knowledge about the other variety which a speaker may have.

Cheng suggests a measure of systemic mutual intelligibility where the score for two varieties is increased by the presence of regular sound correspondences with wide applicability ("signals" in his terminology) and decreased by irregular correspondences ("noise") and cognates in one dialect which are confusable with words of the other. However, Cheng's measure is derived simply by adding and subtracting arbitrarily chosen reward and penalty points for signals and noise; it therefore lacks a clear mathematical interpretation.

On the other hand, a sound theoretical foundation for the measurement of systemic mutual intelligibility is readily available from the mathematical Theory of Information (Shannon, 1948) and the "noisy channel model." Information Theory was an essential ingredient of early successes in speech recognition technology and machine translation in the 1970s and has since become one of the most widely applied techniques in computational linguistics.

The noisy channel model involves a source probability distribution $P_s(i)$ and a conditional channel distribution $P_c(o|i)$ which accounts for distortions introduced in the transmission of a message. In the context of understanding between related but distinct language varieties based on phonological predictability, we can apply the noisy channel model by letting P_s be a probability distribution over words of variety X, while $P_c(o|i)$ gives the conditional probability that a given word i of variety X will correspond to word o in Y. The speaker of X hearing a word o seeks to find the word i of X which maximizes $P_s(i) \times P_c(o|i)$. Since we are interested in studying the predictability of related forms in the abstract, as with Cheng's systemic mutual intelligibility, and since we are focusing on phonological variation, we will assume that the value of $P_c(o|i)$ is determined solely by the phonological form of i. In particular, we will understand "phonological form" to mean simply the sequence of segments in the word (more sophisticated representations are also possible).

The power of Information Theory comes from its ability to quantify the difficulty that $P_c(o|i)$ presents for the decoding process with the *conditional entropy* $H(X|Y)$. We propose to use conditional entropy as a measure of the predictability of one variety based on another. Like Cheng's systemic mutual intelligibility, it measures only the amount of information that Y gives about X and abstracts away from any given speaker's knowledge about other varieties. Note that $H(X|Y)$ is an asymmetric measure, in general distinct from $H(Y|X)$. This is in accord with the fact that a pair of varieties may exhibit one-way intelligibility: for instance, Danes can often understand Swedish more easily that Swedes can understand Danish (Hock 1991 p. 381). For the purposes of dialectometry and clustering a symmetric measure is desirable; for this we will use the mutual information $I(X;Y)$ between X and Y, which is the difference between the non-conditional entropy of one variety and the conditional entropy of the translation.

In order to estimate $P_c(o|i)$, we will use the alignments between words computed by Kondrak's algorithm for detecting sound correspondences. Knowing which sounds in one variety correspond to which in another, we can estimate the probabilities of such correspondences in a given phonetic context and then derive $P_c(o|i)$ based on the sounds in o and i. Due to inevitable data sparseness, we need to use a learning technique for deriving these estimates that is able to generalize on the basis of restricted data. We will apply the standard technique of probabilistic decision tree learning with pruning (see Mitchell 1997 ch. 3) with subsequent compilation of the decision trees into stochastic finite state transducers (Sproat and Riley, 1996).

In summary, our approach to the measurement of linguistic diversity in terms of entropy is novel, however it is built up out of well-established techniques in computation linguistics and information theory. We expect it to complement our measures based on Levenshtein distance

by taking sound correspondences and predictability into account.

5 Workplan

The proposed project is planned for a duration of 36 months. The research will require close collaboration by all three project partners over the entire duration of the project.

The general plan with respect to the partners' relationship in the project is as follows:

1. German and Dutch partners transfer state-of-the-art technology for measuring language contacts to Bulgarian and other contact Balkan languages (Macedonian, Serbian, Greek, Romanian);

2. Bulgarian partners ensure the digitized form and the consistency of a large database of dialectal data and adapt the transferred methodology and software;

3. As a result, the German and Dutch partners will benefit from testing their algorithms on new types of languages with non-Germanic phonetic systems. The Bulgarian partners will benefit from the support in data digitization, processing and observation, as well as from collaboration with experts on computational dialectometry.

Hence, having in mind the above partnership architecture, the main tasks for each project partner will be as follows:

5.1 Bulgarian Academy of Sciences

- Digitization of the Bulgarian dialect corpus by OCR methods, data conversion to XML format using the CLARK system (LML) (year 1).
- Manual post correction of digitized data and collection of additional language data for the contact languages (IBL) (year 1).
- Interpretation and validation of the computational results produced in Groningen and Tübingen (year 2 + 3).

5.2 University of Groningen

- Derivation of three types of similarity matrices (year 1 + 2):

    1. among Bulgarian dialects
    2. between Bulgarian dialects and Standard Bulgarian, and
    3. between Bulgarian dialects and relevant contact languages, using computational distance measures and clustering techniques.

- Évaultaion of different variants of Levenshtein distance on Bulgarian data

5.3 University of Tübingen

- Automatic discovery of regular sound correspondences among dialects and automatic identification of cognate forms (year 1). Specifically, we will extract the following correspondences:

    1. among dialects of Bulgaria and Macedonia;
    2. between these dialects and Standard Bulgarian; and
    3. between these dialects and Standard Macedonian.

- Probabilistic modeling of sound correspondences among dialects and between dialects and the standard languages (year 2).

- Application of information-theoretic measures of predictability to the probabilistic models (year 3).

5.4 Joint task by all three project partners

- Comparison of Balkan case study with other language contact situations (Dutch/German and possibly other European case studies) (year 3).

5.5 Cooperation with consulting researchers

In our development of this proposal, we have received feedback and encouragement from specialists in Language Contact and Dialectology, who have furthermore agreed to continue to share their expertise throughout the eventual duration of the project:

- **Prof. Brian D. Joseph**, full professor of Linguistics at the Ohio State University and editor of the journal *Language*, holds the prestigious Kenneth E. Naylor chair of South Slavic Linguistics. He is a world-renowned expert on Balkan Linguistics and has authored numerous publications on the topic, including a monograph on the effect of language contact on verb infinitives in the Balkan *Sprachbund* (Joseph, 1983). His work has also addressed more general theoretical issues in Language Contact and Change, e.g. Joseph (1990); Joseph and Hock (1996)

- **Prof. Hermann Niebaum** is full professor of Low Saxon at the University of Groningen. He is the author of a standard introduction to German dialectology (Niebaum and Macha *Einfürung in die Dialektologie des Deutschen*, 1999, 2002), and editor of the journal *Driemaandelijkse Bladen*. He was active in the campaigns to have Low Saxon and Low German (*Plattdeutsch*) recognized as regional languages under the European charter.

- **Dr. Charlotte Gooskens** is associate professor of Scandinavian languages at the University of Groningen. She has written extensively on Dutch and Scandinavian varieties and currently collaborates in the Scandinavian project studying degrees of communication and mutual intelligibility among the Scandinavian languages (*Linguistic and Extralinguistic Predictors for Inter-Nordic Communication).*

6 Proposed Budget

- 2005:

Personalkosten	Tübingen	BAT IIa/2	27 600 EUR
		stud. Hilfskraft	12 000 EUR
	Groningen	wiss. Hilfskraft	26 300 EUR
	Sofia	3 wiss. Ang. (400 EUR/Monat)	14 400 EUR
		3 stud. Hilfskräfte (100 EUR/Monat)	3 600 EUR
Reisemittel			17 600 EUR
einmalige Kosten	Sofia	PC (Webserver)	750 EUR
		scanner + OCR software	250 EUR
		PC	750 EUR
gesamt			103 250 EUR

- 2006:

Personalkosten	Tübingen	BAT IIa/2	27 600 EUR
		stud. Hilfskraft	12 000 EUR
	Groningen	wiss. Hilfskraft	26 300 EUR
	Sofia	3 wiss. Ang. (400 EUR/Monat)	14 400 EUR
		3 stud. Hilfskräfte (100 EUR/Monat)	3 600 EUR
Reisemittel			17 600 EUR
gesamt			101 500 EUR

- 2007:

Personalkosten	Tübingen	BAT IIa/2	27 600 EUR
		stud. Hilfskraft	12 000 EUR
	Groningen	wiss. Hilfskraft	26 300 EUR
	Sofia	3 wiss. Ang. (400 EUR/Monat)	14 400 EUR
		3 stud. Hilfskräfte (100 EUR/Monat)	3 600 EUR
Reisemittel			17 600 EUR
gesamt			101 500 EUR

Die Arbeitsplatzrechner für die Stellen in Groningen und Tübingen werden aus der Grundausstattung des jeweiligen Instituts zur Verfügung gestellt.

References

Asenova, Petya (1989), *Balkan Linguistics: Main Problems of the Sprachbund of Balkan Languages*, Science and Arts, Sofia, Bulgaria. (in Bulgarian).

Atlas of Bulgarian Dialects: Morphology (in preparation), number IV. In Bulgarian.

Blancquaert, E. and W. Peé, eds. (1925–1982), *Reeks Nederlands(ch)e Dialectatlassen*, De Sikkel, Antwerpen.

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer and P. Roosin (1990), 'A statistical approach to machine translation', *Computational Linguistics* **16**(2), 79–85.

Browne, Wales (2002), *What is a standard language good for, and who gets to have one?*, number 3 *in* 'The Kenneth E. Naylor Memorial Lecture Series in South Slavic Linguistics', Department of Slavic and East European Languages and Literatures, The Ohio State University, Columbus, Ohio (USA).

Chambers, J. K. and Peter Trudgill (1980), *Dialectology*, Cambridge textbooks in linguistics, Cambridge University Press.

Cheng, Chin-Chuang (1996), Quantifying dialect mutual intelligibility, *in* C.-T. J.Huang and Y.-H. A.Li, eds., 'New Horizons in Chinese Linguistics', Vol. 36 of *Studies in Natural Language and Linguistic Theory*, Kluwer Academic Publishers, Dordrecht, chapter 8, pp. 269–292.

Friedman, Victor A. (1993), Macedonian, *in* B.Comrie and G. C.Cobett, eds., 'The Slavonic Languages', Routledge, London, pp. 249–305.

Friedman, Victor A. (2003), *Linguistic emblems and emblematic langauges: on language as a flag in the Balkans*, number 1 *in* 'The Kenneth E. Naylor Memorial Lecture Series in South Slavic Linguistics', Department of Slavic and East European Languages and Literatures, The Ohio State University, Columbus, Ohio (USA).

Gilbers, Dicky, John Nerbonne and Jos Schaeken, eds. (2000), *Languages in Contact*, Vol. 28 of *Studies in Slavic and General Linguistics*, Rodopi, Amsterdam, Atlanta, GA.

Gooskens, Ch. and W. Heeringa (2004), The position of Frisian in the Germanic language area, *in* D.Gilbers, M.Schreuder and N.Knevel, eds., 'On the Boundaries of Phonology and Phonetics', Center for Linguistics and Cognition, Groningen, University of Groningen, pp. 61–87.

Haugen, Einar (1966), 'Semicommunication: The language gap in Scandinavia', *Sociological Inquiry* **36**(2), 280–297.

Heeringa, W. J. (2004), Measuring Dialect Pronunciation Differences using Levenshtein Distance, PhD thesis, Rijksuniversiteit Groningen, Groningen.

Heeringa, Wilbert and John Nerbonne (2001), 'Dialect areas and dialect continua', *Language Variation and Change* **13**, 375–400.

Herson Finn, Viktoria (1996), What is NAŠ: a Theory of Ethnolect in the South Slavic Dialect Continuum, PhD thesis, Ohio State University, Columbus, OH.

Hock, Hans Henrich (1991), *Principles of Historical Linguistics*, Mouton de Gruyter, Berlin.

Joseph, Brian D. (1983), *The Synchrony and Diachrony of the Balkan Infinitive: A Study in Areal, General, and Historical Linguistics*, Cambridge Studies in Linguistics, Supplementary Series, Cambridge University Press.

Joseph, Brian D. (1990), *Morphology and Universals in Syntactic Change: Evidence from Medieval and Modern Greek*, Outstanding Dissertations in Linguistics Series, Garland Publisher.

Joseph, Brian D. and Hans H. Hock (1996), *Language Change, Language History, and Language Relationship. An Introduction to Historical Linguistics*, Trends in Linguistics: Studies and Monographs, Mouton de Gruyter, Berlin.

Kochev, Iv., D. Vakarelska-Chobanska, T. Kostova, El. Kiaeva and M. Tetovska-Troeva, eds. (2001), *Atlas of Bulgarian Dialects: Phonetics. Intonation. Lexicology*, Vol. I - III, Trud Publishing House. In Bulgarian.

Kondrak, Grzegorz (2002), Algorithms for Language Reconstruction, PhD thesis, University of Toronto.

Krech, H. and U. Stötzer (1969), *Wörterbuch der deutschen Aussprache*, Max Hueber Verlag, München.

Labov, William (1972), *Sociolinguistic patterns*, Vol. 4 of *Conduct and Communication*, Univ. of Pennsylvania Press.

Levenshtein, V.I. (1965), 'Binary codes capable of correcting spurious insertions and deletions of ones', *Problems of Information Transmission* **1**(1), 8–17. Russian orig. in *Problemy Peredachi Informatsii* 1(1), 12–25, 1965.

Mirchev, K., ed. (1972), *Atlas of Bulgarian Dialects: Bulgarian Dialects from Aegean Macedonia*, Vol. I, Publishing House of the Bulgarian Academy of Sciences, Sofia, Bulgaria. In Bulgarian.

Mitchell, Tom M. (1997), *Machine Learning*, McGraw-Hill, New York.

Nerbonne, John and Christine Siedle (2003), 'Dialektklassifikation auf der Grundlage Aggregierter Ausspracheunterschiede'. Submitted for publication.

Nerbonne, John, Wilbert Heeringa and Peter Kleiweg (1999*a*), Edit distance and dialect proximity, *in* D.Sankoff and J.Kruskal, eds., 'Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed.', CSLI, Stanford, CA, pp. v–xv.

Nerbonne, John, Wilbert Heeringa and Peter Kleiweg (1999*b*), Edit distance and dialect proximity, *in* D.Sankoff and J.Kruskal, eds., 'Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed.', CSLI, Stanford, CA, pp. v–xv.

Rechnik na makedonskiot jazik (1961–1966). Skopje.

Schalkoff, Robert (1992), *Pattern Recognition: Statistical, Structural and Neural Approaches*, John Wiley, New York.

Schaller, Helmut Wilhelm (1975), *Die Balkansprachen: Eine Einführung in die Balkanphilologie*, Carl Winter Universitätsverlag, Heidelberg.

Shannon, Claude E. (1948), 'A mathematical theory of information', *Bell System Technical Journal* **27**, 379–423, 623–656.

Sproat, Richard and Michael Riley (1996), Compilation of weighted finite-state transducers from decision trees, *in* 'Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics', pp. 215–222.

Stoykov, St., ed. (1966), *Atlas of Bulgarian Dialects: North- Eastern Bulgaria*, Vol. II, Publishing House of the Bulgarian Academy of Sciences, Sofia, Bulgaria. In Bulgarian.

Stoykov, St., Iv. Kochev and M. Mladenov, eds. (1981), *Atlas of Bulgarian Dialects: North-Western Bulgaria*, Vol. IV, Publishing House of the Bulgarian Academy of Sciences, Sofia, Bulgaria. In Bulgarian.

Stoykov, St., K. Mirchev, Iv. Kochev and M. Mladenov, eds. (1975), *Atlas of Bulgarian Dialects: South-Western Bulgaria*, Vol. III, Publishing House of the Bulgarian Academy of Sciences, Sofia, Bulgaria. In Bulgarian.

Stoykov, St. and S. B. Bernshteyn, eds. (1964), *Atlas of Bulgarian Dialects: South-Eastern Bulgaria*, Vol. I, Publishing House of the Bulgarian Academy of Sciences, Sofia, Bulgaria. In Bulgarian.

Trubetzkoy, Nikolai Sergejewitsch (1930), Proposition 16: Über den Sprachbund, *in* 'Actes du premier congrès international de linguistes à La Haye du 10-15 avril 1928', Leiden, pp. 17–18.

Tsanov, B., ed. (in preparation), *Atlas of Bulgarian Dialects: Bulgarian dialects from Aegean Macedonia*, Vol. II. In Bulgarian.

Weinreich, Max (1945), 'Der Yivo un di problemen fun undzer tsayt ('Yivo and the problems of out time')', *Yivo Bleter* **25**, 1–13.

Wolfram, Walt and Natalie Schilling-Estes (1998), *American English. Dialects and Variation*, Blackwell, Oxford.

Erklärungen

- Angaben über die Vorlage des Antrages oder thematisch verwandter Anträge bei anderen Förderinstitutionen:

  Hiermit erklären alle Antragssteller, dass der vorliegende Antrag bei keiner anderen Förderinstitution eingereicht wurde und dass keine thematisch verwandten Anträge bestehen.

- Vorgesehener Bewilligungsempfänger:

  Eberhard Karls Universität Tübingen
  Seminar für Sprachwissenschaft, Abt. Computerlinguistik
  Wilhelmstr. 19
  72074 Tübingen, Germany