Variation in the Aggregate: An Alternative Perspective for Variationist Linguistics

John Nerbonne
Alfa-informatica, University of Groningen
j.nerbonne@rug.nl
P.O.Box 716, NL9700 AS Groningen, Netherlands
Tel. +31 50 363 58 15, FAX +31 50 363 68 55

June 6, 2007

Abstract We argue that an aggregate view of language variation is needed to supplement the usual characterizations based on single features. Aggregate characterizations are compatible with the well-known exceptions in the distributions of individual elements (features), they depend less on the theorist's choice of which features to use as the basis of a characterization, and finally, they enable the characterization of general tendencies in variation data, something which has escaped accounts of dialectology based on single features (even in combination).

1 Introduction

We summarize our arguments for employing aggregating techniques in this introductory section, and elaborate on them in the following several sections. Before launching into the arguments, we clarify what we see as the issue.

1.1 Feature-Based Variation Studies

Language variation is normally analyzed and presented in a bottom-up fashion, one element—i.e., FEATURE—at a time, whether the elements (or features) be sounds, words, morphemes, constructions, or whatever. This mode of presentation illustrates the issues concretely, surely a virtue. Linguists have for example studied the pronunciation of /r/ in the 1960's in New York City, the folk words for 'dragonfly' in Pennsylvania, the realization of the German diminutive suffix along the Rhine, and the orders of auxiliary verbs in continental West Germanic. We refer to such characterizations based on individual features as (SINGLE)-FEATURE BASED, and we emphasize that we use the term 'feature' not only to refer to phonological features such as [+ROUND] or morphosyntactic features such as [+PLURAL], but more broadly. There are countless single-feature-based studies; an immense amount is known about the geographic and social distribution of individual linguistic features; and this information has often been organized into fascinating dialect atlases.

The issue we pose does not concern the TRUTH of statements about single-feature distribuitions. Such statements are, of course, subject to the same risks of being based on flawed observations or poor analytical technique, but the aggregating techniques we champion are also not immune to these. Rather, we argue that aggregating techniques enable answers to the fundamental questions of variationist linguistics, and in particular, dialectology in ways that are simple inaccessbile to single-feature studies. Having said that, we ought to at least mention what see as the fundamental questions of variationist linguistics.

Language users speak primarily in order to communicate, but they employ overlays of variation in form in order to signal geographic and social provenance. By using syllable-final [r] in New York City in the 1960's, speakers signaled their membership in the middle classes; by calling a dragonfly a *darning needle*, Pennsylvanians signaled their northern provenance in the mid-twentieth century; and by using auxiliary verbs before main verbs in subordinate clauses, German speakers identify themselves as Swiss.

1.2 Issue

By focusing exclusively on single features or small combinations of these, variationists, including dialectologists, fail to isolate signals of provenance clearly. The signals are so complex, even misleading that they resist analysis using simple, single-featured methodologies.

Please note that we have not accused dialectology of not identifying important signals of provenance. We rather formulate the charge that single-feature based dialectology fails to isolate these signals analytically and that aggregation is the key methodological step needed to enable analytical progress. Kretzschmar (2006, 400) notes that one prominent dialectologist he collaborated with extensively "[...] could afford to ignore interpretation of the data because he already knew what it meant," and the same is true of many other excellent researchers. They are so familiar with the data and how its signals of provenance are received that they identify these correctly without much difficulty. But by relying on informed intuition rather than analytic technique, they provide no foundation for more abstract questions. We return to this in Section 4 below.

Naturally, the field has been aware of this difficulty, and there are numerous discussions of how single-feature studies are related to more general issues such as the character of DIALECT AREAS. These discussions falter universally on the complexity of distributions of single features, which inevitably have exceptions, and normally contradict each other, at least in detail. Section 2 reviews this discusion, using a recent German dialectology database as illustration.

1.3 Structure

The feature-based approach is thus concrete and in many respects successful. We argue nonetheless that it is unsatisfactory in some respects. The reason which has played the largest role in theoretical discussion is the problem of generalizing from single-feature distributions to characterizations of social or geographical varieties. The single features inevitably contradict each other, at least in detail, and being linguistic features, they tend to have exceptions and sparse distributions. We wish to review these arguments in Section 2 below, since in our experience linguists are not in general aware of how large and genuine the problem is.

A second reason for dissatisfaction with not proceeding beyond feature-based characterizations is methodological. Languages are large and complex, and there are easily tens of thousands, probably hundreds of thousands or more ways for language varieties to differ. If dialectological or variationist theory says only that some linguistic feature distinguishes areas (or social groups), then that theory is wildly underdetermined—it has hundreds of thousands of features to choose from. If combinations of features are appealed to, the range of possibilities rises enormously. We develop this argument in Section 3.1 below. We shall contrast this with a view which aggregates over all available features, which views aggregate differences as characterizing the relations among varieties.

A third reason to be dissatisfied with the single-feature approach is also theoretical, like the second, but appeals to theoretical ambition rather than to methodological scruple. We should like to characterize linguistic variation in more general terms, e.g., characterizing not only which English varieties do or do not pronounce syllable-final /r/, but e.g. also how linguistic variety in general is influenced by geographic distance. In order to do this, we need to move to a more abstract level of characterization, and we argue in Section 4 that the view of variation from an aggregated perspective enables the formulation of more general laws.

In this essay we argue for an alternative approach to the study of variation, in particular that it be studied first and foremost in the aggregate. Our approach is indebted to the DIALECTOMETRY of Séguy (1973) and Goebl (1984), and we are indeed pleased to regard it as dialectometry, but we focus here neither on measurement nor on principles of classification, the common focus in dialectometry, but rather on the aggregating step which both Séguy and Goebl use to great advantage.

2 Need for Abstraction

We noted above that variationist studies which focus on single variables face various impediments when they wish to show that they have characterized the speech of a region or that of a social group. One stumbling block is brutally empirical, and obvious to every student of linguistic variation who has ever inspected a linguistic atlas more than briefly. The geographic distributions of individual linguistic elements—be they phonological features, lexicalizations, allophones, or case restrictions—are never smooth, but rather always fraught with exception. This is the source of the complaint echoed by Bloomfield (1933), that "every word has its own history" (p.328). Variationist linguistics has advanced a great deal since Bloomfield, but still remains focused on individual liguistic features.

It will be instructive to examine some material from a dialect atlas in order to drive home the point that dialect data is fraught with exceptions, and it will allow us to make a minor point in this section in favor of abstraction as well. In what follows we use material from the *Phonetischer Atlas Deutschlands* (PAD), material collected between 1965 and 1991 by Marburg fieldworkers under the supervision of Prof. Joachim Göschel. 201 words from the famous *Wenkersätze* were recorded in 186 sites throughout Germany (Göschel 1992). The pronunciations in these recordings were subsequently transcribed by a team of professional phoneticians, including Prof. Angelika Braun of Marburg. They used a methodology in which two phoneticians transcribed each pronunciation independently, and later compared results to obtain consensus transcriptions. Researchers from the University of Groningen digitized the handwritten IPA material in X-SAMPA in 2003 (Nerbonne & Siedle 2005). The material exclusively concerns pronunciation, but we maintain that other linguistic levels will show similar patterns vis-à-vis exception.

We have a second reason for wishing to review this material, namely, to drive home that point that dialectology already makes use of a number of aggregating steps. In doing this we wish to sharpen the debate about the need for aggregation: in general, dialectology and other variationist studies accept many aggregating steps. The issue is thus not whether to aggregate rather on what scale.

2.1 Single Segments

Before examining the geographic distributions of linguistic features, we note that the PAD is similar to most linguistic atlases in being recorded in almost daunting phonetic detail. One of the simplest words in the atlas is *ich*, (/[ɪx/, [ɪç] in standard German). The final consonant is pronounced [ç] in standard German, and normally analyzed as palatal allophone of the velar fricative /x/, so that we sometimes refer to the stop/fricative distinction as

Table 1: 87 pronunciations of *ich* at the 201 different collection sites of the PAD. Twelve transcriptions are omitted since they seemed to violate IPA specifications, almost all involving what appeared to be the trailing diacritics [-] and [+], presumably denoting retraction and advancement. [ɪç] was recorded 17 times, [ɪk] 13 times, and [ɪ] nine times, but no other pronunciation was recorded more than five times.

ɨ	ɐɪç	ɐɪç̰	ɐ̱ɪç	ʕɪ̜k	ʕɪk	əɪʃ	ə̰͡ɪg	ç	ɛɪ̜ʃ̱	ɛç̰k	ɛ̜g	ɛ̣ɪç̟	ɛ̣ɪʃ̱
ɛ̣ɪk	ɛk	ɛk̃^{h}	ɪ	ɪː	ɪʔ	ɪç	ɪç̟	ɪç̱	ɪɣ	ɪɣ̟	ɪʃ	ɪʃ̟	ɪʃ̱
ɪ̱ç	ɪ̱ç̱	ɪ̱ɣ	ɪ̱g	ɪ̱k	ɪ̱k.	ɪ̱ɕ	ɪ̱ʑ	ɪ̥k	ɪ̜ç	ɪ̜g	ɪ̜g.	ɪ̜j	ɪ̜k
ɪ̜ɕ	ɪ̜x	ɪ̣	ɪ̣ç̱	ɪ̣ː	ɪ̣ːç	ɪ̣ç	ɪ̣χ	ɪ̣g	ɪ̣g.	ɪ̣k	ɪ̣ɕ	ɪ̣ʑ	ɪg
ɪj	ɪj̰	ɪk	ɪk^{h}	ɪɕ	ɪx̟	ʏç̱	ʏʑ	e	e̱͡ɪɣ	e̜ʔ͡k	e̜ç	e̜g	ẹʃ̼
ẹ͡çj	ẹç	ẹɣ	ẹg	ẹj	ẹɕ	eg	ek	ek͡x̰	i	iː	iːç	iːç̰	iç
i̜	i̜ː͡jç	i̜k											

a ‘k/x’ distinction, when perhaps we should always note that the /x/ may be realized as [x(ç)], i.e., as [x] or as [ç]. We find eighty-seven different phonetic transcriptions for this word at the 201 different data collection sites, which we present in Table 1. We note that there are 28 different renderings of the final consonant and 29 different renderings of the vowel. A small number of the transcriptions are distinguished only in that one records a syllable break following the consonant while the other does not, and we do not suppose that this distinction is dialectologically relevant. But eliminating these would not change the overall situation significantly: phonetic atlases contain so great a variety of material that the analyst is forced to categorize to make any sense of the material.

It is worth emphasizing that the example of *ich* is not exceptional. For example, 34 different vowels were recorded in the word *Eis*, even if only three different consonants were recorded. This sort of variation is frequent in dialect atlases. For an example from another data collection, the publicly available LAMSAS dataset contains over 1.100 different vowels at 450 sites (`http://hyde.park.uga.edu/lamsas/`).

Let us characterize the variation in the final consonant of *ich* in the standard way, as a difference between stops and fricatives. Although this is the form normally presented in textbooks on linguistics or on dialectology, a nontrivial step is needed to categorize the approximately 28 variants of the variable found in just this one word. [k, g, c, k^{h}, k^{j}, and g^{j}] and [g̥] are clearly stops, and [x, ç, ɣ, ʝ, χ, ʁ, ɕ] and [ʑ] are clearly fricative and plausible results of frication applied to [k], but there remain fricative allophones which are not straightforward frications of the velar stop ([ʃ, ʃ̱] etc.), the nonfricative

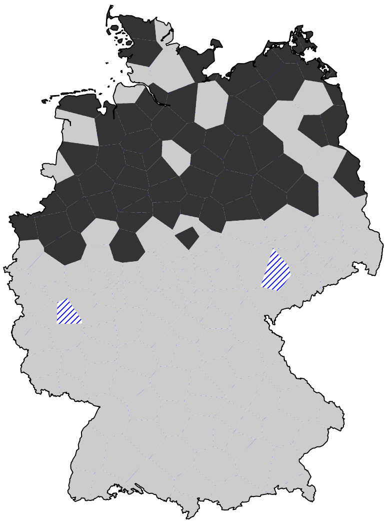


Figure 1: The [k/x(ç)] distinction in the word *ich* in the PAD. The darker the polygon, the greater the concentration of stop variants. There was no data available for the polygons with diagonal lines.

approximant [j], and, finally, cases where no final consonant is realized. Since the problematic cases are in some sense interpretable as LENITIONS of the velar stop, we in fact opt to class all of these with the clear cases of fricatives.

Figure 1 shows the relative concentrations of stop versus fricative variants in the pronunciation of *ich*. We obtained this by first dividing the map of Germany into polygons surrounding collection sites, and then coloring each polygon darker in proportion with the stop variants of the final consonant in *ich*. Once we aggregate over the many subvariants of stops and fricatives, a relatively clear pattern emerges, but one with prominent pockets of exceptions. This is the normal result.

The degree of phonetic detail in Table 1, and that in most dialect atlas collections,[1] suggests that we shall always need to move from low-level characterizations to more abstract levels. This move to a higher level of abstraction involves classifying the different recordings along one or more parameters. And this is what dialectologists have in fact always done with this sort of data, for example, focusing on two sets of variants. While one may always explore alternative abstractions (classifications), it is clear that the step to a more abstract view of the data promises to liberate the analysis

[1] One may ask whether the practice of atlas compilers to transcribe in such narrow detail is sensible. On the other hand Ton Goeman measured the consistency of the two main transcribers for the recent, very large ($> 10^6$ word/phrase transcriptions) Goeman-Taledeman-van Reenen project (GTRP) at $r \approx 0.95$ for consonants, $r \approx 0.9$ for vowels, and $r \approx 0.8$ for diacritics (Goeman 1999, Ch. 3). Perhaps the atlasses are faithful renderings of speech, which, however contains a great deal of subdialectal as well as dialectal variation.

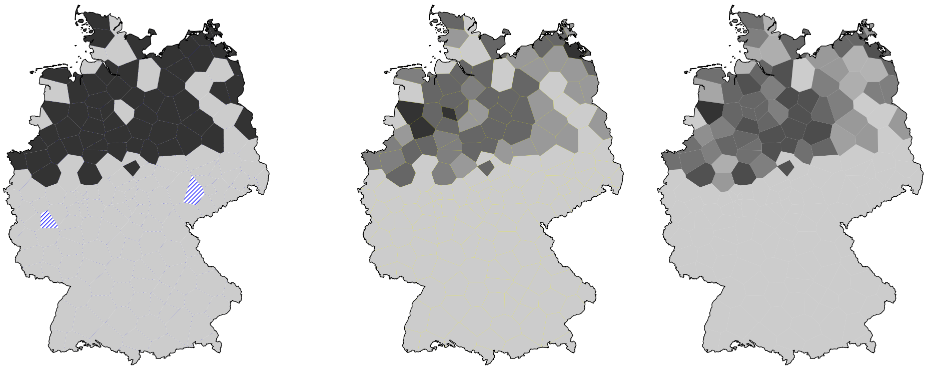
to allow more insightful analyses. But let's note that the classification step is effectively a step in aggregation: many observations are grouped into a single class. With an eye toward future aggregating steps we note that this step is always taken with respect to a single paradigmatic dimension. Thus it involves aggregating among the pronunciations of the final consonant in *ich* or the initial vowel in *Eis*, but it does not require aggregating across such categories.

We sum up this section by noting that good dialectological practice has always aggregated in one fashion, abstracting from dauntlingly detailed recordings to more abstract renderings of selected differences. Please note that we do note argue tendentiously that this aggregating step justifies all others, only that dialectology has made extensive use of this sort of aggregation in any case. This does not mean that *all* aggregation is sound, but it certainly does mean that *some* aggregation is standard practice. Far from criticizing this practice, we argue below that good dialectological analysis needs to adopt techniques of aggregation more extensively.

2.2 Multiple Occurrences of Variables

In fact, a second step of aggregation is likewise common, that of aggregating over the occurrences of variables in different words. To continue using the example we began in Section 2 above, we need to collect various words in which the variable occurs and aggregating over the variants used in their pronunciations. In our data collection, this includes the words *ich* /ɪç/, *dich* /dɪç/, *auch* /aux/, and *gleich* /glaɪç/. (In fact we likewise have *schlechte* /ʃlɛçtə/ and *schlechten* /ʃlɛçtən/ in the dataset, but these are never pronounced with a /k/, so they are not used in the present example.) The increase in scope complicates the set of variants in that we now find not only the palatovelar stops and fricatives noted above in Table 1, but also the rhotics [r, ʀ, ʀ̆] and the voiced alveolar fricative [ʒ]. We have again opted to classify these with the fricatives because they might be understood as lenitions.

It is useful to compare increasingly inclusive patterns of variation, and this is presented in Fig. 2. The leftmost map is identical to the map in Fig. 1 and is based on the final consonant in the single word *ich*. The middle map includes the variation in the final consonant of a second word, *gleich*, and the addition of this word immediately smooths the distribution a bit, for example, filling in sites where data was missing. The third and rightmost map is based on all five words in which we find variation. The rightmost map shows clearly that the lenis variants completely dominate the south, but also that the north is quite variable. The darkest areas have high concentrations of plosive variants ([k], etc.), and the lighter ones are mixed. Ideally, we would extend such a series to include as many words as possible, benefiting from the statistical stability of large data sets. We contend that



ich [k] (dark) vs. [ç/x] (light) *ich, gleich* all five occurences

Figure 2: The stop/affricate for the variation [x(ç)/k] ([ç] in High German. Occurrences of the variable increase from left to right, yielding more regular depictions of the distribution.

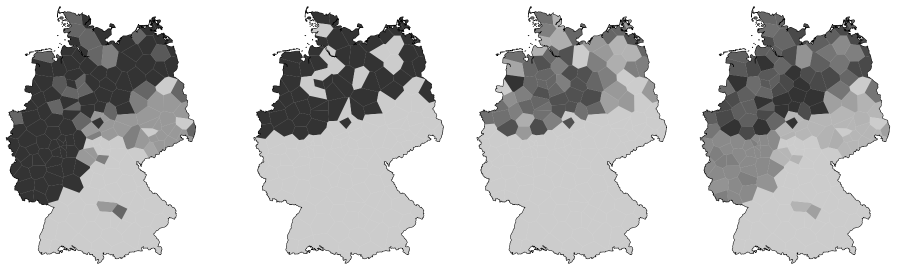
such maximally comprehensive maps will best predict the pronunciation of other words with this variable.

To return to our main argument, note that none of this makes sense without a second sort of aggregation, namely the sort which classifies the variants not only of a single segment of a single word, but also the sort which classifies variants of a single variable as it occurs in multiple words. This sort of aggregation, too, is common throughout dialectology.

2.3 Phonological Features

In the search for more robust generalizations, one may look to increasingly abstract characterizations, e.g. the well-known characterization of variation involving a single phonological feature, such as the famous "second sound shift" in German, the distinction between [p/p͡f, t/t͡s] (where we shall include [s] as a variant of the affricate [t͡s] and [k/x(ç)]. These are all instances of [stop/affricate], and it is striking that such a simple linguistic distinction characterizes German dialect areas as reliably as it does. Figure 3 compares the distribution of these three distinctions.

Indeed the commonality is striking, so that the characterization of dialect areas which aggregates over these three variations is quite good. Even if we include words such as *zwei* 'two' and *zwölf* 'twelve' which varied in the past, but for which the southern variant dominates to the complete exclusion of the expected variant in [t], we obtain a fairly clear delineation.



[p] (dark) vs. [p͡f] [t] vs. [t͡s] [k] vs. [x(ç)] aggregate view

Figure 3: The stop/affricate distinction resulting from the second sound shift. Although the patterns are similar, they certainly do not overlap perfectly. The simple aggregation on the right depicts the degree of overall differences in German varieties more faithfully.

3 Choice of "Features"

But there is a great deal more systematic geographic variation which further aggregating steps may incorporate. We extract a number of features from the PAD and sketch their geographic distribution in Figure 4. We select well-discussed features of German dialectology (König 1994, Niebaum & Macha 2006), including the characterization of the stop/affricate series examined above (for reference). In addition, we include maps sketching the distribution of the following:

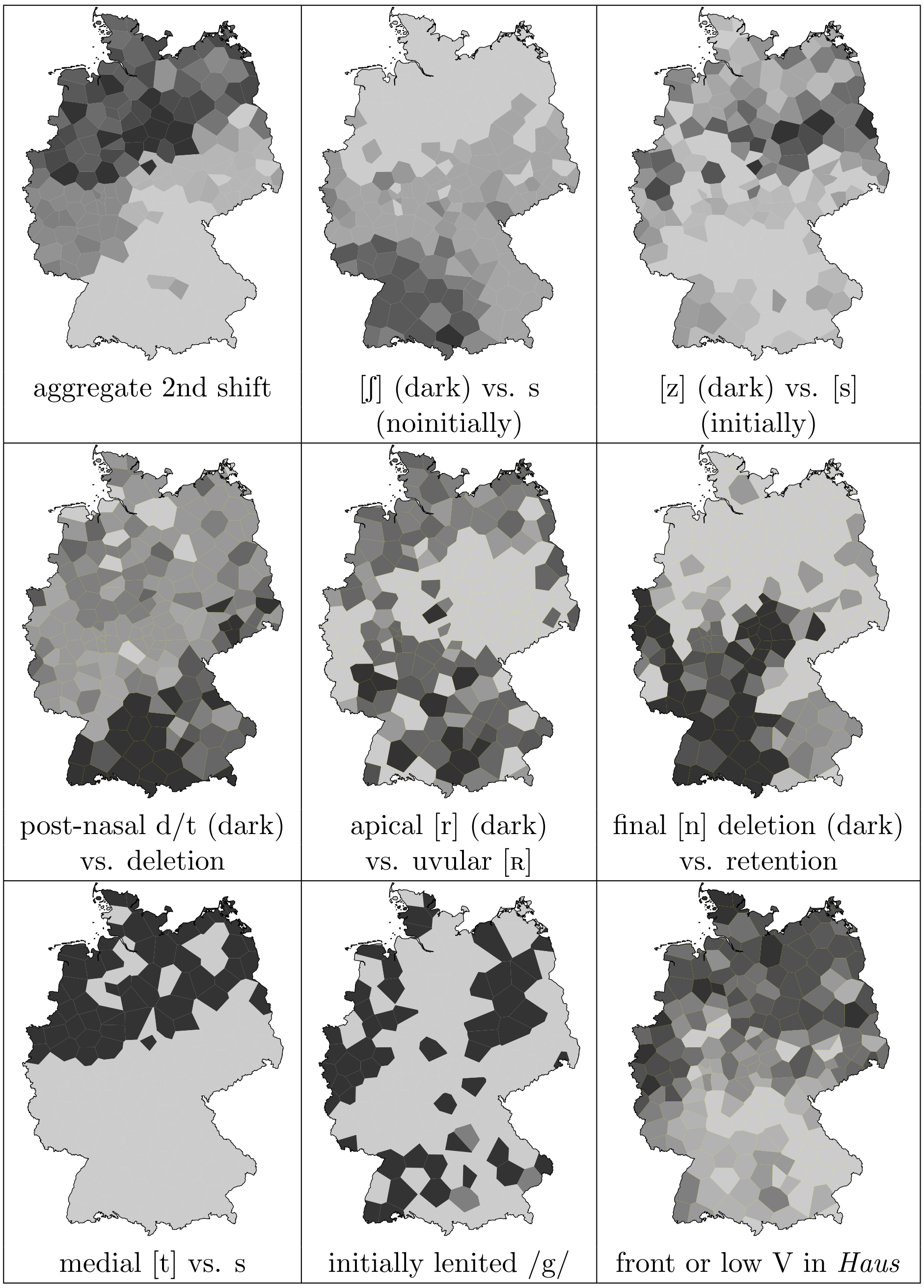
palatalization of non-initial /s/ in words such as *Wurst* 'sausage', *fest* 'firm', *gestern* 'yesterday', *ist* 'is' and *selbst* 'self'. Top row, middle in Fig. 4.

s/z word initially in words such as *Sonntag* 'Sunday', *selbst* 'self', *Seife* 'soap', *sie* 'she', *sieben* 'seven', *so* 'so' and *sollen* 'should'. Top row, right column.

t,d → ∅ / n _____ /t/ and /d/ are not always pronounced after /n/; thus we find many pronunciations of *unten* 'underneath', *anderen* 'others'and *gefunden* 'found (part.)' with no traces of a medial alveolar stop. The same phonological environment is present in *Winter* 'winter', but the t/d is only rarely suppressed when *Winter* is pronounced. See middle row, left column in Fig. 4.

apical vs. dorsal pronunciations of /r/ i.e., [r,ɾ] vs. [ʀ] in words such as *Brot* 'bread', *Bruder* 'brother', *Ohren* 'ears', or *wäre* 'be (subjunctive)'. Middle row, middle column.

retention/deletion of final nasal (in unstressed syllables in [ən]) in words such as *machen* 'make', *treiben* 'drive', *trinken* 'drink', *wachsen* 'grow', and *werden* 'become'. Middle row, left column.



aggregate 2nd shift

[ʃ] (dark) vs. s (noinitially)

[z] (dark) vs. [s] (initially)

post-nasal d/t (dark) vs. deletion

apical [r] (dark) vs. uvular [ʀ]

final [n] deletion (dark) vs. retention

medial [t] vs. s

initially lenited /g/

front or low V in *Haus*

Figure 4: The distribution of a range of pronunciation features, clearly overlapping only imperfectly. See text for further explanation.

lenition of medial t i.e., [t] vs. [s] in *Wasser* 'water'. This is part of the second shift (top left), but note how fragmented the distribution is. Bottom row, left column in Fig. 4.

g → ɣ,j / # _____ /g/ is often lenited to a fricative [x,ɣ] or even to an approximant [j] in participles such as *geschlafen* 'sleep (part.)' but also in *gut* 'good'. Bottom row, middle column.

vowel in *Haus* 'house' Vowels occur in so many varieties that simple characterizations are perhaps always misleading. We encountered 322 different vowels (different combinations of base segment and diacritics) in the six words *Haus* 'house', *braune* 'brown', *verkaufen* 'sell', *auch* 'also', *Frau* 'woman' and *auf* 'on'. We divided these into vowels with mid to high back onsets, such as [u, ɯ, ʊ, o, ɤ, ɔ] or [ʌ] and those with front or low onsets, such as [ɑ, ɒ, a, ə, æ, ɛ, œ, ø, ɪ, y] and [ʏ]. We admit immediately that other divisions here are as plausible, but also this division is reflected geographically. Bottom row, right column.

Fig. 4 is important for several reasons. First, it illustrates that individual features are often at odds with one another in detail, making any one of them unsuitable as a sole defining element in linguist geography. We need to find a way to aggregate over many features if we wish reliably to detect the relations different varieties have to one another. Second, Fig. 4 illustrates that, in spite of the conflicts in detail, the member of a dialect community has many, often redundant signals as to the geographic provenance of a dialect speaker. Dozens of words in our small sample alone indicate roughly whether a speaker is from southern or northern Germany.

We assume that dialect speakers are sensitive enough to linguistic variation to be able to detect a large number of signals and that they are intelligent enough to combine these—albeit subconsiously. Heeringa (2004) provides an overview of the analytical and aggregating techniques that are needed to combine the information in the many variables present in an atlas such as the PAD, and Nerbonne & Kretzschmar (2006) provide references to more recent developments. It would take us too far afield to present all of the material in this essay, but Fig. 5 presents a defensible analysis.

It is important to add, however, that there are many aggregating techniques, and that there are debates about the best techniques of analyis (Nerbonne & Kretzschmar 2006). The purpose of the present essay is to defend the need for aggregation, but neither to present aggregating techniques in detail, nor to take a stand on which are best.

3.1 Keeping it Simple

We charge dialectology based on single features with discriminating too little in the features it chooses. Given the current state of the art, in which re-

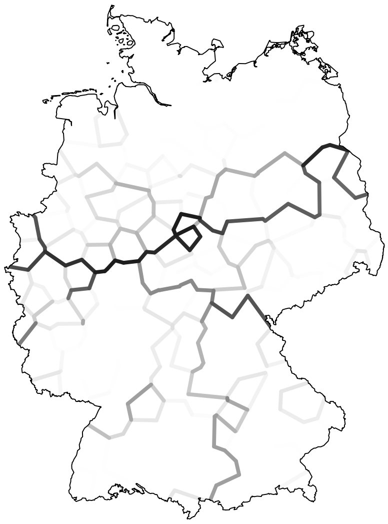


Figure 5: The aggregate view of pronunciation variation in the PAD. Darker lines correspond to better founded distinctions.

searchers choose arbitrarily among linguistic features that are hypothesized to be associated with extralinguistic variables, very little is shown when *some* variable or other can be shown to associate strongly with some extralinguistic property.

If this seems exaggerated, consider that there are 20 to 100 phonemes in a typical variety, each of which typically has five to ten allophones, depending on the level of detail one is willing to examine. The distribution of allophones is governed by 20 or more phonological processes. Varieties may differ in their phoneme inventories, their range of allophones, and in the rules governing the distribution of the allophones, and of course, in combinations of these. Nor is the situation simpler at other linguistic levels: Miller estimates that adults have vocabularies of approximately 50,000 lexemes (Miller 1991). Even in morphologically poor languages such as English, these lexemes are subject to modification by 100 or more bound morphemes, some of which have effects in combination which may be peculiar to certain varieties. Large syntactic descriptions typically contain hundreds of phrasal rules, and theorists increasingly concede that a great deal of syntactic structure requires even more specific licensing of constructions (Fillmore & Kay 1999), i.e. phrasal patterns with often idiosyncratic restrictions to specific (combinations of) lexical items. The number of possible hypotheses is multiplied again by the incorporation of frequency information in analyses. In frequency-based accounts, we do not need to demonstrate that the presence or absence of a feature is associated perfectly with the extralinguistic variable, we may instead appeal to the frequency with which the feature occurs.

Aggregating accounts postulate that extralinguistic variables are associated with aggregate differences in entire varieties, not merely with specific linguistic variables. We are motivated to postulate this in part in considering the cognitive problem of detecting the signals of provenance. If signals were not robust, i.e. likely to be present and detectable in many speech events, then they simply would not function. Thus given the information in the aggregate views in Fig. 3 and Fig. 5, we hypothesize that even speakers within pockets of exception appearing in single feature distributions such as Fig. 1 will provide signals of provenance.

4 General Characterizations

If there are larger, simpler trends present in linguistic variation, then single-feature based approaches seem ill-equipped to search for them. This is due both to their variety and to the fact that there are exceptions, but especially because the notion "linguistic variety"—the collection of speech patterns used in a community—plays no role in analyses. Aggregate analyses proceed, on the contrary, by characterizing the relations between varieties.

As an example of a general theoretical question in dialectology which aggregate studies seem poised to answer, consider the relation of geography to linguistic variation. In particular Peter Trudgill has been at pains to point out that dialectology should strive toward more general accounts of how variation is distributed geographically (Trudgill 1974), but single-feature studies have a decided mixed record in this regard (Bailey, Wikle, Tillery & Sand 1993, Wikle & Bailey 1997, Boberg 2000, Horvath & Horvath 2001)—most of the literature consists only of criticisms of Trudgill's "gravity hypothesis". We suggest that the problem is the level of analysis.

Aggregating analyses such as Séguy (1971) have long noted that there is a simple, lawlike relation between geographic distance and linguistic variation of exactly the sort Trudgill sought, and about which the other authors cited have been sceptical (arguing that the relation is more complex than Trudgill postulated). In general linguistic variation increases as a sublinear function of geographic distance, as Fig. 6 illustrates.

Aggregating views likewise offer new perspectives on the validation of claims in variationist linguistics (Gooskens & Heeringa 2004). If we are claiming to characterize signals of provenance, then we should be prepared to test whether the signals are genuine and effective (i.e., whether they are perceived as such), and Gooskens and Heeringa validate some aggregating techniques using perceptual data. We are aware of the work in psychoacoustics with similar aims (Clopper, Levi & Pisoni 2006), but aggregating techniques can contribute as well to validation.

We further predict that aggregating techniques may serve as the basis for new approaches to classic and important issues in the theory of language variation. For example, dialectological handbooks agree that linguistic vari-

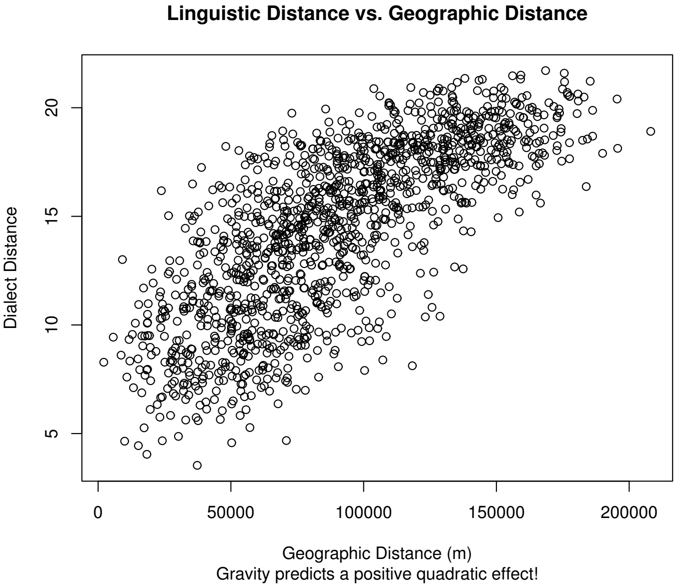


Figure 6: Linguistics variation is a sublinear function of geographic distance just as Séguy (1971) demonstrated, but this characterization requires an aggregating step. The curve above is based on Dutch data from Nerbonne & Heeringa (2007), essentially replicating Séguy's analysis on a novel data set.

ation serves to identify some speakers in contrast to others (Chambers & Trudgill 1998, [^{1}1980]), noting that variation serves simultaneously as a mark of solidarity with some, and a mark of differentiation with respect to others. Single-feature studies seem unable to evaluate the relative strengths of these dynamics, because there is no characteristic of the aggregate relation between varieties. Aggregating techniques should be brought to bear.

Other questions include the nature of the geographic influence. Are areas (or regions) the organizing elements in dialectology, leading us to expect a partition of sites, or should one rather analyze variationist data in terms of continua? Are human families important mediating factors in transmitting variation (Manni, Heeringa & Nerbonne 2006)? These sorts of questions again require general characterizations of variation, which single-feature studies have not produced.

As a final example, we note that we criticized approaches based on single features above (Section 3.1). But there is a good deal of linguistic interest on the nature of the variation which serves to distinguish geographic and social groups. The inclusion of an aggregating perspective should not mean that such questions are ill posed, but rather that they may be posed against the backdrop of the aggregated analysis. In this view one should ask which linguistic variables are responsible for the aggregate distinctions we encounter.

5 Previous Aggregate Views

There has certainly been a good deal of attention paid to aggregating approaches, most notably by Goebl (1984). In fact the techniques are also found in the handbooks (Chambers & Trudgill 1998, [^{1}1980], § 9.4.1). But Goebl's focus is generally on the the taxonomic methodology he has developed and applied so extensively, and Chambers and Trudgill focus on the issue of "quantifying linguistic variables", rather than on the opportunity for aggregation, which is at the heart of the benefits of dialectometry. There does not seem to be a single work attempting to defend the need for aggregation in a focused way, which has been our goal here.

5.1 Bundling Isoglosses

Haag (1898) (discussed by Schiltz (1996)) proposed a quantitative technique in which the darkness of a border between two adjacent sites was reflected by the number of differences counted in a given sample, and similar maps have been in use since. This appears to be the first published proposal of how one operationalize the idea of "bundling isoglosses", and it clearly implies aggregating over a variety of features, so it is an important early recognition of the need for aggregation.

Since for many dialectologists, the search for isogloss bundles is the final methodological wisdom in seeking geographic determinants of variation (dialect borders and areas), let us emphasize that our plea here is more general in several ways. First, we should not restrict the application of aggregation to situations in which clear borders exist. Aggregation is a very useful step in characterizing dialect continua as well (Heeringa & Nerbonne 2001). Second, we should prefer to emphasize that there are many approaches to operationalizing the sort of aggregation we have in mind so that we need not rely on counting isoglosses. For example, we have quantified the occurrence of contrasting elements in the maps above, and we have developed numerical characterizations of pronunciation difference which lend themselves well to aggregation (Nerbonne, Heeringa & Kleiweg 1999, Heeringa 2004). Third, and finally, there are technical advances in pattern recognition and classification which enable us to seek borders even in cases where the *local* differences between two sites do not suggest them (Nerbonne, Kleiweg & Manni 2007). For example, clustering can take non-local differences into account and thus detect borders even where differences are gradual.

5.2 Martinet and Labov

We have not discussed Martinet's and Labov's work on the complicated chains of vowel shifts which often occur in series (Labov 1994), a body of work with the admirable ambition of seeking very general laws. Like the approach we argue for here, it attempts to seek characterizations at a higher level of aggregation. But the focus of Martinet's and Labov's work is historical, whereas ours is synchronic. It should also be clear that we have a much less structured notion of aggregation in mind in this essay.

6 Conspectus and Prospectus

The essential aggregating step is common only up to a certain degree in variationist linguistics, but we have argued here that its more general application solves important analytical problems. The key problem is the problem of extracting a reliable signal of provenance from variationist data. Single-feature studies risk being overwhelmed by noise, i.e., missing data, exceptions, and conflicting tendencies, which are common in this and most areas of linguistics. We aggregate in order to obtain a clearer signal.

We repeat here the qualification that "aggregation" is a very general term which needs to be operationalized carefully. We have not attempted in this essay to identify features that are particularly suitable, nor to address technical issues such as weighting data, how much data is needed or which techniques are most suitable for analysis. We refer interested readers to Heeringa (2004) for examples of this sort of work.

Not only does aggregation enable an answer to the problem of rebarbative data, but it also enables us as dialectologists to reduce the hypothesis space within which associations between linguistic and extralinguistic variables must be found. While existing practice seems to allow any single variable to serve as the putative linguistic base of an extralinguistic association, we have postulated that linguistic signals of provenance should be detected and analyzed in the aggregate, reducing, we hope significantly, the number of potential hypotheses.

Finally, we claim that aggregate analyses provide a level at which very general laws concerning linguistic variation might be formulated. This section was quite programmatic, but dialectology is in sore need of more general theoretical work, and aggregating analyses are promising.

Acknowledgments

We thank Peter Kleiweg for programming and graphics (`http://www.let.rug.nl/kleiweg/L04/`), Hermann Niebaum for suggestions for improvements, and Tette Hofstra for insisting that difficult questions be asked.

References

Bailey, Guy, Tom Wikle, Jan Tillery & Lori Sand. 1993. "Some Patterns of Linguistic Diffusion." *Language Variation and Change* 3(3):241–264.

Bloomfield, Leonard. 1933. *Language.* New York: Holt, Rhinehart and Winston.

Boberg, Charles. 2000. "Geolinguistic Diffusion and the U.S.-Canada Border." *Language Variation and Change* 12(1):1–24.

Chambers, J.K. & Peter Trudgill. 1998, [^{1}1980]. *Dialectology.* Cambridge: Cambridge University Press.

Clopper, Cynthia, Susannah V. Levi & David Pisoni. 2006. "Perceptual similarity of regional varieties of American English." *Journal of the Acoustical Society of America* 119:566–574.

Fillmore, Charles & Paul Kay. 1999. "Grammatical Constructions and Linguistic Generalizations: the *What's X Doing Y?* Construction." *Language* 75(1):1–33.

Goebl, Hans. 1984. *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF. 3 Vol.* Tübingen: Max Niemeyer.

Goeman, Antonie. 1999. T-deletie in Nederlandse dialecten PhD thesis University of Amsterdam.

Gooskens, Charlotte & Wilbert Heeringa. 2004. “Perceptual Evaluation of Levenshtein Dialect Distance Measurements using Norwegian Dialect Data.” *Language Variation and Change* 16(3):189–207.

Göschel, Joachim. 1992. Das Forschungsinstitut für Deutsche Sprache “Deutscher Sprachatlas”. Wissenschaftlicher Bericht Das Forschungsinstitut für Deutsche Sprache Marburg: .

Haag, Karl. 1898. *Die Mundarten des oberen Neckar- und Donaulandes (schwäbisch-alemannisches Grenzgebiet: Baarmundarten).* Reutlingen: Buchdruckerei Egon Hutzler.

Heeringa, Wilbert. 2004. Measuring Dialect Pronunciation Differences using Levenshtein Distance PhD thesis Rijksuniversiteit Groningen.

Heeringa, Wilbert & John Nerbonne. 2001. “Dialect Areas and Dialect Continua.” *Language Variation and Change* 13(3):375–400.

Horvath, Barbara M. & Ronald J. Horvath. 2001. “A Multilocality Study of a Sound Change in Progress: The Case of /l/ Vocalization in New Zealand and Australian English.” *Language Variation and Change* 13(1):37–57.

König, Werner. 1994. *DTV Atlas zur deutschen Sprache.* München: Deutscher Taschenbuch Verlag. ^{1}1978.

Kretzschmar, William A. 2006. “Art and Science in Computational Dialectology.” *Literary and Linguistic Computing* 21(4):399–410. Special Issue, J.Nerbonne & W.Kretzschmar (eds.), *Progress in Dialectometry: Toward Explanation.*

Labov, William. 1994. *Principles of Linguistic Change. Vol. 1: Internal Factors.* Oxford: Blackwell.

Manni, Franz, Wilbert Heeringa & John Nerbonne. 2006. “Are Family Names just Words? Comparing Geographic Patterns of Surnames and Dialect Variation in the Netherlands.” *Literary and Linguistic Computing* 21(4):507–528. Special Issue, J.Nerbonne & W.Kretzschmar (eds.), *Progress in Dialectometry: Toward Explanation.*

Miller, George. 1991. *The Science of Words.* New York: Scientific American Library.

Nerbonne, John & Christine Siedle. 2005. “Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede.” *Zeitschrift für Dialektologie und Linguistik* 72(2):129–147.

Nerbonne, John, Peter Kleiweg & Franz Manni. 2007. Projecting Dialect Differences to Geography: Bootstrapping Clustering vs. Clustering with Noise. In *Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classification Society*, ed. Lars Schmidt-Thieme, Hans Burkhardt & Reinhold Decker. Berlin: Springer. Submitted. Prepubl avail. at http://www.let.rug.nl/nerbonne/papers/.

Nerbonne, John & Wilbert Heeringa. 2007. Geographic Distributions of Linguistic Variation Reflect Dynamics of Differentiation. In *Roots: Linguistics in Search of its Evidential Base*, ed. Sam Featherston & Wolfgang Sternefeld. Berlin: Mouton De Gruyter. Accepted to appear. Prepub. avail. at http://www.let.rug.nl/nerbonne/papers/.

Nerbonne, John, Wilbert Heeringa & Peter Kleiweg. 1999. Edit Distance and Dialect Proximity. In *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed.*, ed. David Sankoff & Joseph Kruskal. Stanford, CA: CSLI pp. v–xv.

Nerbonne, John & William Kretzschmar, eds. 2006. *Progress in Dialectometry: Toward Explanation.* Vol. 21(4) Oxford University Press. Special Issue of *Literary and Linguistic Computing.*

Niebaum, Hermann & Jürgen Macha. 2006. *Einführung in die Dialektologie des Deutschen, 2te, neubearbeitete Auflage.* Tübingen: Niemeyer. ^{1}1999.

Schiltz, Guillaume. 1996. German Dialectometry. In *Data Analysis and Information Systems: Statistical and Conceptual Approaches. Proc. of 19th Meeting of the Gesellschaft für Klassifikation, Basel, March 8–10, 1995*, ed. Hans-Hermann Bock & Wolgang Polasek. Berlin: Springer pp. 526–539.

Séguy, Jean. 1971. "La relation entre la distance spatiale et la distance lexicale." *Revue de Linguistique Romane* 35(138):335–357.

Séguy, Jean. 1973. "La dialectometrie dans l'Atlas linguistique de Gascogne." *Revue de Linguistique Romane* 37(145):1–24.

Trudgill, Peter. 1974. "Linguistic Change and Diffusion: Description and Explanation in Sociolinguistic Dialect Geography." *Language in Society* 2:215–246.

Wikle, Thomas & Guy Bailey. 1997. "The Spatial Diffusion of Linguistic Features in Oklahoma." *Proceedings of the Oklahoma Academy of Science* 77:1–15. avail. at `digital.library.okstate.edu/OAS/`.