Data-driven Dialectology

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

October 30, 2008

Abstract Most studies of language variation proceed from the geographic or social distribution of single elements (features), and find it difficult to proceed further. Data-driven dialectology, and more generally, data-driven variationist studies, begin instead from an aggregate view of language variation and reap immediate benefits in dealing with well-known exceptions in the distributions of single features and in avoiding the need to select which features to use as the basis of characterizations. But the major advance is the opportunity to characterize general tendencies in linguistic variation.

Keywords Language variation, dialectology, sociolinguistics, data-driven linguistics

1 Introduction

The Gershwins celebrated linguistic variation famously in *Let's call the whole thing off*: "You say tomato [tə.ˈmeɪ.ɾoʊ], and I say tomahto [tə.ˈma.toʊ], You say potato [pə.ˈteɪ.ɾoʊ], and I say potahto [pə.ˈta.toʊ]." The systematic study of linguistic variation profits in popularity from the fascination for "accents" and dialects, but it also interacts with historical linguists, who see in variation the pool of mutations from which language change arises, and with theoretical linguists in search of a broader base of material to test their ideas. Variationist linguistics proceeds from this natural fascination to solid and well-documented scholarship (Chambers & Trudgill 1998, [^{1}1980], Milroy & Gordon 2003, Niebaum & Macha 2006).

Variationist linguistics studies the differences in a single language as it is used in different areas (dialectology), or by people of different classes, occupations, sexes, or races (sociolinguistics). The name "variationist linguistics" has become popular as more and more attention has come to be paid to the many non-geographic factors which determine differences.

Most work in variationist linguistics focuses on the geographic or social distribution of single features or small numbers of single features. The Gershwins' lyrics mention about five varying sounds in nine different words,[1] and many linguistics articles study even smaller numbers of varying elements. This article presents a data-driven alternative which has developed under the name DIALECTOMETRY in which large numbers of features are aggregated when varieties are characterized (Séguy 1973, Goebl 1984). We argue below that students of linguistic variation implicitly accept the need to aggregate over many linguistic differences but differ on the level of abstraction at which aggregation is to be employed. Seen from this perspective, the focused contribution of present essay is to reflect on why higher levels of aggregation make sense.

We summarize the 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

Linguists normally analyze variation in one element—i.e., FEATURE—at a time, whether the elements (or features) be sounds, words, morphemes, con-

[1] For the cultural historians, we note that the others were 'either' and 'neither', 'pajama' and 'after', 'vanilla' and 'parilla, and 'oyster'.

structions, or whatever. 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 (see note 3 for references). We refer to such characterizations based on individual features as (SINGLE)-FEATURE BASED,[2] 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. As a passage from *Judges* suggests (see note), single features are often sufficient for the detection of signals of social and geographical belonging. There are moreover 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.

We nonetheless argue that aggregating techniques supplement the existing feature-based analytic arsenal significantly and enable answers to some fundamental questions of variationist linguistics that are inaccessible to single-feature studies. Having said that, we need of course to explain what we see as the fundamental questions of variationist linguistics. We turn to this directly.

Speakers employ overlays of variation in form in order to signal geographic and social provenance (even if they speak primarily to communicate). The study of linguistic variation focuses on this variation and how it signals 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 or, at least, Southern.[3]

Without presuming to define variationist linguistics, we proceed uncontroversially from the assumption that one of its tasks is to characterize the signals of provenance language users provide.[4] One way of re-stating our

[2] They might less neutrally be referred to as SHIBBOLETH-BASED as the earliest commentary on linguistic variation is *Judges* (12:6), which focuses on a single feature:

> [...] Then said they unto him, Say now Shibboleth: and he said Sibboleth: for he could not frame to pronounce it right. Then they took him, and slew him at the passages of Jordan: and there fell at that time of the Ephraimites forty and two thousand.

[3] Labov (2006) analyzed New York city's loss of syllable-final /r/, LAMSAS provides the documentation of the lexicalization of 'dragonfly' (Kretzschmar 1994), and Abraham (2008) discusses the German verb clusters.

[4] Of course we realize that some speakers consciously modify their pronunciation so as

main thesis is that these signals need to be studied in the aggregate. At higher levels of aggregation, signals become more reliable, signals may be compared to one another with respect to their relative strength, and, last but not least, general laws of linguistic variation may be stated.

1.2 Structure

Section 2 presents an archive of dialect data, demonstrating the enormous amount of variation often left unmentioned in textbooks and even in scholarly articles on language variation, which normally abstract away from this to some extent, implicitly introducing aggregation at a very low level.

Section 3 develops the argument that single features inevitably contradict each other, at least in detail, and that they tend to have exceptions and sparse distributions (like all linguistic distributions). The point of discussing this admittedly well-known point in more detail here is two-fold. First, in our experience linguists underestimate how large and genuine the problem is. To make the point concrete, we continue the analysis begun in Section 2. Second, the problematic status of single-feature distributions motivates moving to higher levels of aggregation.

Section 4 below is devoted to the criticism that single-feature studies are methodologically weak in allowing too much freedom of choice concerning *which* features are said to figure in the definition of dialect areas.[5] We contrast this with a view which aggregates over a large number of features, which views aggregate differences as characterizing the relations among linguistic varieties. Section 5 presents the case that an aggregated perspective enables the formulation of more general laws. Finally Section 6 discusses how the present view vis-à-vis aggregation differs from earlier views *inter alia* on "bundling isoglosses" and on attention to structural effects.

Our primary point is that linguistic variation ought to be studied in the aggregate. Our approach is massively indebted to the DIALECTOMETRY of Séguy (1973) and Goebl (1984), as noted above, and we are 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.

to mask their real provenance, but we regard them too as providing signals of provenance, only misleading ones, and will not take care to consider them separately in what follows.

[5] And we claim that the problem of the choice of features re-emerges even if one interprets the data as signaling, not areas, but rather dialect continua (Heeringa & Nerbonne 2001) or membership in another extralinguistic group. We wish to remain judiciously vague about the nature of the extralinguistic provenance which is signaled.

2 Low-Level Aggregation

By focusing exclusively on single features or small combinations of these, variationists, including dialectologists, sometimes fail to isolate signals of provenance clearly. The signals are often so complex, even misleading, that they resist analysis using simple, single-featured methodologies. This is one reason to employ the aggregating techniques of dialectometry. This and the following section follows aggregating steps from very specific to quite general levels.

The main point of this section is to show how abstraction leads to more satisfying characterizations of how provenance is signaled in linguistic variation, and to reflect on this. This step of abstraction naturally implies aggregating data into the more abstract categories. This section follows the data from large collections (the archives on which dialect atlases are based) to characterizations of its distribution in maps. The purpose of following the analysis this way is to emphasize how common certain steps of abstracting and aggregating are. We hope to make more ambitious aggregating steps more palatable in this way.

Inspecting the archive of linguistic material behind an atlas, one is struck by the amount of variation that never makes it to the beautiful pages of the atlas itself. We examine material from a dialect data collection in order to drive home the point that the data is extremely varied. In what follows we use material from the *Phonetischer Atlas Deutschlands* (PAD), material collected between 1965 and 1991 by Marburg dialectologists 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 notation 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 and conflicting indications.

Our second reason for wishing to review this material is to drive home the point that dialectology already makes use of a number of steps of abstraction that implicitly aggregate. 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.

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 [i] 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	ɛkʰ	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	ɪkʰ	ɪɕ	ɪx̟	ʏç̱	ʏʑ	e	e̱͡ɪɣ	e̜ʔ͡k	e̜ç	e̜g	ẹʃ̫
ẹç͡j	ẹç	ẹɣ	ẹg	ẹj	ẹɕ	eg	ek	ek͡x̰	i	iː	iːç	iːç̰	iç
i̜	i̜ːj͡ç	i̜k											

2.1 Phonetic Tokens in Single Words

The PAD is similar to most linguistic atlases in being recorded in 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 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* in the PAD, even if only three different consonants were recorded. This sort of detail 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/).

We shall 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, $\mathrm{k^h}$, $\mathrm{k^j}$, and $\mathrm{g^j}$] 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, ʃ̠] etc.), the non-fricative 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.

The degree of phonetic detail in Table 1, and that in most dialect atlas collections,[6] suggests that we shall always need to move from low-level characterizations to more abstract levels. The 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 to allow more room for insight. But let us 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.

2.2 What is signaled?

We turn to an examination of the geographic distribution of the final stop variants found in *ich* /ɪx/. 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

[6] 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 atlases are faithful renderings of speech, which, however contains a great deal of sub-dialectal as well as dialectal variation. But this is a point about which serious criticism is certainly possible.

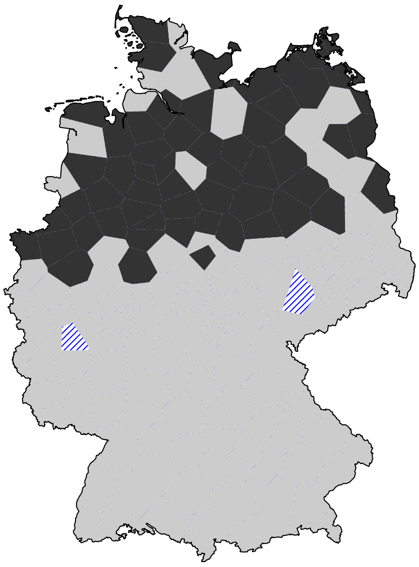


Figure 1: The [k/x(ç)] distinction in the word *ich* in the PAD. The darker the polygon, the greater the concentration of stop variants. There was no data available for the polygons with diagonal lines.

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 obtained when mapping any single linguistic feature, even after abstracting over a great deal of detail in variation.

The pockets of exceptions in distributions such as that in Figure 1 has often been remarked on in dialectology. 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 linguistic features whose distributions are inevitably rough.

We sum up this section by noting that good dialectological practice has always aggregated in this fashion, abstracting from extremely detailed recordings to more abstract renderings of selected differences. We do *not* 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.

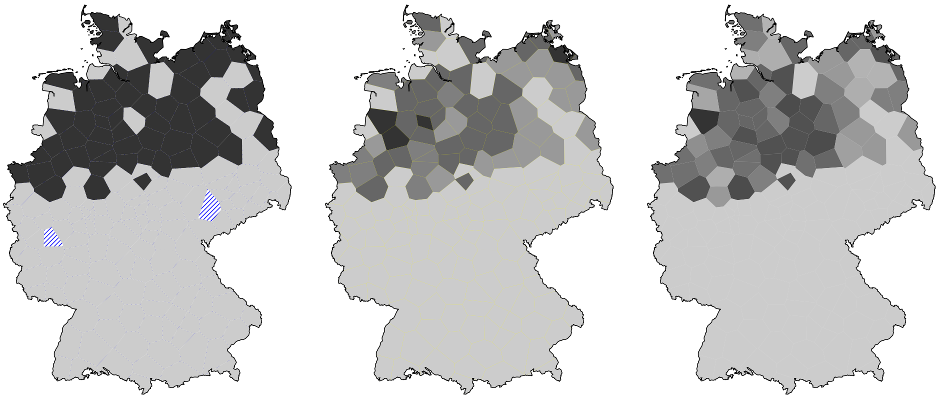
2.3 Phonetic Tokens in Multiple Words

In fact, a second step of aggregation is likewise standard, 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 aggregate over the variants used in their pronunciations. In our data collection, this includes the words *ich* /ɪx/ (standardly pronounced [ɪç]), *dich* /dɪx/, *auch* /aux/, and *gleich* /glaɪx/. (In fact we likewise have *schlechte* /ʃlɛxtə/ and *schlechten* /ʃlɛxtən/ in the dataset, but these – etymologically different – tokens of /x/ are never pronounced alternatively 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, representing increasingly more inclusive aggregations, 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 "smoothes" the distribution a bit, for example, filling in the sites marked by diagonal lines 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 such maximally comprehensive maps best indicate what the linguistic variation is signaling, in this case whether the speaker comes from the north of Germany or not.

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 (phoneme) as it occurs in multiple words. This sort of aggregation, too, is common throughout dialectology.

We note nonetheless that this aggregating step risks historical confusion, that of confusing etymologically different elements such as the phoneme /x/ as in *ich* vs. *schlecht*. So while we illustrate various aggregates based on



ich [k] vs. [ç/x] *ich, gleich* all five occurrences

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

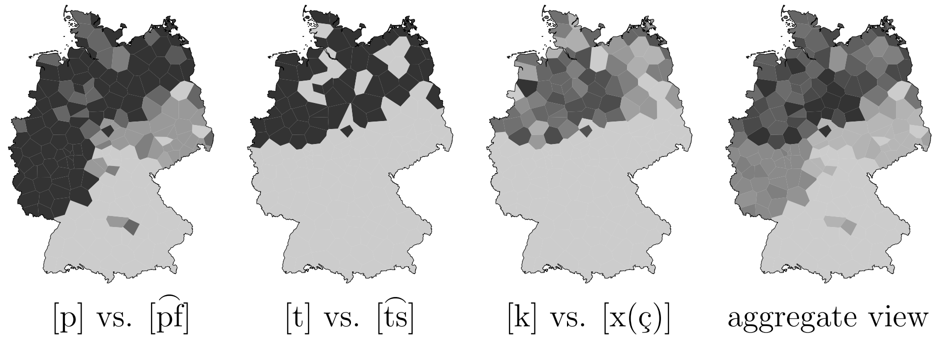
standard German pronunication, we restrict our general technique to the comparison of single words (etymologically unassailable), and then aggregates across the entire vocabulary sample. We could imagine that signals of provenance are detectable across etymologically different elements, but the topic cannot be pursued here.

2.4 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.

Naturally, the step toward the abstracter characterization aggregates over more linguistic material, and so it is not surprising that the signal of provenance associated with it is more reliable. We return to this in Sec. 5 below.



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

Figure 3: The stop/affricate distinction resulting from the second sound shift, where the relative frequency of stops is shown in darker shading. 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 Two, Three, Many "Features"

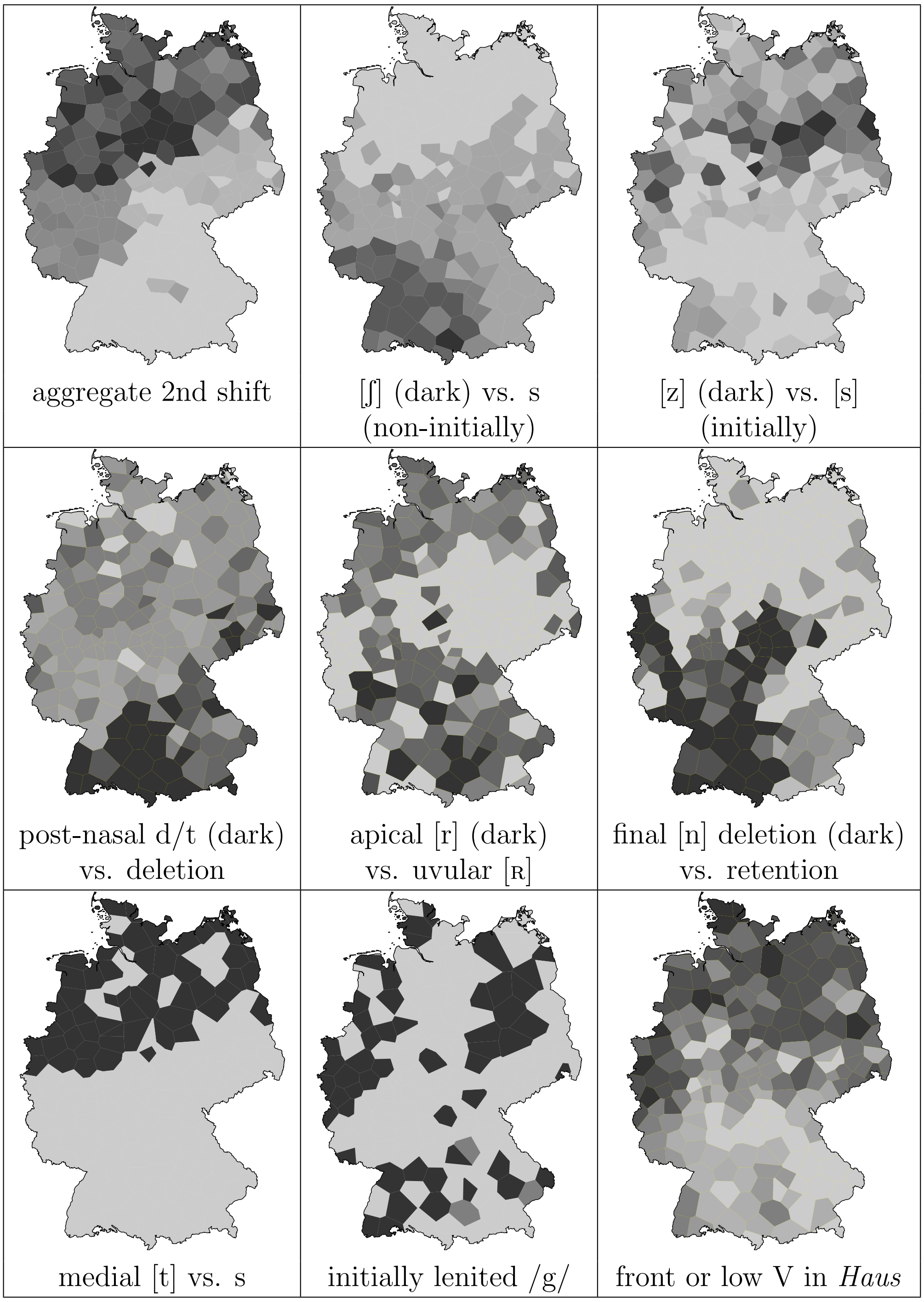
But there is a great deal more systematic geographic variation which further aggregating steps may incorporate. We extract a number of features from the PAD and sketch their geographic distribution in Figure 4. We select well-discussed features of German dialectology (König 1994, Niebaum & Macha 2006), 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 $\rightarrow \emptyset$ / 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



aggregate 2nd shift

[ʃ] (dark) vs. s (non-initially)

[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 geographically conditioned, but overlapping only imperfectly. See text for further explanation.

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.

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 variant pronunciations 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 note that this division is geographically coherent. Bottom row, right column.[7]

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. This illustrates the complaints of Bloomfield noted above, and many researchers since (Wagener 1988). We suggest aggregating over many features in order to detect reliably the relations different varieties have to one another. Second, Fig. 4 also 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 subconsciously. At this point we

[7] One colleague noted critically that this step involves aggregating over etymologically different elements since the vowels in *Haus* and *verkaufen* have different sources. The point is well taken. See discussion at the end of Section 2.3.

suppress a full explanation of alignment software which automatically aligns corresponding phonetic segments as this work has been presented in detail elsewhere (Kruskal 1999, Nerbonne, Heeringa & Kleiweg 1999).[8]

It would take us too far afield to present the workings and the analysis of the alignment algorithm in detail, but the overall effect is readily sketched. For each pair of field work collection sites (varieties), we align the pronunciations of each word in the list of words (or phrases) elicited. The result of alignment is illustrated as follows:

k	œ		s	t	ə	'crust'
k	ɔ	r	s	t		
	1	1			1	(sum = 3)

We note especially the non-aligning points, as these contribute to the pronunciation distance between the two varieties. The alignment algorithm is sensitive enough to find the optimal alignment, i.e. the one in which the sum of differences is minimal, after which it is child's play to total up the number of non-aligning points (Kruskal 1999). If we have a great deal of material, it is sufficient to use exactly this rough measure. Refining the alignment procedure to use phonetically more sensitive measures of segment distance is the subject of ongoing research (Heeringa, Kleiweg, Gooskens & Nerbonne 2006), but the very simplest measure assays pronunciation distances reliably.

The procedure is applied to all $\binom{186}{2} = 17.390$ pairs of sites, comparing 201 words in each site comparison, resulting in $3,5 \times 10^6$ word comparisons. Since the mean length of words (in phonetic segments) is 5, the overall comparison is based on $1,75 \times 10^7$ segment-pairs in this procedure (in fact, the algorithm examines many non-corresponding segments as well, which we ignore at present). It is this very comprehensive comparison which allows us to obtain reliable results using rough comparisons.[9]

The mean distance between varieties is obtained from the pronunciation distances of the words in the two samples, yielding a large distance table. At this point, there are several techniques available for analyzing the resulting distance table. We note two here. First, we may apply clustering to seek groups in the data. In fact we apply clustering using a bootstrap procedure (or equivalent) in order to ensure stability in results. The borders which then emerge are sketched on a map (Nerbonne, Kleiweg & Manni 2007) (see

[8] In particular Heeringa (2004) provides a detailed presentation 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.

[9] We discuss the validation of our measurements below briefly, and at more length in (Nerbonne & Heeringa to appear).

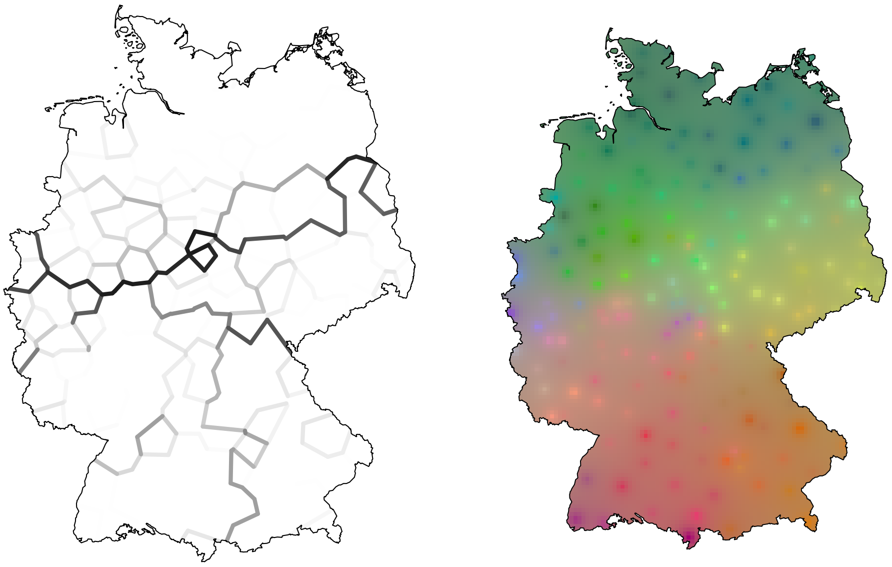


Figure 5: Two visualizations of aggregate pronunciation variation in the PAD. On the left, darker lines correspond to better founded distinctions. On the right, the major dimensions of a multi-dimensional scaling analysis of the aggregate pronunciation differences are represented as intensities of red, green and blue (in order of significance).

Fig. 5, left). Alternatively, we apply multi-dimensional scaling (MDS) to obtain a low-dimensional representation of the data set. It turns out that a three-dimensional solution represents the input data well, in the sense that the distances between sites in the input matrix correlate well with distances measured between the sites as given by their coordinates in the three dimensions inferred by MDS ($r \approx 0.89$). We therefore use the three-dimensional solution, and sketch the map by associating each dimension with an intensity in the red, green and blue color scheme. Fig. 5 (right) presents these visualizations of the results of the alignment-based comparison as well. These are different visualizations of the same aggregate view of the pronunciation differences within Germany based on the assessment of proportion of matching segments, as just described.

It is important to add, however, that there are many aggregating techniques, and that there are debates about the best techniques of analysis (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.

4 Keeping it Simple

A second reason to proceed beyond single features or shibboleths 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, for example by reference to "bundles of isoglosses", the range of possibilities rises enormously (see Section 6.1 below for a discussion of isogloss bundles).

In this section we charge single-feature variationist studies with discriminating too little in the features they are willing to entertain. Given the current state of the art, in which researchers 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.

Sociolinguistics introduced the use of frequency in examining variation, esp. in its well-known "varb-rule" (variable rule, or logistic regression analyses, see Paolillo (2002)). 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, allowing us to smooth over some exceptions, but it also allows additional degrees of freedom to the analysis. The number of possible hypotheses is multiplied again by the incorporation of frequency information in analyses.

Aggregating accounts postulate that extralinguistic variables are associated with aggregate differences in entire varieties, not merely with specific linguistic variables. The motivation for this postulate stems from 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. The information in the aggregate view in Fig. 5 shows that even speakers within pockets of exception appearing in single feature distributions such as Fig. 1 will provide signals of provenance.

5 General Characterizations

In this section we sketch issues to which we claim that aggregate analyses contribute. They progress from relatively concrete to more speculative.

The present essay has focused on pronunciation as this has received the lion's share of attention in dialectology and variationist studies, but studies on lexical and syntactic variation also underscore the value of the aggregate perception (Nerbonne & Kleiweg 2003, Spruit 2008, Gooskens & Heeringa 2006).

Forgetting (bundles of) isoglosses

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 variation is organized into areas of relative homogeneity in some case and in continua of increasing differences in others. Single-feature studies have tried to identify the single features or "bundles" of features responsible for this, but there is no consensus on how to evaluate the relative strengths of different features, however (Chambers & Trudgill 1998, [^{1}1980], 96-97):

> It is undeniable that some isoglosses are of greater significance than others [...and...] that some bundles are more significant than others [...]. Yet, in the entire history of dialectology, no one has succeeded in devising a satisfactory procedure or a set of principles to determine which isoglosses or which bundles should outrank some others. The lack of a theory or even a heuristic that would make this possible constitutes a notable weakness in dialect geography.

It should not be surprising that we suggest that aggregating techniques should be brought to bear. The purpose of identifying (bundles of) isoglosses is to identify dialect areas or dialect continua, more generally the geographic elements that are associated with linguistic variation. If the aggregate perspective is correct, we should seek evidence for this sort of geographic influence on variation in the aggregate of many variables. So we maintain that Fig. 5 represents the geography of German linguistic variation. See also Sec. 6.1 below for further discussion.

Identifying important features

But using the aggregate to identify the important geographic parameters does not mean that we should ignore the importance of some features vis-à-vis others. Single-feature based dialectology fails to interpret individual features with respect to global patterns and is therefore unable to assess the importance of the individual signals it studies. Aggregation is the key step needed if one is to assess which features most reliably indicate global patterns.[10]

For example, in the German example we developed above, let us calculate a place × place distance matrix for each of the 201 words in the PAD sample. We then examine which word matrices correlated most strongly with each dimension of the multi-dimensional scaling analysis (Fig. 5, right). It turns out that *Zeiten* 'times', *sein* 'be (inf.)', *bleib* 'stay (imp.)' and *weisse* 'white (infl.)' are at the top of the list, all showing strong correlations ($r > 0.5$), suggesting that the variation in the stressed vowel (standard German [aɪ], but South [i]) is the single strongest indicator of provenance among the 201 words in our sample. Prokić (2007) explores more systematic analysis of the aligned segments with the goals of identifying the linguistic factors in aggregate analysis.

There are more sophisticated techniques in use as well. Shackleton (2005) use principal component analysis in an aggregate analysis to sketch the relations between English and American dialects, and Nerbonne (2006) uses factor analysis to identify the recurrent features in a phonological analysis of

[10] Kretzschmar (2006, 400) sounds a bit disparaging about one prominent dialectologist he collaborated with extensively when he reports that the colleague "[...] could afford to ignore interpretation of the data because he already knew what it meant," but Kretzschmar's respectfully intended point is that this researcher was so familiar with the data that he easily identified the signals correctly. We suspect that many other experienced colleagues have such excellent intuitive sense of the variation they study that they judge the significance of individual features quite well. Nonetheless, relying on informed intuition rather than analytic technique provides no foundation for more abstract questions.

LAMSAS (Kretzschmar 1994). Hyvönen, Leino & Salmenkivi (2007) apply independent component analysis to Finnish lexical data.

Linguistic distance as a function of geography

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 the variety of features and to the fact that there are exceptions, but especially because such accounts characterize only the relations among varieties 'with respect to a single feature'. There is no attempt to aggregate over a large number of such relations. But this means that the notion "linguistic variety" *simpliciter* — the collection of speech patterns used in a community — plays no role in standard analyses. Aggregate analyses proceed, on the contrary, by assaying the relations between varieties based on a sample of speech habits.

Another reason to aggregate more vigorously in variationist linguistics is the opportunity to formulate more general characterizations of variation. Variationist linguistics has been aware of the difficulty of working from single-feature distributions to more general characterizations of varieties, dialects, sociolects, dialect areas and the like, and there are numerous discussions of how single-feature studies are related to more general characterizations. These discussions falter universally on the usual complexity of distributions of single features, which inevitably have exceptions, and normally contradict each other, at least in detail.

As an example of a general theoretical question in dialectology which aggregate studies are 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 decidedly 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 that Trudgill's "gravity hypothesis" needs to consider not only geography (and population size), but also social and political relations. The criticisms are well taken, but sidestep the question of *how* geography influences variation. We suggest that the problem is the level of analysis, in particular that previous studies have focused on a small number of variants.

Aggregating analyses such as Séguy (1971) have long noted a simple, law-like relation between geographic distance and linguistic variation of exactly the sort Trudgill sought, and about which the other authors cited have been skeptical (arguing that the relation is more complex than Trudgill postulated)

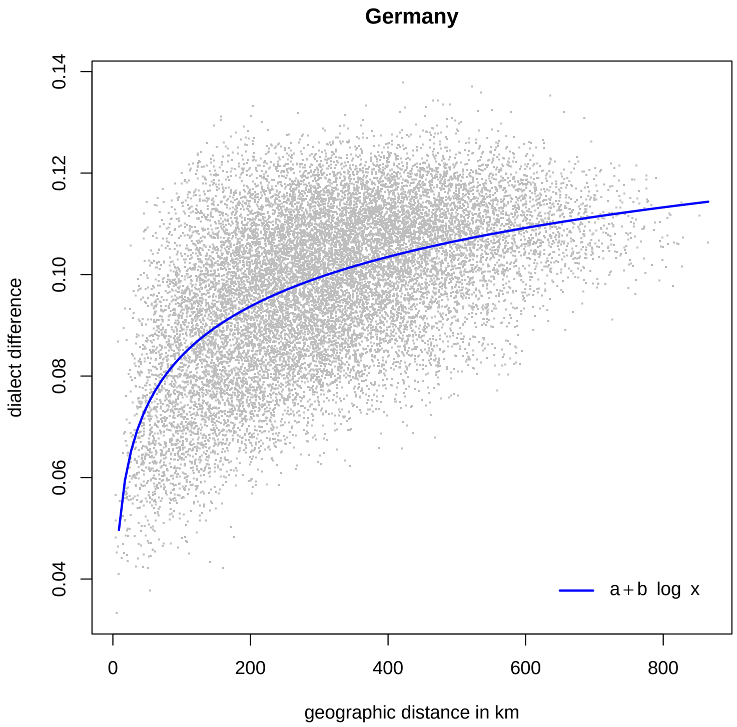


Figure 6: The Séguy curve. Linguistics variation is a sublinear function of geographic distance just as Séguy (1971) demonstrated, but the characterization requires an aggregating step. The curve is based on the German data in this paper, and shows a logarithmic model. The sublinear curve explains about 30% of the variation in the data ($r \approx 0.56$), essentially replicating Séguy's analysis on a novel data set.

(Nerbonne & Heeringa 2007). In general linguistic variation increases as a sublinear function of geographic distance, as Fig. 6 illustrates.[11]

Reliability and Validation

Working with a large aggregate also allows us to analyze the reliability of data formally, and we have done this for several large data sets. The reliability of the German data set as analyzed by Cronbach's α is at least 0.95. A

[11] We exclude exceptional "breaks" in the linguistic landscape such as those due emigration or colonization. The only exception thus far is to the sublinear pattern is Spruit (2008), who finds that a linear model gives a slightly better fit than a sublinear model when examining syntactic distance.

common threshold in the social sciences in 0.7. See Heeringa et al. (2006) for explanation of the statistical calculations.

Aggregating views likewise offer new perspectives on the validation of claims in variationist linguistics, which Gooskens & Heeringa (2004) have indeed already initiated. Whoever claims to characterize signals of provenance 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 the aggregate pronunciation measure used in this article using perceptual data, focusing on Norwegian data. There is of course other work with validating aims (Clopper, Levi & Pisoni 2006), but with less general measures of varietal difference.

All accounts of variation — aggregating or not — should in principle be subjected to some sort of behavioral validation. The criticisms in Section 4 as well as the more complicated analyses associated with aggregating analyses convince us that we should not be content with just fitting models to data.

New Vistas

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.

6 Previous Aggregate Views

There has certainly been a good deal of attention paid to aggregating approaches, most notably by Goebl (1984), who has pursued this work for three decades (Goebl 2006). 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.

6.1 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 might 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 (*pace* Séguy, 1973:14). We note that isogloss bundling, strictly speaking, only makes sense for geographic variation, whereas aggregating makes sense for many sorts of variation. But we take the counting of features that are different to constitute the essential insight. While this was a step forward, it is clear that more had to be done.

The most important difference is that we do not rely on a specific, potentially biased choice of which isoglosses to bundle. Instead, we envisage using all available material ideally in a randomly chosen sample.[12] After all, there is an enormous number of potential subsets of features for bundles to consist of. Without the discipline of working with an independently specified set of features, the "isogloss bundling" technique runs the risk of being even less restrictive about its admissible hypotheses than the single-feature approach.

Second, 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). Third, there are many ways to operationalize aggregation that are unlike counting isoglosses. For example, we have quantified the occurrence of contrasting elements in the maps above, using a numerical characterization of pronunciation difference. In the simplest versions of the alignment technique, the results are more or less counts of isogloss differences. But in many more complex versions, for example those where segment differences are specified via phonetic differences (derived from phonetic feature systems), or whether they are weighted by frequency, this interpretation is impossible. Fourth, 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). In these cases isoglosses may e.g. run on the one or on the other side of a given data collection

[12] Bolognesi & Heeringa (2005) worked with a randomly chosen subset of Sardinian words in a new survey of Sardinian variation. This is not a common tack, however.

site (or set of sites). But we are still capable of detecting similar groups. For example, clustering can take non-local differences into account and thus detect borders even where differences are gradual. For a second example, Monmonier analysis seeks borders only after discounting the general effect of geography on linguistic distances (Manni, Heeringa & Nerbonne 2006).

6.2 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. Further, we have a much less structured notion of aggregation in mind in this essay.

Labov is also the one of the developers of variable rule analysis, a major step forward in sociolinguistic analysis (Paolillo 2002). It should be clear, however, that variable rules are a means of conducting analyses of single features or small numbers of features.

7 Conspectus and Prospectus

The essential aggregating steps are 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 more reliable 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 or small subset of variables 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.

There are innumerable future tasks, as always. We certainly need to continue to hone techniques, both with respect to linguistic sensitivity and with respect to isolating the most important linguistic components in aggregate tendencies. We have been deliberately vague about the many different sorts of aggregates which may be examined, and we should prefer to restrict the hypotheses we entertain. A major further challenge lies in confronting aggregate analyses aimed at identifying historical and typological relatedness (Nerbonne 2007) and developing techniques capable of separating the different sorts of effects.

There are intriguing opportunities to use aggregate analyses in conjunction with the detection of other signals of cultural and genetic relations. Which cultural and linguistic signals tend to be associated with each other, and to what degree? We are just now catching glimpses of what might be possible (Manni, Heeringa & Nerbonne 2006).

The work behind the wonderfully large data collections that dialectologists have bequeathed us has unfortunately not been continued in other areas of variationist linguistics, meaning that it is difficult to obtain enough data to test hypotheses outside of geographic variation. If we are correct in arguing for the importance of aggregate analyses, we need much more comprehensive collections of variationist data for which other potential correlates of variation are noted. It would be fantastic to see larger collections compiled and made available.

Acknowledgments

An earlier version of this paper appeared as “Variation in the Aggregate: An Alternative Perspective in Variationist Linguistics” in K. Dekker, A. MacDonald and H. Niebaum (eds.) *Northern Voices: Essays on Old Germanic and Related Topics Offered to Prof. Tette Hofstra*, Leuven: Peters, 2008, 365–382.

We thank Peter Kleiweg for programming and graphics (`http://www.let.rug.nl/kleiweg/L04/`), and Jack Chambers, Charlotte Gooskens, Jack Hoeksema, Stephan Kleiner, Sebastian Kürschner, Therese Leinonen, Alasdair MacDonald, Hermann Niebaum, Jelena Prokić, and Martijn Wieling for

critical discussion and suggestions for improvements. As usual, they should get credit, but deserve no blame for remaining errors and problems.

References

Abraham, Werner. 2008. Gradience in the V-cluster. In *Describing and Modelling Variation in Grammar*, ed. Andreas Dufter, Jörg Fleischer & Guido Seiler. Berlin: Mouton De Gruyter p. to appear.

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.

Bolognesi, Roberto & Wilbert Heeringa. 2005. *Sardegna fra tante lingue. Il Contatto linguistico in Sardegna dal Medioevo a oggi.* Cagliari: Condaghes.

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.

Goebl, Hans. 2006. “Recent Advances in Salzburg Dialectometry.” *Literary and Linguistic Computing* 21(4):411–436. Spec. Issue, *Progress in Dialectometry: Toward Explanation* ed. J. Nerbonne & W. Kretzschmar, Jr.

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

Gooskens, Ch. & W. Heeringa. 2006. "The Relative Contribution of Pronunciational, Lexical and Prosodic Differences to the Perceived Distances between Norwegian Dialects." *Literary and Linguistic Computing* 21(4):477–492.

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.

Heeringa, Wilbert, Peter Kleiweg, Charlotte Gooskens & John Nerbonne. 2006. Evaluation of String Distance Algorithms for Dialectology. In *Linguistic Distances*, ed. John Nerbonne & Erhard Hinrichs. Shroudsburg, PA: ACL pp. 51–62. Proc. of a workshop held at the joint meeting of ACL and COLING, Sydney, July, 2006.

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.

Hyvönen, Saara, Antti Leino & Marko Salmenkivi. 2007. "Multivariate Analysis of Finnish Dialect Data—An Overview of Lexical Variation." *Literary and Linguistic Computing* 22(2):271–290.

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.*

Kretzschmar, William A., ed. 1994. *Handbook of the Linguistic Atlas of the Middle and South Atlantic States.* Chicago: The University of Chicago Press.

Kruskal, Joseph. 1999. An Overview of Sequence Comparison. In *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, ed. David Sankoff & Joseph Kruskal. Stanford: CSLI pp. 1–44. ^{1}1983.

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

Labov, William. 2006. *The Social Stratification of English in New York City.* Cambridge: Cambridge University Press. (1st ed., 1966).

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.

Milroy, Lesley & Matthew Gordon. 2003. *Sociolinguistics: Method and Interpretation.* Oxford: Blackwell.

Nerbonne, John. 2006. "Identifying Linguistic Structure in Aggregate Comparison." *Literary and Linguistic Computing* 21(4):463–476. Special Issue, J.Nerbonne & W.Kretzschmar (eds.), *Progress in Dialectometry: Toward Explanation.*

Nerbonne, John. 2007. "Review of April McMahon & Robert McMahon *Language Classification by Numbers.* Oxford: OUP, 2005." *Linguistic Typology* 11.

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

Nerbonne, John & Peter Kleiweg. 2003. "Lexical Variation in LAMSAS." *Computers and the Humanities* 37(3):339–357. Special Iss. on Computational Methods in Dialectometry ed. by John Nerbonne and William Kretzschmar, Jr.

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. Accepted. 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. to appear. Measuring Dialect Differences. In *Theories and Methods*, ed. Jürgen Erich Schmidt & Joachim Herrgen. Language and Space Berlin: Mouton De Gruyter.

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.

Paolillo, John C. 2002. *Analyzing Linguistic Variation: Statistical Models and Methods.* Stanford: CSLI.

Prokić, Jelena. 2007. Identifying Linguistic Structure in a Quantitative Analysis of Dialect Pronunciation. In *Proceedings of the ACL 2007 Student Research Workshop.* Prague: Association for Computational Linguistics pp. 61–66.

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 dialectométrie dans l'Atlas linguistique de Gascogne." *Revue de Linguistique Romane* 37(145):1–24.

Shackleton, Jr., Robert G. 2005. "English-American Speech Relationships: A Quantitative Approach." *Journal of English Linguistics* 33(2):99–160.

Spruit, Marco René. 2008. Quantitative Perspectives on Syntactic Variation PhD thesis University of Amsterdam.

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

Wagener, Peter. 1988. *Untersuchungen zur Methodologie und Methodik der Dialektologie*. Marburg: N.G. Elwert.

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/`.