

Crosstalk in Humanities Computing

John Nerbonne¹

June 27, 2008

1 Introduction

In this essay I would like to stimulate *crosstalk*, or interdisciplinary investigation among the humanities, and more specifically to show how humanities computing might prompt this by providing common means of analysis. The main practical point we emphasize is that techniques from humanities computing may be applied to linguistic variation but also to other cultural variation which does not involve language. The two areas of analysis—linguistic and nonlinguistic culture—have to be compared if we wish to understand whether culture is transmitted along similar paths and in similar ways. Then the essay continues to the more polemical point that interdisciplinary work is worthwhile in itself and therefore that humanities computing is more worthwhile for further enabling the interdisciplinary work.

The somewhat negative term ‘crosstalk’ is used deliberately here to acknowledge that efforts at interdisciplinary investigation are inevitably going to involve difficulties in communication, even misunderstandings. The potential benefits of interdisciplinary work outweigh these problems; in fact the unavoidable misunderstandings may even prove stimulating.

Cultural variation and cultural diffusion underscore the main argument

that such crosstalk might take place at a more fruitful level than before. Culture—meaning material culture, linguistic culture, religious culture, etc.—is at the core of several humanities disciplines, and it is plausible to hypothesize that the distributions of different cultural expressions follow similar patterns. The essay illustrates how the study of linguistic culture has benefited from the use of computational techniques suggests several avenues along which crosstalk might be pursued. The computing perspective suggests where this interdisciplinary work might progress beyond older views.

1.1 Background

Since this paper addresses a general humanities audience, it makes sense to state some background assumptions about computers and the humanities. Willard McCarty recently published a book-length essay on the proper view of the relation between the two, arguing that the deployment of computing in the humanities will deepen our understanding of humanities themselves, and ultimately lead to innovations in scholarly form and content.² Others plea for a focus on questions of digital culture, or an emphasis on pedagogical applications of computers.³

It is unwise to embrace these views exclusively, however. Humanities computing should pursue the research of the humanities using computing: computing in service of the humanities.⁴ In this view computationalists should seek answers to scholarly questions in linguistics, history and art history by using the computer, exploiting especially its ability to process large amounts of data and the transparency of its processing. From this perspective humanities computing be viewed, not as a discipline, but rather as a federation of disciplines, whose practitioners find it opportune to collaborate for reasons of some common problems. Our ability to deal with

large amounts of data marks the distinctive contributions we can make to humanities scholarship.

This pedestrian view is opposed to the more revolutionary views of McCarthy on humanities computing, but it recognizes the exciting developments in the various humanities disciplines which only information and communication technology (ICT) could enable—even if they do not suggest very new visions of the disciplines they arise in. We can now parse tens of millions of sentences in an effort to understand syntactic structure, and Gertjan Noord provides examples of points where this has uncovered new facts in syntax.⁵ Wilbert Heeringa has applied computational techniques in dialectology, and has established a perspective from which dialect areas and dialect continua make sense.⁶ In addition Heeringa together with Charlotte Gooskens initiated a discussion on the validation of techniques in dialectology, a reflective counterpart to the empirical and analytic progress which ICT enabled.⁷

There are by now countless more examples of well-respected contributions to many subfields in history, art and architectural history, linguistics, archaeology, musicology, and biblical studies. This means that humanities computing is now contributing to humanities scholarship, and it should be clear that its potential for further contribution is even greater.

1.2 Present Essay

The present essay, intended for a general humanities audience, will proceed from the thinking we have developed based on our work on dialects in order to suggest that this sort of work might interact with studies involving other humanities subfields, including at least social and economic history, archaeology, and architectural history. The goal is to sketch opportunities

for crosstalk in the humanities, stimulated by our common interest in culture and our common computing infrastructure.

2 The Humanities

One of the exciting parts of older discussions about the humanities is the unity they see in the study of how human culture is constituted and develops linguistically, socially and materially.⁸ Even if Wilhelm von Humboldt's ideas of language influencing thought seem overly focused on national(ist) tendencies, and even if they remain controversial nearly 200 years later,⁹ still they concern how cultural patterns are reflected in thought, language, material culture, forms of interaction and organization, as well as in music, art and literature. Of course, the reflections are imperfect, and the mutual influences are not determinate, so there is likewise large variation in how much the different expressions of culture correlate.

But it would be wonderful to go beyond these general remarks in order to understand at least some of these interactions more thoroughly. The undertaking will almost certainly have to be collaborative, as no single discipline commands all the needed knowledge, analytical skill and interpretive traditions. Linguistics is in a position to contribute to such a collaborative undertaking, and this essay aims to stimulate adventurous minds in other disciplines along these lines.

3 A Common Problem: Cultural Transmission

For present purposes it is useful to understand *culture* broadly, more or less as the economists urge, namely as all of the activity of a group that is not motivated directly by the satisfaction of natural needs, such as the need to

eat, to sleep, to reproduce, to protect oneself, and to stay warm. In this sense culture may be contrasted with nature, even if it may often be unclear whether cultural or natural forces are at work in a given phenomenon (such as manners of greeting, ideals of physical beauty, or sounds of aggression). Given this broad definition it is clear that several humanities disciplines study culture in one or another guise. Archaeology studies ancient cultures through their material remains; history focuses on cultural activity (including political, social and economic activity) which has left written records, and the literary, musical and visual arts tend to focus on culture in the narrower sense, the culture devoted to producing art in its different forms.

Since everyone wants “cobblers to stick to their lasts,” I shall focus here on linguistics, which is most frequently motivated by curiosity about those elements of language that might be common to all languages, and that therefore might plausibly be postulated to belong to humans by virtue of their genetic make-up, a fascinating and important aspect of language.¹⁰ But even if some very general linguistic properties are part of our innate, natural endowment, it is nonetheless clear that a great deal of language is culturally conditioned. All those properties that differ across languages—words, sounds, grammatical structures—must involve some choice beyond what is naturally given, even if there is little consensus among linguists about the division of labor concerning what is innate and what is culturally added.

At a certain level of abstraction, most of the humanities disciplines study culture, therefore, even if their foci differ, and even if they include non-cultural perspectives as well. We have deliberately not included law in our list, since it is no longer normally regarded as a humanities discipline,¹¹ but some aspects of legal history might be included as well.

The study of culture is difficult, and G.H. von Wright has even argued that the humanities need to strive after *understanding* rather than *explanation*—in part motivated by the need for sensitive interpretation in the humanities.¹² There nonetheless remains a role for simple, empirical studies of the sort that may be operationalized to the extent needed to benefit from a computational approach. Cultural differences may be assayed in a number of legitimate ways, and it is straightforward to operationalize a number of these. In linguistics we can get a rough, but nonetheless valid picture of variation in a group from very simple facts about whether linguistic habits are the same or different. The variation may involve pronunciation, or word choice, or the use of an alternative grammatical pattern, and the group may be the tribes in a given area, or the villages of a more modern state, or even several social groups within a single community (the usual focus of contemporary sociolinguistics).

In order to be more concrete, we shall examine one linguistic example in more detail below, but let us first note that the other sorts of culture studied in the humanities also can (and do) operationalize indications of common culture. Archaeology studies patterns of shared material culture,¹³ and some history likewise focuses on levels of culture such as technology and diet, which are likely to submit to operationalizable empirical techniques, and the arts are rich with studies of differences in poetic, musical, and architectural technique. Given a shared interest in cultural variation, and a shared opportunity to gauge cultural variation empirically, these disciplines are in an excellent position to collaborate in studies of cultural variation and cultural transmission, and we turn to this below. (Please do not understand this as a plea to view any of the disciplines as concerned exclusively with cultural variation, but rather as an appeal to consider what aspects might

be considered this way fruitfully.)

Culture must be transmitted by social contact, so it is plausible that the various expressions of culture might follow similar lines, being similarly affected by geography, family ties, and communication opportunities, among others.

3.1 Linguistic Variation

There are many sophisticated ways of studying linguistic variation, but one very simple technique is useful here for illustration. We take the example from the study of dialect geography, which studies how language varies in space. The language level involved is that of vocabulary, which is easily understood, and also reflects cultural influence in immediately apparent ways. For example, the word ‘stoop’ is used in the northeast of the US to refer to the stairs leading to the entryway to a house. The word is a borrowing introduced by Dutch immigrants (cf. Dutch *stoep*, ‘sidewalk’), and its distribution reflects their settlement patterns. This sort of data is available in dialect atlases for many language areas, and often in large quantity.

The basic idea for a more general analysis is due to Jean Séguy and is very simple: data atlas field workers recorded the responses to questions aimed at detecting common vocabulary for a range of dialect sites.¹⁴ We then compare each pair of sites, recording how many answers are the same and how many are different. For this purpose we ignore questions for which there is no answer at one or both of the sites, treating the categories of ‘not asked’ and ‘no response’ both as missing data (see below). The proportion of answers that is the same might be referred to as the LEXICAL PROXIMITY of the sites and the proportion of answers that is different is the LEXICAL DISTANCE. For example, given the data in the table below, we should conclude that

there's a lexical distance of 0.25 between Brownsville and Whiteplain since 75% of their responses was the same for the fields for which responses are available, and 25% were different.

Site	Vocabulary Item				
	<i>dog</i>	<i>hat</i>	<i>horse</i>	<i>toilet</i>	<i>smallest finger</i>
Brownsville	<i>dog</i>	<i>hat</i>	<i>horse</i>	<i>bathroom</i>	<i>pinkie</i>
White Plain	<i>dog</i>	<i>cap</i>	<i>horse</i>	<i>bathroom</i>	—

Naturally, several refinements and alternatives of this very basic technique are needed to handle the large reservoirs of dialect atlas data we have available in Linguistics. One needs to refine the basic procedure to find ways of recognizing different forms of the same word, of treating reports in which alternative word choices are given, and of analyzing data sets in which the questions vary (a bit). Alternative methods are conceivable for treating missing responses differently, for example, to regard the differing responses to the question about the *smallest finger* above as contributing to lexical difference (in the current calculation, it does not).

The data for the present study comes from the *Linguistic Atlas of the Middle and South Atlantic States*, restricting our analyses to the data collected by Guy Lowman, who was responsible for 71% of the 1,062 interviews.¹⁵ Our analysis followed the same technical choices explained in earlier work, which will not be repeated here.¹⁶

Given this basic data, we can analyze areas of relatively dense overlap, and transition areas where the gradient of overlap changes. Figure 1 sketches one analysis of the degree to which the vocabulary is shared, one which is divided into areas to facilitate comparison to traditional, non-computational studies. This sort of analysis of dialect data illustrated in Fig 1 is well accepted linguistically, and computational work now allows more exact formu-

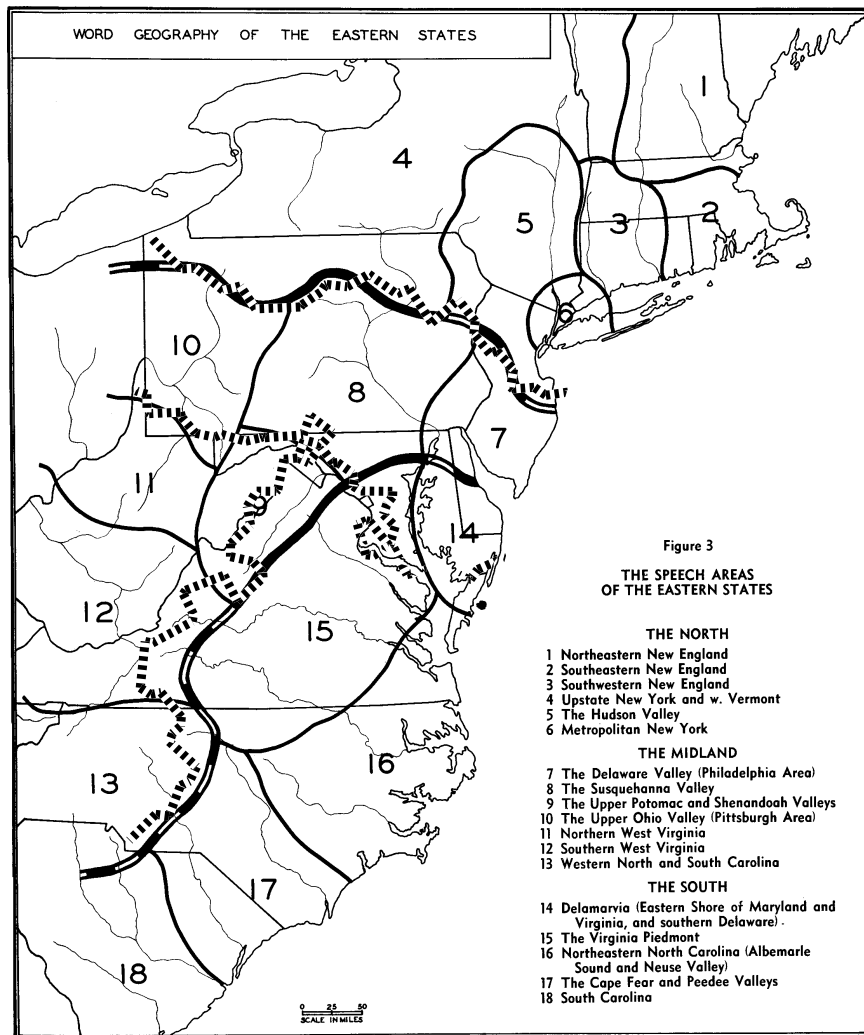


Figure 1: Hans Kurath's dialect division in *Word Geography*, 1949, based on the data we used, is shown in solid lines. Our four-way division is superimposed in the broken line. The agreement with Kurath is striking, but we also see a significant North-South division, much as other commentators have proposed.

lations of the methods, and, most importantly, more comprehensive analyses of larger data sets. Linguists routinely claim that linguistic culture mirrors other aspects of culture, and our computationally derived analyses of linguistic culture should be able to reflect on such claims with the benefits of their larger data reserves and more exactly formulated analyses.

We forego discussion of alternatives here¹⁷ in order to emphasize the point about linguistic and non-linguistic culture being parallel. The reasoning goes as follows: the areas which share a great deal of vocabulary almost certainly acquired this through a shared history. The people in these areas must have been in close contact to develop the vocabulary together, or must have moved from the same area, or must have been in common contact with the groups who introduced the vocabulary. But then it is likely that there was transmission not only of linguistic culture, but also of non-linguistic culture. Since this is the subject of other disciplines, we have an opportunity for crosstalk.

3.2 Crosstalk

The very simple analysis above can lead to very interesting crosstalk among the humanities disciplines. If we had similar collections of nominal data concerning history, material culture or folk art, it would be interesting to compare these to the linguistic analyses. These might be shared building patterns, common tools or dress, common religious services, children's songs, dances, or many other cultural artefacts or customs. Here we effectively follow in the footsteps of early dialectologists, especially in France, who were interested in whether the French *langue d'oc/langue d'oïl* split were mirrored in law, architecture and folk art.¹⁸ But we would not simply replicate earlier work if we adopted a quantitative, computational approach and formulated

our hypotheses accordingly. For example, we could then measure the degree to which these expressions of culture are associated.

3.3 Linguistics and Kinship

It is difficult to obtain the necessary data, but let us discuss one recent investigation which illustrates some of the possibilities.. Franz Manni and colleagues compare linguistic and genetic variation along the lines suggested here.¹⁹ Genetic variation is not normally thought of as cultural variation, but the choice of partner is clearly culturally mediated, as are the customs determining where couples and families stay—in particular whether newly married couples settle near the bride’s family or near the groom’s. So perhaps genetic variation gives us a toehold on a cultural area, that of kinship, after all.

Whether or not one accepts the idea of genetic variation as reflecting cultural patterns, there remains an interesting question of the degree to which linguistic and genetic differences pattern similarly, and the techniques for analyzing the similarity of the patterning are the same here as they would be for clear cases of cultural difference.

As noted above, the leading hypothesis on which this sort of analysis is based is that culture of all sorts spreads via the same mechanisms. The hypothesis, formulated as generally as this, is undoubtedly false in some circumstances, but it is worthwhile following it up in more detail, to try to understand which sorts of culture do follow similar lines, and which do not, and why.

In the concrete case at hand, Manni shows that linguistic and genetic differences correlate moderately ($r = 0.4$). This result, by itself, might suggest that linguistic culture is passed on through families, at least to some

extent. But in a multiple regression design, Manni also demonstrates that the “hidden variable” geography is entirely responsible for the modest correlation between genetic and linguistic differences. This technical correction likewise suggests a correction in interpretation: in general, people have acquired both their genetic makeup and also their speech patterns from people who are quite nearby.

Of course the opportunities for interdisciplinary investigations go much further. Besides examining the areas of relative linguistic similarity as we do above in Fig. 1, we might observe other indications of cultural contact between settlements, for example, regular trade connections, long-term service conducted in one town by someone from another, or marriage records. Rather than try to complete this list, we simply state that we hope that experts in other humanities disciplines might fill it in further. What candidate cultural markers are there whose distribution we can examine?

It would be exciting to compare work in other areas about cultural distributions—for example, those involving music, diet, family and church customs, tools and technologies, stories, and many more. Naturally we should wish to examine distributions not only with respect to geography, but also with respect to other candidate influences. Through an examination of many such distributions we may hope to learn more concretely how culture is transmitted and how.

3.4 Cultural Dynamics

In addition, it is instructive to examine the spread of culture from a more abstract perspective. Here we examine questions such as how important geography is in the determination of cultural similarity, and, by reasoning from current distributions back to plausible mechanisms of diffusion and

differentiation, we enable questions about cultural dynamics. Figure 2 provides a view of the American vocabulary data from two more abstract, and complementary perspectives.

First, we examine lexical distance (computed from the same data as before) as a function of geographical distance in the graph on the left hand side of Fig. 2. We note that the data is quite noisy, reflected in the wide cloud of points. So the cultural differences we encounter are not extremely orderly, at least not with respect to geography. Further, there is a clear positive slope, reflecting the increasing lexical distance of increasingly distant settlements, as expected. If we experiment with regression models, we find that we explain lexical distance best as a linear ($r = 0.670$) or logarithmic ($r = 0.664$) function of geography, which suggests that geography is playing a major role. Naturally, a large role for geography is not incompatible with there being other important forces, but alternative candidates will most likely be weaker than geography, or will overlap with it (collinearly).

Second, and incidentally, we see very similar curves in population genetics where genetic differences are plotted against geographic distances. Population genetics view graphs such as those in Figure 2 as an instance of what they call “isolation by distance”.²⁰

Third, the linear/sublinear form of the dependence of lexical distance on geography suggests a fairly gentle dynamic behind the diffusion of lexical variation.²¹ Strongly differentiating forces would suggest a steeper rise in the regression curve.

We experiment with a second graph to complement the first. To obtain it, we have first calculated, for each site s , the degree of correlation $r_s(d, l)$ between geographic distance d and lexical distance l for all of the other sites in the data set. We then plot $r_s(d, l)$ as a function of d for all sites s . Finally,

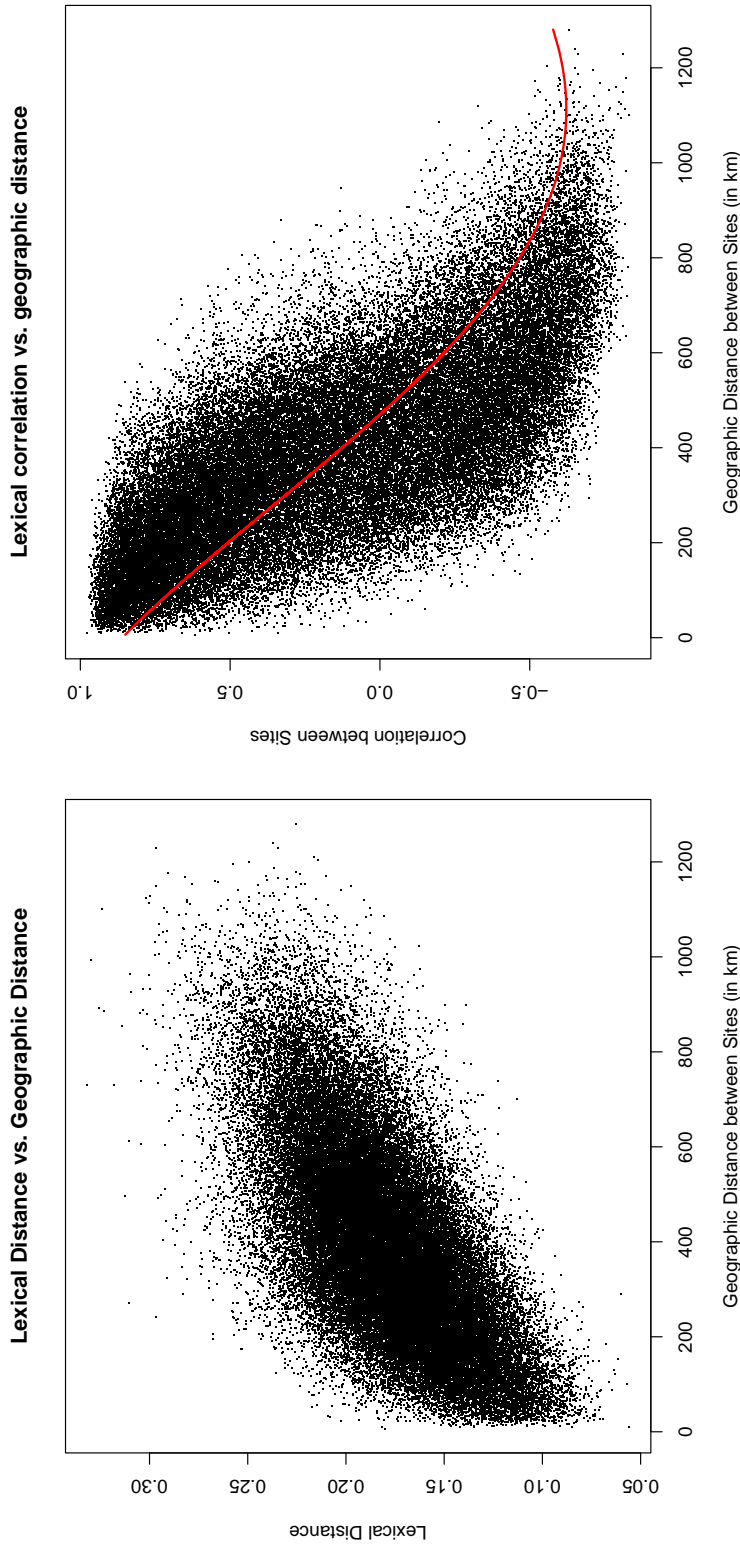


Figure 2: Lexical distance in LAMSAS as a function of geography (left), and the correlation of lexical distance as a function of geography (right). Note that we regard the negative correlations as an artefact of the variable being probed (lexical distance), and the zero as (roughly) the extent of influence. Both maps are based on the analysis in Nerbonne and Kleiweg (2003).

we perform a regression analysis, which we use to draw the slope of the curve shown on the right side. The negative slope of the regression line is naturally expected, since lexical distances will vary more as distance increases. We note in particular that the regression line reaches zero at approximately 400 km, meaning that we can no longer predict the lexical distance between a pair of sites simply by knowing the lexical distance of other sites separated by the same physical distance. (The negative correlation found among the more distant sites is essentially an artefact of our using distances as the dependent variable. Whenever we examine the correlations of linguistic (vocabulary) distances with respect to two distant sites, there will be a large number of sites to compare, most of which are much closer to one site than another.) The interesting supplementary information we obtain in the autocorrelation graph is an estimation of the geographic extent of the influence exerted by one town on another. Because the regression line is drawn on the basis of these relatively distant towns as well, we do not report the r resulting from the regression line—the more distant towns are mathematically influential, but as we have argued, linguistically irrelevant.

4 Conclusions and Prospects

Cultural transmission and cultural diffusion are core topics in several humanities disciplines, and we would like to know whether the various foci of study—material culture, linguistic culture, religious culture, etc.—follow similar patterns. This essay illustrates how the study of linguistic culture has benefited from the use of computational techniques, and it invites colleagues in other humanities disciplines to think along these lines. In order to be convincing about the usefulness and interest in these techniques, we have discussed a concrete computational analysis of linguistic variation, but

it is important to abstract away from the linguistic details in order to accept the invitation.

For example, we have emphasized the geographic distribution of culturally mediated differences in this paper, and we have ignored the very important temporal perspective. There is a completely mundane reason for this, namely, that data on the geographic distribution of linguistic variation is plentiful (in the form of dialect atlases), while data spanning a long period is rather more difficult to find. Where historical data is available in quantity, it is an attractive target for analysis.²²

Furthermore, we have ignored the “data bottleneck” here, the fact that it is still difficult to find data which has been collected, and to obtain data once it has been found. Fortunately, humanities organizations are aware of this, and have begun discussing the compiling of catalogs of available datasets, their technical maintenance in the face of evolving standards and practice, their accessibility, the need to respect confidentiality of respondents, as well as many other important issues. We need to collaborate in these efforts as researchers, and to plead for their importance to our funding agencies.

Many have spoken more recently of a “crisis in humanities”.²³ Without pausing to attempt to evaluate how serious that crisis is, and certainly without pretending to offer solutions for it, let us stress that co-operative research efforts such as the one advocated above have the potential to excite colleagues in the humanities and in the rest of academia. The effort at “crosstalk” will at the very least require us to become more familiar with each other’s work. We have argued above that this sort of crosstalk will provide opportunities for us to study essential themes from a variety of perspectives, and that is the central argument of the paper. But note that this should make us better spokesmen for the humanities as a whole, since we

will need to engage each other more broadly if interdisciplinary efforts are to succeed.

Notes

¹We are indebted to the Netherlands Organization for Scientific Research (NWO) for support (“Determinants of Dialect Variation”, 360-70-120, P.I. J.Nerbonne), to Peter Kleiweg for programming, and to anonymous referees for treating an earlier version of this rather roughly.

²Willard McCarty, *Humanities Computing* (Houndmills and New York: Palgrave MacMillan, 2005).

³See issues of *Computers and the Humanities* or *Literary and Linguistic Computing* for these views. University administrations are particularly fond of understanding computing in the humanities primarily as computer-assisted instruction, or, worse, as “computer literacy”.

⁴John Nerbonne, ‘Computational Contributions to the Humanities’, *Linguistic and Literary Computing*, 20/1 (2005): 25–40. Invited plenary talk to joint ACH/ALLC meeting, June 11, 2004, Gothenburg. See www.let.rug.nl/~nerbonne/paper.html.

⁵Gertjan van Noord, ‘Error Mining for Wide-Coverage Grammar Engineering’, in *Proc. of 42nd Meeting of the Association for Computational Linguistics* (Barcelona: ACL, 2004), pp. 446–453.

⁶Wilbert Heeringa, *Measuring Dialect Pronunciation Differences using Levenshtein Distance*, Ph.D. thesis, Rijksuniversiteit Groningen (2004).

⁷Wilbert Heeringa & Charlotte Gooskens, ‘Norwegian Dialect Examined Perceptually and Acoustically’, *Computers and the Humanities*, 37/3 (2003): 293–315.

⁸Peter Berglar, *Wilhelm von Humboldt in Selbstzeugnissen und Bilddokumenten* (Reinbek bei Hamburg: Rowohlt, 1970); Wilhelm Dilthey, *Einleitung in die Geisteswissenschaften: Versuch einer Grundlegung für das Studium der Gesellschaft und der Geschichte* (Stuttgart: Teubner, ¹1883, 1973)

⁹Stephen Levinson, Sotaro Kita, Daniel Haun & Björn Rasch, ‘Returning the tables: Language affects spatial reasoning’, *Cognition*, 84 (2002): 155–188.

¹⁰Steven Pinker, *The Language Instinct* (New York: W. Morrow and Co., 1994).

¹¹Humanities scholars who look to Hegel for the foundations of *Geisteswissenschaften*

would include law as part of the expression of *Geist*, ‘spirit’.

¹²Georg Henrik von Wright, *Erklären und Verstehen* (Frankfurt: Athenäum, 1974).

¹³Peter Bellwood and Colin Renfrew in fact anticipate my argument below for the case of archaeology and linguistics. Peter Bellwood & Colin Renfrew (eds.), *Examining the Farming/Language Dispersal Hypothesis* (Oxford: Oxbow, 2002).

¹⁴Jean Séguy, ‘La relation entre la distance spatiale et la distance lexicale’, *Revue de Linguistique Romane*, 35/138 (1971): 335–357.

¹⁵The data is made available online by Prof. William Kretzschmar (at <http://us.english.uga.edu/lamsas/>).

¹⁶John Nerbonne & Peter Kleiweg, ‘Lexical Variation in LAMSAS’, *Computers and the Humanities*, 37/3 (2003): 339–357. Special Iss. on Computational Methods in Dialectometry ed. by John Nerbonne and William Kretzschmar, Jr..

¹⁷Interested readers may consult earlier work, such as Heeringa, *Measuring Dialect Pronunciation Differences*, and John Nerbonne, Wilbert Heeringa & Peter Kleiweg, ‘Edit Distance and Dialect Proximity’, in David Sankoff & Joseph Kruskal (eds.), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, 2nd edn (Stanford, CA: CSLI, 1999), pp. v–xv.

¹⁸Jules Gilliéron, *Atlas linguistique de France* (Paris: Champion, 1902–10).

¹⁹Franz Manni, Wilbert Heeringa, Bruno Toupance & John Nerbonne, ‘Do Surname Differences Mirror Dialect Variation?’ *Human Biology*, 80/1 (2008): 41–64.

²⁰Mark A. Jobling, Matthew E. Hurles & Chris Tyler-Smith, *Human Evolutionary Genetics: Origins, Peoples and Diseases* (New York: Garland, 2004), pp. 142–143.

²¹John Nerbonne & Wilbert Heeringa, ‘Distributions of Linguistic Variation Reflect Dynamics of Differentiation’, in Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in Search of its Evidential Base* (Berlin: Mouton De Gruyter, 2007), pp. 267–297.

²²April McMahon & Robert McMahon, *Language Classification by Numbers* (Oxford: Oxford University Press, 2005).

²³Michael Bérubé & Cary Nelson, *Higher Education Under Fire: Politics, Economics, and the Crisis of the Humanities* (New York: Routledge, 1995); Stanley Chodorow, ‘Taking the Humanities off Life Support’, *American Council of Learned Societies Occasional Paper*, 40 (1997), avail. at www.acls.org/aclpubs.htm; Robert Weisbuch, ‘Six Proposals to Revive the Humanities’, *Chronicle of Higher Education*, 26 (March, 1999): B4–5;

George Steiner, 'The Humanities: At Twilight?' *Poetry Nation Review (P.N.Review)*,
25/4 (March–April, 1999): 18–24.