Chapter 9

Liberating Dialectology

J. K. Chambers
University of Toronto

Until the 1980s, dialectologists simply accepted the fact that their investigations would produce "a superfluity of data", as Kretzschmar, Schneider & Johnson (1989: v) pointed out. "Even smaller surveys have had to settle for selective analysis of their data," they went on to say, "because the wealth of possibilities overran the editors' time and the human capacity for holding in mind so much information at once." That changed in the 1980s, when computers offered relief from the sorting problem, and advances in multivariate statistical methods began making inroads into the analysis of complex corpora. One of the disciplines that emerged at the intersection of computation and quantitative analysis was dialectometry, merging the joint progress in both areas.

1 The Groningen Initiative

What is sometimes missing in dialectometry is the humanistic interpretation that is so integral to the study of language, this most human of all attributes. Our sophistication in applying quantitative models to large corpora has greatly outpaced our capability for evaluating our results in terms of variable strength and social significance. One of the initiatives of John Nerbonne and his Groningen protégés has been, in Martijn Wieling's phrase (2012: 3), "increasing the dialectology in dialectometry". Dialectology in the twenty-first century is quantitative, variationist and social. Dialectometry applies quantitative models to large corpora, often to archival dialect data that were collected years before we had the capability for sorting the profusion of information let alone analyzing it. We now know more about those data than the dialectologists who gathered the data could ever imagine.

The great strength of dialectometry so far has come from the rigor it has been able to impose on the geographical distribution of linguistic variables. For the first hundred years after Georg Wenker inaugurated the systematic study of dialect in Germany in 1876, dialectology was largely qualitative and impressionistic. Its few

governing concepts had their roots in common sense rather than empirical testing. A prime example is the "dialect continuum" (Chambers & Trudgill 1998: 5-7), the idea that "dialects on the outer edges of a geographical area may not be mutually intelligible, but they will be linked by a chain of mutual intelligibility". The concept has its origin in the common observation that people who live near to one another normally speak more similarly than people who live further away. The idea of the dialect continuum was routinely invoked, selectively corroborated, and never problematized. The first empirical test of the concept came decades after it had become common coin. Heeringa & Nerbonne (2001) isolated a string of 27 Dutch towns and villages and calculated the aggregate pronunciation difference from one to the next (the "Levenshtein distance"). The dialect continuum survived their scrutiny but in a greatly nuanced conceptualization, with points of relative coherence and disruption correlated to some extent with cultural boundaries. The simple unexamined assumption of the dialect continuum emerged as a layered concept, and we can now imagine it being further examined in terms of internal forces, social and linguistic explanation, and comparative typology.

The next step, a crucial one surely, will be to rationalize gradations in the continuum by correlating them with social factors such as population, mobility and services. The Groningen initiative actively seeks this better balance. In a state-of-the-art summary, Wieling & Nerbonne (2015: 258) quote Lord Kelvin: "When you can measure, ... you know something; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind." Kelvin's axiom is nicely illustrated, it seems to me, in applying actual measures to the dialect continuum, which has given us a richer concept than ever before. But Wieling and Nerbonne remind us that there is a further step. They go on to say that Kelvin "never suggested that measuring was sufficient", a succinct reminder for all dialectometrists. The axiom expresses a crucial step for students of language. Measuring dialect has been a liberating force in dialectology.

2 The pioneer

A unique measure of how far we have come is afforded by looking backward to some moments when a few solitary souls recognized that dialectology could be dialectometric — quantifiable (not qualitative), relative (not absolute) and dynamic (not static). I would make the case that Jan Czekanowski (1882–1965) was the father, albeit an estranged one, of contemporary dialectometry. Indeed, he was practising dialectometry almost 45 years before Jan Séguy coined the term dialectométrie (1973). I will outline Czekanowski's accomplishment, and document his momentary influence when his methods were applied by a couple of Americans to a purportedly intractable case of dialect heterogeneity. Czekanowski's quantitative studies of dialect made little impression in his lifetime, and ultimately his insights were bypassed. Looking back reveals something about the core values of our discipline. Some fifty years after his death, dialectology has been reformed in exactly the terms he envisaged.

Jan Czekanowski is always identified as a Polish anthropologist but he was affili-

Figure 1: Jan Czekanowski.

ated with institutions in many nations (Payne 2014-2015), sometimes without physically moving. He was born in Głuchów, Poland, in 1882, and he began his primary education in Warsaw but completed it in Latvia. In 1902, he attended the University of Zurich and studied under the physical anthropologist Rudolf Martin (Oetteking 1926). Anthropological fieldwork took him to the Royal Museum in Berlin and to the Congo in Africa. He published the materials he collected in Africa in five volumes in 1910, while he was curator of the Ethnology Museum in St. Petersburg, Russia. In 1913, he became professor of anthropology at the University of Lwów, which was in Poland at the time (and in Ukraine from 1939). The photograph is probably from his early years at Lwów. He joined the University of Poznan in 1937 and retired there in 1946, when he was 64.

Czekanowski is remembered heroically because he convinced German "race scientists" in 1942 that the Karaim, a Polish-Lithuanian ethnic group that practised Judaism and used Hebrew as their liturgical language, were Turkic, not Semitic (`https://en.wikipedia.org/wiki/Jan_Czekanowski`). The Karaim were accordingly spared from the Holocaust.

Czekanowski spent much of his academic life working on the racial categories of the human species, a predilection of physical anthropologists of his day, and judging by the sources I have found (and cited here) his abiding reputation among anthropologists is based on that work. Among linguists he is not well known nowadays, and the American linguists who noticed his work in the 1950s were affiliated with anthropology departments, a common affiliation at the time. Czekanowski's achievement in linguistics, in retrospect, was no mere dalliance. He tried to establish the genetic and

cultural ties among branches of language families, that is, between dialect groups, and, against the grain of the times, he sought quantitative evidence for determining the relative strength of those ties. This line of research apparently started for him with inquiries that skirted linguistic matters. In the 1920s, he began with attempts at classifying cultural relatedness of ethnic groups by counting the cultural features they shared. He soon realized that shared linguistic features provide a solider basis. In 1927, he compared Polish dialects based on the morphological features they shared. In 1928, he broadened his purview and compared Indo-European dialects using the same methods. In 1929, he narrowed his view to Slavic dialects by the same method.

Czekanowski's method seems surprisingly sophisticated when we recall that he was working at a time when very few anthropologists and even fewer linguists counted anything at all. He began by compiling a list of linguistic features, and then for each pair of dialects he calculated correlation coefficients using a formula known as Q6, based on four factors: (1) the number of features present in both, (2) the number of features absent in both, (3) the number present in the first but absent in the second, and (4) the number absent in the first but present in the second.

He then arranged the dialects on a matrix according to their correlation coefficients. Figure 2 shows the pairwise comparisons of the Slavic dialects (Czekanowski 1931). It is a marvel of quantitative methods in its anticipation of multivariate statistics and a feat of calligraphy in its anticipation of computer graphics. The black squares show near-perfect correlations (> .80) and the groupings of black squares represent branches of the family tree as a kind of intuitive cluster analysis. He uses Old Church Slavonic as the base (= *Starocerkiewnoslow* in Figure 2, an abbreviation of *Starocerkiewnoslowianski*, according to my colleague Alexei Kochetov). By tracing coordinates on the axes of the figure, one sees, for instance, that Slovak and Czech (Slowacki and Czeski) share many features (> .80) and that Slovak and Malorussian (= *Małoruski*, a philological name for old Ukraine) share almost none (< .10). Despite its apparent complexity, the figure is easy to read and very revealing.

It is tempting to linger over Czekanowski's diagram, admiring its elegance and its perspicacity. Indeed, those were the features that captivated the American anthropologists Alfred L. Kroeber and Douglas Chrétien, who praised it in esthetic terms (1937: 84): "If the symbol values are chosen judiciously, the diagram becomes an exceedingly effective and rapidly grasped representation of the stronger relationships, wherein the salient features of the classification force themselves upon the eye and the mind through the automatic clustering of symbols." But Czekanowski was apparently too far ahead of his time and his work made no immediate impact.

3 Spreading the word

Kroeber and Chrétien's advocacy of Czekanowski's methods did, however, make an impression in the more adventurous reaches of dialect study, if only momentarily. As American dialect research moved away from the long-settled Atlantic coast into the heterogeneous interior, the traditional atlas-oriented practitioners found themselves at a loss to identify patterns. Two of those practitioners, Alva Davis and Raven

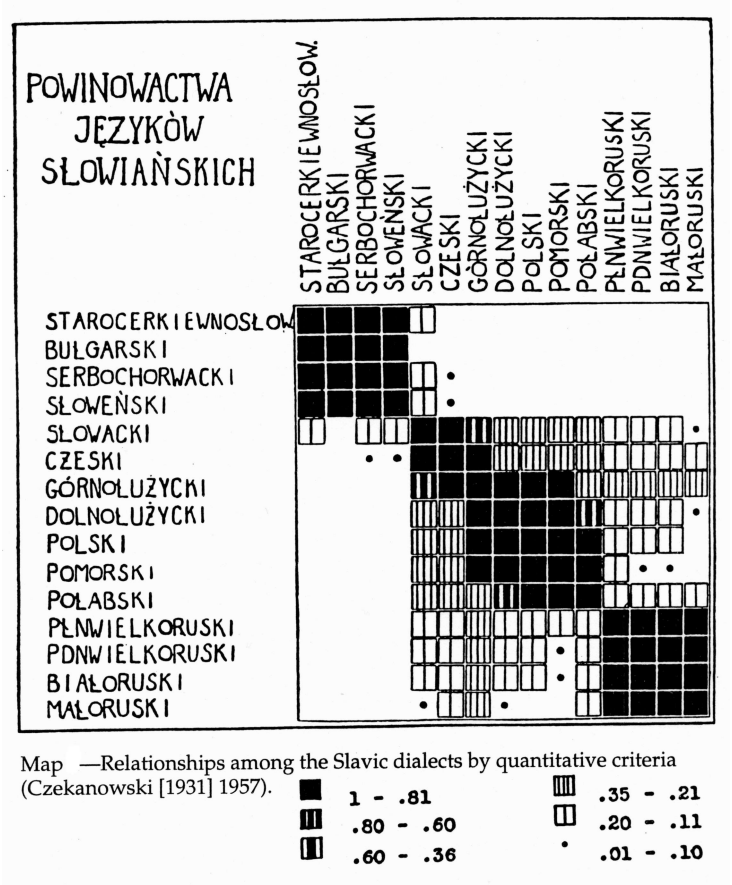

Figure 2: Czekanowski's quantitative arrangement of ancient Slavic dialects (1931).

McDavid, admitted defeat in a widely read article on "transition areas" in the journal *Language*. They wrote, "One is at a loss to give convincing reasons for the restriction of some items and the spreading of others" (1950: 186). In reaction, two of Chrétien's students, David Reed and John Spicer, offered a rebuttal in the same journal (Reed & Spicer 1952: 348), pointing out that "the speech patterns of transition areas grow much clearer when viewed as quantitative rather than qualitative phenomena", an axiom that two decades later would become the mantra of sociolinguistics and dialectology.

Reed and Spicer applied Czekanowski's method to Davis and McDavid's elicited features from ten subjects in five towns, and compared each pair of speakers using the Q6 formula to derive correlation coefficients. The quantification revealed some rough geographical generalizations, such as: the closer the town, the higher the coefficient, and the two southern towns are most similar to one another. But in the

end their mapping schema was very complicated and unrevealing because they did not choose a base dialect but presented all five as bases, and the variable patterns were hardly discernible because they required making inferences from five-way relationships. Their analysis definitely fails Kroeber's esthetic test: the schema does not provide a "rapidly grasped representation of the stronger relationships" and the "salient features" do not "force themselves upon the eye and the mind". With hindsight, we know that retaining geographical mapping obscured the patterns rather than revealing them, and bivariate statistics, the pairwise comparisons, required too many inferences for impressionistic interpretation.

Fifty years later, I applied multivariate correspondence analysis to Davis and McDavid's data matrix and discovered the pattern fairly readily. Figure 3 places the ten speakers in quadrants according to their shared features. They clearly fall into three clusters, with P1 and P2 in one quadrant, five others together in the lower quadrant, and a messy little group of three spread along the left side. Identification of Davis and McDavid's dialect features shows that they are characteristically either Northern or Midland in their point of origin. Speakers in the top cluster (P1, P2) use essentially Northern features, and speakers in the bottom cluster (O2, V1, V2, US1 and US2) use essentially Midland features. The messy group along the side is mixed, a transitional group that uses some Northern and some Midland features.

Although the statistical input for the correspondence analysis encoded no geographic information whatsoever, the subcategories cohere strikingly to their relative locations on the map of northwestern Ohio: speakers P1 and P2, essentially Northern in their dialects, live in the northernmost town, and four of the five essentially Midland speakers live in the two southern towns. The mixed dialect speakers live in the two towns in between.

Why does geographical distance match statistical distance? It is hardly a mystery. As every dialectometrist knows, people who live close together tend to speak more like one another than people who live further away. We have come a long way since Jan Czekanowski, but dialectometrists like John Nerbonne have helped to reshape the study of dialects in exactly the terms that Czekanowski envisioned. Dialect distances, like geographic distances, are measurable things. Having established that beyond a doubt, dialectometrists are now taking the next step and increasing the dialectology in dialectometry.

References

Chambers, J. K. & Peter Trudgill. 1998. *Dialectometry.* 2nd edn. Cambridge, UK: Cambridge University Press.

Czekanowski, Jan. 1931. *Róznicowanie sie dialectów prastowinskich w swietle kryterjum ilociowego [Differentiation of ancient Slavic dialects...].* Prague: First Congress of Slavic Dialectology, 1928.

Davis, Alva & Raven I. McDavid. 1950. Northwestern Ohio: a transition area. *Language* 26. 186–89.

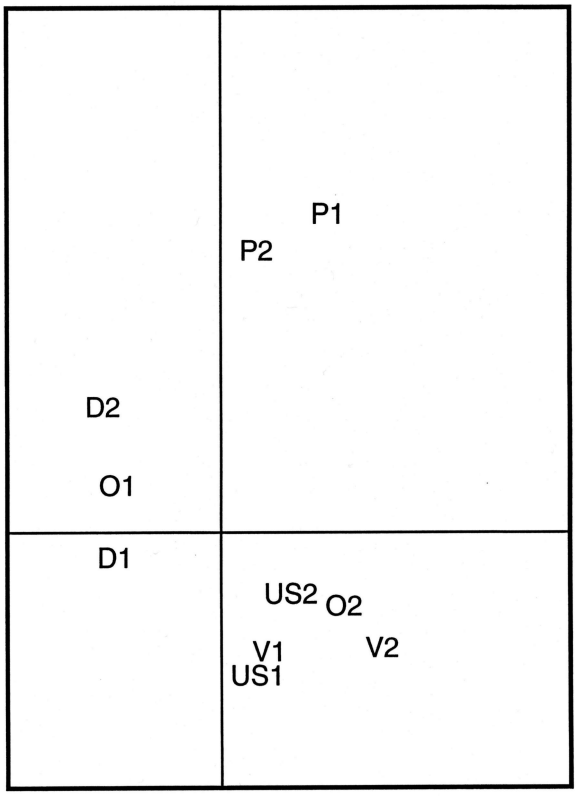

Figure 3: Correspondence analysis of ten subjects (P1, P2, etc.) from five towns (P, D, O, US and V) in northwestern Ohio (Chambers & Trudgill 1998: 144-48).

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

Kretzschmar, William, Edgar Schneider & Ellen Johnson (eds.). 1989. *Journal of English Linguistics* 22, special edition: *Computer Methods in Dialectology.*

Kroeber, Alfred L. & C. Douglas Chrétien. 1937. Quantitative classification of Indo-European languages. *Language* 13. 83–103.

Oetteking, Bruno. 1926. http://onlinelibrary.wiley.com/store/10.1525/aa.1926.28.2.02a00070/asset/aa.1926.28.2.02a00070.pdf?v=1&t=iph1ihz6&s=4b5bff98f70837980b38d3c478c4d4aeb73c1229 (June 2016).

Payne, Stephen. 2014-2015. *The Info List — Jan Czekanowski.* http://www.theinfolist.com/php/SummaryGet.php?FindGo=Jan%5C%20Czekanowski (June 2016).

Reed, David W. & John L. Spicer. 1952. Correlation methods of comparing dialects in a transition area. *Language* 28. 348–59.
Séguy, Jan. 1973. *Atlas linguistique de la Gascogne.* Vol. 6: *Notice explicative.* Paris: Centre national de la recherche scientifique.
Wieling, Martijn. 2012. *A quantitative approach to social and geographical dialect variation* (Dissertations in Linguistics 103). Groningen: University of Groningen.
Wieling, Martijn & John Nerbonne. 2015. Advances in dialectometry. *Annual Review of Linguistics* 1. 243–264.