Berkeley Linguistics Society. 2013. 245-258. http://dx.doi.org/10.3765/bls.v39i1.3884
Published by the Linguistic Society of America

Measuring Linguistic Distance in Athapaskan

CONOR SNOEK, CHRISTOPHER COX
University of Alberta

Introduction

The following paper is a contribution to the study of language relationships in Athapaskan.[1] Despite many years of dedicated scholarship and research on Athapaskan languages, the field still lacks a definitive sub-grouping. Instead, the grouping of languages has relied largely on classification primarily on the basis of geographic and cultural proximity. This grouping no doubt has its merits, and it seems indeed likely that the larger geographic divisions—into Pacific coast, Southern (or Apachean), and Northern branches—also reflect longer and sustained historical relationships among their constituent speaker communities. At a higher level of resolution among languages within shared geographic areas, however, this form of grouping remains unsatisfactory. The reasons for this difficulty in grouping Athapaskan languages are outlined in more detail in the following section.

Since the last attempts to establish sub-grouping in Athapaskan (cf. Mithun 1999), the field has benefited from a wider availability of data through published grammars, dictionaries, and articles, as well the greater ease of accessibility to digital archives containing field notes and other relevant primary materials. Additionally, computer-aided techniques of data analysis have been developed, making it possible to treat larger sets of data and more readily visualize these data with graphs and maps. Here, we present the results of applying statistical clustering and mapping techniques in grouping Athapaskan languages on the basis

[1] As Rice (2012: 249) notes, the terms "Athapaskan" and "Athabaskan" (and further variants thereof) have long been used to refer to this language family, but the term "Dene" has also come to be favoured more recently in some communities. These terms are used here interchangeably.

of phonological similarity. We want to argue for the usefulness of applying such techniques to Athapaskan, and point toward future work that will integrate greater and more varied bodies of data that we believe will lead to a reliable sub-grouping of Athapaskan languages and bring greater understanding of the history of the Athapaskan-speaking peoples.

1 Classificatory problems in Athapaskan

Athapaskan represents one of the largest Indigenous language families in North America, comprising approximately forty languages spoken from western Alaska to northern Mexico. While having one of the most extensive geographical ranges of any Indigenous language family in North America, the distribution of Athapaskan is not contiguous. Athapaskan languages appear in three distinct areal clusters: one on the Pacific coast, with a group of eight languages centered in present-day Oregon and California; another in the American southwest, Oklahoma, and Texas, representing seven Apachean languages; and a geographically larger group of 23 or more languages in northwestern Canada and Alaska, with the majority of these spoken in Alaska and the Yukon (Krauss and Golla 1981, Mithun 1999).[2]

Despite the considerable geographical separation that exists between these clusters, all Athapaskan languages share a recognizable typological profile, retaining the heavily prefixing polysynthetic verbal morphology and coronal-heavy phoneme inventories characteristic of the family as a whole. Notably, this linguistic conservatism holds even in cases of extensive historical contact with neighboring non-Athapaskans: Athapaskan languages on the whole show few signs of significant morphological, phonological, or lexical influence from non-Athapaskan sources (Sapir 1925:185). In general, the degree of differentiation encountered between Athapaskan languages suggests relatively recent division into these branches, perhaps as late as 500 B.C.E. (Krauss and Golla 1981:68).

The high degree of geographical dispersion between members of the language family, combined with the relatively low degree of linguistic differentiation between languages, has raised questions as to the internal classification of Athapaskan, both within and between the aforementioned geographical clusters. In the case of Pacific Coast Athapaskan, recent assessments have called into question the treatment of this grouping as resulting from a single historical wave of southward migration and subsequent linguistic diversification, rather than a loose geographical grouping of communities whose separation occurred prior to their entry into the region (Golla 2011, Spence 2013). By comparison, classifications of Southern Athapaskan have generally been treated in the linguistic literature as a single historical unit, with later differentiation into

[2] In some cases, both Krauss and Golla (1981) and Mithun (1999) group several related languages (e.g., Tahltan [tht], Kaska [kkz], and Tagish [tgx]) into a single unit, thus lowering their estimates of the total number of distinct Northern Athapaskan languages.

distinct languages. While relationships between languages in both of these clusters have been suggested to be amenable to comparative reconstruction, the same cannot be said of Northern Athapaskan. For these languages, Krauss and Golla (1981:68) argue, "linguistic relations [...] cannot be adequately described in terms of discrete family-tree branches," with isoglosses for historical changes not forming clear bundles, but rather cross-cutting one another in ways that prove problematic for coherent classification. As a result of essentially constant intergroup communication, Krauss and Golla (1981:68-9) propose that Northern Athapaskan be treated as a "dialect complex," with the "areal diffusion of separate innovations from different points of origin" both obscuring earlier idiosyncratic historical developments and undermining attempts to establish consistent subgroups on the basis of such criteria alone.

In sum, while little disagreement exists over a broadly geographical classification of Athapaskan languages into three main branches, the status of (and prospects for) further internal classification on the basis of shared historical innovations within these branches remain in question. Areal diffusion of linguistic features through networks of regular contact between neighboring Athapaskan groups in at least Northern Athapaskan presents a situation not unlike a traditional dialect continuum, where the linguistic boundaries between adjacent varieties are sometimes similarly blurred as a result of contact. Given this similarity and the general geographical orientation of Athapaskan language classification, it might be expected that methods developed to study areal linguistic variation and dialect classification may be of some service in approaching internal classification in Athapaskan, as well. Such methods and their application to Athapaskan are considered in greater detail below.

2 Dialectometric approaches

As noted above, the situation described by Krauss and Golla (1981) for Northern Athapaskan bears some similarity to problems found in the analysis of dialect continua in traditional dialectology. While dialectology offers many methodological options for the interpretation of complex linguistic geography, we concentrate here specifically on quantitative, multivariate methods drawn from recent research in dialectometry (Goebl 2006, Nerbonne et al. 2011). These methods aggregate substantial amounts of dialect data in order to facilitate large-scale comparisons in which contemporary statistical methods might be applied. This approach has several notable strengths: first, aggregating multiple linguistic variables has the potential, as Nerbonne et al. (2011) suggest, to "strengthen the signal of speaker provenance," highlighting significant trends in the patterning of isoglosses which might otherwise be overlooked in manual inspection of the same data or obscured by apparently contradictory differences between individual dialect features. Second, such methods encourage the use of aggregation and classification algorithms that can be replicated between studies, situating such research to benefit from a growing literature on the interpretation of such data and

from continued methodological advances in this area. Third, dialectometric methods profit from the increasing availability of computational resources for classification and visualization, allowing more data to be weighed in consideration when evaluating possible linguistic groupings than would otherwise be possible. All of these reasons present incentives for considering potential applications of dialectometric methods to Athapaskan classification.

In a dialectometric analysis, a distance measure is applied to a set of linguistic features for some number of languages, producing for each feature a square matrix of linguistic distances between all unique pairs of languages. In order to estimate the distances between Athapaskan language features, we used a simple Levenshtein distance, which computes the minimum number of insertions, deletions and substitutions required to transform one string into another (Levenshtein 1969). Difference is evaluated on a binary basis, producing a count of 1 if the characters are different and 0 if they were the same. In cases where two characters are distinguished by a diacritical mark only (e.g. for tone marking or aspiration), the distance is counted as 0.5. In this study, Levenshtein distances were calculated for each pair of phonemically transcribed word forms, as illustrated by the comparison of two words for 'back' (body part) in Dene Sųłiné /nené/ and Dena'ina /tʰanəq/ in (1) below. The total distance between these two features is given in the last column.

(1) Example of Levenshtein distance calculation

Dene Sųłiné	n	e	n	é		
Dena'ina	tʰ	a	n	ə	q	
	1	1	0	1	1	4

Once these distance matrices have been computed for all available words, the overall linguistic distance between two languages is then calculated as the average of the distances between all corresponding word pairs (Heeringa 2004:145). This results in a distance matrix that can be fed into a clustering algorithm to produce a dendrogram indicating language proximity. The algorithm used to calculate the distances between languages is implemented in the Gabmap application (Nerbonne et al. 2011). In order to compensate for the variability of outcomes from different clustering procedures, the stability of clusters can be checked against an analysis of the same data using Multi-Dimensional Scaling (MDS).

The data compiled for this study form part of an ongoing project at the University of Alberta that aims to build a database of linguistic, cultural, and biological information on the Athapaskan languages and their speech communities. The origins of this project lie in the work of Sally Rice and Jack Ives, who sought to bring together linguistic and archaeological information for research into Athapaskan prehistory, especially with a view to shedding light on the migration of Apachean peoples. Rice and Ives named the database the Pan-

Athapaskan Comparative Lexicon (PACL; Snoek 2012).[3] PACL is envisioned as a dynamically expanding project which will allow individuals from multiple communities and institutions to access and contribute information. At present, the most developed aspect of the database is a set of comparative lexical lists which have been annotated for morphological and semantic information.

Drawing on both published resources and unpublished field notes, we selected three lexical domains from PACL—kinship terms, numerals, and body parts—to serve as the basic set of comparative items for the Athapaskan languages included in this study. While we view this list as partial in the sense that the addition of more lexical domains will eventually be necessary for reliable comparison, we nevertheless consider the choice of these items a potential improvement over the strategy of using Swadesh lists, as these items represent culturally meaningful categories. In this respect, we are following suggestions made by Matisoff (1975:134) in his work on Tibeto-Burman, wherein he argued for the adaptation of the Swadesh list approach to the cultural context of the languages he was studying. We consider this a particularly fruitful approach for our case, especially because, with the possible exception of Haida, the membership of the Athapaskan language family is uncontroversial, with only the relationships of the member languages to each other remaining unclear. In the spirit of Matisoff (1975), then, we have sought to make the basis of comparison a culturally meaningful set of lexical items.

3 Exploring Athapaskan classification

For this analysis, 105 comparative lexical items were assembled from the three PACL lexical domains, with 52 body part terms (e.g., "finger", "heart", "teeth"), 30 kinship terms (e.g., F, M, FB, MB), and 23 numerals (e.g. "one", "two", "three", "two persons"). These items were sampled for 22 Athapaskan varieties representing 15 distinct languages. These languages were chosen to represent members of all three major geographical divisions for which adequate information in all three lexical domains was available, while taking care to include several well-documented dialect distinctions within particular languages (e.g., between the varieties of Dena'ina represented in Kari 2007) as a means of checking the ability of these methods to correctly identify such subgroups. (2) presents the geographical range of the varieties in the sample visually, while (3) provides additional information about each variety and the sources of information consulted on these lexical items.

[3] http://www.linguistics.ualberta.ca/en/Research/Projects/PanAthapaskanComparativeLexico.aspx

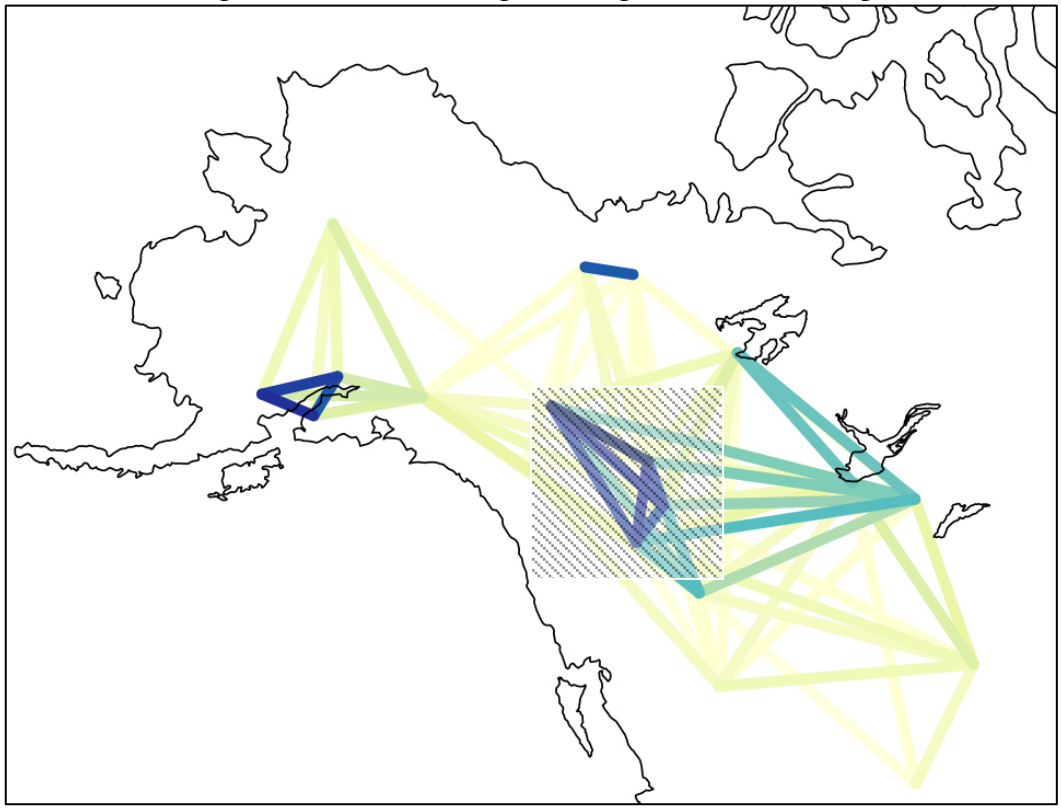
(2) Geographical distribution of the languages in the sample

(3) Languages in the sample, with varieties given in parentheses. Codes refer to ISO 639-3 language identifiers, while row numbers correspond to geographical points in (2) above.

#	*Language (Variety)*	*Code*	*Sources*
1	Ahtna	aht	Kari (1990)
2	Carrier (Central)	crx	Antoine et al. (1974)
4	Dena'ina (Inland)	tfn	Kari (2007)
5	Dena'ina (Outer Inlet)	tfn	Kari (2007)
6	Dena'ina (Upper Inlet)	tfn	Kari (2007)
7	Dene Sųłiné	chp	Elford and Elford (1998), Cook (2004)
9	Gwich'in (Gwichya)	gwi	GSCI and GLC (2005)
8	Gwich'in (Teetl'it)	gwi	GSCI and GLC (2005)
11	Jicarilla Apache	apj	Opler (1936), Phone, Olsen, and Martinez (2007)
12	Kaska (Frances Lake)	kkz	Kaska Tribal Council (1997)
15	Kaska (Good Hope Lake)	kkz	Kaska Tribal Council (1997)
13	Kaska (Liard)	kkz	Kaska Tribal Council (1997)
14	Kaska (Pelly)	kkz	Kaska Tribal Council (1997)
16	Koyukon	koy	Jetté and Jones (2000)
17	Navajo	nav	Young and Morgan (1987)
19	North Slave (Bearlake)	scs	Bloomquist (1978), Rice (1989)
18	North Slave (Mountain)	scs	Rice (1989), Kaska Tribal Council (1997)
20	Sekani (Kwadacha)	sek	Kaska Tribal Council (1997)
22	Southern Tutchone (Kluane)	tce	Tlen (1990)
21	South Slave (Katl'odehche)	xsl	Rice (1989), SSDEC (2009)
24	Tolowa	tol	Bommelyn (1995)
25	Tsuut'ina	srs	Cook (1984)

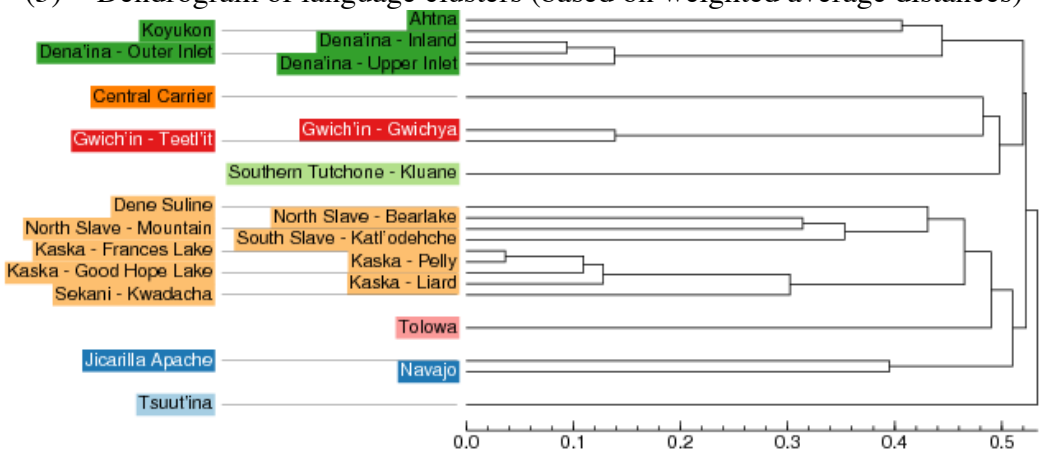
The lexical data were subsequently imported into Gabmap, which was used to compute Levenshtein distances for each of the lexical items. As Nerbonne et al. (2011) note, Gabmap allows for inspection not only of the distribution of individual lexical items, but also of the aggregate distances computed over the entire set of lexical items. These aggregate distances can then be visualized in several forms, including as a beam map, as seen in (4) below.

(4) Beam map, with darker lines representing closer relationships



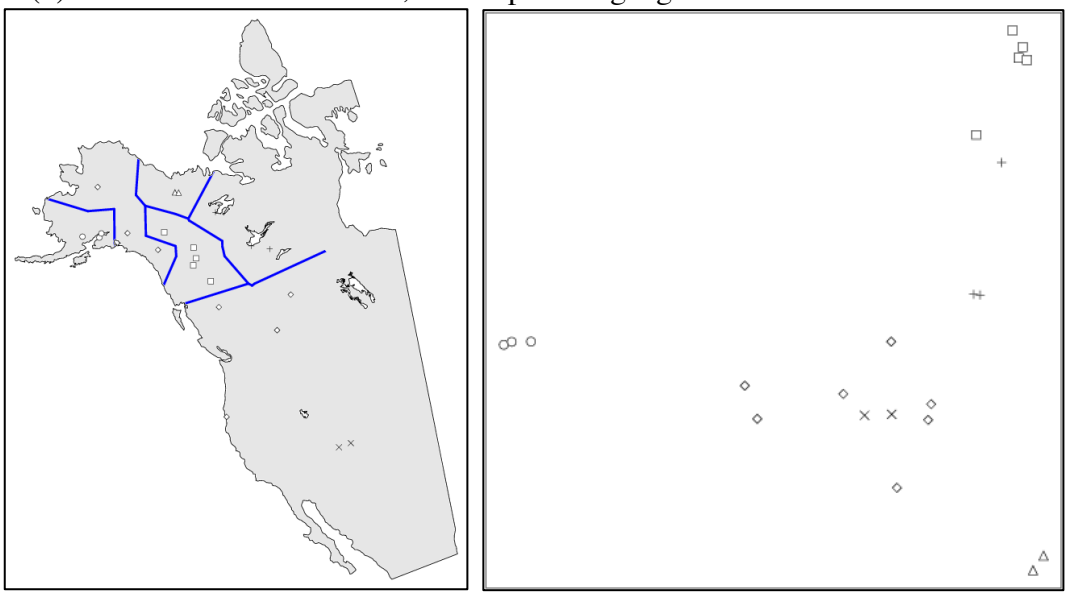
In (4), we can observe very dark lines connecting communities on the central Alaskan coast representing dialects of Dena'ina. The area marked by diagonal hatching is constituted by dialects of Kaska. To the north, dark lines connect the two dialects of Gwich'in. The algorithm identifies these dialect chains quite clearly. Furthermore, it is interesting to note the proximity of the Kaska dialect chain to the Slave languages to the east and Sekani to the south. This relationship is visible in the cluster dendrogram in (5) below.

(5) Dendrogram of language clusters (based on weighted average distances)



In the above dendrogram, Sekani and the four varieties of Kaska form one half of a larger cluster, with the Slave languages and Dene Sųłiné forming the other. Comparing the cluster validation in (6) with the above dendrogram, however, shows that the relationships between Slave and Dene Sųłiné are less tightly knit than among the Kaska dialects. The two Gwich'in dialects present another loose cluster, with Southern Tutchone forming a group of Athapaskan languages spoken in what is today the Yukon Territory and adjacent areas of the Northwest Territories. This is an interesting result, as Gwich'in has been traditionally viewed as being closer to the Alaskan languages, and Southern Tutchone is spoken in regions geographically much closer to the northern end of the Kaska dialect chain.

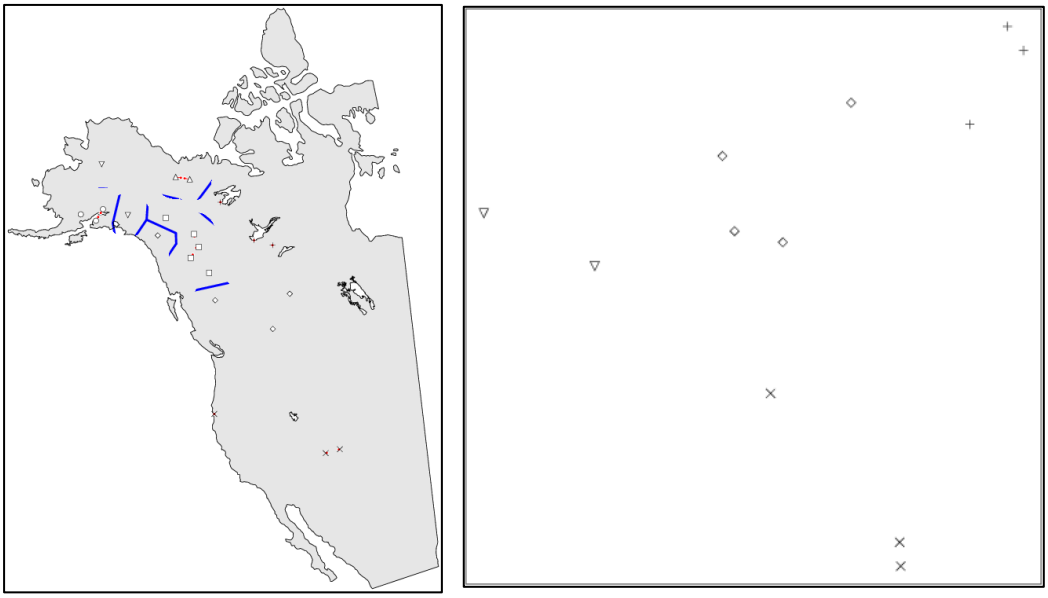
(6) MDS cluster validation, all sampled languages



The Alaskan languages form their own cluster, as would be expected both from studies with traditional methods as well as on the basis of regional association. Equally unsurprising is the coherence of the Apachean cluster (here, Navajo and Jicarilla Apache). Tolowa is isolated as the only representative of the Pacific coast languages. Finally, Tsuut'ina also forms its own branch, which is in line with prior classifications, where it is identified as a sharply-defined outlier among the (Northern) Athapaskan languages (cf. Krauss and Golla 1981:84). Comparison of the results of this clustering against an MDS analysis of the same data presented in the right panel of (6) reveals that only three of these clusters can be viewed as immediately reliable groupings. These clusters are the Kaska dialect chain, the Gwich'in languages, and the dialects of Dena'ina.

Leaving aside these more robustly attested groupings briefly to inspect the remaining languages in detail, we find evidence for a smaller cluster consisting of the three Slave languages, a weaker Alaskan subgroup made up of Ahtna and Koyukon, and a clear north-south division separating the representatives of the Pacific Coast and Southern branches from the Northern Athapaskan languages. These smaller clusters are represented graphically in (7) below.

(7) MDS cluster validation, all sampled languages except Sekani-Kaska, Dena'ina, and Gwich'in



It must not be forgotten, however, that this provisional sample represents only a third of the languages of the Athapaskan family, and that other relationships could emerge when further data are brought into the analysis. Indeed, the sparseness of representation of languages in the Pacific Coast and Southern branches may be expected to present a challenge for any form of general classification, whether based on manual comparison or aggregate analysis of phonological differences. Given the scope of the present sample, we consider these results to be reasonable and view them as promising enough to warrant further expansion to include both further lexical domains and additional members of the language family.

4 Prospects and conclusions

Although the results presented in this study are necessarily limited in scope, given the restricted size of the sample in terms of both languages and semantic domains, we nevertheless find them to provide sufficient motivation for continued investigation of the application of similar computational methods to outstanding problems in Athapaskan classification. Given the apparent complexity of the Northern Athapaskan situation in particular, it would seem important to identify methods which neither whitewash attested points of differentiation between varieties, nor allow individual points of deviation to exert undue influence on the overall classification under development. Inasmuch as the problem of linguistic classification is a multivariate one, so too should multivariate methods be considered that are capable of giving balanced attention to the full range of linguistic phenomena which form the empirical basis of classification.

In the case of dialectometric studies, quantitative, statistical methods and accompanying visualizations often serve this purpose, facilitating the identification of significant trends in the data even when seemingly opposing patterns are also attested. Yet, current tools for dialectometry are also capable of providing detailed information on the distribution of individual items, opening these data to further comparative analysis and to other forms of visualization and thus serving a range of quantitative and qualitative purposes.

While we have found these methods to be useful for Athapaskan, it bears noting that some arguments have been made against the use of Levenshtein distances in linguistic classification (Greenhill 2011). In the present study, we would argue that the application of this distance measure is not entirely inappropriate. As Greenhill (2011:693) notes, the apparent congruence of Levenshtein-distance-based sub-grouping with the results of traditional dialectology is likely due to the greater accuracy of Levenshtein distances between languages of relatively low phylogenetic difference. This would appear to be the case at least in most of Northern Athapaskan, where a dialect continuum-like configuration of varieties is reported; and arguably within the Southern and Pacific Coast branches, as well, given the relatively shallow degree of linguistic differentiation found in each. This may have contributed to the good approximation of our results to prior sub-groupings derived through traditional

methods. However, we do not claim to have produced a definitive classification, and view these results more as a stepping stone to further work on this formidable problem. Beyond the distance measures and clustering algorithms provided by dialectometric services such as Gabmap, the visualization of lexical and phonological data that such systems offer presents researchers with another excellent tool for the exploration of areal linguistic phenomena and linguistic classification.

References

Antoine, Francesca, Catherine Bird, Agnes Isaac, Nellie Prince, Sally Sam, Richard Walker, and David B. Wilkinson. 1974. *Central Carrier Bilingual Dictionary*. Fort Saint James, BC: Carrier Linguistic Committee.

Bloomquist, Chuck. 1978. *Slavey Topical Dictionary (Ft. Franklin Dialect)*. Unpublished ms.

Bommelyn, Loren. 1995. *Now You're Talking Tolowa*. Arcata, CA: Humboldt State University Center for Indian Community Development.

Cook, Eung-Do. 1984. *A Sarcee Grammar*. Vancouver: University of British Columbia Press.

Cook, Eung-Do. 2004. *A Grammar of Dëne Sųłiné (Chipewyan)*. Algonquian and Iroquoian Linguistics: Special Athabaskan Number, Memoir 17. Winnipeg: Algonquian and Iroquoian Linguistics.

Elford, Leon W. and Marjorie Elford. 1998. *Dene (Chipewayn) Dictionary*. Prince Albert, SK: Nothern Canada Mission Distributors.

Goebl, Hans. 2006. Recent Advances in Salzburg Dialectometry. *Literary and Linguistic Computing* 21(4):411-435.

Golla, Victor. 2011. *California Indian Languages*. Berkeley / Los Angeles / London: University of California Press.

Greenhill, Simon J. 2001. Levenshtein Distances Fail to Identify Language Relationships Accurately. *Computational Linguistics* 36(4):689-698.

Gwich'in Social & Cultural Institute and Gwich'in Language Committee. 2005. *Gwich'in Language Dictionary*. Fifth edition. Tsiigehtchic / Fort McPherson, NT: Gwich'in Social Cultural & Cultural Institute.

Heeringa, Wilbert Jan. 2004. *Measuring Dialect Pronunciation Differences Using Levenshtein Distance*. Groningen Dissertations in Linguistics 46. Groningen: University of Groningen. Online: http://irs.ub.rug.nl/ppn/258438452

Ives, John W. 2003. Alberta, Athapaskans and Apachean Origins. In Jack W. Brink and John F. Dormaar, eds., *Alberta, a View from the New Millenium*, 256-289. Medicine Hat, AB: Archaeological Society of Alberta.

Jetté, Jules and Eliza Jones. 2000. *Koyukon Athabaskan Dictionary*, ed. James Kari. Fairbanks, AK: Alaska Native Language Center.

Kari, James. 1990. *Ahtna Athabaskan Dictionary*. Fairbanks, AK: Alaska Native Language Center.

Kari, James. 2007. *Dena'ina Topical Dictionary*. Fairbanks, AK: Alaska Native Language Center.

Kaska Tribal Council. 1997. *Guzāgi K'ū´gé': Our Language Book: Nouns. Kaska, Mountain Slavey and Sekani*. Watson Lake, YT: Kaska Tribal Council.

Krauss, Michael E. and Victor Golla. 1981. Northern Athapaskan Languages. In William C. Sturtevant and June Helm, eds., *Handbook of North American Indians 6: Subarctic*, 67-85. Washington, DC: Smithsonian Institution.

Levenshtein, Vladimir I. 1969. Bounds for Codes Ensuring Error Correction and Synchronization. *Problemy Peredachi Informatsii* 5(2):3-13.

Mahli, Ripan S., Angelica Gonzalez-Olivier, Brian M. Kemp, and Kari B. Schroeder. 2008. Distribution of Y Chromosomes Among Native North Americans: A Study of Athapaskan Population History. *American Journal of Physical Anthropology* 137:412-424.

Matisoff, James A. 1978. *Variational Semantics in Tibeto-Burman*. Philadelphia: Institute for the Study of Human Issues.

Mithun, Marianne. 1999. *The Languages of Native North America*. Cambridge: Cambridge University Press.

Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg, and Therese Leinonen. 2011. Gabmap—a Web Application for Dialectology. *Dialectologia (Special Issue II)*:65-89.

Opler, Morris E. 1936. The Kinship Systems of the Southern Athapaskan-speaking Tribes. *American Anthropolo*gist 38(4):620-633.

Phone, Wilhelmina, Maureen Olson, and Matilda Martinez. 2007. *Dictionary of Jicarilla Apache: Abáachi Mizaa Iłkee' Siijai*, eds. Melissa Axelrod, Jule Gómez de García, Jordan Lachler, and Sean M. Burke. Albuquerque, NM: University of New Mexico Press.

Rice, Keren. 1989. *A Grammar of Slave*. Mouton Grammar Library 5. Berlin / New York: Mouton de Gruyter.

Rice, Keren. 2012. Linguistic Evidence Regarding the Apachean Migration. In Deni J. Seymour, ed., *From the Land of Ever Winter to the American Southwest: Athapaskan Migrations, Mobility, and Ethnogenesis*, 249-270. Salt Lake City: University of Utah Press.

Snoek, Conor. 2012. PACL: A Database for Linguistic Research and Language Revitalization. Paper, Alberta Graduate Conference, University of Alberta, Edmonton, AB, May 3-5, 2012

South Slave Divisional Education Council. 2009. *Dene Yatıé K'éé Ahsíı Yats'uuzı Gha Edıhtł'éh Kátł'odehche. South Slavey Topical Dictionary, Kátł'odehche Dialect*. Fort Smith, NT: South Slave Divisional Education Council

Spence, Justin. 2013. A Computational Assessment of Pacific Coast Athabaskan. Paper, Society for the Study of the Indigenous Languages of the Americas (SSILA) Winter Meeting, Boston, MA, January 4, 2013.

Tlen, Daniel. 1993. *Kluane Southern Tutchone glossary: English to Southern*

Tutchone. Whitehorse, YT: Northern Research Institute.
Young, Robert and William Morgan. 1987. *The Navajo Language: A Grammar and Colloquial Dictionary*. Albuquerque, NM: University of New Mexico Press.

Conor Snoek
Department of Linguistics
University of Alberta
3-02 Assiniboia Hall
Edmonton, Alberta, Canada T6G 2E7

snoek@ualberta.ca

Christopher Cox
Department of Linguistics
University of Alberta
3-02 Assiniboia Hall
Edmonton, Alberta, Canada T6G 2E7

christopher.cox@ualberta.ca