Berkeley Linguistics Society. 2013. 182-194. http://dx.doi.org/10.3765/bls.v39i1.3880
Published by the Linguistic Society of America

Large-scale Vocabulary Surveys as a Tool for Linguistic Stratigraphy: A California Case Study

YORAM MEROZ

As early as the 19th century, linguists have sought to classify California's hundred-odd languages and dialects, with view to understanding the area's prehistory. Early linguists were quick to recognize shallower genetic groups, but later studies have made relatively little further headway in understanding the historical connections between the area's diverse languages. Very few higher subgroups are universally accepted, and the largest subgrouping proposals, the Penutian and Hokan hypotheses, remain uncertain and controversial a century after they were first proposed (Dixon and Kroeber 1913a, 1913b).

Deeper subgrouping hypotheses have commonly been proposed on the basis of lexical lookalikes between languages, which on further study might turn out to be cognates. If a set of languages are related at a great time depth, few cognates will be available to be compared, and if more recent contact affected those languages, later loanwords may outnumber the cognates. For a linguistically complex area such as California, the history of later language contact must be well-understood before deeper relationships can be established with confidence. Moreover, loanword studies may reveal prehistorical contacts, and if relatively recent, may be more easily apparent in the data.

Lexical surveys, in California and elsewhere, have typically concentrated on basic vocabulary, the part of the vocabulary most resistant to replacement through either internal change or borrowing. Such surveys highlight genetic over contact relationships. To detect borrowings, a complementary type of survey is called for, one covering words which are more prone to borrowing.[1]

[1]Heggarty (2010) is a related statistical approach, which separately considers conservative and

Few studies of borrowing patterns in California exist. Some are confined to particular languages or families, and aim at detecting vocabulary borrowed from neighboring languages (Klar 1977, for Chumashan; Whistler 1977, for Wintuan; Turner 1983, for Salinan; Loether 1998, for Mono, Sierra Miwok, and Yokuts). Such studies are valuable, but by concentrating on only a small number of languages, they risk mistaking widely diffused words for local borrowings.

Other studies have studied lexical diffusion over a broad range of languages, but considreing only a few lexical items at a time (Nichols 1998, for the word for 'mountain lion'; Golla 2011:227–229, for words for six animal species, the bow and arrow, and some numerals). The wanderwörter identified in these studies are too few to recognize regular patterns in their distributions.

Some wide-ranging lexical items have been noted for California and beyond, in the context of evaluating deep subgrouping hypotheses. Campbell (1997) mentions some widely occurring words, arguing against their use as evidence for particular subgroupings (e.g. 'nose' and 'mouth' in the context of Hokan, p. 294, and 'goose' in the context of Coahuiltecan, p. 298). However, he does not attempt a systematic survey of such widespread forms.

Bowern et al. (2011) quantify the degree of lexical borrowing in several linguistically complex areas. In California, this study surveys 46 languages, using a standard list of 204 words, and presents statistics for the rate of borrowing in each language. Since the wordlist is selected from basic vocabulary items, the observed rate of borrowing is less than would be expected for more borrowable vocabulary, such as words for areally restricted flora and fauna. The only recurrent pattern in California mentioned in that study is heavy borrowing from Yokuts into Bankalachi/Toloim, a neighboring Uto-Aztecan language.

This paper presents the results of a comprehensive search for lookalikes among words for plants and animals in California languages. Words in this domain are typically more prone to borrowing than basic vocabulary, especially when speakers of a language move and encounter different species. A survey of such vocabulary is especially suited to identifying and highlighting old language contact. Since a language may be spoken far away from where its ancestor was once in contact with another language, and since words may spread far from their source through intermediate languages, this study does not exclude any languages in the area from being ultimately interconnected though old contact events.

1 Sources and Methods

This study is based on the vocabularies of C. Hart Merriam, naturalist, ethnographer, and amateur linguist, who between 1902 and 1938 conducted an exhaustive lexical survey of languages throughout California. As part of his

borrowable vocabulary to distinguish genetic connection from contact, in the case of Quechua and Aymara.

survey, Merriam used a standard form listing about 420 species of plants and animals, to which he often added additional ones by hand. He collected 156 such vocabularies, representing languages and dialectal varieties from throughout California and the neighboring Great Basin and the Arizona desert.[2]

This work employs a subset of Merriam's vocabularies, edited and published by Robert Heizer (Merriam 1979). Although less complete than the manuscript version, the published version could be digitized more easily and rapidly. The published edition was scanned, the scanned images converted to text through a commercial optical character recognition program, and the resulting text files edited and corrected by hand using the published edition as a guide. Additional species, which Merriam added as necessary to his forms, are not used here. The collected vocabularies were then imported into a database program for easy retrieval by either species or language.[3] In total, the database includes some 16,000 lexical items in 122 languages and dialects, representing 420 species of animals and plants. Of these, about 250 species are represented in enough languages to be useful for the comparative purposes of this work.

Merriam was not a trained linguist, and insisted on using a transcription system of his own, loosely based on that used for transcribing pronunciation in English dictionaries. His transcriptions were neither accurate nor consistent, and ignored some phonetic distinctions. Nevertheless, they are usually adequate for this study, which does not attempt to obtain exact sound correspondences.

This vocabulary database was arranged by species and printed out, and the comparisons carried manually. Similar words within each species were noted, as were words for closely related species. Phonetic similarities were judged subjectively and marked as 'likely', 'possible', or 'farfetched'. In this paper, only lookalikes marked 'likely' are used. In general, very short forms were disfavored, as were pairs of words with unexplained mismatching segments. While this procedure leaves out what may later turn out to be related words, it is necessary for reducing chance resemblances.

This subjective comparison not ideal. The search for lookalikes has missed some candidates which were found on later inspection, and others are no doubt still unnoticed. A reliable automatic cognate detection algorithm, if one is devised, would provide a more objective and complete collection of potential historically related words.

As a final step, the sets of lookalikes—each set corresponding to a species—were compared, and recurring patterns of forms shared between languages were noted. Again, this is a process that may eventually be automated, for the sake of

[2]Merriam's vocabulary manuscripts are kept at the Bancroft Library in Berkeley. Digital images of the vocabularies are available online, through the Internet Archive (http://www.archive.org). Merriam also procured vocabularies in non–natural history domains; those are not utilized here.

[3]All the materials used for this study will be posted on website of the Survey of California and Other Indian Languages, at Berkeley (http://linguistics.berkeley.edu/Survey).

demonstrable objectivity.

The following section discusses some of the recurring patterns of vocabulary sharing between disparate language families, noted in Merriam's vocabulary database.

2 Results

As mentioned above, Merriam used an idiosyncratic and inconsistent transcription system. His system rarely marks phonemic distinctions not present in English, such as glottalization and the /q/-/k/ distinction. He does often transcribes /x/ with a distinctive sign (<ch̲>), but at other times uses <k> or <h> for /x/; he often notes retroflex stops (for example using <tr> for /ṭ/). For ease of reading, I use here my interpretations of his forms, rather than quote them verbatim.

Some languages and families are represented in the database by a large number of closely related dialects: 16 Yokuts varieties, 6 of Patwin, 7 of Palaihnihan, etc. This enables a more fine-grained view of the distribution of particular words, and helps guard against relying on any one informant as a representative of a language as a whole.

In the examples given below, each common taxon name is followed by its number in the published edition.

2.1 General patterns

The similarity judgments used in this study are subjective. As mentioned above, some effort was made to reduce chance similarities. It is reassuring to see that not all language groups are represented equally in non-genetic lookalike lists, suggesting that chance lookalikes are not a significant part of the sets. Roughly, Coast Range languages (Athabascan, Algic, Yuki, Costanoan, Salinan, Chumash) and Yuman languages share relatively few words with external groups. Central Valley languages (Yokuts, Miwokan, Wintuan) share relatively more with their neighbors. This is consistent with previous studies, and with the observation that more mobility and therefore language contact would be expected in the Central Valley than in more isolated mountainous areas.

Onomatopoeias and other sound-symbolic words are often considered unreliable for hypothesis formation when comparing vocabularies, since similar sound-symbolic motivation can independently produce similar words in disparate languages. In the present database, onomatopoeias occur as words for many animal species, especially birds. Nevertheless, with enough attention to formal detail, many of these word sets convey useful information. For example, 'osprey' (76) is represented in 75 vocabularies, including the following words, arranged by family and language, which could all plausibly be of sound-symbolic origin:

(1)			
	Athabascan:	Mattole	saki
	Algic:	Wiyot	tsaktsakw
	Yukian:	Coast Yuki	čučuka
		Wappo	tsuku
	Shastan:	Shasta	čuču
		Konomihu	čuču
	Palaihnihan:	Apwurakeyi	toktokisi
		Atsugewi	toktokisi
	Yana:	Yana	čiči
	Maiduan:	Chico Maidu	tsitsi
		N. Maidu	čawtata
	Wintuan:	Patwin (5 varieties)	tuktuk
	Miwokan:	N. Sierra Miwok	tuktuku
		Lake Miwok	tuktuk
	Yokutsan:	Chukchansi	šošu
		Gashowu	šošu
		Choinimini	šukšu
		Nutunutu	saksux
		Tachi	soksox
		Chunut	soxsu
	Numic:	Wobonuch Mono	soksok
		Entimbich Mono	šokšo
		Waksachi Mono	šokšu

While all these forms are broadly similar, they are generally more similar within families than across them. Among cross-family lookalikes, the Yokutsan-Monache similarities parallel those of many other lookalike sets, which are interpreted here as loans from Yokutsan languages into various Western Mono varieties, as also noted by Loether (1998). Likewise, the Patwin forms are identical with those of Lake Miwok but altogether different from that of their nearest relative, Wintu /kule/, suggesting a loan from Miwok into Patwin. Other such loans were noted by Whistler (1977), as discussed further below.

2.2 Bankalachi

Bankalachi, or Toloim, is a dialect of Tübatulabal (Uto-Aztecan), spoken around Deer Creek, in the foothills of the southernmost Sierra Nevada. Jane Hill (in Bowern et al. 2011) has previously noted a high rate of borrowing into Bankalachi, amounting to about 20% of the basic vocabulary, and attributes it to ongoing language shift. In the vocabularies studied here, which consist of the more borrowable natural history terms, some 80% of the Bankalachi words are borrowed from Yokuts languages.

Nearly all the borrowings match most closely the form in Yawlamni ('Yawelmani'), a Valley Yokuts language. Historical Yawlamni territory, however, is where the Kern River enters the Central Valley, some 50 km to the south. This suggests that the contact between Yawlamni and Bankalachi was not recent, but occurred at a time when the groups lived closer to each other.

Three words have a Yokuts source other than Yawlamni:

(2) 'toad' (245)

Bankalachi	koyetwuk
Nutunutu	koyotawuk
Yawlamni	okoko

(3) 'scorpion' (276)

Bankalachi	itetiš
Nutunutu	itatit
Yawlamni	petetič

(4) 'sycamore' (308)

Bankalachi	kolek
Palewyami	kolak
Yawlamni	kočik
other Yokuts	kotik / koṭik / kotsik

Palewyami was spoken along Poso Creek, 30 km to the south of Bankalachi territory. Nutunutu was spoken north of Tulare Lake, 80 km to the northwest. The evidence of loanwords in Bankalachi indicates a complex linguistic history in the San Joaquin Valley.

2.3 Pomoan-Yokutsan

The Pomoan languages belong to the coastal ranges north of San Francisco Bay. Pomoan is one of the branches of the putative Hokan language family, though no language family has been clearly demonstrated to be related to it. Surprisingly, in Merriam's vocabularies, several lookalikes are shared between Pomoan and Yokuts languages and no others, except for obvious later local loans. Yokuts is one of the proposed branches of Penutian, but genetically unrelated to Pomoan. Several geographical barriers and hundreds of kilometers separate the two families:

(5) ‘flying squirrel’ (51)

N. Pomo	keple
E. Pomo	kepla
Choinimni, Wikchamni	kapalala

(6) ‘kingbird’ (132)

N.E. Pomo	tapičoroka
Yawlamni	tapičlela
Chunut	tapičlala

(7) ‘mallard’ (194)

S. Pomo	watata
C. Pomo	wadawada (‘merganser’, 193)
Chukchansi, Choinimni, Telamni	watwat

(8) ‘spider’ (274)

N. Pomo (Tabate)	mča
N. Pomo (Kayaw)	misa
Chukchansi, Gashowu, Telamni	meča
Tachi	metsa
Wikchamni	muča

(9) ‘yerba santa’ (364)

C. Pomo, E. Pomo	tekale (< -q^{h}ale ‘tree’?)
Yawlamni (Tinlini)	taxal

To my knowledge, there is no claim that these two language families or their ancestors were ever near each other. A less obvious historical scenario will be needed to explain these sets, if they are confirmed to not be accidental.

This example demonstrates the utility of using broad surveys of borrowing-prone words for detecting unexpected relationships in an area of a complex linguistic history. Linguistic surveys based on basic vocabulary, aimed at detecting genetic relationships, might not show enough borrowed vocabulary to detect this relationship.

2.4 Patwin borrowings

Patwin belongs to the Wintuan language family, located along the western side of the Sacramento Valley. Whistler (1977) reconstructed words for flora and fauna in Proto-Wintuan, and used these to show that its homeland was near the California-Oregon border. Patwin, the southernmost of the Wintuan languages, is located at the southern end of the valley, and borders Miwok territories. Whistler

proposes a number of borrowings from Miwokan languages into Patwin, and thus argues that the Patwin entered the southern Sacramento Valley after Miwokan speakers had already been established there.

While I agree with Whistler's conclusions, a few of his proposed etyma turn out to have a more complex history. I demonstrate this with the following three species. For each one, I show Whistler's (1977:162) proposed etymology, followed by Merriam's data:

(10) 'incense cedar' (290) / 'juniper' (292).

W77 Proto Miwok *mo·n 'cedar' : Patwin mon 'juniper'	
Yana	muniyi ('juniper')
Nomlaki	mun ('juniper')
Patwin	mun / munmun / mon ('juniper')
S. Maidu	monimča ('cedar')
Konkow (Huncut Creek)	monimča ('cedar')
N. Maidu	manimča ('cedar')
N. Sierra Miwok	monogo ('cedar')

Whistler proposes that this word was borrowed from a Miwokan language into Patwin. Its wider distribution argues against that scenario. The word seems to have started its spread somewhere to the north, entered the Wintuan languages Nomlaki and Patwin, the Maiduan languages, and finally Miwokan.

(11) 'condor' (81). W77 Patwin mo·lok : Proto Sierra Miwok *mol·ok

Wintu	moluk
Nomlaki	molok
Patwin (6 varieties)	molok / moluk
N. Sierra Miwok	moluko
Coast/Lake/Plains Miwok	moluk
Maiduan (5 languages)	moluk / molok / moluko / moloko

The word is present in all branches of Wintuan, and is not merely a loan from Miwokan into Patwin. The connection with Maiduan is less clear, but I surmise that the word was borrowed into Maiduan from a Wintuan language, or that both borrowed it from some other common source.

The word also appears as N.E. Pomo moluk, probably a Patwin borrowing, and as Telamni Yokuts limik, perhaps a S. Sierra Miwok loan, with metathesis.

(12) ‘fly’ (265). W77 River Patwin homo·tay : Proto E. Miwok *homo·-

Hammawi	hamomuma
Maidu (2 varieties)	hamelulu / emalula
Konkow (2 varieties)	emelulu-m / hemelulu
Patwin (Colusa)	homotai
N. Sierra Miwok	homomiyu
Plains Miwok	homomiye

This widespread species has forms akin to homo- in one Patwin dialect and in Miwokan, as in Whistler, but also in Maiduan and in Hammawi (a Palaihnihan variety close to Achumawi), but nowhere else in the collection. A Palaihnihan language could be a source for the word, though the path from it to Miwokan and Maiduan languages is still to be elucidated.

2.5 Pomoan and Palaihnihan

A number of words in the database are shared between Pomoan and Pit River languages, and no others:

(13) ‘grizzly bear’ (1)

Apwarukeyi, Atsugewi	piriki
E., N.E., S.E. Pomo	puraka

(14) ‘red fox’ (10)

Apwarukeyi, Atsugewi	kwaw
N.E. Pomo	kawka
N. Pomo, C. Pomo	kaw
E. Pomo	kakaw

(15) ‘wolf’ (14)

Astakiwi, Atwamwi, Achumawi	tsimu
Hammawi, Mahdesi	čimu
N.E. Pomo	čomeka
N. Pomo	tsimeya / čimyu / smewa
C. Pomo	smewa
S. Pomo	tsemyuwa
E. Pomo	čemu
S.E. Pomo	sumu

(16) ‘cottontail’ (63) / ‘snowshoe rabbit’ (64) / ‘black-tail jackrabbit’ (65)

Achumawi	kalak (‘snowshoe rabbit’)
N.E. Pomo	takalika (‘cottontail’), makalakaka (‘jackrabbit’)
N. Pomo, C. Pomo	makala (‘jackrabbit’);
(note also:) Nomlaki	takalal (‘cottontail’ < Pomo?)

(17) ‘western tanager’ (130)

Apwarukeyi, Atsugewi	waswosa
S. Pomo	wašwaš

(18) ‘yellow-breasted chat’ (131)

Mahdesi	waswasa
N. Pomo	waswas

(19) ‘ruddy duck’ (201)

Hammawi, Atsugewi	tanana
N., E. Pomo	tana

(20) ‘trout’ (248)

Achumawi	selepi
Hammawi, Astakiwi, Atwamwi, Mahdesi	salepi
N. Pomo	šalobi

(21) ‘centipede’ (277)

Mahdesi	hustoyi
N. Pomo	hošutil

(22) ‘gray pine’ (283)

Hammawi	tutsxale
Atwamwi	tutsxalo
Achumawi	totsxalo
Mahdesi	tuxale
N.E. Pomo	tutekale
N. Pomo	kotekale / ketexale
E. Pomo	kotexale

(23) ‘sugar pine’ (282)

Achumawi	asawyo
Apwarukeyi	atsowo
N. Pomo	šuye

The Pomoan languages are spoken in the Coast Range, at the southwest corner of the Sacramento valley. The Palaihnihan languages are spoken in the Pit River basin, at the northeast corner of the valley, some 250 km away. The lookalikes given here, if confirmed, can be explained only through a genetic relationship, or through old contact.

The Pomoan and Palaihnihan languages have in the past been hypothesized to be related, as members of the putative Hokan family; but, to my knowledge, no one has ever proposed linking the two groups in a closer relationship than Hokan as a whole. In contrast, the data here shows a close relationship between the two groups, since no comparably large set of lookalikes has been found containing members of the two languages and some additional ones.

Gursky (1974) is the largest published comparative list of potential Hokan etymologies.[4] Out of the 30 sets in Gursky's list which refer to basic (non–natural history) vocabulary and which contain Palaihnihan and Pomoan words, 12 do not contain examples from other language families. That would normally be a strong argument for a genetic connection between the two groups, assuming that the forms were plausibly related. However, that set is suspect. Although Gursky used both Achumawi and Atsugewi dictionaries to construct his lists, all of the exclusive Pomoan-Palaihnihan sets contain Achumawi examples, and none contain Atsugewi, although these two branches of Palaihnihan are fairly closely related. The explanation for that is apparently that Gursky used Olmsted's (1966) Achumawi dictionary as his source. As Nevin (1998:10) notes, and as Gursky later recognized, Olmsted's dictionary has inadvertently mingled Pomoan lexical materials among the Achumawi ones; and in fact, the exclusive Pomo-Achumawi matches in Gursky's list all show a suspiciously near-exact phonetic match. I conclude that there is no close genetic connection between Palaihnihan and Pomoan, and that the lookalikes in Merriam's lists indicate borrowing.

A more detailed analysis of the data should be able to show the direction of borrowing, and perhaps offer clues as to where the borrowing took place. For now, a reasonable hypothesis is that languages belonging to either or both of these families were spoken in the Sacramento Valley, in what is now Wintuan territory.

3 Conclusion

Although much work in California and elsewhere in North America has been directed at finding genetic groupings, searches for old language contact have been few and localized. This study aims at detecting prehistoric language contact in California by systematically searching for loanwords in lists of natural history words, a semantic domain particularly prone to borrowing.

This paper presents some representative results of this study. In the case of Bankalachi and Patwin, it confirms and elaborates observations made by earlier

[4]Gursky has published several addenda to his original publication, which were not used here.

researchers. For Bankalachi, several Yokuts varieties are identified as sources of borrowing, not all contiguous with it in historical times. For Patwin, Miwokan is confirmed as the source of some loanwords, as first shown by Whistler (1977), but some connections with Maiduan, Yana and Palaihnihan are identified as well.

Two new contact situations have been identified here, one between Yokuts and Pomoan, the other between Pomoan and Palaihnihan. In both cases, the language families are now far apart; these results therefore provide new clues to ancient population movements.

This study has been exploratory, and is far from exhausting the potential of the method and of the existing materials. Future work should include augmenting Merriam's vocabularies by transcribing the ones not in Heizer's compilation, and adding other published and unpublished materials; in particular, ethnobotanical studies are rich in detailed plant vocabularies, and will add names for species not compared here. More accurate transcriptions from other sources will help distinguish accidental lookalikes from significant ones. With detailed knowledge of the languages involved and with more accurate data, there is a great potential for discovering loan translations as well. The study area can and should be extended to languages further north.

The method illustrated here should be applicable in any linguistically diverse area, and similar studies elsewhere should be likewise fruitful in uncovering old language contact.

References

Bowern, Claire, Patience Epps, Russell Gray, Jane Hill, Keith Hunley, Patrick McConvell, and Jason Zentz. 2011. Does Lateral Transmission Obscure Inheritance in Hunter-Gatherer Languages? *PLoS One* 6(9):e25195.

Campbell, Lyle. 1997. *American Indian Languages: The Historical Linguistics of Native America*. Oxford: Oxford University Press.

Dixon, Roland B., and Alfred L. Kroeber. 1913a. Relationship of the Indian Languages of California. *Science* 37:255.

Dixon, Roland B., and Alfred L. Kroeber. 1913b. New Linguistic Families in California. *American Anthropologist* N.S. 15(4):647–655.

Golla, Victor. 2011. *California Indian Languages*. Berkeley: University of California Press.

Gursky, Karl-Heinz. 1974. Der Hoka-Sprachstamm: Eine Bestandsaufnahme des lexikalischen Beweismaterials. *Orbis* 23:170–215.

Heggarty, Paul. 2010. Beyond Lexicostatistics: How to Get More out of 'Word List' Comparisons. *Diachronica* 27(2):301–324.

Klar, Kathryn A. 1977. *Topics in Historical Chumash Grammar*. Ph.D. dissertation. Berkeley: University of California, Berkeley.

Loether, Christopher P. 1998. Yokuts and Miwok Loanwords in Western Mono. In Hill, Jane, P. J. Mistry, and Lyle Campbell, eds., *The Life of Language: Papers in Linguistics in Honor of William Bright.* pp. 101–122. Berlin and New York: Mouton de Gruyter.

Merriam, C. Hart. 1979. *Indian Names for Plants and Animals among Californian and other Western North American Tribes.* Assembled and annotated by Robert F. Heizer. Socorro, NM: Ballena Press.

Nevin, Bruce E. 1998. *Aspects of Pit River Phonology.* Ph.D. dissertation, Philadelphia: University of Pennsylvania.

Nichols, Michael. 1998. An Old California Word for 'Mountain Lion/Wildcat'. In Hinton, Leanne, and Pamela Munro, eds., *Studies in American Indian Languages: Description and Theory*. University of California Publications in Linguistics 131. pp. 241–247. Berkeley and Los Angeles: University of California Press.

Olmsted, David L. 1966. *Achumawi Dictionary*. University of California Publications in Linguistics 45. Berkeley and Los Angeles: University of California Press.

Turner, Katherine. 1983. Areal and Genetic Affiliations of the Salinan. *Kansas Working Papers in Linguistics* 8(2):215–246.

Whistler, Kenneth W. 1977. Wintun Prehistory: An Interpretation Based on Linguistic Reconstruction of Plant and Animal Nomenclature. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society* 3:157–174.

Yoram Meroz

yoram.meroz@gmail.com