Note

"The trials and joys of comparative dictionary making"

Michael Fortescue

Études/Inuit/Studies, vol. 31, n°1-2, 2007, p. 213-221.

The trials and joys of comparative dictionary making

Michael Fortescue*

Résumé: Les vicissitudes de la préparation de dictionnaires comparatifs

Les méthodes d'élaboration de dictionnaires comparatifs des familles de langues autochtones de l'Arctique et du sub-Arctique, ainsi que les motivations qui sous-tendent une telle élaboration, sont illustrées par l'expérience de l'auteur avec la préparation du dictionnaire comparatif des langues esquimaudes (Fortescue et al. 1994), du dictionnaire comparatif du tchoukche-kamtchatkien (Fortescue 2005) et du dictionnaire comparatif du wakashan (Fortescue 2007). L'arrière-plan historique de ces trois projets est esquissé, et si l'intervention portera surtout sur ces projets, nous ferons néanmoins aussi quelques observations générales sur la construction de dictionnaires comparatifs de langues qui sont soit en danger de disparaître, soit dotées de peu de profondeur historique. Sera effleurée également la question plus vaste de l'aire linguistique qui lie les trois projets.

Abstract: The trials and joys of comparative dictionary making

The motivation behind and the methods used in putting together comparative dictionaries for language families in the Arctic and Sub-Arctic is illustrated from the author's experience in working on the *Comparative Eskimo Dictionary* (Fortescue et al. 1994), the *Comparative Chukotko-Kamchatkan Dictionary* (Fortescue 2005) and *Wakashan Comparative Dictionary* (Fortescue 2007). The historical background behind these three projects is sketched and, although the focus is on them in particular, some general observations concerning the making of historical/comparative dictionaries for language families that are either endangered or possess little historical depth of attestation are presented. The broader areal framework linking the three projects is also touched upon.

* Department of Scandinavian Studies and Linguistics, University of Copenhagen, Njalsgade 120, DK-2300 Copenhagen S., Denmark. fortesq@hum.ku.dk

Introduction

In the course of the last two decades I have had the privilege of working successively on three large-scale comparative dictionaries, the *Comparative Eskimo Dictionary* (Fortescue et al. 1994), the *Comparative Chukotko-Kamchatkan Dictionary* (Fortescue 2005), and now the *Wakashan Comparative Dictionary* (Fortescue 2007)[1]. I have sometimes been asked why and how such rather esoteric dictionaries are made. Encouraged by the guest editor of the present issue of *Études/Inuit/Studies*, I shall endeavour to provide some answers, illustrated mainly from my experience with the *Comparative Eskimo Dictionary* (CED). In doing so, I will try not to dwell too much on the realm of purely personal experience.

The "why" question

Let me start with the "why" question. This can be answered in good forensic fashion from the perspective of motivation and opportunity. Objectively, the major motivation is of course the aim of furthering our knowledge of the world's languages—comparative dictionaries with protolanguage reconstructions open, as it were, a window on the conceptual past of the people that speak them. The opportunity to pursue this goal (in terms of funding, available materials, and the involvement of researchers able to devote sufficient time to such labour-intensive undertakings) is not always at hand.

The Comparative Eskimo Dictionary

It was, however, at hand when I first joined up with fellow Eskimologists Steve Jacobson and Larry Kaplan to work on the CED. This project combined ongoing comparative work at the Alaska Native Language Center (ANLC), encouraged by then director Michael Krauss, with my own work at the Institute of Eskimology in Copenhagen, where I had available to me the extensive comparative files of the late Erik Holtved. These were based largely on outdated and unreliable sources (especially for the Yupik languages) but their alphabetised organisation provided a unique starting-point for a more up-to-date dictionary, ensuring that no major cognate sets would be missed on the way. The newer Alaskan material being produced at ANLC was more accurate but lacked data on the Canadian and Greenlandic dialects.

As in the case of the other two comparative projects mentioned above, the descriptive coverage of the languages and dialects concerned had reached a point where a comprehensive treatment of the family was possible, with only a relatively small amount of fresh field-work still needed to fill in the remaining holes. This was also true

1 The Chukotko-Kamchatkan or Paleosiberian language family of the Russian Far East covers a northern branch, containing Chukchi, Koryak, Kerek and Alutor, and a southern branch, now only containing Itelmen of western Kamchatka. The Wakashan family of British Columbia also has a northern branch (the best known language being Kwakwala, formerly Kwakiutl) and a southern branch (the best known language being Nuuchahnulth, formerly Nootka).

of the Aleut material cited in the CED, for which we relied heavily on Knut Bergsland's on-going work—it was of course his seminal work in the 1950s to 1970s (summed up in Bergsland 1986) that largely clinched the genetic unity of the Eskimo-Aleut family in the first place. He assisted us greatly in the integration of the relevant forms, as well as with general advice on the comparative task at hand. Owing to the far-flung nature of the Eskimo-Aleut family, with its many local varieties, a huge amount of data, in a variety of orthographies and of varying degrees of accuracy, had to be marshalled: team-work was essential. Holtved's own project, though it resulted in several thousand carefully collated library filing cards was too ambitious and started perhaps too late in his life ever to have been completed without the assistance of computers. As it was, three researchers spending the bulk of their research time on the project for about 10 years, with the excellent technical back-up and publishing resources of ANLC, was needed to complete the project.

The Chukotko-Kamchatkan dictionary

The second project, the Chukotko-Kamchatkan dictionary, was a natural development out of the first, and in a way presupposed it, since Chukchi has had such far-reaching effects on Siberian forms of Yupik and effects of Yupik on Chukotkan languages were in turn also suspected. Moreover, the question of a possible deep genetic relationship lay—and many would say still lies—open. The lexical coverage of this neighbouring family had reached a relatively high level of completeness (with certain exceptions) thanks to the comparative spade-work of Irina Muravyova (1976) and others, comparable to the case of Eskimo-Aleut in the early 1980s. The task was somewhat less daunting than with the latter family, covering a smaller number of extant languages and dialects, but presented additional problems due to the poor quality of earlier work on all but the surviving western dialect of Itelmen (heavily influenced by Koryak and Russian). It was a project that appealed to my long-term interest in the relationships between languages on both sides of Bering Strait, whether genetic or areal. The genetic unity of the family was still a contested matter, and a comparative dictionary was essential to delineate the common grounds of the family (now largely accepted as such) in a common protolanguage stage. I did some limited field work with Chukchi dialects myself, but as was the case with the other two projects I have benefited greatly from the cooperation of other scholars in the area. I was especially lucky—at a rather late stage in the project—in having made available to me Alexander Asinovsky's unique files on Kerek, the fourth Chukotian language, now extinct. In putting together the Chukotko-Kamchatkan dictionary I was able to apply the methods—not to mention shortcuts—that I had gained experience with through working on the CED, whose format (which had proved quite successful) I also adopted.

The Wakashan Comparative Dictionary

The latest, Wakashan project, also falls naturally within the framework established by the CED and represents a further investigation of contact between the various peoples of the North Pacific Rim. Linguists, archaeologists, and anthropologists have often suggested links between speakers of these languages and Eskimo-Aleuts: it has not gone unnoticed that all these people (and none of their neighbours apart from the coastal Chukotians) are people adapted to coastal life, who developed complex whaling societies and speak purely suffixing polysynthetic languages. Thanks to the pioneering work of Franz Boas and Edward Sapir, amongst others (*cf.* especially Sapir and Swadesh 1952), this was one of the few Indian language families that was ripe for a full-blown comparative treatment. Only Ditidaht (formerly Nitinat) of southern Vancouver Island remained rather poorly described as regards its lexicon, and this I have attempted to remedy by recent fieldwork of my own with the last handful of fluent speakers. The completion of this dictionary will hopefully contribute to untangling the different linguistic and cultural strands that have entered the Northwest American area—in this case presumably from the north.

In none of the three dictionaries is there any speculation about distant genetic relations to other language families: they are quite neutral in this respect, following the rules of the (comparative) game to the letter. They simply represent the prerequisite for undertaking further investigations of the genetic affinities between the languages of the Old World and those of the American northwest. It is also to be hoped that they will prove to be useful tools for anyone—including anthropologists and archaeologists as well as linguists—who is interested in complex areal questions of cultural relations and roots such as those presented by Chukotka or the American northwest coast. Of the various hypotheses that may eventually be supported by correlating the results of the three projects one could be the presence of a common linguistic substratum beneath Aleut and/or Itelmen, perhaps directly relatable to Wakashan during pre-Eskimo (Ocean Bay tradition?) times. This might explain the marked divergences—lexical as well as structural—of these languages from the other branches of respectively the EA and CK families.

The "how" question

Now for the "how" question. The first concern in constructing a comparative dictionary is of course the marshalling of all available data. This includes, besides primary published data on the individual languages and any previous comparative work on the family concerned, all archival material, however poor in quality, since this can be crucial when dealing with languages lacking any historical depth of attestation. In the case of the CED this was facilitated by the wealth of material accumulated at ANLC and at the Institute of Eskimology. Based on a survey of the materials available, some fundamental decisions then have to be made which will have a profound effect on the following steps.

First, it has to be decided how to divide up the family into languages and (groups of) dialects, if that has not been fully established. Then the criteria for setting up cognate sets of related forms must be agreed upon, and, finally, a working hypothesis as regards the phoneme system of the proto-language must be set up such that regular sound changes can be established for developments from that to the modern languages. Related to this latter step is the need for orthographical consistency. In the case of the CED, the first decision resulted in a division of the Inuit continuum of dialects (these broadly corresponding to the traditional ethnographic groups) into five overall dialect regions, as defined by certain major isoglosses. From each of these one default dialect was selected, as determined by the comprehensiveness and quality of data available, such that only significant (irregular) deviations from the default dialect are indicated for other dialects in the region. This was a perfectly defensible way of preventing an exponential expansion of the dictionary's size, as was the decision to reconstruct Proto-Inuit cognate sets only when represented in at least two of these overall divisions. The Yupik branch did not need this treatment, as each Yupik language covers a more limited array of dialects: a proto-Yupik form would simply need to be attested in a minimum of two of the accepted Yupik languages. Cognate forms in each Inuit dialect group and each Yupik language were given their own "line" with additional comments in brackets, to provide a consistent reader-friendly format, with prominent indication of unattested gaps.

It was further decided that as far as possible one common set of orthographical symbols (fully phonemic) would be used, from which representations in all the languages drew, including the proto-language. This facilitated the alphabetical ordering of cognate sets by reconstructed stem forms with derived sets of various proto-stages indented below them, but it was not the only possible solution and is a matter that needs to be addressed from family to family, but here it proved to be expedient. In language families with less profuse derivational potential this two-tier arrangement would be superfluous. As for the phonemic inventory of the reconstructed proto-language itself, this was reached by a combination of trial and error and a judicious application of Occam's Razor (aided of course by insights from previous comparative work). The final set of phonemes agreed upon (and restrictions on their sequencing) was that which functioned most efficiently to explain all surviving modern forms, minimising the need for recourse to special developments such as assimilatory, dissimilatory and analogical changes. This did not happen all at once, prior to the actual working out of a good many cognate sets, however. As Mary Haas described the process (referring to Wakashan):

> Once you have reconstructed the main features of the protolanguage you are in a position to attempt to do some historical linguistics. That means going to it from the other end. In other words, you take the reconstructed forms and trace them down into each daughter language, and this provides a check-up on your reconstructions. This is especially useful if you do it with a daughter language that was not used in making the original reconstructions. In doing this you have to remember that your reconstruction is only an hypothesis, whereas the data from each of the daughter languages is not hypothetical but actual. In this way you may uncover evidence that will make it necessary to change some of your reconstructions. This is a point that is often overlooked (Haas 1979: 10).

To this could be added that the typological likelihood of one's reconstructed sound system needs to be born in mind: rare or unexpected configurations can be a sign that something is amiss and needs rethinking. In practice, rather few adjustments after the initial stage of the project needed to be made. Hypothesising from the start a maximal range of possible significant differences paid off: these are easier to eliminate through generalisations later than to go back and put in differences that were earlier ignored as mere phonetic detail.

Thereafter follows the pure slog of sorting out more and more cognate sets, a matter of sifting through dozens of single-language dictionaries. It is difficult to imagine how this already time-consuming phase of the endeavour could have been done in earlier times without the help of search-and-replace functions on computer files. These are a boon when checking that, for instance, a given form has not already been mentioned elsewhere in one's files. Comparative dictionaries in the past were typically the culmination of a whole scholarly lifetime of devotion to a single language family. The process is—or can be—much faster today. This stage of the gestation of the CED nevertheless lasted several years and necessitated the regular comparison and coordination of the efforts of three collaborators who only worked literally side by side during a period of several months near the beginning of the project. A dialectic between "lumping" and "splitting" tendencies was at play all the time. This resulted *en route* in numerous dubious sets later being split into two, often linked by a non-committal "*cf.*"—a most useful device covering a whole range of degrees of certainty. A liberal sprinkling of question-marks was also called for, and many of them ultimately kept (one can never achieve total certainty in such a work; lack of question marks should raise suspicions).

On the other hand, many cautiously distinguished sets (*e.g.*, a Proto-Yupik and a Proto-Inuit one, where there were slight discrepancies in form or meaning or both) were often unified in the end, typically when some new form was found that clinched the connection. How to deal with deviations from the combination of formal regularity and transparent semantic plausibility is something that only experience with the particular language family concerned can satisfactorily determine—intuition cannot wholly be dispensed with, but nor can knowledge of the specific cultural background of the people speaking the languages. One learns to recognise the pitfalls lurking: the hidden loan-words, the contaminations between similar forms and the "folk etymological" reanalyses that bedevil any lexical work of this nature. There is also (though less seriously) the temptation of reconstructing proto-sets for such anachronisms as an apparent Proto-Eskimo word for "binoculars"! The latter example—a real case—was of course the result of parallel but independent semantic developments from a common root for looking around in different branches of the family.

It should be added here that there is something that counterbalances the "slog" of this stage and renders it far from tedious, and that is the constant "eureka" experience that accompanies the discovery of every new, and sometimes surprising, correlation of forms and meanings that justifies or explains a new cognate set. Thus the origin of one

of the Eskimo words for 'spider' (*nissavarsuk* in West Greenlandic, *nenguryaq* in Alutiiq Yupik, *nigžuaržuk* in Inupiaq), was quite obscure, a derivation of some unknown stem, until one day we came across the stem in Seward Peninsula Inupiaq (the only language where it has survived), namely *negžuq-* 'let down by rope' (specifically down a rock face to collect eggs): it is this kind of dangling spider the word refers to, not the larger kind that scuttles across the ground.

Finally, of course, everything must be checked and cross-checked—by as many different pairs of eyes as possible—and indexes then made with computer assistance. This is a stage that should not be started too early, for changes can generally be expected virtually up to the last minute before publication, especially if more than one compiler is involved.

Along the way, problems of a more specific nature will inevitably arise. There are, for example, special difficulties attendant on reconstructing inflectional morphology, as these are prone—owing to frequency of usage—to phonological attrition and analogical levelling. One would do well to heed Bergsland's (1986) warning to the compilers of the CED not to assume that the proto-language displayed more regular morphology than the modern languages. Idiosyncrasies common to all daughter languages (*e.g.*, such suffix-initial alternations as p/v or t/ð and y/s) must be reconstructed also for the proto-language, although they doubtless derived from a single member of the pair at some still earlier stage. There is no reason why the proto-language should not have had as much morphological irregularity as its daughter languages. Recreating proto-morphology is nevertheless crucial to attempts to argue for more distant genetic relationships between languages.

Another kind of problem which arose with all three dictionaries is that of how to handle recently extinct or moribund languages for which only poor quality material is available or where there has been strong influence on the language from neighbouring ones. In the case of the CED this arose in connection with the now extinct Sirenikski language. Since it was agreed that Sirenikski was probably a third branch of the family, on a par with Yupik and Inuit, a separate "line" was obviously called for for comparative purposes. However, the forms cited were by necessity from heterogeneous sources (and of varying accuracy)—and many were clearly borrowed from neighbouring Siberian Yupik. It was obviously impossible to reconstruct 'Proto-Sirenikski' on a par with PI (Proto-Inuit) or PY (Proto-Yupik), and for practical reasons an artificial compromise entity, "Proto-Sirenikski-Yupik," was introduced to cover those rather frequent cases where Yupik and Sirenikski shared a word that was not attested in the Inuit branch. Similar problems arose with Itelmen, the southern branch of the CK family. Like Sirenikski, the surviving variety of Itelmen has been much influenced by neighbouring languages. The early sources for the other, extinct varieties of Itelmen were inaccurately documented (by travellers and exiles rather than trained linguists) yet of great comparative importance since they reflected earlier forms of the language before massive influence from Russian and Koryak obscured the picture.

Then there is of course the ultimate question of when to stop. The notion that such a dictionary is finally complete will always be illusory. More relevant data will always turn up later (unless one is dealing with entirely dead languages). This was the case with the CED, for example, when all of Duncan Pryde's comparative Western Inuktitut files turned up unexpectedly on the doorstep of the Institute of Eskimology shortly after his death. During the production of the dictionary only the A's from his files were printed out and available to the project. This material has now been sifted through and relevant forms integrated into a preliminary second edition of the CED, which hopefully will see the light of day before too long. It will contain, besides Duncan Pryde's material, many other detailed additions and corrections.

It is to be hoped that the preceding paragraph does not leave the impression that there is little more to do in the way of comparative Eskimo-Aleut linguistics. There are still many dialects for which only rather patchy data is available, and even when we can be reasonably sure that not much more primary data is ever going to be available for a given language or dialect—as is perhaps the case for Aleut after Bergsland's Herculean efforts—there is a great deal more that can be done relating the data we have to the broader picture. This includes delving deeper into the proto-language behind the whole family and its possible relationship, both genetic and areal, to other language families in the vicinity or in Siberia. Often a single form in a single language can be crucial (as in the "spider" example given above). A particularly pressing matter awaiting future investigation is the high percentage of lexical stems in Aleut that have no apparent equivalent in Eskimo languages. Where do they come from? Do they reflect an ancient substratum or were they simply lost over the millennia in the Eskimo branch of the family (while Aleut in turn lost many other originally common stems)? And where do all the myriad affixes of the family come from? At present there are only one or two that can be definitely related to independent stems. In fact, the writing of a truly comprehensive comparative dictionary is a never-ending task and one has to face the fact that whatever the cut-off point, more relevant data will eventually show up, casting new light on existing data.

Conclusion

Many of the trials and joys of working on the CED repeated themselves with the other two dictionary projects, though the process was somewhat speeded up by the experience gained on the first one and by the relatively restricted range of data to be integrated in the later two. Naturally there were differences engendered by, amongst other things, the various "local" philological traditions involved (the material was almost entirely in Russian as regards the CK project). In the case of Wakashan, more intensive fieldwork was required with the moribund language Ditidaht, which forms an important bridge between better known Makah to the south and Nuuchahnulth to the north. Many of the words needed for comparative purposes were hardly used any more on a daily basis, so much jogging of the memories of elderly speakers was necessary—the process, though sometimes frustrating, had its amusing and enjoyable moments for all concerned, especially in group sessions. An unexpected windfall in the form of the

extensive unpublished lexical material written down by John Thomas (now deceased) was made available to me by Barry Carlson of the University of Victoria and this made a considerable difference to the coverage of the language.

If I may be permitted to end on a subjective note, I would like to add that the putting together of a comparative dictionary seems to require a certain kind of mind-set on the part of the scholars involved. Apart from a sense of delight in both the remarkable regularity of changes in language form and the (very human!) vagaries of semantic change through time, this calls for a kind of tenacity bordering on the obsessive. Being addicted to jig-saw puzzles (something I readily admit to myself) may help. I have been lucky enough to have been in the right time and place to indulge that proclivity on a series of, I trust, useful scholarly endeavours in cooperation with numerous colleagues and native speakers who have a much greater expertise in the individual languages and dialects concerned than I have myself.

References

BERGSLAND, Knut

1986 Comparative Eskimo-Aleut phonology and lexicon, *Journal de la Société Finno-Ougrienne,* 82: 7-80.

FORTESCUE, Michael, Lawrence KAPLAN and Steven JACOBSON

1994 *Comparative Eskimo Dictionary with Aleut Cognates,* Fairbanks, Alaska Native Language Center, Research Paper, 9.

FORTESCUE, Michael

2005 *Comparative Chukotko-Kamchatkan Dictionary*, Berlin, Mouton de Gruyter.

2007 *Comparative Wakashan Dictionary*, Munich, Lincom Europa.

HAAS, Mary

1979 Overview, in Barbara S. Efrat (ed.), *The Victoria Conference on Northwestern Languages,* Victoria, British Columbia Provincial Museum, Heritage Record, 4: 1-14.

SAPIR, Edward and Morris SWADESH

1952 *Wakashan comparative vocabulary* , unpublished manuscript, Philadelphia, Boas Collection of the American Philosophical Society Library.