A NOTE ON STATISTICAL METHODS IN COMPARATIVE LINGUISTICS

Summary

It is desirable to distinguish between *lexicostatistics* as a means of proving relationships between languages or linguistic groups not previously known to be related, and *glottochronology* as a means of measuring the time depths of separations between languages or linguistic groups already known to be related.

For proof of relationships *structure statistics*, i.e. the application of statistical methods to comparisons of structural elements, needs to be stressed as an indispensable complement of lexicostatistics.

This is illustrated with tests.

There are, in the litterature on lexicostatistics, or glottochronology as it is also called, two things which have struck me as remarkable and worthy of comment. The first is that nowhere, as far as I can see, a clear distinction is made between what I would call glottochronology in the stricter sense of the word, i.e. a means of dating separations of languages or linguistic groups known to be related, and lexicostatistics as a means of proving genetic relationships between languages or groups not previously known to be related. This fact is, perhaps, best illustrated by the indiscriminate use of the two terms for the same technique. And yet "lexicostatistics" is a more general and neutral expression than "glottochronology" which implies an historical element that is lacking in "lexicostatistics". This merging of two distinct concepts may be partly due to Swadesh's own simultaneous treatment of them, although to the attentive reader the difference should be perceptible enough [1]). Also the subject has been treated largely by anthropological linguists who are often more interested in the time depth possibilities it offers than in its linguistic importance as such.

Apart from purely logical and methodological considerations – the time depth of a relationship can only be estimated after the re-

[1]) Compare for instance pp. 312 sqq. with p. 326 of Swadesh's "Perspectives and problems of Amerindian comparative linguistics" (*Word*, 10, 1954), where "the problem of proof" and "the problem of determining the order of relationship" are treated respectively.

lationship itself has been established – there is this in favour of making the distinction, that in large areas, such as New Guinea for instance, there is as yet no question of measuring the time depths of linguistic splits, but only of patient and prudent attempts at reducing the complex pattern of extreme linguistic diversity. And to this end lexicostatistics as a means of establishing and proving relationships can be of importance, while glottochronology as a means of measuring the time depths of the relationships is, linguistically, premature, as long as we have not even an outline picture of the latter.

There is, however, a second and more important argument in favour of making a distinction between the two concepts and, consequently, between the two terms. This is that many of the objections that can be raised against lexicostatistics are concerned with the assumption – a "discovery" according to Swadesh [2]) – that the basic vocabulary of a language tends to be replaced at an approximately constant rate, and that this rate is uniform for all languages. This assumption has not been proved. It was found correct in a test of only thirteen languages covering a history of little more than 2,000 years [3]), and Gudschinsky [4]) rightly doubts that the reality of the assumption has been adequately demonstrated since Swadesh himself [5]) has for various reasons discarded as invalid ten out of the thirteen test languages, although in the same article he continues to speak of the rate as uniform. I have myself [6]) criticized Swadesh's "discovery", adducing examples where the theory would not work (nota-

---

[2]) "Lexicostatistic dating of prehistoric ethnic contacts" (*Proceedings of the American philosophical society*, 96: 452–463, 1952). Cf. also Swadesh, "Archeological and linguistic chronology of Indo-European groups" (*Am. Anthrop.*, 55: 349 sqq., 1953), and Swadesh, "Mosan I: A problem of remote common origin" (*IJAL*, 19: 26–44, 1953).

[3]) See also R. Lees, "The basis of glottochronology" (*Language*, 29: 113–127, 1953).

[4]) "Three disturbing questions concerning lexicostatistics" (*IJAL*, 22: 212–3, 1956a). Cf. also her "The ABC's of lexicostatistics (glottochronology)" (*Word*, 12: 175–206, 1956b).

[5]) "Towards greater accuracy in lexicostatistic dating" (*IJAL*, 21: 121–137, 1955).

[6]) "Prospects of a 'Papuan' comparative linguistics" (*Bijdragen Kon. Inst.*, 113: 70–91, 1957).

bly as between Dutch and English and between Dutch and German). To this I can now add a comparison between Old Netherlandic (Old Low Frankish) of the so-called Carolingian Psalmfragments of Wachtendonck, dating from the beginning of the 10th century A.D. [7], with New Netherlandic (Dutch). This shows that in ca. 1.05 millennium out of 68 words in the old language, occurring in Swadesh's 100 word test list, fully 63 words, or 92,6%, were retained, or 61 words, or 89,7%, if the words for "many" which is *manage* in ON, and for "hair" which is **lok* in ON (gen. s. *lockis*, acc.p. *locka*) are not counted as retentions because NN *menige* and NN *lok*, though common words, are in every day language usually replaced resp. by *veel*, *vele* (= Eng. "much" and "many": ON *uilo*, *filo-* is rather = Eng. "much" than "many"), and by *haar* (NN *lok* is rather = Eng. "lock of hair"). The other "losses" are: ON *mikil*, NN *groot*, "big"; ON *luttika* (nom.p.), NN *klein*, "small" (although NN *luttel* still exists but in restricted and somewhat modified usage); ON *quethan*, NN *zeggen*, "say" (although ON *farseggon* occurs once in the sense of "excuse"); ON *thu*, NN *jij*, "thou"; and ON **reslân* (ger. *te reslâ:e*, 2 p.p. ind.praes. *reslât*), NN *doden*, "kill" (NN *slaan* = "strike, beat"; *verslaan* = "beat, defeat, slay"). This result, which means a retention rate of 93% or with a very strict count 90,2%, per millennium, is not only very different from the retention rate of approximately 81% per 1,000 years used for Swadesh's old 215 word test list, but also substantially higher than the rate of approximately 86% used for the 100 word list (Swadesh 1955: 127). It lies between the maximum rates recorded for German (90%) and Swedish (94,4%) which latter, according to Swadesh, there is reason to believe has been scored too high by about 5 percentage points (*ibid.*: 130–1). This result is the more significant since the ON Psalmfragments are in an eastern Low Frankish dialect, while modern Dutch is mostly based on a western Low Frankish dialect. This may perhaps account for such a "loss" as e.g. ON *thu*, since NN *jij* is due to Frisian or Ingweonic (i.e. western) substratum influences. Also, since the text is so fragmentary, chance may have played a role in that the simple verb

---

[7] See for this text recently H. K. J. Cowan, "De oudnederlandse (oudnederfrankische) psalmenfragmenten" (*Textus Minores*, ed. E. J. Brill, Leiden, 1957), and the litterature mentioned there.

**seggon* (cf. *farseggon* mentioned above) is not recorded as against *quethan* which is; the closely related Old Saxon, at any rate, has *seggian* as a synonym of *quedan*. Similarly OS has *grôt* (NN *groot*) as a synonym of *mikil*. The retention score for the 68 words used could, therefore, actually prove to be even higher if the material were richer than it is. Also I doubt that the percentage would be appreciably lower if the full test list of 100 words were available instead of only 68 words. In fact, out of an additional 13 words occurring in the older 215 word list only one "loss" was counted: "at", which is expressed by *an* in the ON text and by *te* in NN, *to*, *te* in ON being rather equivalent to Eng. "to". This addition, therefore, does not lower the retention rate. The least one can say is that there are reasons for grave doubt about the constancy and uniformity of the ratio for all languages.

It is not necessary, of course, to accept the validity of the theory in order to draw conclusions concerning *relative degrees* of relationship as soon as the relationships are proved and the percentages of cognates in the compared languages known. These percentages themselves supply the necessary data for that. What is impossible or not proved to be possible is the fixing of the degrees by *dating* the times of separation of the proto-languages. Speaking of "an *indicated* divergence time, representing both the duration and degree of separation" and which is "not necessarily an actual divergence time" (Swadesh 1954: 326), does not help much since the *time* measurement remains with its dubious elements.

Now the assumption of a constant and uniform retention rate is not, as is sometimes said (e.g. Gudschinsky 1956a: 212), the basis of *lexicostatistical* theory, but of *glottochronology* in the sense that I have just attributed to this term. The basis of *lexicostatistics* is the probability calculus. And against the validity or the usefulness of this calculus itself for linguistic purposes no serious doubts, as far as I can see, have been raised. Why should they? There is no reason why the calculus should not be as useful to provide a reliable measure of proof in linguistics as it has been for other disciplines, provided that it is translated in a form appropriate to that end. What is subject to discussion, therefore, are the *requirements* that should be postulated to make the probability calculus applicable for linguistic purposes. This being so, we can, if necessary or so desired, easily raise the

minimum standards for accepting proof. We can, with Fairbanks [8]) raise the phonetic expectancy ratio, and/or, with Kroeber [9]) use longer test lists (although Swadesh, 1955, in spite of Kroeber, *reduced* his original list in order to discard items that had proved to be either non-universal or not sufficiently non-cultural etc.), and/or, with Fairbanks, Gudschinsky (1956*b*) and W. Bright [10]), use stricter rules for judging word agreements if phonological correspondences are not previously known. But none of these criticisms seriously questions the lexicostatistic technique as such.

More serious seems to be Hoijer's critique [11]), according to which "none of the items of the test list are necessarily 'relatively neutral in their cultural implications' ...., to say nothing of their being 'universal and non-cultural'", and which judges it doubtful that any test list, though "taken as a whole (it) may well represent an area of experience common to all mankind, (can) also provide a division of this area into one hundred 'easily identifiable broad concepts which can be matched with simple terms in most languages'". Yet Hoijer himself made almost unqualified use of the technique for his later "Chronology of the Athapaskan languages". His criticism is, without doubt, too severe. He admits himself that "it is true, of course, that we can avoid items which are obviously culture-bound (like most kinship terms, or terms related to a particular political system or set of religious beliefs) and that items can be excluded which have reference to geographic and climatic factors of restricted distribution". Only he adds that "what cannot be avoided is the particular patterning of categories and types that a language and the culture to which it belongs impose upon lexicon and vocabulary..." I cannot see, however, – and Hoijer, in applying the method to Athapaskan, apparently does not think so himself – that this should invalidate any test list, let alone the entire technique. In its extreme consequence it would mean that *no* two comparative lists are ever

---

[8]) "A note on glottochronology" (*IJAL*, 21: 116–20, 1955).

[9]) "Linguistic time depth results so far and their meaning" (*IJAL*, 21: 91–104, 1955), especially p. 97.

[10]) "Glottochronologic counts of Hokaltecan material" (*Language*, 32: 42–48, 1956).

[11]) "Lexicostatistics: a critique" (*Language*, 32: 49–60, 1956). Cf. also Hoijer, "Chronology of the Athapaskan languages" (*IJAL*, 22: 219, 1956).

valid, whatever their composition and however large the number of agreements, and this would unsettle the basis for *all* comparative linguistics. But the final paragraph of his critique seems to indicate that it is the errors by overestimating vocabulary losses that he fears and the resultant errors in retention percentages and times of separation (i.e. errors in *glottochronology*) rather than errors in judgment of relationships.

Nevertheless, Hoijer's objections are important enough to question the advisability of using *lexical* statistics alone. And examples from more than one area increase this doubt. Thus in Malay at least 7 of the 97 "non-cultural" words of the test list used by Swadesh in his second experiment (1954) appeared to be Sanskrit loanwords, thus yielding a percentage of more than 7 of apparent cognates (Cowan, 1957). In the Achehnese language of North Sumatra there were 22% cognates with the Mon-Khmer languages of Indo-China as against 67% cognates with Austronesian or, where this language does not conform with Austronesian as reconstructed by O. Dempwolff, with Malay. The residue of ca. 10% could not be placed, possibly because of insufficient data in Mon-Khmer. For the close relative of Achehnese, the Cham language of Indo-China, the figures were not much different. In the case of these two languages the situation was further complicated because of the distant relationship that undoubtedly exists between the Austronesian and Mon-Khmer families as such, a fact that is also of importance for structural and morphological comparisons (Cowan, *ibid.*). In New Guinea *mainland* Melanesian languages little more than 35% of the vocabulary appeared to be Austronesian, although the distribution of the AUN material is not the same for every single language. This means that a considerable quantity of lexical material must belong to a Papuan substratum. This conclusion finds support in the fact that the Papuan substratum that Capell [12]) reconstructed in the Melanesian languages of South Eastern Papua includes no less than 37% of the words occurring in the 97 word list mentioned above, although these do not all occur in every single language (Cowan, *ibid.*). On the other hand non-AUN ("Papuan") languages sometimes show a considerable number of loanwords from AUN, which in e.g. Lawuna of S. E. Papua amounts to 14% of the

---

[12]) "The linguistic position of south eastern Papua", Sydney, 1943.

test list. A further complication is that sometimes even certain structural features of morphology and syntax are borrowed from AUN by Papuan languages or found in AUN (Melanesian) as Papuan substratum features (Cowan, *ibid.*).

As I pointed out for the Achehnese-Cham case, "the answer to the problem cannot be satisfactorily given by applying Swadesh's methods alone. An approach along broader lines is needed" (*ibid.*, p. 75). "However, as ... Swadesh's methods ... present a workable basis for the lexical comparison of languages that have diverged widely, they deserve(d) to be tried in the Papuan field" (p. 77). For "quite apart from the question whether comparison of grammatical structures alone will ever be sufficient to prove genetic relationships, there is the simple fact that for many languages only vocabularies and no grammars are as yet available" (p. 71). Recently Swadesh himself [13]) expressed similar ideas, saying that "in establishing a genetic relationship, either structure or lexicon may be used as the first clue, but both must be involved in the final proof".

This takes me to the second point: the application of statistical methods for *structural* comparisons. It is, at this point, only fair to recall that some early applications of statistics in linguistics were precisely in the fields of phonology, morphology and syntax [14]). But the tendency seems to have been towards re-adapting the techniques used there to lexical ends. Swadesh (1954: 313) has, in a footnote, expressed credit for Björn Collinder's first effort at a mathematical proof of relatedness, adding that this approach was based rather on structural elements than on vocabulary, as if this deserved less attention, although it is true that Swadesh did briefly point to such recurring elements in Amerindian as *n* for the first person pronoun and *m* for the second (*ibid.* 311–2). Kroeber and Chrétien [15]) improved Jan Czekanowski's "method of differential diagnosis by quantitative correlation determinations" for the classification of IE languages, resting "upon the recognition of isolable and definable features or

---

[13]) "Problems of long-range comparison in Penutian" (*Language*, 32: 22, 1956).

[14]) See the litterature mentioned by Bj. Collinder, "La parenté linguistique et le calcul des probabilités" (*Uppsala universitets årsskrift*, 1948), p. 1.

[15]) "Quantitative classification of Indo-European languages" (*Language*, 13: 83–103, 1937).

outside New Guinea. One of the Vogelkop languages in question, Ajamaru (or Mejbrat, as it is also called), yielded "only one dubious CVC comparison ... and two CV correspondences (with N. Halmahera) ... This makes no case. Relationship can be reconstructed by chain procedure ... though by a very round about way .... The chain relationship which can be constructed, however, finds support in the strongly suggestive and significant fact that the pronouns, including those that do not occur in Swadesh's 97 word list, show a close likeness to those in the NH languages ... notably also ... the pronominal prefixes in use in conjugation". This was shown in a table, and the conclusion added that "the chances that all this is coincidence are practically negligible" (Cowan 1957).

If we now test these pronominal subject prefixes by Swadesh's recently proposed more precise new method (1956: 38–9) we obtain the following results. We arrange the material in a table, this time using the nearly extinct Ibu dialect of North Halmahera only, because this has the more archaic forms as van der Veen [18]) has shown. The prefixes are mostly CV but, apart from Ibu 3 p.s.m. *o-* and 3 p.s. non-hum. *i-*, only the consonants are considered because the vowels are unstable in Ajamaru, dropping before stem vowel and sometimes also before consonant, and because their phonological relations cannot be determined.

person	*Ibu*	*Ajamaru*
1 s.	to-	t(e)-
2 s.	no-	n(e)-
3 s.m.	o-	j(e)-
3 s.f.	mo-	m(e)-
3 s.n.h.	i-	—
1 p.ex.	mi-	—
1 p.in.	ne-	n(o)-
2 p.	ni-	n(e)-, b(o)-
3 p.	na-	n(e)-, m(e)-

In some NH dialects the 3 p.s.m. is *wo-*, the 1 p.p.in. *wo-*, *po-*, *fo-*,

[18]) "De noordhalmaherase taalgroep tegenover de austronesiese talen" (doct. thesis, Leiden, 1915), pp. 180, 184, 187 and 188–9. For Ajamaru see J. F. Elmberg, "Notes on the Mejbrat people of the Ajamaroe district" (Stockholm, 1954, *mimeographed*), and Cowan, 1957.

ho-, and the 3 p.p. *jo-*, *ǧo-*, *du-*. For non-human things in the plural the singular is used in NH and it can, therefore, be omitted from the comparison. In Ajamaru the 2 and 3 p.p. each have two different forms in different conjugations, which will be included in the calculation in order to observe strict objectivity. In Ajamaru no separate exclusive form of the 1 p.p. exists. In this language also there is no special form for the 3 p.s. non-hum., the 3 p.s.f. being used to that end. Apparently the distinction here is between male humans and everything else as against the NH distinction between humans, subdivided in males and females, and everything else. Since, therefore, class contents appear to have shifted a little, it is only fair to count Ibu 3 p.s.n.h. *i-* and Ajamaru 3 p.s.m. *j(e)-* as corresponding because the shift in class content may very well have been accompanied by a shift in forms. This is, moreover, largely confirmed by the fact that the incorporated *object* affix for the 3 p.s.m. actually is *-i-*, *-wi-* in NH. Also the 3 p.p. form *jo-*, *ǧo-* of other NH dialects seems to be related.

We have, then, $9 \times 9 = 81$ total pairs involved in the comparison, with 22 possible agreements if meaning is not considered, viz. I. *to* = A. *t(e)*; I. *no*, *ne*, *ni*, *na* each = A. $3 \times n(e)$ and $1 \times n(o)$; I. *mo* and *mi* each = A. $2 \times m(e)$; I. *i* = A. *j(e)*. This makes a chance expectancy ratio of $\frac{22}{81} = 27{,}16\%$. The actual number of correspondences if meaning is taken into account is, however, 7 out of 9, or 77,77%; that is more than $2\frac{4}{5}$ times what could be expected from chance. Even without admitting I. *i* = A. *j(e)* the actual agreement percentage is 66,66%, that is nearly $2\frac{1}{2}$ times the chance expectancy.

The chance expectancy ratio, as calculated according to Swadesh's new method, was checked empirically with a simple experimental device, namely the use of two sets of 9 "Scrabble" game letters each, corresponding to the two sets of single prefix phonemes involved in the comparison under discussion. The two sets were set apart in two heaps with their backs up, and, after shuffling, two letters, one from each heap, were taken simultaneously and at random, and turned. The number of times that an identical pair ("agreement") showed on turning, was noted. This experiment was repeated a number of times planned in advance, viz. 50, so that $50 \times 9 = 450$ pairs of letters have been turned. The result was 114 agreements, or 25,33% of 450; the average was, therefore, $\frac{114}{50} = 2{,}28$ agreements per experiment, or 25,33% of 9. There is, then, only little discrepancy between the

calculated and the empirically determined chance expectancy ratios; if they differ, the former does so on the safe side rather than otherwise. In this connection it is, perhaps, worth noting that after the first ten experiments the chance ratio was 23,33%; after twenty: 27,77%; after thirty again 27,77%; and after forty: 25,55%. The 114 agreements in 50 experiments consisted of 2 × 5, 6 × 4, 8 × 3, 22 × 2, and 12 × 1. The maximum of potential chance agreement (obtained only twice) was, therefore, 5 out of 9, or 55,55%. This is still substantially less than our actual agreement percentage of 77,77, or 66,66 if I. i = A. $j(e)$ is not admitted.

Against the argument as set out above possibly two methodological doubts may be raised. One is whether in calculating the chance expectancy ratio meaning is really not taken into account, considering that it is *personal* elements we consciously choose to compare. And the other is whether the list of elements is a sufficiently large and objectively chosen sample. As to the first objection I would point out, however, that *any* two test lists, whatever their contents, are necessarily made up for the meanings of their items. What is meant, of course, is the import of meaning when *comparing* the two lists. The second objection seems stronger, but it must be remembered that lexical items consist of phoneme complexes, while the present material consists of single phonemes only. The objectivity requirement was taken into account as well as possible by including in the calculation the 2 and 3 p.p. prefixes of different conjugations in Ajamaru.

However, for confirmation we now shall also test the material with a more elaborate method similar to that of Collinder's. As Vendryes [19]) has said, "le nombre des phonèmes possibles s'étend presque à l'infini .... Mais il s'en faut beaucoup que les langues utilisent à la fois toutes les ressources du langage. Le nombre des phonèmes de chaque langue est au contraire assez limité". Bloomfield [20]) says that "the number of simple primary phonemes in different languages runs from about fifteen to about fifty". The average per language can therefore, be put at ca. thirty. Collinder put the total of possible phonemes hypothetically at 300 and the average per language at 50, but in these he includes complex phonemes. For

---

[19]) "Le langage" (1950), p. 40.

[20]) "Language" (1950), p. 90.

simple primary phonemes we may safely put the number at a total of 100; this prevents a too favourable presentation. Our average of 30 per language is well in keeping with the number in NH (28, see Cowan 1957: 87), while Ajamaru, which does not distinguish between long and short vowels and between voiced and unvoiced plosives, has less (Elmberg 1954: 1, 10).

The number of known linguistic families in the world is estimated at ca. 150, but some 4/5 of these are small Amerindian groups which are, moreover, in a permanent course of reduction. We can, therefore, safely take into account only a number of 30, the more so considering the isolation ever since neolithic times of Papuan by Austronesian to the West, North and East, and by Australian to the South.

These figures are tabulated in a table consisting of 30 horizontal lines (for the 30 linguistic families) and 30 vertical columns (for the average number of phonemes per language), thus forming 900 squares. A pack of 100 numbered cards (for the total of possible phonemes) is prepared, the pack is shuffled and 30 cards of it are turned one after the other for the 30 squares in the first horizontal line, in each of which the corresponding number of the card is noted. This procedure is then repeated 29 times (for the other 29 horizontal lines), each time, remixing the 30 cards with the 70 remaining cards of the pack, and taking good care that the complete pack is reshuffled very well every time. What is the probability that two lines at least will show a minimum of 4 correspondences? For safety we count only 4 correspondences instead of 7, viz. I. *to* = A. *t(e)*; I. *i* = A. *j(e)*; I. *n* = A. *n* only once; and I. *m* = A. *m* only once, because the other 3 correspondences do not contain different phonemes. The results were as follows: in 72 cases one agreement, in 15 cases two agreements, and in 3 cases 3 agreements as between any two lines. Not once 4 agreements occurred, and the maximum of 3 was so rare that we may safely state that this experiment confirms the findings of the other test.

These results make a very good basis for assuming that the agreements between Ibu and Ajamaru are not due to chance; and if not to chance, to what else than common origin, since there can hardly be any suspicion of borrowing here, nor of sound-imitation or cultural binding. Also, no objections like those against non-material features can be raised here. And in any case the results are a positive confir-

mation of the chain relationship that could be arrived at in a very round-about way by applying lexical statistics. This case shows that sometimes a connection can be demonstrated by structural elements even when the vocabulary does not give the slightest suggestion of relationship. Incidentally, a case like this makes an attempt at glottochronological dating illusory, for what is the use of a vocabulary retention rate where no vocabulary retention is apparent?

Tolsteegplantsoen 26^{II}, Utrecht H. K. J. COWAN