ENRICO CAMPANILE - ANTONIO ZAMPOLLI

PROBLEMS IN COMPUTERIZED HISTORICAL LINGUISTICS: THE OLD CORNISH LEXICON *

This work represents an attempt to utilize the computer in solving problems in historical linguistics.

The *corpus* upon which it operates is not a language but a recently published etymological dictionary of Old Cornish.[1] Any observations regarding the scarcity or inaccuracy of the data utilized are, therefore, irrelevant, as far as the present paper is concerned.

As the dictionary in question was compiled according to the usual methods employed with such works, a detailed explanation of methodology is unnecessary. It should also be noted that Old Cornish is known only through glosses to Latin words, and that in this case « Cornish gloss » is equivalent to « Cornish word ».

With the help of the computer, we have attempted to solve the following problems:

a) To establish the percentage of words with and without Indo-European etymology in the Cornish lexicon. (Let us stress that this study concerns not a language but an etymological lexicon; hence, the presence or absence of Indo-European etymology should not be construed as a definitive characteristic of a Cornish word. Such statistics are, in fact, relevant only to the present state of research on the subject).

b) To establish the degree of certainty concerning the material of Indo-European etymology.

c) To evaluate the extent of the connection between elements of Indo-European etymology existing in the Cornish lexicon and the other Indo-European linguistic groups according to the degree of certainty of each individual etymology.

d) To establish, on the basis of existing etymological studies of Old Cornish, the lines future research should follow.

---

* The computational part of this research has been conducted by A. Zampolli, the historical linguistics part by E. Campanile.

[1] E. CAMPANILE, *Profilo etimologico del Cornico antico*, Pisa, 1974. Also in *SSL*, XIII (1973), p. 1.

The reader will observe that the first problem is purely statistical (though it has an obvious diachronic premise), that the second aims at attaining qualitative data (though they are expressed quantitatively), that the third concerns the area of Indo-European dialectology, and that the fourth has its own specific heuristic and methodological significance.

In order to accomplish these goals, the contents of the etymological dictionary were put on cards, each of which contained the following entries:

a) a non-Cornish word (with an indication of the language to which it belongs);

b) the Cornish word related in the dictionary to the item under *a*);

c) the type of relationship existing between item *a*) and item *b*); and whether this relationship is affirmed, denied or uncertain;

d) the indication that item *b*) is or is not a nominal compound (this being the only type of compound found in Old Cornish);

e) in the event that item *b*) is a nominal compound, a breakdown of the elements contained in it; [2]

f) the page from which the foregoing material was taken.

With regard to item *c*), the possible types of relationships have been described (see below) according to the information supplied, either explicitly or implicitly, by the etymological dictionary and have been rated according to the following numerical system:

1 = the relationship between the two words is etymologically certain.
2 = » » very probable.
3 = » » probable
4 = » » doubtful
5 = » » not very probable
6 = » » improbable
7 = » » non-existent
8 = the Cornish word was borrowed from item *a*)
80 = » » is a calque from item *a*)
82 = a relationship exists between the Cornish word and item *a*), but the nature of the relationship cannot be determined exactly (that is, whether it is a matter of kinship or loan).[3]

---

[2] Every element has been given either in the Cornish form (if it is attested elsewhere in the text or if it is not attested only be cause of lack of documentation), or in the common Celtic form or in the Indo-European form; certain diacritic signs indicate which possibility has been chosen.

[3] The distinction between borrowed words and co-radicals is that provided by the etymological dictionaries and handbooks of historical linguistics. Since the difference

9 = the Celtic co-radical of the Cornish word (this rating prevails over ratings 1,2 and 3 because the prime object of the present research is Indo-European etymology rather than the Celtic connections of Cornish).

To these eleven ratings will be added that of 0 which will not indicate the relationship between Cornish and non-Cornish voices, as in the case of the other ratings, but will serve instead to distinguish the non-Cornish words (actually, Cymric) which, due to the various vicissitudes of the handwritten tradition, have crept into the authentic Cornish glosses and which, as such, do not form part of the present study.

The following items, taken from the etymological dictionary, and their ratings illustrate the preceding principles:

FORN gl. *fornax l. clibanus* 920. Come il bret. *fo(u)rn* (ant. bret. *gufor(n)* gl. clibani), il cimr. *ffwrn* e l'irl. *sorn*, è prestito dal lat. *furnus*. HV, 179; VG, 221; LH, 274; VB, 190.

FRIIC gl. *nasus* 30. Formazione in -IC (con originario valore, forse, diminutivo), da compararsi con bret. *fri* « naso». Non è da escludersi un rapporto con formazioni (originariamente onomatopeiche) in **sr-* designanti il russare e il naso; cf. gr. ῥέγχω, arm. *ṙngunk'* etc. IEW, 1002.

FROT gl. *alueus* 737. Identico a bret. *froud* « torrente », cimr. *ffrwd* « corrente », irl. *sruth* (gen. *srotha*) « fiume, corrente », gall. Φρουδις (leggi Φρουτυς), tutti da **sprutu-*. Ma il confronto con lit. *spriaũnas* « fresco », ted. *spröde* « secco » non è semanticamente convincente. Il termine sopravvive anche nell'ital. dial. *froda* « torrente » (REW, 3545), VG, 35; Pokorny, *Celtica* 3, 1956, 308; LH, 541; Meid, IF 65, 1960, 39; IEW, 994.[4]

---

between the two concepts exists only as a chronological distinction, the problem is, therefore, irrelevant. Cf. V. PISANI, *Parenté linguistique*, in « Lingua », (1952), p. 3 (or *Saggi di linguistica storica*, Torino, 1959, p. 29) and *Variazioni sul problema indoeuropeo*, in *Lingua e culture*, Brescia, 1969, p. 21.

[4] FORN gl. *fornax l. clibanus* 920. Like the Breton *fo(u)rn* (OBr. *gufor(n)* gl. clibani), the Cymr. *ffwrn* and the Irish. *sorn*, was borrowed from the Latin *furnus*. HV, 179; VG, 221; LH, 274; VB, 190.

FRIIC gl. *nasus* 30. Formation in -IC (originally, perhaps, diminutive), is comparable to Breton *fri* « nose ». They may also have kinship with formations, originally onomatopoeic, in **sr-* which designate both snoring and nose; cf. gr. ῥέγχω, arm. *ṙngunk'* etc. IEW, 1002.

FROT gl. *alueus* 737. Identical to Bret. *froud* « brook », Cymr. *ffrwd* « stream », Irish *sruth* (gen. *srotha*) « rover, stream », Gaul. Φρουδις (read Φρουτυς), all from **sprutu-* But the comparison with Lit. *spriaũnas* « cool », Germ. *spröde* « dry » is not semantically convincing. The term survives in Italian (dial.) *froda* « brook » (REW, 3545). VG, 35; Pokorny, *Celtica* 3, 1956, 308; LH, 5, 1; Meid, IF 65, 1960, 39; IEW, 994.

*

These three paragraphs gave the following 15 cards:

br. *fo(u)rn* [5]	9 [6]	*forn*	47
a. br. *gufor(n)*	9	*forn*	47
cim. *ffwrn*	9	*forn*	47
irl. *sorn*	9	*forn*	47
lat. *furnus*	8	*forn*	47
br. *fri*	9	*friic*	47
gr. ῥέγχω	3	*friic*	47
arm. *ṙngunk'*	3	*friic*	47
br. *froud*	9	*frot*	47
cim. *ffrwd*	9	*frot*	47
irl. *sruth*	9	*frot*	47
gall. Φρουτυς	9	*frot*	47
lit. *spriaũnas*	5	*frot*	47
ted. *spröde*	5	*frot*	47
ital. dl. *froda*	1	*frot*	47

All the words with an index of 0 were eliminated prior to the operation. The analysis of compounds was found to be a particular problem. When the rating was carried out, the section of the compound with a kinship with the non-Cornish word *a*) was indicated (and hence a numerical rating was given). For example, the following paragraph:

HEWUIL gl. *uigil* 401. Composto dal prefisso celt. **so-* « bene, buono » (ant. bret. *ho-*, *hu-*, *he-*, ant. cimr. *hi-*, *he-*, *hu-*, irl. *su-*, *so-*) simile ma non identico al scr. *su-*, gr. ὑ- (in ὑγιής da **su-g*^{w}*iḭ̯ēs* « che vive bene ») e da **guil* « veglia » (= cimr. *gŵyl* « festa », bret. *goel* « id. », irl. *féil* « id. », tutti dal tardo lat. *uēlia*, per *uigilia*). HV, 140; VG, 214; LH, 463 e 659.[7]

yielded the following 14 cards:

a. br. *ho-*	= 9 *hewuil*	(°°*so* °*guil*)	64
a. br. *hu-*	= 9 *hewuil*	(°°*so* °*guil*)	64

[5] Column reserved for information concerning nominal compounds.

[6] Column reserved for the analysis of nominal compounds.

[7] HEWUIL, gl. *vigil* 401. Composed by the Celtic prefix **so-* « well, good » (Old Bret. *ho-*, *hu-*, *he-*, Old Cymr. *hi-*, *he- ho-*, *hu-*, Irish *su-*, *so-*), similar but not identical to Scr. *su-*, Gr. ὑ- (in ὑγιής from **su-g*^{w}*iḭ̯ēs* « that lives well ») and by **guil* « vigil » (= Cymr. *gŵyl* « feast », Bret. *goel* « id. », Irish *féil* « id. », all from late Latin *uēlia*, equal to *vigilia*). HV, 140; VG, 214; LH, 463 and 659.

a. br. *he-*	= 9 *hewuil*	(°°*so* °*guil*)	64
a. cim. *hi-*	= 9 *hewuil*	(°°*so* °*guil*)	64
a. cim. *ho-*	= 9 *hewuil*	(°°*so* °*guil*)	64
a. cim. *hu-*	= 9 *hewuil*	(°°*so* °*guil*)	64
irl. *su-*	= 9 *hewuil*	(°°*so* °*guil*)	64
irl. *so-*	= 9 *hewuil*	(°°*so* °*guil*)	64
scr. *su-*	= 5 *hewuil*	(°°*so* °*guil*)	64
gr. ὑ-	= 5 *hewuil*	(°°*so* °*guil*)	64
cim. *gŵyl*	— 9 *hewuil*	(°°*so* °*guil*)	64
br. *goel*	— 9 *hewuil*	(°°*so* °*guil*)	64
irl. *féil*	— 9 *hewuil*	(°°*so* °*guil*)	64
lt. volg. *u̯ēl̥ia*	— 8 *hewuil*	(°°*so* °*guil*)	64

(Note: in the preceding table, the sign = indicates that the kinship of word (*a*) is with the first part of the Cornish compound; the sign – indicates that the kinship is with the second part; the sign °° indicates that the given form of the first member of the dissolved compound is referable to the common Celtic period; and the sign ° indicates that the word does not happen to be attested).

But, from the point of view of historical linguistics, it is evident that, while *guil* has not been attested as an autonomous form merely because no documentation happens to be available on the subject, *he-* existed (and always has existed) only as a member of a compound. Nevertheless, while *guil* could possibly be included among the autonomous lexical elements of our text, *he-* could only be found among the morphemes. And finally, the compound *hewuil*, as a creation of the Cornish (or Celtic) age, has no precise equivalents in other Indo-European languages, and any equivalents that happen to exist may be considered *a priori* only the result of chance.

The task of analyzing compounds is further complicated by the presence of words (the Latin *credere*, for instance) that from a diachronic point of view are compounds while from a synchronic point of view they are not.

For the reasons just stated, we decided to eliminate the compounds from the present analyses and to make them the object of a separate study.

Thus, in addition to the words with a rating of 0, entries containing the signs = and-or – have also been discarded.

After the words with a rating of 0 and the nominal compounds were discarded, the surviving Cornish material consisted of 745 elements that, in relation to our first problem, were subdivided in the following way.

words of Indo-European etymology [8]	284	38 %
words borrowed from other languages [9]	254	34 %
calques from other languages [10]	0	0 %
uncertain kind of kinship [11]	0	0 %
words without Indo-European etymology [12]	207	28 %
	745	100 %

With regard to the second problem, the 284 words of Indo-European etymology were divided according to the degree of probability. The breakdown is as follows:

words of certain etymology [13]	238	84 %
words of very probable etymology [14]	23	8 %
words of probable etymology [15]	23	8 %
	284	100 %

In order to solve the third problem, all the entries containing non-Cornish words correlated to one of the 284 Cornish words of Indo-European etymology were taken into consideration. These entries (742 in all) were subdivided into 17 groups according to the linguistic kinship of the language to which the word in item (*a*) belongs:

1 = Tocarian A and B
2 = Sanskrit, Avestan, Persian (Aryan group)
3 = Armenian

---

[8] Words carrying at least one kinship index of 1,2 or 3.

[9] Words carrying a kinship index of 8. The indices 80 and 82 are found only among the nominal compounds.

[10] Words carrying a kinship index of 80. This is found only among compounds.

[11] Words carrying a kinship index of 82. This is found only among compounds.

[12] Words carrying a kinship index of 4 and/or 5 and/or 6 and/or 7 (eventually with a 9).

[13] Words carrying at least one kinship index of 1.

[14] Words carrying no index of 1 and at least one of index 2.

[15] Words carrying no indices of 1 or 2 and at least one index of 3.

4 = Hittite
5 = Phrygian
6 = Greek
7 = Macedonian
8 = Illyrian
9 = Old Slavonic, Old Czech, Russian, Ukranian, Serbian, Middle Bulgarian (Slavonic group)
10 = Albanian
11 = Lithuanian (old and modern), Old Prussian, Lettish (Baltic group)
12 = Old English, Middle English, Danish, Old Icelandic, Old Gutniac, Dialectal Norwegian, Swedish, Old High German, Middle High German, German (modern), Longobard, Gothic, Burgundian (Germanic group)
13 = Ligurian
14 = Oscan, Umbrian
15 = Latin, Vulgar Latin, Medieval Latin, Italian, Dialectal Italian, Old French, Catalan, Old Spanish, Spanish (Latin and neo-Latin group)
16 = Breton, Middle Breton, Old Breton, Cymric, Middle Cymric, Old Cymric, Old Irish, Modern Irish, Ogamic, Scottish, Gaulish, Galatian, Latin-Gaulish, Latin-British, Vannelais (Celtic group)
17 = Finnish, Vogulian

The reader will notice that not all Indo-European languages are represented here. This is due to the fact that not all Indo-European languages are represented in the etymological dictionary that provided the material for the present work. On the other hand, there are two non-Indo-European languages in group 17 because one Cornish word is thought to have a kinship with non-Indo-European words.

Each of the 742 words has an etymological kinship with Cornish words that is either certain (rating 1), very probable (rating 2) or probable (rating 3). These words were arranged into linguistic groups with the rank of 1 going to the group that had at least one exponent with a rating of 1, the rank of 2 going to the group with at least one exponent with a rating of 2, and the rank of 3 to the group with neither rating. Here are the results:

	r. 1	r. 2	r. 3	tot.	% r. 1	% r. 2	% r. 3	% of tot.
GR. 1	12	1	0	13	0.9231	0.0769	0.0	0.0175
GR. 2	95	9	9	113	0.8407	0.0796	0.0796	0.1523
GR. 3	24	0	3	27	0.8889	0.0	0.1111	0.0364
GR. 4	11	0	1	12	0.9167	0.0	0.0833	0.0162
GR. 5	0	0	0	0	0.0	0.0	0.0	0.0
GR. 6	95	6	8	109	0.8716	0.0550	0.0734	0.1469
GR. 7	0	1	0	1	0.0	1.0	0.0	0.0013
GR. 8	1	0	0	1	1.0	0.0	0.0	0.0013
GR. 9	45	4	2	51	0.8824	0.0784	0.0392	0.0687
GR. 10	11	3	2	16	0.6875	0.1875	0.1250	0.0216
GR. 11	71	9	5	85	0.8353	0.1059	0.0588	0.1146
GR. 12	154	8	8	170	0.9059	0.0471	0.0471	0.2291
GR. 13	1	0	0	1	1.0	0.0	0.0	0.0013
GR. 14	10	0	0	10	1.0	0.0	0.0	0.0135
GR. 15	108	6	12	125	0.8571	0.0476	0.0952	0.1698
GR. 16	3	1	3	7	0.4286	0.1429	0.4286	0.0094
GR. 17	0	0	0	0	0.0000	0.0	0.0000	0.0
Tot.	641	48	53	742	0.0002	0.0		

It will be observed that the Celtic group appears to be poorly represented in that the existing etymological kinship with Cornish words is normally expressed by the rating 9 (and not, therefore, 1,2 or 3), while the rank of 1,2 or 3 has been attributed only to those words which, though still within the Celtic group, are part of other linguistic traditions (words of the Celtic substratum in the romance languages, for example).

An analogous operation was then carried out with all the material having a rating of 4,5,6 or 7 (that is, negative etymologies). The results are as follows:

	r. 4	r. 5	r. 6	r. 7	tot.	% r. 4	% r. 5	% r. 6	% r. 7	% of tot.
GR. 1	0	0	0	0	0	0.0	0.0	0.0	0.0	0.0
GR. 2	3	4	9	10	26	0.1154	0.1538	0.3462	0.3846	0.1300
GR. 3	0	0	5	1	6	0.0	0.0	0.8333	0.1667	0.0300
GR. 4	0	0	0	0	0	0.0	0.0	0.0	0.0	0.0
GR. 5	0	0	0	0	0	0.0	0.0	0.0	0.0	0.0
GR. 6	3	5	8	11	27	0.1111	0.1852	0.2963	0.4074	0.1350
GR. 7	0	0	0	0	0	0.0	0.0	0.0	0.0	0.0
GR. 8	0	0	0	1	1	0.0	0.0	0.0	1.0000	0.0050

GR. 9	2	0	2	1	5	0.4000	0.0	0.4000	0.2000	0.0250
GR. 10	0	2	0	0	2	0.0	1.0000	0.0	0.0	0.0100
GR. 11	2	5	2	6	15	0.1333	0.3333	0.1333	0.4000	0.0750
GR. 12	1	6	5	5	17	0.0588	0.3529	0.2941	0.2941	0.0850
GR. 13	0	0	0	0	0	0.0	0.0	0.0	0.0	0.0
GR. 14	0	0	0	0	0	0.0	0.0	0.0	0.0	0.0
GR. 15	2	5	15	30	52	0.0385	0.0962	0.2885	0.5769	0.2600
GR. 16	2	6	8	33	49	0.0408	0.1224	0.1633	0.6735	0.2450
GR. 17	0	0	0	0	0	0.0	0.0	0.0	0.0	0.0
TOT.	15	33	54	98	200					

From all this material it was possible to draw the following conclusions:

1) In the Cornish lexicon there are 254 (34 %) lexical loan-words, there are 284 (38 %) words with Indo-European etymologies, and 207 (28 %) without any known etymology.

2) The vast majority of the words with an Indo-European etymology (238 out of 284 = 84 %) have an etymology that is certain, as far as is known at the present state of research on the subject. Another 16 % have etymologies that are either very probable (23; 8 %) or merely probable (23; 8 %).

3) With regard to etymological kinships with non-Celtic Indo-European linguistic groups, the closest connections are with German (0.2291), Latin (0.1698), Indo-Aryan (0.1523), Greek (0.1469) and with Baltic (0.1146). Such results appear to be extremely important in that they confirm the innovative character of the occidental lexicon (kinships with German, Latin and, at least in part, Baltic) existing side with the preservation of archaic elements in lateral areas (kinship with Indo-Aryan), thereby showing strong kinships with the central area of the Indo-European world (Greek and, at least in part, Baltic) which have yet to be adequately assessed.

4) The highest percentages of now unacceptable relationships suggested by scholars in the past are those with Latin (0.2600), Greek (0.1350) and Indo-Aryan (0.1300). This, together with the fact that these same groups have also yielded a very high percentage of acceptable etymologies, suggests that these areas have been exhausted. As working hypothesis, new etymological comparisons ought now to be considered particularly with German and Baltic, which combine a high yield with a more tolerable percentage of acknowledged errors (0.0850 and 0.0750 respectively).

5) Of the 745 Cornish words which have supplied the material for the present study, as many as 671, almost 90 %, bear at least an index of 9; that is, have one or more Celtic co-radicals. This confirms the « compact » character of the Celtic lexicon.

Moreover, there are Cornish words which have one or more indices of 9 to the exclusion of any other index (139; 18 %). These are words that have co-radicals exclusively in the Celtic world. On the heuristic level, this verification gives rise to a question that is at the same time a working hypothesis: are they substratum words?

The same question and the same working hypothesis also arise with the words where one or more indices of 9 accompany the indices 4, 5, 6, 7: these are words with Celtic co-radicals formerly thought to be of Indo-European etymology but now refuted in the dictionary, They are 60.

Our analysis, therefore, seems to suggest, too, that future linguistic research will find rich material for substratum studies in Cornish and, more generally, in Celtic.