Celtic Words in English Dictionaries and Corpora[1]

Mitsuhiko Ito
Toyohashi University of Technology; Professor Emeritus

The researcher has collected Celtic words from several English dictionaries and a few English etymological dictionaries and found that there are about 300 words in present English dictionaries. He has studied what words and how many of the 300 words native speakers of English know. The research method was giving matching tests of words and definitions and having subjects write appropriate words to definitions. The subjects were all adult voluntaries. Main purposes of the present study are: (1) to survey what words of the 300 words appear in BNC and Wordbnaks, and (2) to survey if well known words by native speakers are highly frequent words in BNC and Wordbanks. Two main results have deduced from the present study. One is that not all of the 300 words appear in BNC and Wordbanks and some words appear in the two Corpora and some others appear in either of the Corpora and the others do not appear in both of them. The other is that well known words in the research do not necessarily come to the top frequent positions of the Corpora.

1. Introduction

The researcher has collected Celtic words from several English dictionaries and a few of English etymological dictionaries and found that there are 312 words in present English dictionaries (Ito, 2005). He has studied what words and how much of the 312 words native speakers of English know. The research method was giving matching tests of words and definitions and having subjects write appropriate words to definitions (Ito, 2006). One question arose after the study: is the result similar to frequencies of words in corpora?

2. Purposes of the present study

Main purposes of the present study are: (1) to survey what words of the 312 words appear in BNC and Wordbnaks, and (2) to survey if well known words by native speakers are highly frequent words in BNC and Wordbanks. Other relevant topics might be touched on through discussion based on data collected from the Corpora and previous work by Ito (2006).

3. Procedure of the research

The researcher has collected 312 words. Therefore, firstly, the 312-word list is made for the next step of research. Secondly, the researcher examines if each word of the 312-word list can be found in BNC and Wordbanks respectively. Thirdly, if he can find words of the 312-word list in BNC or Wordbanks, he records frequencies of those words in each corpus. Finally, he compares frequencies of words of 312-word list based on questionnaire with those of words based on the Corpora. Thus, the researcher obtains results for discussion.

3.1. 312-Word list collected from dictionaries

Three hundred and thirteen words which the researcher collected from English dictionaries are as follows:

[1] This study was supported in part by a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (Foundation Studies (C) ; No. 20520438).

ach, airt, ambassador, anchor, andiron, ap-, arrah, arpent, ass, bal, ball, bandle, bannock, banshee, bard, basin, bawn, beak, beal, beet, Beltane, ben, betony, billet, bijou, bin, biretta, birlinn, bodkin, bodrag, bog, bonaght, bonnyclabber, booly, bothy, bouge, bowssen, bracken, bragget, brail, brasserie, brat, brattach, bray. brehon, brier, brisance, brock, brogan, brogue, brut, buckeen, bulge, bullace, bushel, butcher, caber, cade, cader, cailleach, cain, caird, cairn, cam, cambrel, cammock, cannach, cant, cantankerous, capercailye, car, carpenter, carrow, carry, caschrom, cashel, cateran, caubeen, ceilidh, change, char, char, clabber, clairschach, clachan, clan, claye, claymore, cleave, coarb, coble, coccagee, cockabondy, colleen, collop, commot, coracle, corcass, corgi, cork, corm, coronach, corrie, cosher, coshery, costean, cot, coyne, crag, crannog, cranreuch, creagh, creaght, crine, cro, cromlech, cross, crostarie, crottle, crowd, cudden, cuddy, culdee, currach, cwm, Dail Eireann, dalt, deasil, dartre, dillue, dolmen, dour, down, drape, druid, drum, duan, dudeen, dulse, dun, duniwassal, eisteddfod, elvan, embassy, encumber, eric, esker, fail, fewterer, filibeg, fiorin, flannel, flummery, font, frown, gad, galliard, galloglass, gallon, galore, garron, garrote, garter, gelt, gillie, glean, glen, glib, gob, gob, goblet, gombeen, gossan, gouge, gralloch, grave, gravel, grig, grouse, growan, guillem, gull, gwiniad, hog, hubbub, ieroe, inch, ingle, iron, javelin, keen, keen, kern, kex, kibe, killas, kish, kistvaen, knock, kyle, lance, lawn, lay, league, lech, lee, leprechaun, linn, loch, lochan, lough, lough, loy, lozenge, lymphad, marl, menhir, messan, metheglin, minaudiere, mine, minion, mooch, morgay, morglay, mormaor, muley, mullein, musha, mutton, och, ohone, ollamh, ouch, ovate, oy, palfrey, partan, peat, pen, pendragon, pennill, peulvan, pibroch, piece, pikelet, pillion, plaid, pollan, port, poteen, ptarmigan, puffin, punt, quaich, rap, rapparee, rath, rich, ruche, sagum, say, scraw, sennachie, shamrock, shanty, shebeen, sivvens, skene, slat. Slob, slogan, sock, socket, sonsy, sorran, sowens, spalpeen, spleuchan, sporran, spreagh, stannum, strath, tais(c)h, tan, tanist, tarrier, termon, tinchel, tocher, toman, tope, tor, Tory, towan, trews, trouse, truant, tuath, tun, union pipes, usquebaugh, vassal, vendace, vergobret, vug, warren, weem, whisky, wirra, wrasse (Ito, 2008).

3.2. Words found in BNC and Wordbanks

The researcher examined if each of the 312 words is found in BNC and Wordbanks respectively. Then, he found 144 words in BNC and 128 words in Wordbanks.

144 Words in BNC:

ach, ambassador, anchor, andiron, arpent, ass, bal, bannock, banshee, bard, basin, beak, beal, beet, Beltane, betony, billet, bijou, bin, birlinn, bodkin, bog, brasserie, brogan, brogue, brut, bulge, bushel, butcher, cade, cader, cailleach, cairn, cantankerous, car, carpenter, carry, cateran, ceilidh, change, clachan, clan, claymore, coble, colleen, coracle, corgi, corm, corrie, cromlech, cross, currach, cwm, Dail Eireann, dolmen, dour, drape, druid, dulse, eisteddfod, elvan, embassy, encumber, , fail, filibeg, flannel, flummery, frown, gallon, galore, garron, garter, gillie, glean, glen, goblet, gossan, gouge, gralloch, gravel, grouse, guillem, gull, hog, hubbub, ingle, iron, javelin, keen, keen, kern, killas, lance, lawn, leprechaun, linn, loch, lochan, lozenge, lymphad, menhir, mine, minion, morglay, mormaor, muley, mullein, mutton, och, palfrey, peat, pendragon, pibroch, piece, pillion, plaid, poteen, puffin, quaich, rath, rich, say, sennachie, shamrock, shanty, shebeen, slogan, socket, sowens, spalpeen, stannum, strath, tan, tocher, toman, tor, Tory, trews, usquebaugh, vassal, vug, warren, weem, whisky.

128 Words in Wordbanks:

ach, ambassador, anchor, arrah, ass, bal, bannock, banshee, bard, basin, bawn, beak, beet, Beltane, betony, billet, bijou, bin, biretta, bodkin, bog, bothy, brasserie, brogue, brut, bulge, bullace, bushel, butcher, caber, cade, cader, cairn, cam, cantankerous, car, carpenter, carry, cashel, change, clan, claymore, cleave, coble, colleen, coracle, corgi, corm, corrie, crannog, cromlech, cross, currach, Dail Eireann, dolmen, dour, drape, druid, eisteddfod, embassy, encumber, fail, flannel, flummery, frown, gallon, galore, garter, gillie, glen, goblet, gouge, gravel, grouse, guillem, gull, hog, hubbub, ingle, iron, javelin, keen, kibe, lance, lawn, leprechaun, linn, loch, lochan, lozenge, marl, menhir, metheglin, mine, minion, mooch, mullein, mutton, och, peat, pendragon, piece, pillion, plaid, pollan, poteen, ptarmigan, puffin, rath, rich, ruche, say, shamrock, shanty, slogan, socket, tan, toman, tor, Tory, trews, truant, vassal, vendace, vug, warren, whisky, wrasse.

4. Results

4.1. 163 Words Identified By Native Speakers of English

Three hundred and thirteen words are divided into four groups according to provenance: Celtic words from Irish, words from Scottish Gaelic, Celtic words from French, and Celtic words of Brythonic origin. The subjects of the questionnaire were adults and mainly educated people who are English teachers at college with higher education. Nationalities of the subjects are Americans, and English or British. The subjects are 25 males and 25females (Ito, 2006).

Matching test type of questionnaire and filling test type of questionnaire were made for the four groups of the words. Matching test requires subjects to identify definitions with an appropriate word among words in a group. Filling test requires subjects to fill a blank with one word with clues of initial two letters for each definition.

Out of 312 words, 163 words were identified by native speakers' of English by matching test type questionnaire..and 123 words were identified by filling test type of questionnaire. The 163 words from matching test type of questionnaire are used in the present test, because the 123 words are included into the 163 words (Ito, 2006).

As mentioned above; subjects for questionnaire were 50; therefore, maximum of frequency of identification by the subjects is 50 and minimum is one. The researcher made a list of the 163 words with frequency by the subjects, and added frequencies of BNC and Wordbanks. The 163-word list is as follows:

airt, ambassador, andiron, arpent, arrah, bannock, banshee, bard, basin, beak, Beltane, betony, bijou, billet, bin, bog, bonaght, bonnyclabber, brail, brasserie, brogan, brogue, brut, buckeen, bulge, bullace, bushel, butcher, caber, cade, cailleach, cairn, cantankerous, car, carpenter, carrow, carry, change, clabber, clan, claymore, cockabondy, colleen, coracle, corgi, corm, coronach, corrie, crannog, crine, cromlech, crostarie, currach, cwm, dalt, deasil, dillue, dolmen, dour, drape, druid, duan, dudeen, dulse, duniwassal, eisteddfod, elvan, embassy, encumber, esker, fewterer, filibeg, flannel, flummery, frown, galliard, gallon, galore, garron, garter, gillie, glean, glen, goblet, gossan, gouge, grave, gravel, grig, grouse, growan, gull, ingle, javelin, keen, keen, kern, kibe, killas, kistvaen, kyle, lance, lawn, leprechaun, loch, lochan, loy, lozenge, marl, mavourneen, menhir, metheglin, minaudiere, mine, minion, mooch, morglay, mormaor, mullein, mutton, och, ohone, ouch, palfrey, pendragon, pennill, pibroch, piece, pikelet, pillion, plaid, poteen, ptarmigan, puffin, quaich, ruche, sennachie, shamrock, shanty, shebeen, slogan, socket, sowens, spalpeen, spleuchan, strath, tais(c)h, tanist, tocher, toman, tor, Tory, trews, truant, union pipes, usquebaugh, vassal, vendace, vug, warren, whisky, wirra, wrasse.

4.2. Frequencies by BNC and Those by Wordbanks

However, some of them are not found either in BNC or Wordbanks and some others are not found in both of the corpora. Out of the 163 words, 112 words are found in BNC, and 85 words are found in Wordbanks. Thus, 163 words were deduced to 112 words to compare with the frequencies in BNC.

The 112-word list with frequencies by the subjects and those in BNC and Wordbanks is as follows:

item	R	BNC	WB	item	R	BNC	WB	item	R	BNC	WB
ambassador	48	1346	1273	dolmen	33	26	6	mine	35	1150	644
andiron	29	2	0	dour	28	152	125	minion	11	79	30
arpent	4	2	0	drape	17	182	25	mooch	4	21	7
bannock	17	9	6	druid	48	91	74	morglay	1	1	0
banshee	36	47	18	dulse	5	1	0	mormaor	3	1	0
bard	18	143	241	eisteddfod	21	48	70	mullein	3	8	4
basin	33	1511	389	elvan	1	3	0	mutton	49	162	85
beak	50	388	136	embassy	43	1491	1557	och	13	208	23
Beltane	12	8	2	encumber	37	4	21	palfrey	18	8	0
betony	5	3	1	filibeg	2	1	0	pibroch	6	2	0
bijou	28	26	12	flannel	41	243	174	piece	45	14145	7467
billet	6	98	18	flummery	4	5	7	pillion	12	47	27
bin	28	1251	591	frown	46	2129	447	plaid	21	116	127
bog	45	522	151	gallon	44	852	529	poteen	9	6	9
brasserie	31	73	80	galore	23	125	122	ptarmigan	27	14	1
brogan	12	2	0	garron	1	13	0	puffin	32	80	21
brogue	2	94	38	garter	45	116	62	quaich	2	7	0
brut	2	1	0	gillie	10	13	2	sennachie	9	2	0
bulge	1	226	75	glean	30	190	25	shamrock	39	91	143
bushel	27	45	23	glen	24	367	46	shanty	21	118	110
butcher	48	476	275	goblet	47	181	55	shebeen	7	4	0
caber	20	8	8	gossan	1	2	0	slogan	7	801	539
cailleach	3	1	0	gouge	19	45	5	socket	21	929	270
cairn	32	198	31	gravel	47	1245	303	sowens	2	1	0
cantankerous	17	40	24	grouse	38	203	82	spalpeen	2	2	0
car	48	33936	23643	gull	22	513	61	strath	3	24	0
carpenter	47	436	210	ingle	8	1	2	toman	2	43	1
carry	40	30552	13119	javelin	45	78	67	tor	14	120	79
change	46	32558	14230	keen	16	4	6	Tory	42	1986	1597
clan	43	539	398	keen	11	2	0	trews	9	8	0
claymore	18	14	17	killas	1	8	0	truant	43	58(a)	59(a)
colleen	36	13	6	lance	39	222	236	usquebaugh	5	2	0
coracle	26	23	10	lawn	49	1356	1088	vassal	31	235	22
corgi	45	51	15	leprechaun	42	51	29	vug	2	13	0
corm	1	11	37	loch	38	1023	95	warren	44	227	56
corrie	7	45	0	lochan	25	84	2	whisky	48	1840	565
cromlech	4	5	7	lozenge	30	88	15				
cwm	10	6	0	menhir	3	7	4				

Table 1: Frequencies in BNC and Wordbanks

As mentioned above, numbers under the ravel of R comes from earlier research which was conducted on native speakers' of English. Subjects are 50 in number; therefore, maximum number is 50 and minimum is one (Ito, 2006).Numbers under the ravels of BNC and WB are frequencies from one million word text in each corpus (Shogakukan BNC, 2008; Shogakukan Wordbanks, 2008).

5. Discussion

Since the main purpose of the present study is to examine if the results obtained in earlier study (Ito, 2006) are in accord with the frequencies of the identified words in BNC and those in Wordbanks, statistical analysis is applied to the results.

Correlation between frequencies in BNC and those in Wordbanks is 0.97, which means that both corpora show the same tendency to exhibit popularity among native speakers' of English. However, correlation of the results of Receptive Knowledge and those taken from BNC exhibited 0.31, which does not show any correlation between the two croups. Then, rank order is applied to the list: the list is as follows:

item	R	BNC	item	R	BNC	item	R	BNC	item	R	BNC
ambassador	4	11	change	12	2	glean	42	34	palfrey	61	79
andiron	44	96	clan	21	19	glen	52	25	pibroch	84	96
arpent	89	96	claymore	61	71	goblet	11	36	piece	14	4
bannock	64	78	colleen	33	73	gossan	106	96	pillion	70	59
banshee	33	59	coracle	50	69	gouge	60	61	plaid	55	43
bard	61	39	corgi	14	56	gravel	9	13	poteen	77	87
basin	36	8	corm	106	77	grouse	30	32	ptarmigan	48	71
beak	1	24	corrie	81	61	gull	54	21	puffin	38	51
Beltane	70	79	cromlech	89	89	ingle	80	105	quaich	98	85
betony	86	94	cwm	75	87	javelin	14	53	sennachie	77	96
bijou	45	66	dolmen	36	66	keen	67	91	shamrock	29	47
billet	84	45	dour	45	38	keen	73	96	shanty	55	42
bin	45	12	drape	64	25	killas	106	79	shebeen	81	91
bog	14	20	druid	4	47	lance	28	30	slogan	81	18
brasserie	40	54	dulse	86	105	lawn	2	10	socket	55	16
brogan	70	96	eisteddfod	55	58	leprechaun	24	56	sowens	98	105
brogue	98	46	elvan	106	94	loch	30	15	spalpeen	98	96
brut	98	105	embassy	21	9	lochan	51	50	strath	93	68
bulge	106	29	encumber	32	91	lozenge	42	49	toman	98	64
bushel	48	61	filibeg	98	105	menhir	93	85	tor	68	41
butcher	4	22	flannel	26	26	mine	35	14	Tory	24	6
caber	59	79	flummery	89	89	minion	73	52	trews	77	79
cailleach	93	105	frown	12	5	mooch	89	70	truant	21	55
cairn	38	33	gallon	19	17	morglay	106	105	usquebaugh	86	96
cantankerous	64	65	galore	53	40	mormaor	93	105	vassal	40	27
car	4	1	garron	106	73	mullein	93	79	vug	98	73
carpenter	9	23	garter	14	43	mutton	2	37	warren	19	28
carry	27	3	gillie	75	73	och	69	31	whisky	4	7

Table 2: Word List in Rank Order

Correlation between receptive knowledge results and BNC results according to rank order analysis is 0.75, which means that the two types of results are highly correlated to each other.

6. Conclusion

The conclusion is stated below as a summary of the previous sections.

1. Degree of recognizing English words by questionnaire is similar to the degrees of frequencies by BNC and Wordbanks. This is obtained by rank order correlation not by comparing frequencies between the results of questionnaire and those of BNC and Wordbanks.

2. All of the words collected from dictionaries are not found in BNC and Wordbanks, which implies that dictionaries cite words which even large corpora do not include.

3. Even educated adult native speakers' of English do not know all of the words in English dictionaries.

4. Identifying some of the original 312 words was impossible in the corpora, because they have completely different meanings in one and the same spelling under the same part of speech. For example, *down* was impossible to identify to obtain frequency of the word of Celtic origin because one *down* means 'soft, fine, fluffy feathers,' and the other *down* means 'a gently rolling hill.' Thus, corpora are not always almighty for lexical studies.

References

Ito, M. (2005). *Celtic words in English and Spanish*. Toyohashi: San-ai Kikaku.

Ito, M. (2006). *Celtic words in modern English*. Toyohashi: San-ai Kikaku.

Ito, M. (2008). 'Celtic words in BNC and WordBanks Online'. Paper presented at Kwansei Gakuin University at Forum on Lexical Studies 2008 held by JACET Reading Study Group and English Lexical Study Group.

Shogakukan (2008). *Shogakukan BNC Corpus* [on line]. Tokyo. Shogakukan.

Shogakukan (2008). *Shogakukan Wordbanks Corpus* [on line]. Tokyo. Shogakukan.