THE 4-SET PROBLEM

PHILIP M. COHEN
Aliquippa, Pennsylvania

23751 different sets of four letters can be formed from the 26 letters of the alphabet, if order is ignored and repetition allowed. (This is the same as the number of ways of selecting four different objects from a group of 26+4-1.) Is it possible, for each set, to find a word containing its letters?

A few special cases have been investigated before. Dmitri·Borgmann in the February 1969 **Word Ways** considered the sets with all letters the same, from AAAA **anabata** to ZZZZ **pizzazz**. Subsequent research, up to February 1983, has left no holes but XXXX, for which we have nothing but complicated, un-English-looking chemical terms, 'xoxxoxo' from Finnegans Wake, and coinages like the adjective 'sixty-six-sixty-sixths'. In the February 1971 **Word Ways** Murray Pearce looked at the sets of four consecutive letters, from ABCD **cabda** to ZABC **zebraic**. He found dictionary examples for all but three, which we here fill in from other sources: UVWX **Vireux-Wallerand** (populated place, OSFR), VWXY **Xsmwdribvnwlxy** (play title, NUCL), and WXYZ **Xiawaziyu** (town, NIA). (Abbreviations for references are explained in the bibliography.) In the February 1972 **Word Ways** Ralph Beaman gave WXYZ **waxy maize**, from NI, but we restrict ourselves in this project to terms without internal spaces. Dmitri Borgmann asked in Beyond Language (Scribner's, 1965) and later in **Word Ways** about JQXZ, the set containing the four rarest letters in English, which was ultimately found by Darryl Francis in **Xiq-Xhafej** (populated place, OSAL) and given in the August 1970 **Word Ways**. Leslie Card and Ross Eckler in the August 1972 **Word Ways** considered the smaller problem of 3-sets, tackling only what they judged to be the 182 hardest of the 3276 sets, and finding all but 20, which were filled in in the May 1982 **Word Ways**. During 4-set work, the other 3094 were all found in NI words, the last to fall being GXX **sextuplexing** and WWZ **zwetschenwasser**.

In 1970 I decided to try the general 4-set case. It fit my criteria for a good logological project: it had a clearly delimited goal, but one too distant to be reached in one burst of action, so it could be pecked away at for years. It forced research beyond the bounds of NI, and could serve as a setting for many of the odd words I came across in my logological reading. And the chance of ultimate success was far greater than that of my earlier long-term project, finding words containing each of the 17576 trigrams from AAA to ZZZ. (On the other hand, it was far less than the chance for a related project: finding words containing all 17576

ordered 3-sets with separation allowed, so that **xen**on contains XNN but not NXN or NNX. This has never been tried but would be an interesting follow-up to the 4-set project.) 1 wrote out all 23571 sets in four notebooks and began filling slots from memory and with lists of 4- and 5-letter N1 words. Unfortunately, the drudgery outweighed the pleasure, and the prospect of checking longer words later was daunting: ambidextrously alone contains 1001 different sets. 1 soon put the notebooks aside.

In 1981 Alan Frank visited me and I mentioned the project. He had a database of all words from the <u>Official Scrabble Players Dictionary</u>; turning a computer loose on it, he got about 21500 sets in short order. Then, working himself on missing sets from the 7315 not containing the hard letters J,K,Q,V,W,X,Z, he soon had all but seven. 1 was able to find BFHH, FFHM, BFHP, FMPP, CFYY and FFYY – can you? (See Answers and Solutions). This left only YYYY unsolved in N1. (We had **fyyryryn** from O, **polysyndactyly** from D25, and others; he happened on **dacryocystosyringotomy** some months later.) With a computer to do the scutwork of extracting sets from words and updating the list, the project could now move into high gear.

Alan told me why the project appealed to him: "It's somewhat mindless at times when 1 want it to be, but 1 can think about it profitably if 1 want to do that. For example, thinking of possible places to find UUVV, and finally coming up with 'vulturewise'. It's a good computer project. If 1 find a word with BBHWXX, the computer will tell me if that improves any sets. Theoretically, all sets are findable with normal English words; it's only happenstance that ex-Jacquardizer is not in the dictionary, whereas the fundamental structure of English assures having to go to outrageous lengths to find the trigram QPX. 1 like looking through dictionaries, in moderation. 1 like working on projects with other people; usually better-executed than ones done alone, and more fun."

The programs, for simplicity, accepted only all-upper-case letter input. Once we went beyond Scrabble-legal N1 words, it became necessary to introduce tags to indicate orthographic matters like capitalization and hyphens. Since such factors made a word less desirable, we adopted the convention that any word was superior to any other word with a longer tag and should replace it in the database. Extending this idea, since we wanted to keep as close as possible to NI words, abbreviations for sources of non-NI words were incorporated into the tags. Sources having less satisfactory word-stocks would get longer abbreviations, so words from them would have longer tags and be apter to be replaced. Thus all general dictionaries had 1-letter codes, while non-reference works (for example, newspapers and textbooks) had codes of 4 or more letters. Later, secondary tags were introduced to handle diacritics and other information that did not affect the acceptability of a word. Some sample entries, with + separating the word and its tags: BFKQ BREAKFASTEQUIPAGE+O-+@EQ can be decoded as **breakfast-equipage**,

tag length 2, in the Oxford English Dictionary, found unaer "equi-page". CFXZ CAZAUXFRECHET+OSFR*-*+' is Cazaux-Frechét, with tag length 7, in the Official Standard Names Gazetteer for France, two capital letters, an acute accent that is not considered part of the E and does not lengthen the tag. JWWZ WOJC1ESZoW+N1A*+' is Wojcieszów, tag length 4, in the Rand-McNally New International Atlas, not usable as a JOOZ set because the accent on the second O makes it a different letter in Polish (as do umlauts in German or Hungarian); the program treats lower-case letters as non-letters, so it will not try to put this word in the JOOZ slot. We defined the tag of a 4-set as the tag of the best word known to contain that set.

For the most part, we did not try to distinguish levels of ac-ceptibility within sources. Thus rare and obsolete N12 words can supersede everyday, but longer, words, and if a name is trans-cribed three different ways in M79, all three have equal status. However, we added a few tag suffixes in special cases. OBS, for example, is used for words showing obsolete spelling conventions (mainly in O), DEF for words found only in definitions of other words, and 2 for words from 2-word phrases. Examples are MQVW **vmqwhyle**+OOBS+@umquhile, FKPV **fun-provoking**+DEF+@crazy house (in NI3), and CJMX **Jarisch-Herxheimer**+2.

Such a crude and one-dimensional measure of acceptability is of course open to objections. Why, for instance, should N12 and NI3 words be superior to those of other dictionaries, unabridged (like C,F and O) or not? Basically, for expedience; to avoid clut-tering the database and its printout with tags on every word, we say that any untagged word is to be found in one of a specified set of references. N12 and NI3 are a natural set. They are widely available, have a huge word-stock, are few in number (so that untagged words are little trouble to locate), and avoid the weird obsolete spellings of O. Giving preference to N1 reduces the num-ber of tags needed. As another example, is UWWW **chuck-will's-wid-ow**, with its +-'- tag, really on a par with a medical-dictionary term like BBXX hexabromdioxyphenylcarbinmol+D23? Probably not, but there would likely be differences of opinion among logologists as to which is worse.

Crude or not, the system generally works satisfactorily, and anomalies can be handled individually. For example, of the 1997 sets obtainable from **pneumonoultramicroscopicsilicovolcanokoniosis**, shorter untagged words were known for all but KMPV. We nonethe-less used Pervomaisk for KMPV; though capitalized, it was much shorter and less artificial. Alan later eliminated the problem by finding a word better than either; can you think of uncapitalized 10- and 11-letter KMPV words from NI2? (If stumped, see Answers and Solutions.) And for VVVV, we prefer nivvi-nivvi-nak-nak+WD to the shorter **Vivstavarv**+OSSW, though both have tag length 5, because the former is more English and three hyphens aren't that much worse than one.

For really far-out words, intuition breaks down completely. How

should one rank strange coinages from Rabelais, Joyce, and the Guinness Book of World Records (or, worse yet, logologists); esoteric chemical terms full of digits and parentheses; phantom words that slipped into dictionaries as errors or hoaxes; or pre-English Anglo-Saxon words? For the most part, we just assign them whatever very long tag seems fit at the moment. We regard all words with tags above length 10 or so as terrible, and don't worry much about what supersedes what.

Even for less esoteric words, deciding the level of acceptability of words is a problem in any wide-ranging logological project. The problem was greater here because we had to assign exact numerical values. For example, -itides is explicitly given as a plural of the suffix -itis only in N13 and perijejunitis is only in N12, so is perijejunitides acceptable? We decided it was, without even a tag. N12 gives no plural for jack-in-a-box; does that mean the plural is the regular, but odd-looking jack-in-a-boxes, or did they just lack information? O explicitly shows jacks-in-boxes, so we let that go as untagged N12 (not O), with some misgivings on my part. The notes under **number** in N13 indicate that fractions can be written as single hyphenated words if used as adjectives, but does this apply if numerator or denominator is itself hyphenated? I suspect not, and have argued for a long disapproving tag (FRACTION) for semi-sanctioned coinages like VWWX **twenty-six-twenty-seventh**. I've argued that we can't make inferences from one compound to another. **Fifteen-hundred-word** in A Manual of Style does not justify WXXX **six-hundred-sixty-six-word**. **Six-quarter-cattle** in O does not show that CQWX **six-quarter-cow** exists. Alan proposes that anything with a (subjectively estimated) probability of 5 per cent of ever having occurred in print is includable, with an appropriately long tag. l disagree.

On the other hand, I pushed to have the 585-letter surname **Wolfeschlegl...** from Guinness included, though it's clearly a logologically-motivated coinage. It's in a common modern reference, and the name of a real person is more legitimate, less easily foisted upon the world, than any purely literary coinage. And I'm not strongly against pluralized proper names. Even if there is only one place in China named **Jiangjunqiao**, JJQS **Jiangjunqiaos+NIAPL** could mean "places like Jiangjunqiao." A tag suffix like PL would be in order, though.

Like our tagging principles, our word-hunting strategies differ. Alan concentrates on coining plausible words containing needed sets and trying to find them, in one case spending several hours in the library on an unsuccessful search for **extremely-low-frequency**. I'm more apt to plow word by word through a reference or, if pickings are slim, through the most promising parts, like ex-, j-, q-, sq- and x- words. This is far slower, but surer. For example, we both did a lot of coining in trying to find an NI word for GKTX: **ex-knight, ticket-fixing, textbooking, knitting-box, tackle-box**, etc. We found **extinguisherlike+O, mixing-stack+F, and striking-box+O**, but nothing in Nl until an exhaustive search of N13 turned up **alkoxylating**. The search for GMXX **maxixeing** was

likewise fruitless; unlike so many other dances, this was never a verb in any reference checked. Now VWWW is proving intractable. **Willow-woven, widow-wives, window-swivel** and other compounds are not to be found. **View window** is a two-word term in N13, so perhaps it can be found hyphenated somewhere, but ultimate success seems as likely to come from some mindless search.

As the project advanced, the data-processing and output-generating programs have grown more numerous and complex, deserving an article of their own. Alan will use them as examples in an article on computer logology.

With this background, we can now turn to some results. As mentioned above, the 7315 non-JKQVWXZ sets were completely solved early on. This simplified searches, since words not containing one of the seven "hard letters" could be ignored (except when trying to shorten an already-found set). Then, one by one, other letters were removed from the hard-letter list, adding to what could be passed over. First were sets with K, an additional 1570 for a total of 8885. For four of them, only tagged words are known: BFKK **knife-backed** (**flanker back** is two words in N13 addenda), FKKK **Kirkfieldbank**+TIG, GKKK **kirk-skailing**+O, and KKKP **Pakokku**. Next was W, 16 tagged in 10626 sets; the 12 new tagged sets are BBWW **wibbly-wobbly,** BCWW **twice-blowing,** BPWW **whippoorwill's-boots,** BWWY **yellow-brown,** CWWW **twice-widowed,** FWWW **widow-sawfly**+O, KWWW **swallow-woodpecker**+O, UUUW **Shukulumbwe,** UUWW **wau-wau,** UWWW **chuck-will's-widow,** WWWW **wow-wow,** and WYYY **y-wryly.** Z was more intractable; adding Z sets gives us a total of 12560, more than half of all sets, but 85 of them are tagged. This is so many that I will only list the 14 of the 69 added ones with tags longer than 1: BBFZ **Fitzgibbon**+F, BFKZ **furze-break**+O, BFWZ **zebra-wolf**+O, BWWZ **zebra-swallowtail**+F, FFFZ **Fitzjeffrey**+NUCL, FFWZ **zwolffingerdarm**+D24, FFZZ **frizzle-frize**+O, FMMZ **Fitzsimmons**+A, FMWZ **Fitzwilliam**+F, FPWZ **zephyr-flower**+F, FWWZ **Safarewiczowa**+M80, PWWZ and WWWZ **powwow-wizard**+O, and UWWZ **Wurgwitz**+NIA. One could remember the needed K and W sets, but the Z sets are too numerous, and it remains necessary to keep a list for checking when a likely word with Z is found. This is even more true of the V sets; adding them brings us to 14950 solved sets (62.94% of the total), including 20 new ones with tags longer than 2: BBVZ **Bezzubov**+M79, BVVV **Novobarvinovka**+OSNU, BVVW **Traveling-wave-tube**+EDS, BVWZ **Bovenwezet**+OSBE, FFKV **Fefelovka**+OSNU, FKVV **Varfolomejevka**+NIA, FVVZ **Voelfling-les-Bouzonville**+OSFR, FVWZ **Menzenschwand-Vorderdorf**+OSWG, HVVV **Verchovcevo**+NIA, HVWZ **Hazelnut-weevil**+C, KVVV and VVVZ **lvanovo-Voznesensk,** KVWZ and VWYZ **skywave-synchronization**+EDS2, MVVZ **Adzvavom**+NIA', PVWZ **Onze-Lieve-Vrouwpolder**+OSBE, VVVV **nivvi-nivvi-nak-nak**+WD, VVWZ **Wez-Velvain** +OSBE, VWWW **twenty-two-twenty-sevenths**+FRACTlON, and VWZZ **Zwevezele**+NIA. Some might say that VWWW is still a gap. Two more letters are close to being eliminated: only FKXX, VWXZ, VXXZ and WXXZ remain for X, and FJJW, FJJY, FJJZ, JJJW and JJVW for J. Q, however, is still far off; even only considering non-JX sets with only one Q, there are still twelve missing: BFQW, BQVW, FFQZ,

FQVV, FQVZ, FQWZ, KKQZ, KQWZ, PQWZ, QVVW, QVVZ and QVWW. Can readers help us to improve the tagged sets above? It's possible; UUUW, for example, has been improved twice lately, from **Wulanwusu**+N1A and **wound-fungus**+O.

It seemed to me that the order of elimination might serve as a measure of the "friendliness" of letters, their ability to occur with others in a word. Thus the most unfriendly letters, worst listed last, would be KWZVXJQ, and the hardest repeat-free set to find would not be JQXZ (rarest letters) but JQVX; indeed, this remains a hole. But this is no help in estimating the friendliness of the other 19 letters, so Alan wrote a program that would give a numerical measure of friendliness by using the want list. Whenever a letter occurs in a tagged set, the tag length is added to the score, completely missing sets being assigned an arbitrary tag length of 25. Thus the natural DFKX **flex-cracked** adds only 1 to the scores of D,F,K and X. We must go much farther afield for VXYZ **Croix-lez-Rouver**oy+OSBE; since these four are more reluctant to come together, their scores are increased by 8. Note that VVXZ **Vieux-Villez**+OSFR adds 14 (double the tag length) to V. The results agree well with the earlier measure: Extroverts (83-155) EA1SON, Friendly (217-542) LUTRDCG, Neutral (654-1000) HMYPB, Stand-offish (1696-1811) FK, and Hermit (2379-6414) ZVWJXQ. The only large inconsistency is the unfriendliness of W. The reason is that W combines well with the friendly letters, and so is easy to eliminate early, but goes very badly with unfriendly ones, notably Q. A list of 127 tagged or missing sets containing one Q and no other rare letter (JQXZ), shows 67 containing W. (By the way, we do have JQWX **Xiwujieqiao**+OSNCP.) At the other end of the scale, two letters' ranking do not match their commonness. It's understandable with H, which is common in English primarily due to its appearance in a small number of words, but T is surprising. Labials seem to be generally unfriendly, as the placement of WVFBP shows. The unfriendliest 3-set is not surprisingly JQX; the twenty-six foursets containing it have tag-lengths summing to 357, compared to 277 for the runner-up, JXX. As an indicator of how far the project has progressed, among the 126 sets containing only the six hermit letters, we have something for 97.

At present we have untagged words for 21980 of the 23751 sets, or 92.55%. Adding hyphenated and apostrophized words raises this to 22335, or 94.04%, and capitalized N1 words (including names from the N12 Gazetteer and Biography sections) bring us to 22602, or 95.13%. Unpunctuated uncapitalized words from all general reference works come to 22129 or 93.17%. References other than N1 that contributed the most are: O (195), N1A (167), C (68), S (57), R (41), M70 (39), OSAL (32), HNA1 (32), OSFR (27), and R1NG (25). Other sources include 10 other dictionaries (74 entries), 11 other OSN Gazetteers (76), 7 other Mnn's (19), 4 other atlases and maps (15), 21 other reference works (104), 5 other works of all sorts (22), literary inventions by Calvino, Joyce, Rabelais and others (6 works, 32), and 10 semi-justified coinages and unverified citations.

From another viewpoint, the 1695 tagged sets consist of 676 with tag length one, 297 at two, 102 at three, 284 at four, 138 at five, 32 at six, 90 at seven to nine, and 76 terrible ones at ten and up, plus 76 missing outright.

In conclusion, we would like to make an appeal for help to **Word Ways** readers. We list below approximately the worst one per cent of the sets, those most in need of improvement. First is a list of 76 sets for which we have nothing at all; then follows a list of 166 sets with tags of length 7 or more, each followed by its tag length. In addition, capitalization shows how badly a set is needed. Tag length 7 sets are printed as xxxx, 8 as xxxX, 9 as xxXX, 10-14 as xXXX, and 15 or more as XXXX.

NI3 has been searched completely by a human, and computer searches have covered most above-the-line terms in NI2. Obvious compounds have been checked in all the major unabridged dictionaries, and letters J,Q,X and Z have been checked in toto in O. Promising territory includes chemical and other technical terms, gazetteers for a few countries not in the Official Standard Names series, like Poland and the Union of South Africa, and articles and books likely to include useful compound words, such as a book on jukeboxes for JX-- words.

When the stream of improvements has largely dried up, we will make copies of the latest printout (60 to 70 pages) with detailed explanations and bibliography, perhaps as a **Word Ways** monograph.

BFJQ	DKQX	FJMQ	FQVZ	JKQW	JQQZ	JVWX	KQXX	QQVW	QWXX
BFQW	FFQZ	FJPQ	FQWZ	JKQX	JQRX	JVXX	KQXZ	QQWZ	VWXZ
BJQW	FGQQ	FJQV	JJJW	JMQW	JQTW	JXXY	PQQV	QVVW	VXXZ
BJQX	FJJQ	FJXX	JJQV	JMQX	JQVX	KKQX	PQWZ	QVVZ	WXXZ
BKQQ	FJJW	FKXX	JJQW	JPQX	JQWW	KKQZ	QQQV	QVWW	
BPQQ	FJJX	FPQQ	JJQX	JPXX	JQWZ	KQQV	QQQZ	QVWX	
BQVW	FJJY	FQQQ	JJVW	JQQV	JQXX	KQQX	QQRW	QVXX	
CJXX	FJJZ	FQVV	JJWX	JQQW	JQXY	KQWZ	QQTX	QWWX	

bBJQ 14	cFJX 12	ffQW 9	fvwz 7	hQQQ 10	jkqz 7	jrwX 8
bBQX 13	cjjQ 8	fFXZ 10	fVXX 10	hqvx 7	jlxx 7	jrxx 7
bbvX 8	cjqX 8	fhjq 7	fXXZ 10	hqVZ 9	jmqQ 8	jtxx 7
bgjq 7	cJWX 12	fJPZ 12	gKVX 13	hqwX 8	jMQV 10	jUWX 12
bjpq 7	cJXZ 10	fjqx 7	gkxx 7	hWXX 10	jMWX 12	JVWW 15
bJPX 12	cKXX 13	fjqz 7	gpqq 7	jJJQ 14	jppq 7	jVWZ 13
bjqq 7	cqwx 7	FJWW 15	gqqQ 8	jJMQ 10	jpqV 8	jWWX 10
bpqv 7	dhjq 7	fJWX 12	gqqv 7	jjpq 7	jpwX 8	jWXX 10
bQQQ 14	djqQ 8	fkxZ 8	gqvV 8	jJQT 10	jqqQ 8	jWXY 12
bQQX 10	dJQX 10	fqqv 7	gqvx 7	jJQY 14	jQSX 10	JXXX 16
bqvx 7	djqz 7	fqQW 9	gQXX 11	jjqz 7	jQTX 14	jXZZ 14
bqvz 7	dJXX 12	fQQX 12	hjjq 7	jjwY 8	jqvz 7	kkwX 8
bqww 7	dQQX 10	fqwx 7	hjqV 8	jJXX 12	jqxz 7	kKXZ 12
bGQQ 11	dqVX 9	FQXX 17	hjqx 7	jqpQ 8	jQYY 14	knxx 7
bjkq 7	ffqq 7	fvvZ 8	hJWX 12	jkqV 8	jqyZ 8	kPXX 13

kqQQ 9	kXXY 13	ppqQ 8	pwXZ 9	qrwx 7	tvxx 7	vWWW 11
kqVV 9	kxXZ 9	pQQY 10	pxxX 8	qrxx 7	txxx 7	vXXX 11
kqvz 7	kxZZ 9	pqvw 7	qQQY 14	qtwx 7	uvwx 7	wWXZ 10
krxx 7	lqwx 7	pqwW 8	qQRX 10	QVXZ 16	uwXX 9	wXXX 11
kTXX 13	mQQQ 14	pQWX 12	QQVX 16	qvyz 7	uXXX 11	wXZZ 14
kVWX 13	mqwX 8	pqXX 9	QQVZ 16	QXXX 20	vvvx 7	xXXX 11
kVXX 13	mxxx 7	pvWZ 9	QQWX 15	qXXY 11	vVWX 11	xzZZ 9
kwxx 7	nqqQ 8	pvxX 8	qQXY 10	qXXZ 10	vvwz 7	
kWXZ 10	oWXX 10	pVXZ 10	qQXZ 10	qXZZ 13	vvxz 7	

BIBLIOGRAPHY

--- Webster's New International Dictionary of the English Language, Second Edition Unabridged, 1934-58

--- Webster's Third New International Dictionary of the English Language, Unabridged, 1961

A   American Heritage Dictionary, Second Edition, 1982

C   Century Dictionary and Cyclopedia, including Supplements, Cyclopedia of Names, Atlas, 1911

Dxx   Dorland's Medical Dictionary, Xxth Edition

EDS   Modern Dictionary of Electronics, by Rudolf F. Graf, 1962

F   Funk and Wagnalls New Standard Dictionary of the English Language, 1935-53

HNA1   Handbook of American Indians North of Mexico, by Frederick Webb Hodge, Smithsonian Institution of American Ethnology Bulletin 30, Washington, US Government Printing Office, 1907

Mxx   Modern Language Association International Bibliography for the Year 19xx

NIA   Rand McNally New International Atlas, date not given

NUCL   New Union Catalogue of Pre-1956 Imprints, 1968-81

O   Oxford English Dictionary, 1971

OSAL   Official Standard Names Gazetter for Albania, US Board on Geographic Names, 2nd Edition, 1961

OSBE   Official Standard Names Gazetteer for Belgium, 73, March 1963

OSFR   Official Standard Names Gazetteer for France, 83. 1964

OSNU   Official Standard Names Gazetteer for the USSR, 42, 1970

OSWG   Official Standard Names Gazetteer for the German Federal Republic, 47, May 1960

R   The Random House Dictionary of the English Language, 1966

R1NG   The Ring Index, American Chemical Society, by A. M. Patterson et al., 2nd Edition, 1960

S   Supplement to the Oxford English Dictionary, 1972-76-82

TIG   Times Index-Gazetteer of the World, 1965

WD   Dr. Joseph Wright's Dialect Dictionary, 1898-1905