RESTRICTED LETTER SETS

A. ROSS ECKLER
Morristown, New Jersey

In order to understand the limits of recreational linguistics, and in order to identify unsuspected possibilities for research, it is often desirable to map parts of the territory. This was the purpose of "The Word-Surgeon's Compendium" in the August 1976 issue and "Words Within Words" in the May 1978 one, and this is the purpose of the present article. In contrast, the ten-year topical index in February 1978 was a global map of wordplay, containing no descriptive detail to justify the taxonomy used.

Consider an n-letter word in which one letter must be chosen from the set \( \{a_1, a_2, \ldots, a_i\} \), the second from the set \( \{b_1, b_2, \ldots, b_j\} \), and so on, where the \( a_i \)'s and \( b_j \)'s denote letters of the alphabet. This leads to a number of interesting word problems, some of which have been touched on before in Word Ways. It is useful to talk in terms of letters inscribed on the faces of \( n \) dice, as long as one remembers that this is a convenient shorthand for the general problem (\( i \) and \( j \) can be set equal to any integers, not just six).

If one rolls the dice one at a time to form the first, second, etc., letters of a word (picking the die to be rolled each time at random), it is rather unlikely that a word will be produced; \( n \)-letter words are much less common than gibberish like ZTWELC. The dice roll can be relaxed in two ways to make the likelihood of word formation greater:

1. after the dice have been rolled, one is permitted to rearrange the order of letters to form a word
2. after the dice have been rolled, the order is preserved but one is permitted to turn one or more of them over to reveal other faces to form a word

The first generalization was discussed by Dave Silverman in the May 1973 Kickshaws; he showed that it was easy to label three three-sided dice with nine different letters so that a word could always be formed, and conjectured that it was impossible to label four six-sided dice with 24 different letters so that no word from the Pocket Webster could be formed (no matter what sides turn up). Thus, it can be seen that there are two broad classes of interesting problems: assigning letters to maximize or to minimize the probability of word formation on \( m \)-\( n \)-sided dice. What are the largest values of \( (m, n) \) for which it is always possible to form a word? The smallest values of \( (m, n) \) for which it is impossible to form at least one word? Generalizations involving different numbers of faces on different dice, or allowing repeated letters length in a form...
The second generalization leads to harder problems. For a given dictionary, what is the minimum total number of letters needed on all the dice to guarantee that an n-letter word can always be formed? It is assumed that a letter can appear on more than one die, but that the letters on any one die are all different (to make the problem harder, allow each letter to appear on only one die). To avoid trivial solutions to the problem, one must also insist that the n! different orders in which the dice can be selected must lead to n! different words. It can be seen that this problem is a generalization of multiple transpositions; if, for example, there are six transpositions possible for a three-letter word (as is the case for AER using Webster's Second), then the answer to the problem is three. For words only in Webster's Pocket Dictionary, however, five letters are needed: (NE)(A)(TR) yields the six words EAR, ERA, ARE, ANT, TAN and TEA. For four-letter words or more, many extra letters are needed. For Pocket Dictionary four-letter words, the minimum number is at most 11; (EOR)(FAE)(TLV)(SL) gives the 24 words

<table>
<thead>
<tr>
<th>eats</th>
<th>rial</th>
<th>ales</th>
<th>tops</th>
<th>tire</th>
<th>spot</th>
</tr>
</thead>
<tbody>
<tr>
<td>east</td>
<td>rite</td>
<td>also</td>
<td>lose</td>
<td>liar</td>
<td>sale</td>
</tr>
<tr>
<td>eyes</td>
<td>pets</td>
<td>airy</td>
<td>tars</td>
<td>seat</td>
<td>stop</td>
</tr>
<tr>
<td>else</td>
<td>pest</td>
<td>pile</td>
<td>lair</td>
<td>iota</td>
<td>star</td>
</tr>
</tbody>
</table>

Can this number be reduced? Or equalled using no repeated letters? To efficiently explore the problem for words of five or more letters, one may need the aid of a digital computer in which multiple transpositions of many five-letter words have been stored.

A puzzle related to both generalizations appeared in Martin Gardner's "Mathematical Games" column in the December 1977/January 1978 issues of Scientific American: how can one put 18 letters on three dice so that the three-letter abbreviations of the 12 months can always be formed? At first blush, this appears to be impossible, since there are 19 different alphabetic letters present in the months, but by using lower-case letters and noting that d inverts to p and n inverts to u it is possible to solve the problem. This is a special case of a general problem: how few dice of what sizes are needed to spell out all the words of a specified vocabulary? Note that one is allowed both to rearrange and to turn over the dice to find all the words in the class.

Let us now turn from the general problem of restricted letter choice to two special cases which have previously appeared in Word Ways. In these special cases, the labeling of the dice is restricted in some way.

All Dice Labeled The Same

If certain letters are omitted, labeling all dice the same leads to the well-known lipogram problem. Little research work has been
done; most people are content to construct literary lipograms, usu­ally omitting the letter E (see the discussion of Gadsby in the August 1977 Word Ways). Yet even here analysis is possible. For example, if one were to construct a lipogram omitting the rarest letters of the alphabet, how many letters (and which ones) collectively have the same effect on writing as a missing E? The answer appears to be 11, as discussed in "On Writing Lipograms" in the February 1978 issue. The emphasis is on maximizing the number of short words available to the lipogrammatist, where words are either weighted equally or proportional to their occurrence in English-language text. How should the alphabet be split into two parts to make lipogram-writing with both parts as easy as possible?

There are actually two approaches to the lipogram problem—find circumlocutions to express a thought using a restricted vocabulary, or use creative spelling for words excluded by the lipogram rules (such as WUR for WERE if E is banned). The alphabet-split problem has already been explored using the latter approach, in "An Exchange of Complements" in the February 1978 Word Ways. As a variant on this theme, how well can one recognize words from which all the vowels have been removed (see the February 1978 Kickshaws)?

Note that the question of the order in which the dice should be rolled does not enter in this special case; in fact, it is not necessary to construct dice at all to investigate lipogram problems.

All Letters of the Alphabet Allocated to Dice

Suppose that one assigns the alphabet, each letter used once, to a set of n dice, and suppose that many duplicates of each die are available. If one selects a subset of dice in a specified order, can a word be formed by turning over the faces of the various dice? Can more than one word be formed using the same subset of dice in the same order? (Note that the alphabet need not be equipartitioned among the dice; in fact, it is impossible to so except for 2-sided or 13-sided dice.)

This leads at once to "A Readable Polyphonic Cipher" which appeared in the February 1975 Word Ways. In a polyphonic cipher, each symbol stands for one or more letters, leading to potential ambiguity in decoding a word. The simplest and most well-known example is the telephone dial, in which the cipher 227 leads to either BAR or CAP. Given that one wishes to compress the alphabet into m different symbols, how small can one make m and still have a chance of overcoming the ambiguity? What is the best allocation of letters to accomplish this? The November 1975 article showed how a nine-symbol allocation based on common bigrams can be easily read; the follow-up article "Another Polyphonic Cipher" in the May 1978 issue showed how to compress the symbols to six but required the aid of a computer to identify the commonest word (in English-language text) corresponding to each sequence of symbols. It is believed that this cipher can immediately identify better than 90 per cent of the words in a typical sentence. Longer words are almost always unique, as Dave Silverman found when he used a two literal line cipher.

These problems in communication between English language words and symbols have been previously explored using the alphabet-split approach in "An Exchange of Complements" in the February 1978 Word Ways. As a variant on this theme, how well can one recognize words from which all the vowels have been removed (see the February 1978 Kickshaws)?

Note that the question of the order in which the dice should be rolled does not enter in this special case; in fact, it is not necessary to construct dice at all to investigate lipogram problems.
found when searching for seven-digit telephone numbers which have two literal interpretations (such as PYGMIES and SWINGER).

These polyphonic ciphers are not proposed as codes for secret communication, but are useful in understanding the redundancy of the English language.

Polyphonic ciphers are somewhat related to the problem of smudgy letter identification in a bad Xerox copy of a typewritten document. At the 1977 Buffalo convention of the National Puzzlers' League, Will Shortz introduced a word game based on a polyphonic cipher in which the letters of the alphabet were divided into four groups: vowels (aeiou), consonants with ascenders (bdhkl), consonants with descenders (gpqw), and consonants with neither (cmnrs). The audience was challenged to decipher a series of words encoded in this fashion, which (except for the vowel-consonant distinction) can be modeled by a series of squares on the line, rectangles extending above the line, and rectangles extending below the line -- a remarkable mimicry of the smudgy Xerox copy!