THE MATHEMATICS OF WORDS

A. ROSS ECKLER
Morristown, New Jersey

From the standpoint of an algebraist, words are nothing but ordered sequences of letters, each letter being a member of an allowable alphabet of such letters. Any sequence of letters is permitted, and he is often interested in deriving theorems about the existence or non-existence of certain letter-patterns.

A survey of the mathematics of letter-sequences of this type is contained in a recently-published book, Combinatorics on Words (Addison-Wesley, 1983), Volume 17 of the Encyclopedia of Mathematics and its Applications. Although the book is touted as an elementary text, its results are understandable only to those having at least an undergraduate major in mathematics. The first chapter, for example, immediately defines such concepts as morphisms, free monoids, conjugacy, equidivisibility and left factors. The book is, in fact, a series of definitions and theorems presented at a highly abstract level with no hint of real-world applications. To give the **Word Ways** reader a slight appreciation of its contents, I describe below a couple of easily-grasped concepts, the **square-free** word and **cadences** of words.

A **square-free** (or **nonrepetitive**) word is defined as a word which does not contain any repeated sequences of the form **aa, abab, abcabc,** etc.; in logological terms, the word contains no internal palindromes. What is the shortest word-length such that all words of that length are not square-free? For an alphabet of size 2, the answer is 4; all words of length 4 or more must contain **aa, bb, abab** or **baba**. However, for an alphabet of size 3, there exist arbitrarily long words which do not contain such patterns as **aa, caca,** or **bacbac**. A sequence of ever-longer square-free words can be constructed by iteratively replacing **a** with **ab** and **b** with **ba**:

a, ab, abba, abbabaab, abbabaabbaababba, ...

and then replacing the sequences **a, ab, abb** with **c, b** and **a,** respectively:

c, b, ac, abcb, abcacbac, abcacbabcbacabcb, ...

It can be mathematically proved that no matter how long this word is made, it will never contain any repeated patterns.

Mathematicians have generalized the concept of a square-free word to an **Abelian square-free** (or **strongly nonrepetitive**) word. This is defined as a word which does not contain any repeated sequences after allowing for permutations; for example, **bacadcabda** can be rearranged to form **bacadbacad** by rearranging its last five letters. Again, what is the shortest word-length such that all words of that length are not Abelian square-free? For an alphabet of size 3, the answer is 8; there are no Abelian square-free words longer than 7 letters, and the only seven-letter Abelian square-free words are **abcbabc, abacaba** and **abacbab**. For an alphabet of size 4, the answer is not known, but it is conjectured that arbitrarily long Abelian square-free words can be constructed; mathematicians have constructed ones up to 1600 letters long with the aid of a computer. For an alphabet of size 5, it has been proved that there is no limit to the size of Abelian square-free words; one construction is given in P. A. B. Pleasants, "Non-Repetitive Sequences", Proceedings of the Cambridge Philosophical Society 68 (1970), pp. 267-74.

A word contains a **cadence** (more precisely, an **arithmetic cadence** of specified length) if it has a sequence of **n** identical letters that are spaced at equal intervals within the word. For example, the underlined letters in **baaabc, cabacaa** and **abbacaabc** are all cadences of length 3, and there are no other ones of this length present. For alphabets of various sizes, what is the shortest possible word-length such that all words of that length have cadences of length **n**? The answer is trivial for cadences of length 1 or 2 for all alphabet sizes, and for any cadence length for a one-letter alphabet. However, the answer to this question is known for only four non-trivial cases: for alphabets of size 2, cadences of length 3, 4 and 5 must appear in all words of lengths 9, 35, and 178, respectively, and for alphabets of size 3, cadences of length 3 must appear in all words of length 27. Rephrasing the first result, the longest words containing only two different letters and no cadences of length 3 are only 8 letters long: the only examples of such words are **aabbaabb, abbaabba** and **ababbaba**.

It is of interest to logologists to apply these concepts to English words. Are there any seven-letter Abelian square-free words in English? The only known ones are **Sereser** (or **Sarasar**) from the Douay Bible, **patapat** (defined as a repeated patting), **cachaça** (white rum), and **tathata** (suchness, in Webster's Third). The longest dictionary words that are not square-free are **electroencephalographically, ethylenediaminetetraacetate** and **honorificabilitudinatatibus**. On the other hand, **Taeniodontidae** is a word containing a seven-letter pattern repeated in a different order; can one find an English word containing eight or more consecutive letters followed by a permuted version of these same letters?

Although **pneumoultramicroscopicsilicovolcanokoniosis** is an Abelian square-free word, it contains one cadence of length 4 (underlined) and four of length 3 (on O, on I, twice on C). Does any other English word have this many cadences of length 3 or more? **Zenzizenzizenzic**, from the OED, does, but this is a rather special case. **Supercalifragilisticexpilidocious** also has a cadence of length 4, and barely misses one of length 5. The longest dictionary word with no cadences of length 2 or more is, obviously, the isogram **dermatoglyphics** or **uncopyrightable**; what is the longest dictionary word with no cadences of length 3 or more? It appears to be the 31-letter chemical term **dichlorodiphenyltrichloroethane** from the Random House Unabridged dictionary.

How long a cadence can be found in an English word? Cadences with spacings of two were investigated by Ralph Beaman in the May 1971 **Word Ways**; he called them alternating monotonies. The longest alternating monotony is the word **humuhumunukunukuapuaa,** in Webster's Third; it is very unlikely that this can be equalled in a cadence with a spacing of three or more.

In pair isograms (words containing exactly two of each letter) all letters participate in cadences of length 2. Is it possible to find pair isograms containing one cadence of each spacing? Eight-letter pair isograms with the following patterns have cadences with spacings one, two, three and four:

aabcdbdc	abcacbdd	abacbddc
aabcdcbd	abcbacdd	abbcadcd

However, the only English words known to match one of these patterns are **appeases** and **appearer**. Ten-letter pair isograms with the following patterns have cadences with spacings of one, two, three, four and five:

aabcdcebde	abcbdacdee	abbcdacede	abcacdbeed
aabcdbeced	abcadbdcee	abacdbceed	abbcadedce

However, none of these patterns corresponds to a Websterian word. Can anyone find a non-Websterian example, such as a placename?