ACCIDENTAL OUROBOROS WORD WORMS IN RUNNING TEXT

Mike Keith
Furlong, Pennsylvania

The *word worm* was introduced to readers of Word Ways by Ross Eckler in an article of the same name in the August 1993 issue (p. 131). Word worms are based on the three-dimensional lattice of integer-valued points. Each point in this lattice has 26 neighbors which can be reached by a single orthogonal or diagonal step. These 26 stepping directions are identified with the 26 letters of the alphabet by the assignment shown in this 3x3x3 cube:

A	B	C
D	E	F
G	H	I

Top layer

J	K	L
M		N
O	P	Q

Middle layer

R	S	T
U	V	W
X	Y	Z

Bottom layer

Starting at some point in space, such as $(x,y,z) = (0,0,0)$, a word (or, more generally, any contiguous block of letters in some text) traces out a "worm" through 3-D space by using each successive letter to step in the direction indicated by the diagram above. A word worm usually ends at a different point from where it started, in which case we call it an *open* worm. If it ends at its starting point it is called *closed* (also known as an *Ouroboros worm*).

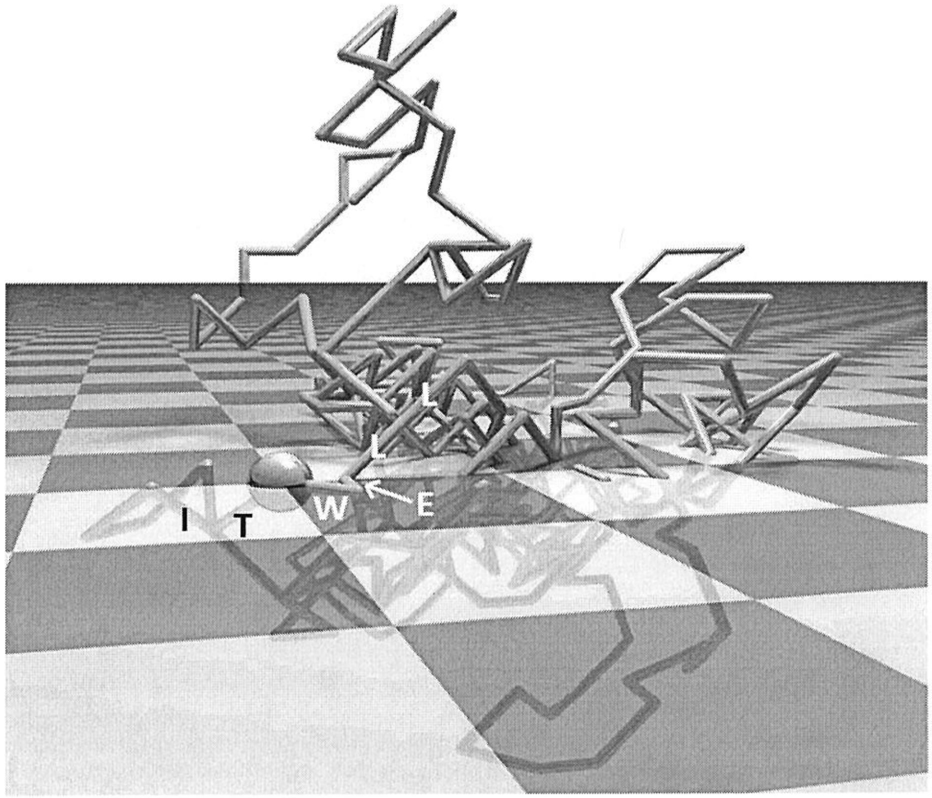
This particular assignment of letters to the 26 positions in the cube is quite natural, but it does have one drawback: it is directionally biased. If we add up each letter of the alphabet's frequency of use in English multiplied by its word-worm displacement in (x,y,z) the result is about (0.01, 0.11, 0.16), which represents the mean expected displacement per letter. So, for instance, the worm for a 100-letter sample of English will, on average, end up about (1, 11, 16) units away from its starting point, implying that accidental Ouroboros worms in running text will tend to be fairly short. But how long can they be?

To get an idea of the possibilities, we scanned the text of 2,500 works of literature from the Project Gutenberg archives, amounting to about 2.5 trillion letters. We looked at every possible window in each book (not just those starting at the beginning of the text) subject to the requirement, imposed for aesthetic reasons, that the window start and end on a sentence boundary.

The longest closed worm that we found, from *The Intrusion of Jimmy* by P. G. Wodehouse, has 313 letters:

> "Well?"
> "I have something to say to you."
> "I'm listening."
> Sir Thomas tried to rally. There was a touch of the old pomposity in his manner when he spoke.
> "Mr. Pitt, I find you in an unpleasant position–"
> Jimmy interrupted. "Don't you worry about my unpleasant position," he said. "Fix your attention exclusively upon your own. Let us be frank with one another. You're in the cart. What do you propose to do about it?"

Here is computer rendering of this worm:



In this picture the sphere marks the location of the start (and end) of the worm and the checkerboard (which is the XZ plane) is translucent so that the portion of the worm which lies below it can be seen. The letters from the first and last word of the text are shown next to their worm segments.

Although this is an impressive example of an Ouroboros worm, we would expect to find even longer examples if the directional letter assignments were unbiased, with a mean displacement as close to (0,0,0) as possible. In order to pursue this idea it is necessary to pick some exact values for the set of letter frequencies that will be used for calculating the bias. We chose to use these numbers, computed from a corpus of 3,104,375,038 letters of running text, given at *http://www.cryptograms.org/letter-frequencies.php*:

A: .08000395	B: .01535701	C: .02575785	D: .04317924	E: .12575645
F: .02350463	G: .01982677	H: .06236609	I: .06920007	J: .00145188
K: .00739906	L: .04057231	M: .02560994	N: .06903785	O: .07591270
P: .01795742	Q: .00117571	R: .05959034	S: .06340880	T: .09085226
U: .02841783	V: .00981717	W: .02224893	X: .00179556	Y: .01900888
Z: .00079130				

There are 26! (about 403 septillion) different ways of assigning the alphabet to the 26 directions, which is far too many combinations to examine even with a computer. We thought it would pleasing if a good, unbiased arrangement could be found having a reasonably short mnemonic that one could use to write down the 3x3x3 cube of letters from memory; this has the side benefit of reducing the size of the search space dramatically. We decided to examine a specific structure that we call the *double-nine* *arrangement*, in which the 26 letters are divided into two nine-letter words, to be placed on the top and bottom layers of the cube in the obvious way, and a set of eight letters for the middle layer.

The two nine-letter words must together consist of 18 different letters of the alphabet, which restricts the number of possibilities to a very small number. We exhaustively tried all pairs of nine-letter words of this type and, for each pair, all 8! = 40,320 ways of arranging the remaining eight letters. The best arrangement we found is this one,

U	N	M
I	X	E
D	L	Y

H	B	Z
W		V
F	G	Q

J	O	C
K	S	T
R	A	P

which can be remembered via this mnemonic, where the 26 underlined letters appear in the same order that they occur in the cube:

> HERBERT WAS UNMIXEDLY AMUSED TO LEARN THAT "HOW BIG IS A ZEBRA IN WEST VIRGINIA?" IS A FREQUENTLY GOOGLED QUESTION AMONG JOCKSTRAP USERS.

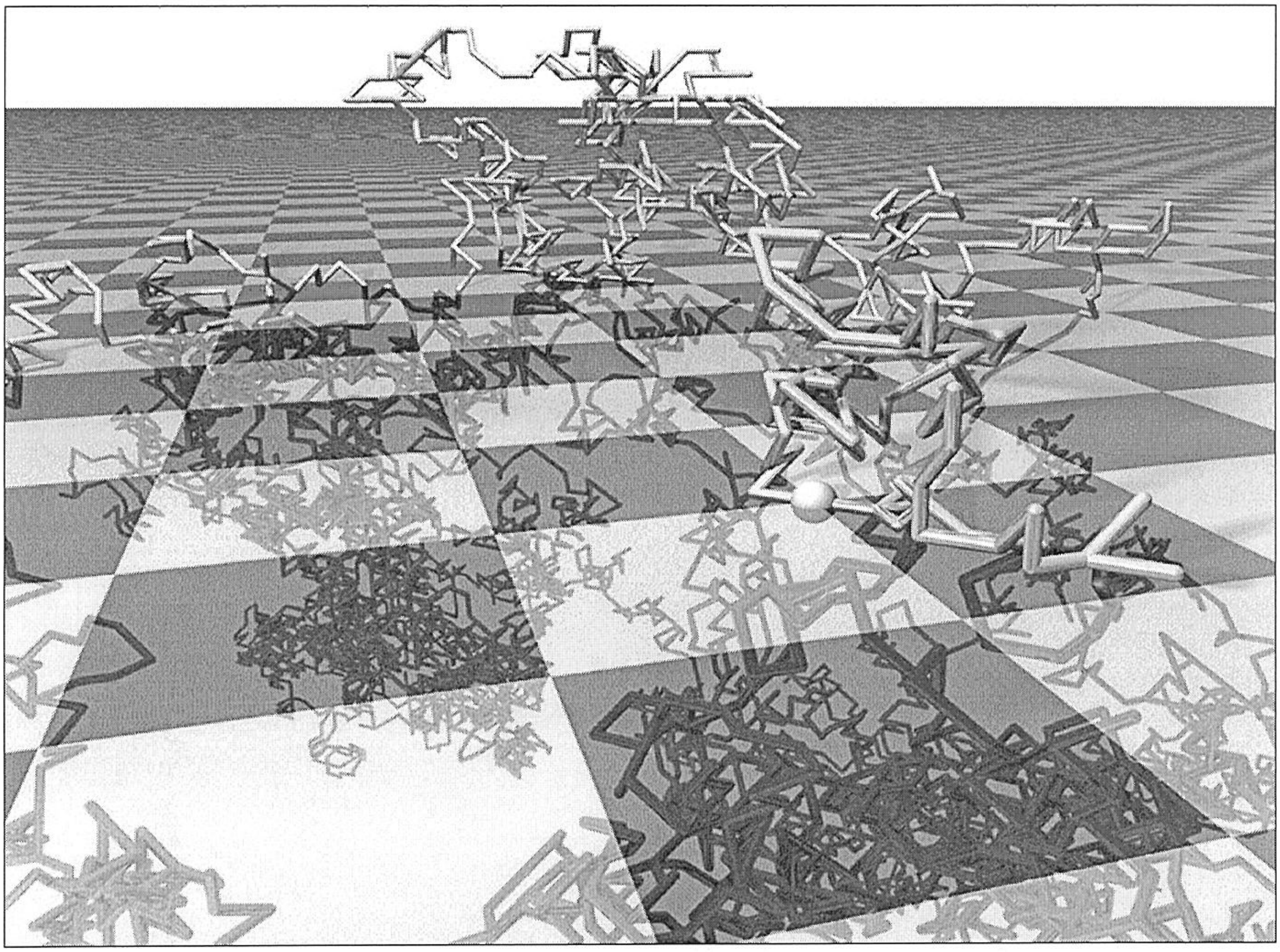
The bias of this scheme, calculated using the letter frequencies above, is (-.000631, .000117, .000244), which is quite small indeed, especially given the severeness of the double-nine constraint.

We reran our search on the Project Gutenberg files using this letter arrangement. The longest accidental closed worm that we found has 6342 letters, from Volume 1, Chapter 8 of *The Prince of India: or, Why Constantinople Fell*, an 1893 novel by Lew Wallace (author of *Ben-Hur*):

> The noise and stir of business at the ancient landing were engaging. With a great outcry, a vessel would be drawn up, and made fast, and the unloading begun. A drove of donkeys, or a string of camels, or a mob of porters would issue from the gate, receive the cargo and disappear with it.
>
> (*... 5836 letters elided ...*)
>
> The master, rich in experience, noticed the deferential manner of the reply, and was agreeably assured on his side.
>
> "Very well. There will be no harm in reserving an opinion," he said. "The good man, as you call him, is making ready a drink with which he has preceded me from his country, and which you must stay and share, as it is something unknown in the West."

Since we expected to find longer closed texts using this unbiased letter arrangement, we added an extra constraint to this search: the closed worm is only allowed to touch the start point at the beginning and end of the worm. It turns out that this restriction is not very onerous, being satisfied by most accidental closed worms. Note that the 313-letter example from P. G. Wodehouse shown earlier also obeys this rule.

Below is a picture showing almost all of this 6342-letter Ouroboros worm. (A small amount of the worm extends beyond the left, bottom, and right edges.) The first and last letters of the text are both T, so the two segments touching the starting point (the sphere) are collinear. As required, these are the only two times that the worm visits the origin.



How much better could we do if we drop the double-nine constraint and allow any arrangement of the alphabet? Before pursuing this question we remark that the path traced out by a word worm from a randomly chosen text is essentially what is known in mathematics as a three-dimensional random walk. The existence of closed worms is directly related to the probability that a random walk returns to its starting point. A surprising theorem, proved by Pólya in 1921, states that, while the probability of this happening in 1D and 2D is one (i.e., an infinitely-long unbiased random walk is guaranteed to return to the start eventually - infinitely often, in fact), the probability in three or more dimensions is strictly less than one, so there is a non-zero chance that the walk will never return. In 3D, the probability of returning to the start point (even once) is about 34%, the chance of no return 66%.

This mathematical result has an implication for the question asked at the start of the previous paragraph. We might think that with a suitably ideal arrangement of letters, arbitrarily long accidental closed word worms would become plentiful, but Pólya's random walk theorem tells us that they will be harder to find than one might intuitively expect.

Undeterred, we nevertheless tried to find an alphabetic ordering that can produce an accidental closed worm longer than 6342 letters, preferably in a well-known a work of literature. We added an extra constraint that the worm must start from the very beginning of the chosen book, but relaxed the requirement that it stop at the end of a sentence, allowing it to terminate on any letter.

Our grand prize winner is this arrangement,

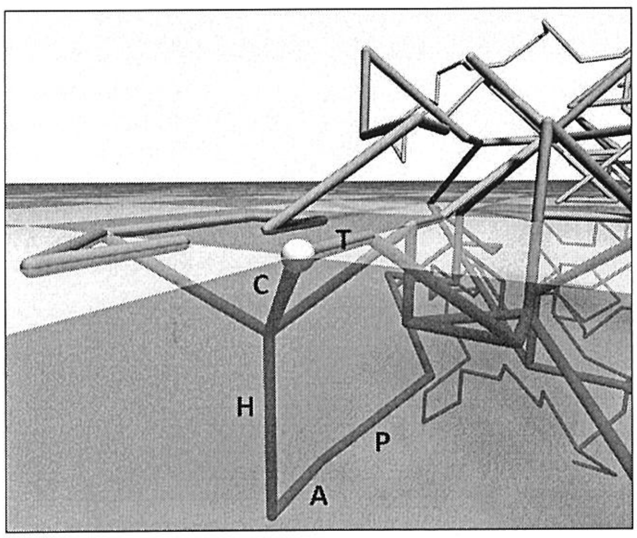
L	J	H
Q	O	G
D	K	U

W	E	P
I		N
Y	T	B

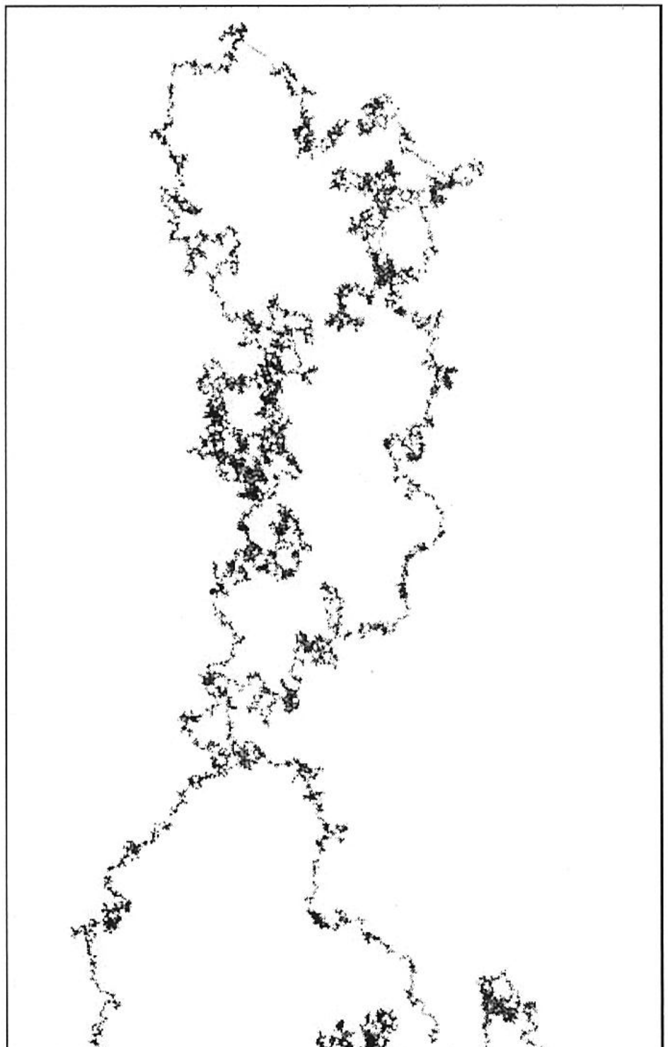
C	X	M
F	A	V
S	Z	R

which leads to a closed worm being formed by the first 408,173 letters (about 44%) of "Moby Dick", from "*Chapter 1 - Loomings. Call me Ishmael...*" to "*...Any man who has gone sailor in a whale-ship will understand this*" in the middle of chapter 50, with the worm ending at the T in "this".

Below is a closeup of the start point of this worm, with the first few letters of "Chapter" and the final T of the worm labelled. As before, the worm touches the origin only at its start and end.



To the right is a picture of the complete worm. It strays very far afield, reaching to a height of almost 8000 units, before eventually dipping below the plane briefly and then returning to the start.



Even better arrangements of the alphabet, leading to still longer accidental closed worms, must surely exist. Finding these is left as an open problem.