

## IS IT PLAGIARISM?

According to the site WordOrigins.Org (<http://www.wordorigins.org/number.htm>), there are about 20,000 English words that are used by educated persons.

### Accidentally Matching 5 Words is "Hitting the Lottery" Twice and then a Hole-In-One

In creating an original sentence, and assuming that the writer is free to choose any of the 20,000 words and use these words in any order, a series of 5 words that exactly match another source would have the "random chance" probability of

$(1/20000) \times (1/20000) \times (1/20000) \times (1/20000) \times (1/20000)$

or 1 chance in 3,200,000,000,000,000,000 (one chance in 3.2 sexillion).

*For the purposes of this illustration, we are ignoring common syntax rules such as the article of "a" or "an" is dependent upon the word that follows.*

This "accidental match" is far less likely than hitting the PowerBall Jackpot twice in a row.

*PowerBall uses 55 numbered balls, of which 5 are picked in random, and then a separate red "PowerBall" is randomly picked from a set of 42 numbered balls. Correctly selecting the 5 lottery balls and the one PowerBall is one chance in 146,107,962.*

The odds of hitting the jackpot twice in a row would be 1 in  $\{146,107,962 * 146,107,962\}$ , or 1 chance in 21,347,536,559,793,444.

Taking 3,200,000,000,000,000,000,000 and dividing it by 21,347,536,559,793,444 leaves you with a remainder of 149,900.

According to the National Hole-In-One Association, the odds of hitting a hole-in-one, even if you are an amateur, is 1 in 33,000.

So, you could hit the lottery jackpot twice in a row and still have the "odds" left over to hit a hole-in-one while golfing, and that is still LESS LIKELY than accidentally matching the same 5 words as another source.

### Smaller Dictionaries

Even if we assume that we are using a smaller subset of the language, which might be the 1000 most common words a person knows, the chance of picking in exact order the same 5 words as another writer would be

$(1/1000) \times (1/1000) \times (1/1000) \times (1/1000) \times (1/1000)$

or only 1 chance in 1,000,000,000,000,000 (one in a quadrillion).

Giving a writer a maximum benefit of doubt, given that writing in a particular field might yield a smaller subset of words, if we reduce the vocabulary list to only 50 distinct words, the chance that a writer will pick exactly the same 5 words in a phrase are:

$(1/50) \times (1/50) \times (1/50) \times (1/50) \times (1/50)$

or 1 chance in 312,500,000 (one chance in 312-million).

Of course, the chances of accidental matches to another writer's work are geometrically smaller the longer the list of words (for instance, a sequential match of 10 words with a vocabulary of 50 items would only occur by chance 1 in 97,656,250,000,000,000 times).

### Asummed Plagiarism

Although a single set match might occur by "chance" once in a paper, when multiple sentences have matching word orders (and these phrases have not been cited as originating from another source), the assumption can be that plagiarism has occurred (or at the least, that sources were not properly cited).

A commercial site that allows educators to analyze the contents of student papers can be found at: [www.turnitin.com](http://www.turnitin.com).

Students can see how to properly cite works by using the *Publication Manual of the American Psychological Association*, the Modern Language Association's *MLA Handbook for Writers of Research Papers*, or online sites such as [www.CitationMachine.net](http://www.CitationMachine.net)

James Falkofske  
©2006-2007