Abstract

It is known that humans can easily read words where the letters have been jumbled in a certain way. This paper examines this problem by associating a distance measure with the jumbling process. Modifications to text were generated according to the Damerau-Levenshtein distance and it was checked if the users are able to read it. Graphical representations of the results are provided.

Introduction

Aoccdrmig to a rscheearch at Cmabrigde Unievrtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rght pclae. The rset can be a toatl mses and you can stil raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey ltteer by istlef, but the wrod as a wlohe.

The above text has circulated on the Web for several years to show how powerful the human mind is in making sense of jumbled spellings. It may be viewed from the perspectives of joint error correction and coding [1] that is done simultaneously and automatically by the mind, or from the point of view of approximate string matching [2]-[6].

It has been proposed that the human brain is able to read the words even when they are jumbled because of the following properties

1. The grammatical structure of the sentence is not disturbed in the above sentence, that is the small words [of 2 or 3 letters] or the function words [by, the, is etc] are not jumbled. Since the grammatical structure is preserved, the user is able to predict the next word in the sentence. The jumbled text not only preserves the grammatical structure, it leaves almost 45-50% of the words correct (In the above paragraph that we took 46% of the words are unchanged.

2. People generally tend to notice the first and last letters more easily than they tend to observe the middle letters. So there is less possibility of finding errors in the middle letters than the initial and last letters.

3. Although the words are jumbled in the paragraph, the jumbled words are not new words, thus making the task of the reader easier.
4. The sound of the original word is preserved in the jumbled words. This also makes reading easy as people tend to read the word by its sound.
5. People read the jumbled text because of the context of the sentence.

The two things that interested me, in this paper, are the use of function words and the context that plays a part in guessing the next word in the sentence. I have decided to remove the function words from the paragraph and then use the same jumbling technique to study the effect of this change. Also, to break the context of the sentence, I have taken 100 independent words that are commonly use in everyday life and then applied the jumbling technique.

**Approximate String Matching**

Approximate string matching is the technique of performing string matching to the pattern of text. The match is measured in the number of operations that are performed to match the exact string. The most common operations that are performed to match the string are insertion, deletion and substitution. The number of operations performed is measured in terms of edit distance [13].

Examples of the operations are shown below:

**Insertion:** monkey $\rightarrow$ monkeys

**Deletion:** monkey $\rightarrow$ money

**Substitution:** monkey $\rightarrow$ donkey

All the above operations the number of edit distances performed are one. Some string matchers also consider transposition of two adjacent letters in the string [14].

**Transposition:** lost $\rightarrow$ lots

Approximate string matching has applications in many fields. Some examples are recovering the original signals after their transmission over noisy channels, finding DNA subsequences after possible mutations, and text searching where there are typing or spelling errors [6].

Most approximate string matchers assume same cost for all the operations performed in string matching, but some matchers do assign different weights to different operations. A more detailed description about edit distance and distance functions are explained in the distance measures section.

**Distance measures**

Edit distance is the number of operations performed to transfer one string into another string. There are different ways of performing the edit distance such as Levenshtein distance [7], Damerau-Levenshtein distance, Hamming distance, Jaro-Winkler distance, Longest common subsequence problem etc.
Levenshtein distance is a metric used to measure the difference between two sequences. This measure between two strings is defined by the number of edit operations used from transforming one string to another. The edit operations may be insertion, deletion and substitution of a single character. Here all the operations cost one unit. Levenshtein distance has a wide range of applications in areas such as spell checkers, dialect pronunciations and used in software’s for natural language translations [6].

As example the Levenshtein distance between Sunday and Monday is 2.

Sunday ->Munday (substituting M for S) ->Monday (substituting O for U).

Damerau-Levenshtein distance is similar to Levenshtein distance except that it includes an extra edit operation called the transposition of adjacent letters. Here all the operations also cost one. Damerau-Levenshtein distance has its applications in fields of fraud vendor name detections, where it can detect the letter that has been deleted or substituted, in DNA, where the variation between the two strands of DNA can be found out by this distance [6].

Hamming distance allows only substitution of letters, which cost one unit. It is applied only to the strings of similar length. It is applied in error detection and correction [6].

In this paper we apply the Damerau-Levenshtein distance to the words to find its effect on reading. This is because Damerau-Levenshtein distance has all the possible edit operations that can be performed.

**Experiment and Analysis:**

In all the experiments that I have conducted, I recorded the time each of 10 readers took to read the text. This time was then averaged.

1. **Removal of function words**

In this section I considered the actual paragraph and then removed all the function words from the paragraph to find the effect on the reader.

**Actual sentence**

According to research at an English university, it doesn't matter in what order the letters in a word are, the only important thing is that the first and last letter is at the right place. The rest can be a total mess and you can still read it without problem. This is because we do not read every letter by itself but the word as a whole.

**After jumbling**

Aoccdrnig to a rscheearch at Cmabrigde Unervtisy, it deosn't mtaer in waht oredr the ltteers in a wrod are, the olny iproootnt tihng is taht the frist and lsat ltteer be at the
The text can be a total mess and you can still read it without problem. This is because the human mind does not read every letter by itself, but the word as a whole.

Without function words

According research English university doesn't matter what order letters word only important thing first last letter right place. Rest total mess still read without problem. This because read every letter itself word whole.

After jumbling the above paragraph without function words

According research English university doesn't matter what order letters word only important thing first last letter right place. Rest total mess still read without problem. This because read every letter itself word whole.

Results for function words

![Graph 1: Time taken for function words](image)

From the graph, we can conclude that the function words don’t affect the reading.
II. Jumbling hundred independent words for testing the importance of context

In order to test the importance of context through which the readers are able to read the jumbled sentence, I took some 100 independent words that are commonly used and then jumbled them with their first and last letters in their original position.

100 independent words considered

- study basis exciting field utilize great interesting expanding many contributed
- normal second below genetic graduate notepad jumble compatible giving camping
- school comment simple scroll action achievement broad paste national essential
- friend dismal diminish greeting divide coming external proud activate recent
- money reader search invite competition scientist elevate programs international consistently
- replace symposium academic followed properties address platform knowledge windows interaction
- product console fraction participate gaining high people retail average dollar
- website waste heading several editing potential fragile spending future shoulders
- burden sector information confused upcoming serious assist substantial quality become
- common maintain require growing humor animal going finance internet women.

Then I jumbled these words with their first and last letters in their original position. The resulting words would be:

- sutdy bisas ecxiting feild utlizie gerat inretesitng exapidnng many cnortubied
- nomral sceond bolew geneitc grdataue nopetad jbmule coapmbitle giivng cmapnig
- sohcol cmoemnt sipmle sclrl atcion aihcveemnt braod psate naitanol esstneial
- firend dsiaml dimiinsh grteeing divdie conimg exetnral puord aitctave rcenet
moeny rdaer saerh iivnte cpmotetioin seicntsit eelavte porargms ietnnraioanl cosnetsntly

rpealce smyspuiam aacdimec floewod pporreites addser pltarofin konelwgde widnwos itinecartoin

porcudt cnolose fcartion paitrpicate gainig high ppoele rteial aevrgae dollar

wseite wtsae headnig sevearl editig ptonetial fargile spennidg fuutre suohedlrs

bruedn scetor inofmritaon cfnoused ucpmoing sreuois assist sbusnatait qlautiy bceome cmmoon matniian reuqrie griwong hmuor ainmal gniog finance ietnnret wemon.

The result for independent words

![Graph 2: Time taken for Independent words](image)

From the above graph, it is clear that the context plays an important role in reading text with jumbled words.
III. Damerau-Levenshtein distance

I applied the Damerau-Levenshtein distance of one to the original paragraph, obtaining the following text:

> Accrding to researh at an Enlgish univiersity, it dosn't matetr in wiat ordier the lettes in a werd are, the only impurtant thng is that the fist and last lette is at the rijght place. The rect can be a totul mess and you can stillt raed it wihout problle. Ths is becase we do not raed evry lette by it slf but the wurd as a whule.

The edit operations performed are addition, deletion, substitution and transposition of neighboring letters. The timings were then recorded.

Graph 3: Time taken for Damerau-Levenshtein distance of one

Then I took the same 100 independent words and used the Damerau-Levenshtein distance of one to find the effect on reading.

On performing the normal Damerau-Levenshtein edit operations on them, the resultant words are:

> stidy basex exsiting fiel utlize gdeat interesing exapidng masy contriuted
sturdy basisd excitng fieldh utilize greate interseting expading mabny contributd
normasl secd beloe genreic gradeuate notead jumble comparible givign campingh
school coment sinple scrllo actyion achievment broiad pastre nacionale essetnial
frind dismnl diminush greating deeide coming extrenal pruud activte recebt
monkey rader scarch invitee competetion scientist elvate progrms internatinal
replase symposium acedemic folowed proerties adrress platgorn knowlgerdge windwos
producrt consxole fractoin particippte gainnig hicgh peoplie reti retaiege dollor
wedsite wasteh headng severol editing potential frggile spenidng futuire shouledrs
burcen secror infomation confusef upcming seious assst substtntial quaity becmoe
comomn maintain requiere groving humuor amnial goig finunce intrenet wmen

When the first letter is kept the same and when the edit operations are performed, the resultant words are:

sturdy basisd excitng fieldh utilize greate interseting expading mabny contributd
normasl secd beloe genreic gradeuate notead jumble comparible givign campingh
school coment sinple scrllo actyion achievment broiad pastre nacionale essetnial
frind dismnl diminush greating deeide coming extrenal pruud activte recebt
monkey rader scarch invitee competetion scientist elvate progrms internatinal
replase symposium acedemic folowed proerties adrress platgorn knowlgerdge windwos
producrt consxole fractoin particippte gainnig hicgh peoplie retaiege dollor
wedsite wasteh headng severol editing potential frggile spenidng futuire shouledrs
burcen secror infomation confusef upcming seious assst substtntial quaity becmoe
comomn maintain requiere groving humuor amnial goig finunce intrenet wmen

When the first and last letters are kept the same and then, when the edit operations are performed the resultant words are:
Study basics exiting field utilize great interesting expanding many contributed
normal second below genetic graduate notepads jumble compatible giving campaigns
school comment simple scroll action achievement broad state national essential
friend dismal diminish greeting dividing coming external produce activate recent
money reader search invite competition scientist elevate programs international consistently
replace symposium academic followed properties address platform knowledge windows interaction
product console fraction participate gaining high people retail average dollar
website waste heading several editing potential fragile spending future shoulders
buden sector information confused upcoming serious assistance substantial quality become
common maintain require growing humor animal going finance internet women

If the edit operations are performed by the neighboring letters in the QWERTY keyboard, the resultant words are:

study basics exciting field utilize great interesting expanding many contributed
normal second below genetic graduate notepads jumble compatible giving campaigns
school comment simple scroll action achievement broad state national essential
friend dismal diminish greeting dividing coming external produce activate recent
money reader search invite competition scientist elevate programs international consistently
replace symposium academic followed properties address platform knowledge windows interaction
product console fraction participate gaining high people retail average dollar
website waste heading several editing potential fragile spending future shoulders
buden sector information confused upcoming serious assistance substantial quality become
common maintain require growing humor animal going finance internet women

The timings for all the paragraphs were recorded and plotted in Graph 4.
From the graph it can be inferred that if the first and last letters are kept the same and if the edit operations are performed the user is able to read the words with little difficulty. The user is also able to read the words if the edit operations are done with the neighbor letters in the QWERTY keyboard.

Graph 4: Time taken for Levenshtein distance of one

**Conclusion**

This paper tried to explain the idea behind the ease of reading jumbled words. The results show that the importance of functional words in reading these words is much less than proposed before. Nevertheless, context plays an important role in helping the user read such words. In addition, this paper applied the Damerau-Levenshtein distance of one to the words and found that the words can be read if the first and last letters are left in their places.
References

1. S. Kak, Joint encryption and error-correction coding, Proceedings of the 1983 IEEE Symposium on Security and Privacy, pp. 55-60 1983.
2. R. Lowrance and R.A. Wagner, An extension of the string-to-string correction problem. J. ACM 22, 177-183, 1975.
3. R. L. Kashyap and B. J. Oommen, An effective algorithm for string correction using generalized edit distances -I. Description of the algorithm and its optimality, Inform.Science, 23(2), 123-142, 1981.
4. T. Okuda, E. Tanaka, and T. Kasai, A method of correction of garbled words based on the Levenshtein metric, IEEE Trans. Comput., C-25, 172-177, 1976.
5. A. Marzal and E. Vidal, Computation of normalized edit distance and applications, IEEE Trans. on Pat. Anal. And Mach. Intel., PAMI-15, 926-932. 1993.
6. G. Navarro, A guided tour to approximate string matching. ACM Computing Surveys 33, 31-88, 2001.
7. V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10, 707-710, 1966.
8. Ayman Jarrous and Benny Pinkas, Secure Hamming Distance Based Computation and Its Applications.
9. S. Kak, Classification of random binary sequences using Walsh-Fourier analysis. IEEE Trans. On Electromagnetic Compatibility EMC-13, pp. 74-77, 1970.
10. S. Kak and A. Chatterjee, on decimal sequences. IEEE Transactions on Information Theory, IT-27, 647 – 652, 1981.
11. S. Kak, Encryption and error-correction coding using D sequences. IEEE Transactions on Computers, C-34, 803-809, 1985.
12. S. Kak, New results on d-sequences. Electronics Letters, 23, 617, 1987.
13. S. Skiena. Algorithm Design Manual (1st Ed.). Springer, 1998.
14. T. Ottmann, P. Widmayer (in German). Algorithmen und Datenstrukturen (4th Ed.). Spektrum Akademischer Verlag. pp. 636–9, 2002.