

An Efficiency Comparison of Document Preparation Systems Used in Academic Research and Development: A Re-analysis.

Christophe Pouzat

<2014-12-28 dim.>

1 Introduction

This is a re-analysis of the data presented in [An Efficiency Comparison of Document Preparation Systems Used in Academic Research and Development](#). My "interest" in this paper was triggered by a discussion on the [emacs org mode mailing list](#). Ignoring the "message" of the paper, what stroke me was the systematic use of bar graphs: a way of displaying data that **should never be used** since when many observations are considered, a box plot is going to do a much better job and when, like in the present paper, few (10 in each of the 4 categories) observations are available, a direct display or even a simple table is going to do a **much better** job. Since it turns out that the data are available both on the Plos web site and on [figshare](#), I decided to re-analyze them.

2 Getting the data, etc.

We get the data with:

```
wget http://files.figshare.com/1849394/S1_Materials.xlsx
```

Using for instance [unoconv](#), we can convert the Excel file into a friendlier csv file:

```
unoconv -f csv S1_Materials.xlsx
```

We then get the data with R `read.csv` function:

```
efficiency <- read.csv("S1_Materials.csv",header=TRUE,dec=",")
```

The description of this table is obtained with:

```
wget http://files.figshare.com/1849395/S2_Materials.txt
cat "S2_Materials.txt"
```

3 Making some figures

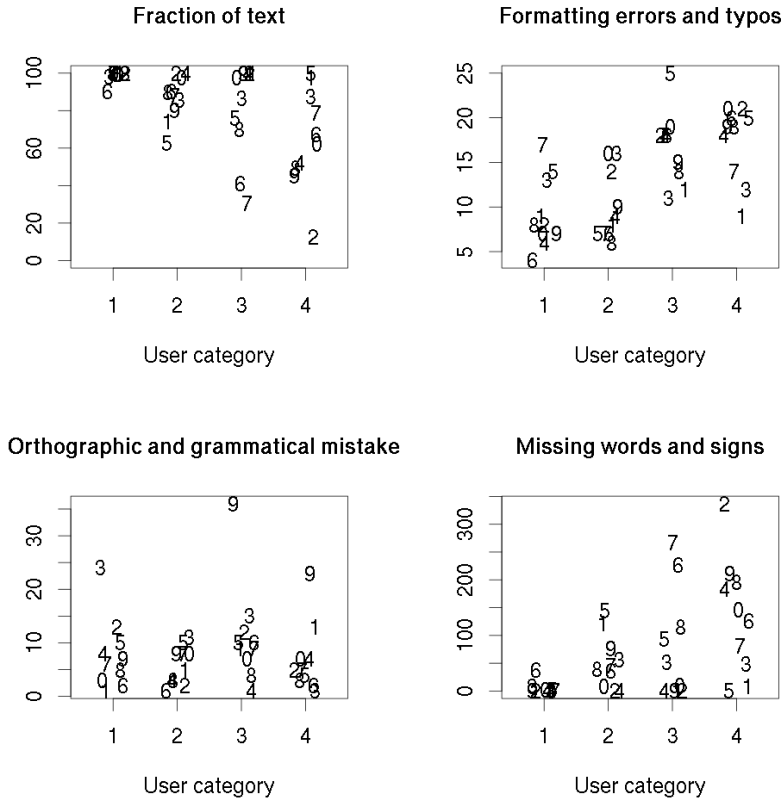
We can now make a figure out of the same data as figures 4, 5 and 6 of the paper but showing the actual data. We start with the "continuous text" exercise. We represent, in each of the four categories, each of the 10 individuals by a number between 0 and 9. Some horizontal jitter has been added to avoid overlaps. Category 1 corresponds to expert Word users; 2 to novice Word users; 3 to expert L^AT_EX users; 4 to novice L^AT_EX users:

```
layout(matrix(1:4,nc=2,byrow=TRUE))
par(cex=2)
plot(c(1,4),c(0,100),type="n",xlim=c(0.5,4.5),ylim=c(0,100),
     xlab="User category",ylab="",main="Fraction of text")
with(efficiency,
     sapply(1:4,
           function(k) points(runif(10,k-0.2,k+0.2),
                              PROZENT1[Kenntnisse==k],
                              pch = paste(0:9))))

with(efficiency,
     plot(c(1,4),c(0,100),type="n",
          xlim=c(0.5,4.5),ylim=range(FEHLERSFT),xlab="User category",
          ylab="",main="Formatting errors and typos"))
with(efficiency,
     sapply(1:4,
           function(k) points(runif(10,k-0.2,k+0.2),
                              FEHLERSFT[Kenntnisse==k],
                              pch = paste(0:9))))

with(efficiency,
     plot(c(1,4),c(0,100),type="n",xlim=c(0.5,4.5),
          ylim=range(FEHLEROFT),xlab="User category",ylab="",
          main="Orthographic and grammatical mistakes"))
with(efficiency,
     sapply(1:4,
           function(k) points(runif(10,k-0.2,k+0.2),
                              FEHLEROFT[Kenntnisse==k],
                              pch = paste(0:9))))

with(efficiency,
     plot(c(1,4),c(0,100),type="n",xlim=c(0.5,4.5),ylim=range(FEHLENDFT),
          xlab="User category",ylab="",main="Missing words and signs"))
with(efficiency,
     sapply(1:4,
           function(k) points(runif(10,k-0.2,k+0.2),
                              FEHLENDFT[Kenntnisse==k],
                              pch = paste(0:9))))
```



Notice that the number of "missing words and signs" exactly mirrors the fraction of written text. We will see that this observation holds for the two following exercises. This "missing words and signs" is always roughly ten times as large as the two other sources of mistakes. This explains the inverse relationship between the blue and pink bars on each of the 3 figures.

Let's keep going with the "table exercise":

```

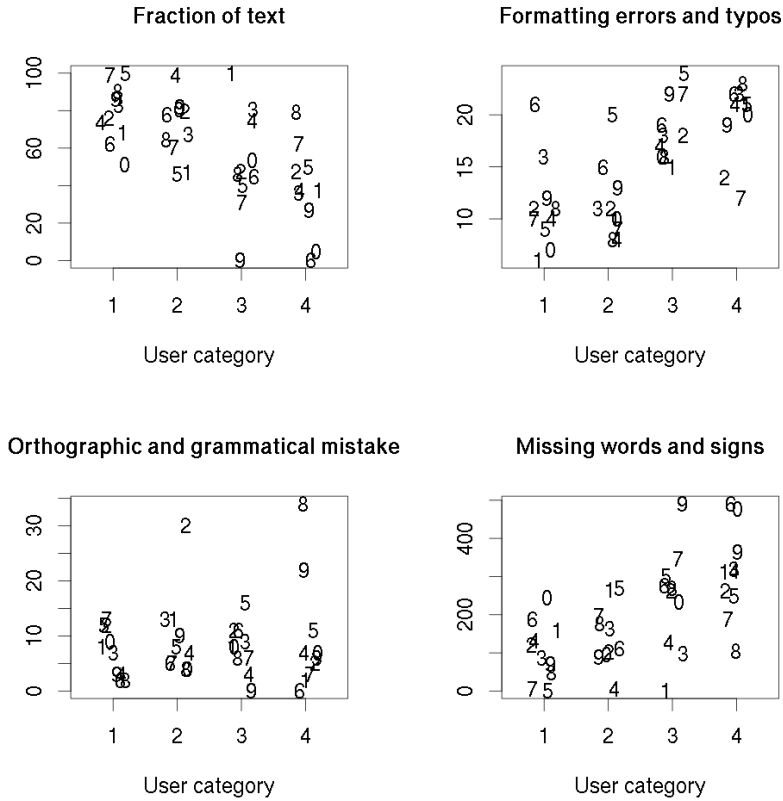
layout(matrix(1:4,nc=2,byrow=TRUE))
par(cex=2)
plot(c(1,4),c(0,100),type="n",xlim=c(0.5,4.5),ylim=c(0,100),
     xlab="User category",ylab="",main="Fraction of text")
with(efficiency,sapply(1:4,
                      function(k) points(runif(10,k-0.2,k+0.2),
                                           PROZENT2[Kenntnisse==k],
                                           pch = paste(0:9))))

with(efficiency,plot(c(1,4),c(0,100),type="n",xlim=c(0.5,4.5),
                    ylim=range(FEHLERST),xlab="User category",
                    ylab="",main="Formatting errors and typos"))
with(efficiency,sapply(1:4,
                      function(k) points(runif(10,k-0.2,k+0.2),
                                           FEHLERST[Kenntnisse==k],
                                           pch = paste(0:9))))

with(efficiency,plot(c(1,4),c(0,100),type="n",xlim=c(0.5,4.5),
                    ylim=range(FEHLEROT),xlab="User category",
                    ylab="",main="Orthographic and grammatical mistakes"))
with(efficiency,sapply(1:4,
                      function(k) points(runif(10,k-0.2,k+0.2),
                                           FEHLEROT[Kenntnisse==k],
                                           pch = paste(0:9))))

with(efficiency,plot(c(1,4),c(0,100),type="n",xlim=c(0.5,4.5),
                    ylim=range(FEHLENDT),xlab="User category",ylab="",
                    main="Missing words and signs"))
with(efficiency,sapply(1:4,
                      function(k) points(runif(10,k-0.2,k+0.2),
                                           FEHLENDT[Kenntnisse==k],
                                           pch = paste(0:9))))

```



We also see a strange thing here: in each of the expert \LaTeX and the novice \LaTeX users we have one individual who did not right anything but still manage to produce 22 "formatting errors and typos" (!) but luckily no orthographic or grammatical error. . .

```
with(efficiency, cbind(c(PROZENT2[Kenntnisse==3][10],
  FEHLERST[Kenntnisse==3][10],
  FEHLEROT[Kenntnisse==3][10],
  FEHLENDT[Kenntnisse==3][10]),
  c(PROZENT2[Kenntnisse==4][7],
  FEHLERST[Kenntnisse==4][7],
  FEHLEROT[Kenntnisse==4][7],
  FEHLENDT[Kenntnisse==4][7])))
```

```
0 0
22 22
0 0
491 491
```

Now for the "equations" exercise:

```

layout(matrix(1:4,nc=2,byrow=TRUE))
par(cex=2)
plot(c(1,4),c(0,100),type="n",xlim=c(0.5,4.5),ylim=c(0,100),
     xlab="User category",ylab="",main="Fraction of text")
with(efficiency,
     sapply(1:4,
            function(k) points(runif(10,k-0.2,k+0.2),
                               PROZENT3[Kenntnisse==k],
                               pch = paste(0:9))))

with(efficiency,
     plot(c(1,4),c(0,100),type="n",xlim=c(0.5,4.5),
          ylim=range(FEHLERSFOR),xlab="User category",ylab="",
          main="Formatting errors and typos"))
with(efficiency,
     sapply(1:4,
            function(k) points(runif(10,k-0.2,k+0.2),
                               FEHLERSFOR[Kenntnisse==k],
                               pch = paste(0:9))))

with(efficiency,
     plot(c(1,4),c(0,100),type="n",xlim=c(0.5,4.5),ylim=range(FEHLEROFOR),
          xlab="User category",ylab="",
          main="Orthographic and grammatical mistakes"))
with(efficiency,
     sapply(1:4,
            function(k) points(runif(10,k-0.2,k+0.2),
                               FEHLEROFOR[Kenntnisse==k],
                               pch = paste(0:9))))

with(efficiency,
     plot(c(1,4),c(0,100),type="n",xlim=c(0.5,4.5),
          ylim=range(FEHLENDFOR),xlab="User category",ylab="",
          main="Missing words and signs"))
with(efficiency,
     sapply(1:4,
            function(k) points(runif(10,k-0.2,k+0.2),
                               FEHLENDFOR[Kenntnisse==k],
                               pch = paste(0:9))))

```

