Lexique

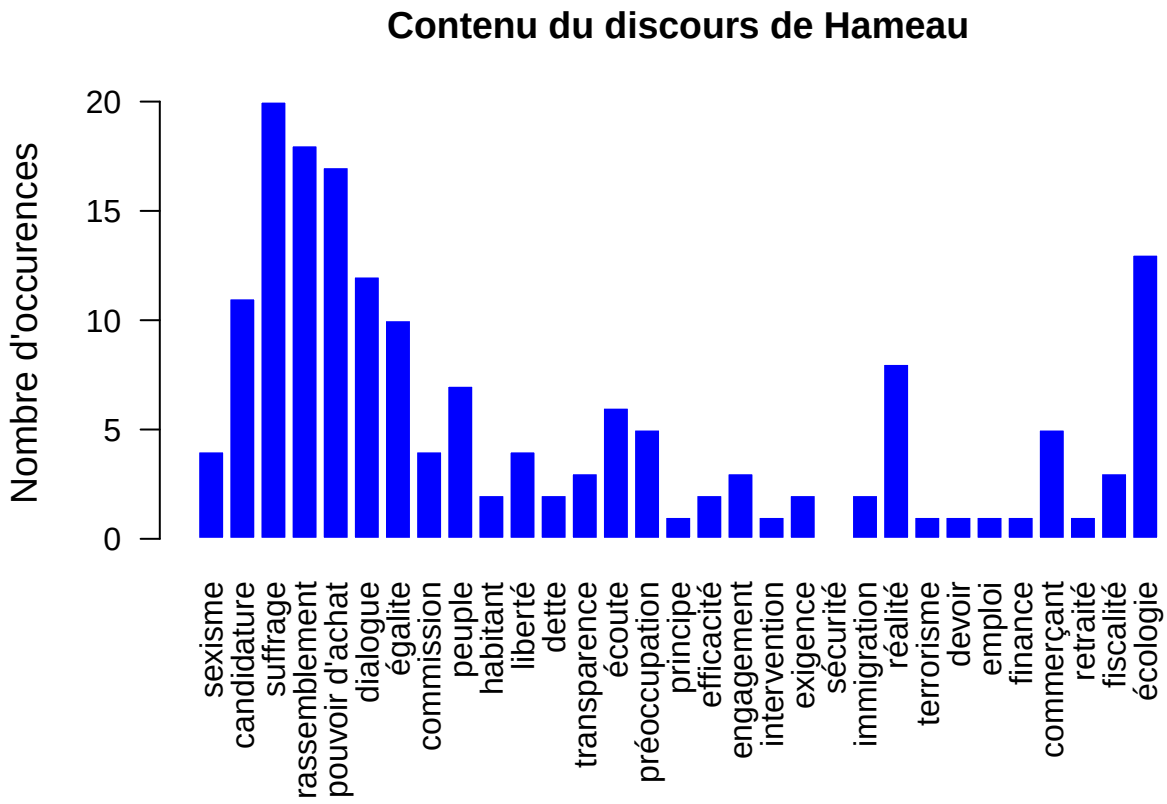
Arthur et Leonardo

12/21/2020

Exercice 2

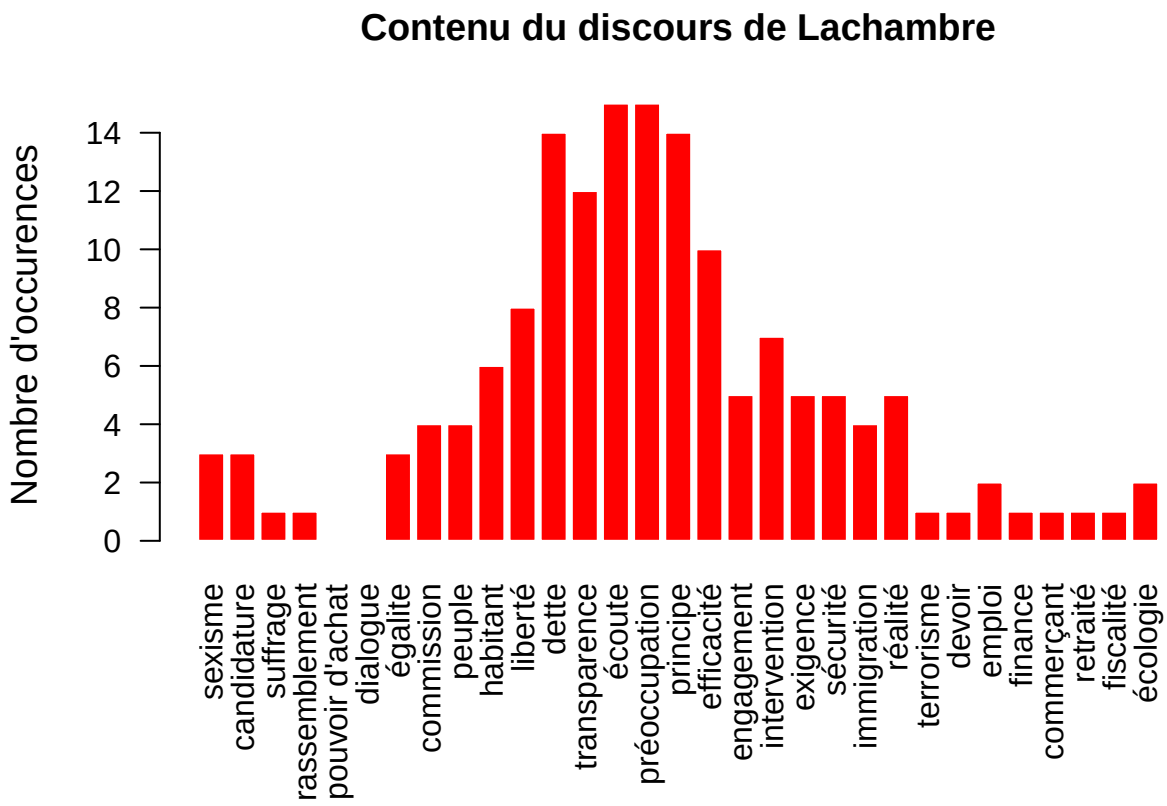
Detailler le contenu du discours de Hameau et de Lachambre

```
par(mar=c(7,4.5,4,1))
barplot( lexique[1:31,"Hameau"],
         names.arg = rownames(lexique)[1:31],
         col = "blue",
         border = "white",
         main = "Contenu du discours de Hameau",
         ylab = "Nombre d'occurences",
         las = 2,
         cex.lab = 1.2)
```



Nous pouvons remarquer que les termes qui apparaissent le plus souvent sont: *suffrage*, *pouvoir d'achat*, *égalité*, *dialogue*, *candidature* et *écologie*. Tandis que les termes apparaissant très peu de fois sont: *sécurité*, *terrorisme*, *devoir*, *emploi*, *finance*, *retraité*, *intervention* et *principe*. Le terme qui apparaît le plus de fois est *suffrage* et celui qui apparaît le moins de fois est *sécurité*.

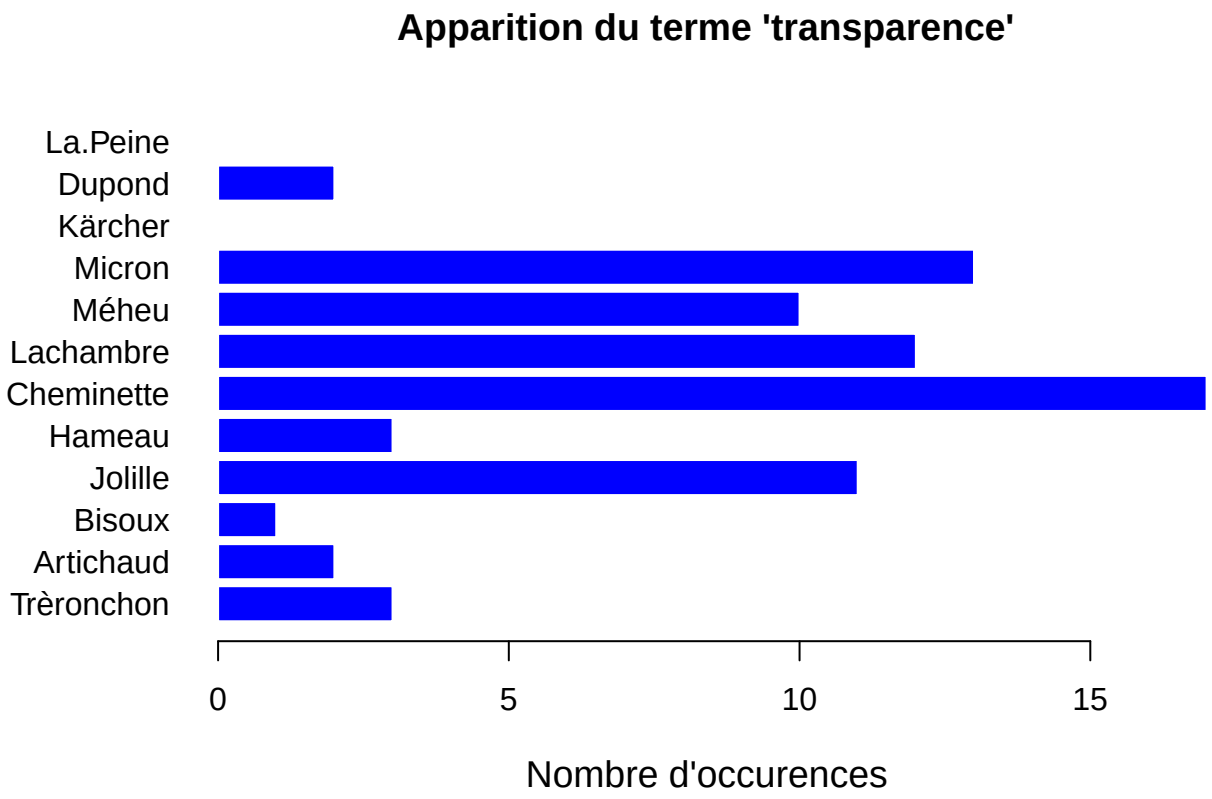
```
par(mar=c(7,4.5,4,1))
barplot(lexique[1:31,"Lachambre"],
        names.arg = rownames(lexique)[1:31],
        col = "red",
        border="white",
        main="Contenu du discours de Lachambre",
        ylab="Nombre d'occurences",
        las=2,
        cex.lab=1.2)
```



Dans le discours de Lachambre, les termes qui apparaîssent peu de fois sont: *pouvoir d'achat*, *dialogue*, *rassemblement*, *suffrage*, *terrorisme*, *devoir*, *finance*, *commerçant*, *retraité* et *fiscalité*. Les termes *pouvoir d'achat* et *dialogue* n'appairaissant pas du tout dans son discours. Les termes qui apparaissent le plus de fois sont: *dette*, *transparence*, *écoute*, *préoccupation*, *principe* et *efficacité*. Les termes *écoute* et *préoccupation* étant ceux qui apparaissent plus de fois que tout les autres (15 fois).

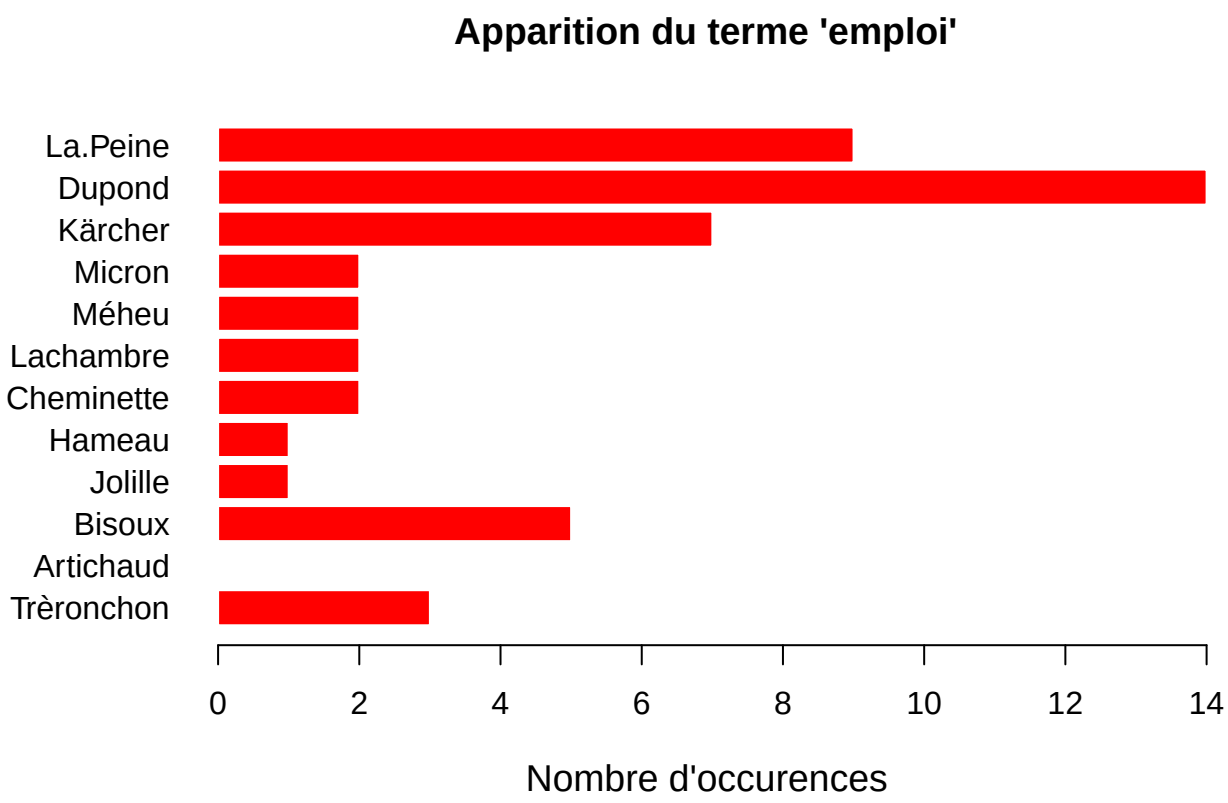
Détailler les apparitions des termes *transparence* et *emploi*

```
par(mar=c(4.5,5.5,4,1))
barplot(as.matrix(lexique["transparence",1:12]),
        names.arg=colnames(lexique)[1:12],
        col="blue",
        border="white",
        main="Apparition du terme 'transparence'",
        horiz = T,
        xlab="Nombre d'occurences",
        las=1,
        cex.lab=1.2)
```



Les personnages politiques utilisant le plus le terme *transparence* sont Micron, Méheu, Lachambre, Cheminette et Jolille. Les autres politiciens utilisent peu ce terme, notamment La Peine et Karcher qui n'ont pas utilisé ce terme une seule fois.

```
par(mar=c(4.5,5.5,4,1))
barplot(as.matrix(lexique["emploi",1:12]),
        names.arg=colnames(lexique)[1:12],
        col="red",
        border="white",
        main="Apparition du terme 'emploi'",
        horiz=T,
        xlab="Nombre d'occurences",
        las=1,
        cex.lab=1.2)
```



Les remarques importantes que nous pouvons déduire de ce graphe sont que Artichaud n'a jamais utilisé le terme 'emploi' dans ses discours. Alors que La Peine, Dupond et Karcher utilisent relativement souvent ce terme (Karcher en 7 occurences, Dupond en 14 et La Peine en 9).Les autres politiciens n'ont pas un nombre remarquable d'occurences du terme *emploi*.

Déterminer les attirances et répulsions entre les modalités

D'après le cours, le test du khi2 n'est valable que si chaque effectif théorique est supérieur à 5 et si l'effectif total est supérieur à 30. Regardons donc le tableau representant le nombre d'occurences théorique de chaque terme pour chaque personnage politique.
Remarque:
Nous afficherons uniquement une partie du tableau pour ne dégrader la lisibilité du document.

```
lex.chi2<-chisq.test(lexique)
```

```
## Warning in chisq.test(lexique): Chi-squared approximation may be incorrect
```

```
lex.chi2$expected[1:5,]
```

```
##                Trèronchon Artichaud   Bisoux   Jolille    Hameau Cheminette
## sexisme          2.869154  2.431950 1.120336  2.718865  2.322648   2.992118
## candidature      5.296900  4.489753 2.068313  5.019443  4.287966   5.523910
## suffrage         9.931687  8.418287 3.878087  9.411456  8.039937  10.357331
## rassemblement   14.014714 11.879138 5.472412 13.280610 11.345244  14.615344
## pouvoir d'achat  7.724645  6.547556 3.016290  7.320021  6.253284   8.055702
```

```
##                 Lachambre    Méheu   Micron  Kärcher    Dupond La.Peine
## sexisme          2.104046 1.366264 1.994745 1.325276  2.773516 1.981083
## candidature      3.884393 2.522333 3.682606 2.446663  5.120336 3.657383
## suffrage         7.283237 4.729375 6.904887 4.587493  9.600631 6.857593
## rassemblement   10.277457 6.673673 9.743563 6.473463 13.547556 9.676826
## pouvoir d'achat  5.664740 3.678403 5.370468 3.568050  7.467157 5.333684
```

Nous voyons que les effectifs ne sont pas assez importants, donc nous ne pourrons pas utiliser ce test pour étudier les attirances et répulsions entre les modalités.

Comment procédér?
Nous savons que :
n(ij) est l'effectif observé pour un terme i et pour le politicien j
t(ij) est l'effectif théorique pour un terme i et pour un politicien j
n est l'effectif total observé
n(i.) est la somme du nombre d'occurences d'un terme i
n(.j) est la somme du nombre d'occurences des termes pour un politicien j

D'après le cours $n(ij) > t(ij) <=> n(ij)/n(i.) > n(.j)/n$ et $n(ij) < t(ij) <=> n(ij)/n(i.) < n(.j)/n$.
Donc nous allons créer une matrice où l'élément se trouvant à la ligne i et à la colonne j prendra la valeur $n(ij)/n(i.) - n(.j)/n$ .De cette manière s'il s'agit d'une valeur négative nous pourrons conclure qu'il y a répulsion entre i et j, et s'il s'agit d'une valeur positive, il y a attirance.
Remarque:
Nous allons multiplier les valeurs par 100 pour améliorer la lisibilité.
Seule une partie de la matrice est affichée pour ne pas dégrader la lisibilité du document.

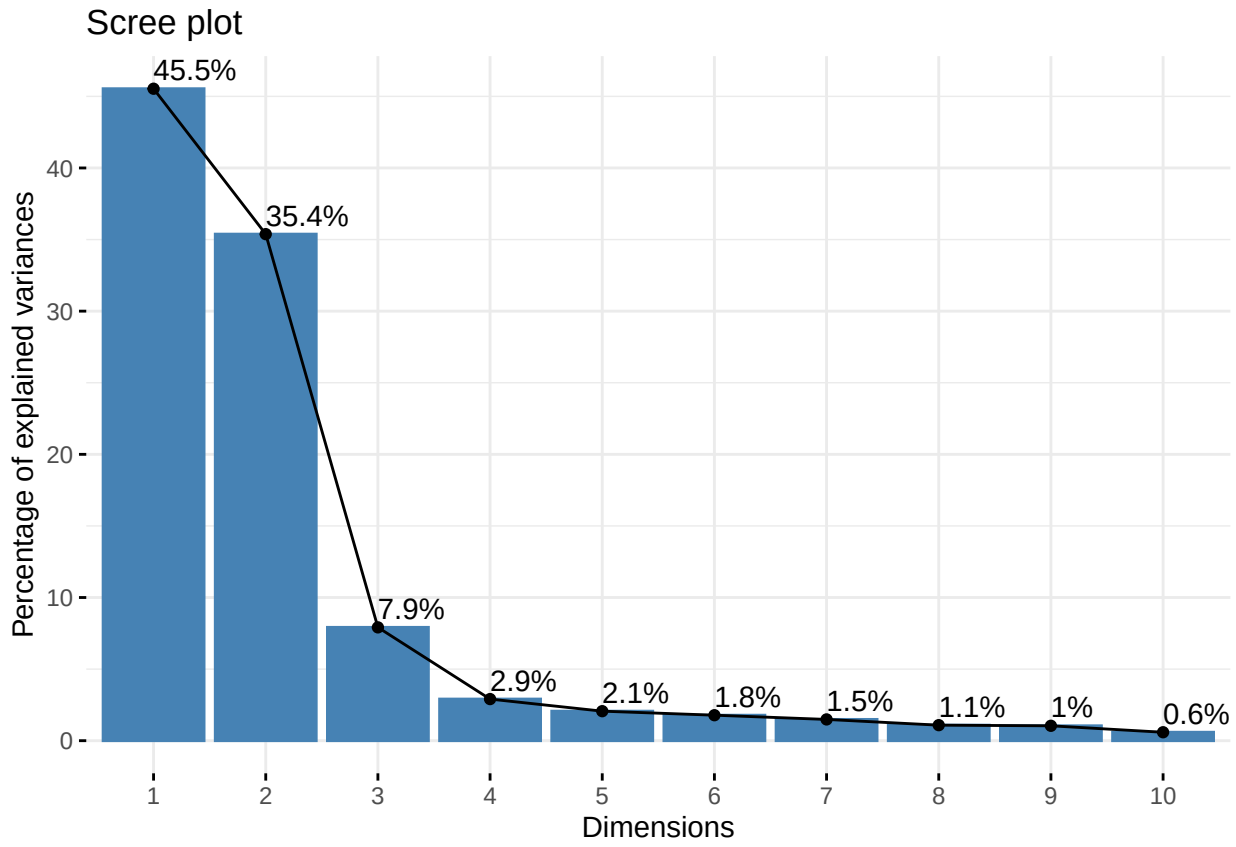
```
M = matrix(nrow = 31, ncol=12)
for (i in 1:31){
        for (j in 1:12){
              M[i,j]<-100*((lexique[i,j]/sum(lexique[i,]))- (sum(lexique[,j])/sum(lexique)))
        }
}

colnames(M)<-colnames(lexique)
rownames(M)<-rownames(lexique)
M[1:5,]
```

```
##                 Trèronchon Artichaud   Bisoux    Jolille    Hameau Cheminette
## sexisme           0.503254  9.877117 7.229476 -10.457173  6.451352  -3.815837
## candidature      13.964792 17.729681 1.941014 -10.457173 13.983403 -11.508145
## suffrage          8.964792 11.757459 7.913236  -8.234951 13.288959  -4.841478
## rassemblement     7.862430  5.606978 3.565030  -8.882370  5.239965  -9.933342
## pouvoir d'achat  17.536221 14.932062 5.691014 -10.457173 15.352451  -7.222431
##                 Lachambre     Méheu    Micron    Kärcher    Dupond    La.Peine
## sexisme          3.445976 -1.408707 -3.825943 -5.0972149 -2.975060  0.07275961
## candidature     -1.842486 -5.254861 -1.422097 -0.9305483 -8.584034 -7.61954808
## suffrage        -6.981374 -3.032639 -4.338763 -5.0972149 -5.111812 -4.28621475
## rassemblement   -7.305084 -5.254861 -6.884695  7.5012103  5.868066  2.61667239
## pouvoir d'achat -8.092486 -2.397718 -6.243525 -3.6686435 -9.238796 -6.19097665
```

Nous pourrions considérer uniquement les valeurs supérieures à 10 et inférieures à -10, pour mettre en évidence les valeurs saillantes. Par exemple: Il y a forte attirance entre le politicien Tréronchon et le terme *pouvoir d'achat*, et une forte répulsion entre Cheminette et le terme *candidature*.

Un autre moyen serait d'effectuer une Analyse Factorielle des Correspondances, choisir un nombre optimal de dimensions, puis d'étudier la distance entre les politiciens et les termes dans l'espace défini par ces dimensions.

```
lex.afc<-CA(lexique, graph=FALSE)
fviz_eig(lex.afc, addlabels=TRUE)
```



Le choix des dimensions à garder se fera grâce au critère de Kaiser qui propose de garder les axes ayant une inertie supérieure à l'inertie moyenne. Calculons l'inertie moyenne:

```
lex.afc$eig
```

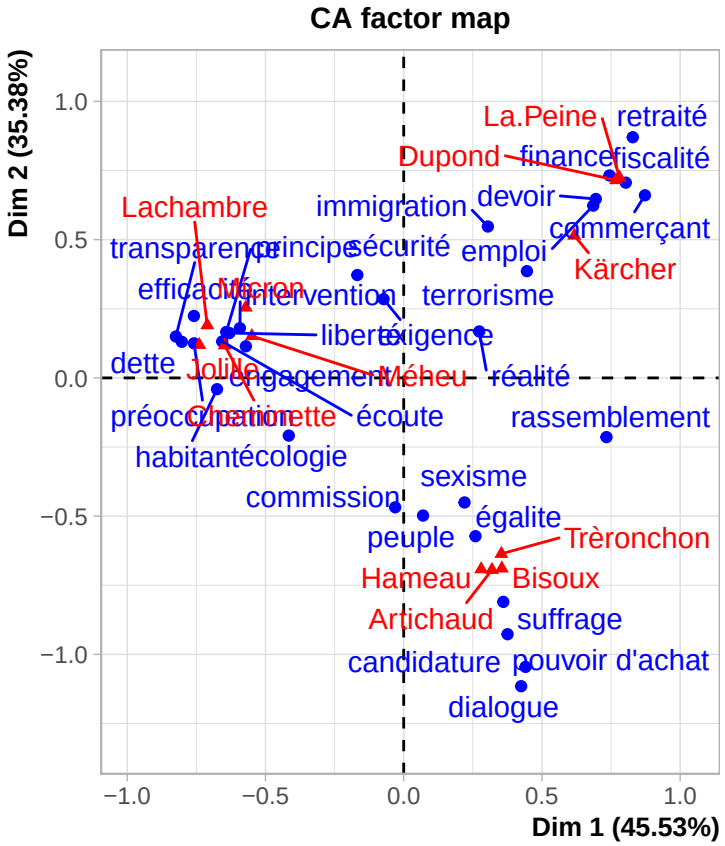
```
##         eigenvalue percentage of variance cumulative percentage of variance
## dim 1  0.351669348             45.5339506                           45.53395
## dim 2  0.273243739             35.3794466                           80.91340
## dim 3  0.061131239              7.9152387                           88.82864
## dim 4  0.022449659              2.9067693                           91.73541
## dim 5  0.015924805              2.0619349                           93.79734
## dim 6  0.013755945              1.7811120                           95.57845
## dim 7  0.011474761              1.4857456                           97.06420
## dim 8  0.008382715              1.0853893                           98.14959
## dim 9  0.008036271              1.0405319                           99.19012
## dim 10 0.004554304              0.5896887                           99.77981
## dim 11 0.001700598              0.2201925                          100.00000
```

```
inertie.moyenne<-sum(lex.afc$eig[,"eigenvalue"])/nrow(lex.afc$eig)
inertie.moyenne
```

```
## [1] 0.07021122
```

Nous garderons uniquement les axes 1 et 2 car leur inertie et supérieure à l'inértie moyenne.

```
plot.CA(lex.afc, axes=c(1,2),col.row="blue",col.col="red" , repel=TRUE)
```



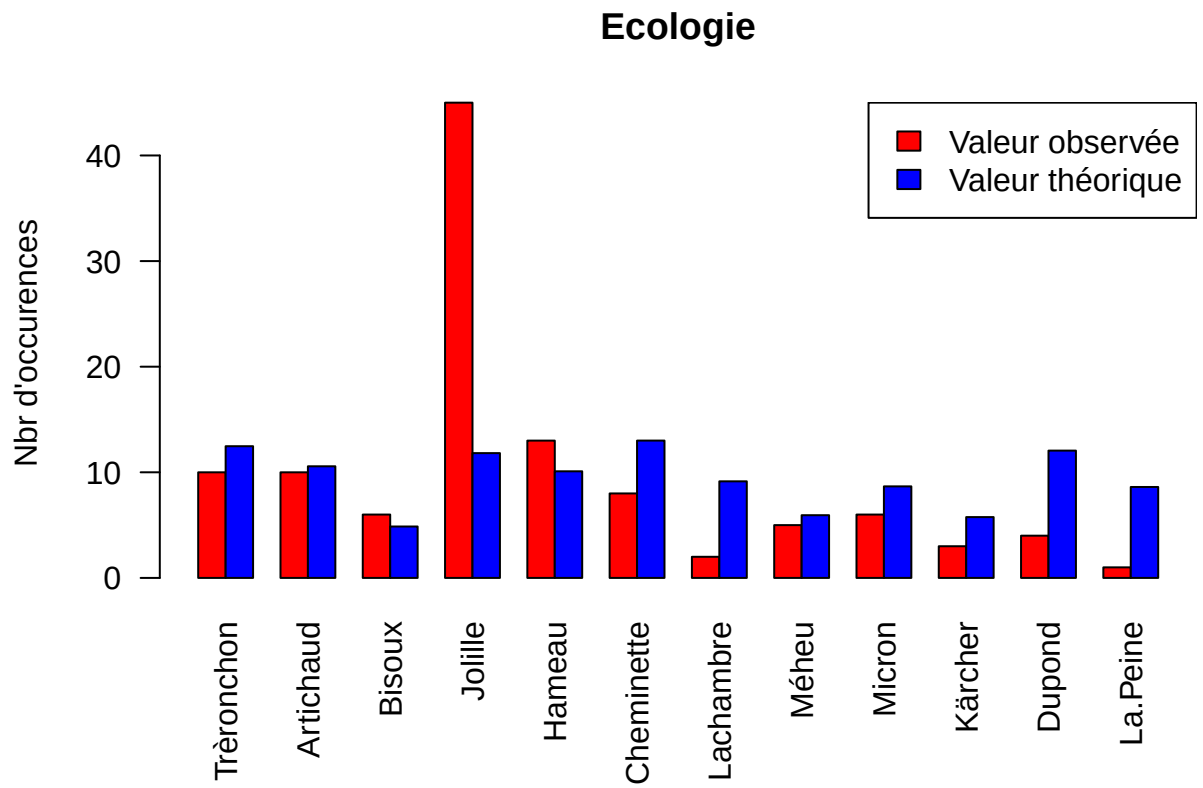
Pour interpreter ce graphique et pour trouver les attirances/répulsions entre les modalités, il suffit de regarder la distance entre les différentes modalités, c-à-d il faut regarder où se situent les politiciens par rapport aux différents termes. Prenons par exemple le terme *candidature*. Sur le graphique, il se trouve près de Tréronchon, Artichaud et Hameau donc nous pouvons conclure qu'il y a attirance entre ces politiciens et ce terme (En moyenne, ils l'utiliseront plus que les autres personnages politiques). Puis nous voyons aussi que le terme *candidature* est situé loin de Cheminette, Méheu, Dupond, Jolille et La Peine, donc il y a répulsion entre ces politiciens et ce terme.

Examiner la projection du terme *écologie* et les relations qu'il entretient avec les autres modalités

D'après le graphique de la question précédente nous pouvons voir que le terme écologie se trouve entre deux groupes de politicens. Le premier contenant Jolille, Cheminette, Méheu, Lachmbre et Micron, le deuxième contenant Hameau, Bisoux, Tréronchon et Artichaud. Le terme se trouve plus près du premier groupe mais pas non plus à une distance permettant de juger avec clarté les attirances et les répulsions. Pour mettre

en évidence les relations nous pouvons soit étudier la matrice M de la question précédente, soit faire un graphique mettant au clair la difference entre l'effectif observé et l'effectif théorique pour le terme *écologie*.

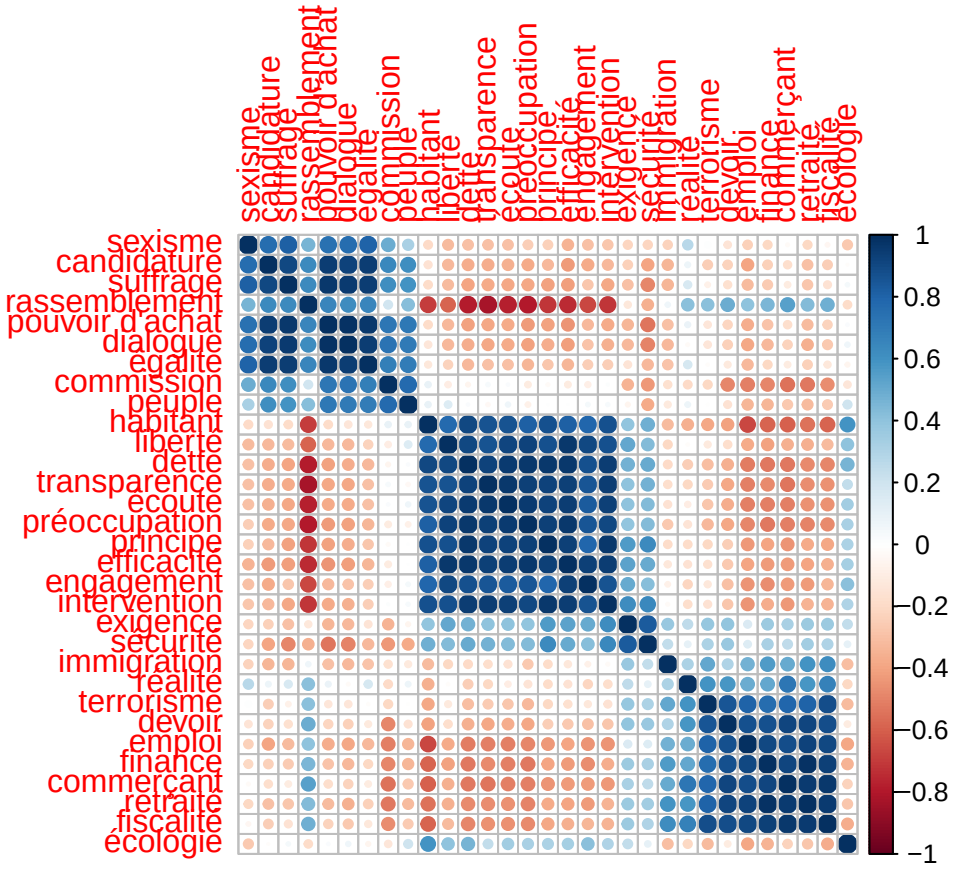
```
par(mar=c(6,4.5,4,1))
mx1<-as.matrix(lexique["écologie",])
mx2<-as.matrix(t(lex.chi2$expected["écologie",]))
mx<-rbind(mx1,mx2)
colours= c('red','blue')
barplot(mx, main="Ecologie", ylab="Nbr d'occurences", las=2,cex.axis = 1,beside=TRUE,col=colours)
legend('topright', fill=colours, legend=c('Valeur observée', 'Valeur théorique'))
```



Nous pouvons voir qu'il y a une forte attirance entre Jolille et *écologie*, une attirance moyenne avec Bisoux et Hameau, puis une répulsion avec tous les autres politiciens. Cela explique pourquoi sur le graphique de la question précédente, la projection du terme se trouvait entre deux groupes.

Nous allons de même tracer la matrice de corrélation entre les différents termes pour voir s'il y a des informations intéressantes:

```
corrplot(cor(t(lexique)))
```



D'après cette matrice, il existe une légère corrélation entre les termes *écologie* et *habitant* et entre *écologie* et *engagement* et une correlation inverse entre *écologie* et *finance*. Cependant ces relations restent négligeables.