Fondements statistiques - Exercice 2

Céline LY and Hugo LAULLIER

20 Décembre 2020

Chargement des librairies et des données

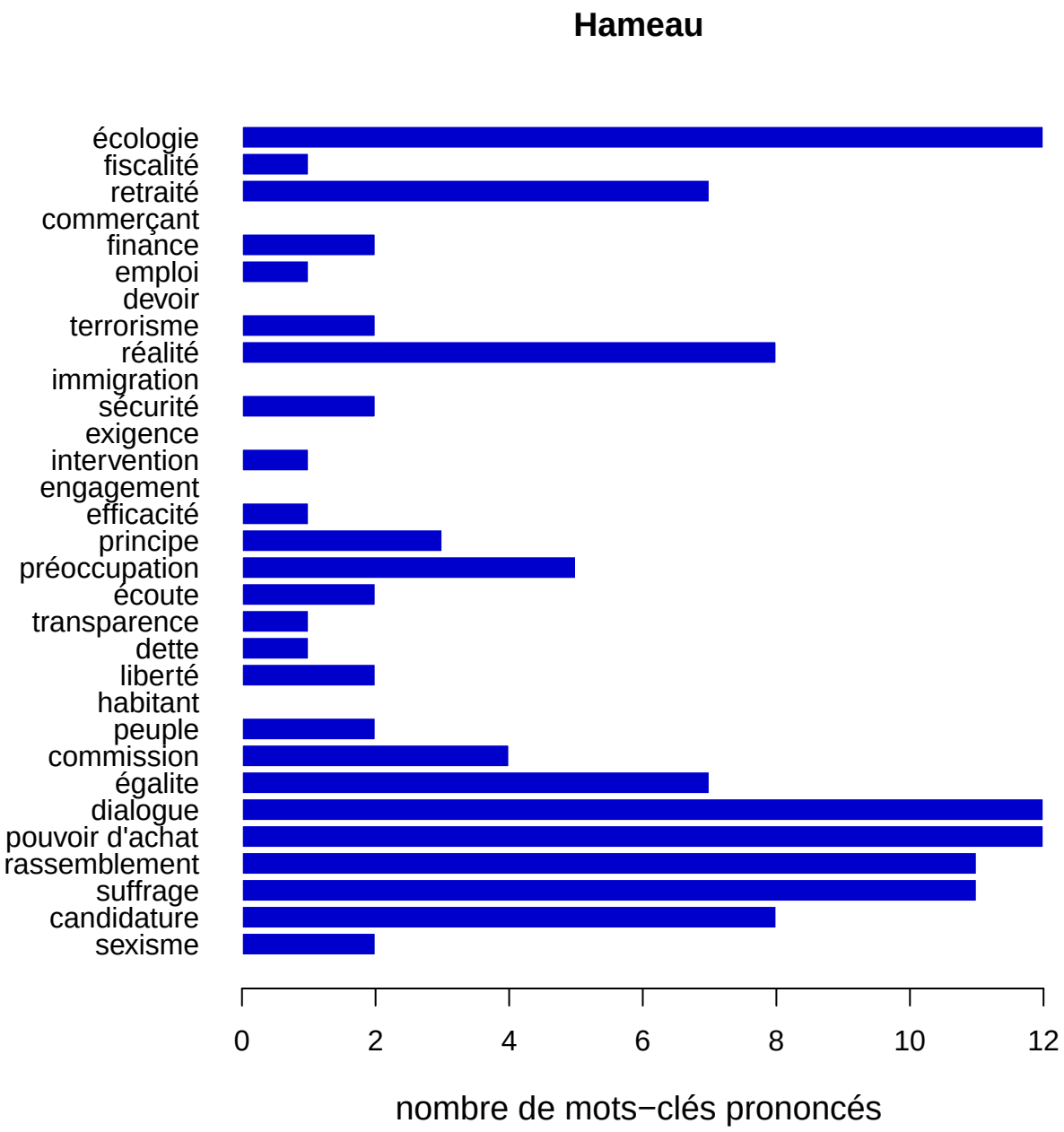
```
## Loading required package: FactoMineR
## Loading required package: factoextra
## Loading required package: ggplot2
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
## Loading required package: corrplot
## corrplot 0.84 loaded
## Loading required package: DescTools
```

Contenu du discours de Hameau et Bisoux

Tout d'abord, nous allons observer le nombre de fois que les mots-clés sont utilisés dans les discours des deux candidats.

Nombre de mots prononcés par Hameau

```
par(mar=c(5,7,4,2))
barplot( t(as.matrix(lexique[1:31, "Hameau"])),
names.arg = rownames(lexique)[1:31],
col="blue3",
border="white",
main="Hameau",
horiz=T,
las=1,
xlab="nombre de mots-clés prononcés",
cex.lab=1.2 )
```

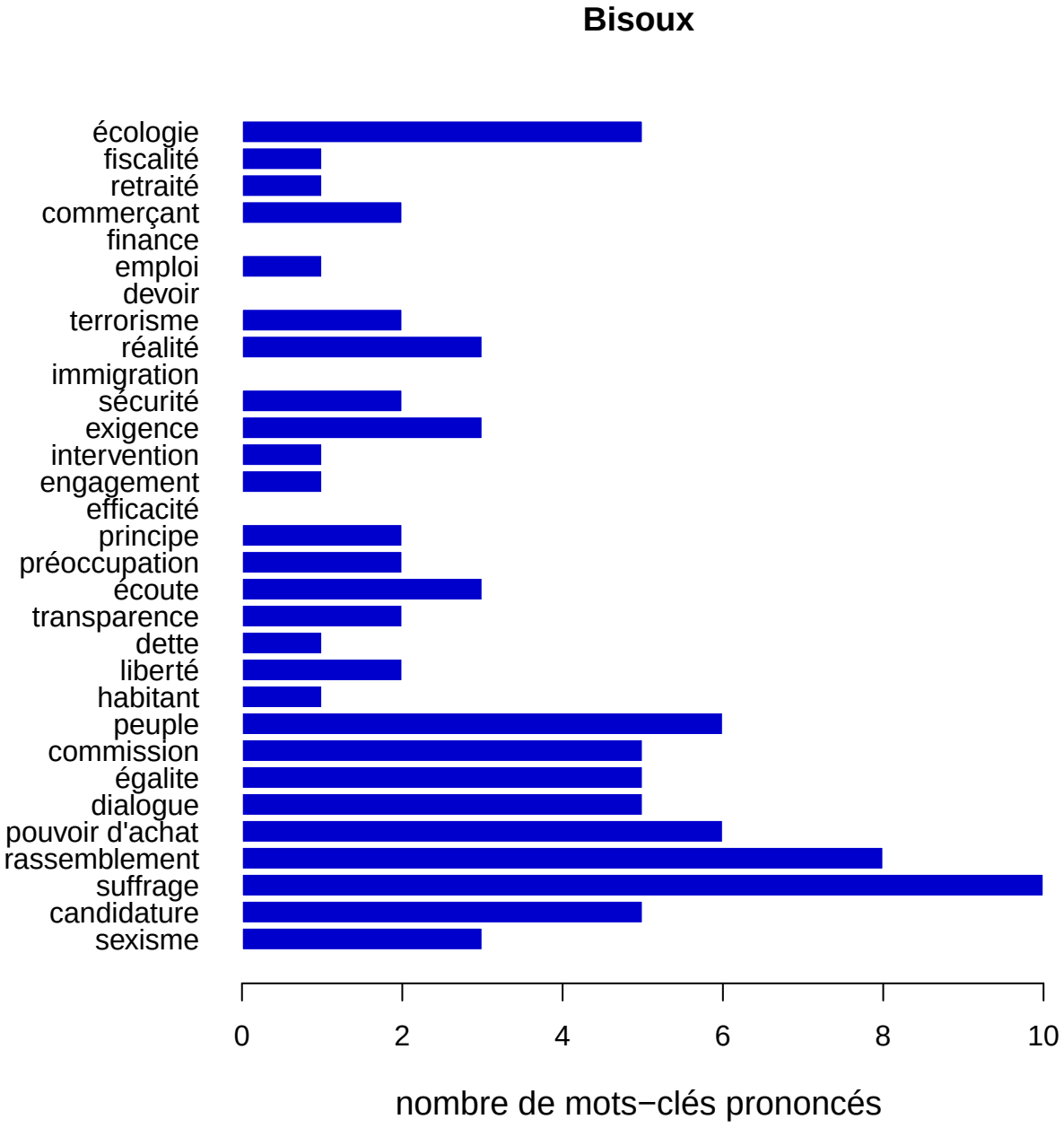


Dans ses discours, Hamaux a prononcé 12 fois les mots écologie, égalité et dialogue, et 10 fois les mots pouvoir d'achat et rassemblement. Nous pouvons aussi noter qu'il n'a jamais parlé de commerçant, de devoir, d'exigence, d'engagement ou encore d'habitant. Il s'agit probablement d'un candidat de gauche.

Nombre de mots-clés prononcés par Bisoux

```
par(mar=c(5,7,4,2))
barplot( t(as.matrix(lexique[1:31, "Bisoux"])),
names.arg = rownames(lexique)[1:31],
col="blue3",
border="white",
```

```
main="Bisoux",
horiz=T,
las=1,
xlab="nombre de mots-clés prononcés",
cex.lab=1.2)
```



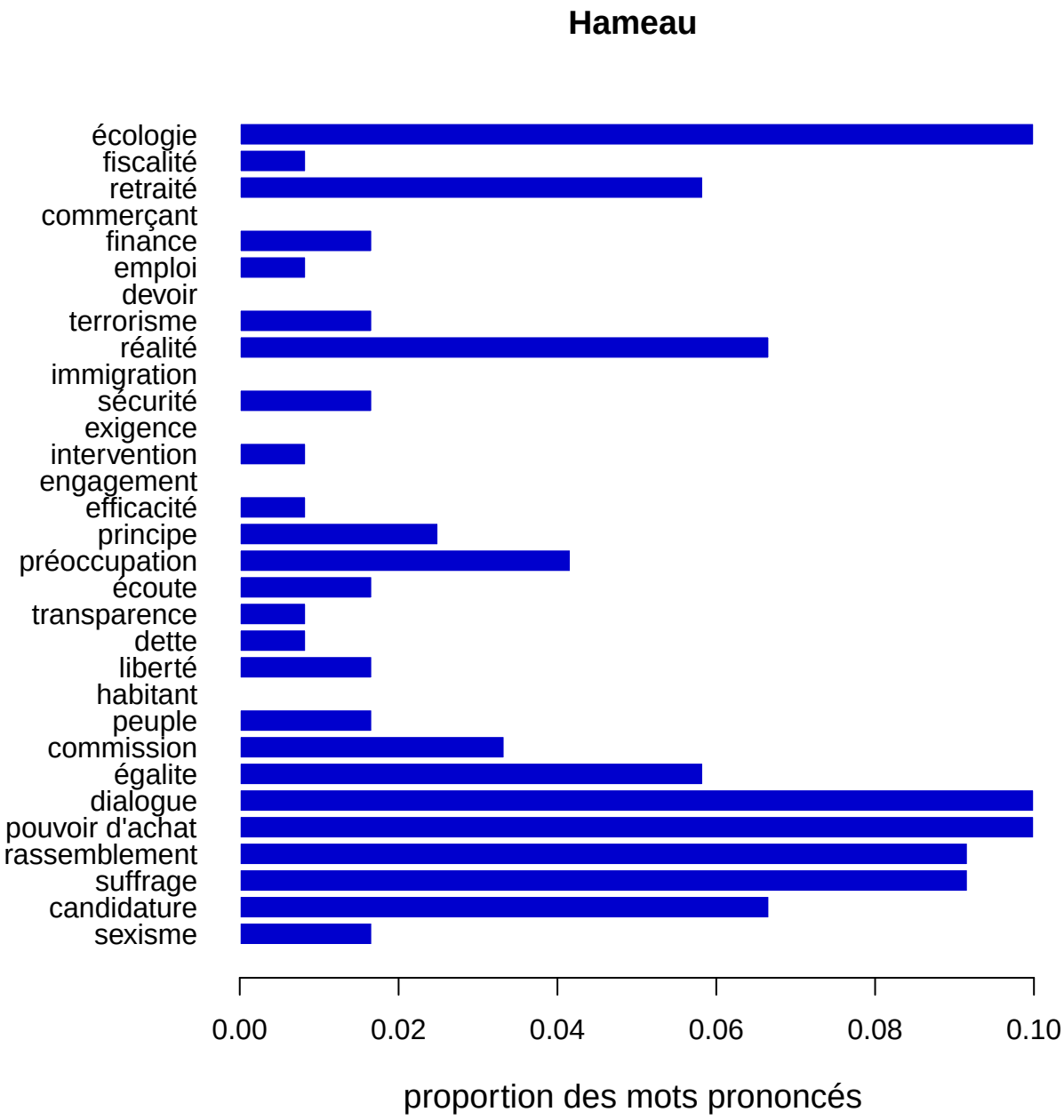
Dans ses discours, Bisoux a prononcé 10 fois le mot suffrage, 8 fois le mot rassemblement et 6 fois les mots peuple et dialogue. Nous pouvons aussi noter qu'il n'a jamais parlé de commerçant, de finance, de devoir, d'immigration ou encore d'efficacité. Il s'agit aussi probablement d'un candidat de gauche.

Afin de bien comprendre le contenu des deux discours et d'avoir des données permettant des compararaisons pertinente, nous pouvons aussi étudier non pas la quantité, mais la proportion des mots-clés présents dans

leurs discours.

Proportion des mots-clés prononcés par Hameau

```
profil_Hameau <- lexique[1:31, "Hameau"]/sum(lexique[1:31, "Hameau"])
par(mar=c(5,7,4,2))
barplot( t(as.matrix(profil_Hameau)),
names.arg = rownames(lexique)[1:31],
col="blue3",
border="white",
main="Hameau",
horiz=T,
las=1,
xlab="proportion des mots prononcés",
cex.lab=1.2)
```

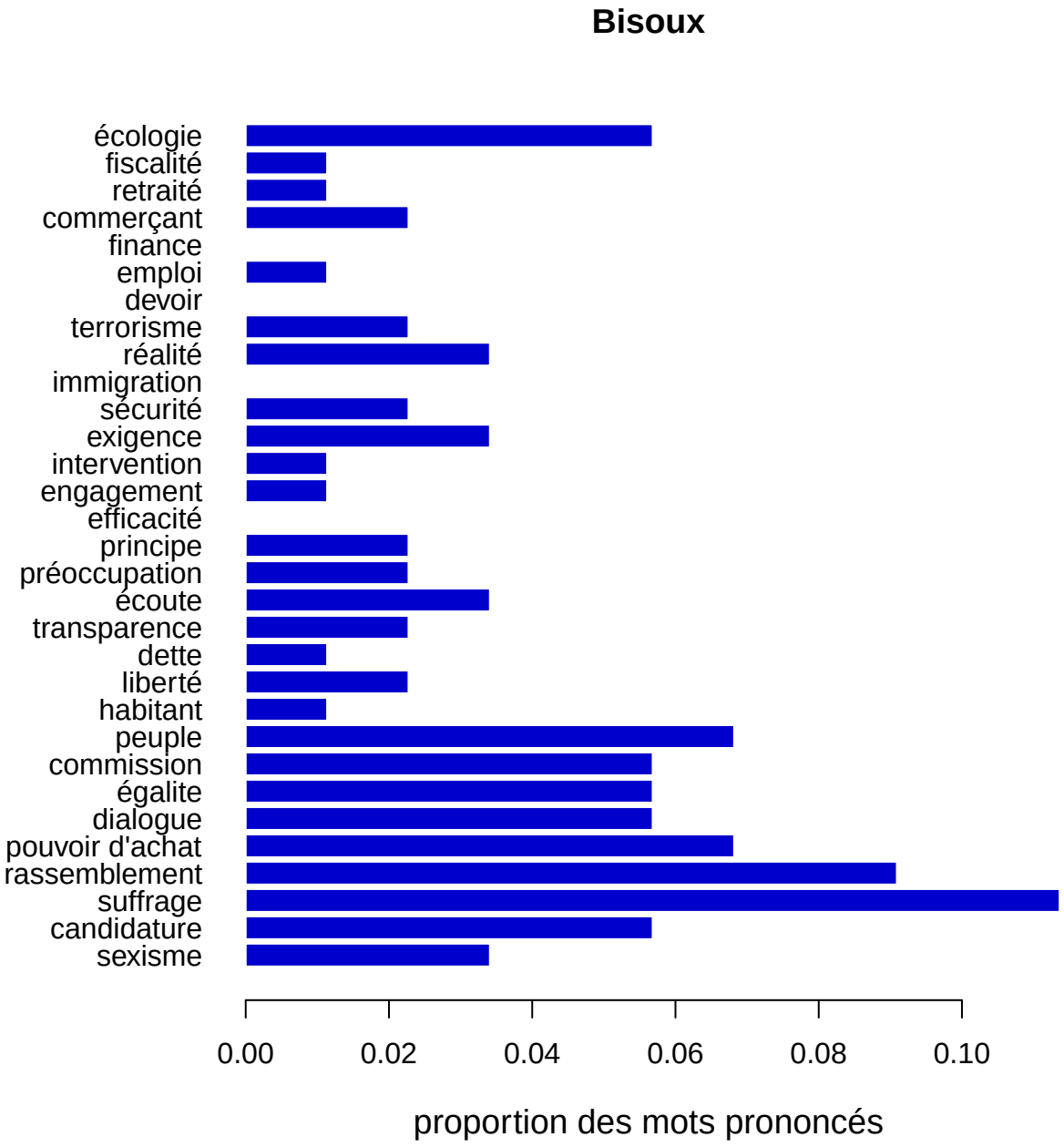


Dans ses discours, Hameau prononce 10% du temps les mots-clés écologie, dialogue et pouvoir d'achat, et à un peu plus que 9% du temps les mots-clés pouvoir d'achat et suffrage.

Proportion des mots-clés prononcés par Bisoux

```
profil_Bisoux <- lexique[1:31, "Bisoux"]/sum(lexique[1:31, "Bisoux"])
par(mar=c(5,7,4,2))
barplot( t(as.matrix(profil_Bisoux)),
names.arg = rownames(lexique)[1:31],
col="blue3",
border="white",
```

```
main="Bisoux",
horiz=T,
las=1,
xlab="proportion des mots prononcés",
cex.lab=1.2)
```

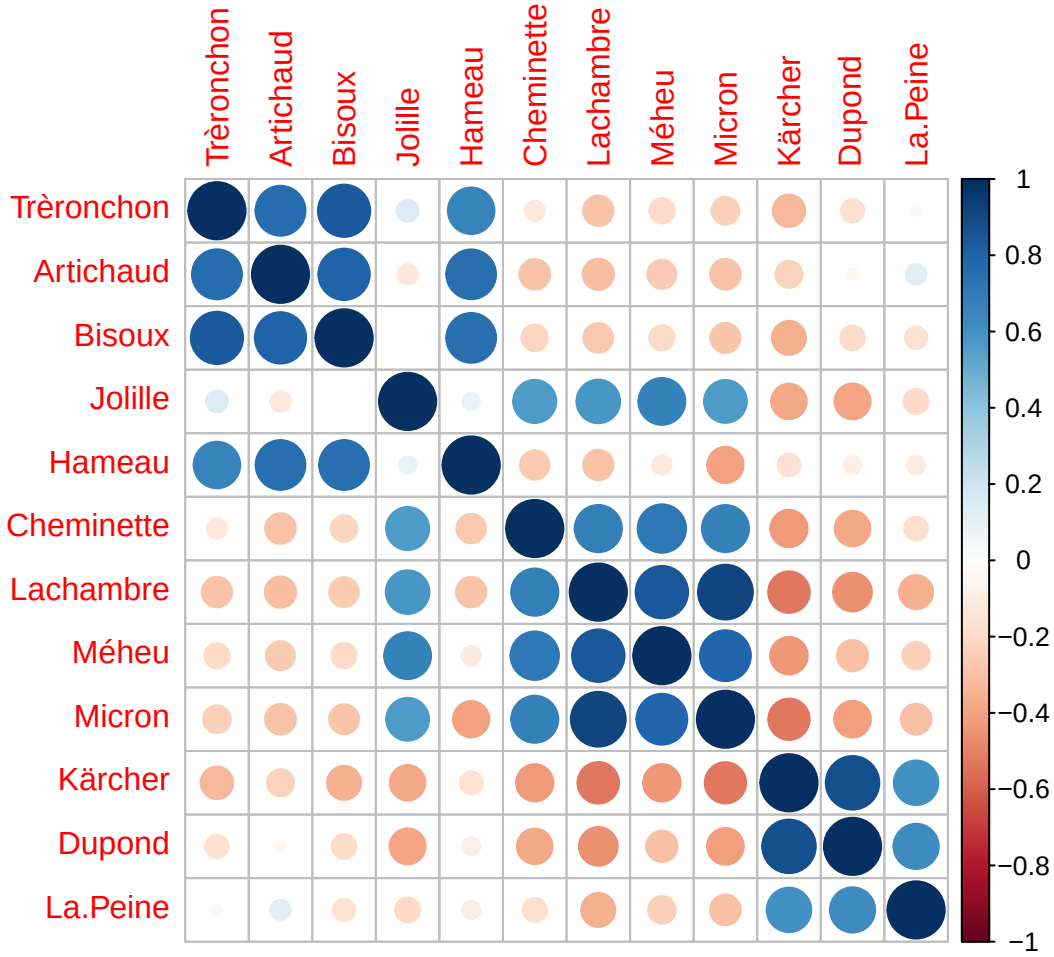


Dans ses discours, Hameau prononce à un peu plus de 11% du temps le mot-clé suffrage, 9% du temps le mot-clé rassemblement, et à environ 7% du temps les mots-clés dialogue et peuple.

Matrice de corrélations

Finalement, nous pouvons nous interesser à la matrice de corrélation, afin de voir si le contenu des discours des deux candidats est semblable.

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



Nous pouvons donc remarqué un fait plutôt intéressant : d'après la matrice de corrélation, le contenu des discours de Hameau et Bisoux sont corrélés. Ces deux candidats semblent partager de manière générale les mêmes idées (très sûrement des idées de gauche).

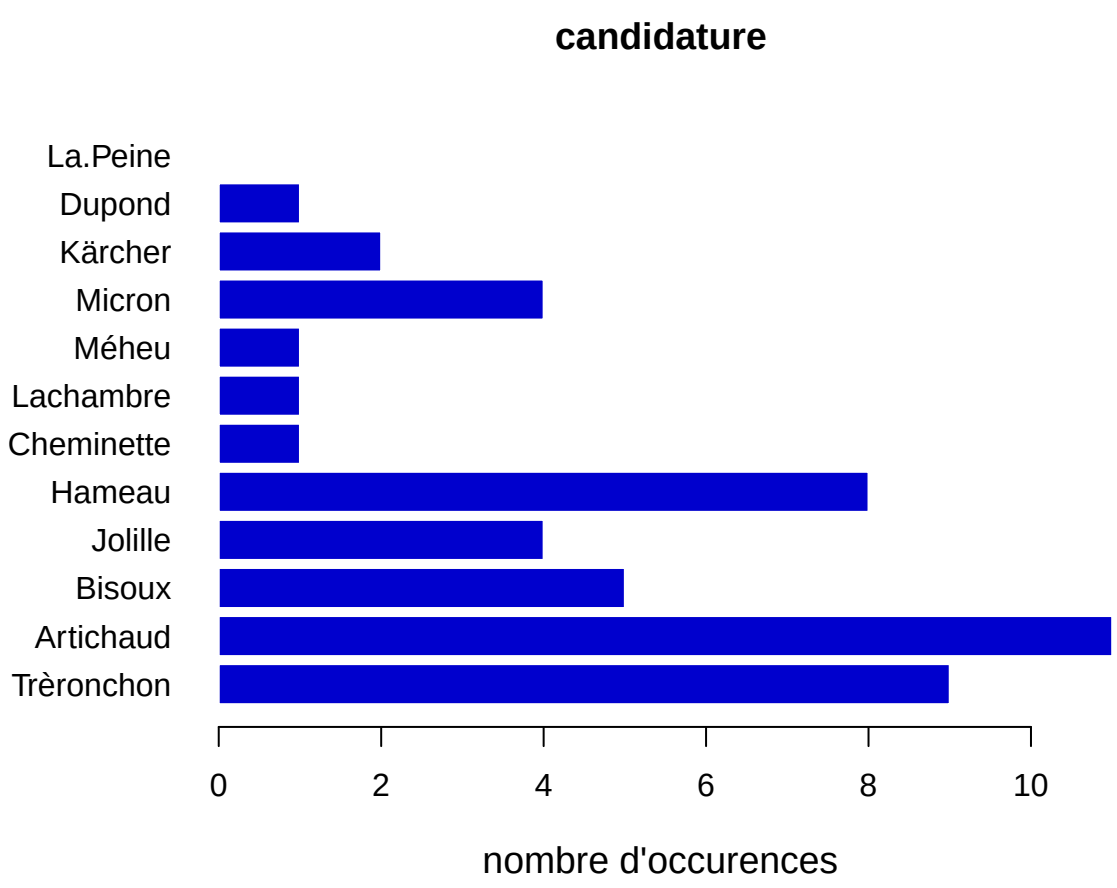
Apparitions des termes "candidature" et "fiscalité"

Nous allons réalisé la même étude qu'avec les candidats. Tout d'abord, nous allons observer le nombre de fois que ces deux mots-clés sont utilisés dans les discours des candidats.

Occurence du mot "candidature" dans les discours de chaque candidat

```
par(mar=c(5,7,4,2))
barplot( as.matrix(lexique["candidature",1:12]),
names.arg = colnames(lexique)[1:12],
col="blue3",
border="white",
main="candidature",
```

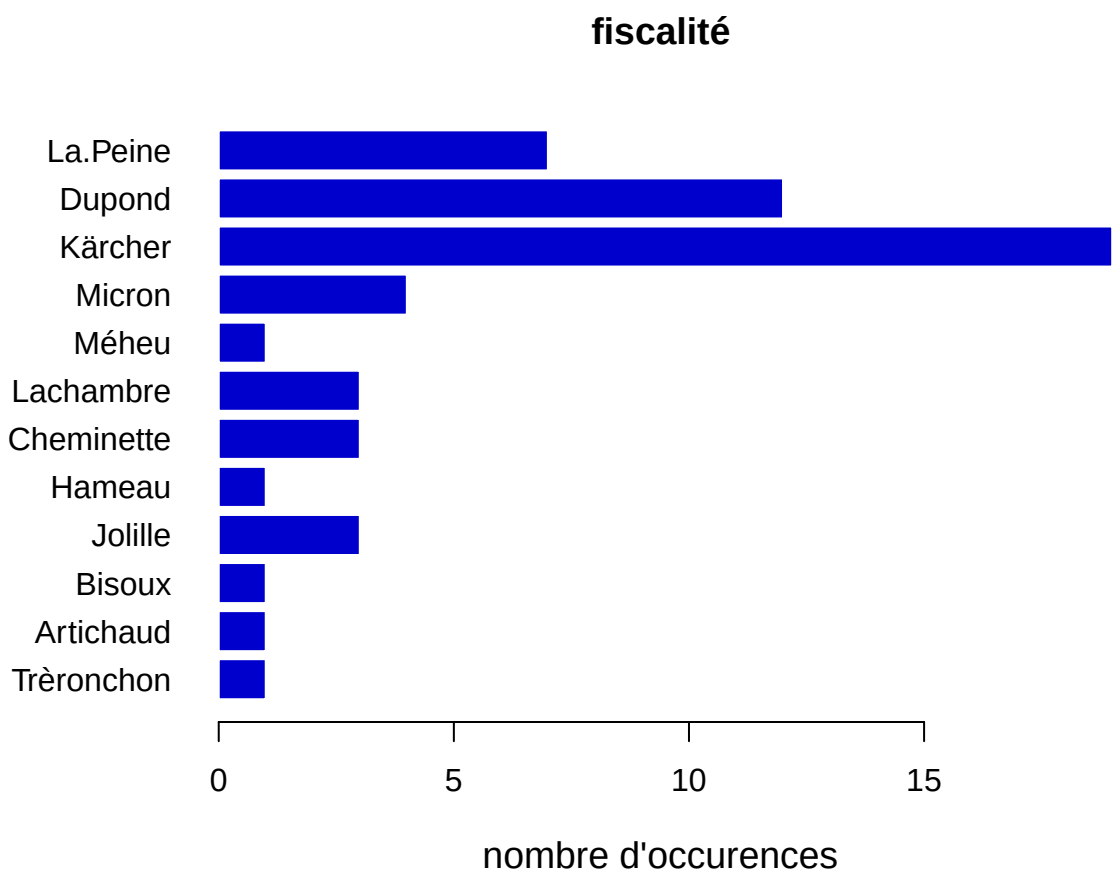
```
horiz=T,
las=1,
xlab="nombre d'occurences",
cex.lab=1.2 )
```



Nous pouvons remarquer que le mot "candidature" est utilisé 11 fois par Artichaud, 9 fois par Tréronchon et 8 fois par Hameau. Nous pouvons aussi noter qu'il n'est utlisé qu'une seule fois par Dupond, Méheu, Lachambre et Cheminette.

Occurence du mot "fiscalité" dans les discours de chaque candidat

```
par(mar=c(5,7,4,2))
barplot( as.matrix(lexique["fiscalité",1:12]),
names.arg = colnames(lexique)[1:12],
col="blue3",
border="white",
main="fiscalité",
horiz=T,
las=1,
xlab="nombre d'occurences",
cex.lab=1.2 )
```



Nous pouvons remarquer que le mot “fiscalité” est utilisé 19 fois par Kärcher, 12 fois par Dupond et 7 fois par La.Peine. Nous pouvons aussi noter qu’il n’est utlisé qu’une seule fois par Méheu, Hameau, Bisoux, Artichaud et Trèronchon.

Comme pour les candidats, afin de bien représenter l’utilisation des deux mots-clés et d’avoir des données permettant des compararaisons pertinentes, nous pouvons aussi étudier non pas la quantité, mais le taux d’occurence de ces deux mots-clés dans les discours des candidats.

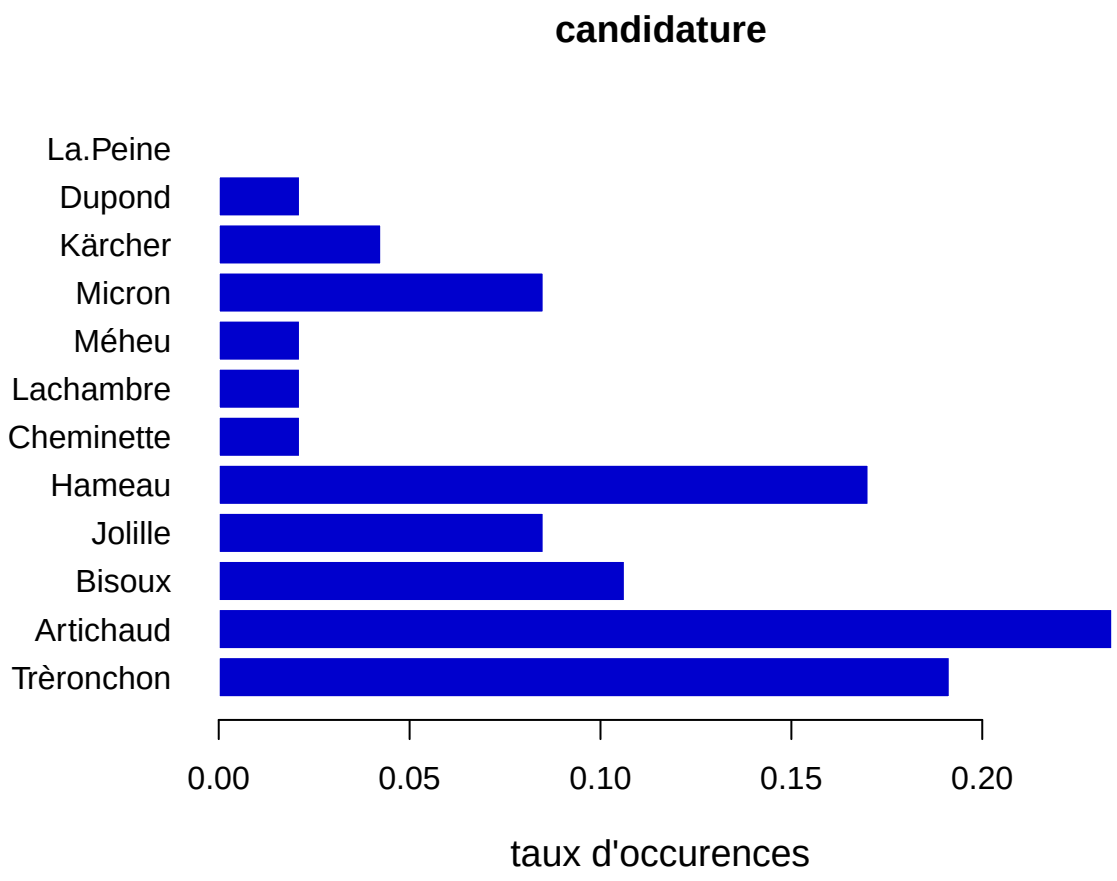
Taux d’occurences du mot-clé “candidature”

```
profil_candidature <- lexique["candidature",1:12]/sum(lexique["candidature",1:12])
profil_candidature
```

```
##             Trèronchon Artichaud   Bisoux    Jolille    Hameau Cheminette
## candidature  0.1914894 0.2340426 0.106383 0.08510638 0.1702128  0.0212766
##             Lachambre     Méheu     Micron    Kärcher    Dupond La.Peine
## candidature 0.0212766 0.0212766 0.08510638 0.04255319 0.0212766        0
```

```
par(mar=c(5,7,4,2))
barplot( as.matrix(profil_candidature),
names.arg = colnames(profil_candidature)[1:12],
col="blue3",
border="white",
main="candidature",
```

```
horiz=T,
las=1,
xlab="taux d'occurences",
cex.lab=1.2 )
```



Nous pouvons remarquer que le mot “candidature” est utilisé à 23% par Artichaud, 19% par Tréronchon et 17% par Hameau. Nous pouvons aussi noter qu’il n’est utlisé qu’à 2% par Dupond, Méheu, Lachambre et Cheminette.

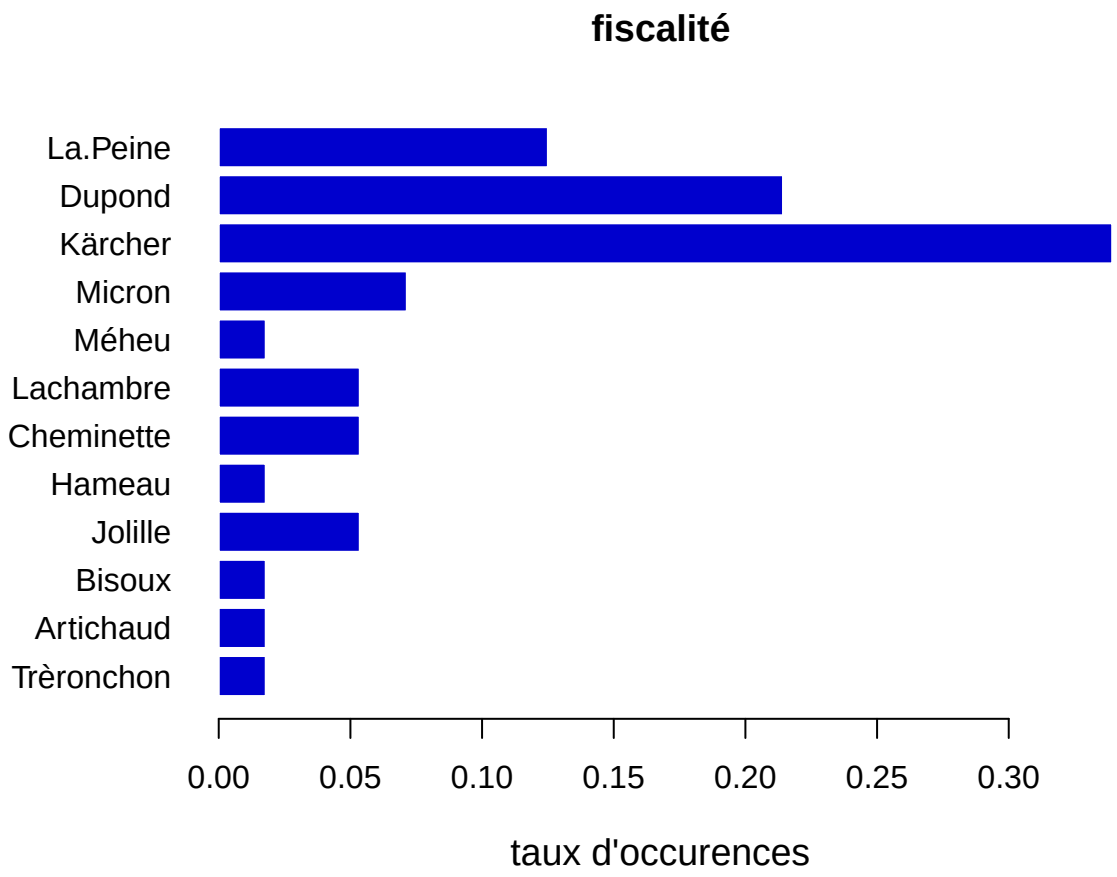
Taux d’occurences du mot-clé “fiscalité”

```
profil_fiscalité <- lexique["fiscalité",1:12]/sum(lexique["fiscalité",1:12])
profil_fiscalité
```

```
##           Trèronchon  Artichaud     Bisoux    Jolille     Hameau Cheminette
## fiscalité 0.01785714 0.01785714 0.01785714 0.05357143 0.01785714 0.05357143
##            Lachambre      Méheu     Micron   Kärcher    Dupond La.Peine
## fiscalité 0.05357143 0.01785714 0.07142857 0.3392857 0.2142857    0.125
```

```
par(mar=c(5,7,4,2))
barplot( as.matrix(profil_fiscalité),
names.arg = colnames(profil_fiscalité)[1:12],
col="blue3",
border="white",
```

```
main="fiscalité",
horiz=T,
las=1,
xlab="taux d'occurences",
cex.lab=1.2 )
```

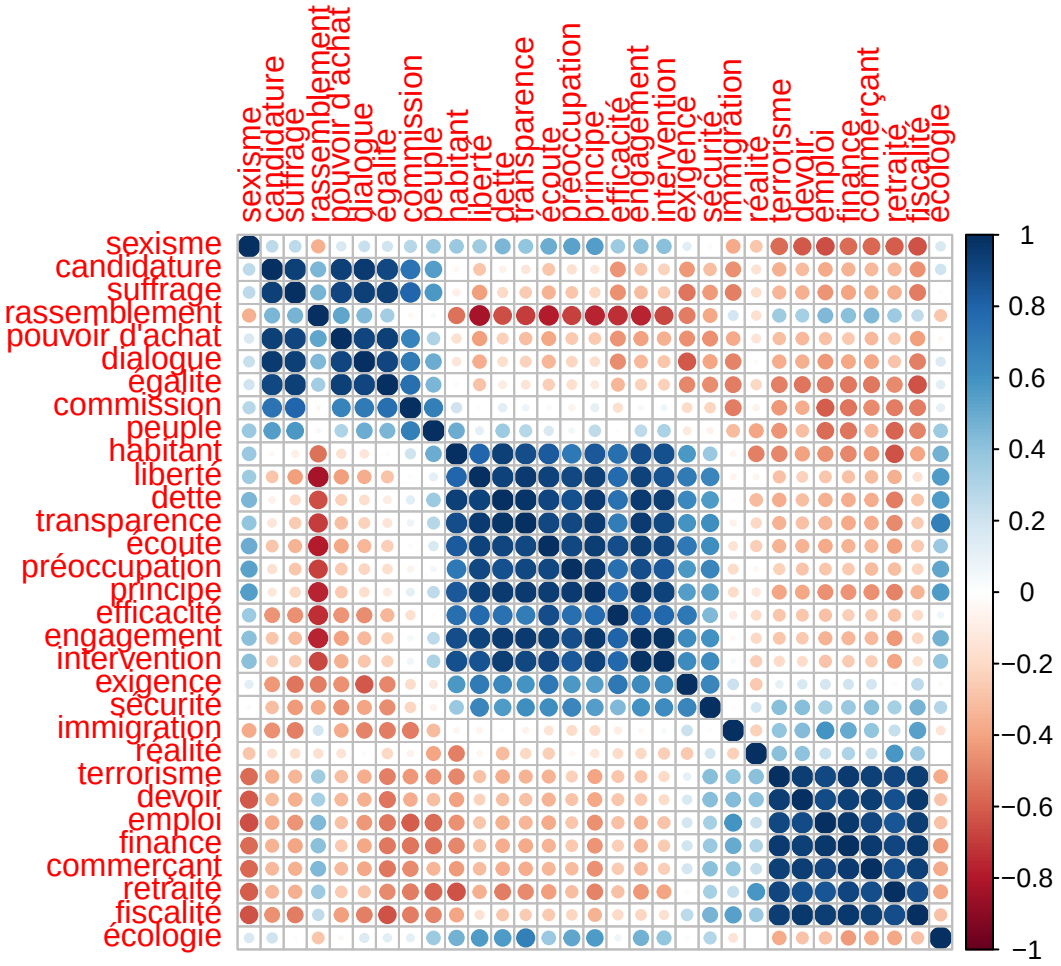


Nous pouvons remarquer que le mot “fiscalité” est utilisé à 34% par Kärcher, 21% par Dupond et 13% par La.Peine. Nous pouvons aussi noter qu’il n’est utlisé qu’à 2% par Méheu, Hameau, Bisoux, Artichaud et Trèronchon.

Matrice de corrélations

Finalement, nous pouvons nous interesser, comme avec les candidats, à la matrice de corrélation, afin de voir si ces deux mots-clés sont autant utilisés chez chaque candidat.

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



Nous pouvons donc remarqué un fait plutôt intéressant : d'après la matrice de corrélation, ces deux mots-clés ont tendance à ne pas être utilisés de la même manière chez chaque candidat.

Détermination des attirances et répulsions entre modalités

Réalisation d'une AFC

Afin de déterminer les attirances et les répulsions entre les différentes modalités, nous allons dans un premier réaliser une AFC. On réalise une AFC, et non pas un ACP, car nos variables sont qualitatives.

```
lex.afc <- CA( lexique,
graph=FALSE, row.sup=c(31))
```

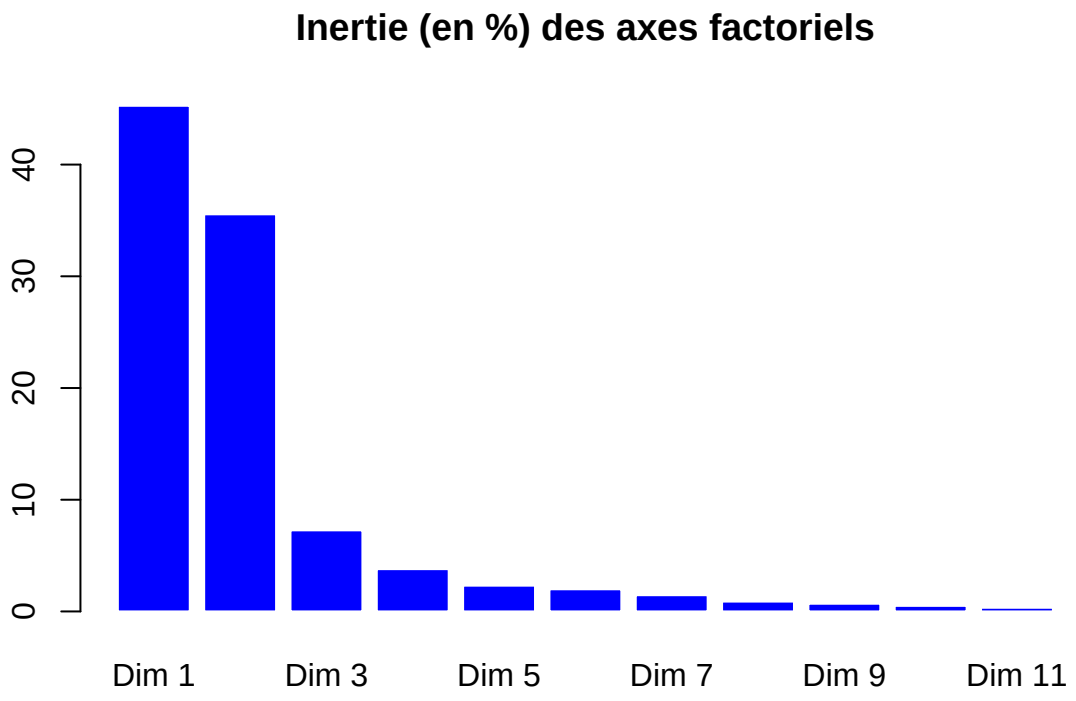
Choix des axes factoriels

```
CramerV( lexique, conf.level =0.99999 )
```

```
##  Cramer V    lwr.ci    upr.ci
## 0.2697954 0.2023465 0.3304305
```

Le coefficient de Cramer est calculé, et un intervalle de confiance à 99% est construit. Cet intervalle ne contenant pas 0, on peut conclure, au seuil de significativité à 0.0001%, que le jeu de donnée contient au moins deux variables dépendantes.

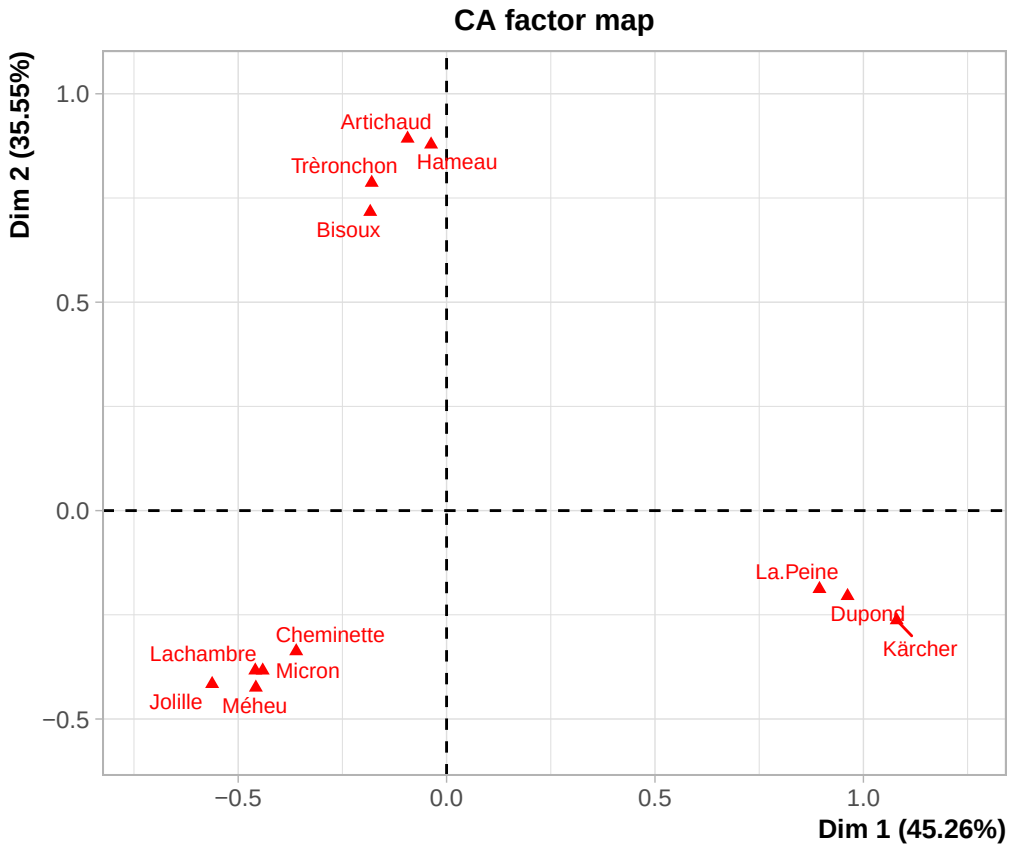
```r
barplot ( lex.afc$eig[,2],
names=paste("Dim",1:length(lex.afc$eig[,2])),
main="Inertie (en %) des axes factoriels",
col="blue",
border="white" )
```



Nous allons donc, par le critère du coude, choisir les deux premiers axes factoriels.

Projection des modalités

```r
plot.CA(lex.afc,
axes = c(1,2),
invisible=c("row.sup", "row"),
cex=0.7)
```



En projettant les modalités, nous remarquons l'apparition de 3 groupes de candidats distincts.

Contributions des modalités aux axes factoriels

Avant de se pencher sur la suite, il peut être intéressant d'être conscient de la contribution des candidats aux axes factoriels.

```
head(sort(lex.afc$col$contrib[,1], decreasing=TRUE), 10)
```

```
##    Kärcher     Dupond   La.Peine    Jolille     Micron  Lachambre      Méheu
## 34.8220359 23.0803687 12.0829827  9.9309051  6.7030174  6.0041110  3.9072260
## Cheminette Trèronchon     Bisoux
##  2.0090178  0.7502421  0.4724323
```

```
head(sort(lex.afc$col$contrib[,2], decreasing=TRUE), 10)
```

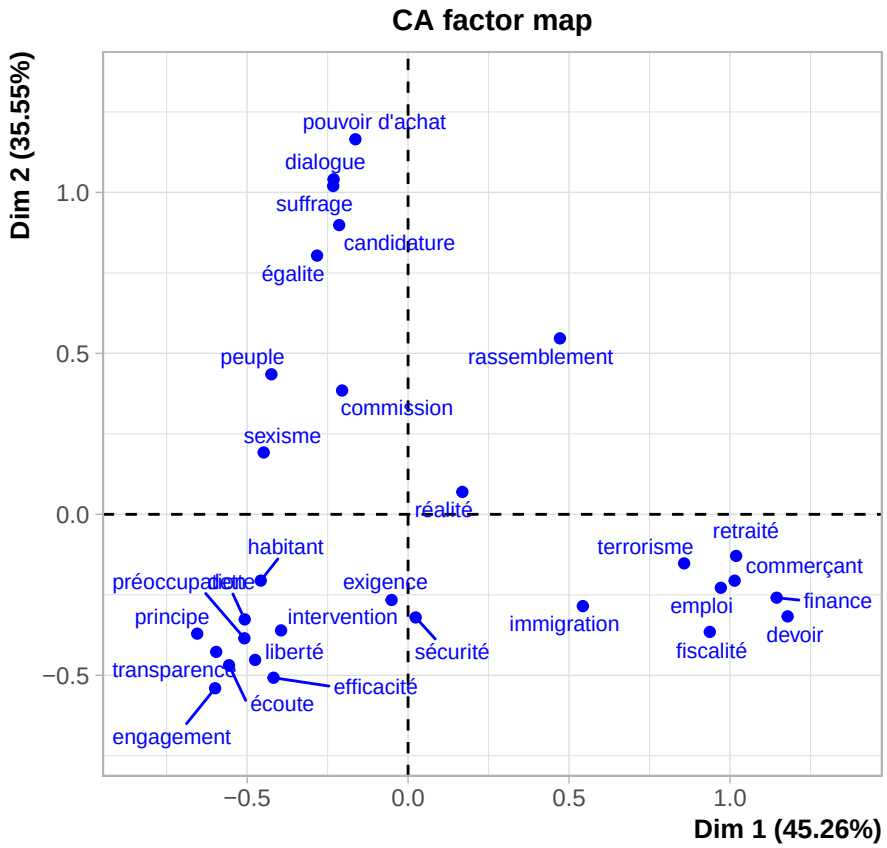
```
##  Artichaud Trèronchon     Hameau     Bisoux    Jolille     Micron  Lachambre
##  24.610118  18.311343  18.002584   9.212026   6.913359   6.452375   5.335972
##      Méheu    Kärcher Cheminette
##   4.283613   2.629495   2.237065
```

Nous pouvons remarquer que Kärcher et Dupond sont les candidats qui contribuent le plus au premier axe, et qu'Artichaud et Tèronchon sont ceux qui contribuent le plus au deuxième axe.

Projection des attributs

```
plot.CA(lex.afc,
axes = c(1,2),
```

```
invisible=c("col.sup", "col", "row.sup"),
cex=0.7)
```



En projettant les modalités, nous remarquons aussi vaguement l'apparition de 3 groupes, qui se trouvent environ aux mêmes endroits que les 3 groupes de candidats identifiés.

Contributions des attributs aux axes factoriels

Nous allons aussi, avant de se pencher sur la suite, nous interesser à la contribution des mots-clés aux axes factoriels.

```
head(sort(lex.afc$row$contrib[,1], decreasing=TRUE), 10)
```

```
##       finance      retraité    commerçant     fiscalité        devoir
##     10.895867      9.335994      8.727323      8.343564      8.252613
##        emploi    terrorisme  transparence      principe rassemblement
##      7.685976      4.989997      4.578394      4.512434      4.228258
```

```
head(sort(lex.afc$row$contrib[,2], decreasing=TRUE), 10)
```

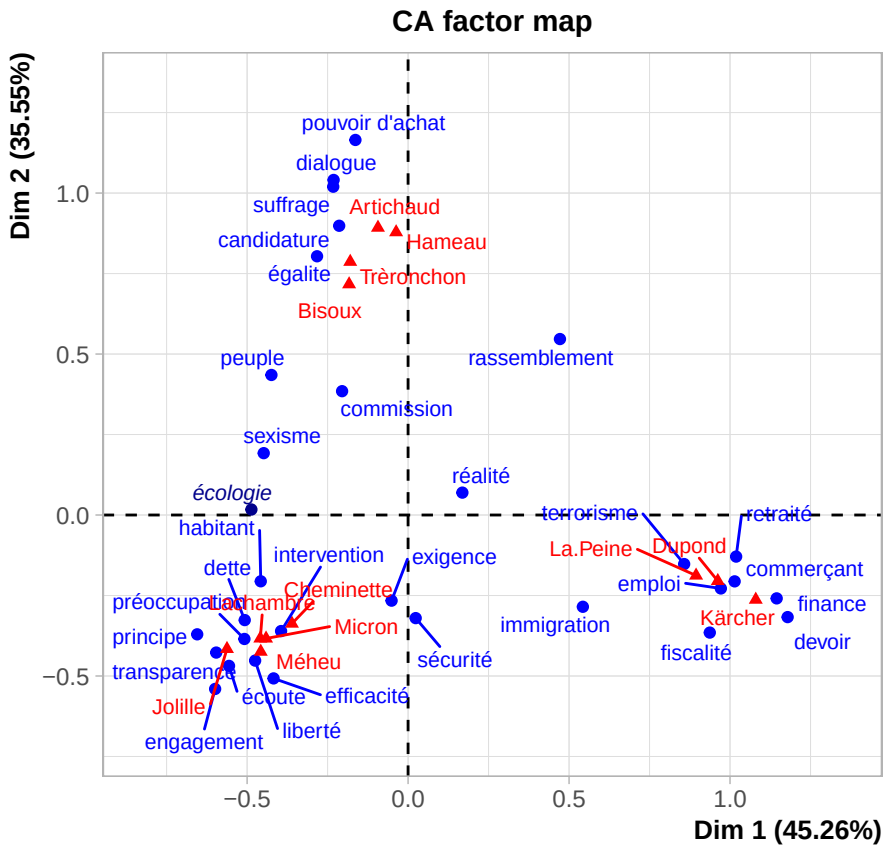
```
## pouvoir d'achat        suffrage        dialogue     candidature   rassemblement
##       17.004735       15.051852       11.934412        8.195884        7.228072
##         égalite      efficacité          écoute      engagement    transparence
##        6.138961        3.616779        3.510377        3.089396        2.997127
```

Nous pouvons remarquer que finance et retraité sont les mots-clés qui contribuent le plus au premier axe, et pouvoir d'achat et suffrage sont ceux qui contribuent le plus au deuxième axe.

Attirances et répulsions entre modalités

En superposant les projections, on peut ainsi étudier les attirances et les répulsions entre les différentes modalités.

```
plot.CA(lex.afc,
axes = c(1,2),
cex=0.7)
```

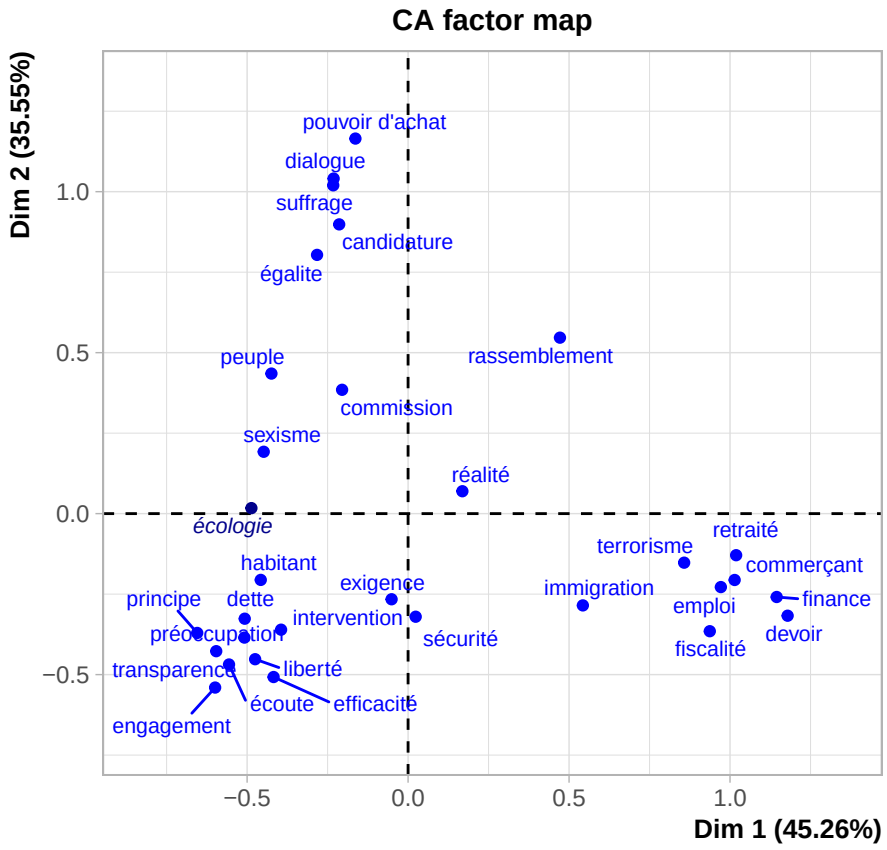


Par exemple, on peut remarquer que Trèronchon attire égalité, canditature et suffrage, tandis que Kärcher les répulsent.

Etude de la projection du terme “écologie” et les relations qu’il entretient avec les autres modalités

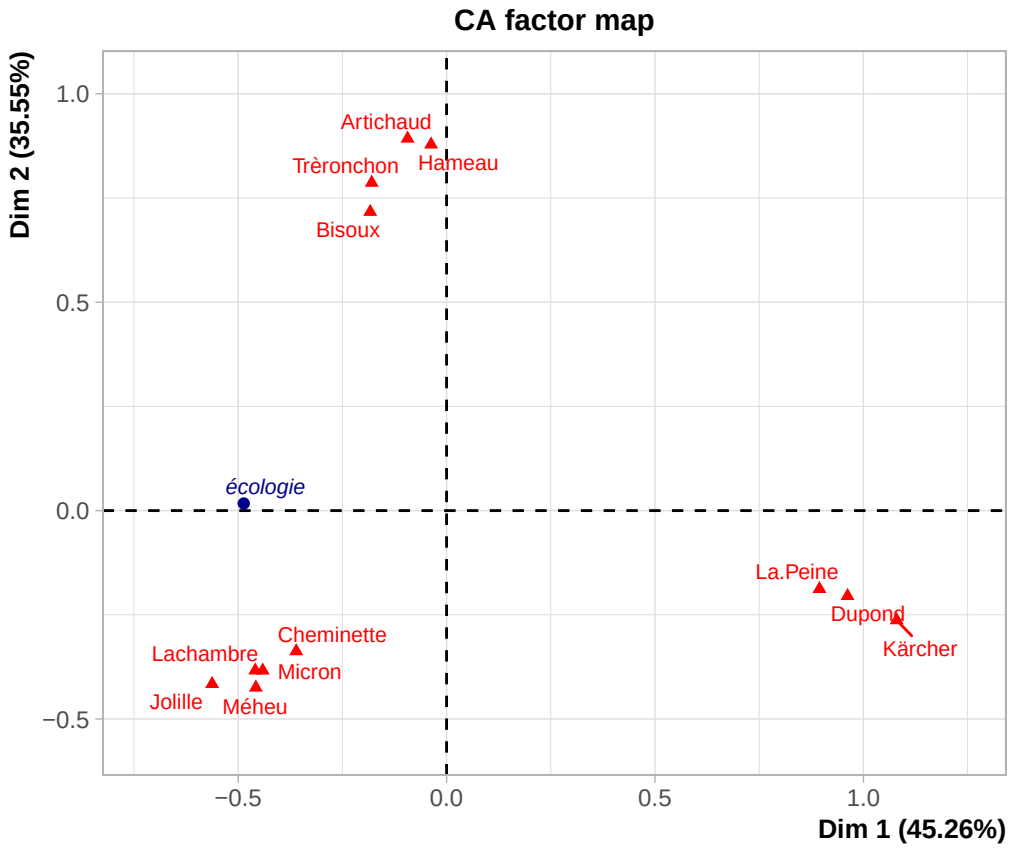
Afin d’étudier ce terme, nous allons le mettre en valeur dans le deux différentes projections.

```
plot.CA(lex.afc,
axes = c(1,2),
invisible=c("col.sup", "col"),
cex=0.7)
```



Sur la projection des attributs, nous pouvons remarquer que les candidats qui utilisent le terme “écologie” ont aussi tendance à utiliser les termes sexisme, habitant, peuple ou encore dette, mais parlent beaucoup moins de pouvoir d’achat, de rassemblement ou de finance.

```
plot.CA(lex.afc,
axes = c(1,2),
invisible=c("row"),
cex=0.7)
```



Sur la projection des modalités, nous pouvons remarquer que le terme “écologie” est d’autant plus attiré par le groupe de candidats en bas à gauche que les deux autres groupes.