Rapport de Textométrie dossier de type 1

Objet : Étudier le type de langage utiliser par les jeunes pour s'exprimer

Realiser par : Darlene VIGNINOU

Nghiêm Đình

PLAN

I-Présentation de l'étude

II-Méthodologie suivie

III-Résultats obtenues

I-Présentation de l'étude

De nos jours, les individus disposent de plusieurs moyens leurs permettant d'exprimer leurs opinions et de faire entendre leurs voix. Le moyen auquel nous nous intéressons ici est un forum "*La voix des jeunes*" destiné uniquement aux jeunes; forum où ils sont libres de partager des articles sur des sujets qui leur tiennent à cœur. Afin de déterminer le type de langage utilisé par les jeunes, nous avons décidé d'étudier les différents types de mots utilisés le plus fréquemment par les jeunes pour exprimer leur opinion dans différents domaines. Pour ce faire nous avons sélectionné plusieurs articles parlant de culture, d'innovation et technologie, de violence, et d'environnement du site "*La voix des jeunes*".

II. Méthodologie

L'analyse de données textuelles est une approche en sciences humaines et sociales pour exploiter, analyser et trouver des sources utiles de connaissances à partir de textes allant des livres aux quotidiens, ou même des lignes de statut publiées sur les sites de réseaux sociaux. Il existe de nombreuses méthodes utilisées pour analyser les données textuelles. Dans le devoir final de textométrie de cette année, nous utiliserons l'analyse quantique des données textuelles, puis utiliserons l'analyse factorielle à correspondant pour approfondir la relation entre les mots. , ou catégorie de textes.

1. Collection des données.
   La collecte des données se fait directement sur le site https://www.voicesofyouth.org/fr/blog. Nous avons sélectionné deux articles par catégorie(), stockés dans un fichier .txt au même format que les textes analysés lors des cours.
   import os
   os.chdir("C://Users//User//Desktop//")
   chemin='Textometrie.txt'
   Le module "Os" fournit un moyen portable d'utiliser les fonctionnalités du système d'exploitation. Dans celui-ci, "os.chdir" nous aide à accéder directement au chemin que nous fournissons. Si nous voulons ouvrir un fichier, nous utilisons l'instruction open().
   import re
   textchaine = re.sub(r"[.:' ;,?=!\n]+()"," ",textebrut)
   Ce module fournit des opérations sur les expressions rationnelles similaires à celles que l'on trouve dans Perl.
   Les motifs, comme les chaînes, à analyser peuvent aussi bien être des chaînes Unicode (str) que des chaînes 8-bits (bytes). Cependant, les chaînes Unicode et 8-bits ne peuvent pas être mélangées : c'est à dire que vous ne pouvez pas analyser une

chaîne Unicode avec un motif 8-bit, et inversement ; de même, lors d'une substitution, la chaîne de remplacement doit être du même type que le motif et la chaîne analysée. Nous appliquons le module **re** pour supprimer des caractères, séparer et joindre des phrases et des paragraphes.

2. Traitement des données.
   Dans la phase de traitement des données textuelles, nous devons comprendre l'objectif du projet. Notre objectif est de découvrir les différences de formulation dans les articles dans différents domaines, afin que nous puissions planifier des articles, des promotions ou du référencement. Tout d'abord, nous devons normaliser les données et filtrer les données inutiles. Pour les données textuelles, ce qui nous donne le plus de

	CULTURE	INNOVATION & TECHNOLOGIE	VIOLENCE & CONFLIT	ENVIRONNEMENT	Total
de	59	45	39	37	180
la	46	23	23	28	120
et	32	19	24	14	89
les	27	21	28	10	86
des	31	23	18	8	80
à	33	23	5	9	70
le	31	8	15	8	62
en	17	11	8	15	51
un	20	12	14	2	48
pour	16	6	14	6	42
que	9	7	13	12	41
nous	1	4	32	2	39
une	11	12	9	4	36
qui	11	6	4	12	33
est	11	9	4	6	30
l	5	11	8	5	29
du	16	5	4	2	27
a	8	8	7	3	26
plus	7	9	5	4	25
sur	3	8	9	5	25

maux de tête, ce sont les mots vides.
Comme nous pouvons le voir dans le tableau ci-dessus, dans les 20 premiers mots les plus utilisés dans les données, la majorité sont des mots vides. Cela rendra notre analyse des données difficile, car il y a beaucoup de données bruitées.
Un outil utile pour gérer les données textuelles est la bibliothèque NLTK (Python). La bibliothèque NLTK fournit des outils tels que l'analyse, la classification, la notation ou encore le prétraitement des données. NLTK est largement utilisé dans le domaine du traitement automatique du langage naturel. NLTK fournit également un outil pour supprimer les mots vides.

```
from nltk.corpus import stopwords
from nltk import word_tokenize
from nltk.tokenize import sent_tokenize
french_stopwords = set(stopwords.words('french'))
filtre_stopfr =  lambda liste: [token for token in liste if token not in
french_stopwords]
```

Cependant, nous avons encore besoin d'une autre étape de filtrage manuel pour supprimer complètement les données indésirables. En fin de compte, nous devrions obtenir une table de données assez propre.

	CULTURE	INNOVATION & TECHNOLOGIE	VIOLENCE & CONFLIT	ENVIRONNEMENT	Total
plus	7	9	5	4	25
comme	7	3	6	2	18
jeunes	4	4	8	0	16
pays	13	1	0	0	14
vie	2	6	1	5	14
canada	13	0	0	0	13
être	2	2	8	0	12
aussi	6	1	4	1	12
peut	0	4	7	0	11
monde	4	3	4	0	11
cela	0	5	2	3	10
tout	2	6	2	0	10
cette	2	4	1	2	9
violence	0	1	8	0	9
internet	1	8	0	0	9
sociaux	2	5	1	0	8
souvent	3	1	3	1	8
social	1	5	0	1	7
théâtre	0	0	7	0	7
autres	1	3	2	1	7

Après avoir filtré les données, nous passerons à l'étape suivante, qui est l'analyse factorielle des correspondances.

3. Analyse factorielle des correspondances.

   L'analyse factorielle des correspondances(AFC)–ou analyse des correspondances pour simplifier -propose une vision synthétique des informations «saillantes» portées par un tableau de contingence. Elle permet de débroussailler rapidement les grands tableaux. Son pouvoir de séduction repose en grande partie sur les représentations graphiques qu'elle propose. Elles nous permettent de situer facilement les similarités (dissimilarités) entre les profils et les attractions (répulsions) entre les modalités. L'AFC est bien une technique factorielle. Les facteurs –les variables latentes –qui en sont issus sont des combinaisons linéaires des points modalités (lignes ou colonnes) exprimés par des profils (lignes ou colonnes).

   Techniquement, on peut voir l'AFC comme une méthode d'ajustement des nuages de profils lignes et colonnes, permettant leur représentation simultanée, c.-à-d. dans le même repère. Nous pouvons aussi la voir sous l'angle de la décomposition orthogonale du X^2 d'écart à l'indépendance. Ce prisme est intéressant parce qu'il montre clairement qu'avant de nous intéresser à la structure des écarts, il est très important de considérer leur importance et se poser la question: y a –t-il de l'information pertinente dans le tableau à analyser? Si le X^2 est trop faible, l'étude subséquente est illusoire.

   Enfin, si l'AFC s'applique en priorité sur les tableaux de contingence (tableau de comptage).Elle est en réalité valable dès lors que les valeurs sont positives, que les notions de marges (sommes en ligne et colonne) et profils (ratios en ligne et colonne) ont un sens dans le tableau que nous analysons.

III- Résultats obtenues

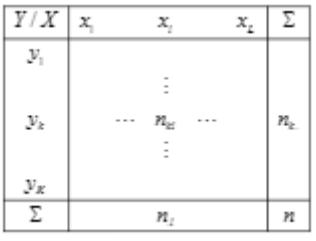
1. Tableau de contingence.

   Un tableau de contingence peut être vu comme un tableau croisé entre deux variables qualitatives. Découvrons s'il existe une relation entre les variables, y a-t-il un contraste dans l'utilisation des mots dans chaque section du site ? Autrement dit, dans

Culture, par exemple, les gens utilisent souvent le mot A plus que le mot B, et quand les gens utilisent le mot A, ils ont tendance à ne jamais utiliser le mot C.
Mais avant d'aborder ces relations, découvrons la répartition des mots dans chaque catégorie. Comme mentionné ci-dessus, notre champ d'étude est les 35 mots les plus utilisés.

X	CULTURE	INNOVATION & TECHNOLOGIE	VIOLENCE & CONFLIT	ENVIRONNEMENT	Total
comme	7	3	6	2	18
jeunes	4	4	8	0	16
pays	13	1	0	0	14
vie	2	6	1	5	14
canada	13	0	0	0	13
Ãªtre	2	2	8	0	12
aussi	6	1	4	1	12
peut	0	4	7	0	11
monde	4	3	4	0	11
cela	0	5	2	3	10
tout	2	6	2	0	10
cette	2	4	1	2	9
violence	0	1	8	0	9
internet	1	8	0	0	9
sociaux	2	5	1	0	8
souvent	3	1	3	1	8
social	1	5	0	1	7
thÃ©Ãctre	0	0	7	0	7
autres	1	3	2	1	7
niveau	4	2	1	0	7
leurs	3	3	0	1	7
fait	3	1	2	0	6
quand	0	1	5	0	6
femmes	0	0	5	1	6
sans	0	3	1	2	6
rÃ©seaux	0	5	1	0	6
si	1	1	0	4	6
faire	1	2	1	2	6
filles	0	0	5	0	5
charte	5	0	0	0	5
car	1	3	1	0	5
dÃ©mocrat	5	0	0	0	5
face	1	1	0	3	5
cours	2	1	0	2	5
Total	96	94	91	35	316

Nous complétons les marges en lignes et en colonnes en faisant la somme. C'est la partie grise de notre panneau supérieur. La gamme totale d'observations est de 316 mots. Par exemple, le mot jeunes apparaît quatre fois dans les catégories Culture, Innovation et Technologie, et huit fois dans Violence et Conflit.

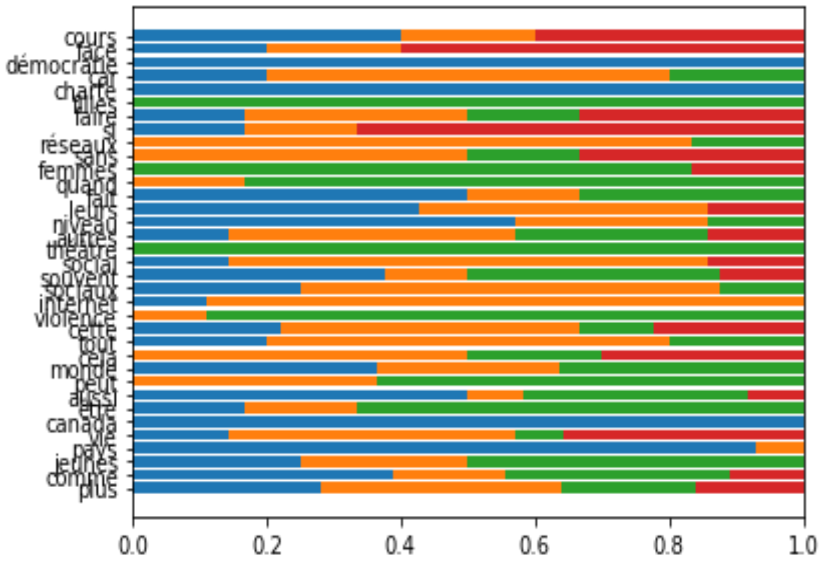


Par exemple on peut calculer la probabilité d'occurrence d'un mot jeunes est de (16/316=0.05) 5%, le mot social est de (7/316=0.022) 2.2%. Pendant ce temps, dans le périmètre des catégories, la culture est la catégorie avec le plus grand nombre d'occurrences de ces 35 mots (96/316), le plus faible est la catégorie de l'environnement (35/316).

2. Analyse des profils lignes.
   Le premier prisme consiste à analyser - les proportions en ligne- du tableau de contingence. Ils sont calculés sous la forme de ratio:

$$P(X = l \mid Y = k) = \frac{n_{kl}}{nk.}$$

Pour le rendre plus intuitif, nous allons créer un graphique de la répartition des mots en quatre catégories.



Dans le graphique, le bleu représente la Culture, le jaune l'Innovation, le vert la Violence et le rouge l'Environnement.

Distance entre profils.

La distance entre profils met en exergue les différences entre proportions en exacerbant celles qui ont trait aux modalités rares. Le même écart est d'autant plus marquant qu'il concerne une modalité peu représentée. On parle de distance du KHI-2:

$$d^2(k, k') = \sum_{l=1}^{L} \frac{1}{\frac{n_{.l}}{n}} \left(\frac{n_{kl}}{n_{k.}} - \frac{n_{k'l}}{n_{k'.}}\right)^2$$

Voyons l'exemple pour préciser les calculs:

	CULTURE	INNOVATION & TECHNOLOGIE	VIOLENCE & CONFLIT	ENVIRONNEMENT	Total
plus	0.28	0.36	0.2	0.16	1
jeunes	0.25	0.25	0.5	0	1
Total	0.304	0.298	0.288	0.11	1

Nous avons appliqué la méthode au dessus:

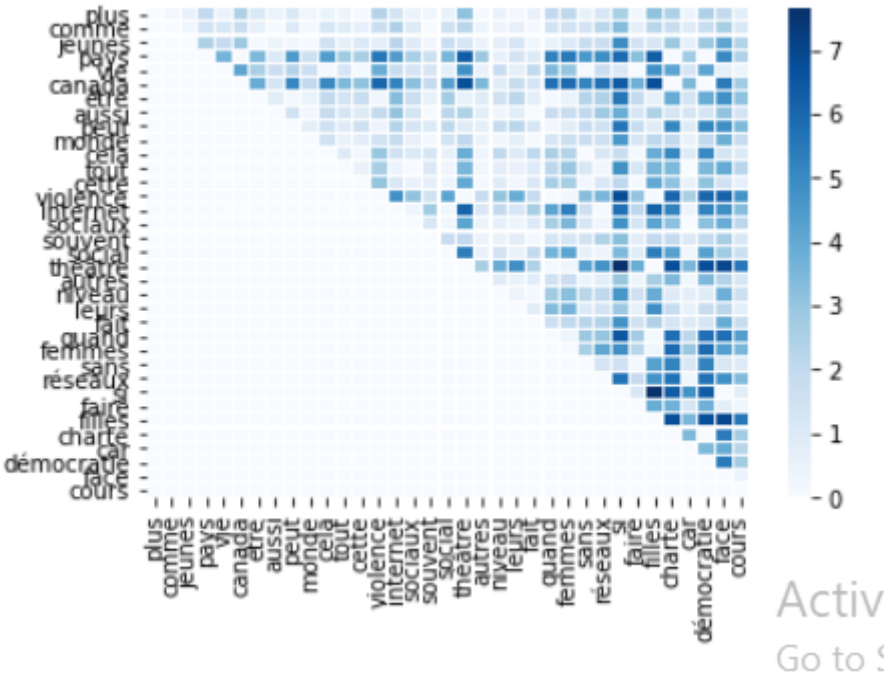
$$d^2(plus, jeunes) = \frac{1}{0.304}(0.28 - 0.25)^2 + \frac{1}{0.298}(0.36 - 0.25)^2 + \frac{1}{0.288}(0.2 - 0.5)^2 + \frac{1}{0.11}(0.16 - 0)^2 = 0.5892$$

En utilisant les bibliothèques numpy et seaborn, nous pouvons observer l'espacement entre les mots. A partir de là, quelques observations peuvent être tirées.

```
In [170]: #distance entre paires de modalités lignes
distPairesLig =numpy.zeros(shape=(prof_lig.shape[0],prof_lig.shape[0]))
#double boucle
for i in range(prof_lig.shape[0]-1):
    for j in range(i+1,prof_lig.shape[0]):
        distPairesLig[i,j] =numpy.sum((prof_lig[i,:]-prof_lig[j,:])**2/prof_marg
        #distPairesLig[j,i] = distPairesLig[i,j]
```

```
In [172]: #affichage sous forme de heatmap
import seaborn as sns
sns.heatmap(distPairesLig,vmin=0,vmax=numpy.max(distPairesLig),linewidth=0.1,cma
```

Nous avons le graphique:



On constate que la distance entre « pays » et « vie » est plus courte que la distance entre « pays » et « filles ». C'est-à-dire que « pays » et « vie » représentent une structure de choix plus proche.

Pour l'analyse des grands tableaux, lorsque le nombre de modalités lignes est élevé, comparer les profils et inspecter les distances par paires deviennent rapidement inextricables. L'AFC a pour objectif de produire un repère factoriel qui rend compte

du positionnement relatif des modalités à l'aide de graphiques «nuages de points», en fournissant des indicateurs permettant d'apprécier la fidélité de la représentation par rapport à la matrice des distances initiales.

Distance à l'origine et l'inertie

Le profil marginal correspond à la structure de choix dans distinction de CSP. Il représente le «profil moyen» c.-à-d. la moyenne pondérée des profils lignes. L'inertie traduit la quantité d'information portée par une modalité. Elle est définie par le produit entre le poids de la modalité et sa distance à l'origine.

In [173]:
```
#distance à l'origine
distoLig =numpy.apply_along_axis(arr=prof_lig,axis=1,func1d=
                                 lambda x:numpy.sum((x
                                                     -prof_marg_lig)
                                                    **2/prof_marg_lig))
#affichage
print(pd.DataFrame(distoLig,index=texto2.index))
#poids des lignes
tot_lig=numpy.sum(texto2.values, axis=1)
print(tot_lig)
poidsLig =tot_lig/numpy.sum(tot_lig)
```

```
                0
plus     0.063776
comme    0.088495
jeunes   0.283967
pays     1.855374
vie      0.853953
canada   2.291667
```

In [174]:
```
#inertie des lignes
inertieLig =distoLig *poidsLig
#affichage
print(pd.DataFrame(numpy.transpose([distoLig,poidsLig,inertieLig]),
                   columns=['Disto2','Poids','Inertie'],index=texto2.index))
```

```
            Disto2     Poids   Inertie
plus      0.063776  0.079114  0.005046
comme     0.088495  0.056962  0.005041
jeunes    0.283967  0.050633  0.014378
pays      1.855374  0.044304  0.082200
vie       0.853953  0.044304  0.037833
canada    2.291667  0.041139  0.094277
```

On peut voir que « plus » à une configuration qui n'est pas trop différente (de l'original), donc même si le poids est plus grand que « paye » ou « vie », la performance sur les données n'est pas aussi claire que « pays » » ou « vie ».

Somme des inerties. La somme des inerties –l'inertie totale -représente la quantité d'information disponible dans les données. C'est un indicateur fondamental. Chaque facteur produit par l'AFC en représente une partie.

In [175]:
```
#total inertie
tot_InertieLig =numpy.sum(inertieLig)
print(tot_InertieLig)
```

```
0.9356211075771296
```

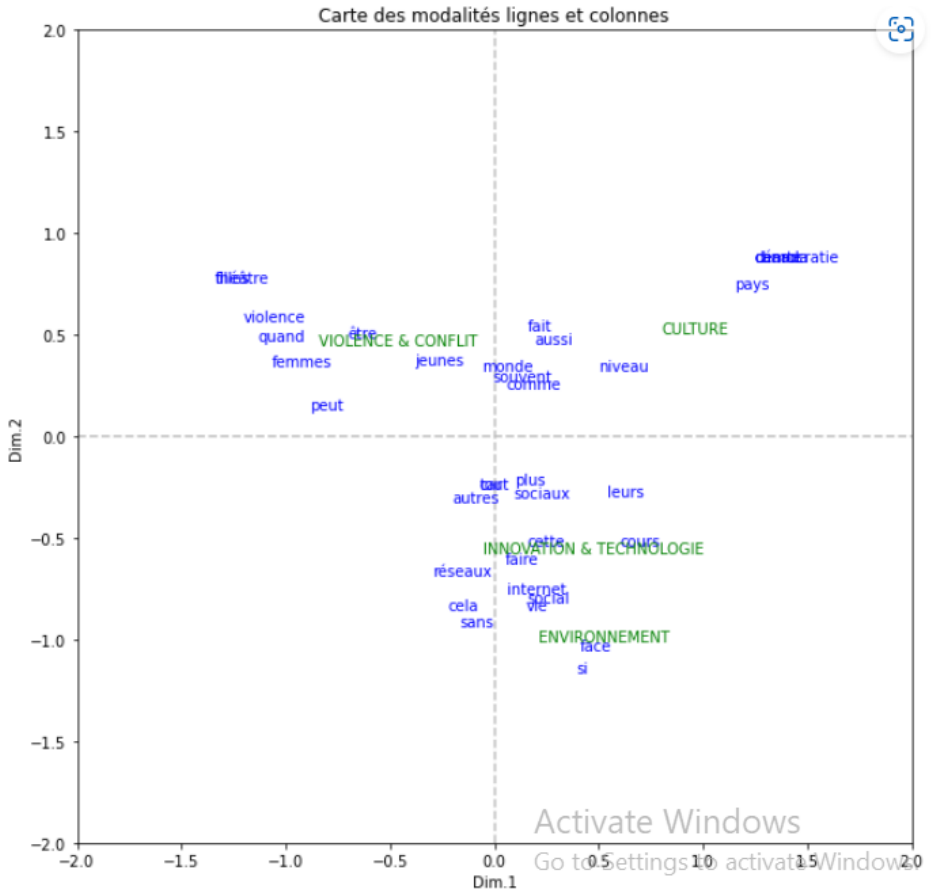
L’analyse factorielle des correspondances.
L’AFC a pour mission de produire une successions de facteurs–des axes factoriels, combinaisons linéaires des profils -qui permettent de positionner les modalités de la manière la plus dispersée possible.Les facteurs sont deux à deux orthogonaux parce que le q eme facteur est élaboré à partir de la partie résiduelle des (q-1) qui le précèdent.De ce point de vue, l’AFC se présente comme une ACP où les individus sont les modalités lignes, décrit par les profils lignes, et pondérés par les effectifs totaux des lignes.

Calculs sous Python.Nous réalisons une AFC avec le package «fanalysis» que nous présenterons plus en détail plus loin. Nous affichons les valeurs propres.

In [176]:
```
#importation de la librairie
from fanalysis.ca import CA
#lancer les calculs
afc =CA(row_labels=texto2.index,col_labels=texto2.columns)
afc.fit(texto2.values)
#information restituée sur les facteurs
print(afc.eig_)
```

```
[[  0.41330516   0.34486083   0.17745511]
 [ 44.17441619  36.85902646  18.96655735]
 [ 44.17441619  81.03344265 100.        ]]
```

Nous voyons que, grâce à l'analyse CA, à partir de quatre variables, nous avons réduit à trois variables, et la représentation des données est jusqu'à 100 %, et avec seulement les 2 premières dimensions, la représentation des données sont jusqu'à 81 %
En effectuant les étapes ci-dessus avec des colonnes, nous obtiendrons un graphique de nuage de données. À partir de là, cela nous aide à avoir une vision visuelle de la relation entre l'utilisation des mots.



Nous voyons que la dimension 1 représente la relation entre les éléments de la culture et de la violence. Dès lors, ces deux écritures ont tendance à se contredire, notamment avec des mots tels que "démocratie" ou "violence". Cela signifie que dans les articles qui utilisent le mot "démocratie", il y a une tendance à ne pas utiliser le mot "violence" ou "quand". Cependant, on voit que ces deux répertoires ne sont pas trop éloignés, et qu'ils partagent quelques mots communs : "fait", "monde", "aussi". Les deux dossiers Violence et Culture ont une relation opposée avec les 2 autres dossiers, Innovation et Environnement, ceci est montré dans la Dimension 2. La Dimension 2 divise notre graphique en 2 moitiés, la moitié supérieure est pour la Dimension 2. pour Violence et Culture, la moitié inférieure pour l'Innovation et l'Environnement. Cette différence s'explique par le fait que les champs de vocabulaire des deux moitiés du graphe sont différents. Sur le graphique, on voit aussi que les deux items Innovation et Environnement partagent un vocabulaire commun, cela s'explique par le fait que les nouvelles technologies se focalisent de plus en plus sur les questions environnementales ?

IV. Conclusion.

L'analyse de données textuelles est aujourd'hui largement utilisée dans les entreprises, notamment dans le service Marketing, ou dans les recherches sociologiques. Avec un exemple simple comme l'analyse basée sur l'AFC, nous avons vu la relation entre les champs lexicaux. Par d'autres méthodes, on peut observer la tendance à utiliser des mots de chaque classe sociale. À partir de ces études, si elles sont appliquées aux entreprises, nous pouvons proposer des stratégies de marketing et des publications médiatiques appropriées qui peuvent atteindre plus d'utilisateurs. Si nous appliquons à la recherche sociale, nous pouvons observer un changement dans la façon dont les gens utilisent le langage par rapport, disons, à il y a 10, 20 ou 30 ans.