L’analyse textuelle avec R.TeMiS

Application à des annonces tirées d’un site de rencontre en ligne

Milan Bouchet-Valat
(Institut national d’études démographiques)

avec Gilles Bastin
(Sciences Po Grenoble, PACTE)

Plan de la présentation

- Introduction
- Le logiciel R.TeMiS : principes et fonctionnalités
- L’importation de corpus de textes
- Le traitement des textes
- Illustration à partir d’un corpus d’annonces Meetic :
  - Statistiques textuelles élémentaires
  - Analyse factorielle

Introduction

Développement lancé en 2011

Limites des logiciels utilisés par les sociologues :

- Propriétaires
- Fortement ancrés dans des contextes théoriques
- Codage manuel du corpus dans un format spécifique
- Isolement des méthodes statistiques et des logiciels *mainstream*

Risques :

- Enfermement dans un environnement du fait des coûts d'entrée élevés
- Surestimation des capacités de l'environnement d'un point de vue épistémologique et manque de réflexivité

Qu’est-ce que R.TeMiS ?

- **R Te**xt **Mi**ning **S**olution
- Environnement graphique sous R
- Création, manipulation et analyse de corpus de textes ; exportation des résultats
- Réduire l’enfermement dans un environnement et promouvoir une approche ouverte et réflexive

Fonctionnalités

Méthodes classiques de l'école française :
Lebart & Salem, *Statistique textuelle*, 1994.
Garnier & Guérin-Pace, *Appliquer les méthodes de la statistique textuelle*, 2010.

- Statistiques descriptives : tris à plat/croisés, séries temporelles, graphiques
- Statistique lexicale : mesures d'occurrence et de cooccurrence de termes, spécificités, bilan lexical
- Analyse de données textuelles : classification ascendante hiérarchique et analyse factorielle des correspondances
- Prise en charge de nombreux formats de corpus
- Exportation des résultats

R

L'environnement statistique R présente des qualités reconnues :

- sa **robustesse** (les procédures statistiques ont été éprouvées par des communautés d'utilisateurs avertis)
- la **transparence** de son code (l'utilisateur pouvant intervenir à chaque étape de l'analyse en modifiant celui-ci)
- sa **gratuité** (les possibilités de traitement — en termes de taille pour un corpus textuel — ne sont pas limitées par la licence acquise)
- son caractère **multi-plateforme**
- enfin sa nature **collaborative** qui permet d'envisager un développement communautaire du logiciel et son adaptation à des usages initialement non prévus

Un environnement réflexif

- Pas d’analyse tout en un de type « boîte noire » : opérations appliquées séparément, pas de cadre théorique imposé
- Génération de code R visible, éditable et extensible
- Grâce à l’utilisation de l’environnement statistique R les traitements disponibles et applicables au corpus sont très nombreux et ne s’arrêtent qu’aux limites de l’imagination sociologique

Détails techniques

- L'interface graphique est un greffon du R Commander (Fox, 2005)
- La gestion des corpus est fournie par le paquet **tm** développé par Ingo Feinerer (Feinerer, 2008; 2011; Feinerer, Hornik & Meyer, 2008)
- Celui-ci a été complété par d'autres paquets classiques de R comme ca pour la représentation des analyses factorielles des correspondances (Nenadic & Greenacre, 2007)
- Enfin des paquets spécifiques ont été développés pour faciliter l'importation de corpus et la lemmatisation

Importation de corpus

- Quatre types de corpus peuvent être importés dans R.TeMiS :
  - des fichiers de texte brut (au format .txt)
  - des fichiers au format tableur (csv, .xls ou .ods), où les lignes correspondent à des individus et les colonnes à des variables descriptives, plus une variable texte
  - des fichiers au format Alceste
  - des résultats de recherche sur Twitter
  - des corpus de presse (ou Web)

L’importation automatique de corpus Web/presse

- Sources gérées : Factiva, Lexis-Nexis, Europresse
- Méta-données présentes : source, date et heure, auteur, section, zone géographique, thèmes, entreprises couvertes…
- Gain de temps : un clic -> 50/100 textes
- Limitation des erreurs de codage
- Traçabilité des opérations menées sur le corpus : toute modification des variables peut être réalisée en R et donc reproduite
- Extension facile lors de l’arrivée de nouvelles données
- Limitation des effets d’enfermement

Traitement du corpus

Lors de l'importation l'utilisateur définit le niveau de traitement lexical du corpus :

- passage des termes en minuscule
- suppression de la ponctuation
- suppression des nombres
- suppression des mots vides (stopwords)
- lemmatisation manuelle ou automatique (paquet SnowballC : algorithme de Porter)

Traitement du corpus

- Possibilité de découper les textes en paragraphes (considérés chacun comme un « document »)
- Prise en compte des différents formats d'écriture de façon moins arbitraire qu'avec un découpage en segments de longueur uniforme
- S'applique aussi éventuellement aux entretiens
- Par défaut les fichiers tabulés sont découpés en autant de documents qu'ils comportent de lignes

Traitement du corpus

Des manipulations peuvent aussi être réalisées après l'importation :

- Choix d'un sous-ensemble de termes, ou élimination de certains termes
- Élimination de documents à partir de termes ou de variables
- Recodage des variables temporelles
- Ces opérations peuvent s'appliquer à des paragraphes, et/ou aux résultats de l'analyse

Concrètement…

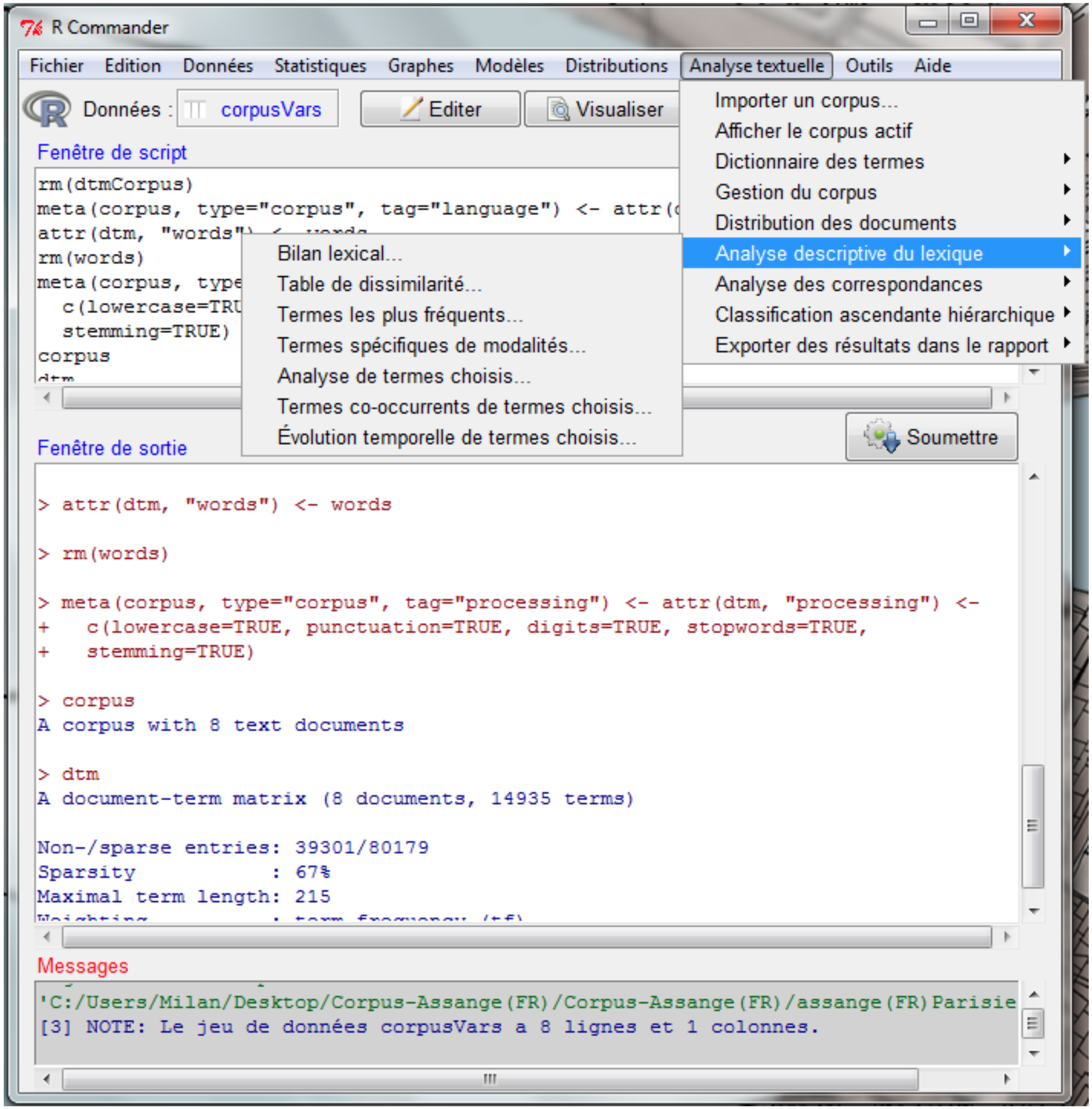


Illustration : corpus Meetic

- Collecté par Marie Bergström et Étienne Ollion par Web scraping des profils
- 21 000 annonces (moitié F/H, réparti par âge)
- Texte de l'annonce :
  - Décrit à la fois l'individu et la personne qu'il recherche
  - 25 mots en moyenne (H/F), 14 sans les mots vides et les nombres. 2 % de vides (conservées comme « annoncevide »).
- Variables associées :
  - sexe, âge, diplôme, revenus, enfants, statut matrimonial, nationalité, appartenance ethnique
  - désir d'enfant, désir de mariage, romantisme…
- Interrogation : quelles sont les variables socio-démographiques qui structurent les différences de vocabulaire dans les annonces ?

L’importation des annonces

- Format tableur :
  - une annonce par ligne
  - 25 variables en colonne
  - le texte du profil est contenu dans une colonne

- Phase de nettoyage semi-manuelle :
  - accents oubliés, mal encodés… → suppression
  - annonces vides → remplies avec un mot-clé « annoncevide »
  - la question des fautes d’orthographe

- Traitements automatisés :
  - Suppression des mots vides
  - Premières analyses sans lemmatisation, puis AFC avec

Termes les plus fréquents par sexe

Femmes	Hommes
annoncevide	*annoncevide*
vie	vie
aime	aime
recherche	si
si	recherche
homme	plus
plus	faire
etre	tout
tout	etre
faire	femme
cherche	bien
personne	cherche
partager	personne
bien	relation
relation	partager
tres	tres
quelqu	homme
femme	peu
humour	simple
peut	quelqu
rencontrer	peut
peu	**comme**
aussi	**bonjour**
vivre	**temps**
simple	**meme**

Termes spécifiques : femmes

	% terme/mod.	% mod./terme	% global	Modalité	Global	Valeur t	Proba.
abstenir	0,15	81	0,09	236	289	Inf	0,00
celui	0,13	78	0,08	195	249	Inf	0,00
ceux	0,12	78	0,08	188	240	Inf	0,00
compagnon	0,10	92	0,06	158	171	Inf	0,00
curieuse	0,12	82	0,07	180	217	Inf	0,00
dynamique	0,24	67	0,18	365	541	Inf	0,00
exigeante	0,05	98	0,02	70	71	Inf	0,00
gaie	0,08	83	0,05	130	155	Inf	0,00
genereuse	0,15	84	0,09	237	281	Inf	0,00
homme	1,00	65	0,75	1525	2323	Inf	0,00
hommes	0,14	84	0,08	209	248	Inf	0,00
independante	0,12	87	0,07	178	204	Inf	0,00
messieurs	0,13	99	0,06	195	196	Inf	0,00
passionnee	0,11	88	0,06	162	184	Inf	0,00
reservee	0,09	93	0,05	145	155	Inf	0,00
rire	0,37	65	0,28	561	855	Inf	0,00
genereux	0,05	25	0,10	80	313	-8,64	0,00
seul	0,05	24	0,11	82	336	-9,41	0,00
sportif	0,04	21	0,09	61	286	-9,81	0,00
celle	0,04	20	0,10	64	309	-10,44	0,00
mesdames	0,00	1	0,04	1	118	-12,06	0,00
compagne	0,01	4	0,06	8	194	-14,01	0,00

Termes spécifiques : sexe

Femmes	
abstenir	**joyeuse**
celui	**active**
ceux	**attentionnee**
compagnon	**maries**
curieuse	inscrite
dynamique	**passez**
exigeante	**franche**
gaie	**merci**
genereuse	**naturelle**
homme	
hommes	
independante	
messieurs	
passionnee	
reservee	
rire	
seule	
souriante	
spontanee	
desolee	

Hommes
mesdemoiselles
divorce
filles
celles
calme
inscrit
toutes
curieux
ouvert
reserve
epicurien
femme
femmes
passionne
genereux
seul
sportif
celle
mesdames
compagne

Termes spécifiques : diplôme

Diplômés du supérieur	Non diplômés du supérieur	
paris	aime	serieux
curieux	*annoncevide*	homme
esprit	gentille	prise
cadre	personne	attentionnee
intelligente	recherche	lol
theatre	serieuse	appelle
curieuse	simple	tres
art	sincere	cherche
	timide	veut
	fidele	fille
	ans	calin
	honnete	merci
	gentil	genereuse
	tete	jai
	calme	jaime
	bonjour	faire
	femme	solitude
	aimerai	amour
	prend	

Table de dissimilarité : diplôme

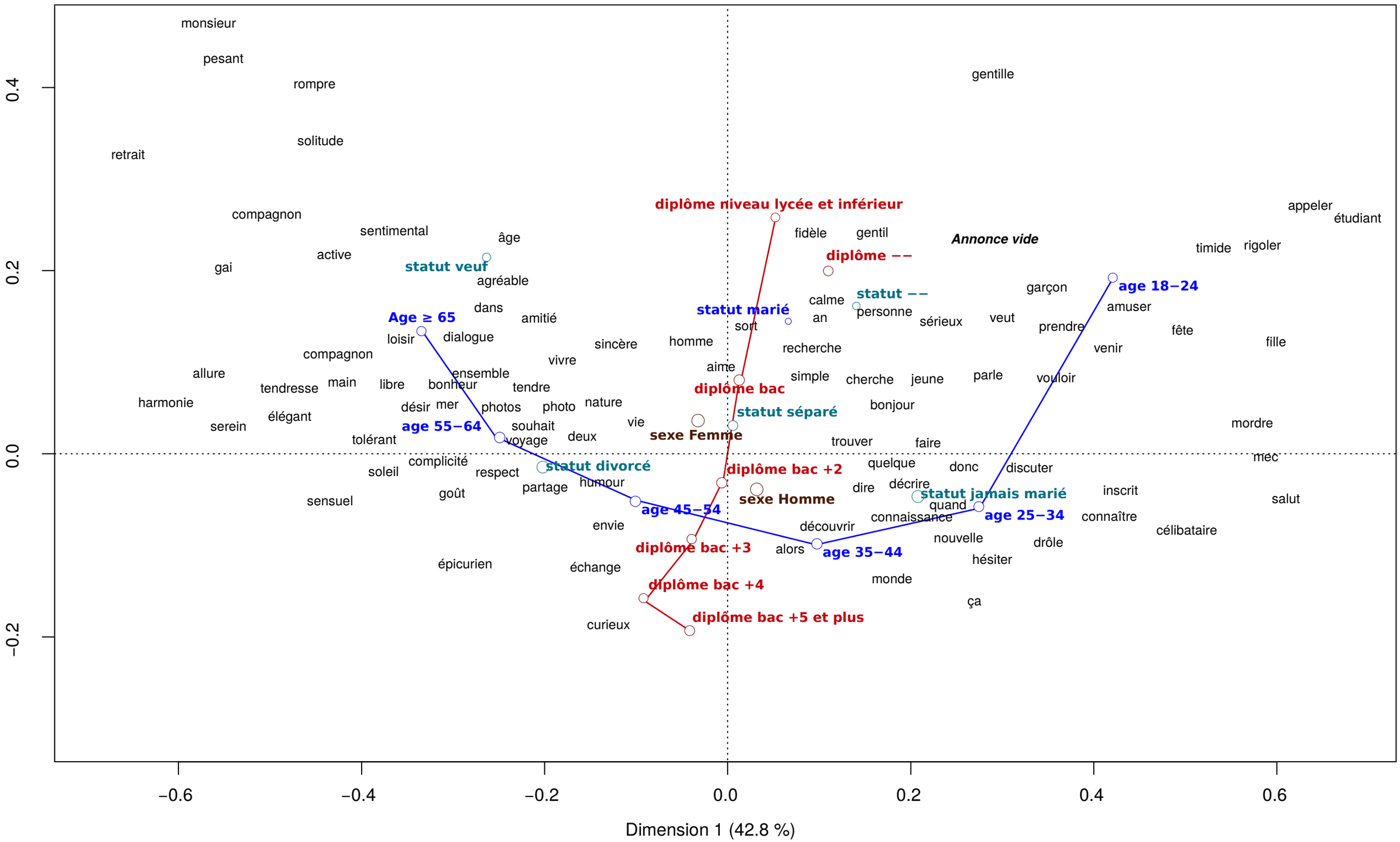
Distance du Khi-2 du vocabulaire utilisé

	--	niveau lycée et inférieur	bac	bac +2	bac +3	bac +4	bac +5 et plus
--	0						
niveau lycée et inférieur	**0,93**	0					
bac	**0,75**	0,86	0				
bac +2	0,79	0,91	0,65	0			
bac +3	0,88	1,00	0,76	0,73	0		
bac +4	1,01	**1,14**	0,91	0,89	0,94	0	
bac +5 et plus	0,92	1,06	0,80	0,77	0,83	0,92	0

Analyse des correspondances

- Deux méthodes :
  - Analyse sur tableau lexical entier
  - Analyse sur tableau lexical agrégé
- Principe de l'analyse sur tableau agrégé :
  - Calcul des occurrences de chaque terme dans chaque modalité des variables retenues
  - Construction d'un tableau avec une ligne par modalité et une colonne par terme
  - Application d'une AFC classique
- Ici, on retient :
  - 4 variables : sexe, âge, statut matrimonial et diplôme (21 modalités)
  - 912 termes (ceux apparaissant dans au moins 0,3 % des documents, soit 64 documents)
- Extraction des radicaux avec l'algorithme de Porter pour éliminer les différences hommes-femmes liées au genre des mots
- Deux premiers axes : 58 % de la variance totale

Analyse des correspondances



Références sur R.TeMiS

- http://rtemis.hypotheses.org
- Bouchet-Valat & Bastin, « RcmdrPlugin.temis, a Graphical Integrated Text Mining Solution in R », *The R Journal*, 5 (1), 2013, p. 188-196. En ligne.
- Bastin & Bouchet-Valat, « Media Corpora, Text Mining and the Sociological Imagination. A Free Software Text Mining Approach to the Framing of Julian Assange in Three News Agencies Using R.TeMiS », *Bulletin de Méthodologie Sociologique*, 122 (1), 2014, p. 5-25. En ligne.