Introduction à la statistique textuelle

Olivier Godechot

Pourquoi compter les mots ?

- Limite des enquêtes par questions fermées
  - La standardisation des réponses est une déformation
  - Alternative : questions ouvertes, entretiens, etc.
- Existence aujourd'hui de sources écrites numérisées de grande taille
  - Internet
  - Base de données de presse (lexis-nexis, europress)
  - Base de données de revue (etc.)
  - Œuvres complètes, etc.
- Limite de nos capacités de lecture analytique
  - Comment analyser des milliers de pages

Qu'est-ce que compter les mots ?

- On travaille sur des chaînes de caractères
  - Les chaînes de caractères semblables = mots semblables
  - Les chaînes de caractères dissemblables = mots différents
    - Ex: Evénement différent de événement
- Pour repérer les mots, il faut repérer les séparateurs de mots :
  - Espace, signes de ponctuation, caractères spéciaux (sauts de paragraphe, de ligne, etc.)
- Choisir ou non de confondre majuscule et minuscule, voyelle avec accents et voyelle sans accents
- Regroupement de mots ou de chaînes de plusieurs mots
  - Segments de deux mots, de trois mots, etc.

Les unités de compte

- Statistiques ➔ corrélation. Deux variables corrélées : Des individus répondent la même chose à plusieurs variables.
- Deux mots corrélés ? Définir un individu ou « unité de contexte élémentaire »
  - Un individu = une personne.
    - Adapté pour réponse courte à question ouverte
    - Pb si réponse longue. Corrélation entre deux mots si énoncés à 15 minutes d'intervalle.
  - Un individu = une phrase
    - Unité de contexte, relativement cohérente
    - Mais phrases de taille différente
  - Un individu = un paragraphe
    - Idem
  - Unité de taille fixe : 10 mots / 12 mots, etc.
    - Taille fixe
    - Pas forcément le même contexte

Comment compter ?

- L'effectif par unité de compte …
- Ou la présence absence par unité ?

Quels mots compter ?

- Les mots les plus fréquents : Les mots outils…(de, le, et, …)
- Les mots les plus singuliers : Les hapax … en général, quelle que soit la taille du texte, 50% des formes graphiques
- Quels mots apportent le plus d'information ? les mots singuliers ou les mots les plus fréquents ?
- Tous les mots : ne rien rater vs beaucoup de mots n'apportent pas d'information. Quantité d'information trop importante
- Une sélection de mots qui ont du sens : information plus précise vs arbitraire de la sélection.

La lemmatisation

- Regroupement de forme lexicale sous un même lemme.
- Regroupement des pluriels et des singuliers, des conjugaisons et des déclinaisons d'un même mot.
- À la main vs automatique (dictionnaire)
- Risque d'erreur important : ex. bus → boire

Les limites de la statistique textuelle

- Le problème de l'homonymie
  - Ex : La petite brise la glace / Le boucher sale la coupe.
- Le problème de la polysémie
- Le problème de la synonymie
- Le problème de la modalisation
  - Formules négatives : travailler et ne pas travailler
  - Nuances
- Arbitraire des modalités de réponse fermées vs arbitraire des catégories du langage
- Le résultat : différence de contenu ou différente manière (sociale) de s'exprimer

Les traitements statistiques sur les mots

- Pas de spécificité statistique – seulement des affinités. On a juste des variables qui sont des mots.
- Les mots surreprésentés. Les phrases typiques
- L'analyse des données
  - Analyse factorielle / Analyse de correspondances multiples
  - Classification ascendante hiérarchique
- La régression
- Les analyses en termes de distance lexicale

Les usages

- Bons résultats : Questions ouvertes dont les réponses sont standardisées et stéréotypées
- Résultats discutables : Entretiens, si textes très différents.
- Les alternatives
  - La lecture simple (utilisation de toute l'information, difficulté à accéder aux phénomènes de nombre, pas d'aspect systématique)
  - Le codage exhaustif (très bons résultats, coût important)
- Les hybridations.
  - Outil possible en complément d'autres techniques
  - Outil élémentaire pour un codage
  - Outil confirmatoire de la lecture simple

Les logiciels disponibles

- Mes macros pour SAS
- Spad-t et spad
- Alceste
- Hyperbase
- Prospero

Problèmes et portée de l'ACM sur variables textuelles

- Sélection des variables. Nombre. Peut-on mélanger des segments de plusieurs mots et des variables d'un seul mot ?
- Problème si les unités servant d'observation comportent peu de vocabulaire.
- Sur quoi porte l'ACM :
  - Variables de fréquences → exclusion des observations qui utilisent aucun des mots
  - Variables de présence:absence → autocorrélation des variables de non utilisation
- Absence de robustesse : Autocorrélation très forte de deux mots (ex. Que & sais-je ? dans ACM sur titres de livres !) D'une population et d'un mot quand elle utilise un seul mot pour réponse : (ex. tranquilité)
- Prendre des individus de grande taille. Changer les individus en classe → analyse factorielle.
- Classification ascendante hiérarchique
  - Sous SAS : Proc Cluster ou Proc Fastclus

Distance lexicale

- Définition des vocabulaires communs entre deux individus.
- Distance = (A – (A∩B))/A) + (B – (A∩B))/B) (méthode Jacquard)
  - = somme de la part des deux vocabulaires distincts. Varie entre 0 (vocabulaire identique) et 200% (vocabulaire différent).
  - Cf. Étienne Brunet, « Peut-on mesurer la distance entre deux textes ? », *Corpus*, *Numero 2 La distance intertextuelle - décembre 2003*, mis en ligne le 15 décembre 2004, http://corpus.revues.org/document.html?id=30
- Possibilité de représentation comme une carte de distance entre individus avec technique de l'étalonnement multidimmensionnel (Multidimmensional scaling = PROC MDS sous SAS).

Les mots spécifiques. Les réponses modales: sous SPAD

- Mots spécifiques
  - Mots les plus sur/sous représentés au terme d'un test de student de différences des proportions entre proportion dans la classe et proportion dans la population d'ensemble.
  - Pour les petits effectifs, un test de student n'est pas forcément adapté. Il peut être judicieux de passer à une loi hypergéométrique.
- Réponses modales
  - Réponses dont la distance modale au vocabulaire de la classe la moins grande
    - Distance du chi-deux :

$$d^2(i,m) = \sum_j \frac{1}{t_{.j}} * \left(\frac{t_{ij}}{t_{i.}} - \frac{c_{mj}}{c_{m.}} \right)^2$$

      Où t_{ij} est le nombre de mots j dans la réponse i, c_{mj} le nombre de mots j dans les réponses de la catégorie m.
    - Technique qui favorise les réponses longues
  - Réponses qui contiennent le plus de formes caractéristiques de la classe.
    - Technique : Moyenne des rangs des formes caractéristiques au sein de la classe m des mots utilisés dans la réponse i.
    - Favorise les réponses courtes

Le fonctionnement des macros Anatext et Vdmot

- Anatext : une macro qui permet de compter les mots (et éventuellement de faire dans certains cas des analyses factorielles)
- Vdmot : une macro qui permet de créer des variables de comptage des formes graphiques

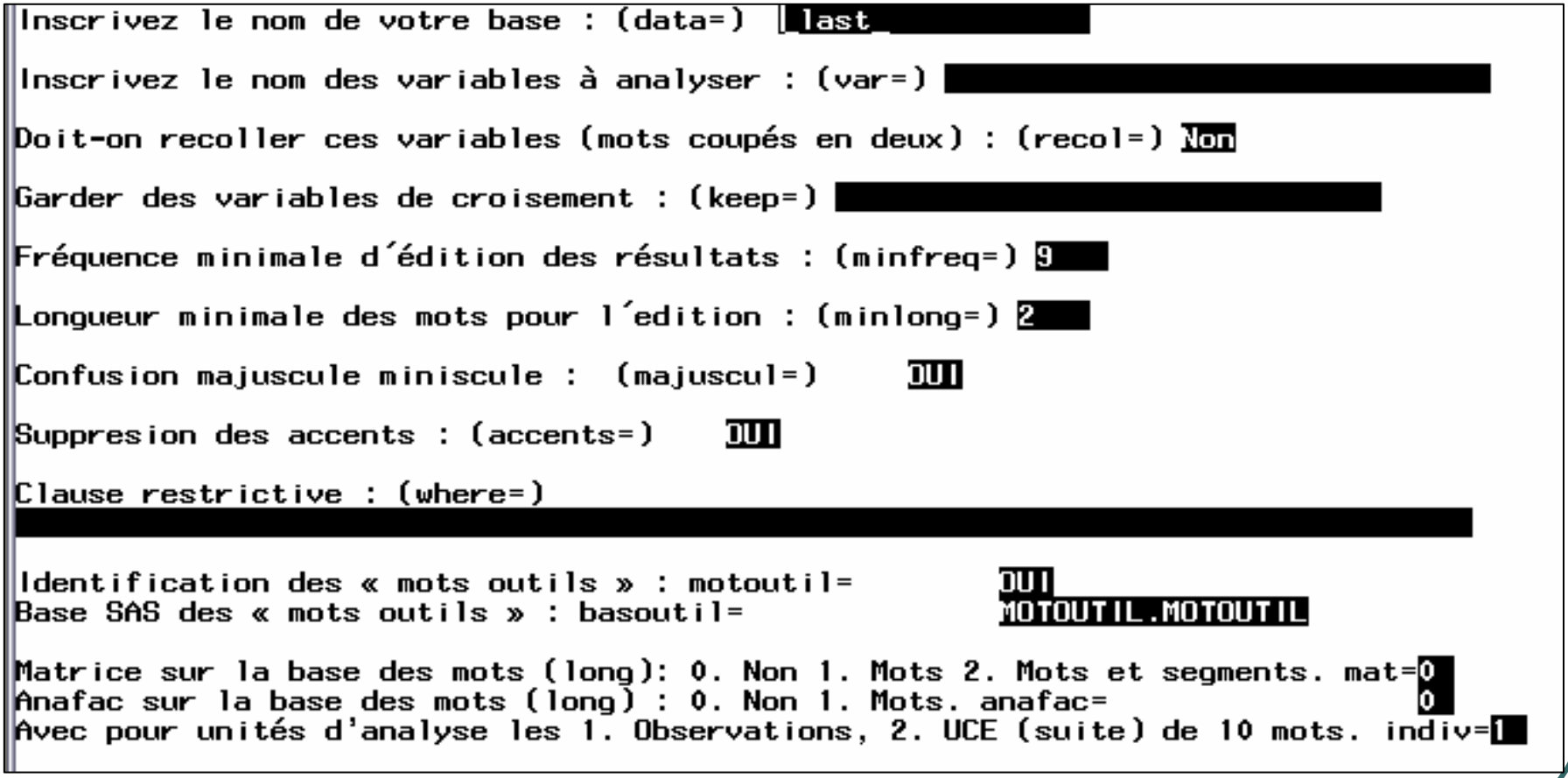
Fonctionnement Anatext

- Soumettre le fichier Anatext.sas
- Soumettre ensuite la commande

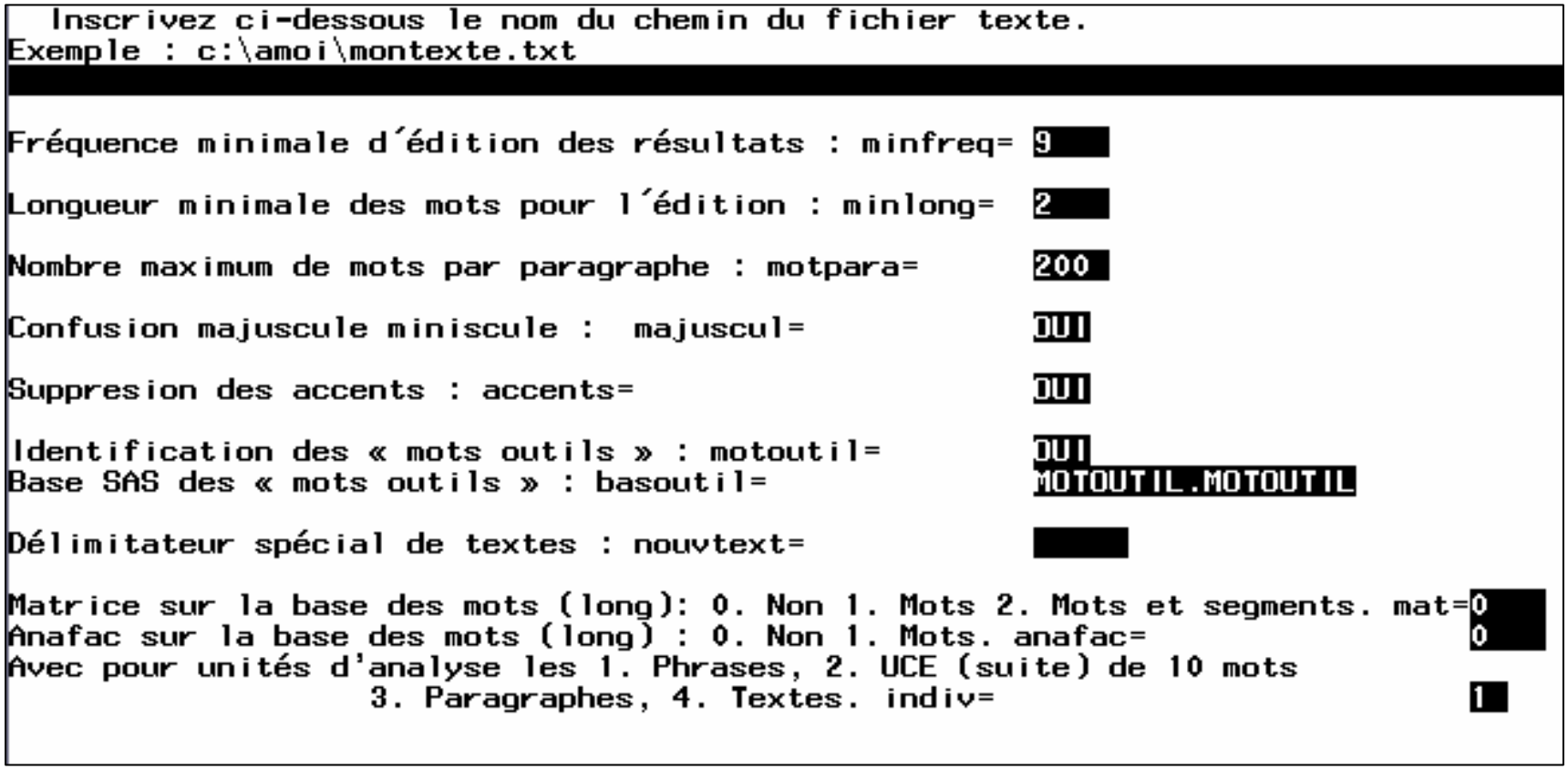
  %anatext;
- Première interface

```
          Bonjour
Type de traitement
  1. Variable textuelle au sein d´une base SAS
  2. Texte

    1

Syntaxe complete :
%Anatext(where, keep, var,fichier,traitem=,data=,
recol=,minlong=,minfreq=,Majuscul=,Accents=,Ecran=,
motpara=,nouvtext=,anafac=,indiv=,motoutil=,basoutil=);

                    Appuyez sur entrée et bonne suite, Olivier Godechot, le 03/02/2000
```

Deuxième interface : cas question ouverte d'un questionnaire



Deuxième interface : cas texte



En mode commande

- Variable textuelle d'une base SAS

```
%Anatext
   (where=, keep=,var=,traitem=1,
   data=_last_,recol=NON,minlong=2,
   minfreq=9,Majuscul=OUI,Accents=OUI,
   Ecran=Oui,mat=0,anafac=0,indiv=1,
   motoutil=OUI,basoutil=MOTOUTIL.MOTOUTIL);
```

- Fichier texte

```
%Anatext
   (fichier=,traitem=2,minlong=2,minfreq=9,Majuscul=
   OUI,
   Accents=OUI,Ecran=Oui,motpara=200,nouvtext=,mat=0
   ,
   anafac=0,indiv=1,motoutil=OUI,
   basoutil=MOTOUTIL.MOTOUTIL);
```

Les sorties

- Sorties imprimés
  - Nombre de mots et formes graphiques
  - Nombre de formes graphiques par taille
  - Nombre de formes graphiques par fréquence
  - Liste et fréquence des formes graphiques (en général et mots outils) (au-dessus du seuil)
  - Liste et fréquence des segments de deux mots
  - Liste et fréquence des segments de trois mots
  - Option : analyse de correspondance multiple sur les variables présence-absence des mots. Graphiques.
- Tables SAS (dans la librairie work)
  - Tableau lexical entier **_mabas1_**
    - Contient mot après mot, les variables conservées, le mot, le segment de deux mots, de trois mots, de dix mots,
  - _mabas2_ tableau de fréquence des mots
  - _mabas3_ tableau de fréquence des segments de deux mots
  - _mabas4_ tableau de fréquence des segments de trois mots
  - Option : _mabas5_ matrice organisée par individu avec variable de fréquence des mots et variable de présence absence des mots.
  - Option : _mabas6_ Résultat de l'analyse de correspondances multiples.
- Log
  - Tous les éléments du programme sont détaillés dans la log.

La macro %vdmot;

- Soumettre le fichier Vdmot.sas
- Soumettre ensuite la commande : **%vdmot;**

```
 Macro de création de variables comptant a/ le nombre d'occurences et
 b/ la présence/absence d une ou plusieurs expressions au sein d une variable texte

Quel est le nom de la table initiale ? (datainit=)
Quel est le nom de la table de sortie ? (dataout=)
Quels sont les noms des variables texte à analyser ? (var=)

Doit-on recoller ces variables (mots coupés en deux) ? (recol=) Non
Quel préfixe pour les variables dichotomiques ? (prefvar=) FG

 Les différentes expressions doivent être séparées par un point.
  Ex : TINTIN. MILOU. LE CAPITAINE HADDOCK. LE CHATEAU DE MOULINSART
 La troncature est signalée par une étoile.
  Ex : *PRENDRE.FAMIL*.*L AMI* (*l ami*==> l amitie, bel ami, quel amiral...)
 Pour coder plusieurs expressions sous la même variable utiliser égal.
  Ex : TRAVAIL=BOULOT=METIER=PROFESSION*. VIVRES=NOURRITURE.
Liste des mots ou des expressions pour faire des variables. (mots=)


Confusion majuscule miniscule :  (majuscul=) OUI
Suppresion des accents : (accents=) OUI
Clause restrictive  (optionnelle)  ? (if ... then do) (iff=)

Anafac sur la base des mots (long) : (anafac=) NON
                                        Appuyer toujours sur Entree !
```

VDmot suite

- La syntaxe détaillée

```
%VDmot (datainit=, dataout=, var=, Prefvar=FG, iff=,
   mots=, Recol=Non, Majuscul=OUI, Accents=OUI,
   Ecran=Oui, anafac=NON);
```

- Les sorties imprimées
  - Comptage exact des mots (Macro pour SAS 9 uniquement)
  - Fréquence d'utilisation (au moins une fois par individu) des mots
  - Option analyse factorielle
- Log
  - Dans la log (le journal), programme SAS de toutes les étapes élémentaires
- Base
  - Base (définie dans dataout) contenant toutes les variables de la table d'origine plus les variables de comptage des mots ainsi que les variables dichotomiques de présence absence.