BNC.query() et BNC.2014.query()

deux outils pour l'étude sociolinguistique de l'anglais contemporain et ses prolongements sociologiques

Guillaume Desagulier[1]

[1]MoDyCo (UMR 7114)
Paris 8, CNRS, Université Paris Nanterre
Institut Universitaire de France
gdesagulier@univ-paris8.fr
https://corpling.hypotheses.org/

AFLA 2020, Université de Cergy-Pontoise, 6 février 2020

1. contexte
2. corpus
3. scripts
4. études de cas
5. discussion

sociolinguistique de corpus

- DEPA Paris 8 > Licence/Master
- sociolinguistique variationniste
- linguistique de corpus quantitative

comparer des corpus : une évidence ?

taille

d'un côté, les corpus sont compilés pour permettre au linguiste de généraliser à partir d'observations.

échantillonnage

de l'autre, les corpus ne sont que des échantillons

l'échantillonnage est spécifique à chaque corpus

conclusion

- comparer des corpus est une pratique courante . . .
- . . .mais qui ne va pas de soi

Brown, LOB, and Kolhapur

`https://corpling.hypotheses.org/284`

Brown (anglais US)

- Francis & Kučera
- compilé en 1964, révisé en 1971, amplifié en 1979

LOB (anglais GB)

- Johansson, Leech & Goodluck
- 1978

Kolhapur (anglais d'Inde)

- Shastri, Patilkulkarni & G.S. Shastri
- 1986

Label	Text category	Brown Corpus	LOB Corpus	Kolhapur Corpus
A	Press: reportage	44	44	44
B	Press: editorial	27	27	27
C	Press: reviews	17	17	17
D	Religion	17	17	17
E	Skills, trades and hobbies	36	38	38
F	Popular lore	48	44	44
G	Belles lettres, biography, essays	75	77	70
H	Miscellaneous (docs, reports, etc.)	30	30	37
J	Learned and scientific writings	80	80	80
K	General fiction	29	29	58
L	Mystery and detective fiction	24	24	24
M	Science fiction	6	6	2
N	Adventure and western fiction	29	29	15
P	Romance and love story	29	29	18
R	Humour	9	9	9
Total		500	500	500

BNC-XML et BNC 2014

(British National Corpus)

	BNC-XML	BNC 2014
année de sortie	1995	2018
période couverte	1960-1993	2012-2016
taille (en tokens)	100M	11.5M
registres	écrit - parlé	parlé
nbre de fichiers	4049	1251

BNC

(British National Corpus)

Bref historique :

- BNC 1.0 : 1re version amorcée en 1991 et sortie en 1995
- BNC World Edition : 2e version sortie en 2001 ; SGML ; étiquetage amélioré par Geoffrey Leech, Roger Garside & Tony McEnery
- BNC-XML : 3e version sortie en 2007 ; XML ; 2 sous-échantillons :
  - BNC Sampler
  - BNCBaby
- 2014 : ultime version sortie en 2018
  - `http://corpora.lancs.ac.uk/bnc2014/`
  - transcriptions orales uniquement

BNC-XML

Un corpus général soi-disant représentatif et équilibré :

- 4 049 fichiers de sous-corpus, env. 100M mots (tokens)
- 8 genres textuels : `ACPROSE`, `FICTION`, `NEWS`, `NONAC`, `OTHERPUB`, `UNPUB`, `CONVRSN`, `OTHERSP`
- 70 sous-genres textuels
  `http://rdues.bcu.ac.uk/bncweb/manual/genres.html`
  - 46 pour l'écrit
  - 24 pour l'oral
- 90 textes écrits - 10 transcriptions de l'oral

BNC-XML

Mais, il s'agit d'un corpus :

- accessible : `https://ota.bodleian.ox.ac.uk/repository/xmlui/`
- normé : en-tête TEI
- annoté : XML
- étiqueté : CLAWS5 (`http://ucrel.lancs.ac.uk/claws5tags.html`)

BNC-XML

en-tête TEI, `A00.xml`

```
<bncDoc xml:id="A00"><teiHeader><fileDesc><titleStmt><title> [ACET factsheets &amp;
newsletters]. Sample containing about 6688 words of miscellanea (domain: social science)
</title><respStmt><resp> Data capture and transcription </resp><name> Oxford University
Press </name> </respStmt></titleStmt><editionStmt><edition>BNC XML Edition, December
2006</edition></editionStmt><extent> 6688 tokens; 6708 w-units; 423 s-units
</extent><publicationStmt><distributor>Distributed under licence by Oxford University
Computing Services on behalf of the BNC Consortium.</distributor><availability> This
material is protected by international copyright laws and may not be copied or
redistributed in any way. Consult the BNC Web Site at http://www.natcorp.ox.ac.uk for full
licencing and distribution conditions.</availability><idno type="bnc">A00</idno><idno
type="old"> AidFct </idno></publicationStmt><sourceDesc><bibl><title> [ACET factsheets
&amp; newsletters]. </title> <imprint n="AIDSCA1"><publisher> Aids Care Education &amp;
Training </publisher> <pubPlace> London </pubPlace> <date value="1991-09"> 1991-09 </date>
</imprint> </bibl></sourceDesc></fileDesc><encodingDesc><tagsDecl><namespace
name=""><tagUsage gi="c" occurs="810"/><tagUsage gi="div" occurs="43"/><tagUsage gi="head"
occurs="45"/><tagUsage gi="hi" occurs="24"/><tagUsage gi="item" occurs="43"/><tagUsage
gi="label" occurs="10"/><tagUsage gi="list" occurs="8"/><tagUsage gi="mw"
occurs="31"/><tagUsage gi="p" occurs="118"/><tagUsage gi="pb" occurs="2"/><tagUsage gi="s"
occurs="423"/><tagUsage gi="w"
occurs="6708"/></namespace></tagsDecl></encodingDesc><profileDesc><creation
date="1991">1991-09 </creation><textClass><catRef targets="WRI ALLTIM3 ALLAVA2 ALLTYP5
WRIAAG0 WRIAD0 WRIASE0 WRIATY2 WRIAUD3 WRIDOM4 WRILEV2 WRIMED3 WRIPP5 WRISAM5 WRISTA2
WRITAS3"/><classCode scheme="DLEE">W nonAc: medicine</classCode><keywords><term> Health
</term><term> Sex </term></keywords></textClass></profileDesc><revisionDesc><change
date="2006-10-21" who="#OUCS">Tag usage updated for BNC-XML</change><change
date="2000-12-13" who="#OUCS">Last check for BNC World first release</change><change
date="2000-09-06" who="#OUCS">Redo tagusage tables</change><change date="2000-09-01"
who="#OUCS">Check all tagcounts</change><change date="2000-06-23" who="#OUCS">Resequenced
s-units and added headers</change><change date="2000-01-21" who="#OUCS">Added date
info</change><change date="2000-01-09" who="#OUCS">Updated all catrefs</change><change
date="2000-01-08" who="#OUCS">Manually updated tagcounts, titlestmt, and title in
source</change><change date="1999-09-13" who="#UCREL">POS codes revised for BNC-2; header
updated</change><change date="1994-11-24" who="#dominic">Initial accession to
corpus</change></revisionDesc></teiHeader>
<wtext type="NONAC"><div level="1" n="1" type="leaflet"><head type="MAIN">
<s n="1"><w c5="NN1" hw="factsheet" pos="SUBST">FACTSHEET </w><w c5="DTQ" hw="what"
pos="PRON">WHAT </w><w c5="VBZ" hw="be" pos="VERB">IS </w><w c5="NN1" hw="aids"
pos="SUBST">AIDS</w><c c5="PUN">?</c></s></head><p>
```

BNC-XML

XML & CLAWS, `A00.xml`, phrase 126

Home care Coordinator, Margaret Gillies, currently has a team of 20 volunteers from a variety of churches providing practical help to a number of clients already referred.

```
<s n="126"><w c5="NN1" hw="home" pos="SUBST">Home </w><w c5="NN1-VVB" hw="care"
pos="SUBST">care </w><w c5="NN1" hw="coordinator" pos="SUBST">Coordinator</w><c c5="PUN">,
</c><w c5="NP0" hw="margaret" pos="SUBST">Margaret </w><w c5="NP0" hw="gillies"
pos="SUBST">Gillies</w><c c5="PUN">, </c><w c5="AV0" hw="currently" pos="ADV">currently
</w><w c5="VHZ" hw="have" pos="VERB">has </w><w c5="AT0" hw="a" pos="ART">a </w><w
c5="NN1" hw="team" pos="SUBST">team </w><w c5="PRF" hw="of" pos="PREP">of </w><w c5="CRD"
hw="20" pos="ADJ">20 </w><w c5="NN2" hw="volunteer" pos="SUBST">volunteers </w><w c5="PRP"
hw="from" pos="PREP">from </w><w c5="AT0" hw="a" pos="ART">a </w><w c5="NN1" hw="variety"
pos="SUBST">variety </w><w c5="PRF" hw="of" pos="PREP">of </w><w c5="NN2" hw="church"
pos="SUBST">churches </w><w c5="VVG" hw="provide" pos="VERB">providing </w><w c5="AJ0"
hw="practical" pos="ADJ">practical </w><w c5="NN1" hw="help" pos="SUBST">help </w><w
c5="PRP" hw="to" pos="PREP">to </w><w c5="AT0" hw="a" pos="ART">a </w><w c5="NN1"
hw="number" pos="SUBST">number </w><w c5="PRF" hw="of" pos="PREP">of </w><w c5="NN2"
hw="client" pos="SUBST">clients </w><w c5="AV0" hw="already" pos="ADV">already </w><w
c5="VVN" hw="refer" pos="VERB">referred</w><c c5="PUN">.</c></s></p><p>
```

BNC-XML

XML & CLAWS, `A00.xml`, un token

```
<w c5="NN2" hw="volunteer" pos="SUBST">volunteers </w>
```

Chaque mot est délimité par une balise d'ouverture (`<w...>`) et de fermeture (`</w>`).
La balise d'ouverture contient :

- un code (`w` pour "word") ;
- une étiquette POS issue du tagset CLAWS5 (`NN2` : nom pluriel)
- le lemme (`hw`, i.e. "head word")
- la valeur de la catégorie associée à l'étiquette (`SUBST`, i.e. `NN2` est un substantif)

BNC-XML

composante démographique : 908 fichiers

Une grande partie de la composante orale du corpus contient des renseignements quant aux locuteurs :

- genre
- catégorie d'âge
- catégorie sociale

Ces renseignements sont dans l'en-tête TEI.

```
<person ageGroup="X" xml:id="PS5VL" role="unspecified" sex="m" soc="UU" dialect="NONE"
educ="X"><persName>js</persName> </person><person ageGroup="X" xml:id="PS5VM"
role="unspecified" sex="m" soc="UU" dialect="NONE" educ="X"><persName>t</persName>
</person>
```

→ exploitable pour la sociolinguistique, mais dans quelle mesure ?

BNC 2014

Voulu comme le successeur du BNC-XML

- projet en cours (objectif : 100M tokens)
- U. de Lancaster & U. de Cambridge
- Tony McEnery et al. (2017). « Compiling and analysing the Spoken British National Corpus 2014. Special issue of ». In : *International Journal of Corpus Linguistics* 22.3

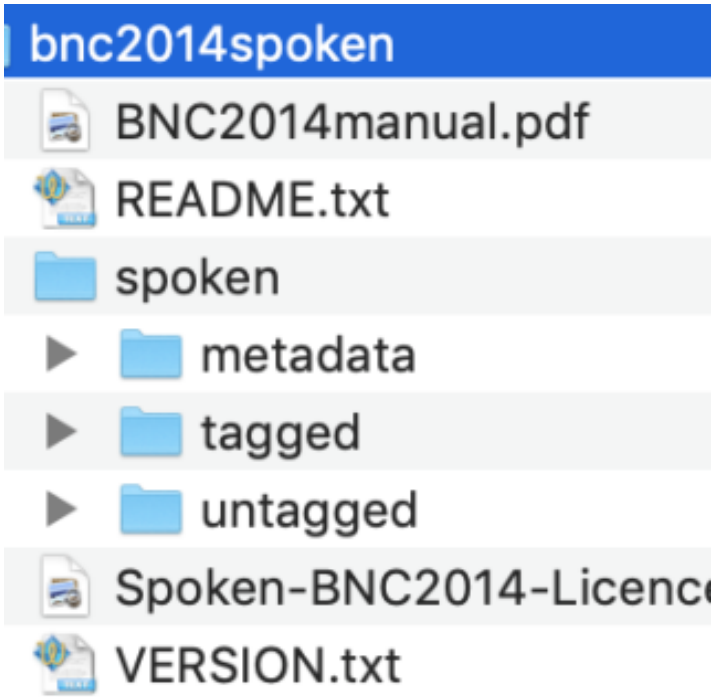
BNC 2014

Un corpus qui a servi de support à plusieurs études sociolinguistiques :

- Vaclav Brezina et al. (2018). *Corpus Approaches to Contemporary British Speech : Sociolinguistic Studies of the Spoken BNC2014*. Routledge
- Jacqueline Laws et al. (2017). « A diachronic corpus-based study into the effects of age and gender on the usage patterns of verb-forming suffixation in spoken British English ». In : *International Journal of Corpus Linguistics* 22.3, p. 375-402
- Tanja Hessner et Ira Gawlitzek (2017). « Totally or slightly different ? : A Spoken BNC2014-based investigation of female and male usage of intensifiers ». In : *International Journal of Corpus Linguistics* 22.3, p. 403-428
- Robert Fuchs (2017). « Do women (still) use more intensifiers than men ? : Recent change in the sociolinguistics of intensifiers in British English ». In : *International Journal of Corpus Linguistics* 22.3, p. 345-374
- Andreea S Calude (2017). « Sociolinguistic variation at the grammatical/discourse level : Demonstrative clefts in spoken British English ». In : *International Journal of Corpus Linguistics* 22.3, p. 429-455.

BNC 2014

structure générale des fichiers



BNC 2014

XML & CLAWS7, `S2A5-tgd.xml`, énoncé 1

```
<text id="S2A5">
<u n="1" who="S0024" trans="nonoverlap" whoConfidence="high">
<w pos="AT1" lemma="a" class="ART" usas="Z5">an</w>
<w pos="NNT1" lemma="hour" class="SUBST" usas="T1:3">hour</w>
<w pos="RRR" lemma="later" class="ADV" usas="T4">later</w>
<pause dur="short"/>
<w pos="VV0" lemma="hope" class="VERB" usas="X2:6">hope</w>
<w pos="PPHS1" lemma="she" class="PRON" usas="Z8">she</w>
<w pos="VVZ" lemma="stay" class="VERB" usas="M8">stays</w>
<w pos="RP" lemma="down" class="ADV" usas="Z5">down</w>
<pause dur="short"/>
<w pos="RG" lemma="rather" class="ADV" usas="A13:5">rather</w>
<w pos="JJ" lemma="late" class="ADJ" usas="T4">late</w>
</u>
```

BNC 2014

XML & CLAWS7, `S2A5-tgd.xml`, un token

```
<w pos="NNT1" lemma="hour" class="SUBST" usas="T1:3">hour</w>
```

balises d'ouverture et de fermeture d'énoncés : `<u ...> </u>`

pour chaque mot :

- un code (`w` pour "word") ;
- une étiquette POS issue du tagset CLAWS7 (`NNT1` : nom singulier à référence temporelle)
- le lemme (`lemma`)
- la valeur de la catégorie associée à l'étiquette (`class="SUBST"`, i.e. `NNT1` est un substantif)
- une étiquette sémantique issue de l'annotateur USAS (`class="T1:3"`,)

Pourquoi créer de nouveaux outils de requête ?

Il existe déjà :

- CQPweb
  - `https://cqpweb.lancs.ac.uk/`
- BNCweb
  - `http://bncweb.lancs.ac.uk/`
- BNC@BYU
  - `https://www.english-corpora.org/bnc/`
- etc.

Pourquoi créer de nouveaux outils de requête ?

`BNC.query()` & `BNC.2014.query()` répondent à des besoins spécifiques :

- croiser la distribution des variables linguistiques et les métadonnées des informateurs dans une optique sociolinguistique
- fournir une base de code R customisable & réutilisable
- intégrer un module de quantification et de statistiques

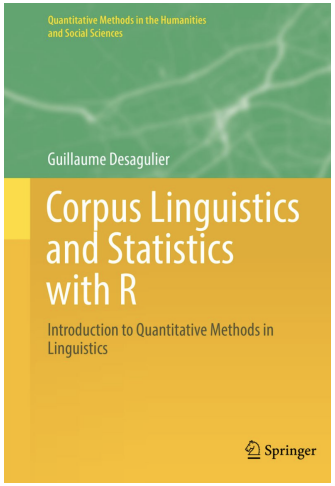
À un niveau plus local (cursus de licence anglais LLCE) :

- initier les étudiants à la programmation sous R
- développer une autonomie relative (création et adaptation d'outils)
- initier aux dangers de la quantification naïve

BNC.query() & BNC.2014.query()

Une architecture algorithmique commune

- packages :
  - `gsubfn` (traitements de chaînes de charactères)
  - `stringi` (traitements de chaînes de charactères)
  - `plyr` (manipulation de data frames)
  - `ggplot2` (graphiques)
- intégration de modules présentés séparément dans Desagulier (2017)



BNC.query() & BNC.2014.query()

lancer les scripts

BNC.query()

```
> rm(list=ls(all=TRUE))
> source("https://www.nakala.fr/nakala/data/11280/7607a789")
> BNC.query()
```

BNC.2014.query()

```
> rm(list=ls(all=TRUE))
> source("https://www.nakala.fr/nakala/data/11280/de1d18d2")
> BNC.2014.query()
```

BNC.2014.query() pas à pas

formuler une requête

```
> rm(list=ls(all=TRUE))
> source("https://www.nakala.fr/nakala/data/11280/1f3b7b18")
> BNC.2014.query()
BNC.2014.query() - version 0.2 - Jan 7, 2019 - CC-BY-NC-ND 4.0
Type your expression (without quotes, punctuation, leading or trailing spaces): (hi|hello)
```

variable(s)	expression régulière (R)
"hello"	`hello`
"hello" et "hi"	`(hello|hi)`
"good morning"	`good morning`
"good morning", "good afternoon", "good evening"	`good (morning|afternoon|evening)`
"I think" et "you see"	`(I think|you see)`

BNC.2014.query() pas à pas

tagset : http://ucrel.lancs.ac.uk/claws7tags.html

RRQ	wh- general adverb (where, when, why, how)
RRQV	wh-ever general adverb (wherever, whenever)
RRR	comparative general adverb (e.g. better, longer)
RRT	superlative general adverb (e.g. best, longest)
RT	quasi-nominal adverb of time (e.g. now, tomorrow)
TO	infinitive marker (to)
UH	interjection (e.g. oh, yes, um)
VB0	be, base form (finite i.e. imperative, subjunctive)
VBDR	were
VBDZ	was

BNC.2014.query() pas à pas

tagset : http://ucrel.lancs.ac.uk/claws7tags.html

```
> rm(list=ls(all=TRUE))
> source("https://www.nakala.fr/nakala/data/11280/1f3b7b18")
> BNC.2014.query()
BNC.2014.query() - version 0.2 - Jan 7, 2019 - CC-BY-NC-ND 4.0
Type your expression (without quotes, punctuation, leading or trailing spaces): (hi|hello)
Enter the CLAWS7 part-of-speech tag of the word you are looking for (if you don't know it, enter \w+): UH
```

BNC.2014.query() pas à pas

conversion en XML

```
> rm(list=ls(all=TRUE))
> source("https://www.nakala.fr/nakala/data/11280/1f3b7b18")
> BNC.2014.query()
BNC.2014.query() - version 0.2 - Jan 7, 2019 - CC-BY-NC-ND 4.0
Type your expression (without quotes, punctuation, leading or trailing spaces): (hi|hello)
Enter the CLAWS7 part-of-speech tag of the word you are looking for (if you don't know it, enter \w+): UH
the expression you are looking for is <w pos="UH" lemma="\w+" class="\w+" usas="\w+">(hi|hello) ?</w>
```

BNC.2014.query() pas à pas

localisation du corpus

```
> rm(list=ls(all=TRUE))
> source("https://www.nakala.fr/nakala/data/11280/1f3b7b18")
> BNC.2014.query()
BNC.2014.query() - version 0.2 - Jan 7, 2019 - CC-BY-NC-ND 4.0
Type your expression (without quotes, punctuation, leading or trailing spaces): (hi|hello)
Enter the CLAWS7 part-of-speech tag of the word you are looking for (if you don't know it, enter \w+): UH
the expression you are looking for is <w pos="UH" lemma="\w+" class="\w+" usas="\w+">(hi|hello) ?</w>
Indicate the path to the folder that contains the BNC 2014 corpus files and press ENTER: /bnc2014spoken
                                                                                        /spoken/tagged
```

BNC.2014.query() pas à pas

sélection du (sous-)corpus

```
> source("https://www.nakala.fr/nakala/data/11280/1f3b7b18")
> BNC.2014.query()
BNC.2014.query() - version 0.2 - Jan 7, 2019 - CC-BY-NC-ND 4.0
Type your expression (without quotes, punctuation, leading or trailing spaces): (hi|hello)
Enter the CLAWS7 part-of-speech tag of the word you are looking for (if you don't know it, enter \w+): UH
the expression you are looking for is <w pos="UH" lemma="\w+" class="\w+" usas="\w+">(hi|hello) ?</w>
Indicate the path to the folder that contains the BNC 2014 corpus files and press ENTER: /bnc2014spoken/sp
oken/tagged

Do you want to run a query over the whole spoken BNC 2014 or just a sample?

1: the whole corpus (1251 files)
2: a random sample (100 files)

Selection: 1
```

BNC.2014.query() pas à pas

vérification du chargement du corpus

```
Indicate the path to the folder that contains the BNC 2014 corpus files and press ENTER: /bnc2014spoken/sp
oken/tagged

Do you want to run a query over the whole spoken BNC 2014 or just a sample?

1: the whole corpus (1251 files)
2: a random sample (100 files)

Selection: 1
 [1] "/bnc2014spoken/spoken/tagged/S23A-tgd.xml" "/bnc2014spoken/spoken/tagged/S24A-tgd.xml"
 [3] "/bnc2014spoken/spoken/tagged/S24D-tgd.xml" "/bnc2014spoken/spoken/tagged/S24E-tgd.xml"
 [5] "/bnc2014spoken/spoken/tagged/S263-tgd.xml" "/bnc2014spoken/spoken/tagged/S26N-tgd.xml"
 [7] "/bnc2014spoken/spoken/tagged/S27D-tgd.xml" "/bnc2014spoken/spoken/tagged/S28F-tgd.xml"
 [9] "/bnc2014spoken/spoken/tagged/S29Q-tgd.xml" "/bnc2014spoken/spoken/tagged/S29X-tgd.xml"
there are 1251 spoken files! Press ENTER to run the query (this may take a while)...
```

BNC.2014.query() pas à pas

c'est parti !

```
ged

Do you want to run a query over the whole spoken BNC 2014 or just a sample?

1: the whole corpus (1251 files)
2: a random sample (100 files)

Selection: 1
 [1] "/bnc2014spoken/spoken/tagged/S23A-tgd.xml" "/bnc2014spoken/spoken/tagged/S24A-tgd.xml"
 [3] "/bnc2014spoken/spoken/tagged/S24D-tgd.xml" "/bnc2014spoken/spoken/tagged/S24E-tgd.xml"
 [5] "/bnc2014spoken/spoken/tagged/S263-tgd.xml" "/bnc2014spoken/spoken/tagged/S26N-tgd.xml"
 [7] "/bnc2014spoken/spoken/tagged/S27D-tgd.xml" "/bnc2014spoken/spoken/tagged/S28F-tgd.xml"
 [9] "/bnc2014spoken/spoken/tagged/S29Q-tgd.xml" "/bnc2014spoken/spoken/tagged/S29X-tgd.xml"
there are 1251 spoken files! Press ENTER to run the query (this may take a while)...
Loading required package: proto
  |============                                                     |  13%
```

BNC.2014.query() pas à pas

génération d'un premier fichier de données

```
speaker_id  word
s0094   hello
s0095   hello
s0032   hello
s0095   hello
s0032   hello
unkfemale   hello
s0032   hello
unkmale hello
unkfemale   hello
s0095   hello
s0032   hello
s0095   hello
s0021   hello
unkfemale   hello
```

BNC.2014.query() pas à pas

croisement avec les métadonnées

```
Do you want to run a query over the whole spoken BNC 2014 or just a sample?

1: the whole corpus (1251 files)
2: a random sample (100 files)

Selection: 1
 [1] "/bnc2014spoken/spoken/tagged/S23A-tgd.xml" "/bnc2014spoken/spoken/tagged/S24A-tgd.xml"
 [3] "/bnc2014spoken/spoken/tagged/S24D-tgd.xml" "/bnc2014spoken/spoken/tagged/S24E-tgd.xml"
 [5] "/bnc2014spoken/spoken/tagged/S263-tgd.xml" "/bnc2014spoken/spoken/tagged/S26N-tgd.xml"
 [7] "/bnc2014spoken/spoken/tagged/S27D-tgd.xml" "/bnc2014spoken/spoken/tagged/S28F-tgd.xml"
 [9] "/bnc2014spoken/spoken/tagged/S29Q-tgd.xml" "/bnc2014spoken/spoken/tagged/S29X-tgd.xml"
there are 1251 spoken files! Press ENTER to run the query (this may take a while)...
Loading required package: proto
  |======================================================================================================| 99%
Press ENTER to save the results in the following file: <interim.results.BNC.2014.txt>...
Here is the path to your output file: /Users/guillaumedesagulier/interim.results.BNC.2014.txt
Press ENTER to save the results in the following file: <data.final.BNC.2014.txt>...
```

BNC.2014.query() pas à pas

croisement avec les métadonnées

Métadonnées : https://www.nakala.fr/data/11280/c750d6f9

```
    speaker_id   age gender                          city
561      S0579 90_99      M                     Tavistock
321      S0327 70_79      F                        London
153      S0155 40_49      M                     Cambridge
74       S0074 60_69      F Hampton-in-Arden, West Midlands
228      S0231 30_39      M
146      S0148 19_29      F                       Reading
       dialect social_grade   education
561    British            A 2_secondary
321    English            B  4_graduate
153  Yorkshire            A  4_graduate
74    Southern            D 2_secondary
228   Southern            E 2_secondary
146 Bristolian            B 3_sixthform
```

BNC.2014.query() pas à pas

croisement avec les métadonnées

genres :

```
  F   M   X
365 305   1
```

classes d'âge :

```
  0_10   11_18   19_29   30_39   40_49   50_59   60_69
     7      42     250      89      76      77      65
 70_79   80_89   90_99 Unknown
    33      19       4       9
```

catégories socio-professionnelles (classification NRS) :

```
     A       B      C1      C2       D       E unknown
   101     149      73      14      53     260      21
```

BNC.2014.query() pas à pas

enregistrement du ficher de données final

```
1: the whole corpus (1251 files)
2: a random sample (100 files)

Selection: 1
 [1] "/bnc2014spoken/spoken/tagged/S23A-tgd.xml" "/bnc2014spoken/spoken/tagged/S24A-tgd.xml"
 [3] "/bnc2014spoken/spoken/tagged/S24D-tgd.xml" "/bnc2014spoken/spoken/tagged/S24E-tgd.xml"
 [5] "/bnc2014spoken/spoken/tagged/S263-tgd.xml" "/bnc2014spoken/spoken/tagged/S26N-tgd.xml"
 [7] "/bnc2014spoken/spoken/tagged/S27D-tgd.xml" "/bnc2014spoken/spoken/tagged/S28F-tgd.xml"
 [9] "/bnc2014spoken/spoken/tagged/S29Q-tgd.xml" "/bnc2014spoken/spoken/tagged/S29X-tgd.xml"
there are 1251 spoken files! Press ENTER to run the query (this may take a while)...
  |================================================================================================== |  99%
Press ENTER to save the results in the following file: <interim.results.BNC.2014.txt>...
Here is the path to your output file: /Users/guillaumedesagulier/interim.results.BNC.2014.txt
Press ENTER to save the results in the following file: <data.final.BNC.2014.txt>...
Here is the path to your final output file: /Users/guillaumedesagulier/data.final.BNC.2014.txt
Press ENTER to see a snapshot of your results (first 20 lines)...
```

BNC.2014.query() pas à pas

prévisualisation du ficher de données final

	speaker_id	word	age	gender	city	dialect	social_grade
1	S0002	hello	19_29	F	Birmingham	Midlands	B
2	S0004	hello	30_39	M	Birmingham	Northern	C2
3	S0005	hello	80_89	F	Birmingham	Midlands	E
4	S0006	hello	80_89	M	Birmingham	Midlands	E
5	S0006	hello	80_89	M	Birmingham	Midlands	E
6	S0007	hello	19_29	M	Birmingham	Midlands	D
7	S0008	hello	60_69	M	Dereham, Norfolk	Norfolk	E
8	S0008	hi	60_69	M	Dereham, Norfolk	Norfolk	E
9	S0008	hello	60_69	M	Dereham, Norfolk	Norfolk	E
10	S0008	hello	60_69	M	Dereham, Norfolk	Norfolk	E
11	S0008	hello	60_69	M	Dereham, Norfolk	Norfolk	E
12	S0008	hello	60_69	M	Dereham, Norfolk	Norfolk	E
13	S0008	hello	60_69	M	Dereham, Norfolk	Norfolk	E
14	S0012	hello	70_79	M	Mattishal, Norfolk	Norfolk	E
15	S0012	hello	70_79	M	Mattishal, Norfolk	Norfolk	E
16	S0012	hello	70_79	M	Mattishal, Norfolk	Norfolk	E
17	S0012	hello	70_79	M	Mattishal, Norfolk	Norfolk	E
18	S0012	hi	70_79	M	Mattishal, Norfolk	Norfolk	E
19	S0012	hello	70_79	M	Mattishal, Norfolk	Norfolk	E
20	S0012	hello	70_79	M	Mattishal, Norfolk	Norfolk	E

>

BNC.2014.query() pas à pas

choix de la visualisation (ici : graphe de Cohen-Friendly)

```
Do you want to plot the results? (0 to exit)

1: barplot
2: association plot
3: no plot (+ archive working files)

Selection:
```

BNC.2014.query() pas à pas

choix de l'effet (ici : l'âge)

```
Do you want to plot the results? (0 to exit)

1: barplot
2: association plot
3: no plot (+ archive working files)

Selection: 2
What effect are you interested in?

1: gender
2: age
3: social class

Selection: 2
```

BNC.2014.query() pas à pas

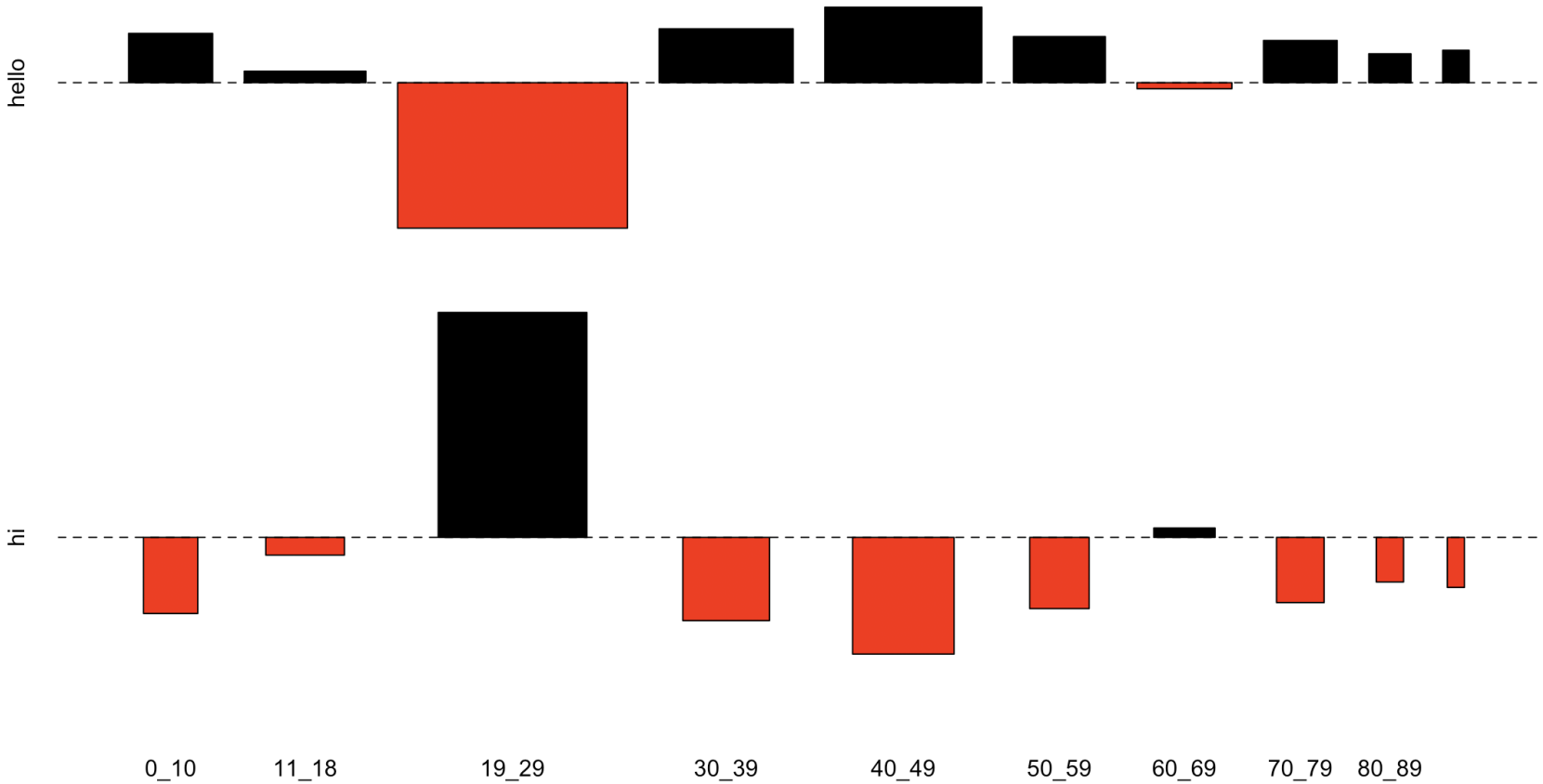
tabulation & résultat du test de χ^2

$$\chi^2 = \sum_{i=1}^{n} \frac{(O - E)^2}{E} \qquad \text{(i)}$$

```
What effect are you interested in?

1: gender
2: age
3: social class

Selection: 2
      0_10 11_18 19_29 30_39 40_49 50_59 60_69 70_79 80_89 90_99
hello   77   139   377   186   261    90    80    59    20     9
hi      14    52   305    46    57    19    36    11     3     0

        Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  tab.of.counts
X-squared = 137.03, df = NA, p-value = 0.0004998
```

BNC.2014.query() pas à pas

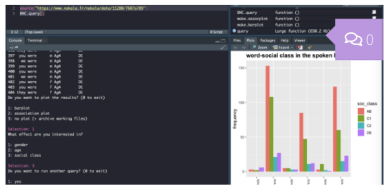
graphe de Cohen-Friendly



tutoriels

Pour `BNC.query()`

`https://corpling.hypotheses.org/2252`

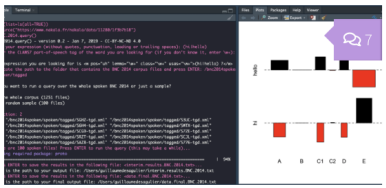


BNC XML / CORPORA / R 08/01/2019

BNC.query(). An interactive R script for a sociolinguistic exploration of the spoken component of the BNC-XML

Pour `BNC.2014.query()`

`https://corpling.hypotheses.org/1632`



BNC 2014 / CORPORA / R 03/01/2019

BNC.2014.query(). An interactive R script for a sociolinguistic exploration of the spoken component of the BNC-2014

la variation liée au genre

Arrière-plan théorique :

Robin Lakoff (1973). « Language and woman's place ». In : *Language in society* 2.1, p. 45-79

Deborah Tannen et al. (1990). *You just don't understand : Women and men in conversation*. Morrow New York (genderlect)

Arrière plan méthodologique :

Paul Rayson et al. (1997). « Social differentiation in the use of English vocabulary : some analyses of the conversational component of the British National Corpus ». In : *International Journal of Corpus Linguistics* 2.1, p. 133-152

Hans-Jörg Schmid (2003). « Do women and men really live in different cultures ? Evidence from the BNC ». In : *Corpus Linguistics by the Lune*. Sous la dir. d'Andrew Wilson et al. T. 8. Frankfurt : Peter Lang, p. 185-221

la variation liée au genre

Rayson et al. (1997)

Table 2.4 Main lexical differences between sex, age and social class categories in the BNC (adapted from Rayson et al. 1997)

Sex		Age		Social class	
Male	Female	Under 35s	Over 35s	ABC1	C2DE
fucking, er, the, yeah, aye, right, hundred, fuck, is, of, two, three, a, four, ah, no, number, quid, one, mate, which okay, that, guy, da, yes	she, her, said, n't, I, and, to, cos, oh, Christmas, thought, lovely, nice, mm, had, did, going, because, him, really, school, he, think, home, me	mum, fucking, my, mummy, like, na, goes, shit, dad, daddy, me, what, fuck, wan, really, okay, cos, just, why	yes, well, mm, er, they, said, says, were, the, of, and, to, mean, he, but, perhaps, that, see, had	yes, really, okay, are, actually, just, good, you, erm, right, school, think, need, your, basically, guy, sorry, hold, difficult, wicked, rice, class	he, says, said, fucking, ain't, yeah, its, them, aye, she, bloody, pound, I, hundred, well, n't, mummy, that, they, him, were, four, bloke, five, thousand

la variation liée au genre

Schmid (2003)

Categories believed to be more typically used by males	Swear-words	*gosh, bloody, shit, damn*	*fuck, fucking*
	Car and traffic	*bus, train, car*	*traffic, crane, windscreen, miles per hour*
	Work	*holiday*	*boss, job, office, meeting, file, colleague*
	Sport	*tennis*	*football, ball, shot, rugby, referee, darts, match, sports*
	Public affairs	-	*reform, government, council, election, Tories, tax, war, Labour*
	Abstract concepts	-	*idea, difference, option, problem, fact, focus, quality*

les gros mots à l'aune du genre

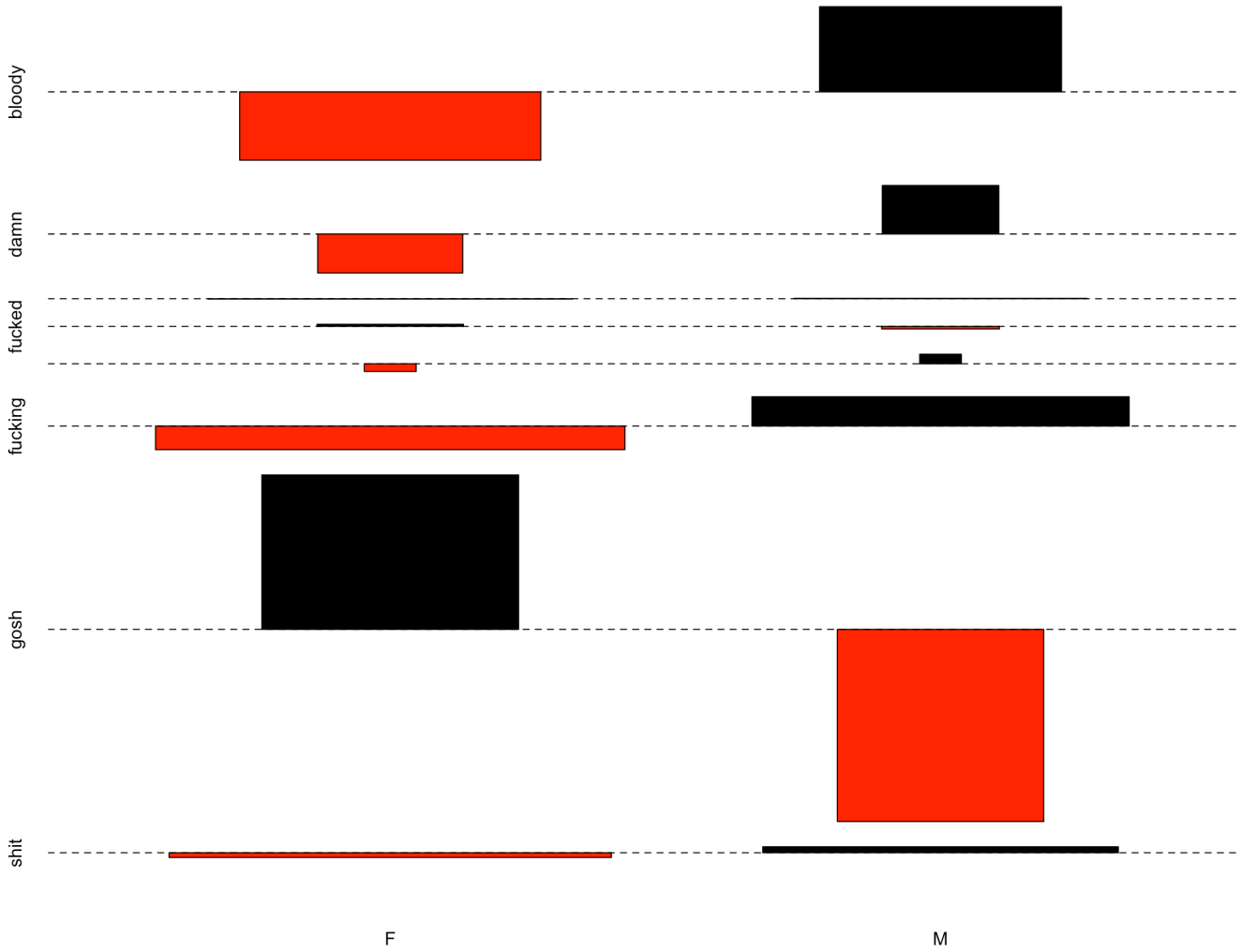
- hypothèse de recherche de Schmid (2003) : les locutrices emploient moins de gros mots que les locuteurs dans le BNC-XML ("démographique").
- résultats de Schmid (2003) : les locutrices emploient certains gros mots plus que les hommes (*gosh*, *bloody*, *shit*, *damn*) mais les mots les plus forts (*f****, *f*******, *f*******) restent l'apanage des locuteurs.

les gros mots à l'aune du genre

```
Selection: 2
What effect are you interested in?

1: gender
2: age
3: social class

Selection: 1
           F    M
bloody   770  678
damn     174  162
fuck    1288  833
fucked   210  133
fucker    24   19
fucking 2079 1440
gosh     851  202
shit    1888 1238

        Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  tab.of.counts
X-squared = 228.75, df = NA, p-value = 0.0004998
```

Il est possible de regrouper les *f*-words !

les gros mots à l'aune du genre

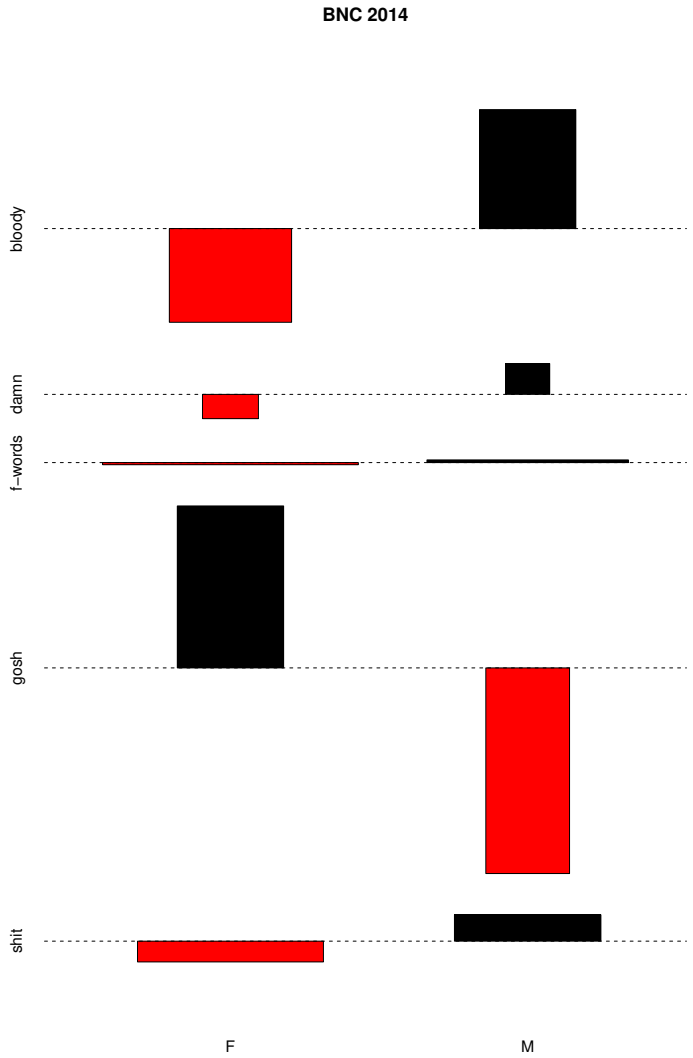


les gros mots à l'aune du genre

Une fois les fichiers sont archivés. On peut les récupérer, les modifier et procéder à d'autres analyses à l'aide de fonctions dans R.

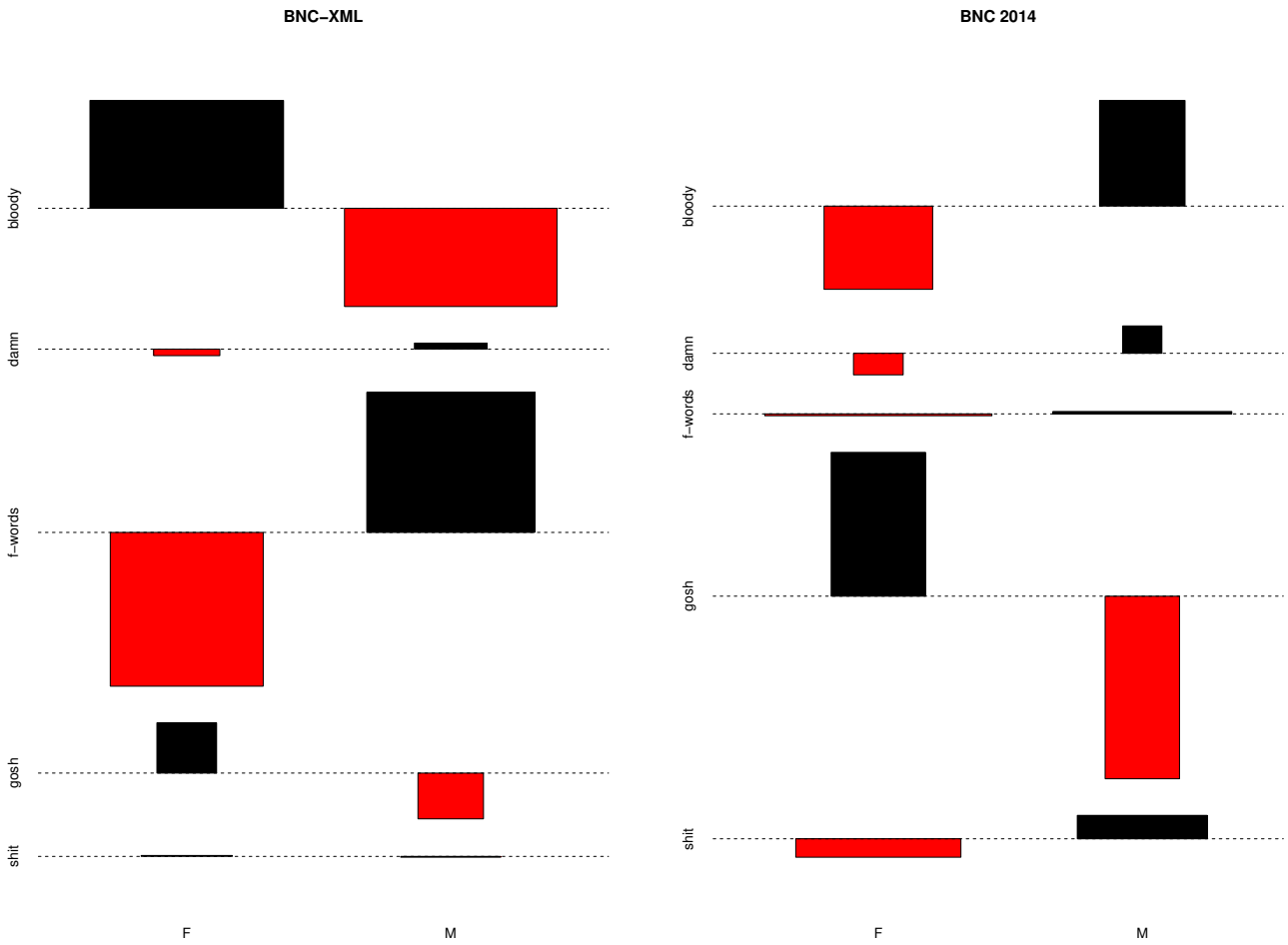
function call	what it does
`make.barplot()`	opens the barplot menu
`barplot.gender()`	makes a barplot based on gender
`barplot.age()`	makes a barplot based on age groups
`barplot.social.class()`	makes a barplot based on social grades
`make.assocplot()`	opens the association plot menu
`assplot.gender()`	makes an association plot based on gender
`assplot.age()`	makes an association plot based on age groups
`assplot.soc.class()`	makes an association plot based on social grades

les gros mots à l'aune du genre



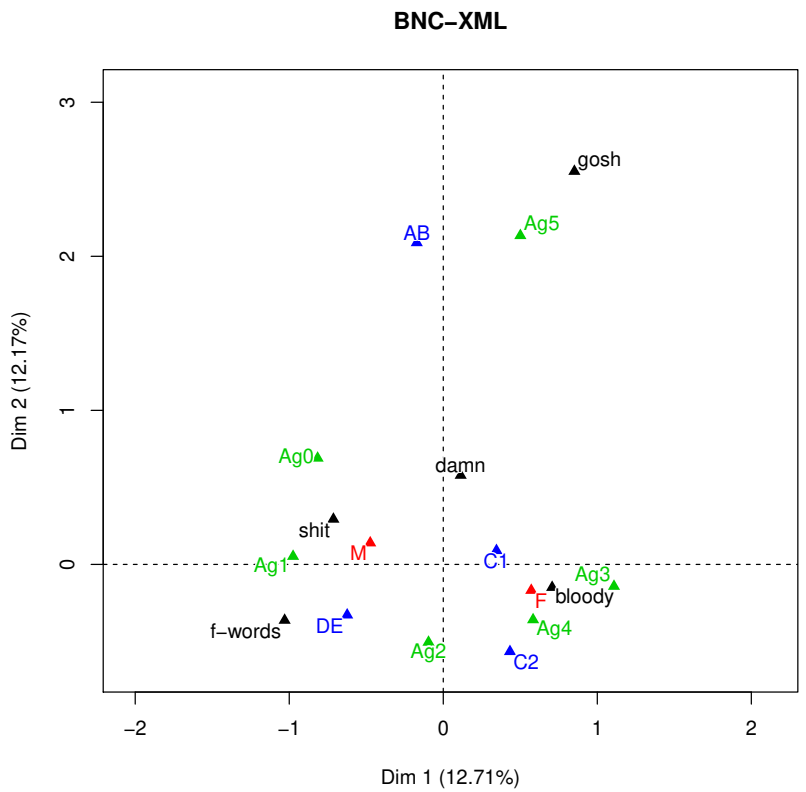
les gros mots à l'aune du genre

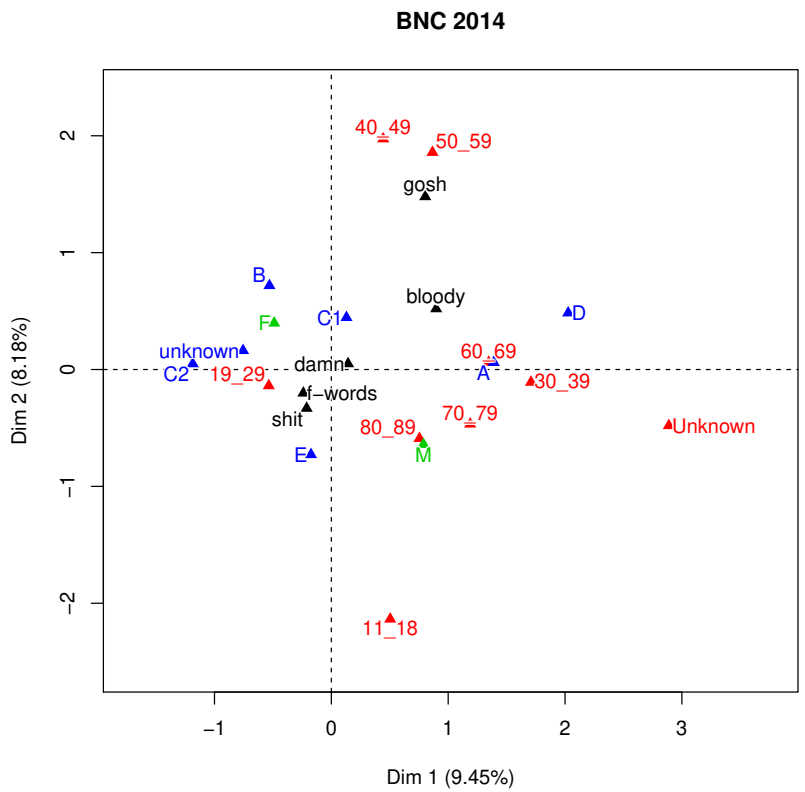
comparaison BNC-XML/BNC 2014



les gros mots à l'aune du genre

comparaison BNC-XML/BNC 2014 (ACM)





question de la représentativité

La représentativité d'un corpus concerne la variabilité. Elle pose question dans les deux corpus.
D'un côté, l'effort a été fait pour ce qui est de la représentativité en termes de

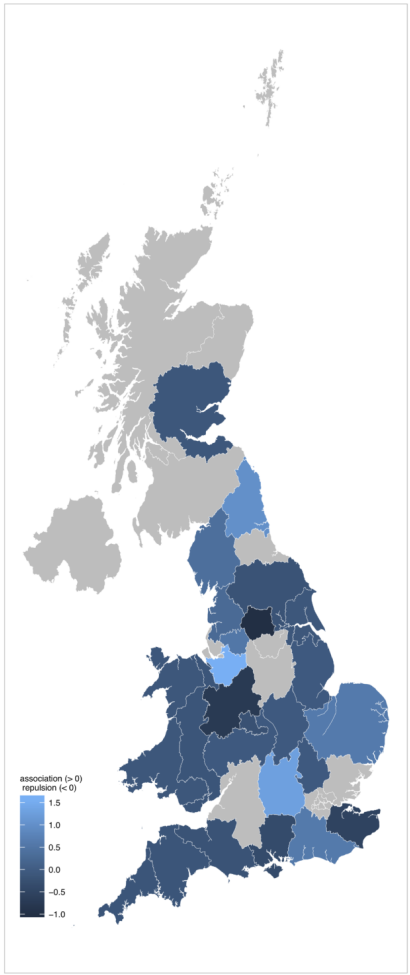
- genres
- de classes d'âge
- de catégories socio-professionnelles

De l'autre, toutes la granularité géographique et dialectale laisse à désirer

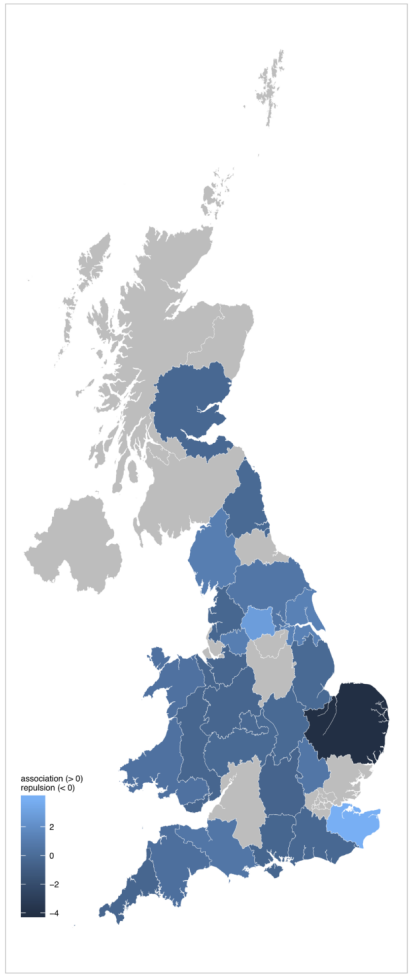
question de la représentativité

un exemple `https://corpling.hypotheses.org/2714` (BNC 2014)

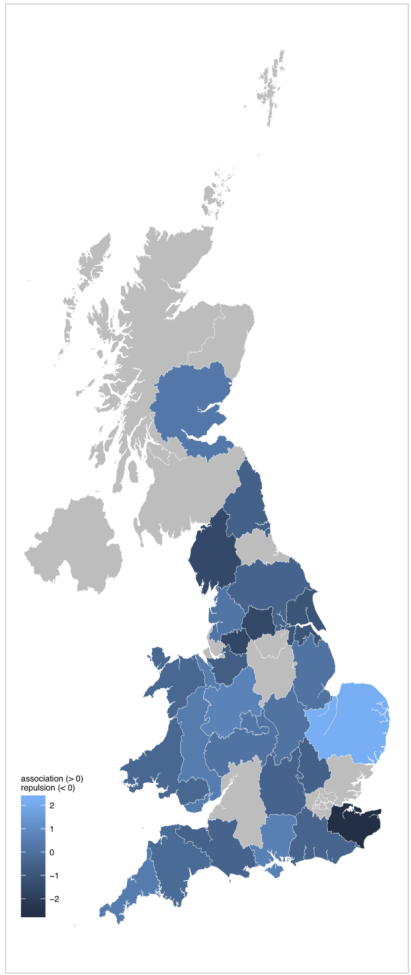
couch in the BNC 2014



settee in the BNC 2014

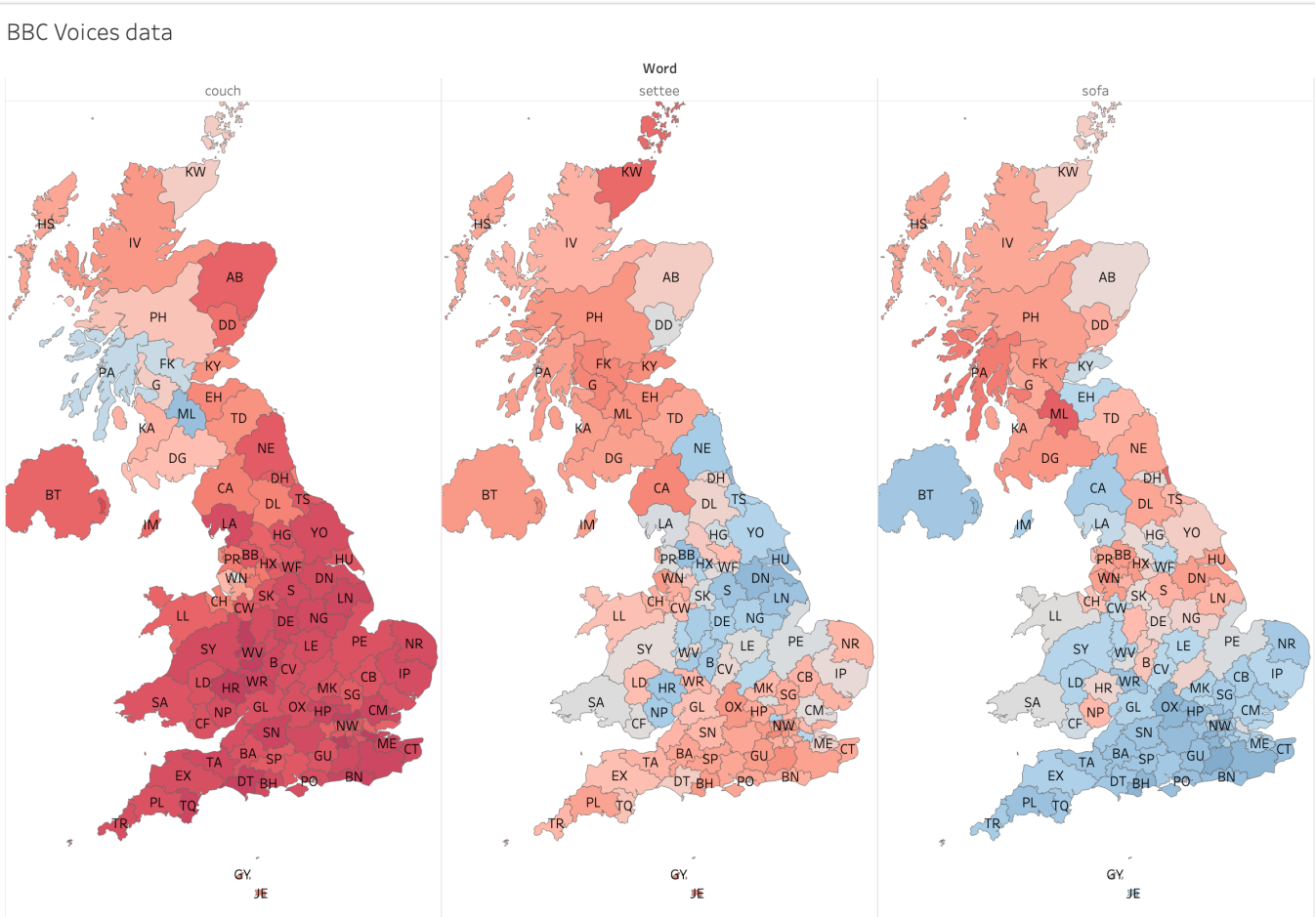


sofa in the BNC 2014



question de la représentativité

un exemple `https://corpling.hypotheses.org/2853` (BBC Voices)



améliorations futures

Quelques pistes en cours de développement :

- intégration du BNC-XML et du BNC 2014 dans la même interface
- intégration d'un module graphique pour la confection de réseaux lexicaux & constructionnels
- module d'analyse des correspondances pour l'exploration multifactorielle des données
- outil générique applicable aux corpus normés TEI/XML

merci pour votre attention !

https://corpling.hypotheses.org/

References I

- Brezina, Vaclav, Robbie Love et Karin Aijmer (2018). *Corpus Approaches to Contemporary British Speech : Sociolinguistic Studies of the Spoken BNC2014*. Routledge.
- Calude, Andreea S (2017). « Sociolinguistic variation at the grammatical/discourse level : Demonstrative clefts in spoken British English ». In : *International Journal of Corpus Linguistics* 22.3, p. 429-455.
- Desagulier, Guillaume (2017). *Corpus Linguistics and Statistics with R*. New York : Springer.
- Fuchs, Robert (2017). « Do women (still) use more intensifiers than men ? : Recent change in the sociolinguistics of intensifiers in British English ». In : *International Journal of Corpus Linguistics* 22.3, p. 345-374.

References II

- Hessner, Tanja et Ira Gawlitzek (2017). « Totally or slightly different ? : A Spoken BNC2014-based investigation of female and male usage of intensifiers ». In : *International Journal of Corpus Linguistics* 22.3, p. 403-428.
- Lakoff, Robin (1973). « Language and woman's place ». In : *Language in society* 2.1, p. 45-79.
- Laws, Jacqueline, Chris Ryder et Sylvia Jaworska (2017). « A diachronic corpus-based study into the effects of age and gender on the usage patterns of verb-forming suffixation in spoken British English ». In : *International Journal of Corpus Linguistics* 22.3, p. 375-402.
- McEnery, Tony, Robbie Love et Vaclav Brezina (2017). « Compiling and analysing the Spoken British National Corpus 2014. Special issue of ». In : *International Journal of Corpus Linguistics* 22.3.

References III

- Rayson, Paul, Geoffrey N Leech et Mary Hodges (1997). « Social differentiation in the use of English vocabulary : some analyses of the conversational component of the British National Corpus ». In : *International Journal of Corpus Linguistics* 2.1, p. 133-152.
- Schmid, Hans-Jörg (2003). « Do women and men really live in different cultures ? Evidence from the BNC ». In : *Corpus Linguistics by the Lune*. Sous la dir. d'Andrew Wilson, Paul Rayson et Tony McEnery. T. 8. Frankfurt : Peter Lang, p. 185-221.
- Tannen, Deborah et al. (1990). *You just don't understand : Women and men in conversation*. Morrow New York.