

TEXT MINING FOR SOCIOLINGUISTIC RESEARCH:

*Lexical Borrowing In
French Print Media*

*Gyula Zsombok • zsombok2@illinois.edu
University of Illinois at Urbana-Champaign
Department of French and Italian*



Research focus

- Lexical borrowings and prescriptivism in French newspapers
- For each variable, there are at least 2 variants:
 - English pure borrowing, e.g. **email**
 - French prescribed term, e.g. **courriel**
 - (Other variant, e.g. **mail**)
- Does prescriptivism affect the probability of use?
- Do topics differ in genres/audience?

Lexis Nexis: Search for data

LexisNexis® Academic

Academic Search

Search by Subject or Topic ▼

Enter Search Terms

Search

Advanced Options ▶

Date

From

to

Or
Select

Source

Build Your Own Segment Search ?

Select a Segment ▼

Useful when you need to narrow your search to something specific, like a legal citation, an author's name, or a word in the headline. For example: HEADLINE(Congress w/5 budget)

Lexis Nexis: Data collection

- | | Results |
|----|---|
| 1. | Analyse; L'Algérie confrontée à ses multiples paradoxes
Le Monde, 1 janvier 2011 samedi, EDITORIAL - ANALYSES; Pg. 17, 874 mots |
| 2. | PIXELS; BONNE ANNÉE 2011 AUX AMATEURS !
Le Monde, 1 janvier 2011 samedi, LE MONDE TÉLÉVISION; Pg. 30, 675 mots |
| 3. | Drôles de sondages
Le Monde, 1 janvier 2011 samedi, EDITORIAL - ANALYSES; Pg. 17, 457 mots |
| 4. | France; Professeur d'optimisme, professeurs de lucidité
Le Monde, 4 janvier 2011 mardi, EDITORIAL - ANALYSES; Pg. 18, 987 mots |
| 5. | Le renouveau de la communauté juive
Le Monde, 4 janvier 2011 mardi, DERNIÈRE HEURE; Pg. 24, 894 mots |
| 6. | Ecologie; Court terme, long terme
Le Monde, 5 janvier 2011 mercredi, EDITORIAL - ANALYSES; Pg. 15, 446 mots |
| 7. | Analyse; 2011, l'an I de la réforme du statut du parquet ?
Le Monde, 5 janvier 2011 mercredi, EDITORIAL - ANALYSES; Pg. 15, 865 mots |
| 8. | Lettre de Wall Street: Plus c'est gros |

Lexis Nexis: Example article

2 of 37 DOCUMENTS

Le Figaro

Jeudi 26 Mars 2009

Les curieuses demandes d'un magistrat à la famille d'un mort;
JUSTICE Le juge d'instruction a chargé une psychologue d'interroger les proches de la victime sur leurs « principaux traits de caractère » et leur « niveau d'information sexuelle ».

AUTEUR: Durand-Souffland, Stéphane

RUBRIQUE: FRANCE; Société; Pg. 9 N° 20110

LONGUEUR: 510 words

Python code: Read in the files



```
#Import re (Regular Expressions), os (Operating system), glob  
import re, os, glob  
  
#Set path name, lexical borrowing name, splitting pattern  
pathname = r"C:/Users/Gyuszi/Dropbox/Conferences/2016/2016 NWAV/Workshop/  
variable = "email"  
split_pattern = r"\s*[0-9]*\s*of\s*[0-9]*\s*DOCUMENTS?\s*"  
  
#Change current directory and find all .txt files there  
os.chdir(pathname)  
files = glob.glob("*.txt")
```

Python code: Read in the files



```
textlist0 = []
for element in text2:
    splittext = re.split(split_pattern, element)
    textlist0.append(splittext)

#Delete the very first, whitespace element of each list
for element in textlist0:
    del element[0]

#Merge all list together
textlist = sum(textlist0, [])
```

Python code: Preprocessing



```
preprocessing = []
for element in textlist:
    extranewlines = re.sub(r"\n\n?\W*", r"\n", element)
    copyright = re.sub(r"(\n?Copyright.*) (\n?)", r"\2", extranewlines)
    tousdroits = re.sub(r"(\n?Tous droits réservés) (\n?)", r"\2",
    encart = re.sub(r"(ENCART:\W|HEADLINE:\W)", r"", tousdroits)
    metainfo = re.sub(r"[A-Z-]+:.*\n?", r"", encart)
    startingday = re.sub(r"^(.*\n) ([Ll]undi| [Mm]ardi| [Mm]ercredi| [Jj]eudi|
    endingday = re.sub(r"(\W*) ([Ll]undi| [Mm]ardi| [Mm]ercredi| [Jj]eudi|
    subJDN = re.sub(r"Journal\W?du\W?Net,\W?JDN\W?Solutions",
    preprocessing.append(subJDN)
```


Python code: Date issue



January 2 2001

January 02 2001

2 janvier 2001

02 janvier 2001

Mardi 2 janvier 2001

2 janvier 2001 mardi



2001-01-02

Python code: Date issue



```
preprocessing = []
for element in textlist:
    extranewlines = re.sub(r"\n\n?\W*", r"\n", element)
    copyright = re.sub(r"(\n?Copyright.*) (\n?)", r"\2", extranewlines)
    tousdroits = re.sub(r"(\n?Tous droits réservés) (\n?)", r"\2",
    encart = re.sub(r"(ENCART:\W|HEADLINE:\W)", r"", tousdroits)
    metainfo = re.sub(r"[A-Z-]+:.*\n?", r"", encart)
    startingday = re.sub(r"^(.*\n) ([Ll]undi|[Mm]ardi|[Mm]ercredi|[Jj]eudi
    endingday = re.sub(r"(\W*) ([Ll]undi|[Mm]ardi|[Mm]ercredi|[Jj]eudi|
    subJDN = re.sub(r"Journal\W?du\W?Net,\W?JDN\W?Solutions",
    preprocessing.append(subJDN)
```

Python code: Date issue



```
#Reorder DAY MONTH YEAR dates into YEAR MONTH DAY  
subdates = []  
for element in preprocessing:  
    subdate1 = re.sub(r"(\n) (\d{,2}) \W* ([JjFfMmAaSsOoNnDd] \w{0,8}) \W* (\d{4})  
    subdate2 = re.sub(r"(\n) ([JjFfMmAaSsOoNnDd] \w{0,8}) \W* (\d{,2}), ?\W*  
    subdates.append(subdate2)
```

2 janvier 2001  2001 janvier 2

Python code: Date issue




```
#Finalize YEAR MONTH DAY line with new formatting
subdates2 = []
for element in subdates:
    if re.search(r"\n\d{4}\W*[Jj]an.i?.ry?\W*\d{1,2}\n", element):
        januaryelem = re.sub(r"(\d{4})\W*([Jj]an.i?.ry?)\W*(\d{1,2})",
            subdates2.append(januaryelem)
    elif re.search(r"\n\d{4}\W*[Ff]..r..ry?\W*\d{1,2}\n", element):
        februaryelem = re.sub(r"(\d{4})\W*([Ff]..r..ry?)\W*(\d{1,2})",
            subdates2.append(februaryelem)
    elif re.search(r"\n\d{4}\W*[Mm]arc?.\W*\d{1,2}\n", element):
        marchelem = re.sub(r"(\d{4})\W*([Mm]arc?.)\W*(\d{1,2})", r"\1-03-\3",
            subdates2.append(marchelem)
    elif re.search(r"\n\d{4}\W*[Aa].ril\W*\d{1,2}\n", element):
        aprilelem = re.sub(r"(\d{4})\W*([Aa].ril)\W*(\d{1,2})", r"\1-04-\3",
            subdates2.append(aprilelem)
```

2001 janvier 2  2001-01-2

Python code: Date issue



```
#Correct some of the date formats
subdates3 = []
for element in subdates2:
    subdate5 = re.sub(r"(\d{4})-(\d{2})-(\d{1,1})", r"\1-\2-0\3",
    subdate6 = re.sub(r"(\d{4})-(\d{2})-0(\d{2})", r"\1-\2-\3", subdate5)
    subdates3.append(subdate6)
```

2001-01-2  2001-01-02

Python code: Assign values



```
#Add genre variable
professionalgenres = r"01net|Journal\W*du\W*Net"
generalgenres = r"Le\W*Monde|Le\W*Figaro"

addgenre = []
for element in subdates3:
    if re.search(professionalgenres + r"\n\d{4}-\d{2}-\d{2}\n", element):
        professionalgenre = re.sub(r"(\n) (\d{4}-\d{2}-\d{2}\n)",
        addgenre.append(professionalgenre)
    elif re.search(generalgenres + r"\n\d{4}-\d{,2}-\d{,2}\n", element):
        generalgenre = re.sub(r"(\n) (\d{4}-\d{2}-\d{2}\n)",
        addgenre.append(generalgenre)
```

Python code: Assign values



```
#Add variable (e.g. e-mail, fax) and type (e.g. English, Prescribed, Other)
addtype = []
for element in addgenre:
    if re.search(r"*/English", pathname):
        englishtype = re.sub(r"(\W*) (.*\n.*\n\d{4}-\d{2}-\d{2}\n)",
            addtype.append(englishtype)
    elif re.search(r"*/Prescribed", pathname):
        prescribedtype = re.sub(r"(\W*) (.*\n.*\n\d{4}-\d{2}-\d{2}\n)",
            addtype.append(prescribedtype)
    elif re.search(r"*/Other", pathname):
        othertype = re.sub(r"(\W*) (.*\n.*\n\d{4}-\d{2}-\d{2}\n)",
            addtype.append(othertype)
```


Python code: CSV preparations



```
#Delete all ";" character, because that will be used as the comma separator
csvlist = []
for element in addtype:
    semicolonreplace = re.sub(";", ",", element)
    csvlist0 = re.sub(r"(.*)\n(.*)\n(.*)\n(.*)\n(\d{4}-\d{2}-\d{2})\n",
    csvlist1 = re.sub(r"\n", r" ", csvlist0)
    csvlist.append(csvlist1)
```


Python code: Check for loss



```
#Verify the length of the lists, so to see if nothing was lost in the process
print("The length of textlist: ", len(textlist))
print("The length of preprocessing: ", len(preprocessing))
print("The length of subdates: ", len(subdates))
print("The length of subdates2: ", len(subdates2))
print("The length of subdates3: ", len(subdates3))
print("The length of addgenre: ", len(addgenre))
print("The length of addtype: ", len(addtype))
print("The length of csvlist: ", len(csvlist))
```

Python code: Write output



```
#Delete new line characters from the string  
finalcsvlist = csvlist  
  
#Insert header of the CSV file  
finalcsvlist.insert(0, "VARIABLE;TYPE;SOURCE;GENRE;DATE;ARTICLE")  
  
#Create CSV output  
csvoutput = open("courrielWorkshop.csv", "w+")  
print("\n".join(finalcsvlist), file = csvoutput)  
csvoutput.close()
```

Excel: Read in output



The screenshot shows the Microsoft Excel ribbon with the 'Data' tab selected. The ribbon includes the following groups and options:

- Get External Data:** From Access, From Web, From Text (highlighted), From Other Sources, Existing Connections.
- Get & Transform:** New Query, Show Queries, From Table, Recent Sources.
- Connections:** Refresh All, Connections, Properties, Edit Links.
- Sort & Filter:** Sort (A-Z, Z-A), Filter, Clear All Filters, Reapply, Advanced Filter.

A tooltip for the 'From Text' option is displayed, containing the text: **Get Data From Text** and **Import data from a text file.**

The spreadsheet area below the ribbon shows a grid with columns A through K and rows 1 through 5. Cell A1 is selected, and the tooltip is positioned over it.

Excel: Read in output



G The Text Wizard has determined that your data is Fixed Width.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

Delimited - Characters such as commas or tabs separate each field.

Fixed width - Fields are aligned in columns with spaces between each field.

Start import at row: File origin:

My data has headers.

Preview of file C:\Users\Gyuszi\Dropbox\Conferences\2016\2016 NWA\Workshop\Prescribed\courrielWorkshop.csv

1	VARIABLE;TYPE;SOURCE;GENRE;DATE;ARTICLE
2	email;prescribed;01net;professional;2007-03-26;Sécurité externalisée : BT France baisse s
3	email;prescribed;01net;professional;2007-04-17;L'antispam de MailInBlack personnalise les

Excel: Read in output



Delimiters

 Tab Semicolon Comma Space Other: Treat consecutive delimiters as oneText qualifier: " ▼

Data preview

VARIABLE	TYPE	SOURCE	GENRE	DATE	ARTICLE
email	prescribed	01net	professional	2007-03-26	Sécurité externalisée : BT France baisse
email	prescribed	01net	professional	2007-04-17	L'antispam de MailInBlack personnalise l
email	prescribed	01net	professional	2007-04-27	Remplacez votre messagerie Web par une s
email	prescribed	01net	professional	2007-05-09	Le nouvel Hotmail disponible en version
email	prescribed	01net	professional	2007-05-11	Le "paperboard" passe au numérique Polym

Excel: Read in output



1	VARIABLE	TYPE	SOURCE	GENRE	DATE	ARTICLE
521	email	prescribed	Le Monde	general	10/19/2011	TIMBRES, La philatélie à l'heure numériq
522	email	prescribed	Le Monde	general	10/20/2011	Ne dites pas à David Guetta qu'il est dev
523	email	prescribed	Le Monde	general	10/21/2011	La vie des entreprises TÉLÉVISION, Démi
524	email	prescribed	Le Monde	general	10/22/2011	Médiateur, Electeurs et lecteurs Beauco
525	email	prescribed	Le Monde	general	10/28/2011	DÉCRYPTAGE, Le " cercle vicieux " de qui
526	email	prescribed	Le Monde	general	10/30/2011	Juste un mot, Attention Difficile de reste
527	email	prescribed	Le Monde	general	11/5/2011	Expression, La langue française Ma patri
528	email	prescribed	Le Monde	general	11/9/2011	Courrier des lecteurs Maternité Salariée,
529	email	prescribed	Le Monde	general	11/12/2011	LES COULISSES DE LA PAILLASSE, Du mul
530	email	prescribed	Le Monde	general	11/15/2011	Le mauvais pari de l'austérité La Grèce e
531	email	prescribed	Le Monde	general	11/15/2011	Quand un ex-directeur de l'AEM part " pa
532	email	prescribed	Le Monde	general	11/17/2011	Le Languedoc-Roussillon gère mieux les
533	email	prescribed	Le Monde	general	11/18/2011	REPORTAGE, A Wigan, Ben, 25 ans, maît
534	email	prescribed	Le Monde	general	11/24/2011	Solaire : la bataille s'intensifie entre la Cl

R code: Read in file



```
1 library(readxl)
2 library(XML)
3 library(rvest)
4 library(tm)
5 library(irlba)
6 library(SnowballC)
7 library(Matrix)
8 library(wordcloud)
9 library(rpart)
10 library(DT)
11 library(topicmodels)
12
13 #Set working directory and read in file
14 setwd("C:/Users/Gyuszi/Dropbox/Conferences/2016/2016 NWAV/Workshop")
15 courriel = read_excel("workshop-courriel.xlsx")
16 corp1 = Corpus(VectorSource(courriel$ARTICLE))
17
```


R code: Text cleaning



```
17
18 #Process the text within each document
19 #Remove white space
20 IC1 = tm_map(corp1, stripwhitespace)
21 #Remove punctuation
22 IC1 = tm_map(IC1, removePunctuation)
23 #Convert all characters to lowercase
24 IC1 = tm_map(IC1, tolower)
25 #Ensure that the format is plain text
26 IC1 = tm_map(IC1, PlainTextDocument)
27 #Remove numbers
28 IC1 = tm_map(IC1, removeNumbers)
29 #Remove stopwords
30 IC1 = tm_map(IC1, removeWords, stopwords("french"))
31 #Stem the documents
32 IC1 = tm_map(IC1, stemDocument, "french")
33
```


R code: Text cleaning



```
28
29
30
31
32
33
34 #Remove additional words
35 additional_removeWords = c("abrev", "adjt", "afin", "after", "ains",
36                             "auss", "aussi", "autr", "avoir", "cas",
37                             "deux", "entre", "être", "laquel",
38                             "surtout", "tout", "tres", "trop", "xix",
39 IC1 = tm_map(IC1, removeWords, additional_removeWords)
40
41
42
43
44
```

R code: Document term matrix



```
41
42 #Create Document term matrices
43 ICTdm_dtm_tf1 = DocumentTermMatrix(IC1, control = list(weighting =
44                                                         tolower = FALSE
45 ICTdm_dtm_tf1 = ICTdm_dtm_tf1[!(is.na(iconv(rownames(ICTdm_dtm_tf1),
46                                                         to = "UTF-8")))), ]
47
48 #Set parameters for Gibbs sampling
49 burnin = 4000
50 iter = 2000
51 thin = 500
52 seed = list(2003,5,63,100001,765)
53 nstart = 5
54 best = TRUE
55 #Set the number of topics
56 k = 5
```

Excel: Topics



1		Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	
2	1	sit	franc	mond	droit	dun	
3	2	servic	dun	cest	courriel	comm	
4	3	moyen	million	quil	fait	polit	
5	4	rÃ©seau	mois	journal	dun	prÃ©sident	
6	5	don	bien	franc	internet	pay	
7	6	dun	fair	comm	rÃ©serv	amÃ©ricain	
8	7	utilis	compt	tous	cour	premi	
9	8	courriel	march	grand	comm	depuis	
10	9	lign	pai	dun	jour	droit	
11	10	system	depuis	lecteur	quil	part	
12	11	web	nest	temp	demand	pouvoir	
13	12	euros	prix	trouv	group	social	
14	13	dÃ©velop	seul	fait	person	contr	
15	14	mis	banqu	fair	gÃ©nÃ©ral	national	
16	15	fonction	cart	quand	sit	gouvern	
17	16	messag	annÃ©	quelqu	courri	derni	
18	17	internet	societ	bien	envoi	fait	
19	18	propos	lanc	livr	jug	europÃ©en	
20	19	lentrepris	nouvel	nest	prÃ©cis	ministr	

Topic differences and genre

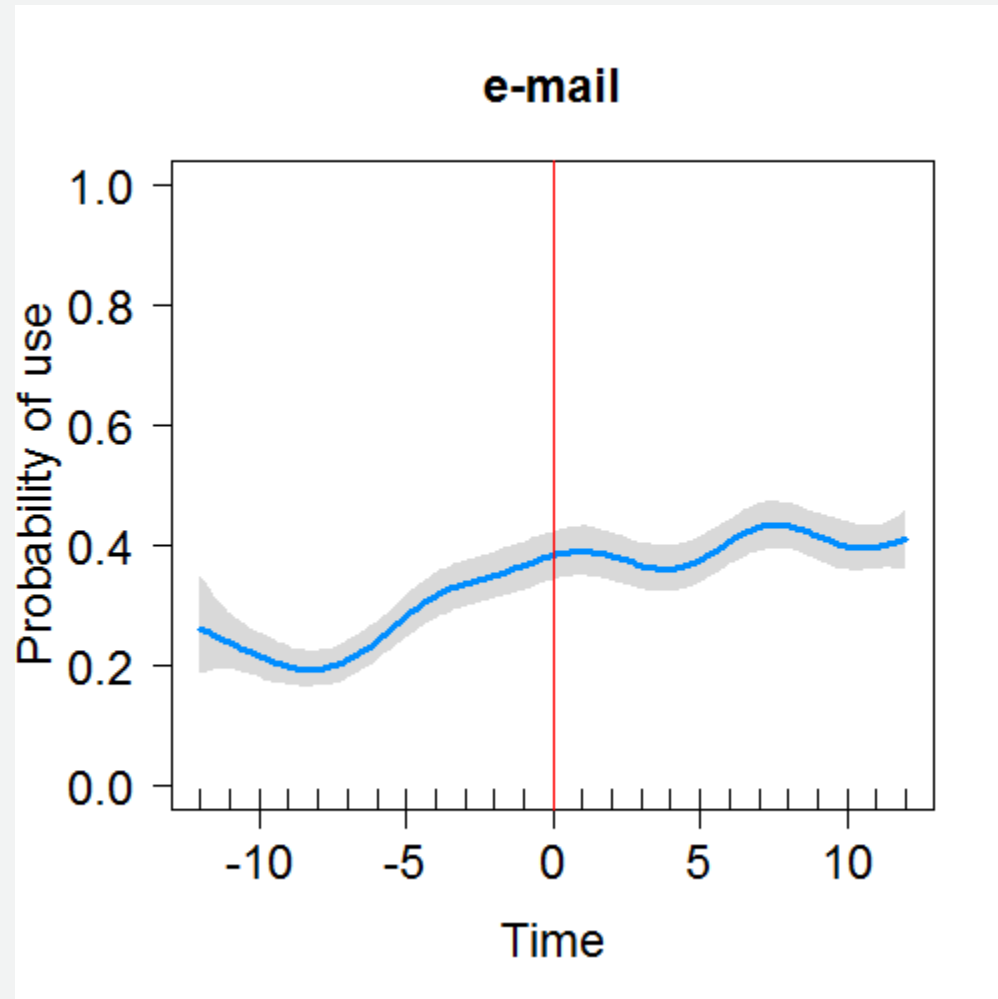


Professional genre



General genre

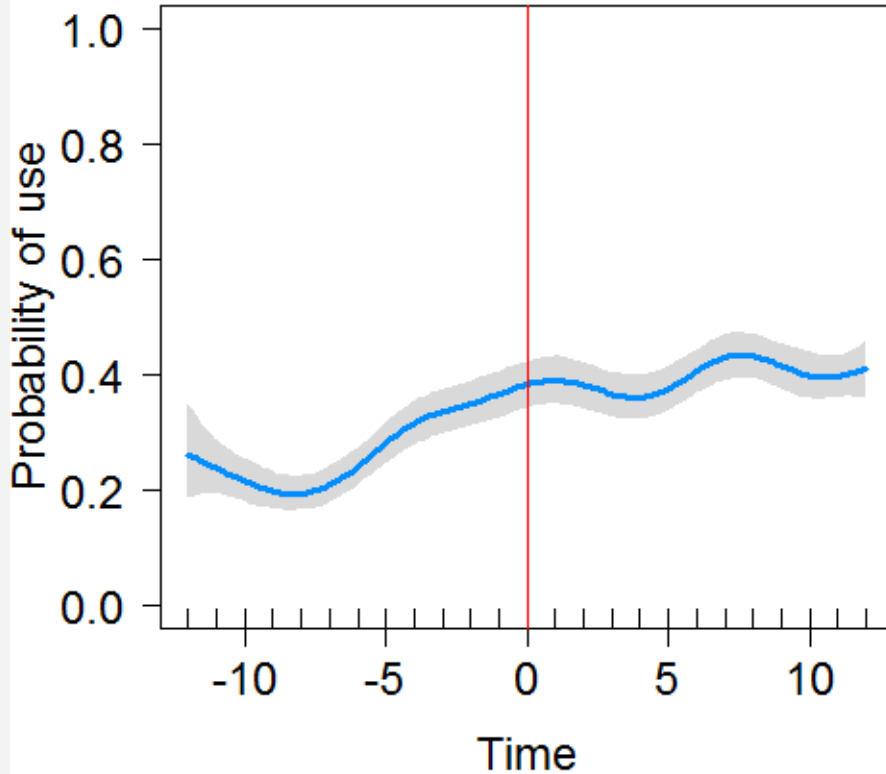
GAM: e-mail, 2000-2006



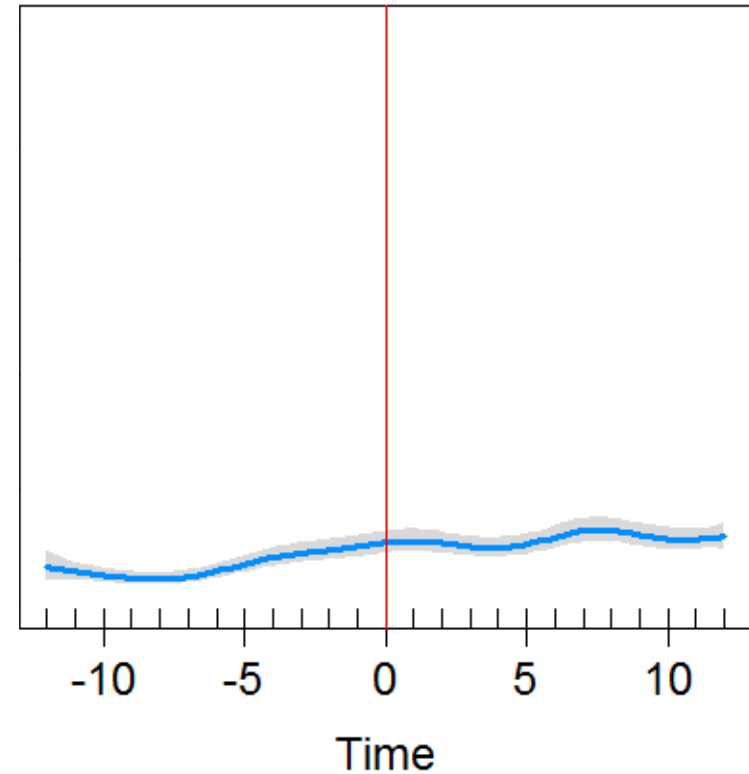
GAM: e-mail and genre, 2000-2007



General



Professional



Thank you!

4:00 PM, Saturday, November 5, 2016, Segal Room (1400-1430)

The effect of prescriptivism on the use of localized terminology in French newspapers

This presentation was supported by the
Graduate College Conference Travel Award
at the University of Illinois at Urbana-Champaign
