Productivité morphologique et lexiques fréquencés

Gauvain Schalchli

Plan

- l’objet de la morphologie dérivationnelle
- La notion de productivité
- l’indice de Baayen
- lexiques fréquencés et productivité
- Exploration de lexique3 avec Zipf et la médiane
- la flexion
- les ion
- conclusion : corpus, lexiques fréquencés et TAL
  - psychoglaff

Le programme de la morphologie dérivationnelle

Methodology of Suffixal analysis (Corbin 1987)

- source dictionnairique : listes de mots, pas de corpus d'énoncés
- principe de panchronie : pas de prise en compte des dates d'attestation
- principe de séparation morphologique : pas de prise en compte de la flexion (lemme)
- Morphèmes (affixes) VS règles

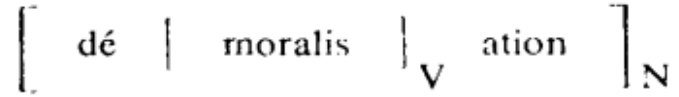
Exemples de « découvertes » dérivationnelles

- « « — Débarcadère, déblocus, déplaisir, désordre peuvent être mis en rapport sémantique avec des verbes, mais la terminaison de ces noms peut difficilement être considérée comme un suffixe, puisqu'elle ne se retrouve que sur les noms non préfixés par dé- correspondants {embarcadère, blocus, plaisir, ordre). Le rapport sémantique entre les noms et les verbes est à la charge de règles sémantiques qui ne peuvent être couplées avec des règles morphologiques : une règle, même non productive, de formation des mots ne peut se limiter à la description d'une unité. » (Corbin 1976, p. 105)

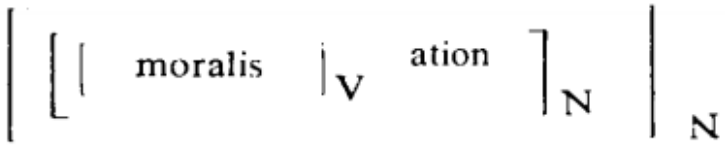
La « grammaire » ou « système » dérivationnel (Corbin 1976, p. 104)

Un nom comme *démoralisation* peut, a priori, recevoir trois analyses :

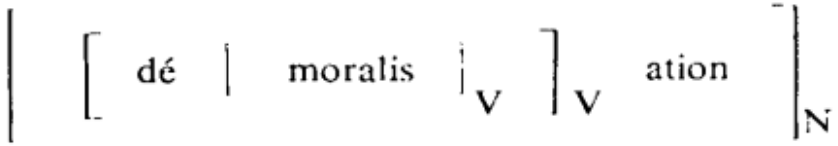
1. Parasynthèse :



2. Préfixation :



3. Suffixation :



On peut rejeter d'emblée la première analyse, parce qu'elle nécessite une règle ad hoc : on a besoin par ailleurs, de toute façon, d'une règle de préfixation par *dé-* et d'une règle de suffixation par *-ation*. Le choix entre la deuxième et la troisième solution ne peut se faire qu'en accord avec l'économie générale des règles. Il a déjà été dit (cf. § 3.2.3.) que les noms porteurs du suffixe *-ation* étaient pratiquement tous dérivés de verbes. La troisième analyse est conforme à cette généralité. De ce fait, la deuxième analyse apparaît plus coûteuse, parce que nécessitant une règle supplémentaire.

Les questions de recherche classiques sur la suffixation en morphologie dérivationnelle

- Quels sont les constituants, procédés morphologiques? -> morphèmes, opérations, processus
- Quelles sont les règles? -> conditions et contraintes d'application des procédés
- Polymorphie -> allomorphie radicale ou suffixale
- compositionnalité sémantique, contraintes sémantiques
- Compétition, blocage
- Productivité -> fonctionnement des patrons morphologiques

TAL et morphologie dérivationnelle

- morphoder pour le TAL
  - moteurs de règles dérivationnelles (MorTal, DERIF)
  - lexiques électroniques (Morphonette, Demonext, etc)
- TAL pour la morphoder
  - traitement automatique de corpus (extraction, nettoyage, filtrage) -> webaffix
  - annotation de données
  - deep learning -> sémantique distributionnelle
  - clustering (Hathout 2009)
  - statistiques (Lstat)

La productivité morphologique

la polysémie de la notion de productivité (Dal 2003)

- Dal et Namer 2016: morphogie = productivité
- « On peut répartir les différentes définitions que la notion de productivité constructionnelle […] a reçues en trois types, selon qu'elles l'appréhendent sous un angle qualitatif, sous un angle quantitatif, ou qu'elles conjoignent qualité et quantité. » (p. 5)
- « […] que ce soit dans l'approche qualitative pure (§ 1.2.1.) ou dans celle qui conjoint qualité et quantité (§ 1.2. 2.), la notion est binaire : un procédé constructionnel ne peut être qu'apte ou inapte à former de nouveaux mots. C'est ainsi que, pour Zwanenburg (1983 : 29-30), qui se réclame de Halle (1973) et de Corbin (1976), il existe une « opposition discrète entre procédés productifs et improductifs » (on retrouve la même opposition dans Booij (1977) et Corbin (1987 : 177-8,) à ceci près que, pour elle, l'opposition trace une ligne entre processus disponibles et non disponibles) » (p. 11)

L’approche qualitative (Dal 2003)

- « aptitude d'un procédé à former de nouvelles unités lexicales » (p. 6) -> H. Marchand
- notion binaire
- « problème de l'apparition sporadique de nouveaux dérivés mettant en œuvre des procédés réputés ne pas être productifs » (p. 6) => truth = true + -th -> coolth, Aronoff & Anshen 1998:243)
- « un patron productif est nécessairement régulier [condition nécessaire mais pas suffisante] » (p. 7)
- « démarcation entre la notion de productivité et celle de créativité morphologique, définie comme la création de nouvelles unités lexicales sans recourir à des règles » (p. 6)

=> « on se retrouve ainsi dans la position inconfortable de devoir considérer productifs des procédés jugés par ailleurs non productifs » (p. 7)

Mesurer la productivité

- « […] dans l'approche strictement qualitative vue au § 1.2.1., il ne devrait pas y avoir de sens à développer des calculs mesurant la productivité, puisque la découverte d'une seule unité lexicale construite ne figurant pas dans les dictionnaires suffit pour décréter productif en synchronie le procédé qui l'a formée. De la même façon, même si ce n'est pas pour les mêmes raisons, dans l'approche qualitativo-quantitative exposée au § 1.2.2., mesurer la productivité devrait être non pertinent, puisque, utilisée intuitivement, la notion d'infini s'oppose à celle de dénombrabilité. » (Dal 2003, p. 11)
- « […] les diverses approches quantitatives qui ont cours ont en commun de voir dans la productivité une notion scalaire susceptible […] d'être affectée d'un indice chiffré : la notion est alors conçue comme un continuum (Aronoff & Anshen 1998 : 243), pouvant aller du non productif à l'entièrement productif, en passant par toutes les valeurs intermédiaires. » (Dal 2003, p. 11)

approche quantitative (1) : type frequency des produits

- nombres de mots correspondant à un patron morphologique
- applicable à des dictionnaires (première génération) -> sous-ensemble des entrées
- « rentabilité » (Corbin 1987 -> profitability chez Plag 1999 et Bauer 2001)
- « realized productivity » (Baayen)
- ne définit une aptitude (Dal 2003, p. 10) -> compétence
- confond en outre productions présente et passée (Dal 2003, p. 10)
- s'appuie sur des inventaires nécessairement fluctuants et soumis aux aléas de l'attestation (Corbin 1987) -> dictionnaires, empiricité

approche quantitative (2) : type frequency ratio des produits

- quotient du nombre de mots possibles [mots qu'une règle peut former] par le nombre de mots attestés [qu'on rencontre réellement dans la langue] (Zwanenburg 1983)
- indice d'Aronoff (1976) : « I = V/S', où V correspond au nombre de dérivés attestés porteurs du procédé étudié, et S' au nombre de dérivés que ce procédé peut former) » (Dal 2003, p. 12)
- circularité : « pour pouvoir donner une valeur au dividende du quotient - en admettant que cela soit possible -, il faut au préalable avoir déterminé la productivité de la règle par un autre moyen » (Dal 2003, p. 10) -> problème rhédibitoire
- « difficulté à déterminer ce que sont l'« actuel » et le « possible » en matière de mots construits - et, par ricochet, sur la difficulté à donner une valeur chiffrée aux produits actuels et possibles d'une règle de formation de mots donnée -, ainsi que sur les résultats contre-intuitifs obtenus pour les cas extrêmes. » (Dal 2003, p. 12)

approche quantitative (3) : type frequency des bases

- quantité des bases auxquelles peut s'appliquer un procédé donné (Trask 1993) -> contraintes catégorielles
- saturation du domaine de dérivationnel (van Marle 1985)

approche quantitative (4) : hapax legomena et token frequency des produits (Baayen 1992)

1) quantité des produits de fréquence 1 (N1) -> estimation brute de la productivité

2) fréquence cumulée des produits (N) -> pondération par les produits les plus fréquents

=> ratio N1/N -> potential productivity

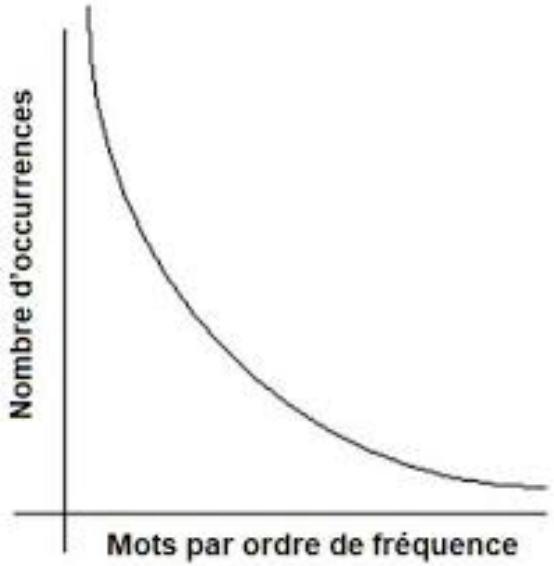
L’indice de Baayen

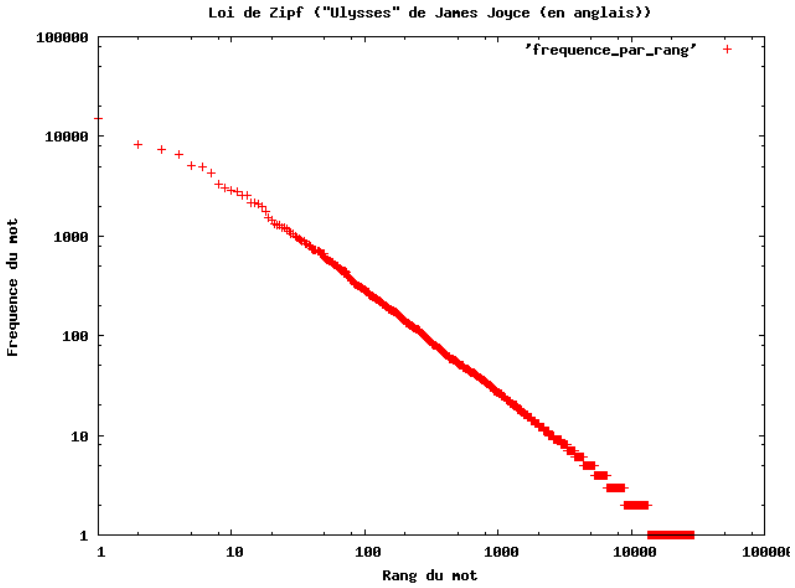
Mesure de productivité et fréquence (Baayen 1992)

- « The notion of morphological productivity has received considerable clarification from the study of the various kinds of restrictions which have been found to condition word formation rules. In a qualitative sense, the productivity of a word formation rule can be said to be inversely proportional to the number of conditioning factors in force (Booij 1977). Nevertheless, the quantitative outcome of the interaction of the – often highly heterogeneous conditioning factors has remained rather obscure.” (p. 110)
- “[…] productivity cannot be simply measured in terms of type frequencies.” (p. 111) -> processus productifs et non-productifs avec même type frequency -> frequency of use (token frequency)

loi de Zipf

- fréquence d’occurrence
- classement décroissant
- loi de puissance : proportion inverse du rang
- F = a * 1/R
- effet de coude





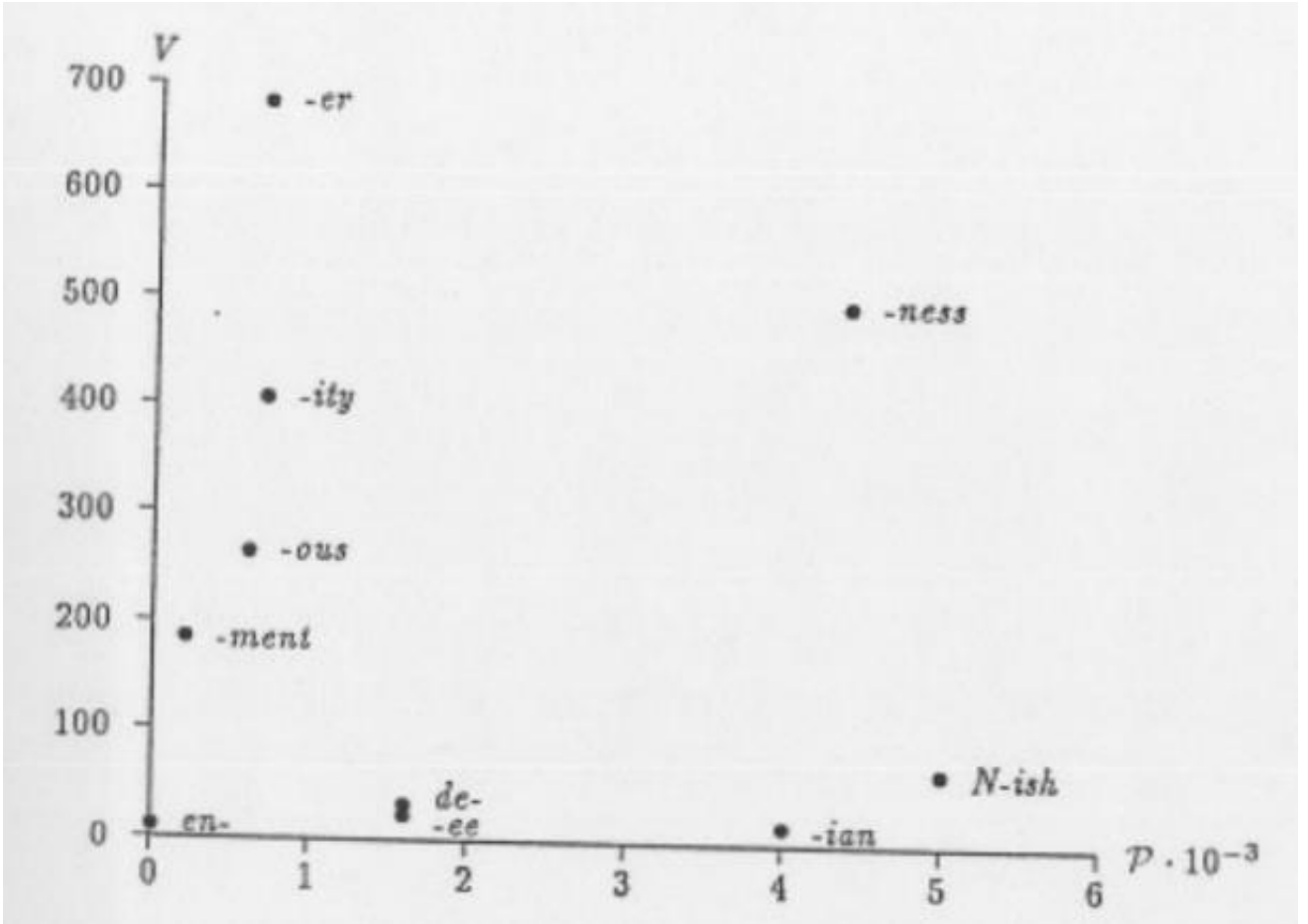
vocabulary growth and productivity

- “growth rate is a measure of likelihood of coming across new types” (p. 113)
- beaucoup plus de mots rares (mais qui apparaissent beaucoup plus rarement)
- les mots rares sont souvent des nouveaux mots -> hapax
- les mots très frequents sont souvent idiosyncratiques -> ponderation
- “probability of encountering a new type not attested before” (Gaeta & Ricca 2006, p. 58)

Indice de Baayen et TAL

- gros corpus -> traitement automatique
- segmentation
- expressions régulières
- étiquetage morpho-syntaxique
- lemmatisation

Critiques de l'indices de l'indice de Baayen (1) : global productivity



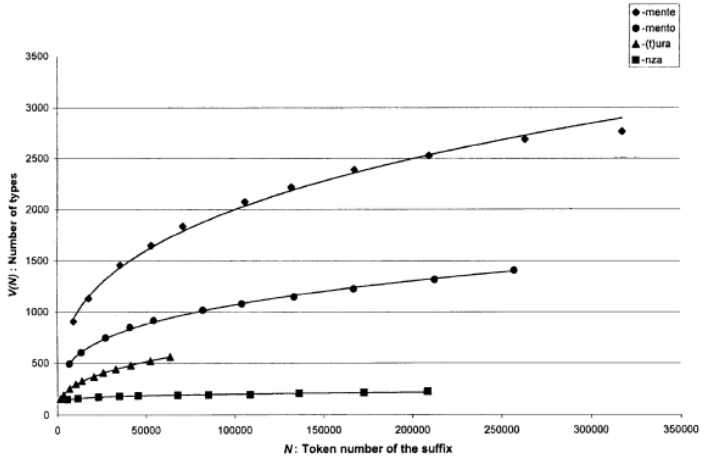
Critiques de l'indice de Baayen (2) : Hay et la fréquence relative

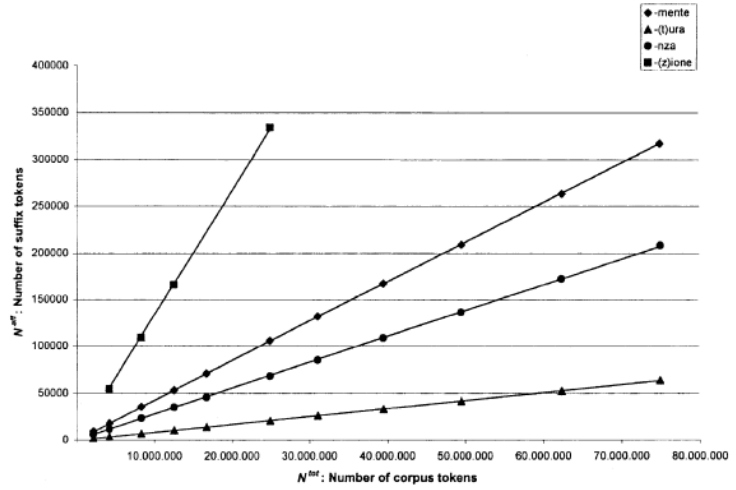
- morphologie = relation Base /dérivé
- effet de la fréquence relative des bases sur l'analysabilité des dérivés
- négligeable?

	dérivé fréquent	**dérivé rare**
base fréquente	?	analysabilité
base rare	non-analysabilité	?

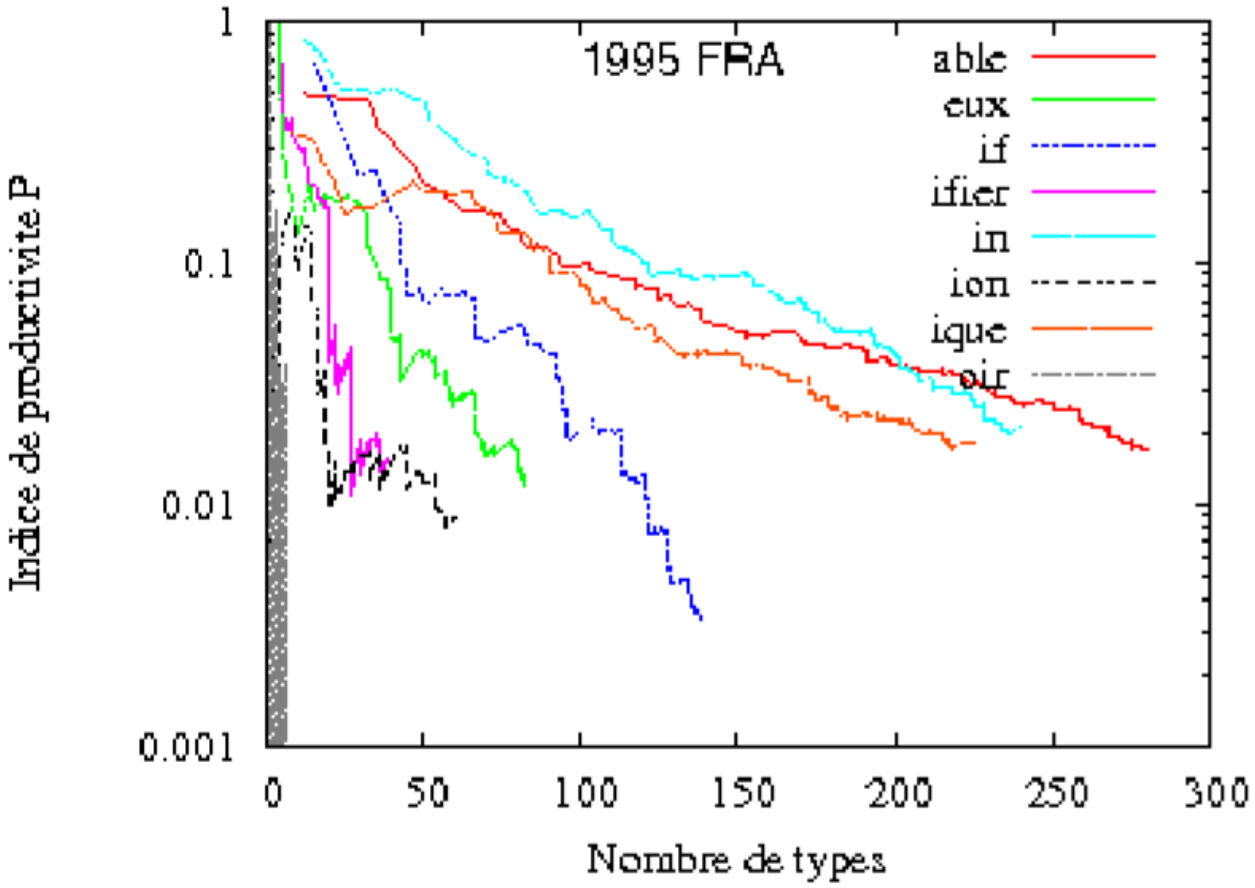
Critiques de l'indice de Baayen (3) : variable-corpus approach (Gaeta & Ricca 2006)

- "[…] in Baayen's procedure […] the index P is always calculated […] referring to the number of tokens N sampled in the whole corpus." (p. 59)
- les différents procédés morphologiques (ex. affixes) ont des fréquences cumulées différentes dans un même corpus
- "the [growth] curves V(N) display different length" (p. 59)
- comparer les indices à fréquence cumulée constante





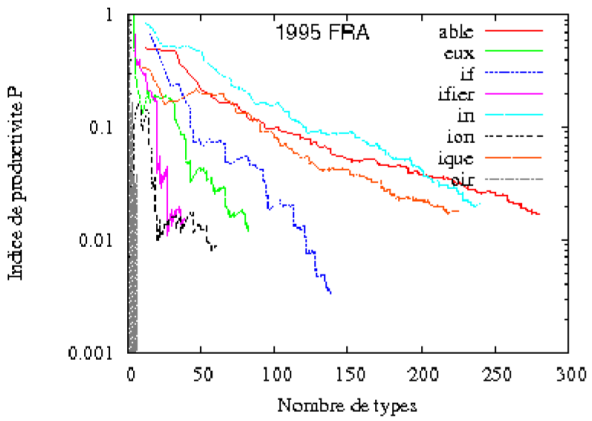
Critiques de l'indice de Baayen (4) : Baayen n'a pas fait centrale



v = d/t
accélération = dérivée de vitesse
delta V / delta T -> différentiel entre 2 points
si v est constante -> a = 0
si a est constante -> n'importe quelle paire de point ($T_n - T_0$)
si a varie -> D_0 est une moyenne

Moi non plus j'ai pas fait centrale (merci Gilles, erreurs sont ma part)

Critiques de l'indice de Baayen (5) : performance (van Marle 1992) et robustesse



able 300
ique 200
if 150
eux 100
ion 50
ifier 50
oir(e) 0

LEXIQUE383films

SUF	nbTyp	productivité
ique_ADJ	911	0,0037
able_ADJ	450	0,0013
if_ADJ	237	0,0012
ion_NOM	1556	0,00066
ifier_VER	72	0,00040
eux_ADJ	350	0,00016
oir(e)_NOM	152	0,00012

PsychoGLàFF_LM10

SUF	nbTyp	productivité
ifier_VER	107	0,010
oir(e)_NOM	138	0,0083
able_ADJ	475	0,0061
ique_ADJ	1052	0,0048
ion_NOM	1132	0,00309
if_ADJ	199	0,00308
eux_ADJ	453	9,7E-05

PsychoGLàFF_FrWaC

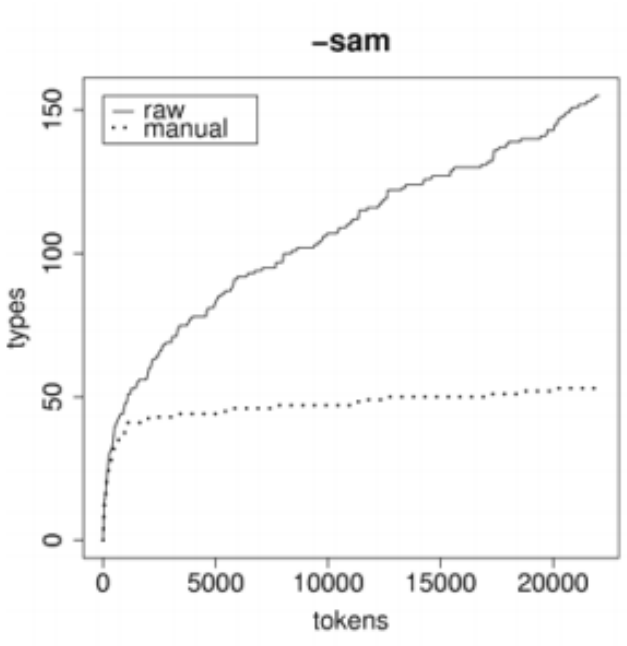
SUF	nbTyp	productivité
ifier_VER	145	0,0034
oir(e)_NOM	240	0,0014
ique_ADJ	1330	0,00079
ion_NOM	1232	0,00062
eux_ADJ	272	0,00053
if_ADJ	212	0,00034
able_ADJ	1034	1,9E-05

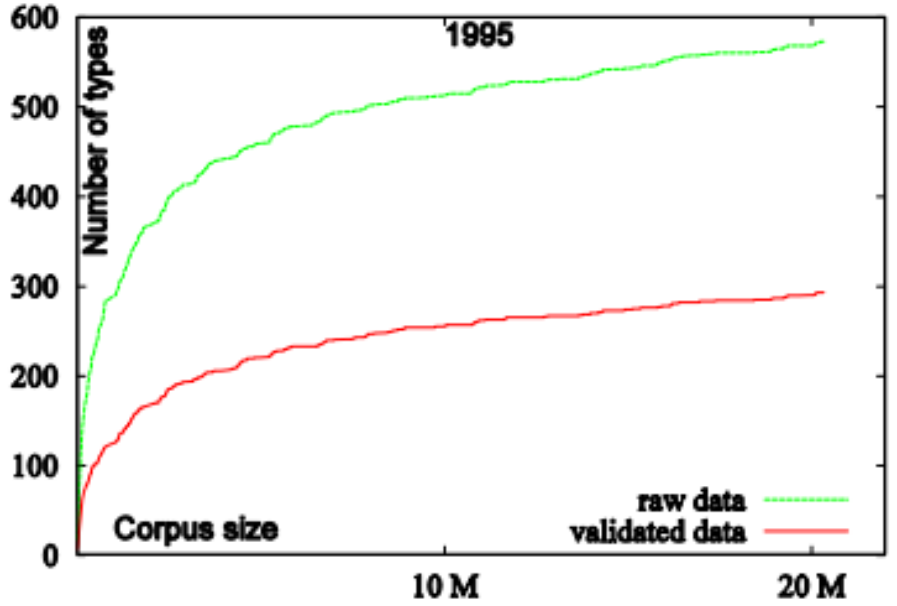
Problèmes:
- sensible à la taille
- inversion des extrèmes
- Valeurs trop petites

LM10 VS frWaC:
- -able/-ique
- -eux

LM10 VS frWaC:
- -able
- -if/-eux

Critiques de l'indice de Baayen (6) : coût (humain) de calcul





Courbes de croissance brute et corrigée manuellement pour la suffixation en -sam en allemand (Lüdeling & Evert, 2001) et pour la suffixation -able en français dans Le Monde 1995 (Grabar & al., 2006).

Calculer l’indice de Baayen avec un lexique fréquencé

Fréquence et lexiques électroniques : frequency norms

- Thorndike and Lorge's (1944) -> « word list counted in books » (18M, Brysbaert&New 2009, p. 3)
- Kuçera et Francis (1967) -> « frequency norms […] of preference […] basis of over 40 years of psycholinguistic and memory research in the US » (1,1M, Brysbaert&New 2009, p. 3)
- TLF (Imbs 1971) -> fréquences dans un corpus de textes littéraires (1919 à 1964, 26 millions de mots) -> FRANTEXT
- BRULEX (Content, Mousty et Radeau 1990) -> 35 746 entrées (Petit Robert) + fréquences TLF
- Novlex (Lambert et Chesnet, 2001) -> corpus spécialisé (textes pour enfants, 417 000 mots)
- Lexique 1 (New et al 2001) -> 130 000 items + fréquences Frantext 1950-2000 (31 millions d'items)
- Lexique 3 (New et al 2007) -> corpus de sous-titrage de films (51M)
- Psychoglaff (Calderone Et al 2014) -> Glaff + 3 normes de fréquence (Frantext, LM10, frWaC)
- Lexique4 (New & Schalchli in progress) -> 190 000 mots, corpus Opensubtitles 360M)

Ressources pour mesurer la productivité?

AVANTAGES

- lexique + fréquence
- pas de bruit
- pas de pré-traitement
- informations morphologiques?

INCONVEVIENTS

- filtrage des hapax
- pas de vision longitudinale

Pourquoi des hapax?

- Le nombre d'hapax est un indice statistique (normalisé par la fréquence cumulée)
- Nombre d'hapax = fréquence cumulée des lexèmes de fréquence 1 (ou de RANG Rmax)

$\Rightarrow$est-ce qu'un autre rang pourrait être significatif (f=2, f=3, …)?

$\Rightarrow$est-ce que la cumulation de plusieurs rangs pourrait être significative (N1+N2+N3 …)?

$\Rightarrow$Jusqu'où remonter?

Productivité et interaction entre fréquence d'occurrence et fréquence de type

- indice de Baayen
  - Nombre d'hapax (numérateur de l'indice) = type frequency = token frequency
  - Global productivity = token frequency X type frequency
  - Fréquence cumulée (dénominateur de l'indice) -> impac des types fréquents
- Corbin 87 -> productivité = rentabilité + disponibilité
- Définition classique (Schultink 1961) -> néologismes = peu fréquents (-> hapax de Baayen)

Idée de génie : séparer/combiner les deux facteurs

- Fréquence de type = compter les lexèmes
- Fréquence d’occurrence = distinguer les lexèmes fréquents des lexèmes rares

=> comparer les lexèmes rares et les lexèmes fréquents

Les arguments psycholinguistiques

- Fréquence et irrégularité (supplétion) des formes fléchies (frequency effect?)
- Dual-route model
- Hay et la fréquence relative
- Fréquence et transparence/compositionnalité sémantique
- Fréquence d'occurrence et acquisition
- Fréquence de type et régularité -> régularisation diachronique, surgénéralisation enfantine
- Fréquence et taille -> zipf et principe de moindre effort
- Family effect
- Ambridge et al 2015

Chercher un seuil

- continuum
- Filtrage des hapax
- Variabilité des corpus

Quartilotracter?!

- Exemple des allomorphies en ion (Bonami, Boyé, Kerleroux 2009)

Classe	Description	Exemple		Effectif
1	**Rad3** ⊕ asjɔ̃	*vexation*	/vɛksasjɔ̃/	1093
2	**Rad3** ⊕ kasjɔ̃	*modification*	/modifikasjɔ̃/	95
3	**Rad3** ⊕ jɔ̃	*dispersion*	/dispɛrsjɔ̃/	86
4	**Rad3** ⊕ isjɔ̃	*composition*	/kɔ̃posisjɔ̃/	33
5	**Rad3** ⊕ sjɔ̃	*pollution*	/polysjɔ̃/	50
6	**X** ⊕ jɔ̃	*abstraction*	/abstraksjɔ̃/	277
7	**Pas de base autonome**	*compétition*	/kɔ̃petisjɔ̃/	474

TABLEAU 6—*Classification de surface des noms en* -ion

- Trois seuils!

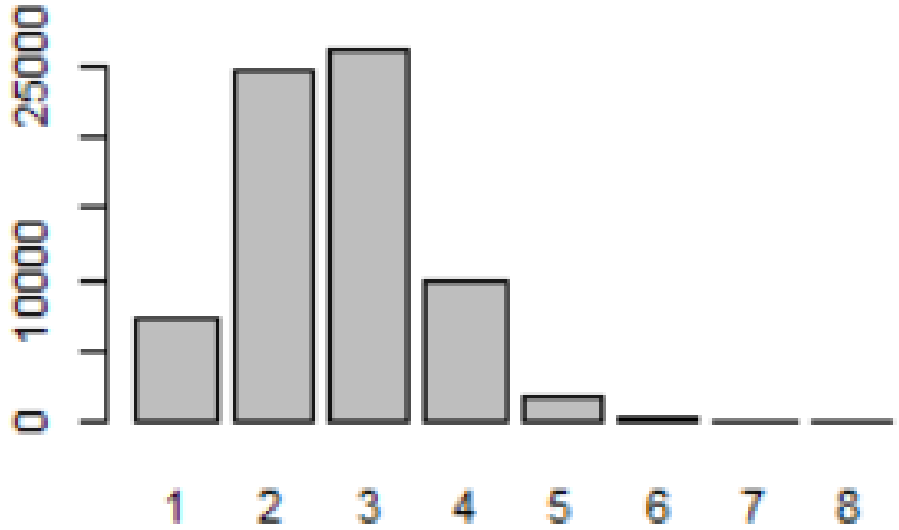
- Schalchli, in progress

Inverse frequency order	PRODUCTIVE CLASSES	UNPRODUCTIVE CLASSES	COMPARISON
1st quartile	25,6%	21,7%	- 4,9%
2nd quartile	23,6%	28,1%	+ 4,5%
3rd quartile	23,3%	24,9%	+ 1,6%
last quartile	27,4%	15,9%	- 11,5%

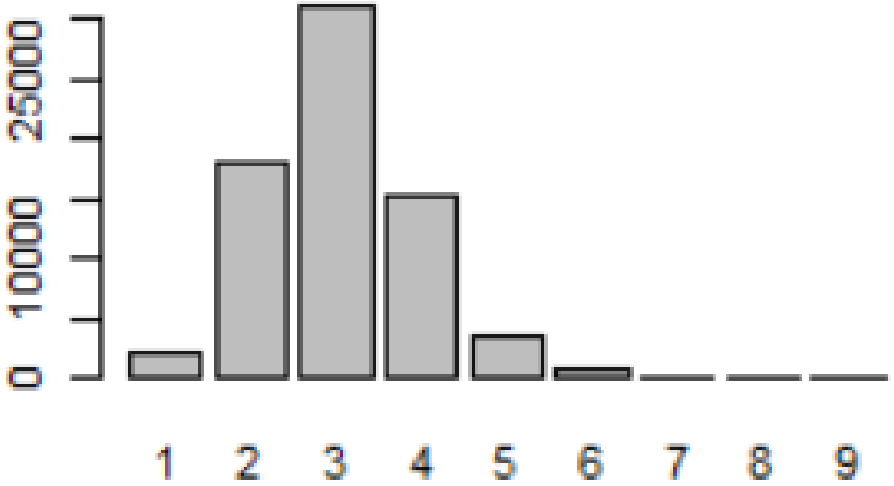
Faire au plus simple (bête?) : la médiane

- 50 % du lexique > médiane > 50% du lexique
- Un seul seuil
- Hypothèse nulle

Hypothèse du seuil médian : Application au Nombre de syllabes (1)

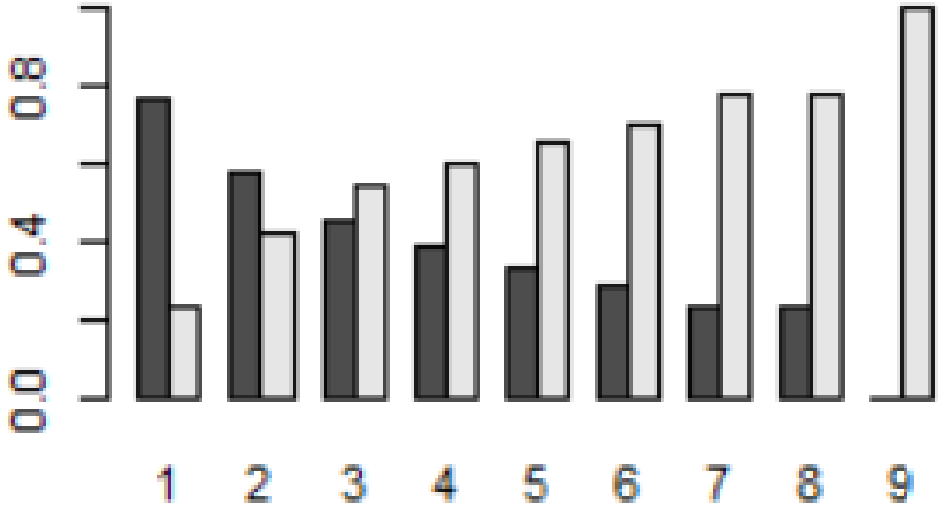


> médiane



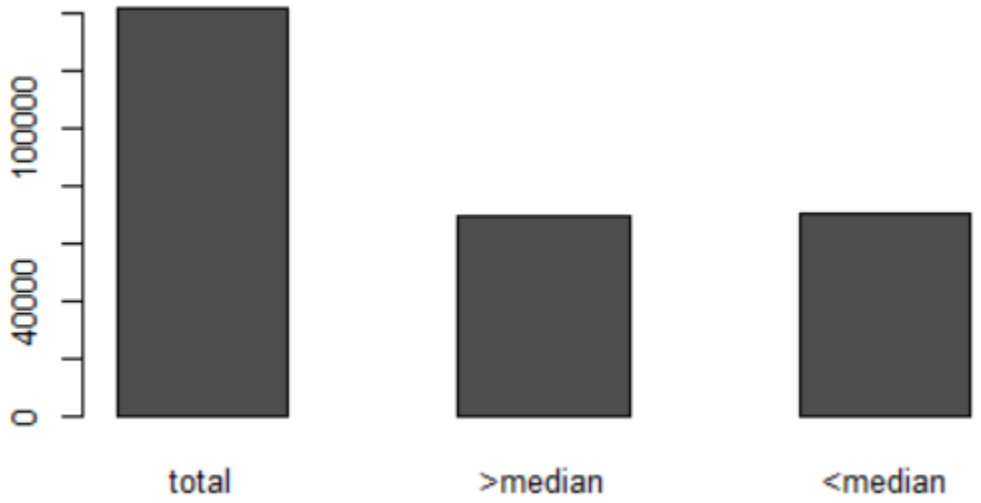
< médiane

Hypothèse du seuil médian : Application au Nombre de syllabes (2)

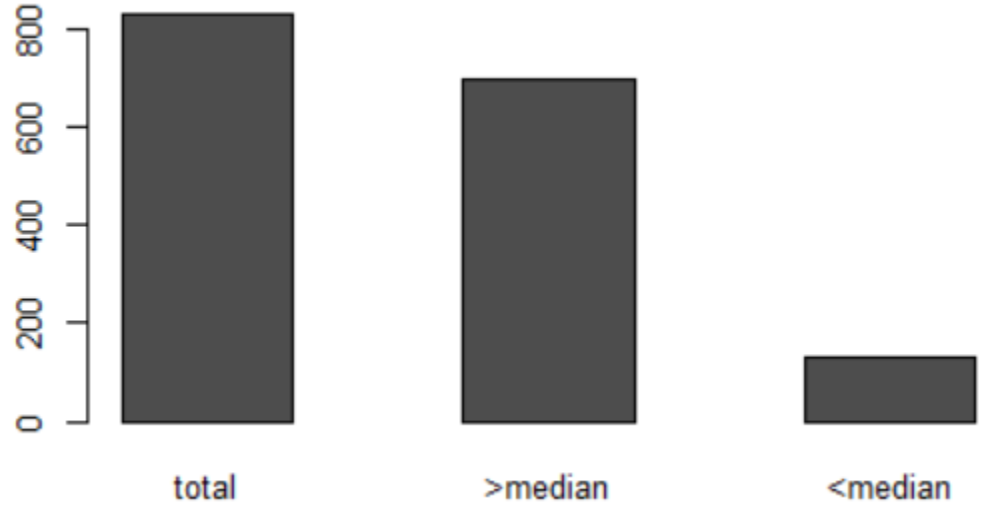


noir : > médiane
blanc : < médiane

Hypothèse du seuil médian : Application aux Catégories : majeures VS mineures

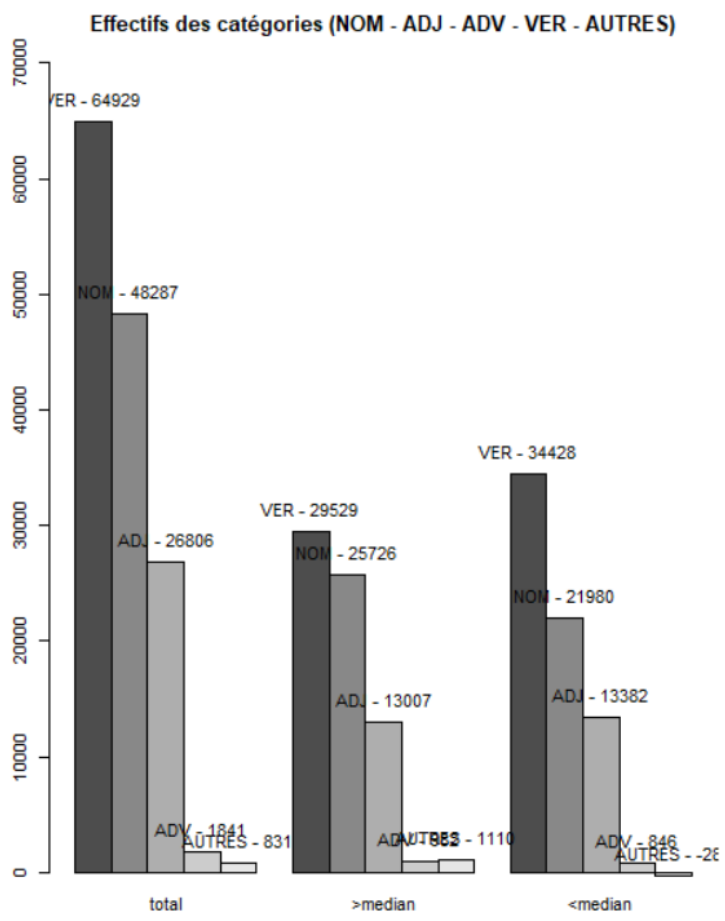


catégories majeures

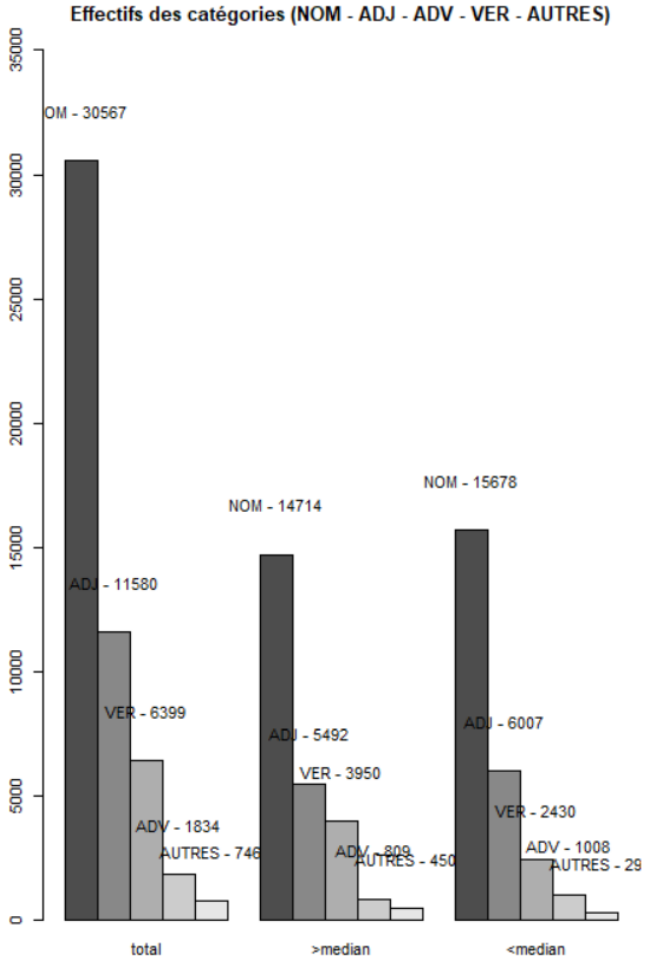


catégories mineures

Catégories – formes fléchies VS lexèmes (1) : totaux

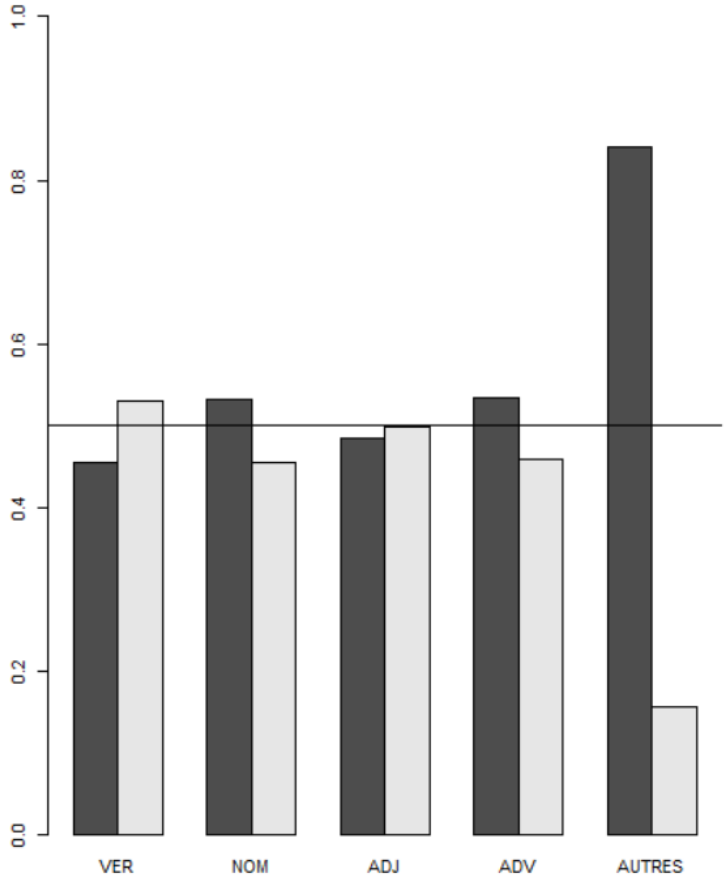


formes fléchies (VER – NOM – ADJ – ADV – autres)

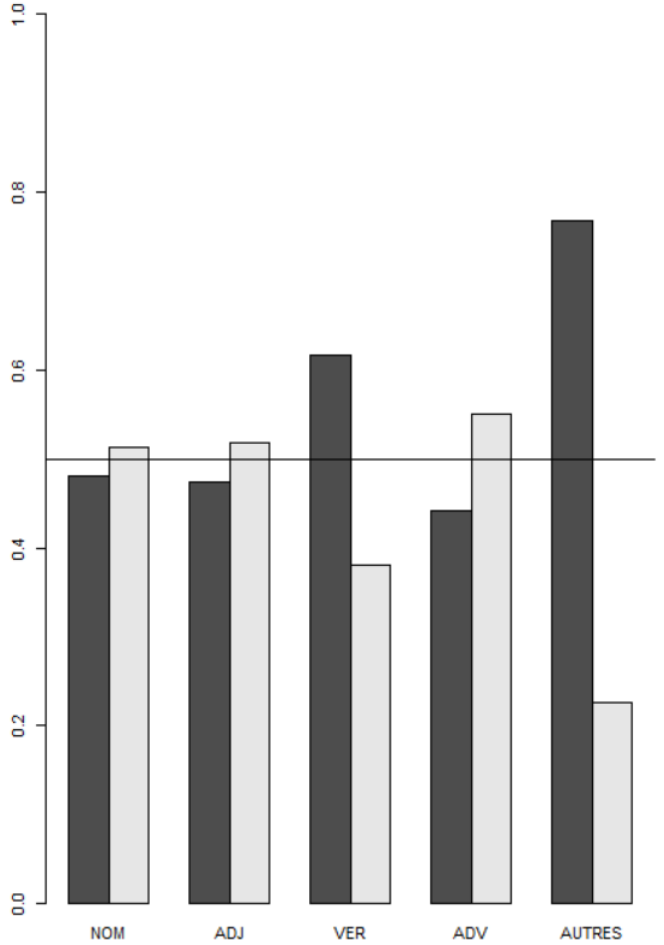


lexèmes

Hypothèse du seuil médian : Application aux Catégories – formes fléchies VS lexèmes : moitiés

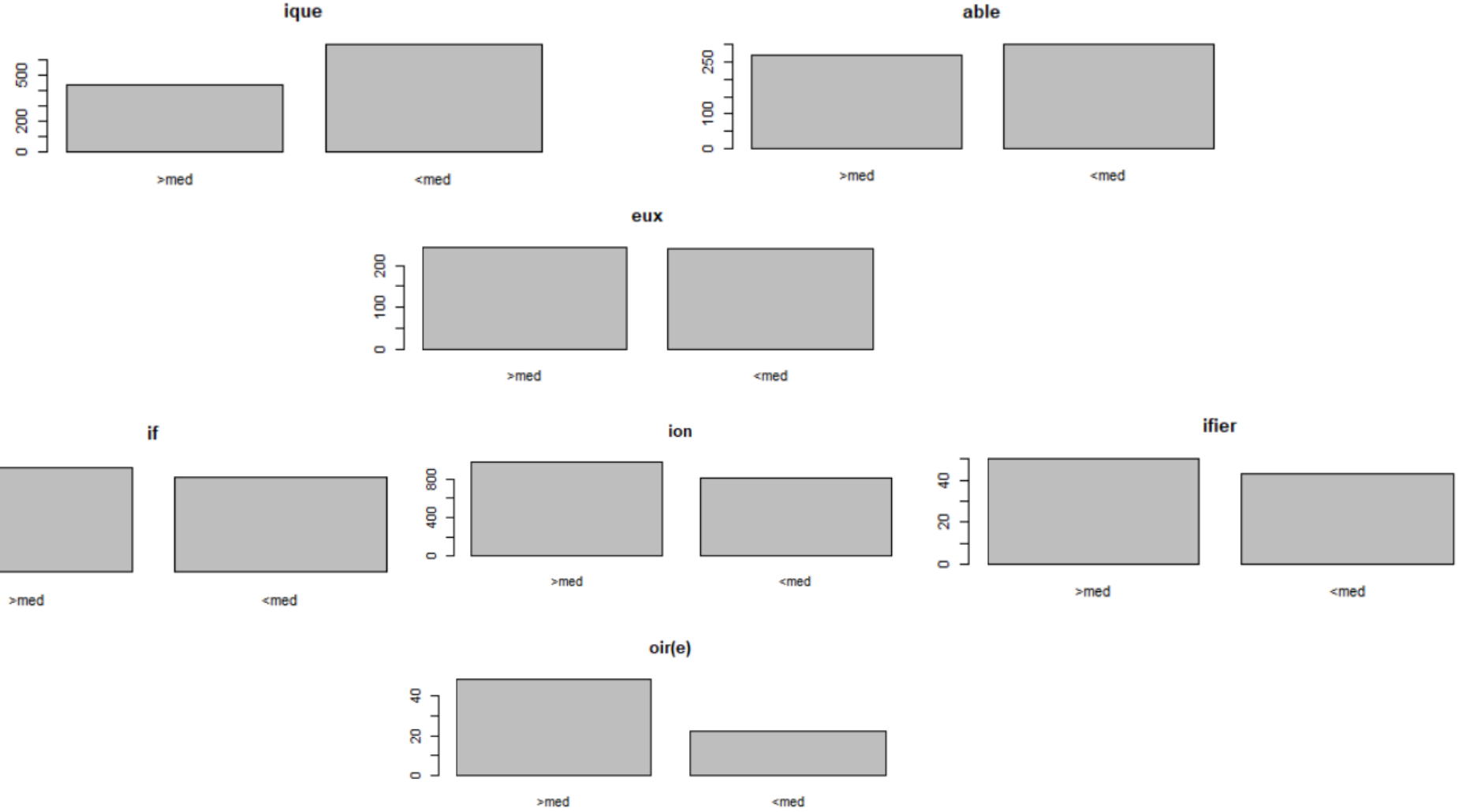


formes fléchies

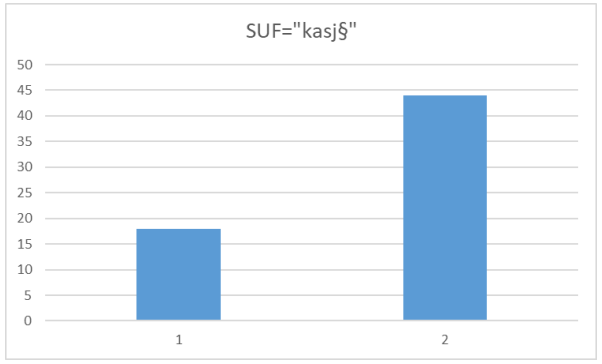


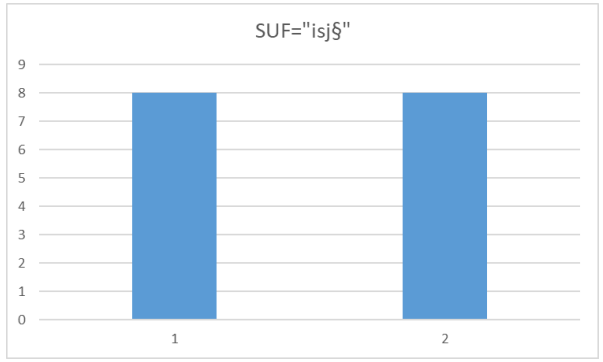
lexèmes

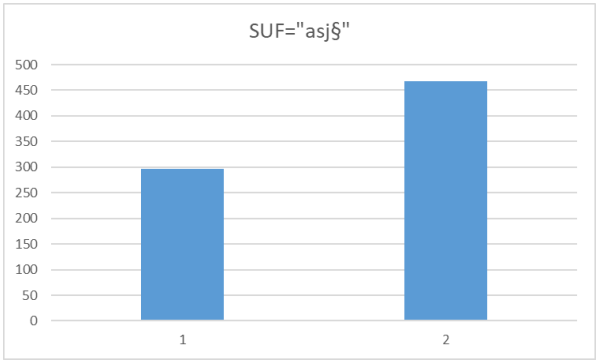
Hypothèse du seuil médian : Application aux Suffixes de Dal et al 2008

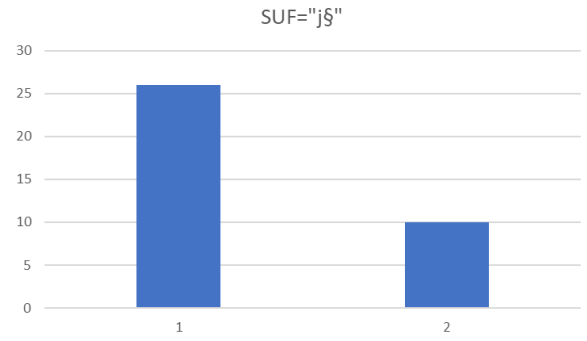


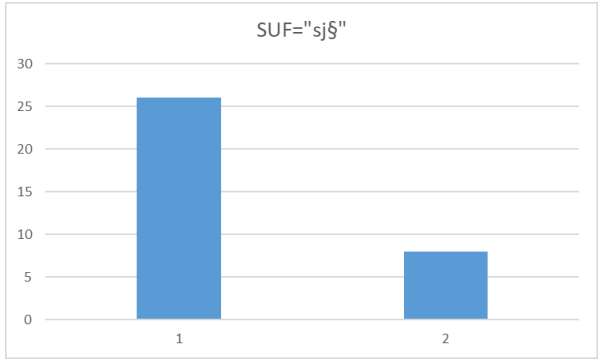
Hypothèse du seuil médian (5) : Application aux données en ion de BoBoKer (2009)



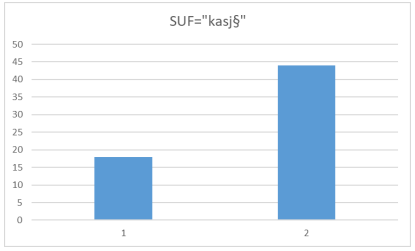


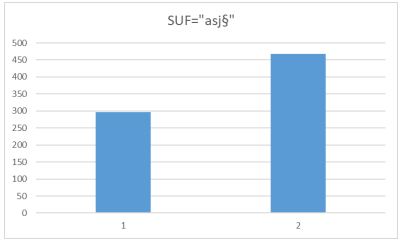


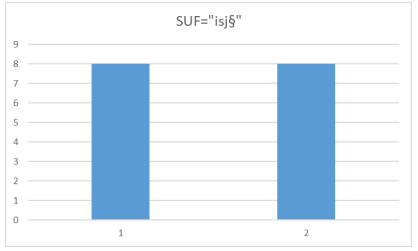


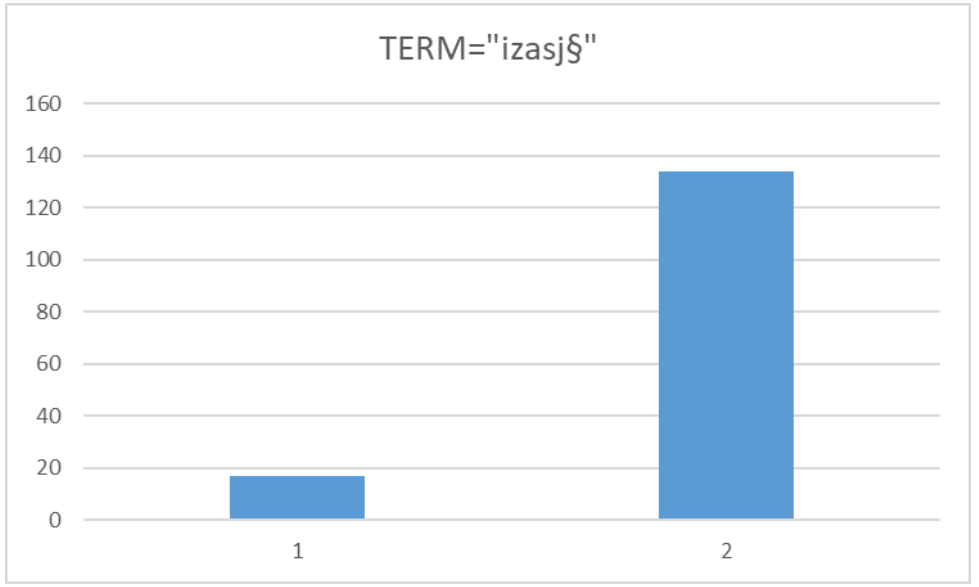


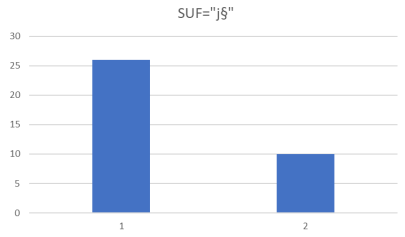
Hypothèse du seuil médian (5) : Application aux données en ion de BoBoKer (2009)

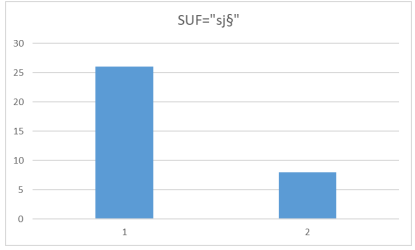












Hypothèse du seuil médian : un nouvel indice ?

- P = V(M2) / V (M1)
- si V(M2) = 0 => P=0
- si V(M2) > 0 => P > 0
- si V(M2) = V(M1) => P = 1 (productivité moyenne)
- si V(M2)< V(M1) => P < 1 (productivité faible)
- si V(M2)> V(M1) => P > 1 (productivité forte)

si V(M1) = 0 => !!! (données pas pertinentes)

Indice du seuil médian : Application aux données en ion de BoBoKer (2009)

TERMINAISON	Effectifs dans M1	Effectifs dans M2	Ratio
SUF="asj§"	297	386	1,2996633
SUF="j§"	26	10	0,38461538
SUF="sj§"	26	8	0,30769231
SUF="kasj§"	18	39	2,16666667
SUF="isj§"	8	6	0,75

TERM="izasj§"	**17**	**134**	**6,41176471**

Conclusion

- une stratégie
- un indice
- des résutats encourageants
- persévérer (généraliser, systématiser, affiner)
- autres domaines (composition, préfixation, …)

Ouverture: un projet TAL

- DERIF (Namer 1) -> un système expert
- Lstat (Namer 2) -> un moteur de productivité à partir de corpus (projet abandonné)
- Lstat + DERIF -> néo DERIF

(Namer 1 + Namer 2) = SuperNamer