Modélisation statistique et morphologie

Olivier Bonami[1] et Juliette Thuilier[2]

[1]Université Paris Diderot, Laboratoire de linguistique formelle
[2]Université Toulouse Jean Jaurès, CLLE-ERSS

Post-décembrettes - 5 décembre 2015

Une observation

- La grande majorité des travaux publiés en morphologie par des chercheurs français
  1. ne comporte aucune indication quantitative ; ou
  2. se contente d'indications impressionnistes ; ou
  3. se contente de statistiques descriptives.
- C'est dommage !
- C'est d'autant plus dommage que la communauté française prend particulièrement au sérieux l'utilisation de données extensives.
- Explications possibles :
  1. Manque de culture mathématique
  2. Culture des distinctions catégoriques
  3. Culture de la modélisation déterministe
  4. Culture de la recherche d'une cause unique

Plan

1. De la morphologie extensive à la morphologie quantitative (Olivier)
2. Une étude type : *-iser* et *-ifier*
   - (i) Constitution des données (Olivier)
   - (ii) Étude des corrélations (Juliette)
   - (iii) Modélisation multivariée (Juliette)
3. Nouvelles méthodes, nouvelles questions (Olivier)

De la morphologie extensive à la morphologie quantitative

Statistique(s)

Une statistique quantité qui résume une série d'observations

- Note moyenne à mon dernier examen de morphologie
- Proportion de francophones natifs inscrits

Statistique(s) descriptive(s) Étude des caractéristiques d'une population fermée à l'aide d'une ou plusieurs statistiques

- Moyenne, médiane, écart-type, quantiles, etc.
- Tableaux de contingence, corrélation

Statistique(s) inférentielle(s) Étude de l'extensibilité des observations sur un échantillon d'une population à la population générale

- Quelle confiance accorder à l'idée que le fait d'être natif aide pour réussir au cours de morphologie ?

Modèle statistique Modèle mathématique à visant reproduire le mécanisme ayant généré un ensemble d'observations

- La note à l'examen comme fonction linéaire de l'age, du sexe, et de la langue maternelle

Observer l'impossible

- Tribout (2010, 310) : « Les noms instrumentaux [sont] inexistants lorsque les noms dérivent du thème 13 »

Th.	Exemple	Nb. instr.	Total	Proportion
0	*rallonge*	17	156	10,9%
12	*conduite*	3	136	2,2%
13	—	0	49	0.0%

- Supposons que l'estimation de la proportion d'instrumentaux est correcte pour les converts du thème 12.
- Si la proportion réelle est la même pour le thème 13, quelle est la probabilité de n'observer aucun exemple d'instrumental ?

$$\underbrace{(1-0,022)\times\cdots\times(1-0,022)}_{\text{49 fois}} = 0,978^{49} \approx 0,34$$

- On ne peut donc exclure que la proportion réelle soit de 2.2%.
- Type de raisonnement qui fonde l'inférence statistique fréquentiste : évaluer la probabilité des observations étant donné une hypothèse.

Observer l'impossible, suite

Th.	Exemple	Nb. instr.	Total	Proportion
0	*rallonge*	17	156	10,9%
12	*conduite*	3	136	2,2%
13	—	0	49	0.0%

- En fait l'estimation de la proportion réelle pour le thème 12 est elle-même en question
  - On sait qu'elle n'est pas nulle.
  - En revanche, on ne sait pas si elle est exactement de 2,2%.
- La question raisonnable est donc : Étant données nos observations, quelle confiance accorder à l'idée que les proportions d'instrumentaux sont différentes pour les converts à partir du thème 12 et du thème 13 ?
- Test exact de Fisher : *p*-valeur de 0,56.
- La probabilité d'observer un décalage de proportions au moins aussi grand dans un échantillon de cette taille si les proportions sont égales dans la population générale est de 0,56.
- En l'occurrence, rien ne permet d'exclure cette possibilité.

L'extensibilité des tendances

- Tribout (2010, 309) : « Il est notable que les noms processifs sont proportionnellement plus nombreux parmi les noms dérivés du thème 12 »

Th.	Exemple	Nb. proc.	Total	Proportion
0	*approche*	75	156	48,1%
12	*découverte*	78	136	57,4%
13	*assassinat*	10	39	20,4%

- Pour décider si l'observation est pertinente, on veut décider si on peut rejeter l'hypothèse selon laquelle les proportions sont identiques.
  - Test exact de Fisher : p-valeur de 0,13.
  - Si les proportions sont identiques dans la population générale, alors il y une chance sur 8 d'observer un décalage au moins aussi grand de proportions dans un échantillon de cette taille.

L'extensibilité des tendances, suite

- Qu'en est-il de la différence de proportions entre les processifs dérivés du thème 12 et les autres ?

Th.	Exemple	Nb. proc.	Total	Proportion
0 ou 12	*approche*	153	292	52,4%
13	*assassinat*	10	49	20,4%

- Test exact de Fisher : *p*-valeur de 0,000035.
- Si les proportions sont identiques dans la population générale, alors il y moins d'une chance sur 10000 d'observer un décalage au moins aussi grand de proportions dans un échantillon de cette taille.
- NB : La *p*-valeur dépend simultanément :
  - de la taille des échantillons
  - de l'écart entre les proportions observées
  - du nombre de situations distinguées
- Le fait d'avoir réuni beaucoup de données aide, mais ne garantit pas qu'on va obtenir des différences statistiquement significatives.

Identifier l'anormal

- La description morphologique utilise fréquemment une notion d'irrégularité :
  - ☞ « Il y a dans nos listes 452 cas où un verbe possède au moins deux des trois dérivés concernés. Dans 98% des exemples [=443 cas] la même séquence précède *-ion*, *-eur* ou *-if*. » (Bonami et al., 2009, 116)
  - ▸ Interprétation de Bonami *et al.* : ces 2% d'exceptions sont des formations irrégulières considérées comme morphologiquement inanalysables.
- Interprétation possible de l'irrégularité : les hasards de l'histoire introduisent de la variation aléatoire dans le système qui fait que l'impossible régulier devient de l'improbable.
- Le problème est de se mettre d'accord sur la quantité de variation aléatoire, et de prendre en compte cette variabilité dans l'interprétation des observations.

☞ Noter la proximité entre les chiffres de Tribout et ceux de Bonami *et al.*

L'étude des corrélations

- Plénat and Roché (2003) : la proportion de troncation de la voyelle finale avant le suffixe *-esque* varie en fonction de la longueur de la base.

σ	tronqués	total	prop.
1	0	3	0%
2	32	92	35%
3	122	154	79%
4+	70	84	83%

- On voudrait savoir, au vu de ces chiffres :
  - Quelle est l'ampleur de l'influence de la longueur de la base ?
  - Quel est la nature de la fonction reliant les deux grandeurs ?
  - À quel point est-on assuré de la réalité du phénomène ?
- La régression logistique est l'outil de choix pour étudier cette question :
  - Permet de trouver une fonction optimale qui lie la probabilité qu'une propriété soit vérifiée à une ou plusieurs variables explicatives.

L'étude des corrélations - suite

- Plénat and Roché (2003) notent l'influence du choix de la voyelle.

	i	y	u	e	ø	o	a
tronqués	46	1	4	15	0	73	85
total	91	13	16	20	0	91	102
proportion	51%	8%	25%	75%	—	80%	83%

- Les questions habituelles se posent sur la significativité des différences.
- De plus, corrélation entre les deux facteurs :

	i	y	u	e	ø	o	a
long. moy.	3,0	2,7	2,4	2,8	—	3,1	3,0

- Les deux facteurs sont ils réellement *tous deux* pertinents ?

☞ Régression logistique multiple

  - Évaluation de l'effet de chaque facteur en présence de l'autre.

Une étude type : *-iser* et *-ifier*

- Nous avons fait le choix d'illustrer *une* technique de modélisation statistique sur la base d'*une* phénoménologie familière : rivalité entre deux procédés constructionnels.
- Choix de *-iser* vs. *-ifier* :
  - Études précédentes de Lignon (2013) et Namer (2013) qui nous donnent plusieurs hypothèses à tester.
  - Plus facile de travailler sur un choix binaire.
- Il ne faut pas perdre de vue le fait que dans la réalité ces deux procédés s'inscrivent dans un système plus large.

Procédé	Base A	Base N
a-	*agrandir*	*aligner*
é-	*élargir*	*effiler*
en-	*embellir*	*enflammer*
conversion	*mûrir*	*liasser*
-iser	*banaliser*	*canaliser*
-ifier	*densifier*	*momifier*

Questions de recherche

1. La catégorie syntaxique de la base influence-t-elle le choix du suffixe ?
2. Les propriétés phonologiques de la base influencent-elles le choix du suffixe ?
   - Longueur
   - Structure syllabique
   - Identité segmentale
3. Les propriétés morphologiques de la base influencent-elles le choix du suffixe ?
   - Nature de la dernière opération constructionnelle
   - Composition de la famille morphologique
4. Les propriétés syntaxiques et sémantiques de la base influencent-elles le choix du suffixe ?
5. Les propriétés syntaxiques et sémantiques souhaitées pour le dérivé influencent-elles le choix du suffixe ?
6. Les facteurs influençant le choix du suffixe sont-ils stables dans le temps ?

Une étude type : *-iser* et *-ifier*

Constitution des données

Les données sélectionnées

Intersection des verbes documentés dans le Glàff (Hathout et al., 2014) et dans Google ngrams (Michel et al., 2010).

- Glàff fournit une validation humaine de l'utilisabilité du mot et une transcription phonémique
- Google ngrams fournit une validation de l'attestation du mot, une information sur sa fréquence, et une estimation de sa date d'apparition.

Source	taille
Frwac	4796
Glàff	3213
Google ngrams	2660
Lexique	395
Glàff ∩ Gng	**1263**

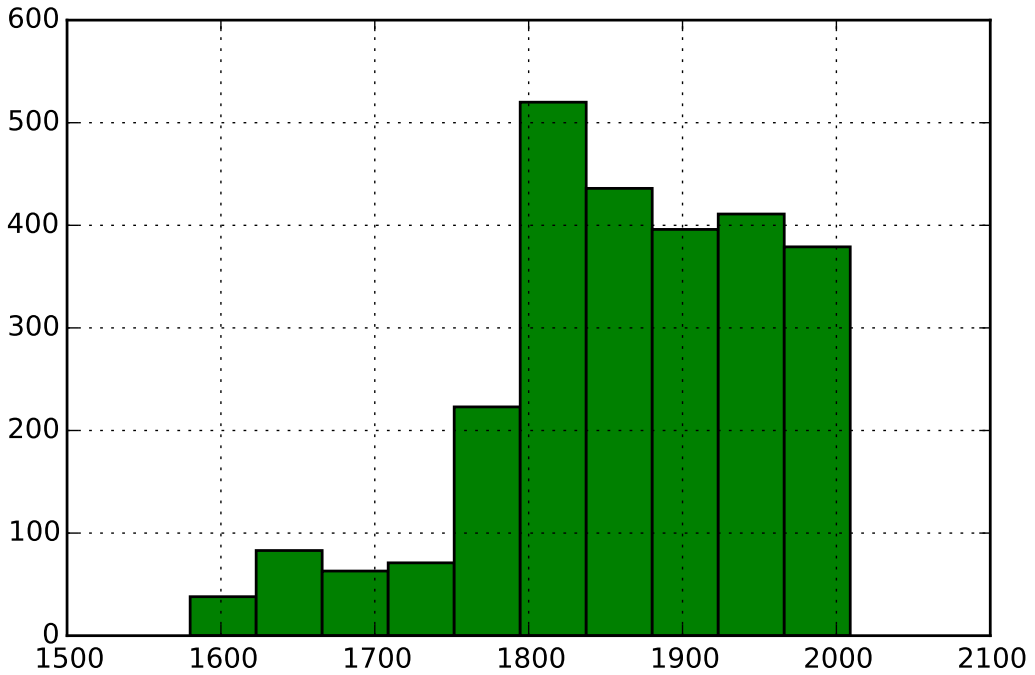
Tri manuel

- On ne veut que des exemples dont l'analysabilité en français comme des verbes formés par suffixation ne fait pas de doute.
  - Élimination de tous les verbes préfixés
    réunifier, surnaturaliser, surcapitaliser, immortaliser, etc.
  - Élimination de tous les verbes non-suffixés
    baliser, briser, rectifier, etc.
  - Élimitation de tous les verbes sans base française indiscutable
    lignifier, pulvériser, randomiser, etc.

Source	taille
Glàff ∩ Gng	1263
Après filtrage	**795**
Lignon 2013 TLF	*940*
Lignon 2013 web	*1564*

Datation

- Google n-grams comme source de dates d'attestation des mots étudiés.
- Date choisie : centre de la première période de 10 ans pendant laquelle le mot est attesté 10 fois.
  - Évite les faux positifs dus à la qualité de l'OCR.
  - Fournit une date d'utilisation soutenue, pas une attestation isolée.



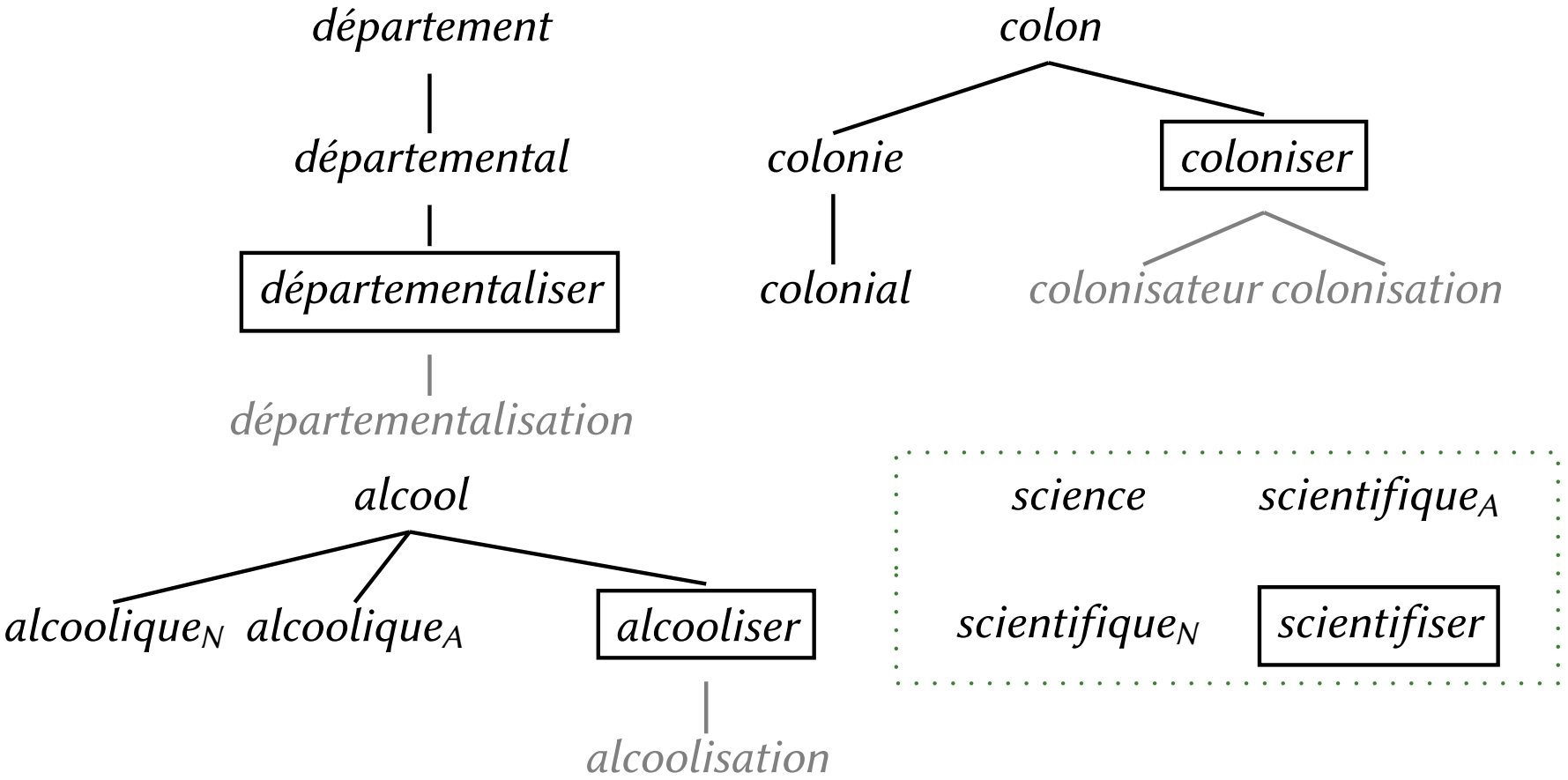
Répartition des dates d'attestation des mots en *-iser* et *ifier* dans Google ngrams

Catégorie de la base

- On s'attend à une influence de la catégorie de la base sur le choix du suffixe
  - ☞ Intuition : toutes choses égales par ailleurs, *-ifier* préfère les noms
- Problème : il n'y a généralement pas de base unique facilement identifiable
  - Souvent nom et adjectif sont tous deux de bons candidats :
    - ★ *département*, *départemental* ⇝ départementaliser
    - ★ *colonie*, *colonial* ⇝ coloniser
  - Souvent la forme ne permet pas de départager les deux hypothèses :
    - ★ *cannibale*, *cannibale* ⇝ cannibaliser
    - ★ *calcium*, *calcique* ⇝ calcifier
    - ★ *idéologie*, *idéologique* ⇝ idéologiser
- On sait (Namer, 2013) que l'examen de la forme n'est pas suffisant pour déterminer une base au sens classique.
- Cependant une classification fine à base sémantique
  - est très coûteuse
  - est peu reproductible
  - comporte des risques de biais incontrôlés

La famille morphologique ascendante

- Ordre partiel structurant les familles morphologiques :
  Y est un descendant de *X* si le radical de *Y* comporte des traces explicites de tous les morphes du radical de *X*.
- La famille morphologique ascendante exclut ses descendants.



Codage des propriétés de la famille morphologique

- En l'absence de certitudes sur les bases (Lignon et al., 2014 ; Strnadová, 2014), les propriétés de la famille morphologique ascendante prennent la place des « propriétés de la base ».
- Notre annotation :
  1. Présence d'au moins un adjectif dans la FMA
  2. Présence d'au moins un nom dans la FMA qui ne descend pas d'un adjectif

Classe	exemple	Effectif
N seulement	*avaliser*	116
Adj seulement	*béatifier*	60
Les deux	*alcooliser*	619

- NB :
  - Une annotation de la catégorie de l'ascendant immédiat donne des résultats contre-intuitifs.
  - Une annotation qui cherche le meilleur candidat à être la base d'un point de vue formel laisse encore 341 cas indécidables

Classe morphologique de l'adjectif

- Lignon (2013) suggère que, quand le verbe est dérivé d'un adjectif, la classe morphologique de l'adjectif a une incidence sur le choix de l'affixe.
- Pour prendre en compte cette information, nous nous sommes appuyés sur la base d'adjectifs construits Dénom (Strnadová, 2014).

☞ Codage du suffixe éventuel de l'adjectif le plus proche dans la FMA.

Verbe	Adj. proche	Classe
départementaliser	*départemental*	*-al*
coloniser	*colonial*	*-al*
communiser	*communiste*	*-iste*
britanniser	*britannique*	*-ique*
scientifiser	*scientifique*	*-ique*
alcooliser	*alcoolique*	*-ique*
avaliser	—	—
béatifier	*béat*	—

Propriétés phonologiques du radical

- On s'attend à une influence de la phonologie de la base sur le choix du suffixe (Lignon, 2013)
- Dans la majorité de nos exemples, il n'y a pas une unique forme qui est candidate à être le radical de la base :

Classe	exemple	Effectif
N seulement	*avaliser*	116
Adj seulement	*béatifier*	60
Les deux, N=Adj	*cannibaliser*	136
Les deux, N≠Adj	*alcooliser*	483

- Deux solutions possibles :
  1. Limiter l'étude aux 312 cas où un radical unique est identifiable.
  2. Étudier les propriétés phonologiques du radical d'affixation tel qu'il se présente dans le dérivé.
- On adopte la solution 2.
- Avantage : indépendance de l'étude de l'influence de la phonologie et de celle de la famille morphologique.

Propriétés phonologiques du radical

- On travaille à partir des transcriptions phonémiques syllabées du GLÀFF.
- Propriétés codées :
  1. Longueur du mot en syllabes
  2. Identité de la dernière voyelle du radical
  3. Identité de la séquence de consonnes s'il y en a une

Lexème	Radical	Long.	Voy.	Cons
athéiser	a.te	2	e	NULL
choséifier	ʃo.ze	2	e	NULL
absolutiser	ap.sɔ.ly.t	3	y	.t
centraliser	sɑ̃.tʁa.l	2	a	.l
coloniser	kɔ.lɔ.n	2	ɔ	.n
algébriser	al.ʒe.bʁ	2	e	.bʁ
anarchiser	a.naʁ.ʃ	2	a	ʁ.ʃ

- Noter les décalages avec les propriétés phonologiques des lexèmes candidats à être base.

Une étude type : *-iser* et *-ifier*

Étude des corrélations

La table de données

- Combien de *-iser* et *-ifier* dans nos données ?

	-iser	*-ifier*	Total
Nbre de vbs	703	92	795
Prop.	88.4%	11.6%	100%

  - 9 doublons : *turquifier*/*turquiser* ; *électrifier* / *électriser* ; *estérifier* / *estériser* ; *étanchéifier*/*étanchéiser* ; *éthérifier* / *éthériser* ; *fluidifier* / *fluidiser* ; *nanifier* / *naniser* ; *terrorifier* / *terroriser* ; *typifier* / *typiser*
  - cf. données de Lignon (2013) : TLFi = 16% de *-ifier* (12 doublons) ; Web = 21.4% de *-ifier* (258 doublons)

- Comment se présentent les données ?

Lexème	IFIER	Long.	Voy.	Cons	Date	FMA	Classe
athéiser	False	2	autre	autre	1801	both	non-suff
choséifier	True	2	autre	autre	1960	N	pas-adj
absolutiser	False	3	fermée	alvObs	1905	both	non-suff
centraliser	False	2	autre	son.	1790	both	-al
densifier	True	1	nas.	alvObs	1841	N	non-suff

Age du dérivé

- Dates d'attestation vont de 1580 jusqu'aux années 2000

1580 *favoriser*, *fortifier*, *justifier*, *réaliser*, *sacrifier*, *spécifier*

2000et+ *stendhaliser*, *galliciser*, *chimériser*, *métropoliser*

  - On s'attend à ce que la productivité de *-ifier* diminue avec le temps (cf. Plag 1999 sur l'anglais)

Proportion de verbes en *-ifier* et *-iser* en fonction de la date estimée de première attestation.

  - Corrélation significative entre la date d'attestation et le choix du suffixe (p-valeur $< 0,0001$)

Propriétés phonologiques

1 Longueur du radical en syllabes

- D'après Lignon (2013), le suffixe *-ifier* attire massivement les bases monosyllabiques en raison de la différence de structure segmentale des deux suffixes : « The segmental structure of *-ifier* can explain this remarkable attraction for monosyllabic bases. Indeed, -ifier adjunction (without inflectional endings) increases the base length by two syllables when the base ends in a consonant (the majority of cases) [...] In contrast, *-iser* constructs derived verbs whose length is one syllable longer for consonant-final bases » (Lignon, 2013, 116)
- Confirmation des observations de Lignon (2013)

★ le suffixe *-ifier* a tendance à apparaître majoritairement avec des radicaux monosyllabiques
★ le suffixe *-iser* est proportionnellement beaucoup moins fréquent avec un radical monosyllabique
★ Corrélation statistiquement significative (p-valeur $< 0,0001$)

	1 syll	2 syll	+de 2 syll
-iser	38	351	315
	38,8%	91,9%	99,6%
-ifier	60	31	1
	61,2%	8,1%	0,4%
Totaux	98	382	316
	100%	100%	100%

Propriétés phonologiques

2 Dernière consonne du radical

- On s'attend à ce que les alvéolaires obstruantes favorisent le suffixe *-ifier* → cf. contrainte de dissimilation (Lignon, 2013 ; Plénat, 2000)
- Lignon (2013) observe que les phonèmes /l/, /ʁ/ et /n/ sont particulièrement fréquents avec le suffixe *-iser*

⇒ 3 catégories

(i) Alvéolaire obstruante [t, d, s, z]
(ii) Sonante [l, m, n, ɲ, ʁ]
(iii) Autres

- Confirmation des observations de Lignon (2013)

★ Corrélation statistiquement significative ($\chi^2 = 42,16$; df $= 2$; p-valeur $< 0,0001$) : il y a moins de 0,01% de chance que l'on observe une telle répartition des données si les deux variables étaient indépendantes

	Alv. Obs	Sonantes	Autres
-iser	131 76,2%	520 93,2%	52 80%
-ifier	41 23,8%	38 6,8%	13 20%
Totaux	172 100%	534 100%	96 100%

Propriétés phonologiques

- Dernière consonne du radical semble avoir un effet sur le choix du suffixe : est-ce que d'autres propriétés phonologiques interviennent ?

3 Groupe consonantique en fin de radical

- Variable binaire
  - (i) GroupeCons = True, lorsque il y a au moins 2 consonnes en fin de radical
  - (ii) GroupeCons = False, lorsque il y a 0 ou 1 consonne en fin de radical
- Observation : la proportion de suffixes *-ifier* augmente lorsqu'il y a un groupe consonantique en fin de radical

	False	True
-iser	659 90,5%	44 65,7%
-ifier	69 9,5%	23 34,3%
Totaux	728 100%	67 100%

★ NB : cela concerne à la fois les radicaux se terminant par une attaque branchante (*am.plifier*, *sa.crifier*) ou par une coda suivie d'une attaque simple (*giscar.diser*. *mor.tifier*)

★ Corrélation statistiquement significative ($\chi^2 = 34,6$; df $= 1$; p-valeur $< 0,0001$)

Propriétés phonologiques

4 Dernière voyelle du radical

- 3 catégories
  - (i) voyelle fermée [i,y,u]
  - (ii) voyelle nasale [ɑ̃, ɛ̃, ɔ̃]
  - (iii) autres
- D'après la contrainte de dissimilation, on s'attend à ce que les voyelles fermées défavorisent le suffixe *-ifier*

	Nasales	Fermées	Autres
-iser	15	122	566
	55,6%	80,8%	91,7%
-ifier	12	29	51
	44,4%	19,2%	8,3%
Totaux	27	151	617
	100%	100%	100%

Fermée Observation contraire à la contrainte de dissimilation : la présence d'une voyelle fermée semble favoriser *-ifier*

Nasale Observation surprenante : la présence d'une nasale favorise fortement *-ifier*, explication ?

- ★ pas de corrélation apparente avec la présence d'un suffixe particulier
- ★ peu de données, à explorer

Propriétés morphologiques

1 Propriétés de la famille morphologique ascendante

- 3 catégories
  - (i) N = seulement un nom dans la FMA
  - (ii) Adj = seulement un adjectif dans la FMA
  - (iii) Both = un N et un Adj dans la FMA

Classe	N	Adj	Both
-iser	95	49	559
	82,6%	81,7%	90,2%
-ifier	20	11	61
	17,4%	18,3%	9,8%
Totaux	115	60	620
	100%	100%	100%

- D'après Lignon (2013), on s'attend à ce que les bases adjectivales favorisent *-iser* au détriment de *-ifier*
- Pour les cas où on est sûrs d'avoir une seule base (N ou Adj), pas de différence observable (p-valeur $= 1$)
- Cela suggère qu'il est préférable de raisonner en termes de famille morphologique

Propriétés morphologiques

1 Propriétés de la famille morphologique ascendante

Classe	N	Adj	Both
-iser	95	49	559
	82,6%	81,7%	90,2%
-ifier	20	11	61
	17,4%	18,3%	9,8%
Totaux	115	60	620
	100%	100%	100%

- La présence d'un adjectif seul ou d'un nom seul dans la FMA favorise le suffixe *-ifier*
- Corrélation statistiquement significative ($\chi^2 = 8,3063$; df $= 2$; *p*-valeur $< 0,02$)
- Cela suggère que la configuration de la famille morphologique a une influence sur le choix du suffixe et met en question l'idée qu'on puisse n'étudier qu'une sous-partie des verbes en *-iser* et *ifier*

Propriétés morphologiques

2 Classe morphologique de l'adjectif

- Regroupés en 8 catégories

(i) -aire
(ii) - al
(iii) - el
(iv) -ien
(v) -ique
(vi) autres : autres suffixes dénominaux
(vii) non-suff. : base adjectivale non suffixée
(viii) pas-adj : pas d'adjectif dans la FMA

- Distribution semblable lorsque l'adjectif n'est pas suffixé et lorsqu'il n'y a pas d'adjectif dans la FMA du verbe (p-valeur $= 0,7$)

	-aire	*-al*	*-el*	*-ien*	*-ique*	autre	non-suff	pas-adj
-iser	39 92,9%	82 98,8%	55 100%	21 91,3%	167 96%	76 80.9%	168 80.4%	95 82.6%
-ifier	3 7,1%	1 1,2%	0 0%	2 8,7%	7 4%	18 19.1%	41 19.6%	20 17.4%
Totaux	42 100%	83 100%	55 100%	23 100%	174 100%	94 100%	209 100%	115 100%

Propriétés morphologiques

2 Classe morphologique de l'adjectif

- Hypothèse de Lignon (2013, 116) : « one can suggest that *-iser* preferentially selects denominal adjectives »
- Argument en faveur de cette hypothèse : présence d'un adjectif dénominal en *-aire*, *-al*, *-el*, *-ien*, *-ique* dans la FMA favorise le suffixe *-iser*

	-aire	*-al*	*-el*	*-ien*	*-ique*	autre	non-suff	pas-adj
-iser	39	82	55	21	167	76	168	95
	92,9%	98,8%	100%	91,3%	96%	80.9%	80.4%	82.6%
-ifier	3	1	0	2	7	18	41	20
	7,1%	1,2%	0%	8,7%	4%	19.1%	19.6%	17.4%
Totaux	42	83	55	23	174	94	209	115
	100%	100%	100%	100%	100%	100%	100%	100%

- A noter : sur les 521 verbes en *-iser* avec une sonante en fin de radical, 201 ont des radicaux issus de bases adjectivales dénominales

$\Rightarrow$ La corrélation entre la présence de sonantes en fin de radical et la présence du suffixe *-iser* n'est qu'un effet collatéral de la préférence du suffixe *-iser* pour les bases dénominales (cf. Lignon 2013)

Une étude type : *-iser* et *-ifier*

Modélisation statistique multivariée

Modélisation statistique multivariée

- compétition entre les deux procédés morphologiques régulée par un ensemble de facteurs non-déterministes (phonologie, morphologie et évolution diachronique du système)
- on souhaiterait savoir comment les facteurs dont nous avons décrit le comportement individuel agissent conjointement et quel est le rôle de chacun, une fois pris en compte le rôle de tous les autres.

→ Modélisation mutivariée : régression logistique

- permet de modéliser le comportement d'une variable binaire, en fonction de plusieurs variables prédictrices
- permet de prendre en compte simultanément l'ensemble des facteurs
- permet d'inférer les propriétés de la population (ici, le lexique) à partir de l'échantillon que constitue la table de données (statistique inférentielle)

Régression logistique

- Modéliser le comportement d'une variable binaire, en fonction de plusieurs variables prédictrices
  - variable binaire : suffixe *-iser* ou suffixe *-ifier*
  - variables prédictrices :
    - ★ âge du dérivé
    - ★ longueur du radical
    - ★ dernière consonne du radical
    - ★ groupe consonantique
    - ★ dernière voyelle du radical
    - ★ FMA
    - ★ Classe morphologique de l'adjectif (CMA)
  - modéliser l'existant : étant donné l'échantillon de verbes que nous avons, quel ensemble de facteurs a le meilleur pouvoir explicatif (statistiquement parlant) ?
- Estimer la probabilité $P(Y = \text{IFIER} \mid X)$ que le suffixe soit *-ifier* étant donné un ensemble X de variables prédictrices
  - pouvoir prédictif du modèle : pour un nouveau cas de compétition (par ex. *zenifier* ou *zeniser*), étant donné la valeur de chaque variable prédictrice, le modèle prédit la probabilité $P(\text{IFIER})$ que le verbe construit soit en *-ifier*

Modélisation multivariée

- Quelques préalables méthodologiques
  - Il est préférable de ne pas construire un modèle contenant des variables avec de nombreux niveaux, s'ils sont peu fréquents (il faudrait un nombre raisonnable d'observations par niveau, donc une quantité beaucoup plus importante de données)
    - ★ regroupement des niveaux de *Classe morphologique de l'adjectif* en 4 catégories : *-ique* / dénominaux avec sonante / autres dénominaux / non-suff + pas-adj
    - ★ 2 niveaux seulement pour *Dernière consonne du radical* : Alvéolaire Obstruante ou Autre
  - on ne peut pas utiliser simultanément les variables *FMA* et *CMA*, car elles sont trop fortement corrélées

	Adj	both	N
-ique	2	172	0
denom_autres	8	67	0
denom_son	6	216	0
non-suff + pas-adj	44	165	115

    - ★ On va faire deux modèles différents : l'un avec la *FMA* et l'autre avec la *CMA*

Modèle FMA

- Le modèle avec l'ensemble des variables phonologiques, la date d'attestation + la FMA
  - Méthode de compactage du modèle :
    - ★ on compare deux modèles, l'un avec et l'autre sans une variable prédictrice donnée ;
    - ★ si l'ajout de la variable n'améliore pas la valeur explicative du modèle, alors on l'exclut du modèle

  ⇒ La variable *Groupe Consonantique*, qui selon notre étude de corrélation favorisait *-ifier*, est éliminée

Modèle FMA

- Le modèle :

	Estimate	Std Error	z value	Pr	
(Intercept)	4.418687	0.764498	5.780	7.48e-09	***
date	-0.003701	0.001406	-2.632	0.00848	**
longueur	-2.847165	0.285133	-9.985	< 2e-16	***
alvObs=True	0.784509	0.313883	2.499	0.01244	*
voyelle=fermee	0.583080	0.335803	1.736	0.08250	.
voyelle=nasale	1.638080	0.543319	3.015	0.00257	**
FMA = both	-0.919231	0.487462	-1.886	0.05933	.
FMA = N	-0.457696	0.581087	-0.788	0.43090	

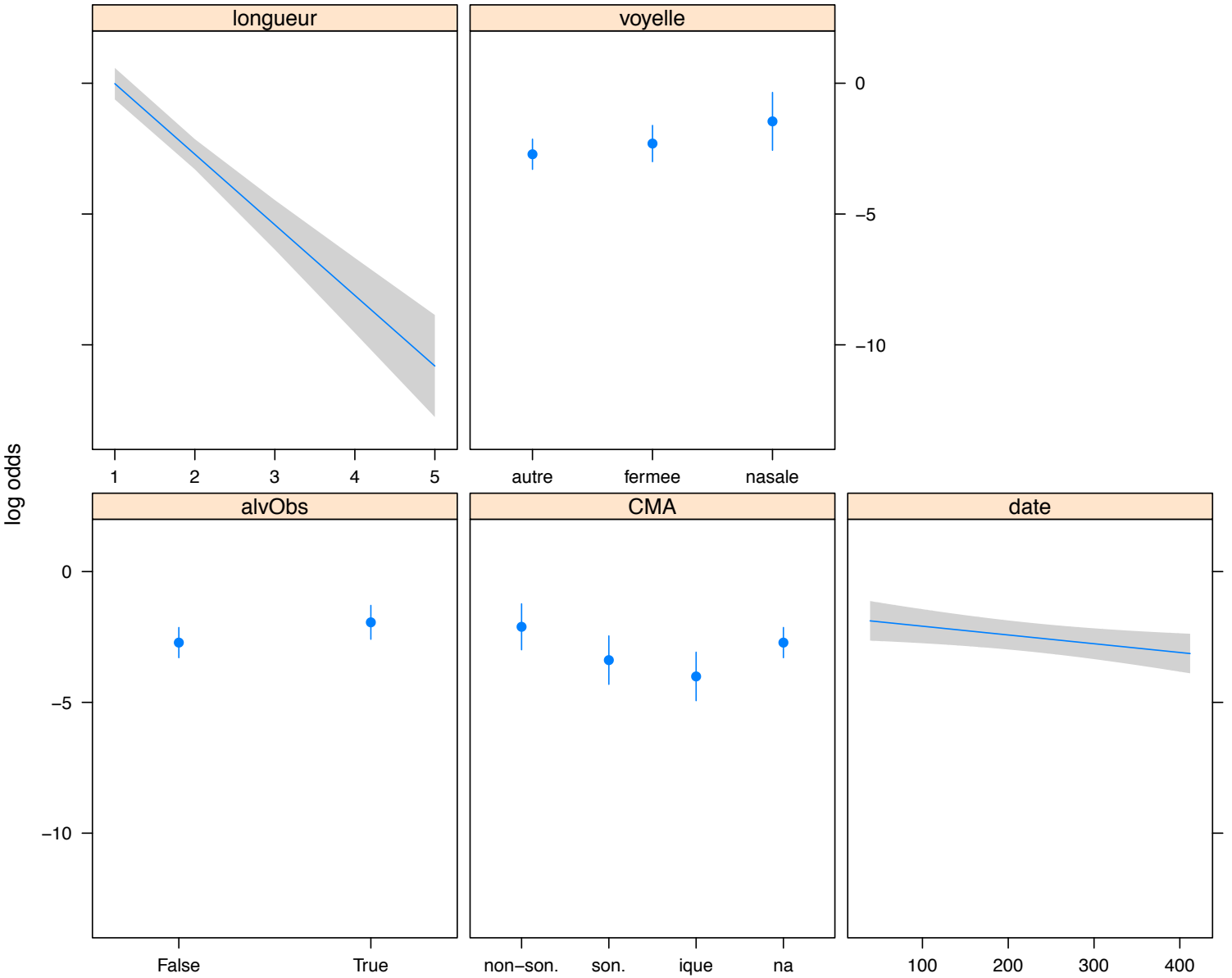
→ La variable *FMA* apparaît très faiblement significative ; d'autres tests statistiques tendent à montrer que son pouvoir explicatif est quasi-nul

Modèle CMA

- Le modèle avec l'ensemble des variables phonologiques, la date + la CMA
  - Méthode de compactage du modèle amène à l'élimination de la variable *Groupe consonantique*
- Le modèle CMA :

	Estimate	Std Error	z value	Pr	
(Intercept)	4.246550	0.710404	5.978	2.26e-09	***
date	-0.003348	0.001413	-2.370	0.01779	*
longueur	-2.698550	0.278013	-9.707	< 2e-16	***
alvObs=True	0.774453	0.325081	2.382	0.01720	*
voyelle=fermee	0.409135	0.338902	1.207	0.22734	
voyelle=nasale	1.257070	0.546489	2.300	0.02143	*
CMA=denom_son	-1.273474	0.614115	-2.074	0.03811	*
CMA = ique	-1.897110	0.583385	-3.252	0.00115	**
CMA = na	-0.603284	0.432394	-1.395	0.16295	

Effets des variables prédictrices – Modèle CMA

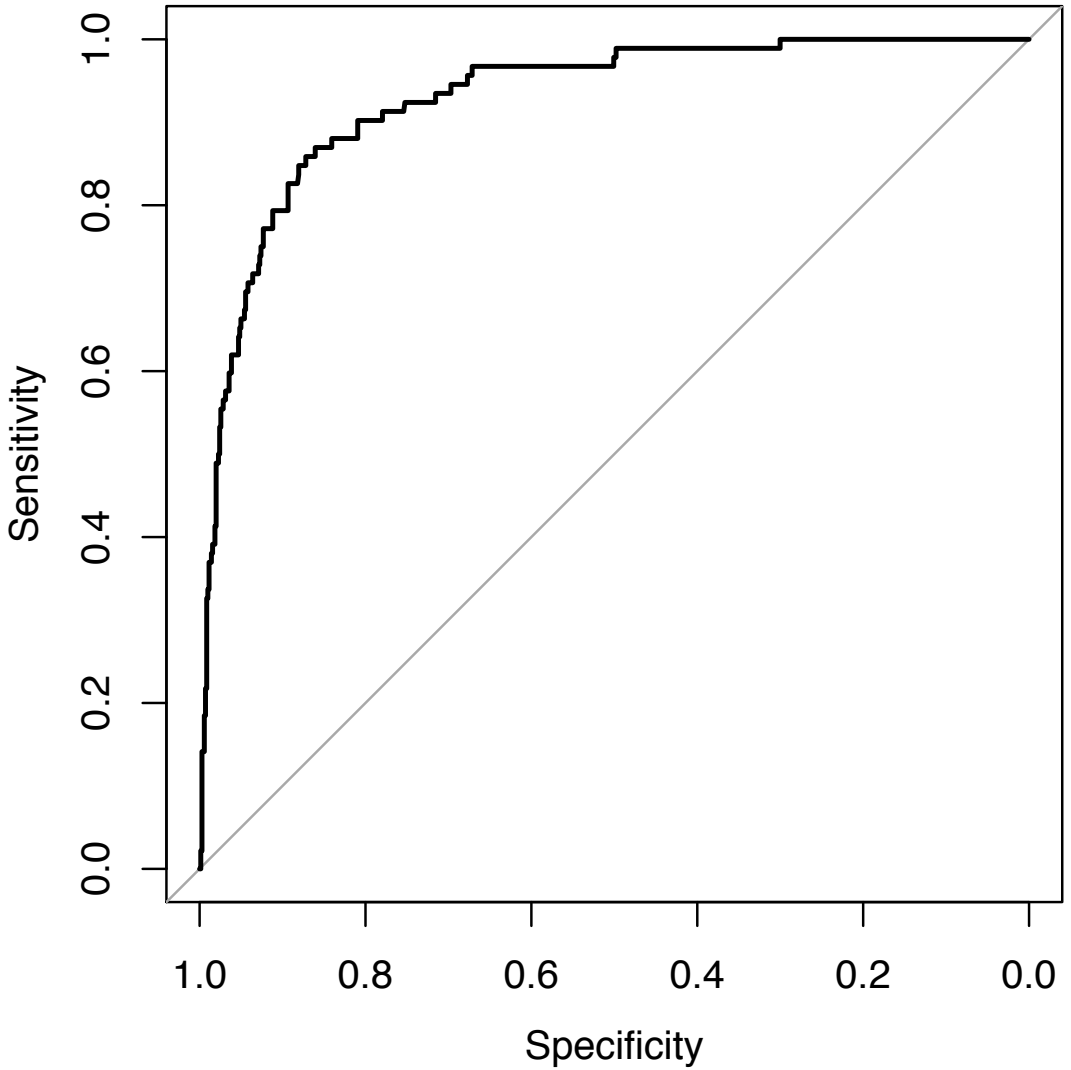


Evaluation du modèle

- Exemple de mesure utilisée pour évaluer la qualité de prédiction
  - Aire sous la courbe ROC : comparaison du nombre de fois où *ifier* est prédit correctement par le modèle, par rapport au nombre de fois où *iser* est prédit incorrectement
  - Pour avoir une bonne qualité de prédiction, il faut maximiser les cas où *-ifier* est correctement prédit et minimiser les cas où *-iser* est incorrectement prédit.
  - Pour le modèle CMA, ROC $= 0,92$
    - ROC $= 0,5$: prédiction aléatoire
    - ROC $= 1$: prédiction parfaite
- Le modèle CMA a une très bonne qualité de prédiction, il présente une bonne adéquation aux données

Evaluation du modèle

- Aire sous la courbe ROC



Modélisation multivariée

- Intérêt au niveau méthodologique :
  - facteurs qui n'ont pas d'effet sur le choix du suffixe, une fois pris en compte l'ensemble des facteurs (cf. groupe consonantique en fin de radical)
  - facteurs corrélées que l'on peut essayer de départager en comparant leur pouvoir explicatif (cf. CMA et FMA)
- Résultats « plus qualitatifs » : compétition entre *-iser* et *-ifier* met en jeu simultanément
  - propriétés du radical (longueur, consonne finale, voyelle finale)
  - dimension diachronique
  - propriété morphologique (classe morphologique de l'adjectif)
- Nous avons créé un modèle qui rend compte de l'effet simultané de ces facteurs, et qui peut prédire avec une bonne précision la probabilité du choix du suffixe *-ifier* pour un nouveau cas de compétition

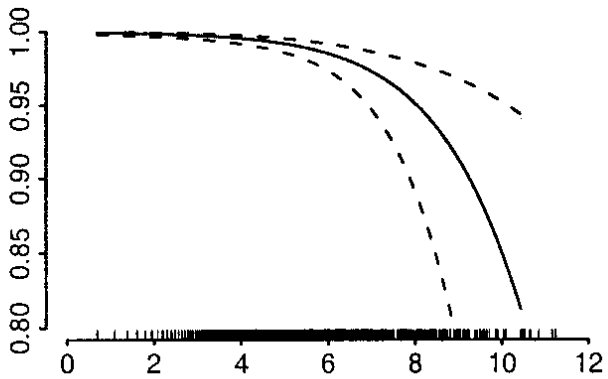
Nouvelle approche, nouvelles questions

Nouvelle approche, nouvelles questions

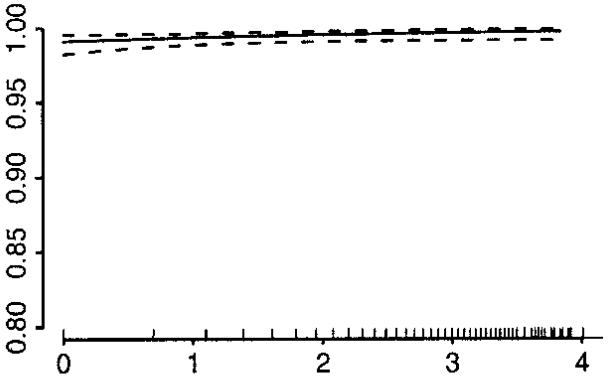
- Dans ce qui précède, on a appliqué la régression logistique à la modélisation de la rivalité entre procédés constructionnels.
- La littérature récente applique des modèles similaires à des questions très différentes.
  - Prédiction du caractère irrégulier d'un verbe en anglais (Baayen and Moscoso del Prado Martín, 2005)
    - ☞ Régression logistique sur des types
  - Différences entre adjectifs ethniques et SPrep (Boleda et al., 2012)
    - ☞ Régression logistique sur des occurrences
  - Prédiction de l'accentuation des composés en anglais (Bell and Plag, 2012)
    - ☞ Régression logistique à effets mixtes sur des occurrences
  - Réalisation des *s* finaux en anglais (Plag et al., 2015)
    - ☞ Régression linéaire à effets mixtes sur des occurrences
  - Distribution des noms apocopés en anglais (Kanwal, 2015)
    - ☞ Régression linéaire à effets mixtes sur des occurrences

Prédire l'irrégularité

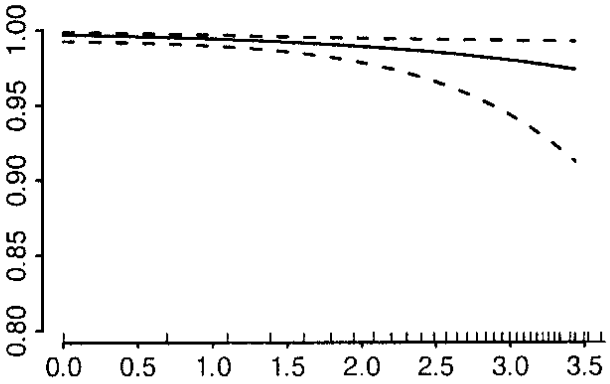
- Baayen and Moscoso del Prado Martín (2005) : En anglais, le fait qu'un verbe soit régulier ou non est partiellement prédictible à partir de ses propriétés fréquentielles, des propriétés de sa famille morphologique, et de ses propriétés sémantiques.



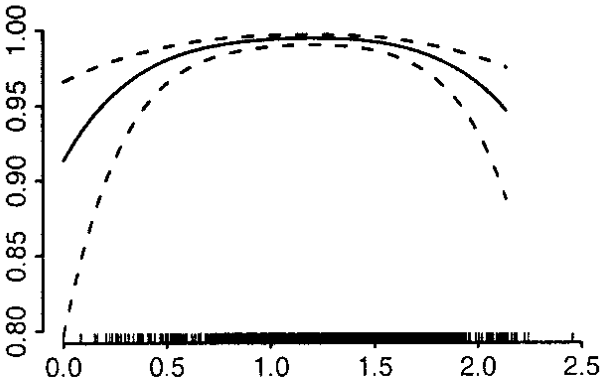
log. de la fréquence



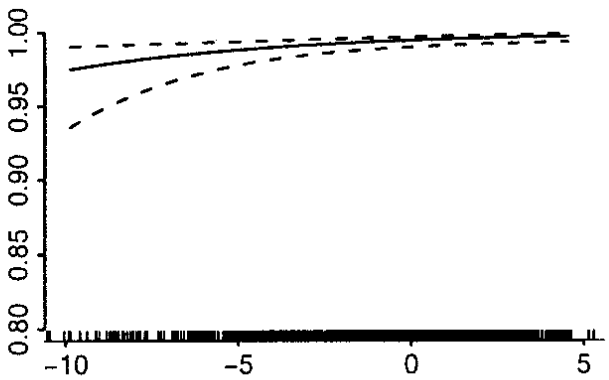
taille de la famille morphol.



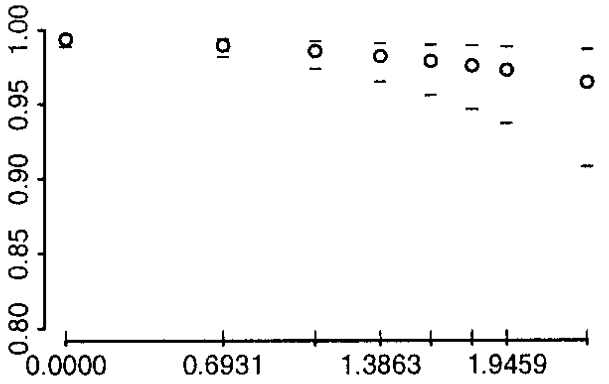
log nb de synsets WordNet



entropie flexionnelle



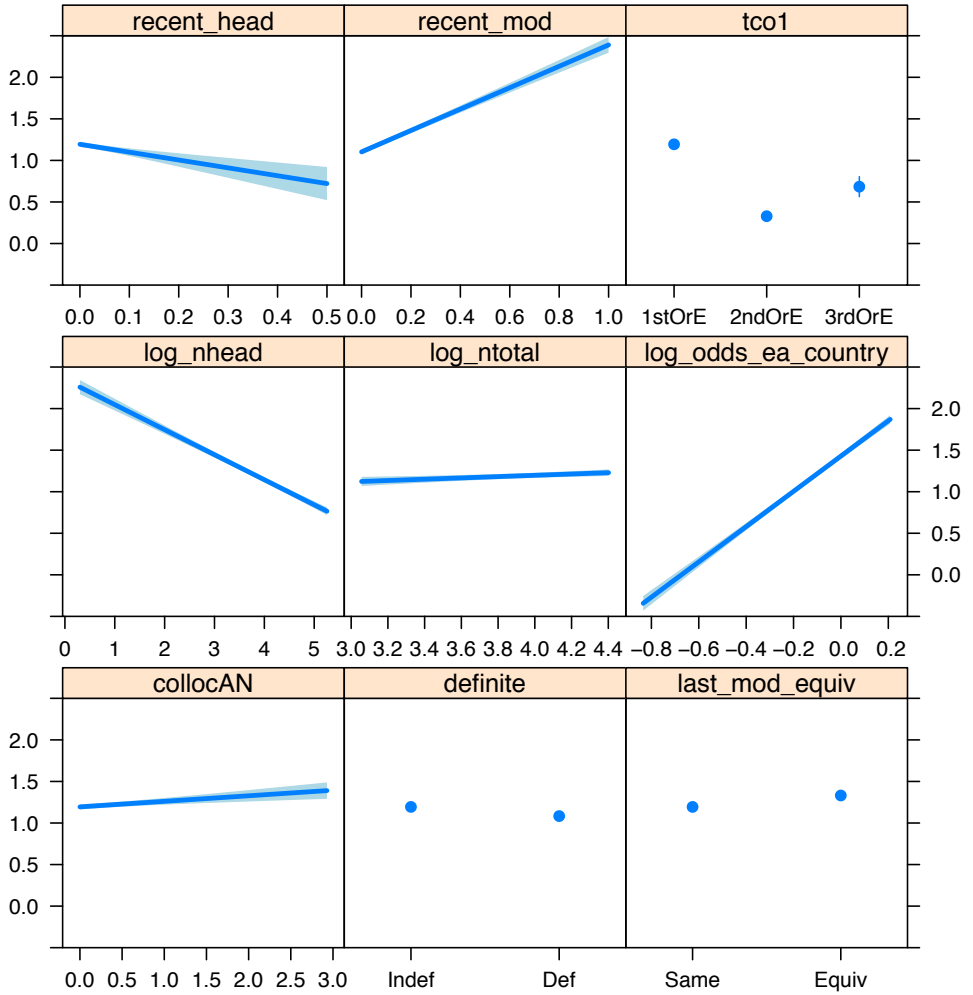
log. ratio nom/verbe



log nb de classes Levin

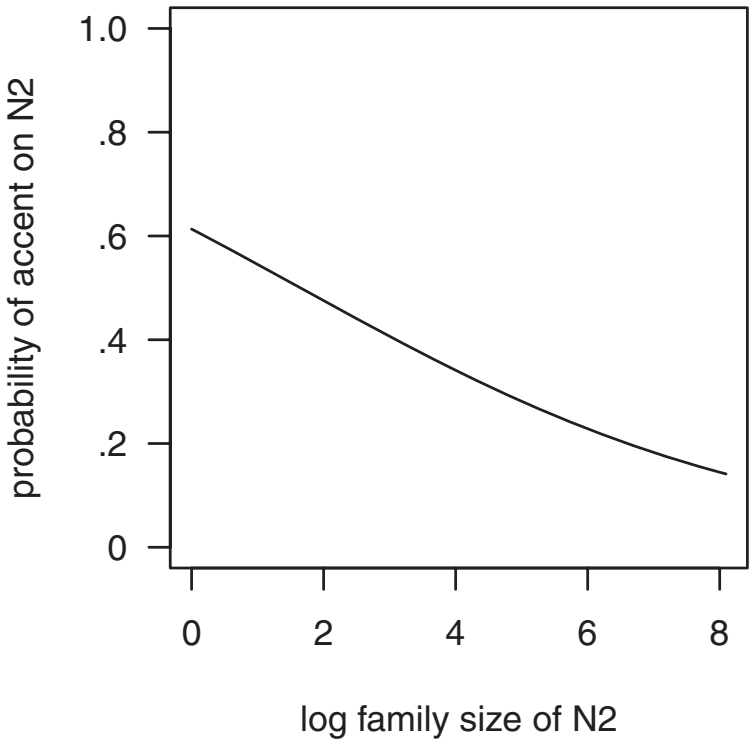
Départager des analyses concurrentes

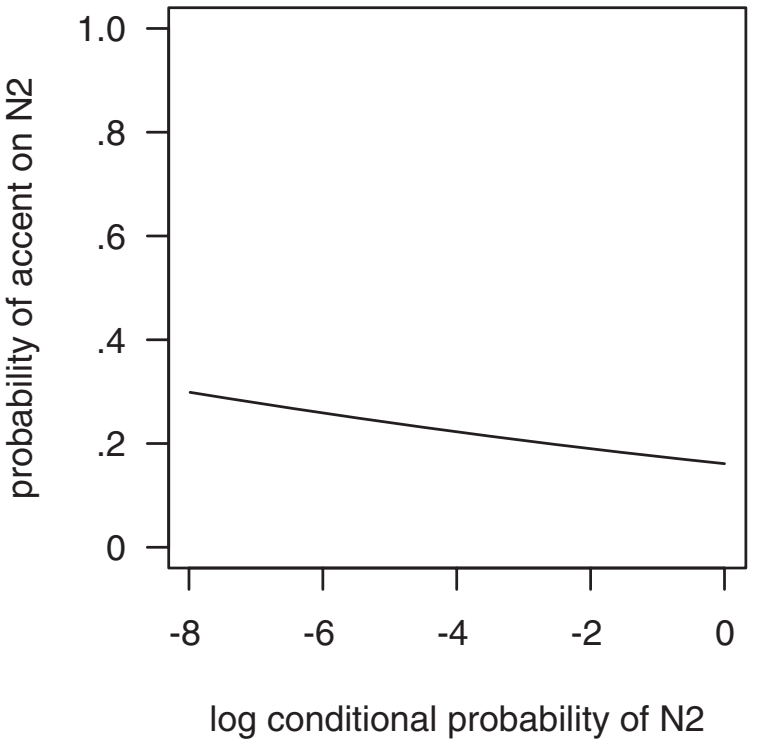
- Boleda et al. (2012) compare les analyses des adjectifs relationnels comme saturateurs (i.e. équivalents à des arguments) ou modifieurs.
- Dans le cas particulier des adjectifs ethniques en anglais, montrent que la distribution des adjectifs ethniques vs. des SP favorise une analyse comme modifieur.



Prédire l'accentuation des composés

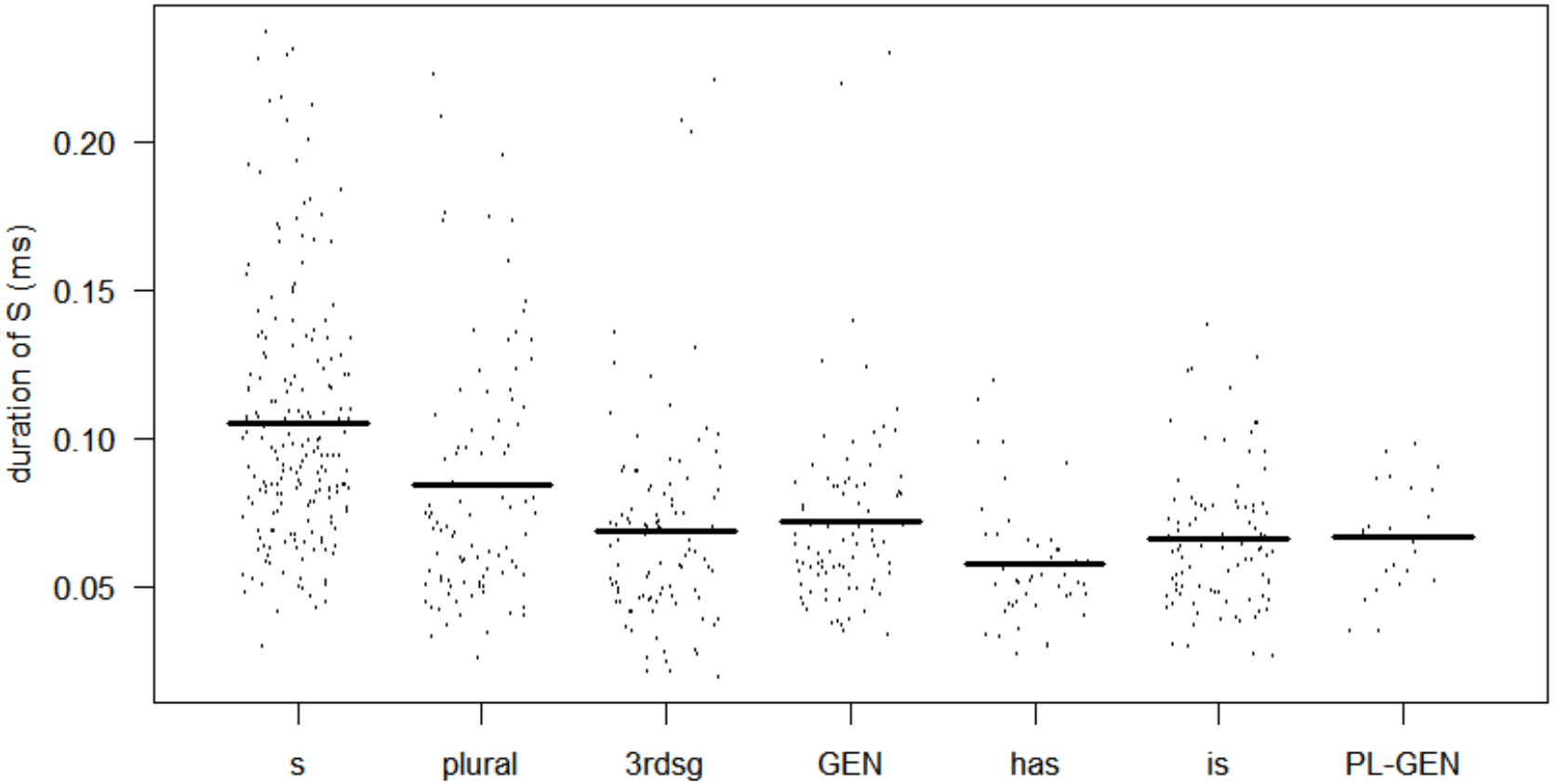
- Bell and Plag (2012) : la place de l'accent dans les composés NN en anglais est partiellement prédictible à partir de différentes variables sémantiques et de l'informativité relative des deux N.





Les corrélats phonétiques de la phonologie

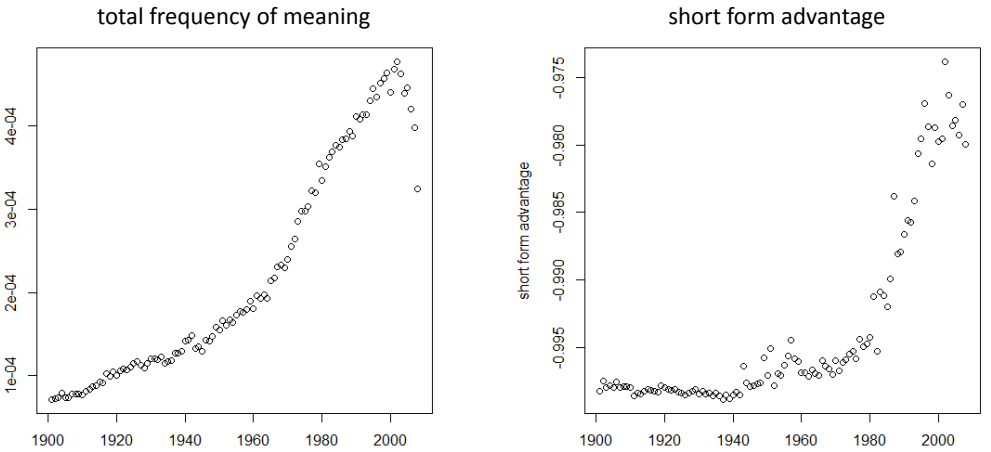
- Plag et al. (2015) : les *s* finaux de mots en anglais sont réalisés différemment en moyenne selon leur statut morphologique
- De manière cruciale, il y a à la fois une différence significative et un recouvrement important des distributions



Examiner l'évolution de l'emploi de concurrents

- Kanwal (2015) : test empirique des effets diachroniques du Principe du Moindre Effort.
- Quand un même sens peut être désigné par 2 formes de longueurs différentes, la préférence pour la forme courte co-varie avec la fréquence d'emploi des deux.
- Confirmé sur deux domaines empiriques : noms apocopés, paires *-ic/-ical*.

info vs. *information*



asymmetric vs. *asymmetrical*

Conclusions

Conclusions

- Les points que nous avons tenté de faire :
  - Même si c'est le plus souvent implicite, la recherche en morphologie s'appuie sur des données quantitatives. On a besoin d'outils statistiques pour juger de ces données.
  - Le fait d'utiliser de grands échantillons améliore la qualité des généralisations, mais ne garantit pas en tant que tel qu'on peut faire une induction de l'échantillon au cas général.
  - Il existe des tests statistiques simples qui permettent de vérifier qu'une telle induction est raisonnable.
  - La régression logistique est un outil abordable pour étudier les phénomènes de tendances, en particulier la rivalité entre affixes.
  - D'autres types de modèles de régression peuvent être mis en œuvre pour aborder des questions morphologiques importantes mais négligées.

Baayen, H. and Moscoso del Prado Martín, F. (2005). 'Semantic density and past-tense formation in three Germanic languages'. *Language*, 81 :666–698.

Bell, M. J. and Plag, I. (2012). 'Informativeness is a determinant of compound stress in English'. *Journal of Linguistics*, 48 :485–520.

Boleda, G., Evert, S., Gehrke, B., and McNally, L. (2012). 'Adjectives as saturators vs. modifiers : Statistical evidence'. In M. Aloni, V. Kimmelman, F. Roelofsen, G. W. Sassoon, K. Schulz, and M. Westera (eds.), *Logic, Language and Meaning - 18th Amsterdam Colloquium, Amsterdam, The Netherlands, December 19-21, 2011, Revised Selected Papers*. Dordrecht : Springer, 112–121.

Bonami, O., Boyé, G., and Kerleroux, F. (2009). 'L'allomorphie radicale et la relation flexion-construction'. In B. Fradin, F. Kerleroux, and M. Plénat (eds.), *Aperçus de morphologie du français*. Saint-Denis : Presses de l'Université de Vincennes, 103–125.

Hathout, N., Sajous, F., and Calderone, B. (2014). 'GLÀFF, a large versatile French lexicon'. In *Proceedings of LREC 2014*.

Kanwal, J. (2015). 'The principle of least effort and diachronic lexical change'. In *Causality in the Language Sciences*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Lignon, S. (2013). '*-ISER* and *-IFIER* suffixation in French : Verifying data to 'verize' hypotheses'. In N. Hathout, F. Montermini, and J. Tseng (eds.), *Morphology in Toulouse, Selected prodeedings of Décembrettes 7*. Munich : Lincom Europa, 119–132.

Lignon, S., Namer, F., and Villoing, F. (2014). 'De l'agglutination à la triangulation ou comment expliquer certaines séries morphologiques'. In F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J. Meinschaefer, and S. Prévost (eds.), *Actes du quatrième Congrès Mondial de Linguistique Française*. 1813–1835.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., andSteven Pinker, J. O., Nowak, M. A., and Aiden, E. L. (2010). 'Quantitative analysis of culture using millions of digitized books'. *Science*, 14 :176–182.

Namer, F. (2013). 'Adjectival bases of French *-aliser* and *-ariser* verbs : syncretism or underspecification ?' In N. Hathout, F. Montermini, and J. Tseng (eds.), *Morphology in Toulouse, Selected prodeedings of Décembrettes 7*. Munich : Lincom Europa, 185–210.

Plag, I. (1999). *Morphological productivity*. Berlin : Mouton de Gruyter.

Plag, I., Homann, J., and Kunter, G. (2015). 'Homophony and morphology : The acoustics of word-final -S in English'. *Journal of Linguistics*.

Plénat, M. (2000). 'Quelques thèmes de recherche actuels en morphophonologie française'. *Cahiers de lexicologie*, 77 :27–62.

Plénat, M. and Roché, M. (2003). 'Prosodic constraints on suffixation in French'. In G. E. Booij, J. de Cesaris, A. Ralli, and S. Scalise (eds.), *Topics in Morphology. Selected Papers from the Thirst Mediterranean Morphology Meeting*. Barcelona : IULA-Universitat Pompeu Fabra, 285–289.

Strnadová, J. (2014). *Les réseaux adjectivaux : Sur la grammaire des adjectifs dénominaux en français*. Ph.D. thesis, Université Paris Diderot et Univerzita Karlova V Praze.

Tribout, D. (2010). *Les conversions de nom à verbe et de verbe à nom en français*. Ph.D. thesis, Université Paris Diderot.