

Ressources et outils en phraséologie informatique

Jean-Pierre Colson

Université catholique de Louvain
(Louvain-la-Neuve)

1. Du côté des ressources : corpus, taille et diatopie

- La taille des corpus est cruciale; en effet, la phraséologie dans son ensemble représente plus de la moitié de n'importe quel texte (Sinclair 1991), mais une structure phraséologique précise nous gratifiera généralement d'une fréquence très faible sur un corpus donné (Moon 1998, Colson 2007)

- Pour rappel : le Web (par exemple via le moteur Google) ne livre aucune donnée fiable en matière de fréquence des unités phraséologiques
- Exemple : quelle est la fréquence de l'expression *ça fait la rue Michel* Internet? (moteur Google, consulté le 18/08/2016)

← → ↻  [https://www.google.be/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=%22%C3%A7a](https://www.google.be/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=%22%C3%A7a%20fait%20la%20rue%20michel%22)



"ça fait la rue michel"



Tous

Vidéos

Images

Actualités

Maps

Plus ▼

Outils de recherche

Environ 3.580 résultats (0,40 secondes)



"ça fait la rue michel"



Tous

Vidéos

Images

Actualités

Maps

Plus ▼

Outils de recherche

Page 12 sur environ 110 résultats (0,35 secondes)

- Les corpus de taille gigantesque sont donc nécessaires à la phraséologie informatique et, à défaut, les meilleures ressources sont fournies par les corpus Web, par exemple via le Sketch Engine (www.sketchengine.co.uk)
- Exemples sur le frTenTen (10 milliards de mots)

ça fait la rue michel ▼ 🔍 📖 French Web 2012 (frTenTen12)

Search in French Web 2012 (frTenTen12) (trial)

Query **ça, fait, la, rue, michel** 15 (0.00 per million) ⓘ

doc#1029909	sur le guidon avec un collier serflex et ça fait la rue michel ça va ça va j'connais l'chemin Invité
doc#1975505	une ou 2 feuilles de journal au bout, et ça fait la rue Michel !" De toute façon les femmes, ce
doc#7525700	bon compte, ou comme on dit par chez moi, ça fait la rue Michel ... enfin exception faite de cette petite
doc#8002438	conclue par cette sentence : "Et voilà, ça fait la rue Michel !". A son air égaré, je sentais bien qu'
doc#8002438	pour n'en faire qu'une phrase : "Et voilà, ça fait la rue Michel ! Oui, bon, CQFD si vous préférez ?". L'
doc#8557658	tout, une bonne sauce salsa bien relevée, ça fait la rue Michel ! Et rendez-vous dans dix jours pour savoir
doc#9259382	finies de mastic silicone en dessous, et ça fait la rue Michel . Le tapis fii est rès fin, et se pose sur
doc#9999797	Lecomte à Paris : "ça fait le compte", donc " ça fait la rue Michel " (Lecomte !). Question : Peut-on
doc#10264021	cher, un peu de coton au fond du sac et ça fait la rue michel . J'ai pas bon ? ça fait complètement
doc#14497163	aurais obtenu comme réponse : 1113122119 Et ça fait la rue Michel . Quelle est de toutes les choses
doc#15731613	pris au sens de compte) . Autrement dit " ça fait la rue Michel " = "ça fait le compte". Paris 4ème
doc#16222059	face au Réal. Après on fait nul à Milan et ça fait la rue Michel à nous les huitièmes :happy2: Oui
doc#17765110	lesquelles il recharge ses lanceurs. Et ça fait la rue Michel ! Enfin surtout rectification ultime
doc#20189912	lecture diagonale, des morceaux choisis et ça fait la rue Michel . " Le zapping a ceci de très commode qu'
doc#20279267	rajoutes juste une couche de couverture et ça fait la rue michel , c'est juste pas trop facile pour les couloirs

- Sur le même corpus (10 milliards de mots), aucune occurrence pour *faire danser l'anse du panier* (faire des profits illicites), 22 pour *ça fait la rue Michel*, 15 pour *adieu paniers, vendanges sont faites*; aucune occurrence au sens figuré pour *emporter la paille* (obtenir un succès remarquable)
- Même des expressions plus courantes, telles *laisser la proie pour l'ombre*, obtiennent des fréquences très basses sur ce corpus gigantesque:

Query **laiss.***, **proie**, **ombre** 15 (0.00 per million)

doc#309346	plus décevant des mirages, que nous avons	laissé la proie pour l'ombre	! </p><p> Et soudain, vivement, inopinément
doc#2182448	raisonnable de sa femme l'engageant à ne pas	laisser la proie pour l'ombre	et à ne pas quitter l'école Dupras, même
doc#2190213	entretien n'est pas précisé. Une fois de plus on	laisse la proie pour l'ombre	et il faudra se battre pour l'achat le
doc#2274440	s'opposent pas vraiment. Qui est fou pour	laisser la proie pour l'ombre	? Au nom de qui ? Au nom du peuple ? </p>
doc#4747027	certaines acteurs de la scène sociopolitique ne	laissent la proie pour l'ombre	en s'épuisant dans des débats de salons
doc#8396595	l'atteigne et ne l'empoigne ! Peut-être	laissait-elle la proie pour l'ombre	, mais ici ce n'était tout simplement pas
doc#11135273	que le " Zorro " de la politique togolaise	laisse la proie pour l'ombre	. Celui qui crie sur tous les toits prôner
doc#11271684	pas nous conseiller d'imiter le chien qui	laisse sa proie pour l'ombre	? </p><p> Et d'ailleurs, encore une fois,
doc#11869462	me suis souvent demandé pourquoi j'avais	laissé la proie pour l'ombre	en revenant en France l'année dernière.
doc#12132278	de l'élite du pays contre la tentation de	laisser la proie pour l'ombre	. Après avoir laminé nos défenses immunitaires
doc#15463478	détermination comme par votre audace de	laisser la proie pour l'ombre	projetée de votre rêve. Quitter le rémunéré
doc#16994064	à Ouaga 2000. </p><p> Cependant ce serait	laisser la proie pour l'ombre	, qu'au moment où les Guinéens sont à la
doc#17236669	illuminer la planète bleue au lieu de la	laisser en proie à l'ombre	? Personne n'aurait pu dire mais tous ce
doc#17384312	regrettent leur vote et qui reconnaissent avoir	laissé la proie pour l'ombre	. Seul ce blog permet le débat car le maire
doc#19082577	nouveau dès qu'il est sorti de l'église. Il	laisse sa proie pour son ombre	. Il est l'emblème de l'ambition stupide

Query **emport.***, **paille** 17 (0.00 per million)

doc#310268	contre eux et ils sèchent. La tempête les	emporte comme une paille	. 25 "À qui m'identifiez-vous qui m'équivaille
doc#794106	(la pierre l'emporte sur le bois qui l'	emporte sur la paille	) permet néanmoins de reconquérir des positions
doc#945658	sécurité en supprimant de la nacelle l'	emport de paille	enflammée pour chauffer l'air ; même si
doc#1284888	ses "remparts personnels" comme le vent	emporte la paille	... Il se remémora.. Ses éclats
doc#1315613	lancer le blé dans les airs pour que le vent	emporte la paille	et qu'il ne reste que le grain. Jésus va
doc#1784018	battage de la moisson permettait au vent d'	emporter la paille	, les feuilles et les détritrus légers et
doc#2505758	, devant le tribunal de Dieu, ils seront	emportés comme la paille	que le vent dissipe, car sur la balance
doc#3954166	ruse pour toucher les pieds d'Hugues et	emporter ainsi la paille	de ses sabots [Vita Hugonis , ch.35].
doc#6458849	maitresse . A peine au ras du sol, elle semblait	emporter toute la paille	à sa suite . Dire que Aaron avait été un
doc#8333915	enseignement grave de la morale, ils se laissent	emporter comme une paille	légère au vent de toute doctrine. Ils se
doc#8814463	pesants résistent à. l'action du vent, qui n'	emporte que la paille	inutile. Aussi l'apôtre saint Paul, après
doc#10885602	balayage ou renoncer. Les cultivateurs qui	emporteront leur paille	devront nettoyer, la propriété de la dite
doc#11308297	régulière et drue, tandis que le vent de la mer	emporte la paille	et la fait tourbillonner. on vanne sur
doc#12218523	empresse de changer le lit du mort, et l'on	emporte la paille	sur un grand chemin pour y être brûlée.
doc#12266468	marchait sur Moscou : il n'y a plus rien à	emporter, ni paille	, ni blé.... Avec le mois de Novembre
doc#14830276	délégué. >> >> >> >> Laurent Gouteng >> >>N'	emportez rien (paille	chaines,cacahuettes)seront à votre >>disposition
doc#15602412	de nos passions si changeantes, et d'être	emporté comme une paille	au moindre souffle de la tempête, On devient

- Du point de vue de la variation phraséologique, les corpus Web ne couvrent que partiellement la diatopie, même pour les langues internationales: seul l'espagnol offre via le Sketch Engine une distinction globale Europe / Amérique

Spanish	American Spanish Web 2011 (esamTenTen11, Freeling v4)	7,475,645,291	 
Spanish	Araneum Hispanicum Maius [2013]	892,299,770	 
Spanish	CHILDES Spanish Corpus	802,743	 
Spanish	DGT, Spanish	57,311,149	 
Spanish	esTenTen [2011, Eu + Am, Freeling v4, sample]	63,615,789	 
Spanish	EUROPARL7, Spanish	54,302,284	 
Spanish	European Spanish Web 2011 (eseuTenTen11, Freeling v4)	2,021,756,831	 
Spanish	OPUS2 Spanish	701,944,027	 
Spanish	Spanish Web 2011 (esTenTen11, Eu + Am, Freeling v4)	9,497,402,122	 
Spanish	Spanish web corpus	97,773,185	 

2. Du côté des outils : entre informatique, statistique, langue et culture

- Le traitement informatique de très gros corpus (> 200 MW) à des fins de phraséologie, notamment pour étudier la variation diatopique, peut se faire de manière efficace (mais plutôt lente) par des « expressions régulières », dans n'importe quel langage informatique
- Exemple: *laisser la proie pour l'ombre*

/laiss\S+\s(\S+\s)[0,2]la(\S+\s)[1,3]proie(\S+\s)[1,4]ombre\b/

- Cette syntaxe très puissante permet d'extraire des corpus toutes les variantes lexicales et phraséologiques recherchées; exemple:
POUSSER LE/LA/LE.... trop loin (corpus NewFrench200MW2, JP Colson 2015)

\bpouss\S+\sl[ae]s?\s\S+\strop\sloin\b



Regex: **\bpouss\S+\s[ae]s?\s\S+\strop\sloin\b**

5	pousser le bouchon trop loin
3	pousser les choses trop loin
2	pousse le bouchon trop loin
2	pousser les exigences trop loin
1	poussé le balancier trop loin
1	poussé le bouchon trop loin
1	poussais le bouchon trop loin
1	poussais le scepticisme trop loin
1	pousse le barillet trop loin
1	pousser la panse trop loin
1	pousser la sévérité trop loin

- Le Sketch Engine permet également de tester la fréquence de variantes régionales d'unités phraséologiques (Exemple: *être plus royaliste que le roi / être plus catholique que le pape*)

Query **plus, royaliste, que, le, roi** 1,150 (0.10 per million) 

Page of 58 [Next](#) | [Last](#)

doc#24493 libanaises " sont championnes dans l'art d'être **plus royalistes que le roi** ". Ils se sont dit " surpris " qu'un film
doc#30272 dit l'adage populaire qu'ils sont devenus **plus royalistes que le roi** . </p><p> Je voudrais rappeler à ces nouveaux
doc#37740 faudra nous poser la question: pourquoi être **plus royalistes que le roi** ? </p><p> Il n'y a pas de solidarité arabe
doc#46581 il n'y avait plus de raison de se montrer **plus royaliste que le roi** . D'autant que la coopération arabe n'a
doc#78978 creuset d'où ils sont issus et en se faisant **plus royaliste que le Roi** lui-même. " Une idée sociale de droite
doc#79108 grandes armées pourquoi voulais-tu être **plus royaliste que le roi** ? Un Chef d'Etat est le protecteur du peuple
doc#107376 s'en tenir. De grâce n'essayez pas d'être **plus royaliste que le roi** et soyons courtois au moins dans nos posts
doc#107376 vous savoir de plus...n'essayez pas d'être **plus royaliste que le roi** ", (Non, ce n'est pas ce que fais. Pourquoi
doc#148337 . Vous me permettez donc de ne pas être **plus royaliste que le roi** . </p><p> En plus du site www.Castaluna.com
doc#158611 dire que certains gens, qui se croyaient **plus royalistes que le roi** , ont agi en son nom pour faire des choses
doc#160457 nous nous demandons si nous ne sommes pas **plus royalistes que le roi** (plus zen que les maîtres zen !) quand
doc#187194 sens de la fronde qui nous rendrait presque **plus royaliste que le roi** . Il faut dire qu'il connaît la musique.
doc#189421 tribu. S'il connaît des gens qui ont été **plus royalistes que le roi** qu'il les dénonce cela dédouanerait mieux
doc#204256 première d'une modération mais ne soyons pas **plus royaliste que le roi** . Nous sommes tous pareils. </p><p> - l'impact
doc#210898 expression de M. de Loménie, lui, qui avait été **plus royaliste que le roi** , il se montra parfois plus libéral que
doc#213229 citoyens guinéens à part entière. Pourquoi **plus royaliste que le roi** lui-même??? Dans l'esprit de la plus grande
doc#241942 ralliés " se sentent morveux : ils se montrent **plus royalistes que le roi** et attaquent ce qui demeure la véritable
doc#254441 dès qu'il le peut ". </p><p> Les ordres sont **plus royalistes que le roi** , bien au chaud derrière leurs rentes de
doc#258003 </p><p> Auteur Sujet: [SPOIL] Teshi est elle **plus royaliste que le roi** ou plus kurumadienne que kuru ? (Lu 5751
doc#261606 dans l'autre sens... ils seront donc encore **plus royaliste que le roi** , et plus pro Srako, que sarko lui-même.

Page of 58 [Next](#) | [Last](#)

Query **plus, catholique, que, le, pape** 247 (0.00 per million) 

doc#114730	automatiquement ressenti. </p><p> Tu veux être plus catholique que le Pape ! Plus Kacher, que la tradition juive l'
doc#318643	Amérindiens non plus. Ah, quand on veut être plus catholiques que le Pape , on devient ridicule. Notre société devient
doc#318643	Finalement, l'intransigeance et l'attitude " plus catholique que le Pape " des groupes écolos va finir par se retourner
doc#377328	<p> Mon voisin est le pharisien de base : plus catholique que le Pape , plus parfait que moi (là, pourtant, je
doc#418878	cache l'arbre. </p><p> Perce qu'à vouloir être plus catholique que le pape on aboutit avec les condos Griffintown...
doc#550429	sont-ils tous des putes? Ou sommes-nous plus catholiques que le pape ?La question en soulève une autre: qui prend
doc#595384	comme il est bien commode de ne pas être plus catholique que le pape et de savoir se cacher la vue pour ne pas
doc#662999	dire que l'homme fort de la Flandre se veut plus catholique que le Pape en matière d'intégration européenne et
doc#674089	réussissait la prouesse d'être à la fois plus catholique que le pape et profondément athée. C'est ce christianisme
doc#759423	<p> Dans mon précédent billet " La France plus catholique que le pape ... j'avais affirmé que la France se préparait
doc#791439	deuxième raison est qu'on ne peut pas être plus catholique que le pape . En effet, si les soldats, caporaux et
doc#932813	dans sa vie précédente. Il veut devenir " plus catholique que le pape ", afin qu'on ne puisse mettre en doute
doc#1098247	que Muhammad ? Les occidentaux diraient " plus catholique que le Pape ". Les manipulations d'Ecritures sont dangereuses
doc#1112754	accords, ne pouvait se permettre d'être plus catholique que le pape . Autrement dit, puisque le Président avait
doc#1210259	<p> 31 août 2009 14h57 </p><p> Ils se croient plus catholiques que le papenon mais attends, ils ne sont meme pas
doc#1257728	la fois dans le centre et dans la marge, plus catholiques que le pape , ajouta la femme en riant fort, des étrangers
doc#1465278	fédéralC'est ce qu'on appelle , être plus catholique que le papeBizarre </p><p> Raymond Saint-Arnaud
doc#1626748	jusque la rupture, le Maroc se montrant plus catholique que le Pape). C'est dans ses relations avec Israël
doc#1663286	des problèmes insurmontables. Sommes-nous plus catholique que le Pape ? La religion verte a t'elle remplacée
doc#1675145	</p><p> Ma mère fut, presque toute sa vie, plus catholique que la pape ! Mais dans sa dernière année de vie, le

- Afin de réduire le temps de traitement des données phraséologiques issues de vastes corpus, il est nécessaire de recourir à des techniques informatiques plus poussées
- L'utilisation du *query likelihood model* (exemple: *Lemur Toolkit*, <http://www.lemurproject.org/lemur/>) permet d'effectuer des requêtes sur un très gros corpus en un temps record (Ex.: *cracher le morceau*)

- Le projet IdiomSearch (J.-P. Colson 2015)

<http://IDIOMSEARCH.LSTI.UCL.AC.BE>

réalise une synthèse entre le *query likelihood model* et l'extraction d'information (*information retrieval*) sur la base d'algorithmes de proximité (*metric clusters*); il s'agit en outre d'une méthode issue des corpus (*corpus-driven*) et pas seulement basée sur les corpus

$$cpr = \frac{n(x_{i_1} x_{i_2} x_{i_3} \dots x_{i_n})}{n(\max_{x_{i_1}, x_{i_2}, x_{i_3}, \dots x_{i_n}} (||x_i - x_j||) \leq W}$$

Corpus Proximity Ratio (CPR), J.-P. Colson 2015

- Interactions possibles entre phraséologie informatique et recherche sociolinguistique sur la variation
 - Recherche de variantes d'unités phraséologiques sur de vastes corpus
 - Constitution de corpus ciblés et extraction automatisée des unités phraséologiques

guagua (noun)

American Spanish Web 2011 (esamTenTen11, Freeling v4) freq = [16,880](#) (1.95 per million)

object_of	subject_of	n_modifier	modifies	y_o
2,614 0.15	1,800 0.11	1,334 0.08	2,106 0.12	486 0.03
comian 7 6.25	llorar 38 6.24	anunciador 15 7.93	manito 34 8.16	cacha 6 8.09
coger 52 5.41	guagua llora	guaguas anunciadoras	manito de guagua	carro 26 6.91
coger una guagua	dormir 24 5.40	omsa 11 7.75	poto 20 7.79	guaguas y carros
mudar 7 4.91	la guagua duerme	fenatrano 7 7.30	como poto de guagua	autobús 6 5.48
comer + 175 4.88	nacer 81 4.70	pan + 121 6.63	llanto 35 6.63	tren 5 4.79
come guaguas	guagua nace	las guaguas de pan	el llanto de una guagua	camión 9 4.07
criar 16 4.74	molestar 6 3.90	llorón 10 6.28	cocedura 6 6.53	perro 6 4.00
nacer + 117 4.49	comer 5 2.71	guagua llorona	coche + 113 6.51	niño 95 3.59
comer+se 8 4.48	aprender 9 2.13	pecho 23 5.53	coche de guagua	guaguas y niños
besar 13 4.44	pesar 6 2.04	guagua de pecho	bautizo 11 6.51	vehículo 5 2.47
bautizar 7 4.01	sentir 9 1.91	volador 22 5.22	parada 92 6.49	adulto 10 1.81
llorar 11 4.00	morir 10 1.89	las guaguas voladoras	parada de la guagua	guaguas y adultos
pa 5 3.67	escuchar 7 1.84	apocalíptico 11 4.95	pañal 24 6.47	hijo 6 0.94
botar 6 3.54	crecer 19 1.59	la guagua apocalíptica	pañales de guagua	persona 5 0.20
matar 27 3.51	la guagua crece	prematureo 25 4.45	cofrador 10 6.29	
matar guaguas	gustar 8 1.53	las guaguas prematuras	chupete 6 6.08	
esperar + 211 3.38	sufrir 12 1.23	diabólico 5 4.39	colado 6 6.06	
esperando guagua	necesitar 18 1.18	habana 5 3.56	shampoo 7 5.95	
cuidar 38 3.15	la guagua necesita	feo 11 3.38	chofer 29 5.64	
callar 5 3.04	correr 6 1.06	guagua fea	el chofer de la guagua	
dormir 6 2.47	bajar 6 1.04	brazo 5 3.33	choferes 19 5.35	
montar 14 2.34	ir 76 0.99	chico 17 3.23	choferes de guaguas	
alquilar 5 2.00	la guagua va	una guagua chica	pote 6 5.34	
cargar 11 1.98	venir 23 0.99	repleto 7 3.01	succión 6 5.27	
parar 8 1.76	la guagua viene	mes 26 2.94	papá 20 5.07	
robar 10 1.23	adquirir 5 0.83	guagua de meses	el papá de la guagua	
salir 26 1.18	traer 6 0.74	enfermo 9 2.36	carita 7 4.47	
quemar 5 1.12	salir 17 0.72	sexo 8 2.34	pata 32 4.45	
morir 8 0.88	pasar 22 0.58	ruso 16 2.01	y la pata de la guagua	

autobus (noun)

European Spanish Web 2011 (eseuTenTen11, Freeling v4) freq = [5,788](#) (2.45 per million)

<u>object_of</u>	<u>subject_of</u>	<u>n_modifier</u>	<u>modifies</u>	<u>y_o</u>
910 0.16	323 0.06	612 0.11	821 0.14	264 0.05
cojer 10 7.72	linear 21 7.02	interurbano 20 7.58	parada + 326 8.68	tranvia 10 9.93
cojer el autobus	un autobus de linea	autobus interurbano	parada de autobus	autobus y tranvia
apedrear 6 7.39	parir 3 5.81	turistico 13 7.26	conductor + 112 7.83	metro 50 9.22
fletar 7 7.01	numerar 4 4.74	autobus turistico	conductor de autobus	de autobus y metro
cojiendo 4 6.82	trasladar 4 4.45	alsa 4 7.07	chofer 8 7.67	renfe 6 9.03
coger+te 3 6.30	pa 3 4.17	emt 9 6.96	estacion 13 7.26	taxi 21 8.79
coger + 199 6.24	recorrer 3 3.23	descapotable 4 6.30	la estacion de autobus	autobus y taxi
coger el autobus	salir 13 2.32	compañia 4 5.45	marquesina 8 7.16	tren 38 8.55
coje 3 5.83	autobus sale	cortesía 3 4.37	atras 3 6.23	autobus y tren
conducir 17 4.69	valer 3 2.17	urbano 95 4.24	paliza 3 5.65	autocar 7 8.28
conducir un autobus	dejar 14 2.15	autobus urbano	carnet 7 4.82	cercanias 3 8.13
parar 13 4.30	esperar 4 1.69	vuelta 10 4.03	ticket 3 4.80	tranvía 6 7.06
parada autobus	pasar 11 1.56	autobus de vuelta	billete 16 4.74	estación 9 5.94
pillar 7 4.00	llegar 12 1.35	ida 5 3.62	un billete de autobus	avión 4 5.69
esperar 51 3.73	ir 15 0.50	circular 5 3.37	pasajero 3 4.05	parada 3 5.24
esperando el autobus	el autobus va	nocturno 11 3.05	estación 32 3.22	coche 4 4.46
quemar 4 2.85	llevar 5 0.42	autobus nocturno	la estación de autobus	hotel 3 3.97
tardar 4 2.78	poner 4 0.26	gratis 7 2.69	línea 69 2.95	comercio 3 3.21
pagar 20 2.36	volver 3 0.20	escolar 24 2.60	líneas de autobus	centro 5 2.11
pagar el autobus	dar 7 0.02	un autobus escolar	as 3 2.11	
alquilar 3 2.22		camino 3 2.52	ventana 3 2.01	
ultimar 6 1.89		lleno 11 2.36	combinación 4 1.51	
poner 52 1.83		un autobus lleno de	hora 17 1.35	
llenar 4 1.83		justo 7 2.36	horas de autobus	
tomar 39 1.62		gratuito 19 2.33	accidente 5 1.35	
tomar el autobus		un autobus gratuito	carne 3 1.33	
llegar 20 1.60		turístico 15 1.77	motor 4 1.07	
llega el autobus		un autobus turístico	ruta 3 1.00	
venir 7 1.56		cercano 11 1.76	servicio 47 0.91	