Utilisation et extraction de lexique fréquencé pour l'étude quantitative de la productivité suffixale en français

Présentation de la mise à jour de Lexique3

plan

- Morphologie dérivationnelle et productivité
- Psycholinguistique et bases de données lexicales
- Une stratégie d’estimation de la productivité
- Lexique 3 et sa dernière mise à jour

La productivité suffixale en français

Notions

- productivité:
  - Définition: Une règle ou un procédé morphologique est productif quand il contribue à la création de nouveaux lexèmes. (cf. Schultink 1961)
  - ex. –estre (terrestre, Dal et Namer 2016); -ustre (lacustre); -aume (royaume, Corbin 1987) ; -our (amour) VS –able, -(at)ion, -ifier, -eur, -ment, -iste/-isme;…
- Approche qualitative
  - dictionnaires -> types lexicaux
  - 1ère attestation (diachronie)
  - néologismes
- Approche quantitative:
  - corpus -> occurrences lexicales
  - fréquence (synchronie)
  - hapax

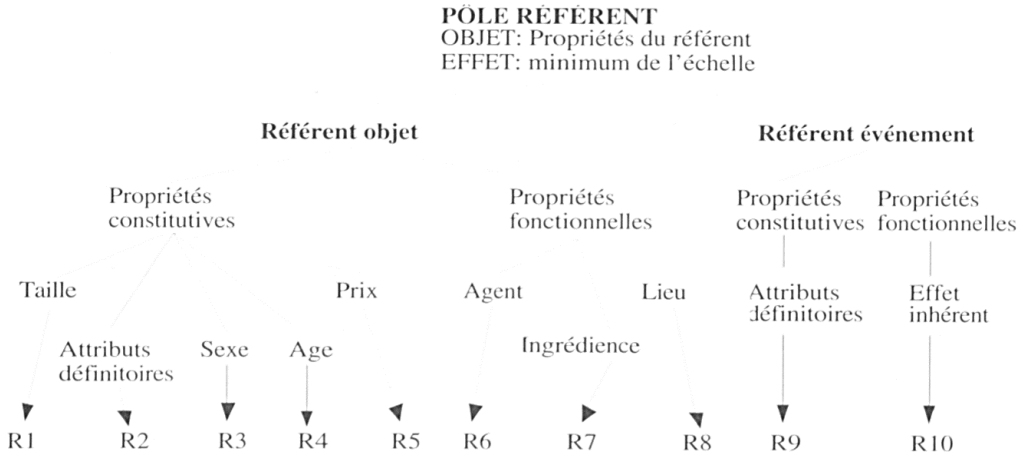
Aliquot-Suengas 2003 : une étude de la productivité à l'ancienne

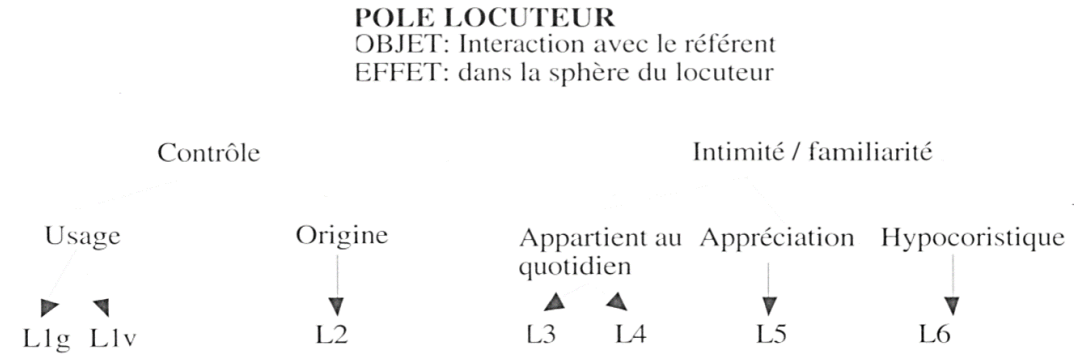
- Objet de l'étude : « la productivité de l'opérateur constructionnel (des opérateurs constructionnels) qui a (ont) la forme -*ade* » (p. 38)
- sources :
  - dictionnaires (Larousse, Robert, TLF)
  - journaux (Le Monde, Libération)
- Méthode:
  - Relevé -> liste de types
  - tri -> sous-ensemble de la liste de départ (exclusion des unités non-construites, sélection de la catégorie de la base)
  - date d'attestation -> chronologie
  - analyse sémantique-> classement
  - comparaison diachronique -> analyse contrastive
- 353 dérivés dénominaux (sur 619 unités construites en –ade)
- « tout à fait disponible dans certains emplois » (p. 52)

Fradin et al 2003 : une étude fine de la productivité en corpus

- objet:
  - comparer la productivité des suffixes *–et* et *–ette*
- corpus:
  - archives électroniques du quotidien Libération
  - de janvier 1995 à décembre 1999 -> 5 ans
  - 87 millions d'occurrences
- méthode:
  - extraction des occurrences -> concordancier
  - classement sémantique -> base de donnée
  - mesures de fréquence -> indices de productivité

Fradin et al 2003: classification sémantique





	EXEMPLE	ENTRÉE	SORTIE
R1	*clochette*	objet'•x_1(nombrable'•x_1	inf•(deg•x_2•taille')•(deg•x_1•taille')
R2	*opérette, cigarette*	entité'•x_1	$\forall P_i$. P_i•x_1, $\wedge$ $\forall DMS_k$. (P_i, (P_i•x_2 => inf•(deg•x_2•DMS_k)•(deg•x_1•DMS_k)) $\vee$ $\neg P_i$•x_2
R3	*merlette*	mâle'•x_1	femelle•x_2
R4	*poulet*	animal'•x_1	inf•(deg•x_2•âge')•(deg•x_1•âge')
R5	*castorette*	fourrure'•x	inf•(deg•x_2•valeur')•(deg•x_1•valeur')
R6	*roitelet*	exercer'•x_1•pouvoir' => (R_1…R_n)	exercer'•x_2 •pouvoir' => ¬(R_1…R_n)
R7	*vinaigrette*	substance'•x_1	substance'•x_1•(dans•x_2) $\wedge$ préparation'•x_2 $\wedge$ ¬majoritaire'•x_1•(dans•x_2)
R8	*couchette*	V•e•x_1…	V•e•x_1•(dans•x_2) $\wedge$ temporaire'•e
R9	*giclette, causette*	V•e_1•x…	V•e_2•x… $\wedge$ $\forall DMS_k$. inf•(deg•e_2•DMS_k)•(deg•e_1•DMS_k)
R10	*réformette*	V•e•x_1 $\wedge$ effet-de'•e_1•e_2	V'•e•x_3 $\wedge$ effet-de'•e_3•e_4 $\wedge$ $\forall DMS_k$. inf•(deg•e_4•DMS_k)•(deg•e_2•DMS_k)

	EXEMPLE	DÉCLENCHEUR	APPORT SÉMANTIQUE
L1g	*serpette*	objet-fonctionnel'•x	maniable'•x
L1v	*camionnette*	véhicule'•x	manœuvrable'•x
L1f	*chinchillette*	imitation-fourrure'•x	abordable'•x
L1h	*talonnette*	partie -vêtement'•x	améliorer'•e•x•y $\wedge$ vêtement'•y
L2	*baladurette*	mesure-adm'•x	favoriser'•e•x•y $\wedge$ mesure-adm•x
L3	*bergeronnette*	espèce-nat'•x	dans•x•y $\wedge$ sphère-de-loc •y
L4	*rillettes*	aliment'•y	goût'•z•y $\wedge$ apprécier'•e•x•z $\wedge$ loc'•x
L5	*fripounette*	personne'•x	avoir-affection-pour'•e•x•y $\wedge$ loc'•x

Fradin et al 2003 : critères d'analysabilité

N°	Période	Analysable	Patron	Exemple	Type
1	synchronie	oui	oui	*clochette, réformette, beurette*	T
2	synchronie	oui	non	a) *vignette* b) *galette* < a fr. *gal* c) *baguette* < it. *bacchetta*	O1
3	synchronie	non	non	a) *houlette, galet, clarinette* b) *squelette, silhouette, disette*	O2
4	diachronie	oui	oui	a) *vignette* b) *galette, meurette* c) *houlette, galet, clarinette*	O3
5	diachronie	non	non	a) *squelette* < grec b) *pamphlet, silhouette* c) *baguette* d) *disette, assiette*	O4

Tableau 3. Opacité/Transparence

- synchronie VS diachronie
- analysabilité formelle
- analysabilité sémantique

Fradin et al 2003 : « statistiques globales »

	occurrences	lemmes	hapax
-et	213 265	570	242
-ette	57 176	575	155
total	270 441	1 145	397

Tableau 4. Statistiques pour les données brutes

- non-mots
- coquilles

	occurrences	lemmes
-et	305	177
-ette	525	62
total	830	239

Tableau 5. Nombre de formes exclues par les critères formels.

- mots étrangers
- emprunts
- apocopes
- homographie

	occurrences	lemmes
-et	152 493	233
-ette	21 112	116
total	173 605	349

Tableau 6. Nombre de formes exclues par les critères morphologiques

	occurrences			lemmes		
	CT	CO	total	CT	CO	total
-et	8 594	29 141	37 735	110	42	151
-ette	23 387	6 730	30 117	340	56	393
total	31 981	35 871	67 852	450	98	546

Tableau 7. Nombre de lexèmes construits.

Fradin et al 2003 : résultats

	hapax		occurrences		productivité	
	-et	*-ette*	*-et*	*-ette*	*-et*	*-ette*
R1	8	15	2 632	9 071	0,0038	0,0018
R2	1	4	74	2 802	0,0116	0,0014
R3	0	17	68	389	0	0,0437
R4	1	0	976	1 039	0,0010	0
R5	0	3	0	5	n.d.	0,6
R6	1	2	28	126	0,0357	0,0159
R7	2	0	13	51	0,1538	0
R8	0	0	0	469	n.d.	0
R9	1	10	1	573	1	0,0175
R10	2	1	5	84	0,4	0,0119
total R	16	52	3 809	14 522	0,0042	0,0036
L1g	3	10	4 757	10 860	0,0006	0,0009
L1v	0	0	22	1 626	0	0
L2	0	3	0	175	n.d.	0,0171
L3	2	6	708	544	0,0028	0,0110
L4	2	3	148	142	0,0135	0,0211
L5	6	11	74	71	0,0811	0,1549
total L	13	33	5 492	13 237	0,0024	0,0025
total	23	80	8 594	23 387	0,0027	0,0034

Tableau 9. Productivité des lexèmes CT.

	hapax		occurrences		productivité	
	-et	*-ette*	*-et*	*-ette*	*-et*	*-ette*
R1	0	1	13 593	1 866	0	0,0005
R2	1	0	1	148	1	0
R3	0	0	0	10	n.d.	0
R4	0	0	23	0	0	n.d.
R5	0	0	0	0	n.d.	n.d.
R6	0	0	0	0	n.d.	n.d.
R7	0	0	0	0	n.d.	n.d.
R8	0	0	0	0	n.d.	n.d.
R9	0	0	0	153	n.d.	0
R10	0	0	0	0	n.d.	n.d.
total R	1	1	13 617	2 177	0,0001	0,0005
L1g	1	1	7 926	1 655	0,0001	0,0006
L1v	0	0	152	73	0	0
L2	0	0	93	0	0	n.d.
L3	0	0	885	21	0	0
L4	0	0	0	314	n.d.	0
L5	0	0	23	725	0	0
total L	1	1	8 290	2 788	0,0001	0,0004
total	2	2	29 141	6 677	0,0001	0,0003

Tableau 11. Productivité pour les lexèmes construits opaques

3 809 14 522

	hapax		occurrences	productivité	
	-et	*-ette*	*-et* et *-ette*	*-et*	*-ette*
pôle R	*16*	*33*	3 809	0,0042	0,0087
pôle L	*13*	*25*	5 492	0,0024	0,0046
total CT	*23*	*55*	8 594	0,0027	0,0064

Tableau 10. Productivité avec des sous-corpus de tailles différentes.

Dal et al 2008: le projet « Productivité morphologique » du GDR 2020

- Objectif :
  - « construire un […] outil permettant de mesurer la productivité » (p. 1527)
- Corpus:
  - articles du quotidien français Le Monde
  - années 1995 et 1999
  - 23 millions occurrences

- 8 affixes:
  - in-
  - -able, -eux, -ique, -if,
  - -ion, -oir(e)
  - -ifier

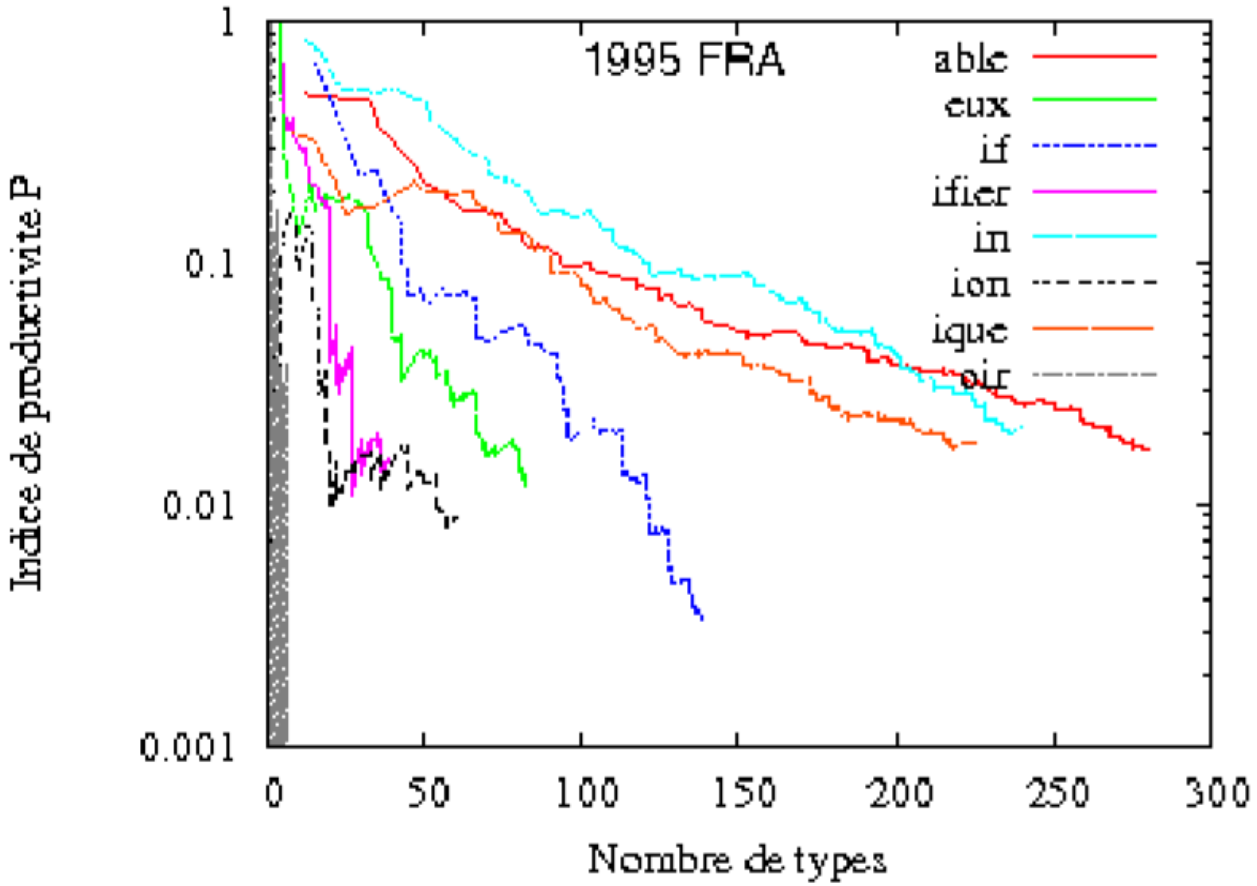
	Rubrique	articles	occurrences	articles	occurrences
		1995		1999	
AGE	Agenda	1 213	490 663	1 776	650 440
ART	Événements culturels	4 242	1 801 044	4 809	2 018 424
FRA	France	6 331	2 704 350	4 264	1 948 253
INT	International	9 276	3 065 884	8 083	3 211 160
LIV	Livres	1 949	1 350 540	2 388	1 280 437
RTV	Programme TV et radio	1 217	718 586	22	5 471
SOC	Société	4 009	1 678 573	2 823	1 260 588
SPO	Événements sportifs	2 362	894 648	2 825	911 162
Total		30 599	12 500 000	26 988	11 000 000

Tableau 1 : Taille des sous-corpus formés par les rubriques du Monde en 1995 et 1999.

Type d'analysabilité	Exemples de lexèmes
1. Analysabilité synchronique	ACCEPTABLE (< ACCEPTER) MANGEABLE (< MANGER)
2. Analysabilité diachronique (diatopique)	ACCEPTABLE (< lat. ACCEPTABILIS) FRIABLE (< lat. FRIABILIS))
3. Analysabilité indirecte	AUTOCASSABLE (< CASSABLE < CASSER) IMMANGEABLE (< MANGEABLE < MANGER)

Tableau 2 : Types d'analysabilité.

Dal et al 2008 : observer la productivité

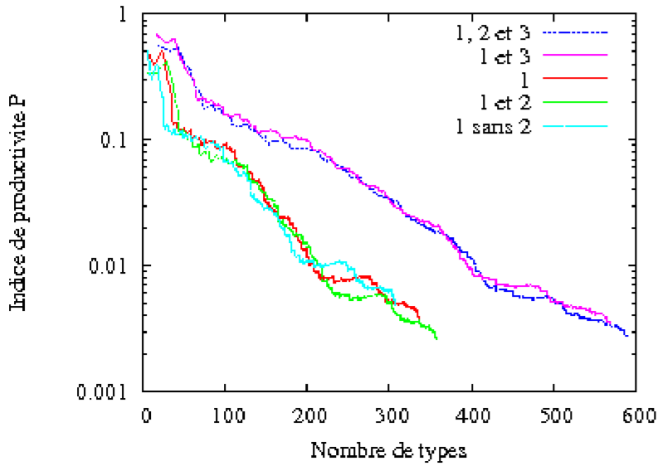


- Productivité = nombre d'hapax / fréquence cumulée du procédé morphologique
- Accroissement du vocabulaire (nombre de lexèmes)
- 3 degrés de productivité:
  - in-, -able, -ique
  - -ion, -if, -ifier, -eux
  - -oir(e)

Dal et al 2008 : productivité différenciée

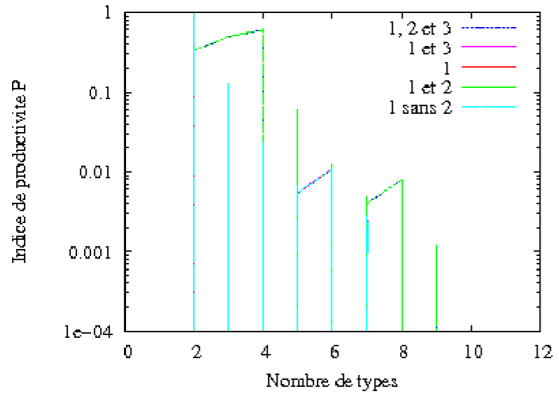
Catégorie(s)	Lexèmes en *-able* retenus
1	ACCEPTABLE, MANGEABLE, ~~FRIABLE~~, ~~AUTOCASSABLE~~, ~~IMMANGEABLE~~
1 ∪ 2	ACCEPTABLE, MANGEABLE, FRIABLE, ~~AUTOCASSABLE~~, ~~IMMANGEABLE~~
1 ∪ 2 ∪ 3	ACCEPTABLE, MANGEABLE, FRIABLE, AUTOCASSABLE, IMMANGEABLE
1 ∪ 3	ACCEPTABLE, MANGEABLE, ~~FRIABLE~~, AUTOCASSABLE, IMMANGEABLE
1 ¬ 2	~~ACCEPTABLE~~, MANGEABLE, ~~FRIABLE~~, ~~AUTOCASSABLE~~, ~~IMMANGEABLE~~

7.3 Productivité de la suffixation en *-able*



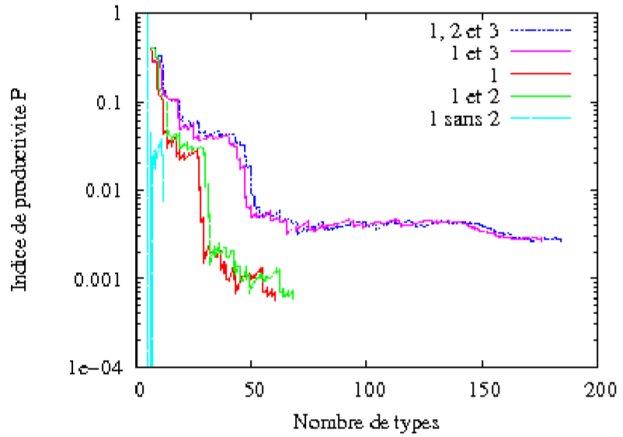
-> autant d'hapax indirects que directs (haut degré)

7.5 Productivité de la suffixation en *-oir(e)*



-> peu d'occurrences (bas degré)

7.4 Productivité de la suffixation en *-ion*



-> plus d'hapax indirects et moins de lexèmes indirects fréquents (degré intermédiaire)

Dal et al 2008 : un projet inachevé

- Un projet ambitieux:
  - 8 membres permanents/réguliers
  - compétences disciplinaires variées
  - clôture en décembre 2007
- Les causes de l'échec:
  - méthode trop qualitative
  - pas de base de donnée de référence
  - pas de critère d'estimation des fréquences

La fréquence et les bases de données lexicales en psycholinguistique

La psycholinguistique et les bases de données lexicales

- Les stimuli lexicaux
  - expérimentations
  - temps de décision lexicale
  - corrélations des différents facteurs (propriétés des locuteurs et propriétés des stimulis)
- La fréquence d'occurrence
  - Le facteur le plus important
  - estimée à partir de corpus

Utiliser les normes de fréquence pour mesurer la productivité morphologique?

- estimation de fréquence -> estimation de productivité?
- base de données de fréquence
  - économiser le corpus?
  - qualité des données
  - Reproductibilité et partage
- fiabilité expérimentale des estimations
  - connaissance lexicale des locuteurs?
  - Expérimenter les hypothèses

Evolution qualitative des bases de données lexicales fréquencées du français

- Brulex:
  - 1986
  - entrées intermédiaires entre lemmes et formes fléchies
  - vocabulaire extrait de la nomenclature du Micro-Robert
  - fréquences littéraires (1950-1964)
  - pas de corpus accessible
  - transcriptions phonologiques
  - informations diverses
- Lexique 1:
  - 2001
  - formes orthographiques et lemmes graphiques
  - corpus littéraire accessible
  - fréquences Web
- Lexique 2:
  - 2004
  - correction des transcriptions phonologiques
- Lexique 3 :
  - 2007
  - corpus plus récent
  - estimations plus proches de la langue parlée
  - entrées formes fléchies morphomiques et champs ISLEM
  - nouvelles corrections phonologiques
- PsychoGLàFF
  - 2014
  - formes fléchies annotées
  - vocabulaire issu de la nomenclature du wiktionnaire
  - fréquences issues de trois corpus très différents

Evolution quantitative des bases de données lexicales fréquencées du français

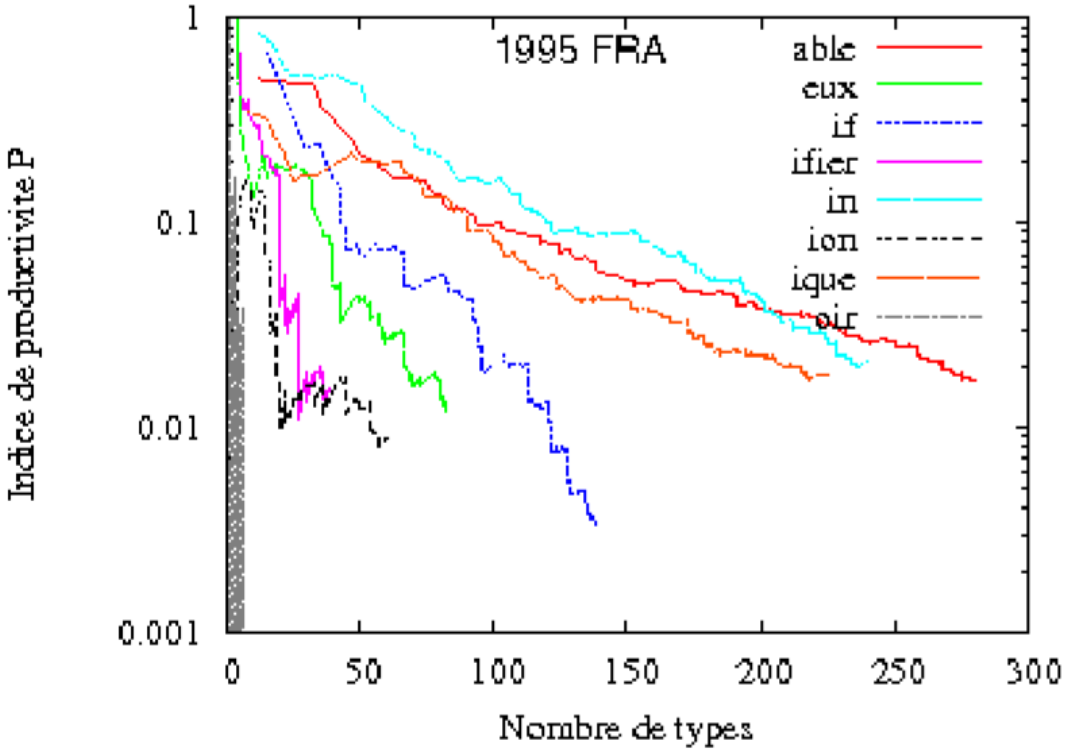
VERSION	NB de graphies	NB de formes fléchies	Nombre de lemmes	Taille du corpus (M de mots)
Brulex	???	???	35 746	8 (Frantext)
Lexique 1 et 2	130 000	???	???	30 (Frantext)
Lexique 3 (films)	128 000	147 000	42 000	50 (sous-titres)
PsychoGlàff	???	337 572	121 021	1600 (FrWaC)

Estimer la productivité morphologique à partir d'une norme de fréquence

Objectifs

- estimer la valeur de la norme de fréquence:
  - estimer la productivité morphologique
  - base de donnée de référence générale
    - répertoire des procédés morphologiques courants
    - échelle de productivité
    - pivot de comparaison pour les études en corpus
- mesure préliminaire
  - pas de classement sémantique
  - pas de tri morphologique
  - critères de sélection grossiers
    - exposant
    - Catégorie
  - indices inexacts (pas de normalisation du nombre d'hapax)

Comparaison Lexique 3 et GDR 2020 (sept suffixes)



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

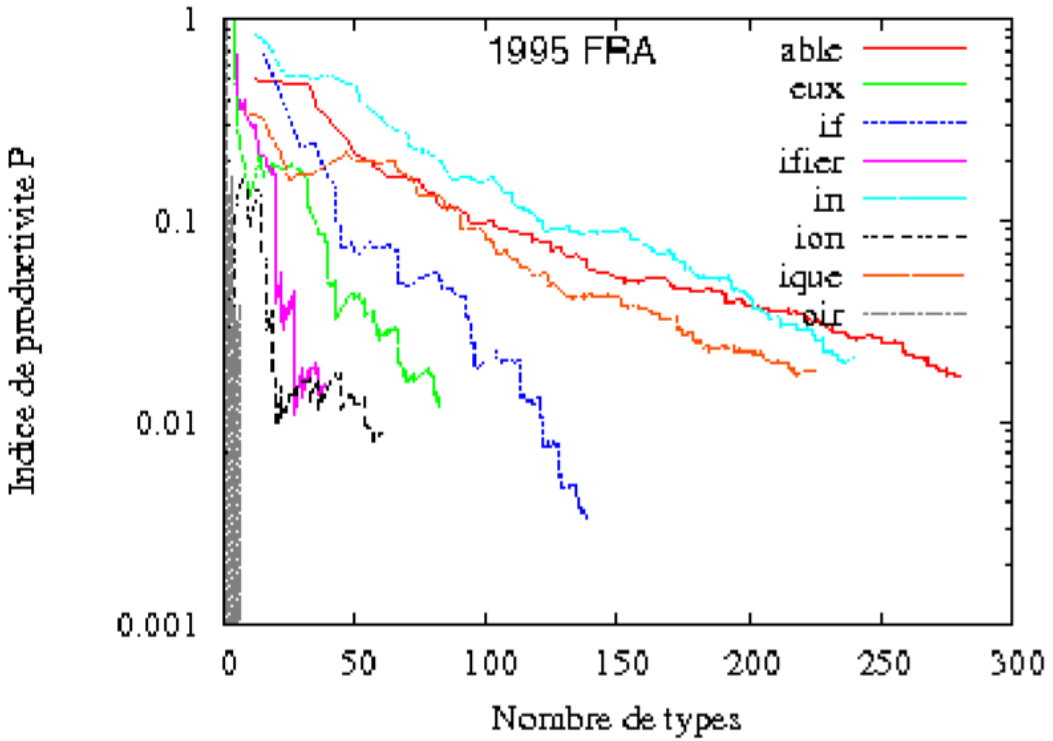
2M de mots

50M de mots

Premiers constats

- L’échelle de productivité est cohérente
  - ique et able sont les plus productifs
  - oir(e) est le moins productif
- Il y a des différences notables dans les mesures
  - effectifs de lemmes
  - Un nouveau découpage?
    - 1) ique, able, if
    - 2) ion, ifier
    - 3) eux, oir(e)
  - indices trop faibles
- Des différences importantes dans les sources:
  - journaux VS sous-titres de films
  - taille des corpus (* 10 à *2 -> 2M, 23M, 50M)

Comparaison de PsychoGLàFF et GDR 2020



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

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:

- inversion des extrèmes
- Valeurs trop petites

Diagnostic : la fréquence de types

LEXIQUE383films

SUF	nbTyp
ion_NOM	1556
ique_ADJ	911
able_ADJ	450
eux_ADJ	350
if_ADJ	237
oir(e)_NOM	152
ifier_VER	72
TOTAL	3728

PsychoGLàFF_LM10

SUF	nbTyp
ion_NOM	1132
ique_ADJ	1052
able_ADJ	475
eux_ADJ	453
if_ADJ	199
oir(e)_NOM	138
ifier_VER	107
TOTAL	3556

PsychoGLàFF_FrWaC

SUF	nbTyp
ique_ADJ	1330
ion_NOM	1232
able_ADJ	1034
eux_ADJ	272
oir(e)_NOM	240
if_ADJ	212
ifier_VER	145
TOTAL	4465

- stabilité entre les corpus
  - films/journaux VS web -> taille?
  - ion/ique; if/oir(e)
- homogénéité avec la productivité
  - ion; ifier

Diagnostic : la fréquence d'hapax

LEXIQUE383films

SUF	nbHap
ique_ADJ	189
ion_NOM	178
able_ADJ	68
eux_ADJ	23
if_ADJ	20
oir(e)_NOM	13
ifier_VER	7
TOTAL	498

PsychoGLàFF_LM10

SUF	nbHap
ique_ADJ	177
ion_NOM	166
able_ADJ	98
eux_ADJ	36
if_ADJ	33
oir(e)_NOM	31
ifier_VER	30
TOTAL	571

PsychoGLàFF_FrWaC

SUF	nbHap
ique_ADJ	120
ion_NOM	113
able_ADJ	64
ifier_VER	40
oir(e)_NOM	39
eux_ADJ	28
if_ADJ	10
TOTAL	414

stabilité:

- films et journaux VS web -> taille?
- Ique/ion/able VS oir(e)/ifier

Diagnostic : fréquence cumulée

LEXIQUE383fil
ms

SUF	nbOcc	productivité
ique_ADJ	1026,6	0,0037
able_ADJ	1012,45	0,0013
if_ADJ	334,34	0,0012
ion_NOM	5383,05	0,00066
ifier_VER	350,78	0,00040
eux_ADJ	2795,32	0,00016
oir(e)_NOM	2178,32	0,00012

50M

PsychoGLàFF_
LM10

SUF	nbOcc	productivité
ifier_VER	15,0325202	0,010
oir(e)_NOM	18,5404116	0,0083
able_ADJ	79,7715428	0,0061
ique_ADJ	185,595208	0,0048
ion_NOM	268,983837	0,00309
if_ADJ	53,5055805	0,00308
eux_ADJ	1852,69897	9,7E-05

200M

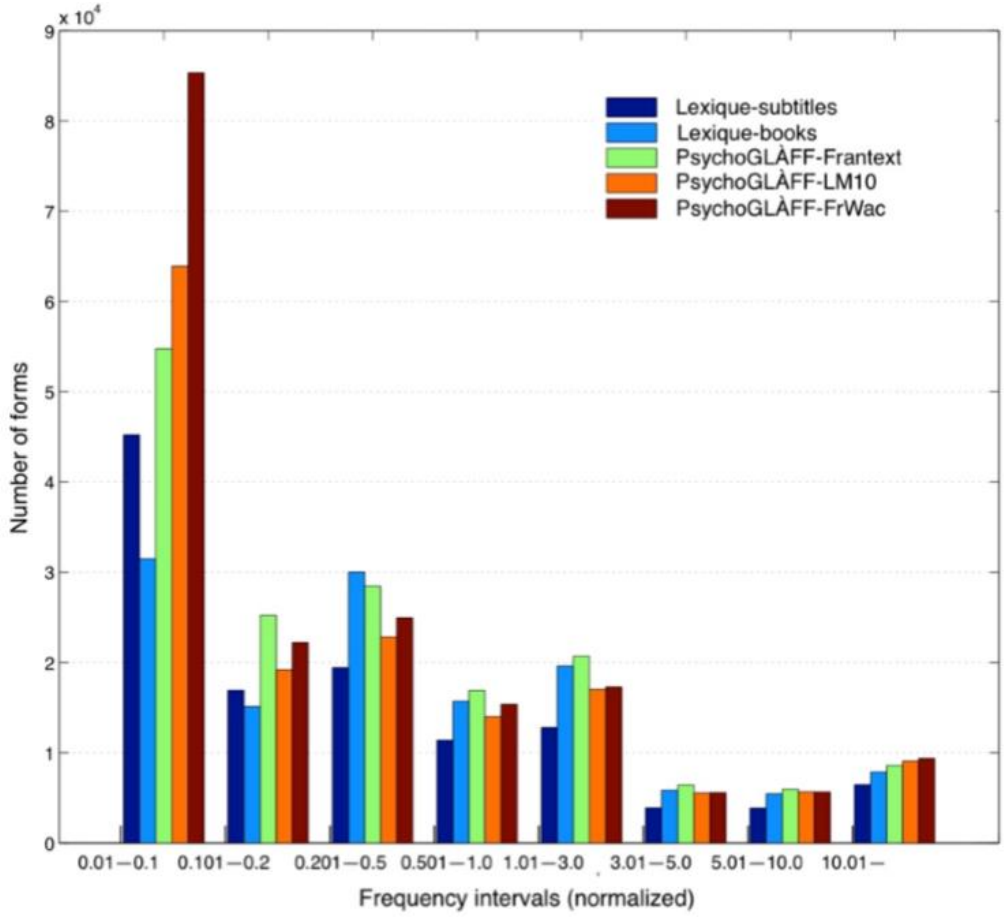
PsychoGLàFF_
FrWaC

SUF	nbOcc	productivité
ifier_VER	7,34634078	0,0034
oir(e)_NOM	16,8716201	0,0014
ique_ADJ	94,7008806	0,00079
ion_NOM	113,317566	0,00062
eux_ADJ	32,8252467	0,00053
if_ADJ	18,2506043	0,00034
able_ADJ	2125,09995	1,9E-05

1600M

- taille des corpus -> lexique3?
- lexèmes de haute fréquence?

Le problème des hapax



-> filtrage des hapax

retour au corpus?

La base de donnée Lexique 3

ACCUEIL / RECHERCHES EN LIGNE / AIDE / CONTRIBUTEURS / BLOG

Lexique

Lexique est une base de données qui fournit, pour 140 000 mots de la langue française, diverses informations. Par exemple, elle fournit notamment les **fréquences** d'occurrences dans différents corpus, la représentation phonologique, les lemmes associés, le nombre de syllabes, la catégorie grammaticale, et bien d'autres informations.

Recherche en ligne dans Lexique 3.83

Recherche hors-ligne

Télécharger Lexique 3.83

RECENT POSTS

Refonte 2019 du site

WorldLex word frequencies is working again !

Lexique is back !

*Note to **non-French speakers:** to read these web pages in your native language, add the google translate extension to your browser .*

Lexique 3 -
10 premières entrées avec affichage par défaut

Columns to display

- [x] ortho
- [x] phon
- [x] lemme
- [x] cgram
- [ ] genre
- [ ] nombre
- [x] freqlemfilms2
- [ ] freqlemlivres
- [x] freqfilms2
- [ ] freqlivres
- [ ] infover
- [ ] nbhomogr
- [ ] nbhomoph
- [x] islem
- [x] nblettres
- [ ] nbphons
- [ ] cvcv
- [ ] p_cvcv
- [ ] voisorth
- [ ] voisphon
- [ ] puorth
- [ ] puphon
- [ ] syll
- [x] nbsyll
- [ ] cv-cv
- [ ] orthrenv
- [ ] phonrenv
- [ ] orthosyll
- [x] cgramortho
- [ ] deflem
- [ ] defobs
- [ ] old20
- [ ] pld20
- [ ] morphoder
- [ ] nbmorph

Lexique3

Show 10 entries

ortho	phon	lemme	cgram	freqlemfilms2	freqfilms2	islem	nblettres	nbsyll	cgramortho
All	All	All	All	All	All	All	All	All	All
a	a	a	NOM	81.36	81.36	1	1	1	NOM,AUX,VER
a	a	avoir	AUX	18559.22	6350.91	0	1	1	NOM,AUX,VER
a	a	avoir	VER	13572.4	5498.34	0	1	1	NOM,AUX,VER
a capella	akapEla	a capella	ADV	0.04	0.04	1	9	4	ADV
a cappella	akapEla	a cappella	ADV	0.04	0.04	1	10	4	ADV
a contrario	ak§tRaRjo	a contrario	ADV	0	0	1	11	4	ADV
a fortiori	afORsjoRi	a fortiori	ADV	0.04	0.04	1	10	4	ADV
a giorno	adZjORno	a giorno	ADV	0	0	1	8	3	ADV
a jeun	aZ1	à jeun	ADV	1.45	0.18	0	6	2	ADV
a l'instar	al5staR	a l'instar	PRE	0.26	0.26	1	10	3	PRE

Showing 1 to 10 of 142,694 entries

Previous 1 2 3 4 5 … 14270 Next

Download filtered data

affichage des champs orthographiques

ortho	nbhomogr	nblettres	cvcv	voisorth	puorth	orthrenv	orthosyll
All	All	All	All	All	All	All	All
a	3	1	V	25	1	a	a

affichage des champs phonologiques

phon	nbhomoph	nbphons	p_cvcv	voisphon	puphon	syll	nbsyll	cv-cv	phonrenv
All	All	All	All	All	All	All	All	All	All
a	9	1	V	20	1	a	1	V	a

affichage des champs grammaticaux

ortho	cgram	genre	nombre	infover	cgramortho
All	All	All	All	All	All
a	NOM	m			NOM,AUX,VER

affichage des champs de fréquence

10 premières entrées de Lexique 3 des formes de l'auxiliaire AVOIR (par freqfilm décroissant)

Lexique3

Show 10 entries

ortho	phon	lemme	cgram	freqlemfilms2	freqfilms2	infover	islem
All	All	^avoir	AUX	All	All	All	All
a	a	avoir	AUX	18559.22	6350.91	ind:pre:3s;	0
ai	E	avoir	AUX	18559.22	4902.1	ind:pre:1s;	0
as	a	avoir	AUX	18559.22	2144.15	ind:pre:2s;	0
avez	ave	avoir	AUX	18559.22	1122.37	ind:pre:2p;	0
ont	§	avoir	AUX	18559.22	1063.32	ind:pre:3p;	0
avoir	avwaR	avoir	AUX	18559.22	674.24	inf;	1
avais	avE	avoir	AUX	18559.22	412.04	ind:imp:2s;	0
avait	avE	avoir	AUX	18559.22	395.71	ind:imp:3s;	0
aurais	oRE	avoir	AUX	18559.22	354.36	cnd:pre:2s;	0
avons	av§	avoir	AUX	18559.22	291.71	ind:pre:1p;	0

Showing 1 to 10 of 46 entries (filtered from 142,694 total entries)

Previous 1 2 3 4 5 Next

Download filtered data

Les problèmes du champs MorphoDer (10 premières entrées lemmes de Lexique 3 en –ion)

Lexique3

Show 10 entries

lemme	cgram	freqlemfilms2	islem	morphoder	nbmorph
.*ion	All	All	1 ... 1	All	All
abdication	NOM	0.05	1	abdiquer-ion → abdiqu-ation	2
abduction	NOM	0.05	1	abduct-ion → abduire-ion	2
aberration	NOM	1.16	1	a-ber-ion → aberration	1 ✗
abjection	NOM	0.51	1	abjection	1
abjuration	NOM	0	1	abjurer-ion	2
ablation	NOM	0.45	1	ablation	1
ablution	NOM	0.48	1	ablution	1
abnégation	NOM	0.91	1	abnégation	1
abolition	NOM	0.45	1	abolir-ion	2
abolitionniste	ADJ	0.03	1	abolitionniste → abolir-ion-iste	3 ✗

Showing 1 to 10 of 2,207 entries (filtered from 142,694 total entries)

Previous 1 2 3 4 5 ... 221 Next

Download filtered data

La chaîne de traitement de Lexique 4

La mise à jour de Lexique 3

Contribuer

Lexique et OpenLexicon sont des projets collaboratifs auxquels tout le monde est encouragé à participer. N'hésitez pas à poser des questions sur le forum, et à proposer des améliorations du code (shiny apps, scripts, …) sur le site github d'Openlexicon.

lexique.org, « Accueil »

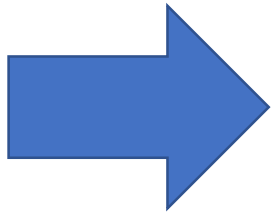
-Morphologie Dérivationnelle (*morphoder*) Ce champs donne la décomposition en morphèmes dérivationnels d'un mot donné. Ainsi *plumage* est décomposé en *plume-age*. Ce champs est le résultat du programme Dérif (Namer, 2003; http://www.cnrtl.fr/outils/DeriF/). Attention pour la version actuelle de ce programme de nombreux suffixes et préfixes étant encore non traités ou traités partiellement). Par exemple, *abandonner* n'est pas ségmenté comme *abandon-er* mais comme un monomorphémique (*abandonner*). Nous sommes donc vivement intéressés par toute contribution concernant ce champs.

-Nombre de morphèmes (dérivationnels) (*nbmorph*) C'est le nombre de morphèmes dérivationnels directement calculé à partir du champs précédent.

Manuel de Lexique3, p. 18

Lexique 4

- Nouveau corpus
- Nouvelle structure
- Nouvelle chaîne de traitement
- Nouvelle nomenclature
- Nouvelles valeurs de fréquence lexicale
- Base ortho
- Base phono
- Contextual diversity
- Nouvelles valeurs des champs de morphologie dérivationnelle



Gauvain

- Mettre à jour Lexique 3
- Pérenniser la chaîne de traitement


Gauvain
Extraction du lexique
Correction du lexique
Acquisition des phonologies
Acquisition des fréquences
Ajouter une autre carte
En cours
Mettre à jour Lexique 3
Ajouter une autre carte

OPUS ... the open parallel corpus

OPUS is a growing collection of translated texts from the web. In the OPUS project we try to convert and align free online data, to add linguistic annotation, and to provide the community with a publicly available parallel corpus. OPUS is based on open source products and the corpus is also delivered as an open content package. We used several tools to compile the current collection. All pre-processing is done automatically. No manual corrections have been carried out.

The OPUS collection is growing! Check this page from time to time to see new data arriving ...
Contributions are very welcome! Please contact <jorg.tiedemann@helsinki.fi >

OpenSubtitles

This is a new collection of translated movie subtitles from http://www.opensubtitles.org/.

IMPORTANT: If you use the OpenSubtitle corpus: Please, add a link to **http://www.opensubtitles.org/** to your website and to your reports and publications produced with the data! I promised this when I got the data from the providers of that website!

This is a slightly cleaner version of the subtitle collection using improved sentence alignment and better language checking.

62 languages, 1,782 bitexts
total number of files: 3,735,070
total number of tokens: 22.10G
total number of sentence fragments: 3.35G

Please cite the following article if you use any part of the corpus in your own work:
P. Lison and J. Tiedemann, 2016, *OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles.* In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)

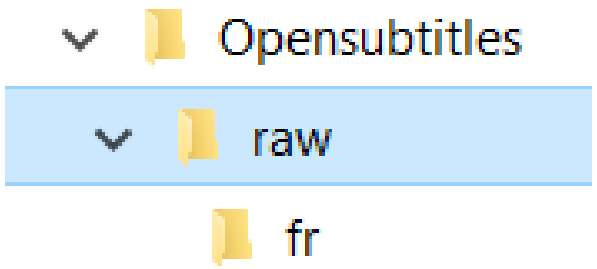
OpenSubtitles v2018 - Intra-Lingual Alignments

The following table lists alignments between subtitles in the same language. There are often various alternative subtitle files for each movie in the collection. Many of them are identical or near identical. We have processed them all and sorted the results in various ways. The resulting files are linked in the table for each language. Here is an explanation of the different columns:

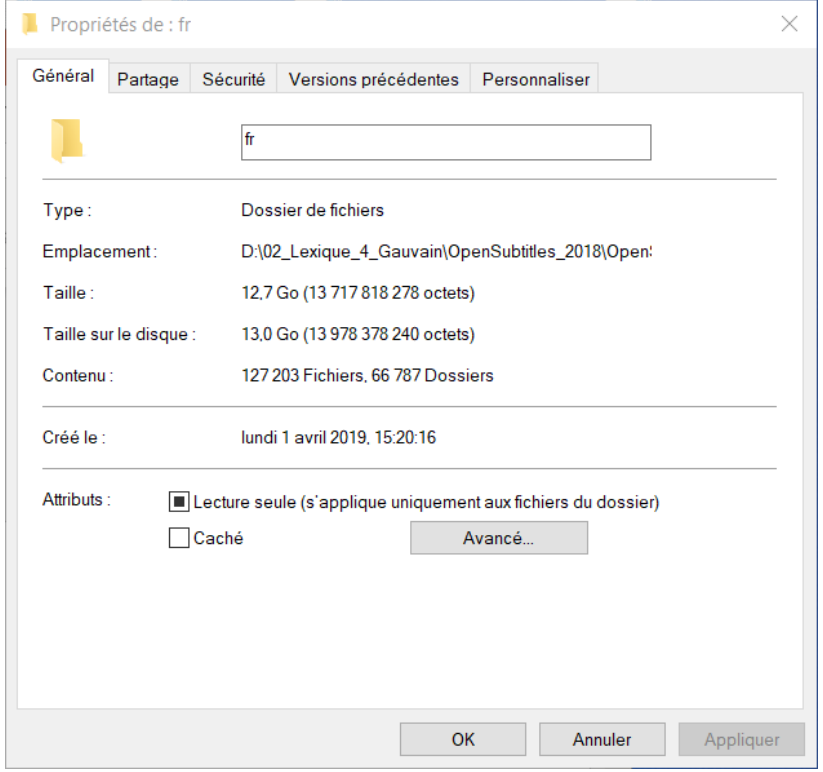
- **corpus:** This is a compressed tar-archive with all movie subtitles for the given language in XML format.
- **all:** All links between alternative subtitle files except the ones classified as "misaligned".
- **insert:** Sentence pairs that differ only by some inserted text on one side
- **misaligned:** Probably misaligned sentences (low lexcial overlap and large differences in lengths).
- **other:** Other types of sentence pairs; probably paraphrases and/or stylistically different subtitles.
- **pct:** Sentence pairs that differ only in punctuation characters.
- **spell:** Sentence pairs that differ in a few characters only that looks suspiciously like misspellings.

Some alignment files exist as XCES only (standoff annotation of sentence alignment) and some of them are also available in TMX format (to make it easier to inspect the actual sentence pairs). If you use the XCES alignment files, then you will also need the corpus, which is linked in the first column.

language	corpus	all	insert	misaligned	other	pct	spell
af	zip	xml tmx		xml	xml tmx		
ar	zip	xml	xml tmx	xml	xml tmx	xml tmx	xml tmx
bg	zip	xml	xml tmx	xml	xml tmx	xml tmx	xml tmx
bn	zip	xml	xml tmx	xml	xml tmx	xml tmx	xml tmx
br	zip	xml			xml tmx		xml tmx
bs	zip	xml	xml tmx	xml	xml tmx	xml tmx	xml tmx
ca	zip	xml	xml tmx	xml	xml tmx	xml tmx	xml tmx
cs	zip	xml	xml tmx	xml	xml tmx	xml tmx	xml tmx
da	zip	xml	xml tmx	xml	xml tmx	xml tmx	xml tmx
de	zip	xml	xml tmx	xml	xml tmx	xml tmx	xml tmx
el	zip	xml	xml tmx	xml	xml tmx	xml tmx	xml tmx
en	zip	xml	xml tmx	xml	xml tmx	xml tmx	xml tmx
eo	zip	xml		xml		xml tmx	xml tmx
es	zip	xml	xml tmx	xml	xml tmx	xml tmx	xml tmx
et	zip	xml	xml tmx	xml	xml tmx	xml tmx	xml tmx
eu	zip	xml	xml tmx	xml		xml tmx	xml tmx
fa	zip	xml	xml tmx	xml	xml tmx	xml tmx	xml tmx
fi	zip	xml	xml tmx	xml	xml tmx	xml tmx	xml tmx
fr	zip	xml	xml tmx	xml	xml tmx	xml tmx	xml tmx



0	1924	1940	1956	1972	1988	2004
1191	1925	1941	1957	1973	1989	2005
1910	1926	1942	1958	1974	1990	2006
1911	1927	1943	1959	1975	1991	2007
1912	1928	1944	1960	1976	1992	2008
1913	1929	1945	1961	1977	1993	2009
1914	1930	1946	1962	1978	1994	2010
1915	1931	1947	1963	1979	1995	2011
1916	1932	1948	1964	1980	1996	2012
1917	1933	1949	1965	1981	1997	2013
1918	1934	1950	1966	1982	1998	2014
1919	1935	1951	1967	1983	1999	2015
1920	1936	1952	1968	1984	2000	2016
1921	1937	1953	1969	1985	2001	2017
1922	1938	1954	1970	1986	2002	
1923	1939	1955	1971	1987	2003	



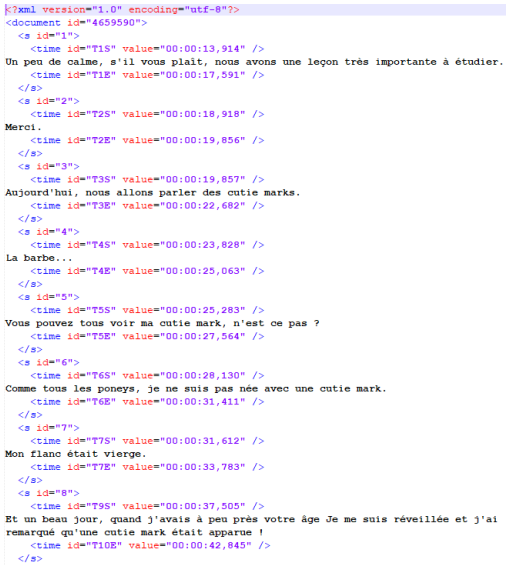

```
<?xml version="1.0" encoding="utf-8"?>
<document id="6328180">
  <s id="1">
    <time id="T1S" value="00:00:00,100" />
Collection de films " ГОСФИЛЬМОФОНДА " U R S S
    <time id="T1E" value="00:00:08,900" />
  </s>
  <s id="2">
    <time id="T2S" value="00:00:09,000" />
LA DAME DE PIQUE (Pikovaïa dama) T/D
  </s>
  <s id="3">
- A. Khanjonkov 1910
    <time id="T2E" value="00:00:18,900" />
  </s>
  <s id="4">
    <time id="T3S" value="00:00:19,000" />
Scénario et réalisation P. Tchardynin Opérateur L.
Foriestié Direction artistique V. Fiestier
    <time id="T3E" value="00:00:28,900" />
  </s>
```

```
  <meta>
    <source>
      <original>Russian</original>
      <genre>Drama,Fantasy,Horror</genre>
      <duration>15</duration>
      <year>1910</year>
      <HD>0</HD>
      <cds>1/1</cds>
      <country>Russia</country>
    </source>
    <subtitle>
      <blocks>14</blocks>
      <date>2015-10-05</date>
      <confidence>1.0</confidence>
      <machine_translated>0</machine_translated>
      <duration>00:15:17,000</duration>
      <version>1</version>
      <language>French</language>
      <rating>1.0</rating>
    </subtitle>
    <conversion>
      <corrected_words>0</corrected_words>
      <encoding>utf-8</encoding>
      <ignored_blocks>0</ignored_blocks>
      <unknown_words>9</unknown_words>
      <truecased_words>6</truecased_words>
      <tokens>160</tokens>
      <sentences>19</sentences>
    </conversion>
  </meta>
</document>
```

sous-titre	ID_sous-titre	ID_année	ID_film	encodage	genre	year	langue
0_1037343	120978	0	1037343	windows-125	Drama;Fantasy;Mystery		French
0_1089124	4588599	0	1089124	windows-125	Animation;Dr	2004	French
0_1171925	4395041	0	1171925	windows-125	Animation;Co	1997	French
0_1171930	4395000	0	1171930	windows-1252			French
0_1209078	4297729	0	1209078	windows-125	Drama		French
0_1210714	4317263	0	1210714	windows-125	Action;Comedy;Crime		French
0_1210715	4317264	0	1210715	windows-125	Action;Comedy;Crime		French
0_1268703	4317267	0	1268703	windows-125	Action;Comedy;Crime		French
0_1297082	4525560	0	1297082	windows-125	Action;Drama	2002	French
0_1392665	4317282	0	1392665	windows-125	Action;Comedy;Crime		French
0_1392666	4317275	0	1392666	windows-125	Action;Comedy;Crime		French
0_1393557	4317276	0	1393557	windows-125	Action;Comedy;Crime		French
0_1393558	4317277	0	1393558	windows-125	Action;Comedy;Crime		French
0_1393559	4317280	0	1393559	windows-125	Action;Comedy;Crime		French
0_1430536	6427163	0	1430536	windows-125	Fantasy;Mystery;Sci-Fi		French
0_1430537	6427162	0	1430537	windows-125	Fantasy;Mystery;Sci-Fi		French
0_1430538	6427164	0	1430538	windows-125	Fantasy;Mystery;Sci-Fi		French
0_1430539	6427159	0	1430539	windows-125	Fantasy;Mystery;Sci-Fi		French
0_1431811	6427161	0	1431811	windows-125	Fantasy;Mystery;Sci-Fi		French

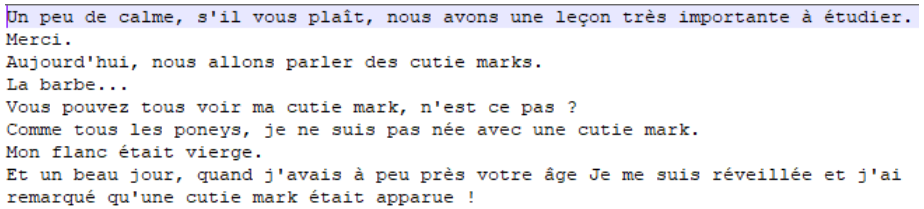
Le fichier n° 18327 (dossier 2011)

TEXTE XML



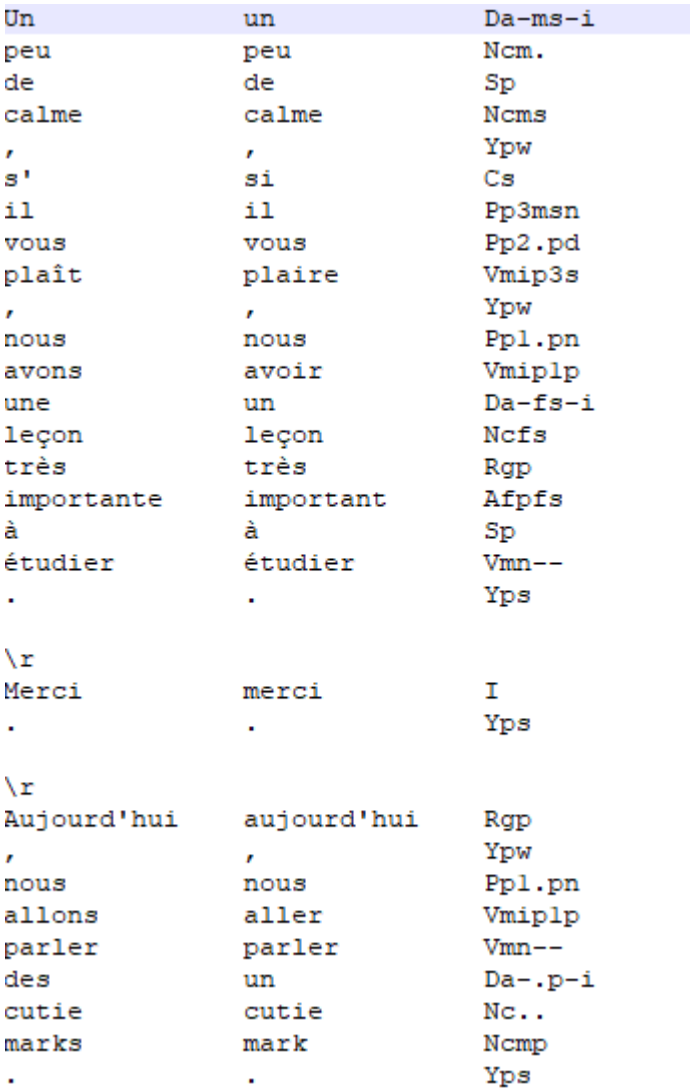
```
<?xml version="1.0" encoding="utf-8"?>
<document id="4659590">
  <s id="1">
    <time id="T1S" value="00:00:13,914" />
Un peu de calme, s'il vous plaît, nous avons une leçon très importante à étudier.
    <time id="T1E" value="00:00:17,591" />
  </s>
  <s id="2">
    <time id="T2S" value="00:00:18,918" />
Merci.
    <time id="T2E" value="00:00:19,856" />
  </s>
  <s id="3">
    <time id="T3S" value="00:00:19,857" />
Aujourd'hui, nous allons parler des cutie marks.
    <time id="T3E" value="00:00:22,682" />
  </s>
  <s id="4">
    <time id="T4S" value="00:00:23,828" />
La barbe...
    <time id="T4E" value="00:00:25,063" />
  </s>
  <s id="5">
    <time id="T5S" value="00:00:25,283" />
Vous pouvez tous voir ma cutie mark, n'est ce pas ?
    <time id="T5E" value="00:00:27,564" />
  </s>
  <s id="6">
    <time id="T6S" value="00:00:28,130" />
Comme tous les poneys, je ne suis pas née avec une cutie mark.
    <time id="T6E" value="00:00:31,411" />
  </s>
  <s id="7">
    <time id="T7S" value="00:00:31,612" />
Mon flanc était vierge.
    <time id="T7E" value="00:00:33,783" />
  </s>
  <s id="8">
    <time id="T9S" value="00:00:37,505" />
Et un beau jour, quand j'avais à peu près votre âge Je me suis réveillée et j'ai
remarqué qu'une cutie mark était apparue !
    <time id="T10E" value="00:00:42,845" />
  </s>
```

TEXTE BRUT

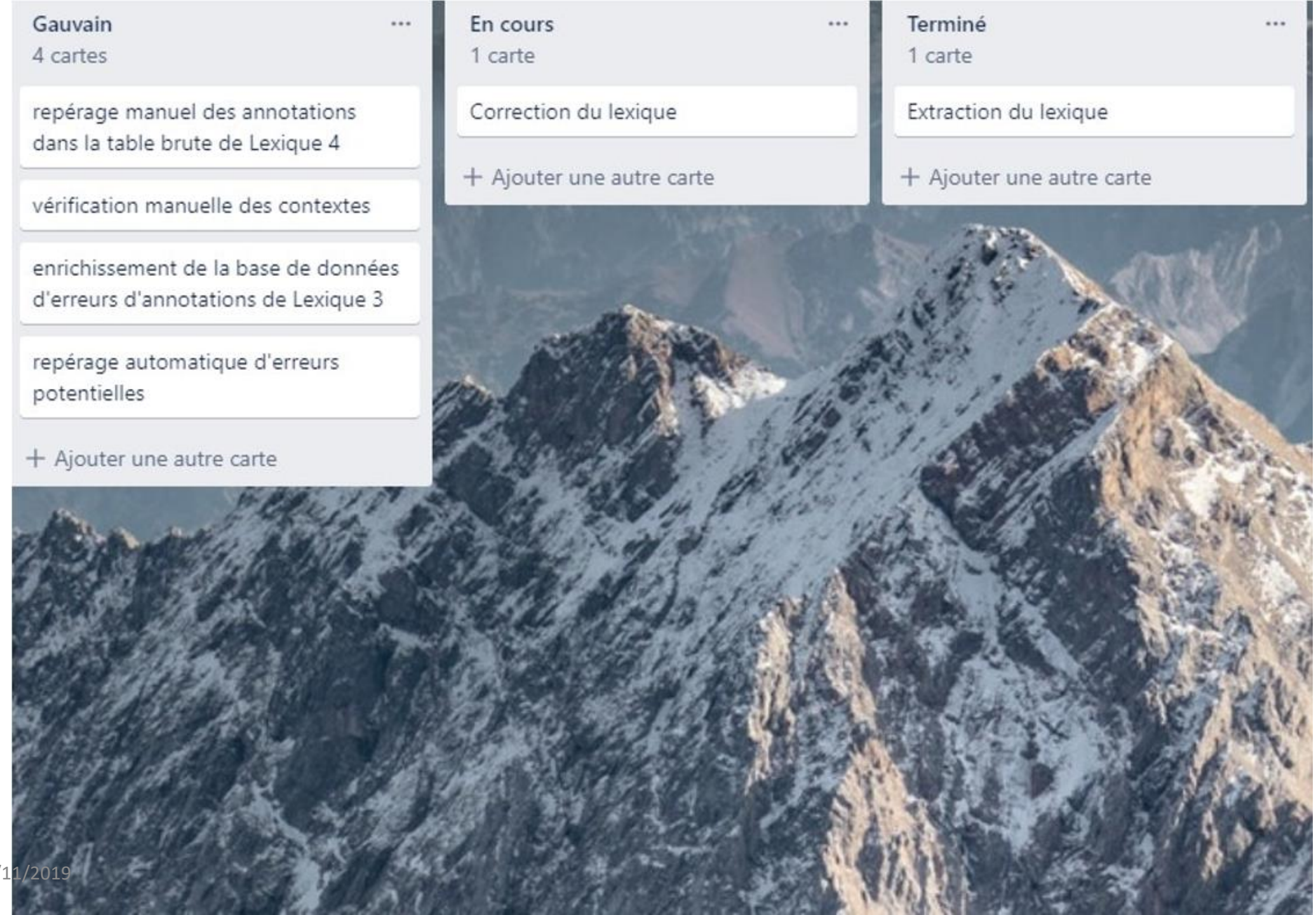


```
Un peu de calme, s'il vous plaît, nous avons une leçon très importante à étudier.
Merci.
Aujourd'hui, nous allons parler des cutie marks.
La barbe...
Vous pouvez tous voir ma cutie mark, n'est ce pas ?
Comme tous les poneys, je ne suis pas née avec une cutie mark.
Mon flanc était vierge.
Et un beau jour, quand j'avais à peu près votre âge Je me suis réveillée et j'ai
remarqué qu'une cutie mark était apparue !
```

TEXTE SEGMENTE ET ANNOTE



Un	un	Da-ms-i
peu	peu	Ncm.
de	de	Sp
calme	calme	Ncms
,	,	Ypw
s'	si	Cs
il	il	Pp3msn
vous	vous	Pp2.pd
plaît	plaire	Vmip3s
,	,	Ypw
nous	nous	Ppl.pn
avons	avoir	Vmiplp
une	un	Da-fs-i
leçon	leçon	Ncfs
très	très	Rgp
importante	important	Afpfs
à	à	Sp
étudier	étudier	Vmn--
.	.	Yps
\r		
Merci	merci	I
.	.	Yps
\r		
Aujourd'hui	aujourd'hui	Rgp
,	,	Ypw
nous	nous	Ppl.pn
allons	aller	Vmiplp
parler	parler	Vmn--
des	un	Da-.p-i
cutie	cutie	Nc..
marks	mark	Ncmp
.	.	Yps


Gauvain
4 cartes
repérage manuel des annotations dans la table brute de Lexique 4
vérification manuelle des contextes
enrichissement de la base de données d'erreurs d'annotations de Lexique 3
repérage automatique d'erreurs potentielles
+ Ajouter une autre carte
En cours
1 carte
Correction du lexique
+ Ajouter une autre carte
Terminé
1 carte
Extraction du lexique
+ Ajouter une autre carte

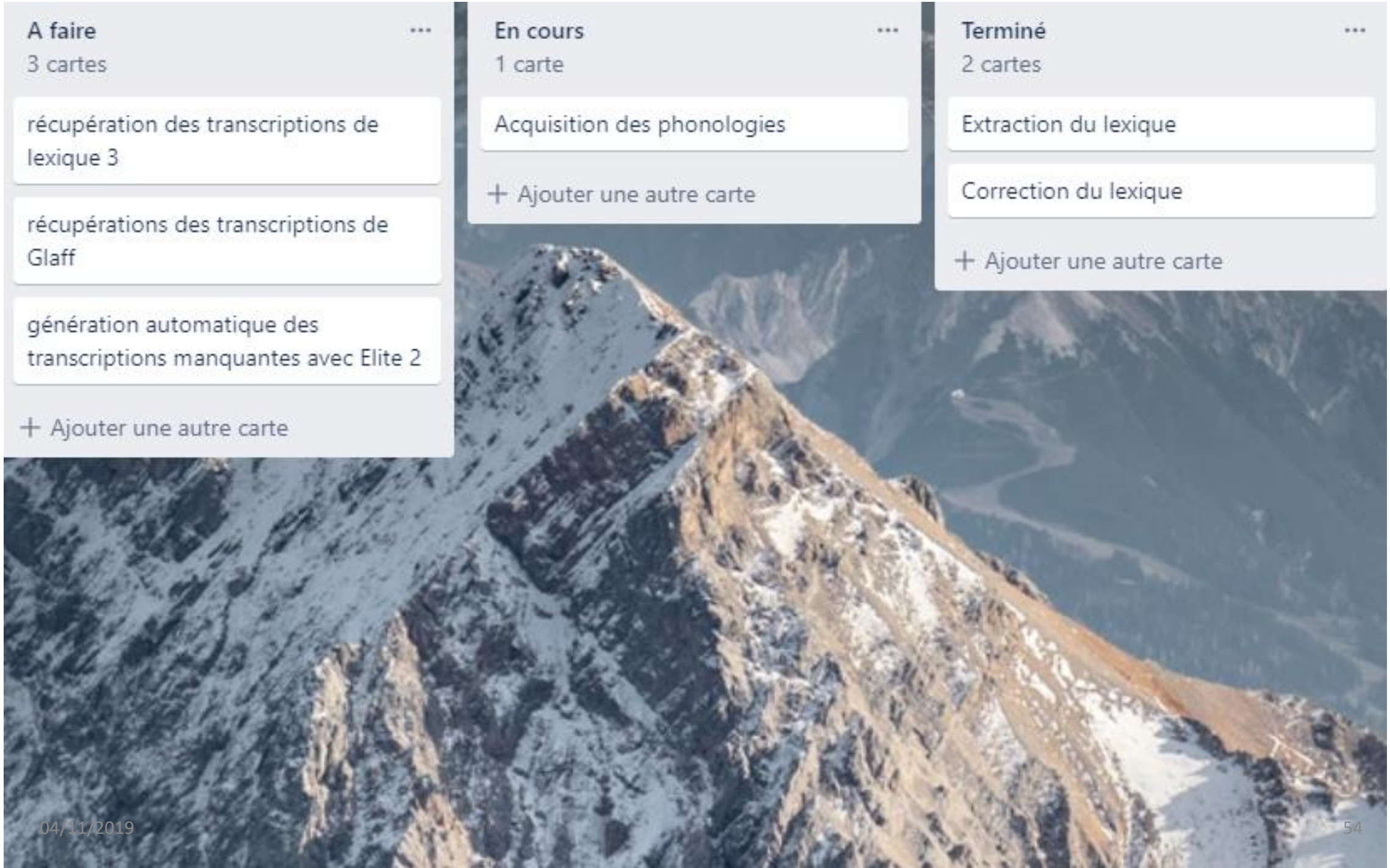
ORTHO	LEMME	CATEGORIE	TYPE	CORRECTIONS	
avouer	luire	VER	LEMME	avouer	
à cloche-pied	à cloche-pied	ADV	LEMCGRAM	à cloche-pied	
déchirés	déchiré	VER	LEMME	déchirer	
dispose	dispos	NOM	TRANSFOVER	disposer	ind:pre:1s;ind:pre:3s;
passer	passer	VER	TRANSFOVER	passer	inf;
détendre	détendre	VER	TRANSFOVER	détendre	inf;
désincarné	désincarné	VER	LEMME	désincarner	
assister	assister	VER	TRANSFOVER	assister	inf;
lamer	mettre	VER	LEMME	lamer	
stérilisés	stérilisé	VER	LEMME	stériliser	
distraire	distraire	VER	TRANSFOVER	distraire	inf;

	ERREURS	PROBLEMES
ANCIEN	2167	4
NOUVEAU	101	3
TOTAL	2268	7

ANCIEN	NOUVEAU
EFFACER	ORTHO
LEMME	CAT
LEMCAT	TOUT
TRANSFOVER	

Algorithme de détection et de correction automatique des erreurs potentielles

- détection:
  - comparaison avec des lexiques de référence (Lefff, Dicollecte)
  - mesure de l'accord avec les références
  - typologie des désaccords entre les références
- correction:
  - validation des entrées correctes
  - Correction des lemmes si accord
  - Correction des catégories si accord


A faire
3 cartes
récupération des transcriptions de lexique 3
récupérations des transcriptions de Glaff
génération automatique des transcriptions manquantes avec Elite 2
+ Ajouter une autre carte
En cours
1 carte
Acquisition des phonologies
+ Ajouter une autre carte
Terminé
2 cartes
Extraction du lexique
Correction du lexique
+ Ajouter une autre carte

base MOTS
(formes fléchies)

base ORTHO
(formes orthographiques)

base PHONO
(formes phonologiques)

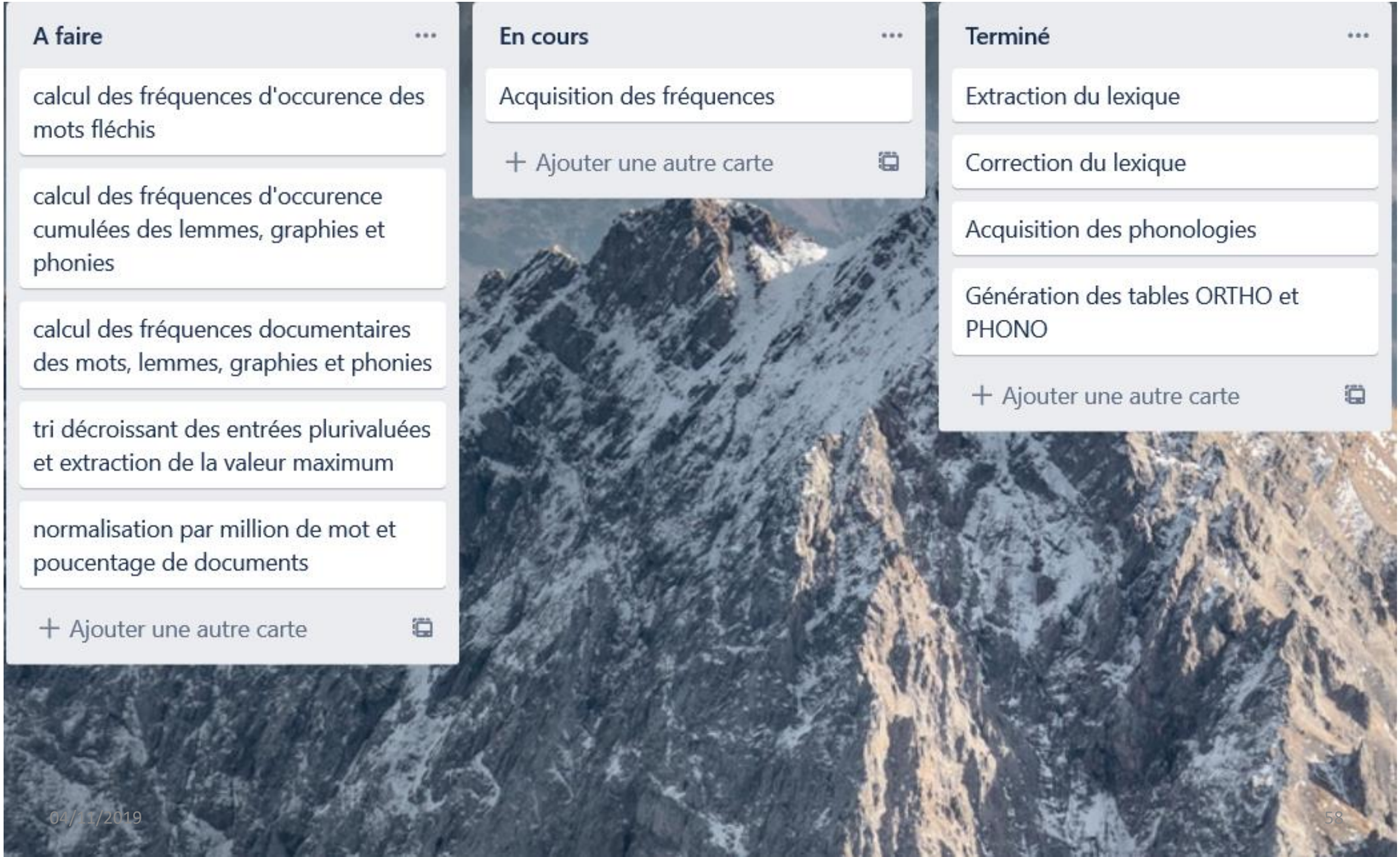
ortho	phon	lemme	cgram	freqfilms2
ai	E	avoir	VER	2475.78
as	a	avoir	VER	1274.31
eu	y	avoir	VER	668.39
avez	ave	avoir	VER	661.88
avait	avE	avoir	VER	596.64
avoir	avwaR	avoir	VER	404.19
ont	§	avoir	VER	381.18

```
lemme          categorie      graphie(s)                      phonie(s)                  frequence(s)
avoir          VER            ai; as;eu;avez;avait;avoir;ont  E;a;y;ave;avE;avwaR;§      2475.78;1274.31;668.39;661.88;596.64;404.19;381.18
```

ortho	phon	lemme	cgram	freqlemfilms2
All	All	All	Al	All
a	a	a	NOM	81.36
a	a	avoir	AUX	18559.22
a	a	avoir	VER	13572.4
a capella	akapEla	a capella	ADV	0.04
a cappella	akapEla	a cappella	ADV	0.04
couvent	kuv	couver	VER	3.59
couvent	kuv@	couvent	NOM	16.83

```
ortho           lemme(s)            catégorie(s)        frequence(s)
a               a;avoir;avoir       NOM;AUX;VER         81.36;18559.22;13572.4
a capella       a capella           ADV                 0.04
a cappella      a cappella          ADV                 0.04
couvent         couver;couvent      VER;NOM             3.59;16.83

-------------------------------------------------------------------------------------------
phono           graphie(s)          lemme(s)            catégorie(s)        frequences
a               a                   a;avoir;avoir       NOM;AUX;VER         81.36;18559.22;13572.4
akapEla         a capella;a capella a capella;a capella ADV;ADV             0.04;0.04
kuv             couvent             couver              VER                 3.59
kuv@            couvent             couvent             NOM                 16.83
```


A faire
calcul des fréquences d'occurence des mots fléchis
calcul des fréquences d'occurence cumulées des lemmes, graphies et phonies
calcul des fréquences documentaires des mots, lemmes, graphies et phonies
tri décroissant des entrées plurivaluées et extraction de la valeur maximum
normalisation par million de mot et poucentage de documents
+ Ajouter une autre carte
En cours
Acquisition des fréquences
+ Ajouter une autre carte
Terminé
Extraction du lexique
Correction du lexique
Acquisition des phonologies
Génération des tables ORTHO et PHONO
+ Ajouter une autre carte

- types de fréquences:
  - fréquences d'occurrences VS fréquence de documents
  - fréquences comptées VS fréquences cumulées
- Types d'unités:
  - mots VS lemmes
  - formes ortho VS formes phono

- fréquence documentaire -> comptage
  - mots, lemmes, ortho, phono
- fréquence d'occurrence -> mots
- fréquences cumulées
  - lemmes, ortho, phono

ortho	lemme	cgram	freqmots	somfreqmots	freqmotmax	freqlemmes	somfreqlemmes	freqlemmax	lemmax	cdortho	cdmots	cdmotmax	cdlemmes	cdlemmax
antidogmatique	antidogmatique	ADJ	0.003	0.003	0.003	0.003	0.003	0.003	antidogmatique	0.001	0.001	0.001	0.001	0.001
zigouille	zigouiller	VER	0.291	0.291	0.291	0.968	0.968	0.968	zigouiller	0.107	0.107	0.107	0.375	0.375
noctambules	noctambule	NOM	0.041	0.041	0.041	0.174	0.174	0.174	noctambule	0.019	0.013	0.013	0.044	0.044
officierais	officier	VER	0.009	0.009	0.009	8.813	8.813	8.813	officier	0.004	0.004	0.004	3.190	3.19
impotents	impotent;impotent	ADJ;NOM	0.012;0.009	0.021	0.012	0.303;0.091	0.394	0.303	impotent	0.010	0.006;0.004	0.006	0.137;0.042	0.137
refoulées	refoulé;refouler;refoulé	ADJ;VER;NOM	0.098;0.041;0.006	0.145	0.098	1.234;2.607;0.196	4.037	2.607	refouler	0.065	0.044;0.019;0.001	0.044	0.527;1.091;0.087	1.091
cochonne	cochon;cochon;cochonner;cochonne	ADJ;NOM;VER;NOM	1.284;0.917;0.066;0.037	2.304	1.284	8.746;21.905;0.120;0.148	30.919	21.905	cochon	0.850	0.508;0.357;0.030;0.009	0.508	3.144;5.925;0.056;0.058	5.925
as	avoir;avoir;as;ai	AUX;VER;NOM;NOM	2093.968;1139.544;13.952;0.006	3247.47	2093.968	19721.841;13002.613;14.085;0.854	32739.393	19721.841	avoir	95.829	93.901;89.508;4.760;0.003	93.901	99.788;99.783;4.808;0.199	99.788

Lexique 4 en scripts

01 – acquisition

- 01 - extraction du corpus.py
- 02 - création de la base de métadonnées.py
- 03 - repérage des problèmes d'OCR.py
- 04 - encodage du corpus en ANSI.py
- 05 - extraction du lexique.py

02 – correction

- 01 - conversion en minuscules.py
- 02 - filtrage des entrées.py
- 03 - récupération des néologismes.py
- 04 - Conversion des catégories.py
- 05 - Separation des infogram.py
- 06 - correction des lemmes des classes fermées.py
- 07 - correction automatique des analyses de Cordial.py
- 08 - correction manuelle des analyses de Cordial.py
- 09 - fusion des doublons graphiques.py

03 – phonologie

- 01 - comparaison des entrées de Lexique4 et Lexique3.py
- 02 - comparaison des entrées de Lexique4_phono1 et Glaff.py
- 03 - comparaison des graphies de Lexique4_phono2 et Glaff.py
- 04 - extraction de la liste des graphies sans phono.py
- 05 - elite2lex.pl
- 06 - appariement des sorties de ELite.py
- 07 - comparaison des graphies de Lexique4_phono3 et Elite.py
- 08 - ajout des champ phono.py

04 – constitution

- A - 01 - génération de la table LEMMES.py
- A - 02 - génération de la table ORTHO.py
- A - 03 - génération de la table PHONO.py
- B - 01 - Extraction du CD Open2018 normalisé.py
- B - 02 - Extraction du CD Open2018 lemmes.py
- B - 03 - Extraction du CD Open2018 ortho.py
- B - 04 - Extraction du CD Open2018 phono.py
- C - 01 - ajout des champs freqLEM, CD et CD_LEM à la base MOT.py
- C - 02 - ajout des CD à la base LEMME.py
- C - 03 - ajout des CD à la base ORTHO.py
- C - 04 - ajout des CD à la base PHONO.py
- D - 01 - normalisation des fréquences et des CD de la table MOTS.py
- D - 02 - normalisation des fréquences et des CD de la table LEMME.py
- D - 03 - normalisation des fréquences et des CD de la table ORTHO.py
- D - 04 - normalisation des fréquences et des CD de la table PHONO.py

- Lexique4 - 01 - acquisition
- Lexique4 - 02 - corrections
- Lexique4 - 03 - phonologie
- Lexique4 - 04 - constitution

dossier type

- Corpus
- batch.ps1
- journal.docx
- lexique.tsv
- readme.md
- script.py

	FICHIERS	SCRIPTS	DOSSIERS	TABLES	ARCHIVES	DOC
LEXIQUE 4	394 008	45	4	4	0	8
ACQUISITION	19	6	6	4	1	2
CORRECTION	32	10	0	20	0	2
PHONOLOGIE	25	9	0	14	0	2
CONSTITUTION	39	20	0	17	0	2

	entrées	champs	valeurs
MOTS :	189 286	12	2 271 432
LEMMES :	68 848	9	619 632
ORTHO :	168 597	15	2 528 955
PHONO :	97 905	17	1 664 385

	formes	lemmes
NOM	67 609	45 692
VER	87 651	7 333
ADJ	31 250	13 132
ADV	1 899	1 890
AUTRES	874	799

Lexique 4 en chiffres

Echantillons de productivité

Lexique4 (316M de mots)

SUF	nbTyp	nbOcc	nbHap	productivité
ique_ADJ	1593	1142,416	302	0,000837
able_ADJ	739	1024,483	140	0,000432
if_ADJ	380	355,13	36	0,000321
ion_NOM	2285	5978,214	305	0,000161
ifier_VER	124	456,146	21	0,000146
eux_ADJ	471	1679,797	58	0,000109
oir(e)_NOM	236	1784,428	35	6,21E-05

Quelques hapax de la poubelle:

- able_ADJ 191
- ion_NOM 2947
- oir(e)_NOM 539

ique 0,0037
able 0,0013
if 0,0012
ion 0,00066
ifier 0,00040
eux 0,00016
oir(e) 0,00012

Lexique383_films
(50M de mots)

Perspectives

- Frantext
  - diachronie (évolution de la productivité, évolution du lexique)
- Le corpus de Lexique4
  - partitions (diachronique, générique)
  - sémantique distributionnelle
  - les contextes de hapax

Références

- Aliquot-Suengas Sophie. La productivité actuelle de la forme constructionnelle -ade . In: Langue française, n°140, 2003. La productivité morphologique en questions et en expérimentations. pp. 38-55;
- **Basilio Calderone, Nabil Hathout et Franck Sajous. (2014).** From GLÀFF to PsychoGLÀFF: a large psycholinguistics-oriented French lexical resource. *Proceedings of the 16th EURALEX Conference*. Bolzano, Italy.
- Content Alain, Mousty Philippe, Radeau Monique. Brulex. Une base de données lexicales informatisée pour le français écrit et parlé. In: L'année psychologique. 1990 vol. 90, n°4. pp. 551-566;
- Georgette Dal, Bernard Fradin, Natalia Grabar, Fiammetta Namer, Stéphanie Lignon, et al.. Quelques préalables au calcul de la productivité des règles constructionnelles et premiers résultats. *Premier Congrès Mondial de Linguistique Française*, Jul 2008, Paris, France. pp.1525-1538, ⟨10.1051/cmlf08184⟩. ⟨hal-00522821⟩
- Fradin Bernard, Hathout Nabil, Meunier Fanny. La suffixation en -et et la question de la productivité. In: Langue française, n°140, 2003. La productivité morphologique en questions et en expérimentations. pp. 56-78;
- New, Boris, Marc Brysbaert, Jean Veronis, and Christophe Pallier. 2007. “The Use of Film Subtitles to Estimate Word Frequencies.” *Applied Psycholinguistics* 28 (4): 661–677.

REMERCIEMENTS

DIRECTION : Boris New
SUPERVISION : Nathalie Gasiglia

INFORMATIONS DIVERSES :
- Gilles Boyé : flexique
- Natalia Grabar : annotation morpho-syntaxique
- Antonio Balvet : annotation morpho-syntaxique
- Anna Kupsc : annotation morpho-syntaxique
- Damy Amiot : filtrage des hapax