

# Entraînement de Talismane pour textes littéraires en français moderne Version 1.1

Sandrine Ollinger  
sandrine.ollinger@atilf.fr

9 juillet 2018



## Table des matières

<b>1</b>	<b>Différences par rapport à la version 1.0</b>	<b>2</b>
<b>2</b>	<b>Commande</b>	<b>2</b>
<b>3</b>	<b>Jeu d'étiquettes</b>	<b>3</b>
<b>4</b>	<b>Segmentation en tokens</b>	<b>4</b>
4.1	Fonctionnement de l'instrument.	4
4.2	Définition des tokens composés	5
<b>5</b>	<b>Lexiques</b>	<b>6</b>
<b>6</b>	<b>Évaluation</b>	<b>7</b>
6.1	Corpus d'évaluation.	7
6.2	Résultats	8
<b>Annexes</b>		<b>12</b>

### Résumé

Ce document présente l'entraînement du module d'étiquetage en parties du discours de Talismane [8] réalisé en 2016 au sein du laboratoire ATILF lors de la campagne de catégorisation de l'ensemble des textes écrits après 1850 de la base Frantext (<https://www.frantext.fr>), dans sa version mise en ligne en avril 2018.

Cet entraînement a été réalisé avec la version 2.6.0b de Talismane.

Il est le fruit de la collaboration de l'ensemble des membres permanents, contractuels et stagiaires suivants : Christophe Benzitoun, Lolita Bérard, Isabelle Clément, Matthieu Delabarre, Ulrike Fleury, Bertrand Gaiffe, Véronique Montémont, Sandrine Ollinger, Etienne Petitjean, Gilles Souvay, Stéphane Tiv et Clémence Urtebize.

Il a également bénéficié de coups de pouce et de conseils avertis de la part de Sacha Diwersy, Achille Falaise, Ingrid Falk, Martin Lent-schat, Jessika Perignon, Franck Sajous, Marie Tonnelier, Assaf Urieli et Marianne Vergez-Couret.

## 1 Différences par rapport à la version 1.0

Cette nouvelle version comporte les modifications suivantes :

- le corpus d'apprentissage a fait l'objet de menues corrections ;
- la segmentation en tokens a été légèrement modifiée (4.2, point 7, page 6).

## 2 Commande

Pour étiqueter en parties du discours un fichier avec notre entraînement, il faut avant tout télécharger Talismane v2.6.0b à l'adresse suivante : <https://github.com/joliciel-informatique/talismane/releases>.

Il est ensuite possible de placer l'archive `frenchFrantextPack-1.0.zip` dans le dossier `talismane-2.6.0b` et de lancer l'annotation d'un texte à l'aide de la commande suivante :

```
java -Xmx2G -jar talismane-core-2.6.0b.jar  
languagePack=frenchFrantextPack-1.0.zip command=analyse  
endModule=postag tokeniserType=simple inFile=test.txt
```

Le modèle de segmentation en phrases utilisé par notre entraînement est celui distribué dans le `frenchLanguagePack-2.5.0`. Les propriétés complètes de l'entraînement sont les suivantes :

```

locale=fr
posTagSet=frantextTagset.txt
tokenFilters=replace :token_filters_fr_simple_final.txt
posTaggerRules=posTagger_fr_rules_final.txt
sentenceModel=sentence_ftb_all_maxent_cut5_v2.zip
posTaggerModel=frantextPosTagger_final.zip
lexicon=lexicons_frantext_ortolang.zip

```

Nous invitons à consulter la documentation de Talismane pour toute information sur les paramètres de l'instrument et les formats d'entrée et de sortie.

### 3 Jeu d'étiquettes

ADJ	adjectif	P+D	préposition + déterminant
ADV	adverbe	PONCT	ponctuation
CC	conjonction de coordination	PRO	pronom
CS	conjonction de subordination	PROREL	pronom relatif
CLO	clitique objet	PROWH	pronom interrogatif
CLS	clitique sujet	P	préposition
DET	déterminant	V	verbe conjugué
ET	mot étranger	VINF	verbe à l'infinitif
I	interjection	VPP	verbe participe passé
NC	nom commun	VPR	verbe participe présent
NP	nom propre	X	mot non traité

TAB. 1 : Jeu d'étiquettes Frantext

Nous avons choisi de nous baser sur le jeu d'étiquettes proposés par Crabbé et Candito [3] et de le modifier à la marge.

Une étiquette **X** a été ajoutée, pour les cas non analysables avec les autres étiquettes.

Sept étiquettes ont été supprimées. Il s'agit d'étiquettes qui étaient associées à peu de formes et d'étiquettes davantage bénéfiques pour l'analyse syntaxique que pour la seule catégorisation ou engendrant des difficultés de catégorisation particulières.

Les étiquettes spécifiques aux clitiques sujets (*je, tu, il, elle, on, ça, nous, vous, ils, elles*) et objets (*le, la, les, lui, leur, y, en*) ont été conservées. Elles rendent compte d'un emploi particulier des pronoms personnels en tant que sujet ou objet de verbes.

## 4 Segmentation en tokens

Nous présentons ici les grands principes de la segmentation en tokens pour laquelle nous avons opté. L'ensemble des règles qui en découlent est fourni avec l'entraînement du module d'étiquetage en parties du discours de Talismane [8], dans le fichier `token_filters_fr_simple_final.txt`.

### 4.1 Fonctionnement de l'instrument

Talismane offre la possibilité de segmenter les phrases en tokens selon deux méthodes distinctes :

- une segmentation simple : à base de règles uniquement, définies à l'aide d'expressions régulières
- une segmentation à base de patrons : par apprentissage supervisé

Dans tous les cas, la segmentation est réalisée en amont de l'étiquetage.

Nous avons décidé d'opter pour la première de ces méthodes. Nous pensions initialement adopter une segmentation maximale et n'effectuer qu'un nombre très limité de regroupement. À l'usage, cette position d'une segmentation maximale s'est avérée insatisfaisante et les regroupements, que nous détaillons ci-après, sont en définitive nombreux<sup>1</sup>.

La segmentation est réalisée à partir de deux ressources :

- un ensemble prédéfinis de caractères délimiteurs<sup>2</sup> :

```
!"#$%&'()*+,-./:;<=>?@[ \ ] ^ _ ` { | } ~ « » - - - - ' " " "
```

as well as the various whitespace characters: the space bar, the tab character, the newline, etc.

---

<sup>1</sup>Ce qui a pour conséquence un temps d'étiquetage relativement lent.

<sup>2</sup>Remarque : les caractères délimiteurs sont généralement isolés lors de la segmentation, mais ce n'est pas le cas de l'apostrophe, que nous avons choisi de rattacher au mot qui précède : *Quand ?* est composé de deux segments, *Qu'* est composé d'un seul segment.

- une liste ordonnée d'expressions régulières qui définit tout ce qui ne doit pas être segmenté.

Cette méthode ne permet aucune gestion de l'ambiguïté. Une suite de caractères est définie comme étant, ou non, un token, indépendamment du contexte.

Il existe donc des regroupements incorrects, qui induisent des erreurs d'étiquetage. Ainsi, dans les exemples suivant, *rendez-vous* est systématiquement regroupé et étiqueté nom commun :

- Vous arrivez en retard à tous vos rendez-vous ? (E115)
- Faites comme lui, rendez-vous sourde à tous les cris, rejoignez la pierre pendant qu'il en est temps. (K244)

## 4.2 Définition des tokens composés

À l'issue des travaux réalisés par divers membres de l'équipe, les unités composées ont été définies comme l'ensemble des formes appartenant aux sous-ensembles suivant :

1. Les formes de lemmes composés comportant une espace ou une apostrophe dans Morphalou 3 [2], après double validation manuelle selon 4 critères :
  - Si l'un des mots graphiques du composé n'existe pas seul (*jeun, afin, fur*) => une seule unité
  - Si le composé est ambigu et que les mots graphiques peuvent se suivre par ailleurs (*bien que*) => plusieurs unités
  - Si le moule syntaxique ne pose pas problème (*à l'aune de*) => plusieurs unités
  - Si le composé est une expression latine (*post scriptum*) => une seule unité.
2. Les formes composées comportant un trait d'union dans Morphalou 3<sup>3</sup>.

---

<sup>3</sup>Morphalou 3 comporte 10 111 lemmes de ce type. Leur vérification manuelle n'a donc pas été réalisée, à la fois pour des questions de temps et de difficulté à maintenir un jugement uniforme sur un nombre aussi important de cas.

3. Les formes composées à l'aide de traits d'union commençant par un préfixe répertorié dans GLAWI [5], à l'exception de *très-* et *mon-*, qui ont été identifiés comme étant source de trop d'erreurs de regroupement.
4. Les formes composées à l'aide de traits d'union commençant par un préfixe selon une liste additionnelle obtenue à partir des formes composées de Morphalou 3 et disponible page 14.
5. Les nombres composés (*trois cent*, *quatre-vingt-dixième*).
6. Les formes débutant par un trait d'union appartenant à la liste de mots grammaticaux disponible page 15.
7. Les noms propres composés répertoriés dans nos lexiques, à l'exception de 239 noms de communes homographes de séquences **DET NC**, telles que *Le Passage*.

## 5 Lexiques

Nous avons choisi d'utiliser principalement le lexique Morphalou 3 [2], dont nous avons exclu la partie « noCategory », après avoir vérifié que son contenu ne nous était pas utile. De plus, certaines entrées ont été exclues des catégories grammaticales restantes (nom commun, adjectif, adverbe, verbe, interjection, mots grammaticaux), parce qu'elles étaient source d'erreurs.

Six lexiques viennent compléter cette ressource :

- Un lexique de classes fermées, nécessaire au fonctionnement de Talismane
- Un lexique de 80 triplets forme-lemme-catégorie constitué pour compléter Morphalou 3 (exemple : *entr'aiderions entr'aider V*)
- Un lexique de Noms Propres qui contient des anthroponymes, des toponymes et des noms de fêtes, constitué à partir des ressources suivantes<sup>4</sup> :
  - Dictionnaire de noms propres Prolex-Unitex 1.2 ([http://tln.li.univ-tours.fr/Tln\\_Unitex.html](http://tln.li.univ-tours.fr/Tln_Unitex.html)) [7]
  - Base Prénoms de l'Université de Savoie (<http://www.lexique.org/public/prenoms.php>)

---

<sup>4</sup>Toutes ces ressources ont été consultées pour la dernière fois le 6 avril 2016

- Découpage administratif communal français issu d’OpenStreet-Map (<http://www.data.gouv.fr/fr/datasets>)
- Départements et régions française de Wikipédia (<https://fr.wikipedia.org>)
- Trois lexiques de mots étrangers :
  - Les 500 mots de la langue anglaise les plus fréquents d’après un travail de l’Université de Leipzig [https://fr.wiktionary.org/wiki/Wiktionnaire:Listes\\_de\\_fr%C3%A9quence#Anglais](https://fr.wiktionary.org/wiki/Wiktionnaire:Listes_de_fr%C3%A9quence#Anglais)
  - Les 500 mots de la langue allemande les plus fréquents d’après un travail de l’Université de Leipzig [https://fr.wiktionary.org/wiki/Wiktionnaire:Listes\\_de\\_fr%C3%A9quence#Anglais](https://fr.wiktionary.org/wiki/Wiktionnaire:Listes_de_fr%C3%A9quence#Anglais)
  - 237 mots latins tirés de [https://fr.wikibooks.org/wiki/Latin/Vocabulaire/Les\\_400\\_mots\\_les\\_plus\\_fr%C3%A9quents](https://fr.wikibooks.org/wiki/Latin/Vocabulaire/Les_400_mots_les_plus_fr%C3%A9quents) et pages liées

## 6 Évaluation

Notre entraînement de Talismane (`frenchFrantextPack-1.0`) a été réalisé à l’aide d’un corpus d’apprentissage composé de cent extraits de textes d’environ 2 000 tokens chacun, issus de la base textuelle Frantext. Ce corpus couvre la période de 1850 à 2015. Il a été échantillonné afin de représenter de façon équilibrée les différentes époques et différents genres en œuvre dans la ressource à catégoriser.

Nous proposons ici une évaluation de ses performances sur un corpus annexe, que nous empruntons à l’équipe de lexicographie en synchronie de l’ATILF<sup>5</sup>.

### 6.1 Corpus d’évaluation

La présente évaluation a été réalisée sur un corpus de 100 phrases extraites de la base d’exemples lexicographiques du Réseau Lexicale du Français [6].

Toutes ces phrases ont été sélectionnées dans le sous-ensemble d’exemples tirés de textes de la base Frantext écrits après 1950. Il a été vérifié qu’aucune

---

<sup>5</sup>L’évaluation de la version 1.1 retourne des résultats identiques à ceux de la version 1.0.

d’entre elles n’appartenait au corpus d’apprentissage utilisé pour mettre au point l’entraînement évalué.

Ces phrases ont été segmentées en tokens selon les choix opérés lors de la campagne de catégorisation de la base Frantext. Elles ont ensuite été étiquetées automatiquement à l’aide de Talismane 2.6.0b et MElt-2.0b12 [4], une fois à l’aide des entraînements pour le français fournis avec ces instruments<sup>6</sup>, une fois avec les entraînements pour la littérature en français moderne que nous avons réalisé.

Une version étalon du corpus d’évaluation a alors été réalisée en comparant les quatre étiquetages obtenus, puis en effectuant une relecture complète<sup>7</sup>.

## 6.2 Résultats

Une fois notre corpus d’évaluation annoté et corrigé, nous l’avons soumis à notre entraînement de Talismane en utilisant la fonction d’auto-évaluation proposée par l’instrument. Les résultats présentés ici sont directement issus de cette auto-évaluation. Ils sont conformes à nos observations manuelles.

Unités testées	2 806
Étiquettes correctes	2 768
Étiquettes incorrectes	38
Précision globale	98,65%
Temps d’exécution	1 min 20 s

TAB. 2 : Évaluation de `frenchFrantextPack-1.0`

Vous trouverez page 10 la matrice de confusion et les mesures de précision, de rappel et de f-scores pour chaque partie du discours, également fournies par l’instrument.

En comparaison, nous avons relevé 106 étiquettes incorrectes produites par l’entraînement pour le français `frenchLanguagePack-2.5.0` fourni avec Talismane lors de la préparation manuelle de la version étalon de notre corpus d’évaluation. Ce qui correspond à une précision de 96,22%. Ces mesures

<sup>6</sup>Ces entraînements ont tous deux été réalisés avec le corpus journalistique French Tree Bank [1].

<sup>7</sup>Merci à Christophe Benzitoun pour sa relecture attentive de la version étalon de notre corpus d’évaluation.



confirment le gain d'un entraînement spécifique pour les textes littéraires en français moderne.

outcome	ADJ	ADV	CC	CLO	CLS	CLS	CS	DET	I	NC	NP	P	P+D	PONCT	PRO	PROREL	PROWH	V	VINF	VPP	VPR	
ADJ	168	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ADV	1	130	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	4	0
CC	0	0	80	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
CLO	0	0	0	90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CLS	0	0	0	0	134	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CS	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DET	0	0	0	0	0	0	346	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NC	3	1	0	0	0	0	1	0	517	5	0	0	0	0	0	0	0	0	0	0	0	0
NP	0	0	0	0	0	0	0	0	5	49	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	1	2	0	0	339	0	0	0	0	0	0	0	0	0	0	0
P+D	0	0	0	0	0	0	0	3	0	0	0	63	0	0	0	0	0	0	0	0	0	0
PONCT	0	0	0	0	0	0	0	0	0	0	0	0	0	375	0	0	0	0	0	0	0	0
PRO	1	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0
PROREL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0
PROWH	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	225	0	0	0	0
VINF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	83	0	0	0
VPP	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	57	0	0
VPR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7

outcome	true+	false+	false-	count	precision	recall	f-score
ADJ	168	5	5	173	97,11	97,11	97,11
ADV	130	1	2	132	99,24	98,48	98,86
CC	80	0	1	81	100	98,77	99,38
CLO	90	0	0	90	100	100	100
CLS	134	0	0	134	100	100	100
CS	37	2	3	40	94,87	92,5	93,67
DET	346	6	4	350	98,3	98,86	98,58
I	4	0	0	4	100	100	100
NC	517	5	10	527	99,04	98,1	98,57
NP	49	7	5	54	87,5	90,74	89,09
P	339	2	3	342	99,41	99,12	99,27
P+D	63	2	3	66	96,92	95,45	96,18
PONCT	375	0	0	375	100	100	100
PRO	25	1	0	25	96,15	100	98,04
PROREL	37	3	1	38	92,5	97,37	94,87
PROWH	2	0	0	2	100	100	100
V	225	0	0	225	100	100	100
VINF	83	0	0	83	100	100	100
VPP	57	4	1	58	93,44	98,28	95,8
VPR	7	0	0	7	100	100	100
TOTAL	2768	38	38	2806	98,65	98,65	98,65
AVERAGE					97,72	98,24	97,97
ACCURACY					98,65		
KAPPA					98,49		

## Références

- [1] Anne Abeillé, Lionel Clément, and François Toussenet. *Building a Treebank for French*, pages 165–187. Springer Netherlands, Dordrecht, 2003.
- [2] ATILF. Morphalou, 2015. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.
- [3] Benoît Crabbé and Marie Candito. Expériences d’analyse syntaxique statistique du français. In *5ème conférence sur le Traitement Automatique des Langues Naturelles - TALN’08*, pages pp. 44–54, Avignon, France, June 2008.
- [4] Pascal Denis and Benoît Sagot. Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Lang. Resour. Eval.*, 46(4) :721–736, December 2012.
- [5] Nabil Hathout and Franck Sajous. Wiktionnaire’s Wikicode GLAWified : a Workable French Machine-Readable Dictionary. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, may 2016. European Language Resources Association (ELRA).
- [6] Veronika Lux-Pogodalla. Intégration relationnelle des exemples lexicographiques dans un réseau lexical. In *TALN (Traitement Automatique du Langage Naturel)*, Marseille, France, July 2014.
- [7] Mickaël Tran and Denis Maurel. Prolexbase : Un dictionnaire relationnel multilingue de noms propres. *Traitement Automatique des Langues*, 47(3) :115–139, 2006.
- [8] Assaf Urieli. *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, Université de Toulouse II le Mirail, 2013.

# Annexes

## Lemmes de Morphalou 3 comportant une espace conservés

a b c	ad vitam aeternam	de plano
à brûle-pourpoint	afin de	de profundis
à califourchon	afin que	de visu
a contrario	alter ego	delirium tremens
à donf	au demeurant	deo gratias
a fortiori	au fur et à mesure	deus ex machina
à jeun	au visé	dies irae
a latere	auto- destructif	don juan
a minima	b.a. ba	don quichotte
à partir de	bel canto	don quichottesque
a posteriori	bernard l'ermite	don quichottisme
a priori	bernard l'hermite	don- juanisme
à priori	bernard lermite	ecce homo
à qui mieux mieux	bernard lhermite	eh ben
à quia	bric-à-brac	eh bien
à tâtons	buen retiro	eh quoi
a tempera	c'est à dire	en catimini
à tue-tête	cardio-péricardo- myopexie de gor	en delà de
à vau-l'eau	carpe diem	en outre
à vrai dire	casus belli	en stand-by
ab absurdo	compact disk	et caetera
ab origine	compte rendu	ex abrupto
ad hoc	cosa nostra	ex aequo
ad hominem	d'ores et déjà	ex cathedra
ad libitum	dare dare	ex nihilo
ad limina	de auditu	ex professo
ad litem	de bric et de broc	ex voto
ad litteram	de commodo	exempli gratia
ad nauseam	de cuius	extra muros
ad nutum	de facto	garde champêtre
ad patres	de guingois	gratis pro deo
ad personam	de gustibus	grosso modo
ad rem	de jure	habeas corpus
ad valorem		hasta luego

hic et nunc	nec plus ultra	punching ball
honoris causa	nihil obstat	quant à
id est	no man's land	quat'z arts
in abstracto	noli me tangere	quod erat demonstran-
in extenso	nota bene	dum
in extremis	numerus clausus	res militaris
in fine	oh la la	rock and roll
in pace	oh là là	rock and roller
in petto	one woman show	sainte nitouche
in praesenti	one-man show	salvé regina
in praesentia	op. cit.	se frotti-frotter
in sillico	opéra bouffe	sedia gestatoria
in situ	operating system	semper virens
in spiritu	opus citatum	sine die
in utero	osso buco	sine qua non
in vitro	over arm stroke	souventes fois
in vivo	p et t.	stabat mater
intra muros	p et t	statu quo
ipso facto	p. et t	stricto sensu
joint venture	parce que	sui generis
kyrie eleison	patati et patata	tandis que
lato sensu	patati patata	te deum
lingua franca	pax americana	terminus a quo
loc. cit.	pax romana	terminus ad quem
locus citatus	per capita	terra incognita
magister dixit	persona grata	terra rossa
manu militari	persona non grata	terza rima
mass media	peu ou prou	tout de go
mea culpa	pique niquer	ultima ratio
mezza voce	plan plan	verbi gratia
minus habens	pole- position	vice versa
missi dominici	post mortem	volens nolens
modern style	post scriptum	vomito negro
modus operandi	pourvu que	vox populi
modus vivendi	pretium doloris	vulgum pecus
motu proprio	prima donna	white spirit
mutatis mutandis	pro domo	y compris
ne varietur	pro forma	

## Lemmes de Morphalou 3 comportant une apostrophe conservés

aujourd'hui	fedda'i	parc'que
baha'i	hors-d'oeuvre	pin's
baha'isme	hors-d'oeuvre	presqu'île
bernard l'hermite	j'm'en-fichiste	presqu'île
bernard-l'ermite	j'm'en-foutisme	presqu'îlette
bernard-l'hermite	je-m'en-fichisme	prud'homal
bin's	je-m'en-fichiste	prud'homie
c'est à dire	je-m'en-foutisme	prud'hommal
c'est-à-dire	je-m'en-foutiste	prud'homme
c't'	jourd'hui	qu'en-dira-t-on
ch'timi	jourd'huis	quat'z arts
chef-d'oeuvre	jusqu'au-boutisme	quat'zarts
chef-d'oeuvre	jusqu'au-boutiste	quelqu'un
chefs-d'oeuvre	l'on	quô'c-ngũ'
ct'	m'amie	R'n'B
d'arrache-pied	m'as-tu-vu	rock'n'roll
d'emblée	m'as-tu-vue	rock'n'roller
d'ores et déjà	m'as-tuvuisme	sot-l'y-laisse
dos-d'âne	main-d'oeuvre	tape-à-l'oeil
entr'aide	main-d'oeuvre	tape-à-l'œil
entr'aimer	mains-d'oeuvre	tout-à-l'égout
entr'apercevoir	mam'selle	traveler's
entr'axe	mam'zelle	traveleur's
entr'égorgement	monte-en-l'air	traveller's
entr'égorger	nid-d'abeilles	trompe-l'oeil
entr'ouvert	nids-d'abeilles	trompe-l'œil
entr'ouvrir	no man's land	vau-l'eau
fedda'i	pac'que	

## Liste additionnelle de préfixes

abat-	adipo-	aides-
abdomino-	adiposo-	aigre-
aborto-	adréno-	aigres-
accroche-	adynamico-	albumino-
acido-	affectivo-	alcalino-

aluminico-	bronchio-	grand-
alvéolo-	bronchiolo-	grande-
ammoniaco-	broncho-	grandes-
ammonio-	broncholo-	grands-
analytico-	cap-	grosso-
anarcho-	casse-	ilié-
anhydro-	cellulo-	ilio-
animo-	centre-	maxillo-
ano-	centro-	nord-
antéro-	cérébello-	occipito-
apico-	cérébro-	ouest-
argilo-	cervici-	physico-
arrière-	cervico-	sud-
artério-	cholédoco-	thyréo-
assyro-	coccy-	tibio-
auriculo-	cortico-	touristico-
avant-	costo-	toxi-
bacill-	cubito-	trachélo-
bacillo-	égypto-	trachéo-
basi-	fisco-	urétéro-
basio-	fluvio-	vésico-
bien-	franc-	vésiculo-
bilio-	fronto-	
bissau-	glucido-	

## Liste de mots grammaticaux

-ce	-leur	-t-il
-ci	-lui	-t-ils
-elle	-m'	-t-on
-elles	-m'en	-t-y
-en	-même	-t'
-il	-mêmes	-t'en
-ils	-moi	-toi
-je	-nous	-tu
-la	-on	-vous
-là	-t-elle	-y
-le	-t-elles	
-les	-t-en	