Présentation d'AnaText

Document réalisé par Agnès Tutin et Olivier Kraif

Anatext est un logiciel de traitement de corpus en ligne développé par Olivier Kraif à l'Université Grenoble Alpes qui offre des fonctionnalités d'exploration de corpus intéressantes et très simples d'utilisation, en particulier en ce qui concerne la lexicométrie, c'est-à-dire les statistiques lexicales. L'outil permet par exemple d'extraire les fréquences des mots du texte, les mots les plus spécifiques ou les suites de mots récurrentes (segments répétés). Contrairement à Antconc, le logiciel effectue un étiquetage morpho-syntaxique du texte, c'est-à-dire qu'il indique la catégorie du mot et son lemme (Cf plus bas).
Plusieurs langues sont disponibles : le français, l'anglais, l'allemand, l'espagnol, l'italien, le russe, le latin, etc.
Le logiciel est librement accessible en ligne sur : http://turing3.univ-grenoble-alpes.fr/anaText/

1. Chargement du fichier

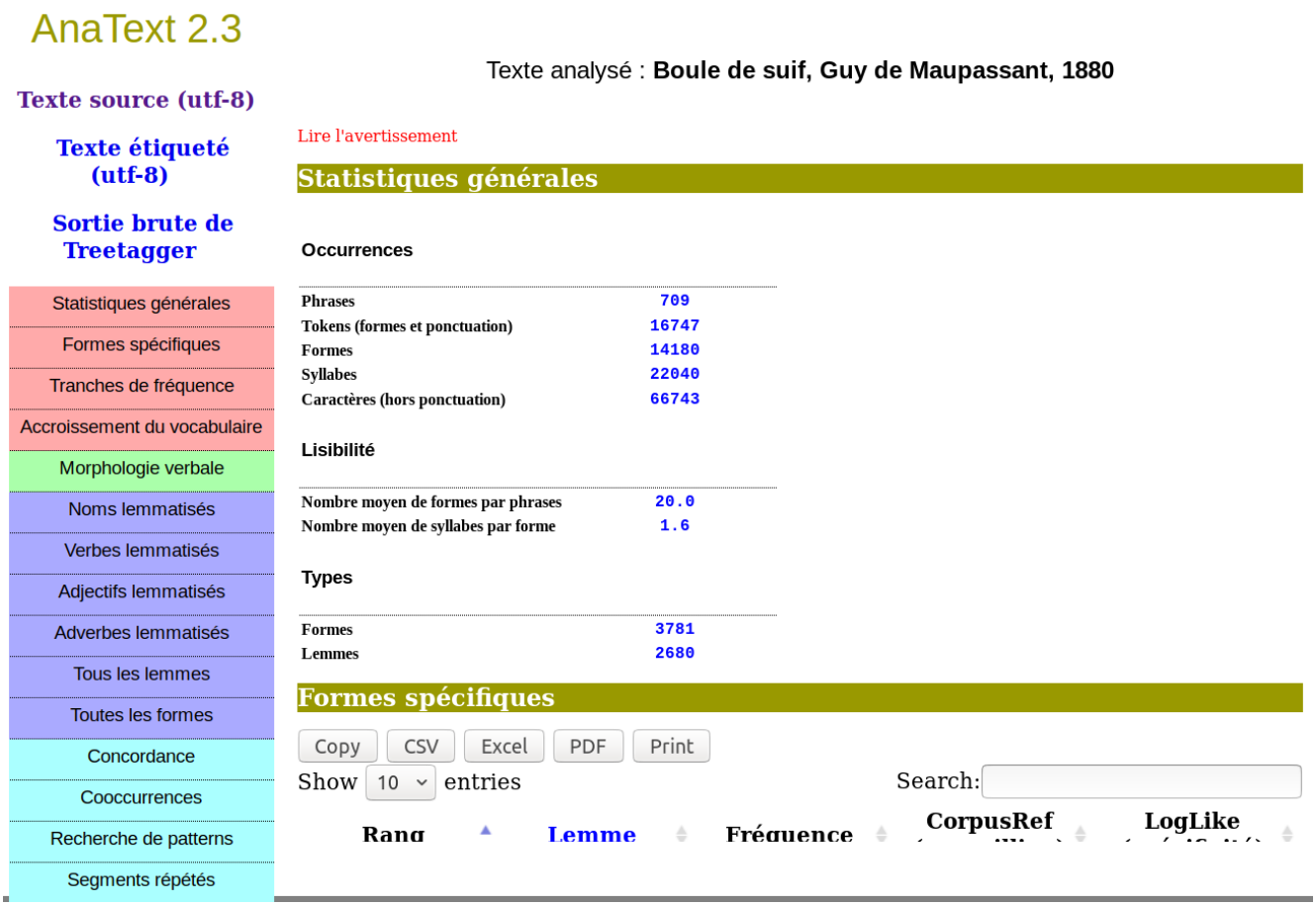
Pour charger le fichier à traiter, il suffit de le copier-coller sur l'interface, comme sur la copie d'écran suivante :

Pour analyser le texte, il suffit de cliquer sur « Analyser le texte ».2 Résultats de l'analyse

2. Etiquetage et lemmatisation

L'outil Anatext effectue un étiquetage morphosyntaxique des textes, c'est-à-dire qu'à chaque mot sont associés a) la catégorie grammaticale la plus probable ; 2) le lemme (la forme normalisée du dictionnaire) du mot, par exemple la forme de l'infinitif pour les verbes. Le logiciel d'étiquetage utilisé est Treetagger.
La figure ci-dessous présente la page qui indique les résultats de l'analyse. Commençons par éclaircir quelques termes :
- lemme : formes canoniques des mots (ce qu'on a dans les dictionnaires). P.ex. pour "examinons ces exemples", on a les lemmes : examiner + ce + exemple
- tokens : le résultat du découpage des phrases en unités graphiques : mots, nombres, ponctuations, symboles, etc.
- formes : les mots tels qu'ils apparaissent dans le texte (on dit aussi "formes fléchies", car ils portent des marques de nombre, genre, conjugaison, etc.)
Quand on compte les mots, on compte deux choses différentes :

- les **occurrences** : le nombre de mots, p.ex. dans "ces mots et ces phrases", on compte 5 occurrences. Le nombre d'occurrences est donc en rapport avec la longueur des textes.
- les **types** : le nombres de mots différents, p.ex. dans "ces mots et ces phrases", on compte 4 types (ces, mots, et, phrases). Le nombre de types indique donc la taille du vocabulaire.



Les calculs sont effectués sur le texte étiqueté (on peut voir l'étiquetage en cliquant sur Texte étiqueté. Par exemple, l'étiquetage pour la phrase : *Pourtant, les heureuses surprises ne sont jamais loin dans cette ville inclassable* apparaît ci-dessous. La première colonne correspond à la forme fléchie du texte, la deuxième colonne est la catégorie, éventuellement suivie d'un trait flexionnel ou une sous-catégorie, la troisième colonne correspond au lemme. Ainsi, la forme *heureuses* est lemmatisée en *heureux* ce qui permettra des calculs de fréquences ou des recherches plus précis (les verbes du français ont beaucoup de formes fléchies différentes).

Forme fléchie	Catégorie	Lemme
Pourtant	ADV	pourtant
,	PUN	,
les	DET:ART	le
heureuses	ADJ	heureux
surprises	NOM	surprise
ne	ADV	ne
sont	VER:pres	être
jamais	ADV	jamais
loin	ADV	loin
dans	PRP	dans
cette	PRO:DEM	ce
ville	NOM	ville
inclassable	ADJ	inclassable

Toutes les fonctionnalités sont observables sur la même page.
Le menu à gauche, avec différentes couleurs, permet d'accéder aux différentes fonctionnalités.

Le tableau ci-dessous indique les principales étiquettes du modèle de treetagger pour le français :

Etiquette	Description
NOM	Nom
NAM	Nom propre
VER	Verbe
ADJ	Adjectif
ADV	Adverbe
DET	Déterminant
DET:ART	Article
DET:POS	Déterminant possessif
PRP	Préposition
PRP:det	Article défini contracté
PRO	Pronom
PRO:PER	Pronom personnel
PRO:REL	Pronom relatif
PRO:DEM	Pronom démonstratif
KON	Conjonction
PUN	Ponctuation
PUN:cit	Guillemet
SENT	Fin de phrase
INT	Interjection
NUM	Numéral

Table 1: Etiquettes du modèle de Treetagger pour le français

Pour la morphologie verbale, en français, Treetagger utilise les étiquettes suivantes :

Etiquette	Description
cond	conditionnel
futu	futur
impf	imparfait
infi	infinitif
ppre	participe présent
pper	participe passé
pres	présent de l'indicatif
simp	passé simple
subi	subjonctif imparfait
subp	subjonctif présent

Table 2: Etiquettes du modèle de Treetagger pour le français - morphologie verbale

Afin d'unifier le système de requête pour toutes les langues, nous proposons un jeu d'étiquettes simplifiées (utilisables dans les fonctions concordances, recherche de patterns, etc.).

Etiquette	Description
NOM	Nom
NPR	Nom propre
VER	Verbe
ADJ	Adjectif
ADV	Adverbe
DET	Déterminant
PRE	Préposition
PRO	Pronom
CON	Conjonction
PON	Ponctuation
PHR	Fin de phrase
INT	Interjection
NUM	Numéral
TOK	Token quelconque
NOSENT	Token quelconque sauf la marque de fin de phrase

Table 3: Jeu d'étiquettes simplifiées (toutes les langues)

3. Statistiques (partie rose du menu)

3.1. Statistiques générales

Plusieurs types de statistiques générales sur le texte sont fournis : le nombre de phrases, de « tokens » (c'est-à-dire des mots fléchis, chiffres et signes de ponctuation), formes fléchies, syllabes et caractères.
Un indice de lisibilité apparaît avec le nombre de mots par phrase, le nombre de syllabes par mot. Enfin, des fréquences sur le nombre de mots et formes différentes apparaissent. Elles peuvent indiquer la richesse lexicale des textes (ce qu'on appelle le ration type/token).

3.2. Spécificités (Formes spécifiques)

Les spécificités sont les mots spécifiques de ce texte calculés par rapport à un corpus de référence, et s'appuie sur les fréquences données sur le site Lexique.org. Deux corpus sont utilisés pour ces fréquences[1] :

- **Frantext** (extrait) : 218 textes littéraires (romans) publiés entre 1950 et 2000 - 14,7 millions d'items
- **Film** : 9474 films ou saisons de séries représentant en tout 50 millions de mots.

Dans Anatext, les spécificités sont calculées sur les lemmes, ce qui apparaît souvent plus intéressant que sur les formes. Sur la copie d'écran suivante, les lemmes les plus spécifiques de Boule de suif sont (par ordre décroissant) : *suif*, *comte*, *prussien*, etc.

1 cf. http://lexique.org/_documentation/Manuel_Lexique.3.2.pdf, p. 6

Formes spécifiques

Copy CSV Excel PDF Print

Show 10 entries Search:

Rang	Lemme	Fréquence	CorpusRef (par million)	LogLike (spécificité)
1	suif	36	1.165	490.426
2	comte	52	45.66	337.414
3	prussien	22	3.855	214.720
4	manufacturier	7	0.035	131.433
5	boule	23	45.94	112.029
6	officier	25	83.905	96.883
7	atteler	8	4.86	57.775
8	neige	14	54.09	50.533
9	voyageur	11	28.6	47.973
10	madame	23	209.135	47.630

Showing 1 to 10 of 101 entries First Previous Next Last

3.3. Tranches de fréquences

Cette fonctionnalité compare les fréquences des formes du texte avec avec celles du corpus de référence. Le vocabulaire du lexique de référence est réparti en tranches :

Tranche 1 : 1000 premiers lemmes les plus fréquents
Tranche 2 : 2000 lemmes suivants
Tranche 3 : 4000 lemmes suivants
Tranche 4 : 8000 lemmes suivants
Tranche 5 : 16000 lemmes suivants
Tranche 6 : tous les lemmes qui restent

On considère 6 tranches, outre les mots inconnus dans la référence. Cette répartition donne une idée de la structure du lexique : très spécialisé (forte représentation des tranches 3-6) ou non (forte représentation des tranches 1-2).

4. Morphologie verbale (partie verte du menu)

Elle indique les temps verbaux rencontrés dans le texte. Par exemple, dans la copie d'écran suivante, c'est l'imparfait et le passé simple qui dominent, mais d'autres temps du passé sont aussi présents (participes passés). Le futur apparaît peu.

Morphologie verbale

Copy CSV Excel PDF Print

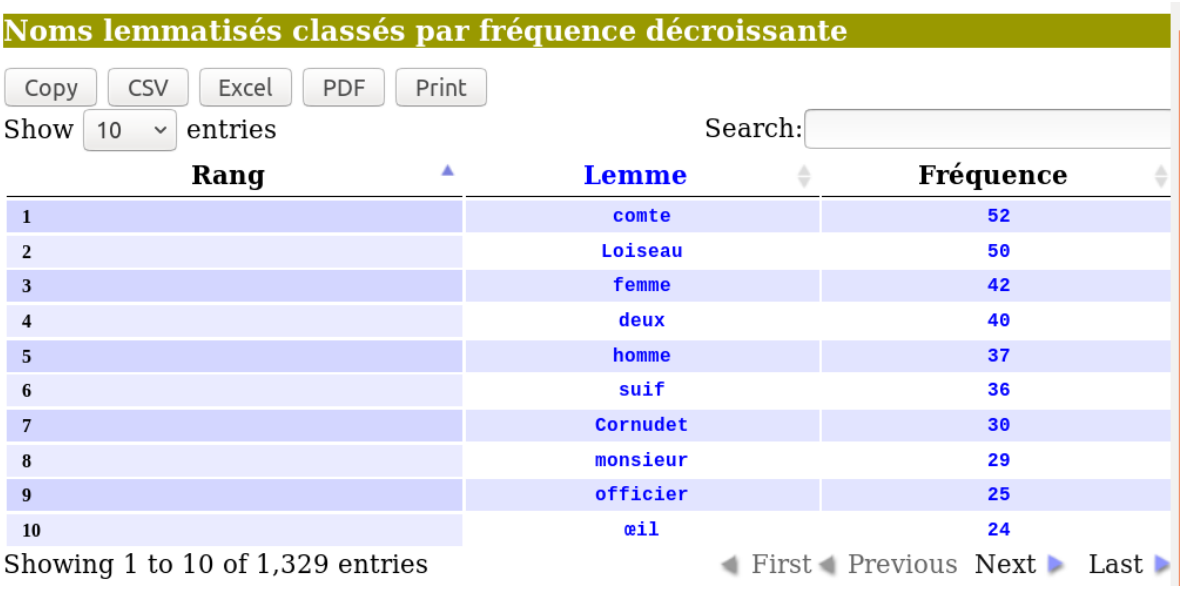
Search:

Temps / mode	Nombre d'occurrences
cond	50
futu	16
impf	661
infi	376
pper	402
ppre	149
pres	312
simp	490
subi	20
subp	14

Showing 1 to 10 of 10 entries

5. Fréquences des lemmes et des formes (partie bleue du menu)

Grâce à l'étiquetage morphosyntaxique, il est possible de calculer non seulement les formes, mais aussi les lemmes du texte en prenant en compte les catégories (par exemple, les noms ou les verbes). Les noms les plus fréquents (par ordre décroissant) sont : *comte, Loiseau, femme, homme …* On observe que tous ne sont pas nécessairement les noms les plus spécifiques du texte (par rapport à un corpus de référence, voir 3.1.2).



Noms lemmatisés classés par fréquence décroissante

Copy CSV Excel PDF Print

Show 10 entries Search:

Rang	Lemme	Fréquence
1	comte	52
2	Loiseau	50
3	femme	42
4	deux	40
5	homme	37
6	suif	36
7	Cornudet	30
8	monsieur	29
9	officier	25
10	œil	24

Showing 1 to 10 of 1,329 entries First Previous Next Last

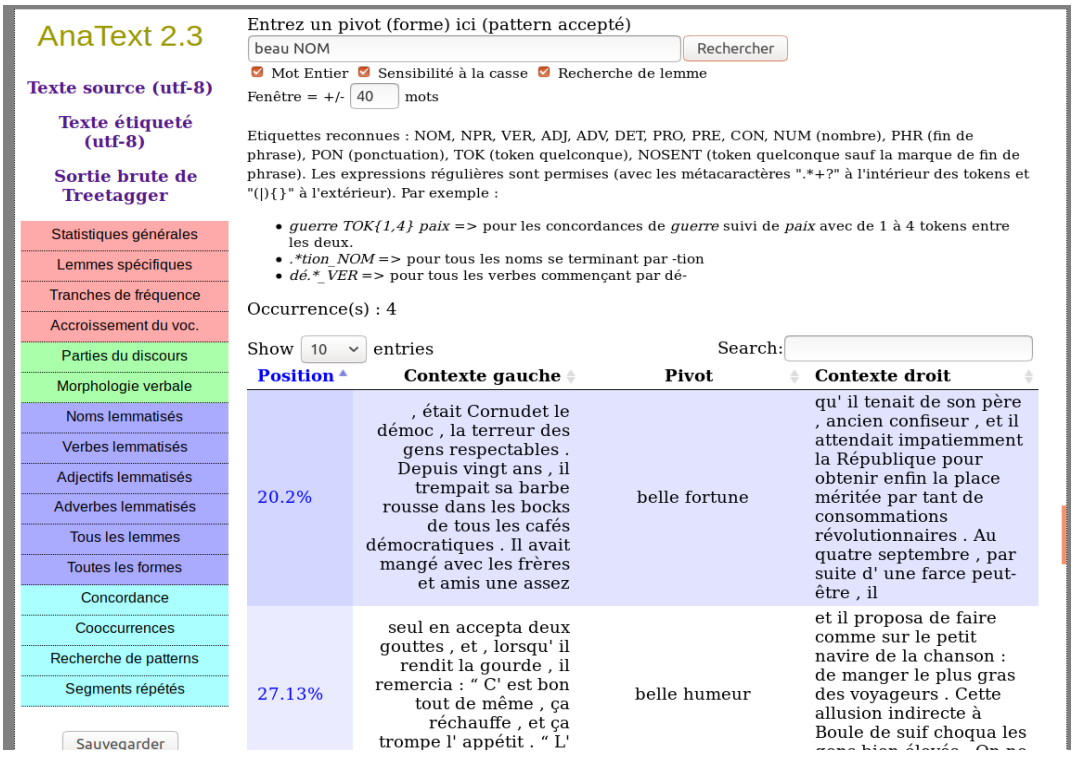
6. Recherche dans les textes (partie turquoise du menu)

Différentes fonctionnalités permettent d'explorer le texte.

6.1. Les concordances

Les concordances peuvent être effectuées sur des formes (comme sur Antconc) mais aussi sur des lemmes ou des catégories (grâce à l'étiquetage morpho-syntaxique). On peut faire des recherches sur des lemmes, des formes et/ou des catégories.

Par exemple, sur la copie d'écran ci-dessous, on effectue une recherche sur les noms précédés du lemme *beau*.



AnaText 2.3

Texte source (utf-8)

Texte étiqueté (utf-8)

Sortie brute de Treetagger

Statistiques générales
Lemmes spécifiques
Tranches de fréquence
Accroissement du voc.
Parties du discours
Morphologie verbale
Noms lemmatisés
Verbes lemmatisés
Adjectifs lemmatisés
Adverbes lemmatisés
Tous les lemmes
Toutes les formes
Concordance
Cooccurrences
Recherche de patterns
Segments répétés

Sauvegarder

Entrez un pivot (forme) ici (pattern accepté)

beau NOM Rechercher

Mot Entier Sensibilité à la casse Recherche de lemme

Fenêtre = +/- 40 mots

Etiquettes reconnues : NOM, NPR, VER, ADJ, ADV, DET, PRO, PRE, CON, NUM (nombre), PHR (fin de phrase), PON (ponctuation), TOK (token quelconque), NOSENT (token quelconque sauf la marque de fin de phrase). Les expressions régulières sont permises (avec les métacaractères ".*+?" à l'intérieur des tokens et "(|){}" à l'extérieur). Par exemple :

- *guerre TOK{1,4} paix* => pour les concordances de *guerre* suivi de *paix* avec de 1 à 4 tokens entre les deux.
- *.*tion_NOM* => pour tous les noms se terminant par -tion
- *dé.*_VER* => pour tous les verbes commençant par dé-

Occurrence(s) : 4

Show 10 entries Search:

Position	Contexte gauche	Pivot	Contexte droit
20.2%	, était Cornudet le démoc , la terreur des gens respectables . Depuis vingt ans , il trempait sa barbe rousse dans les bocks de tous les cafés démocratiques . Il avait mangé avec les frères et amis une assez	belle fortune	qu' il tenait de son père , ancien confiseur , et il attendait impatiemment la République pour obtenir enfin la place méritée par tant de consommations révolutionnaires . Au quatre septembre , par suite d' une farce peut-être , il
27.13%	seul en accepta deux gouttes , et , lorsqu' il rendit la gourde , il remercia : " C' est bon tout de même , ça réchauffe , et ça trompe l' appétit . " L'	belle humeur	et il proposa de faire comme sur le petit navire de la chanson : de manger le plus gras des voyageurs . Cette allusion indirecte à Boule de suif choqua les

6.2. Repérage de cooccurrence

Les mots apparaissant de façon priviligiée avec un mot, un lemme ou une catégorie peuvent être recherchés par une fonction de cooccurrence (avec plusieurs mesures d'association comme le log-like, l'information mutuelle ou le t-score). Par exemple, ci-dessous, on voit apparaître les cooccurrences privilégiées de *femme*.

Recherche de cooccurrences

Entrez un pattern composé de lemmes ou d'étiquettes

femme Rechercher

Fenêtre = -5 / +5 formes ☑ Mot Entier ☐ Sensible à la casse ☑ Recherche de lemme ☐ A l'intérieur des phrases

Etiquettes reconnues : NOM, VER, ADJ, ADV, DET, PRO, PRE, CON, NUM (nombre), PHR (fin de phrase), PON (ponctuation)

Nota bene : dans les indices ci-dessous le logarithme est calculé en base 2

Show 25 entries Search:

Collocatif	Cat	Cooccurrences	LogLike	Im	t-score
le	DET	39	274.013	21.418	3.775
un	DET	19	124.794	18.094	4.262
tout	PRO	9	64.013	14.945	4.743
du	PRE	11	63.36	12.545	4.022
son	DET	10	60.702	12.818	4.2
de	PRE	14	53.789	9.067	2.953
honnête	ADJ	2	26.71	19.868	7.639
alors	ADV	3	25.477	11.493	5.524
et	KON	7	25.102	6.371	2.939
sacrifice	NOM	2	25.008	17.748	7.317
cœur	NOM	2	23.705	16.181	7.054

6.3. Recherche de patrons

La recherche de patrons permet d'extraire et calculer la fréquence de patrons lexico-syntaxiques (avec formes, lemmes et catégorie). Par exemple, ci-dessous apparaît le patron de type Nom + verbe *être* + adjectif.

Recherche de patterns

Entrez un pattern composé de formes ou d'étiquettes

NOM être ADJ Rechercher

☑ Mot Entier ☑ Sensibilité à la casse ☑ Recherche de lemme

Etiquettes reconnues : NOM, VER, ADJ, ADV, DET, PRO, PRE, CON, NUM (nombre), PHR (fin de phrase), PON (ponctuation)

Show 10 entries Search:

Segment	Fréquence
après-midi fut lamentable	1
C' était vrai	1
cidre était bon	1
conversation fut vive	1
demi-heure était nécessaire	1
déjeuner fut tranquille	1
figures étaient pâles	1
force sont inutiles	1
indignation fut vive	1
inquiétude était extrême	1

Showing 1 to 10 of 15 entries First Previous Next Last

6.4. Segments répétés

On peut également extraire des segments répétés, c'est-à-dire les suites de mots les plus récurrentes dans les textes. On peut extraire tous les patrons, ou plus spécifiquement les patrons comportant une forme ou des catégories. Par exemple, la copie d'écran ci-dessous indique les segments répétés comprenant le mot *quartier*.

Recherche de segments répétés

Entrez un pattern composé de formes ou d'étiquettes

prussien Rechercher

Longueur min. 3 - Fréquence min. 2

Etiquettes reconnues : NOM, VER, ADJ, ADV, DET, PRO, PRE, CON, NUM (nombre), PHR (fin de phrase), PON (ponctuation)

Show 10 entries Search:

Longueur	Segment	Fréquence
7	le officier prussien faire demander à mademoiselle	2
7	prussien faire demander à mademoiselle Elisabeth Rousset	2
7	officier prussien faire demander à mademoiselle Elisabeth	2
3	le officier prussien	4
3	de ce prussien	2

Showing 1 to 5 of 5 entries First Previous Next Last