Nouvelle catégorisation de Frantext

Sandrine Ollinger

ATILF - Nancy

3 novembre 2016

Vue d'ensemble

Un travail d'équipe

- Un comité d'experts :
  - Christophe Benzitoun, Isabelle Clément, Mathieu Delabarre, Bertrand Gaiffe, Véronique Montémont, Étienne Petitjean, Gilles Souvay
- Des renforts :
  - Ulrike Fleury (4 mois), Lolita Bérard (12 mois), Stéphane Tiv (2 mois), Clémence Urtebize (5 mois)
- Des conseils judicieux et des coups de pouces heureux :
  - Achille Falaise , Marianne Vergez-Couret, Franck Sajou, Assaf Urieli, Marie Tonnelier, Martin Lentschat

Un projet de longue haleine

- 18 mois de travail

Plan du séminaire

L'existant

Ce qui était avant notre travail

La stratégie

Les choix opérés, le travail réalisé

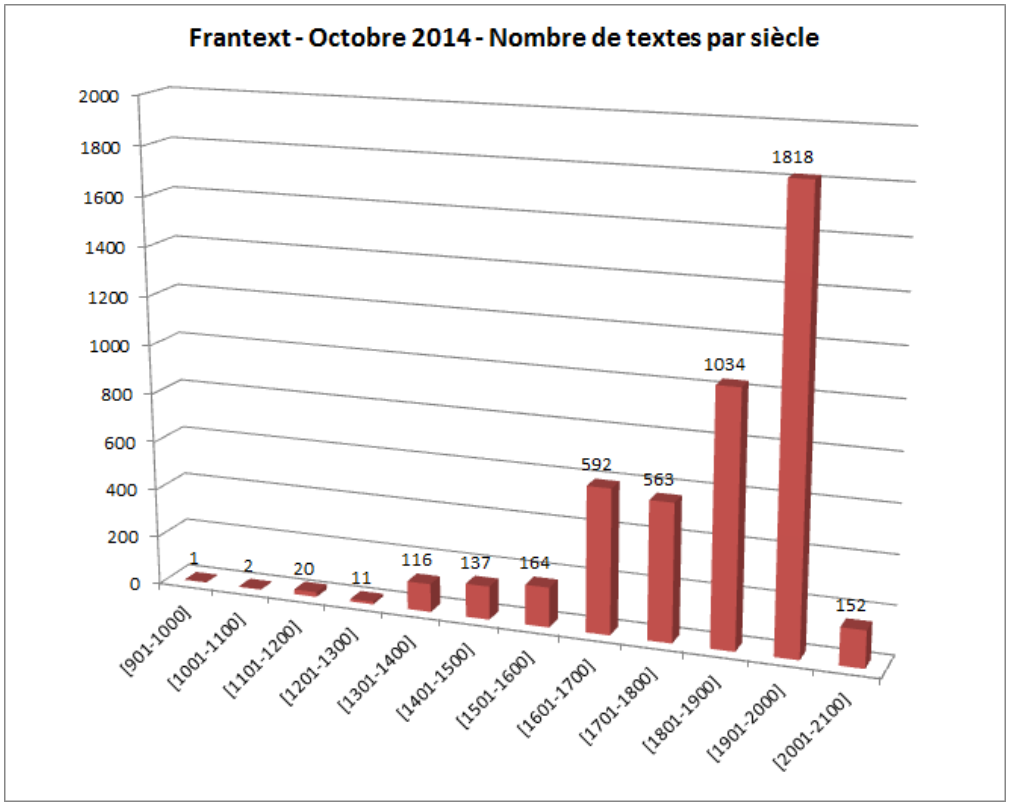
Le résultat

Là où l'on est arrivé et ce qu'il reste à fignoler...

L'existant

Base textuelle FRANTEXT

- 4 613 références
- 277 413 739 mots
- du Xe au XXIe siècle



Périodes de "numérisation"

Avant 2000

- côtes B, K, L, M, P, Q, S, T, Z

Depuis 2000

- côtes E, R

Frantext catégorisé

Étendue

- 1 940 références
- 127 515 681 mots
- du XIXe au XXIe siècle

Format

- XML
- TEI P5
- UTF-8

```
141 sur ses pieds de derrière, des enchevêtrements<lb/>
142 de gens et de chevaux... qu'est-ce que c'est<lb/>
143 que tout ça ? ...<lb/>
144 tout ça, c'est, dans des cadres empire assez<lb/>
145 beaux, le <hi rend="I"> *Napoléon franchissant le</hi><lb/>
146 <hi rend="I"> *Saint-*Bernard, </hi> de *David ; les <hi rend="I"> batailles
147 <hi rend="I"> *Montmirail, </hi> de <hi rend="I"> *Friedland </hi> et d'<hi r
148 d'*Horace *Vernet, et la <hi rend="I"> bataille d'*Austerlitz, </hi><lb/>
149 de *Gérard. C'est entre ce *Napoléon et ces batailles<lb/>
150 que j'ai vécu jusqu'à seize ans.<lb/>
151 Ce matin-là, comme je parais regarder,<lb/>
152 m'intéresser à quelque chose, la jeune femme<lb/>
153 fraîche, qui a des robes qui craquent et que<lb/>
154 j'appelle en moi-même <hi rend="I"> an itroun </hi> dit aux<pb n="6"/>
155 moustaches blanches, inclinées vers moi :<lb/>
156 -elle va peut-être parler ! ... est-ce que<lb/>
157 bonne maman l'a déjà vue ?<lb/>
158 -non ! ... -répond la dame aux belles<lb/>
```

La mission

Ce qui change

- Nouvel outil d'interrogation (Etienne)
- Mise en ligne des textes libres de droits sur la plateforme Ortolang

Ce que l'on souhaite

- Un étiquetage morpho-syntaxique de Frantext (d'ici fin 2015)

La stratégie

Quelles performances des instruments existants sur corpus littéraire ?

Expérience

- 4 étiqueteurs
- 7 textes extraits du corpus de Josette Lecompte
  - 1836-1927
  - 237 135 tokens

	complet	lemmes	POS	POS gros grain
Talismane	71,51%	83,24%	**85,22%**	**95,11%**
MELt	76,21%	80,24%	83,80%	92,58%
TreeTagger	73,59%	**89,67%**	78,47%	94,18%
MarsaTag	**76,81%**	87,10%	82,43%	94,54%

Talismane offre les meilleurs résultats, mais doit pouvoir être amélioré

Choix généraux

Choix généraux

On se concentre sur Frantext moderne

- Textes écrits après 1850

Choix généraux

On se concentre sur Frantext moderne

- Textes écrits après 1850

On définit notre propre jeu d'étiquettes

- Inspiré de Crabbé & Candito (2008)

Choix généraux

On se concentre sur Frantext moderne

- Textes écrits après 1850

On définit notre propre jeu d'étiquettes

- Inspiré de Crabbé & Candito (2008)

On ré-entraîne un instrument

- Talismane (Assaf Urieli 2013)

Jeu d'étiquettes

ADJ	adjectif	P+D	préposition + déterminent
ADV	adverbe	PONCT	ponctuation
CC	conjonction coordination	PRO	pronom
CS	conjonction 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	participe passé
NC	nom commun	VPR	participe présent
NP	nom propre	X	mot non traité

Talismane

Talismane

- Logiciel libre
- Multiplateformes (java)
- Documentation abondante

Talismane

- Logiciel libre
- Multiplateformes (java)
- Documentation abondante

entrée

- texte brut
- XML

Talismane

- Logiciel libre
- Multiplateformes (java)
- Documentation abondante

entrée

- texte brut
- XML

sortie

- CoNLL

Talismane

- Logiciel libre
- Multiplateformes (java)
- Documentation abondante

entrée

- texte brut
- XML

sortie

- CoNLL

modulaire

- Segmentation en phrases
- Segmentation en unités lexicales
- Étiquetage en partie du discours
- Annotation en dépendance syntaxique

Fonctionnement des modules de Talismane

Fonctionnement des modules de Talismane

Segmentation en phrases

- apprentissage statistique version originale

Fonctionnement des modules de Talismane

Segmentation en phrases

- apprentissage statistique version originale

Segmentation en unités lexicales

- apprentissage statistique à partir de patrons
- OU tokenisation par règles (REGEXP) notre choix

Fonctionnement des modules de Talismane

Segmentation en phrases

- apprentissage statistique version originale

Segmentation en unités lexicales

- apprentissage statistique à partir de patrons
- OU tokenisation par règles (REGEXP) notre choix

Étiquetage en partie du discours

- apprentissage statistique
- prise en compte de lexiques nous exploitons Morphalou 3
  - distinction lexiques de classes ouvertes vs. classes fermées
- traitements spécifiques par jeux de règles

Corpus d'apprentissage

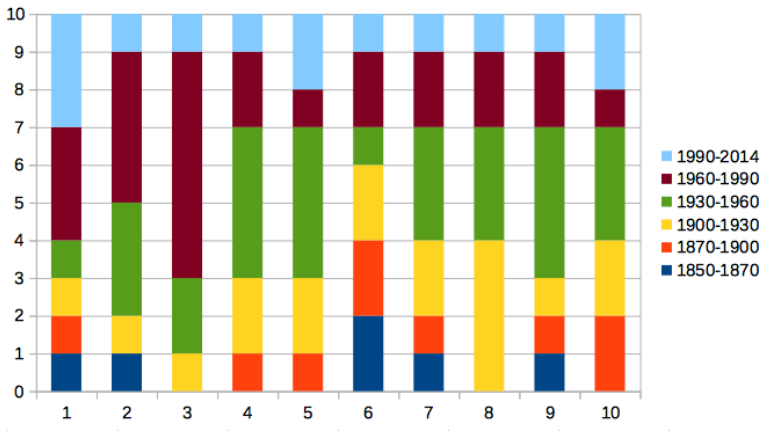
Organisation générale

- 100 échantillons de 2 000 mots
- répartis en 10 tranches de 20 000 mots
- chaque tranche couvre l'intégralité de la diachronie 1850-2014
- genres intégrés au fur et à mesure des tranches
- difficulté syntaxique grandissante

Corpus d'apprentissage

Organisation générale

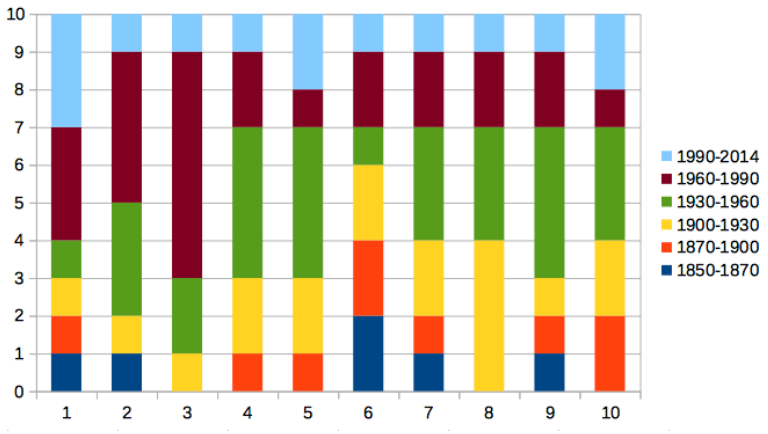
- 100 échantillons de 2 000 mots
- répartis en 10 tranches de 20 000 mots
- chaque tranche couvre l'intégralité de la diachronie 1850-2014
- genres intégrés au fur et à mesure des tranches
- difficulté syntaxique grandissante

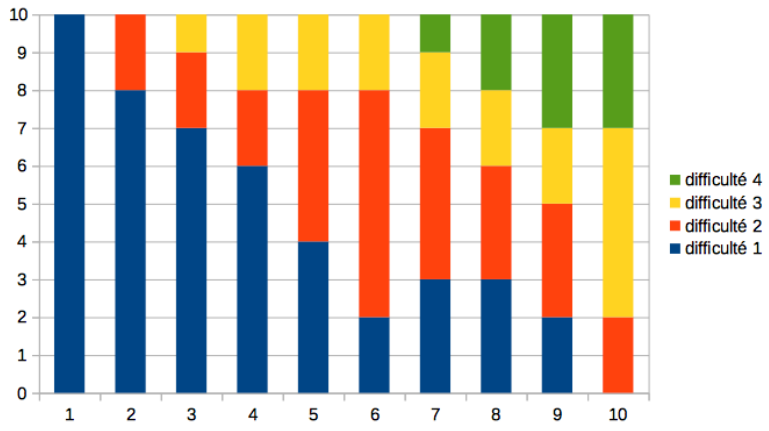


Corpus d'apprentissage

Organisation générale

- 100 échantillons de 2 000 mots
- répartis en 10 tranches de 20 000 mots
- chaque tranche couvre l'intégralité de la diachronie 1850-2014
- genres intégrés au fur et à mesure des tranches
- difficulté syntaxique grandissante





Stratégie (1)

1ère étape

- 20 000 mots étiquetés par Talismane initial
- Correction manuelle
  - 2 équipes
  - 2 correcteurs par texte
  - 10 000 mots par correcteur
- Gestion des désaccords
- Entraînement de Talismane sur 20 000 mots

Stratégie (1)

1ère étape

- 20 000 mots étiquetés par Talismane initial
- Correction manuelle
  - 2 équipes
  - 2 correcteurs par texte
  - 10 000 mots par correcteur
- Gestion des désaccords
- Entraînement de Talismane sur 20 000 mots

réalisée du 15 au 31 juillet 2015

Stratégie (2)

étapes suivantes

- étiquetage des 20 000 mots suivants
- correction
- gestion des désaccords
- Réentraînement

Stratégie (2)

étapes suivantes

- étiquetage des 20 000 mots suivants
- correction
- gestion des désaccords
- Réentraînement

réalisée de août à novembre 2015

Stratégie (3)

tâches annexes

- évaluation de l'amélioration
- optimisation du lexique
- optimisation de la tokenisation
- optimisation du guide d'annotation

Stratégie (3)

tâches annexes

- évaluation de l'amélioration
- optimisation du lexique
- optimisation de la tokenisation
- optimisation du guide d'annotation

réalisée de juillet 2015 à juillet 2016

Pendant que les correcteurs corrigent

Optimisation des lexiques

Morphalou 3

- Exclusion des lemmes peu fréquents pour une même forme
  - documentation Morphalou 3 : 402 formes avec + de 3 lemmes
  - mots repérés au fil des textes
- verbes pronominaux

Lexique interne Talismane : mots de classes fermées

- 1661 suppressions : locutions, nombres, fautes d'orthographe
- ajouts : mots-composés, PROWH, ça en tant que CLS
- modification de lemmes et de catégories

```
ajax:Talismane sollinge$ java -Xmx1G -jar talismane-core-2.4.1b.jar command=seri
alizeLexicon lexiconProps=apprentissageFrantext/lexicons/lexicons_frantext.txt o
utFile=apprentissageFrantext/lexicons/lexicons_frantext.zip posTagSet=apprentiss
ageFrantext/frantextTagset.txt encoding=UTF8 logConfigFile=apprentissageFrantext
/conf/log4j.properties
```

Optimisation de la tokenisation

- adaptation aux choix réalisés lors de la correction
- ajout de mots composés
- ajout de règles de préfixation

```
TokenRegexFilter    \b(([Pp]arce)|([Tt]andis)) que\b
TokenRegexFilter    \b((PARCE)|(TANDIS)) QUE\b
TokenRegexFilter    \b(([Pp]arce)|([Tt]andis)) qu'
TokenRegexFilter    \b((PARCE)|(TANDIS)) QU'
TokenRegexFilter    \bparce qu'
TokenRegexFilter    \bPARCE QU'
TokenRegexFilter    \b[Pp]arc'que\b
TokenRegexFilter    \bPARC'QUE\b
TokenRegexFilter    \b[Pp]arc'qu'
TokenRegexFilter    \bPARC'QU'
TokenRegexFilter    \b[Pp]ac'que\b
TokenRegexFilter    \bPAC'QUE\b
TokenRegexFilter    \b[Pp]ac'qu'
TokenRegexFilter    \bPAC'QU'
TokenRegexFilter    \b[Aa]fin ((de)|(que))\b
TokenRegexFilter    \bAFIN ((DE)|(QUE))\b
TokenRegexFilter    \b[Aa]fin ((d')|(qu'))
TokenRegexFilter    \bAFIN ((D')|(QU'))
TokenRegexFilter    \b[Aa]uprès ((de)|(des)|(du))\b
TokenRegexFilter    \bAUPR[EÈ]S ((DE)|(DES)|(DU))\b
TokenRegexFilter    \b[Aa]uprès d'
TokenRegexFilter    \bAUPR[ÈE]S D'
TokenRegexFilter    \b[Qq]uant ((à)|(au)|(aux))\b
TokenRegexFilter    \bQUANT (([AÀ])|(AU)|(AUX))\b
```

Une fois le corpus d'apprentissage complet

Optimisation des choix d'annotation (1)

Réunions de travail

- Christophe Benzitoun & Véronique Montémont
- Lecture exhaustive de la documentation
- Suggestions d'amélioration de l'ergonomie
- Harmonisation
- Discussions de certains points
- Mise en avant de deux points particulièrement problématiques :
  - Qu'est-ce qu'un nom propre ?
  - Qu'est-ce qu'un mot composé ?

Optimisation des choix d'annotation (2)

Nom propre

- Est un nom propre : prénom, nom de famille, toponyme, nom d'entreprise
- N'est pas un nom propre : ethnonymes et gentilés, particule "De", titres (de films, de livre), noms de journaux, etc.

Optimisation des choix d'annotation (2)

Nom propre

- Est un nom propre : prénom, nom de famille, toponyme, nom d'entreprise
- N'est pas un nom propre : ethnonymes et gentilés, particule "De", titres (de films, de livre), noms de journaux, etc.

Mot-composé

- Formes comportant un trait d'union dans Morphalou 3
- Formes comportant une apostrophe ou une espace dans Morphalou 3 (après révision par C.B. et V.M.)
- Formes commençant par un préfixe
  - liste de préfixes extraite du GLAWI par Martin Lentschat
  - préfixes courants dans Morphalou 3

Conséquences chronophages

- Constitution d'un lexique de NP
- Révision complète du fichier de tokenisation
  - suppression des expressions régulières correspondant à des regroupements abandonnés
  - ajout des expressions régulières correspondant aux nouveaux regroupements souhaités
  - généralisation des expressions régulières des noms communs, adjectifs et verbes pour qu'elles soient sensibles à la flexion
- Révision du corpus d'apprentissage :
  - Tokenisation
    - Regrouper
    - Segmenter
  - Étiquettes NP

Répartition du corpus en train/dev/test

TRAIN	DEV	TEST
28 traités-essais	3 traités-essais	6 traités-essais
22 romans	3 romans	4 romans
14 écrits personnels	1 écrits personnels	3 écrits personnels
10 poésie	/	/
6 théâtre	/	/
= 80 fichiers	= 7 fichiers	= 13 fichiers

Répartition difficulté

- Dev : 2, 3, 4
- Test : 1, 2, 3

Détail Test et Dev

corpus		genre	diff	1850-2014	auteur	titre	date	cote	n	n	n	%	distribution
4	5	Roman	3	1960-1990	Benoziglio	La boite noire	1974	R174					DEV
4	10	Traités/Essais	2	1930-1960	Levi-Strauss	Anthropologie structurale	1958	P652					DEV
5	5	Roman	2	1870-1900	Hugo	Les Misérables	1862	S739					DEV
8	10	Traités/Essais	4	1960-1990	Barthes	Le plaisir du texte	1973	S599					DEV
9	1	EP	4	1990-2014	Roubaud	La Boucle	1948	R001					DEV
10	4	Roman	4	1900-1930	Proust	Du côté de chez Swann	1913	K428					DEV
10	8	Traités/Essais	3	1870-1900	Flammarion	Astronomie populaire	1880	P996					DEV

corpus		genre	diff	1850-2014	auteur	titre	date	cote	n	n	n	%	distribution
1	9	Roman	1	1990-2014	Letessier	Le voyage à Paimpol			##	##		##	TEST
1	10	Traités/Essais	1	1990-2014	Lejeune	Signes de vie	2005	R165	##	##	##	##	TEST
2	8	Roman	2	1960-1990	Marie Chaix	L'âge du tendre	1979	R065					TEST
4	2	EP	3	1990-2014	Des Forêts	Ostinato	1997	R212					TEST
5	2	EP	1	1990-2014	Duperey	Le voile noir	1992	R038					TEST
6	8	Traités/Essais	2	1900-1930	Durkheim	De la division du travail s	1911	K997					TEST
7	1	EP	2	1990-2014	Modiano	Un pedigree	2005	R123					TEST
8	4	Traités/Essais	1	1930-1960	Daudet	Bréviaire du journalisme	1936	P597					TEST
8	8	Traités/Essais	3	1930-1960	Valéry	Variété IV	1938	K555					TEST
9	4	Roman	1	1930-1960	Mauriac	Le nœud de vipères	1933	K246					TEST
9	9	Traités/Essais	2	1960-1990	Françoise Dolt	La cause des enfants	1985	S318					TEST
10	3	Roman	3	1900-1930	Huysmans	A rebours	1903	L486					TEST
10	9	Traités/Essais	3	1930-1960	Schaeffer	A la recherche d'une mus	1952	L777					TEST

Optimisation des lexiques

Nettoyage / Suppressions

- des NP dans le lexique de NC de Morphalou 3
- *uns* comme déterminant
- *comment*, *pourquoi* et *quand* comme adverbe

Ajouts

- création d'un lexique supplémentaire
  - 673 formes, 41 lemmes
- complétion du lexique de classes fermées
- création de lexiques de mots étrangers
  - anglais, allemand, latin
  - issus du Wiktionnaire et de Wikibooks

Optimisation de la prise de décision de Talismane

Contournement de l'apprentissage statistique

- Traitement de *que*
- Adaptation des règles proposées par Assaf Urieli

Évaluation finale

Répartition tokens dans les trois sous-corpus

train

80 extraits

165 784 tokens

9708 ADV
9581 ADJ
8058 CLS
4898 CLO
4463 CC
3925 VINF
3693 VPP
3536 NP
2926 P+D
29226 NC
2472 PROREL
2450 CS
24294 PONCT
2150 PRO
20162 DET
17271 P
15002 V
726 VPR
434 I → ***0,26%***
408 PROWH
237 X
164 ET

dev

7 extraits

14 427 tokens

2674 NC
2098 PONCT
1857 DET
1554 P
1207 V
888 ADJ
872 ADV
590 CLS
372 CC
324 CLO
312 VPP
287 NP
283 VINF
273 P+D
234 PROREL
213 CS
187 PRO
79 VPR
61 ET → ***0,42%***
31 PROWH
23 I
8 X

test

13 extraits

26 492 tokens

4640 NC
3507 PONCT
3302 DET
2923 P
2310 V
1787 ADV
1485 ADJ
1291 CLS
839 CLO
780 CC
671 VPP
638 VINF
493 PROREL
430 P+D
414 CS
376 NP
359 PRO
92 VPR
60 PROWH → ***0,23%***
45 ET
31 I
19 X

Résultat final

Test

- 98,16 de précision
- 97,96 de kappa

Dev

- 98 de précision
- 97,77 de kappa

Comparaison précisions

MElt 97,7 - Talismane FTB 97,55 - Standford 97,28

Catégories farouches

- ET, PROWH et X n'atteignent jamais f-score 90%
- I moins de problème en test qu'en dev

Le résultat

Résumé de ce qui a été fait

- choix catégoriseur pour le français contemporain
  - Talismane d'Assaf Urieli
- choix d'un jeu d'étiquettes
- constitution d'un corpus d'apprentissage
- étiquetage et correction du corpus
- définition de règles de segmentation en unités lexicales
- adapatation de lexiques
- révision du corpus d'apprentissage
- écriture de règles pour améliorer l'étiquetage de *que*
- évaluation de la qualité de l'apprentissage
- étiquetage de **FrantextPost1850**

Étiquetage Frantext Post 1850

- 2 369 références (+ 429)
- 136 495 607 tokens (+ 7%)

- Tout ce qui a été fait est disponible sur un serveur
- Tout est accompagné de documentation
- Tout est reproductible
  - pré-traitement → scripts Ulrike
  - modifications manuelles détaillées
  - Talismane final et instructions
  - post-traitement → script perl
    - ajustement de la segmentation (*ce grand-là*)

Format de sortie

```
1   Il  il  CLS Il  22  19  22  21  99,98
2   a   avoir   V   a   22  22  22  23  99,85
3   plu plaire  VPP plu 22  24  22  27  99,00
4   .   .   PONCT   .   22  27  22  28  100,00

1   Des un  DET Des 22  29  22  32  99,47
2   flaques flaque  NC  flaques 22  33  22  40  99,94
3   d'  de  P   d'  22  41  22  43  99,92
4   eau eau NC  eau 22  43  22  46  99,63
5   reflètent   refléter    V   reflètent   22  47  22  56  99,85
6   les le  DET les 22  57  22  60  99,92
7   derniers    dernier ADJ derniers    22  61  22  69  97,50
8   nuages  nuage   NC  nuages  23  1   23  7   99,91
9   .   .   PONCT   .   23  7   23  8   100,00
```

- La 1re colonne contient un identifiant
- La 2e colonne contient la forme étiquetée par Talismane
- La 3e colonne contient le lemme proposé par Talismane [ou une copie de la forme]
- La 4e colonne contient l'étiquette
- La 5e colonne contient le texte (+/-) original
- Les colonnes 6 à 9 donnent des informations d'emplacements dans les fichiers originaux, en nombres de lignes et de colonnes
- La dernière colonne contient la probabilité qui a amené Talismane à attribuer la partie du discours [dans les cas où la proba vaut 00,00 l'étiquette a été modifié en post-traitement ou le token créé]

En conclusion : ce dont on dispose

- Une version de Talismane pour étiqueter du texte littéraire post 1850
- un ensemble de scripts et instructions pour pré traiter les textes issus de Frantext
- un script pour post traiter les textes
- une documentation des choix de segmentation en unités lexicales
- un corpus d'apprentissage
  - équilibré en périodes et en genre
  - avec des extraits disposant de « notes » de difficultés
  - découpé en sous-corpus test-train-dev
  - couplé à une évaluation des performances de Talismane

Ce dont on ne dispose pas (encore ?)

- des fichiers XML en sortie
  - Bertrand y travaille
- une documentation des choix d'annotation terminée
  - Véronique, Christophe et Isabelle s'organisent
- une évaluation de l'étiquetage de tout Frantext
  - des bribes de stratégie
  - une suggestion de composition de corpus de test
  - rien d'exploitable en l'état
  - Etienne et Christophe ont proposé un projet tutoré aux master1 SCA
- une lemmatisation satisfaisante

Merci