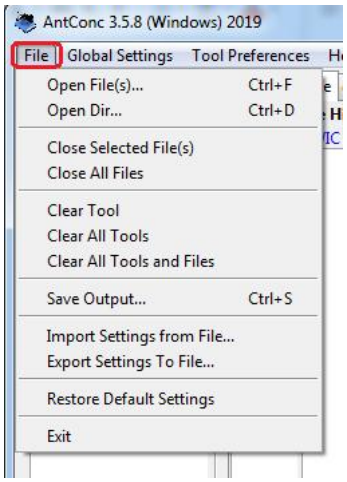
Présentation de AntConc

Plan du cours

1 Débuter un projet

AntConc est un concordancier gratuit, développé par le professeur Laurence Anthony, très utilisé dans l'analyse des corpus en linguistique. Vous allez retrouver la plupart des fonctionnalités que vous avez découvertes avec Frantext. Le principal atout est que vous pouvez travailler sur vos propres corpus.

Le menu *File* permet de choisir, de changer et d'effacer le ou les fichiers à traiter, de sauvegarder les résultats sur le disque dur et d'enregistrer les paramètres.

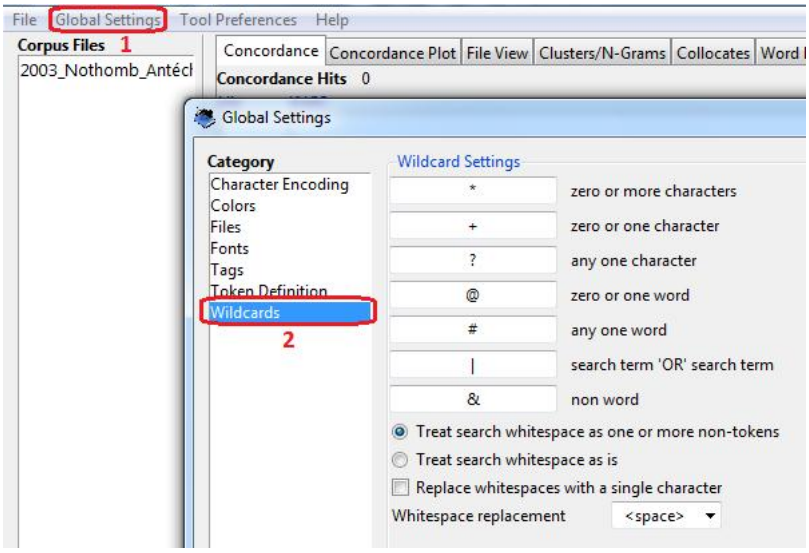


Antconc fonctionne, par défaut, sur des fichiers au format « Texte brut », reconnaissables à leur extension .txt [1]. C'est donc dans ce format que vous devez enregistrer vos documents de travail avant de les confier à AntConc.

Le logiciel offre deux types de configuration :

1. les paramètres généraux (Global Settings) : vous y trouverez la possibilité de régler l'encodage, la coloration syntaxique, les polices : je ne reviens pas sur ces points qui ne sont pas essentiels pour vous. En revanche, j'attire votre attention sur les wildcards (métacaractères) qui

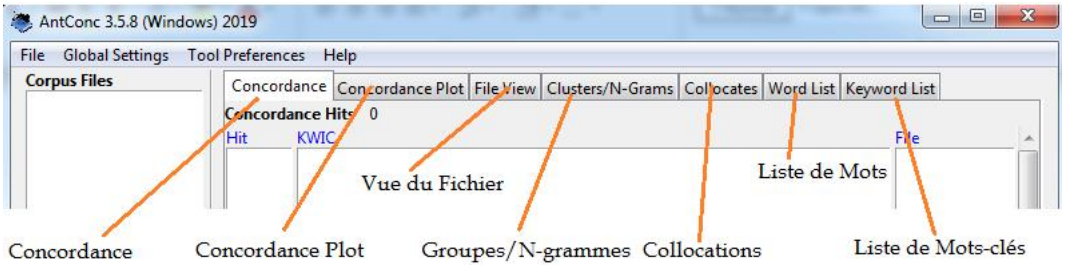
1. Les fichiers Word, de type .doc(x), comportent un en-tête et diverses informations sur la mise en forme, l'auteur, les dates de création et de toutes les versions du fichier ainsi que des statistiques, ce qui les rend plus volumineux que les textes bruts et difficilement exploitables avec Antconc.



vous sont très utiles si vous faites une recherche avec des expressions régulières (Regex).

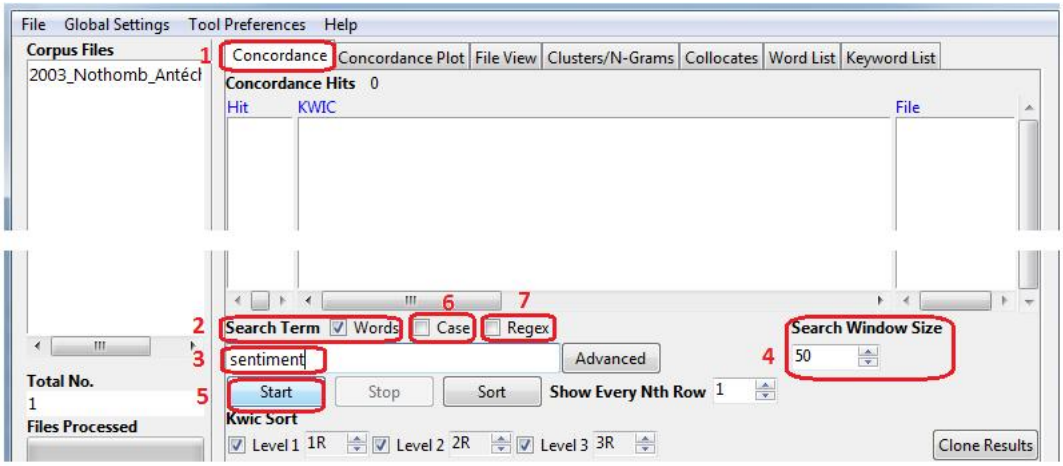
2. les outils (Tool Préférences) : nous y reviendrons au cas par cas.

La fenêtre principale du logiciel donne accès à sept outils : Concordance, Concordance Plot, File View, Clusters/N-grams, Collocates, Word List et Keyword List. Nous allons les explorons successivement dans le cadre de ce cours et du cours de la semaine prochaine.



2 Concordance

Options de recherche Antconc propose trois options de recherche, Words, Case et Regex.



— Avec l'option *Words*, cochée par défaut et seule cochée, la recherche est effectuée sans prise en compte de la casse des caractères. Tous les cas, minuscules et majuscules, sont relevés.
— En revanche, pour prendre en compte la casse, il faut cocher l'option *Case* (ex. : si on recherche le mot « État » et que l'option *Case* est cochée, le mot « état » ne sera pas pris en compte). Pensez à cocher aussi *Words* pour avoir seulement le mot saisi (dans notre exemple, le mot « États » ne sera pas retenu).
— L'option *Regex* permet de faire une recherche avec des expressions régulières. Je vous renvoie sur ce point à la fiche que nous avons étudiée.

Recherche avancée Il est possible de lancer une recherche à partir d'une liste en utilisant le module de recherche avancée (*Advanced*) :

1. si on a créé la liste dans un fichier .txt séparé, il faut cocher la case *Use search term(s) from list below* et télécharger le fichier en utilisant *Load File* ;
2. sinon, il faut cocher la case *Use Context Words and Horizons* et saisir un à un les mots qu'on veut trouver : à chaque saisie, on clique sur le bouton *Add* pour valider le mot. Immédiatement, le mot est placé dans la fenêtre en dessous. On répète l'opération d'ajout autant de fois que de mots voulus.

Tri des résultats (Kwic sort) Le tri est proposé pour observer les environnements des mots-clés recherchés et découvrir, pour un verbe par exemple, les prépositions avec lesquelles il est employé ou ses différentes structures. Ex. : pour le verbe « penser », on trouvera des emplois intransitifs, d'autres avec la préposition « à », d'autres encore suivis d'une proposition subordonnée introduite par la conjonction « que ». « Penser » peut être suivi également d'un verbe à l'infinitif et même d'un nom.

3 Concordance Plot

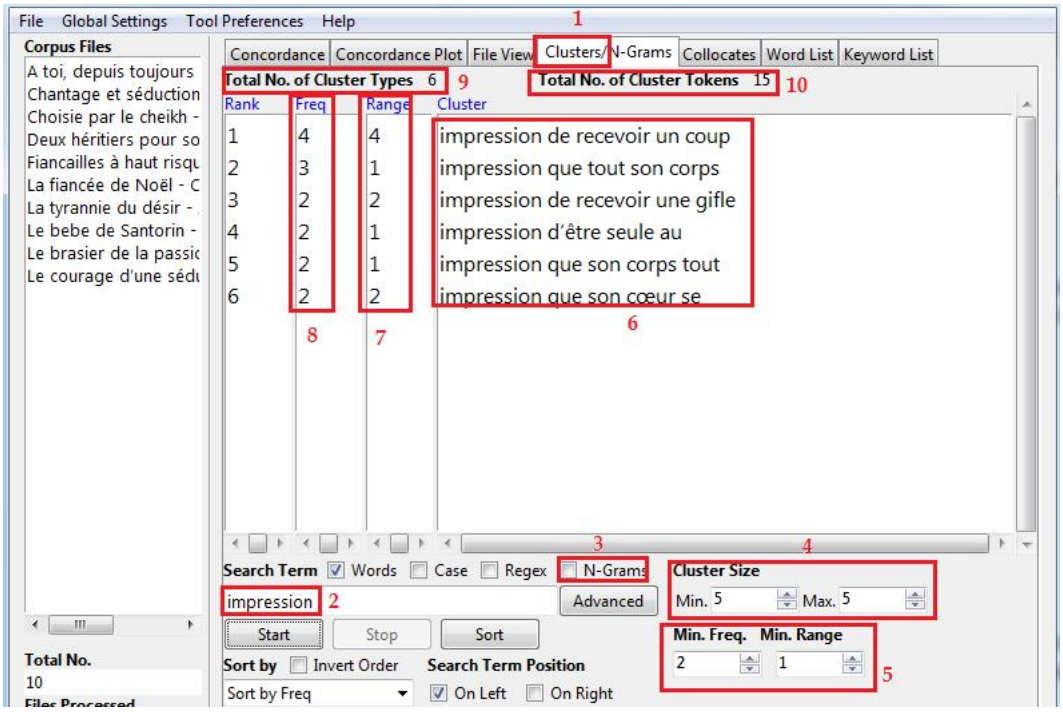
Cet onglet permet de visualiser le résultat de la concordance et les statistiques selon une vue longitudinale originale. Chaque texte analysé est représenté par un rectangle bleu qui correspond à la longueur du texte entier et où toutes les occurrences sont tracées sous forme de barres comparables à celles d'un code à barres indiquant approximativement la position des mots recherchés. On peut ainsi découvrir la répartition des emplois d'un mot ou d'une série de mots sur la longueur du texte.

Toutes les barres sont des hypertextes qui permettent d'afficher, sur simple clic, le volet Concordance et la concordance de l'occurrence en question.

4 Clusters/N-Grams

Clusters (séquences)

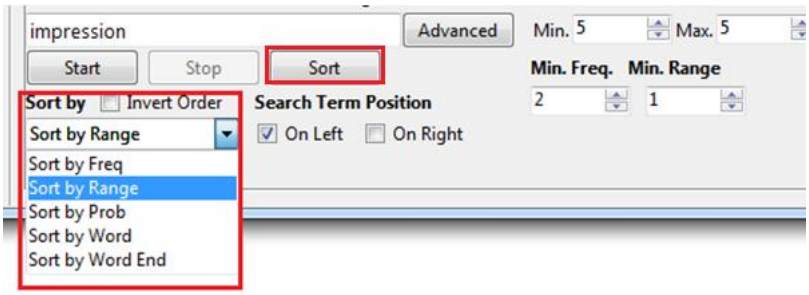
Cet outil produit une liste de groupes de mots contigus qui contiennent la requête tapée dans la zone de saisie. Il s'agit de retrouver les cooccurrents d'un mot ou d'une suite de mots qui pourraient constituer des clichés ou des habitudes de formules chez un ou plusieurs auteurs. Pour pouvoir saisir une requête *Clusters* (1), il ne faut pas cocher la case *N-Grams* (3). Il faut, par contre, régler, avec l'option *Clusters Size*, le nombre de mots des suites à rechercher, y compris le mot saisi, avec la possibilité de choisir un nombre minimum et un nombre maximum.



Dans l'exemple ci-dessus, le mot-clé recherché est impression (2). Le réglage *Clusters Size* (4) indique que la séquence doit comporter 5 mots (le mot-clé accompagné de 4 cooccurrents à sa droite (4+1=5). Deux options supplémentaires (5) servent à régler le minimum de fréquence (Min.Freq.), ici 2, et pour tous les textes, au moins 1 texte (Min.Range 1). Avec un seuil Min.Range de 2, la recherche n'a produit que les suites répétées au moins deux fois dans au moins deux textes différents.

La colonne *Freq* (8) donne le nombre des occurrences par ordre décroissant. La fenêtre précise également qu'il a été trouvé 6 différents types (9) sur un total de 15, celui des Tokens (10).

Quand le résultat est affiché, il est déjà trié par fréquence. D'autres tris sont proposés en plus de celui par fréquence. Un tri par *Range* (7) classera les résultats du plus grand nombre de textes au plus petit selon le minimum de fréquence voulu. Pour changer le tri, il faut utiliser le menu déroulant en bas de la fenêtre :



N-Grams

Il s'agit d'une séquence de taille n (par exemple une suite de 2, 3, 4 lettres ou mots ou plus) qui se trouve dans une séquence de taille plus grande que n (un texte par exemple). C'est la technique utilisée par les moteurs de recherche ou les smartphones pour compléter une saisie commencée en se fondant sur un modèle probabiliste : le programme construit des mots par n-gramme, c'est-à-dire en proposant la combinaison suivante la plus probable, en fonction des lettres saisies.

Contrairement à l'option *Clusters* qui cherche les cooccurrents d'un mot-clé spécifique, l'outil *N-Grams* procède à une recherche à l'aveugle en recherchant les suites de mots, seulement selon la longueur en mots indiquée. Ex. : chez un auteur comme Flaubert, apparaissent certaines séquences de 4 mots comme « de toutes ses forces », « elle baissa la tête », « elle se mit à », ce qui donne une certaine idée des personnages (féminins en l'occurrence ici) et des habitudes langagières du romancier.

5 Collocates

Cet outil permet de rechercher les collocations d'un terme recherché, autrement dit les groupes de mots comportant un terme pivot donné afin de localiser dans les textes des cas d'expressions idiomatiques résultant d'une co-occurrence systématique, c'est-à-dire des structures libres ou figées permises par les règles de combinaisons possibles, grammaticales ou lexicales, dans une langue donnée. C'est le cas notamment des clichés. Cette recherche ressemble

à celle de l'outil *Clusters* avec la différence que, cette fois, les résultats sont réalisés selon des paramètres statistiques.

6 Word List

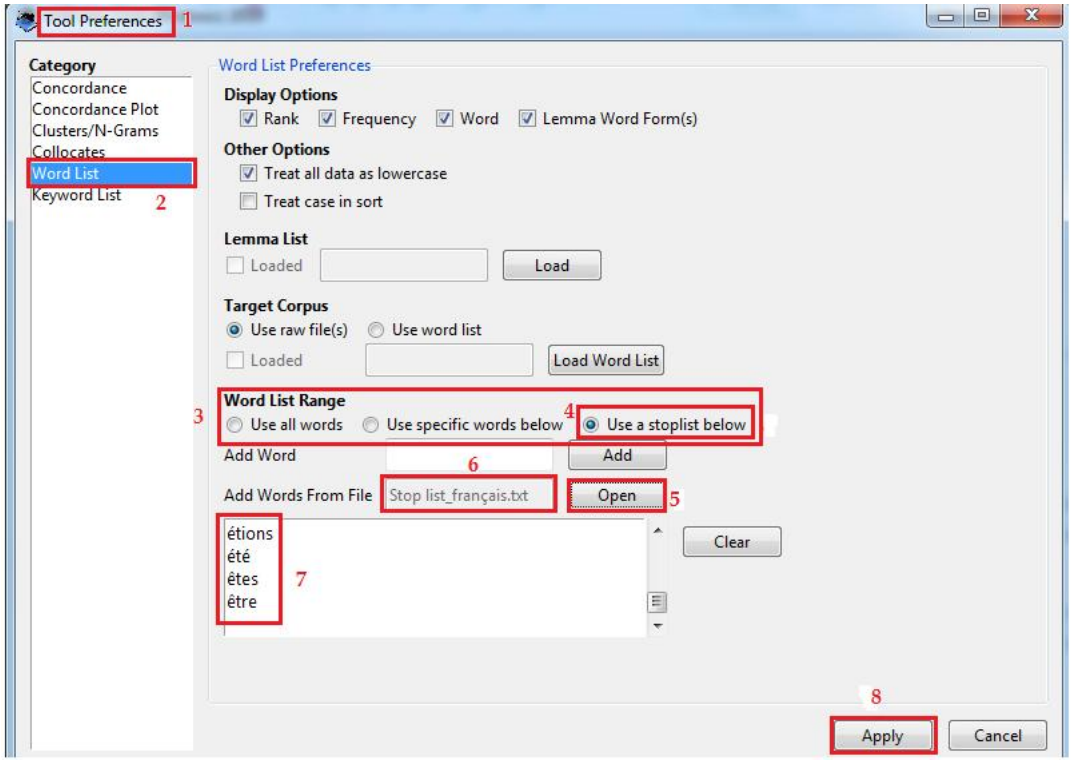
Cet outil permet d'obtenir l'index des fréquences des mots simples employés selon leur forme graphique dans le(s) texte(s) proposés. La liste obtenue classe les mots d'abord du plus fréquent au moins fréquent.

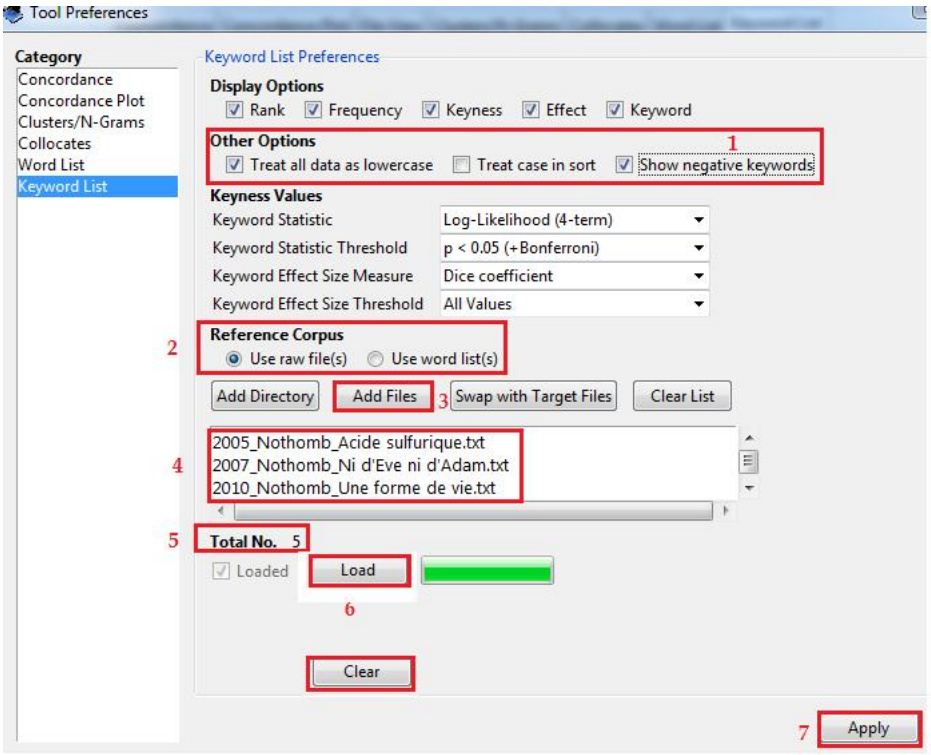
La plupart des textes d'une langue présentent une très grande fréquence des mots grammaticaux. En français, la préposition « de » vient toujours en tête du classement des mots de n'importe quel texte. Pour ne chercher que dans les mots sémantiquement pleins (nom, verbe, adjectif et adverbe) et empêcher le programme de produire l'index des mots grammaticaux, on utilise une *Stop List*, c'est-à-dire une liste créée à l'avance et qui comporte les mots sémantiquement vides. Vous trouverez cette liste sur la plate-forme sous le nom `Stop_list_francais.txt`. Pour charger cette liste, vous devez aller dans les *Tools Preferences* et choisir *Word List* : il faut alors ouvrir le fichier .txt comme le montre la figure de la page suivante.

7 Keyword List (liste de mots-clés)

Ce dernier outil sert à faire des comparaisons entre les mots-clés d'un corpus (un ou plusieurs textes d'un seul auteur, par exemple) et ceux d'un autre corpus de référence (d'un autre auteur) ou d'une liste préétablie, afin de trouver la liste des mots-clés les plus fréquents et donc spécifiques d'un auteur, d'un genre, etc., et ceux qui sont absents ou sous-employés par rapport aux mots du corpus de référence d'un autre auteur. On peut ainsi comparer des auteurs entre eux ou le vocabulaire d'un même auteur dans deux textes différents, des genres littéraires, ou des types de discours.

Pour charger le corpus de référence, qui servira de point de comparaison à notre corpus d'étude, il faut une fois encore aller dans les *Tools Preferences* et choisir *Keyword List* :


Tool Preferences
1
Category
Concordance
Concordance Plot
Clusters/N-Grams
Collocates
Word List
Keyword List
2
Word List Preferences
Display Options
Rank
Frequency
Word
Lemma Word Form(s)
Other Options
Treat all data as lowercase
Treat case in sort
Lemma List
Loaded
Load
Target Corpus
Use raw file(s)
Use word list
Loaded
Load Word List
3
Word List Range
Use all words
Use specific words below
4
Use a stoplist below
Add Word
6
Add
Add Words From File
Stop list_français.txt
Open
5
étions
été
êtes
être
7
Clear
8
Apply
Cancel


Tool Preferences
Category
Concordance
Concordance Plot
Clusters/N-Grams
Collocates
Word List
Keyword List
Keyword List Preferences
Display Options
Rank
Frequency
Keyness
Effect
Keyword
Other Options
1
Treat all data as lowercase
Treat case in sort
Show negative keywords
Keyness Values
Keyword Statistic
Log-Likelihood (4-term)
Keyword Statistic Threshold
p < 0.05 (+Bonferroni)
Keyword Effect Size Measure
Dice coefficient
Keyword Effect Size Threshold
All Values
2
Reference Corpus
Use raw file(s)
Use word list(s)
Add Directory
Add Files
3
Swap with Target Files
Clear List
4
2005_Nothomb_Acide sulfurique.txt
2007_Nothomb_Ni d'Eve ni d'Adam.txt
2010_Nothomb_Une forme de vie.txt
5
Total No. 5
Loaded
Load
6
Clear
7
Apply

Comme avec Word List, au lancement de la recherche, le programme demande au préalable l'indexation des mots du texte pour pouvoir trouver les mots-clés. Il suffit de l'autoriser à le faire en cliquant sur le bouton OK.

8 Sauvegarde des résultats

Tous les résultats qui apparaissent dans la fenêtre principale, à part le texte dans l'onglet *File View*, peuvent être rapatriés et enregistrés sur le disque dur de l'ordinateur dans un fichier au format .txt. Comme Antconc fonctionne dans la mémoire vive de l'ordinateur, il ne conserve pas les opérations après fermeture de la session. Il propose par défaut d'enregistrer les résultats sur le Bureau dans un fichier nommé `antconc_results.txt` (que vous pouvez renommer).

Pour exploiter plus facilement les résultats contenus dans ce fichier, il est conseillé d'utiliser un tableur : sélectionnez le contenu du fichier (Ctrl + A) et copiez-le dans un fichier de tableur. Exemple obtenu :

	A	B	C	D	E	F	G	H	I
1	1	verrai des fleuves aussi b		grands, aussi	sauvages, le	1984_L'Amant - Marguerite Duras.txt			
2	2	en livrée de coton blanc.		grande auto	funèbre de	1984_L'Amant - Marguerite Duras.txt			
3	3	rien comme c'était l'habi		grande auto	était là, long	1984_L'Amant - Marguerite Duras.txt			
4	4	forêt ni dans les villages		grandi autou	de nous. Il n	1984_L'Amant - Marguerite Duras.txt			
5	5	et celui qu'elle appelle s		grande cham	du premier	1984_L'Amant - Marguerite Duras.txt			
6	6	ges, ils occupent des imn		grands comn	des grands r	1984_L'Amant - Marguerite Duras.txt			
7	7	lui. Dès le portail passé e		grande cour	de récréatio	1984_L'Amant - Marguerite Duras.txt			
8	8	forte. Lui aussi il est né e		grandi dans	cette chaleu	1984_L'Amant - Marguerite Duras.txt			
9	9	. Et puis voici les deltas.		grands delta	de la terre. I	1984_L'Amant - Marguerite Duras.txt			
10	10	pas à Vinhlong quand on		grande eau.	Il était chez	1984_L'Amant - Marguerite Duras.txt			
11	11	e sur les photos récentes		grandi entre	nous. Une fo	1984_L'Amant - Marguerite Duras.txt			
12	12	parce qu'il savait que j'er		grande envie	et qu'il voul	1984_L'Amant - Marguerite Duras.txt			
13	13	pu en retrouver son imag		grande et	forte fièvre	1984_L'Amant - Marguerite Duras.txt			
14	14	j'ai dit autrement, quand		grandi et	qu'il est dev	1984_L'Amant - Marguerite Duras.txt			
15	15	me retourne et je vois. C		grande femn	très maigre,	1984_L'Amant - Marguerite Duras.txt			
16	16	'il y avait entre eux, faire		grands, le	regard plus	1984_L'Amant - Marguerite Duras.txt			
17	17	bac, à côté du car, il y a u		grande limou	noire avec u	1984_L'Amant - Marguerite Duras.txt			
18	18	chambres d'asile blanchi		grands lits	en fer noirs	1984_L'Amant - Marguerite Duras.txt			