Les corpus de la formation peuvent être téléchargés sur http://pro.aiakide.net/cours/txm

1 Charger un corpus

Pour charger un corpus au format **.txm**, il faut utiliser… « Fichier→ Charger ».

2 Observer son corpus

Dans « Corpus→ Description », vous avez un aperçu de ce que contient le corpus. Dans **Statistiques générales**, l'élément important à retenir est le nombre de mots. Lorsque vous décrivez un corpus, il faut toujours indiquer combien de mots il contient.

À la base, un corpus ne contient que des mots. Il n'y a pas vraiment de « standard » d'annotation, cela varie souvent en fonction des corpus. Il convient de décrire ces annotations (brièvement !) lorsque l'on décrit un corpus de travail, en détaillant évidemment toujours un peu plus les annotations que l'on utilise.

2.1 Propriétés des unités lexicales

Ces corpus sont annotés en lemmes (étiquette *lemma*) et en parties du discours (*pos* ou *cpos*).

Gardez sous la main la documentation du jeu d'étiquettes de votre corpus, vous pourrez en avoir besoin :

- Jeu d'étiquettes par défaut de TreeTagger pour le français
- Jeu d'étiquettes TreeTagger pour le français oral (perceo) (pages 7 et 8)

2.2 Propriétés des structures

Les structures vont nous permettre de comparer des parties du corpus entre elles.

Corpus Soirée électorale

Il n'existe qu'une seule structure « texte », avec une propriété « locuteur » (*loc*). On peut ainsi comparer les locuteurs entre eux.

Corpus Investitures

Il existe en plus de « locuteur », une propriété « date », qui permet d'étudier l'évolution des discours dans le temps, et une propriété « mandat » pour comparer les discours de « nouveaux » présidents (mandat 1) avec celui de présidents sortants (mandat 2).

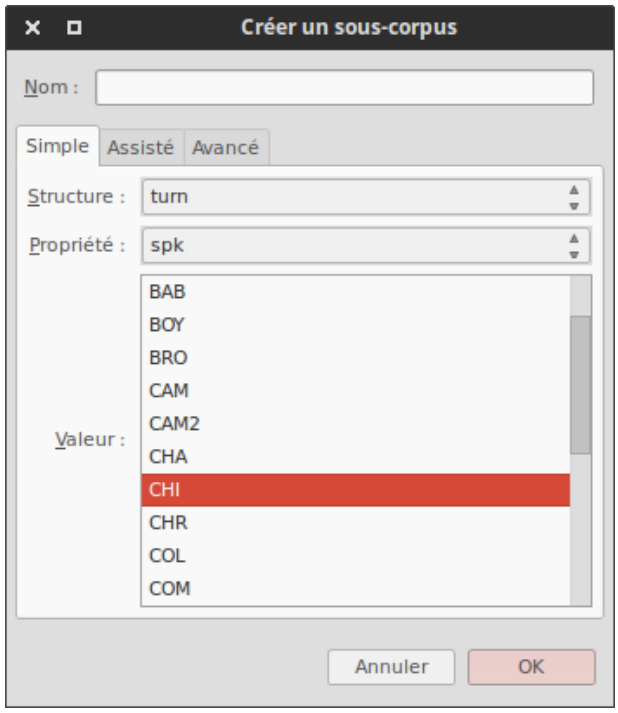
Corpus COLAJE

Le corpus est organisé en plusieurs structures qui s'emboîtent (Épisode→Section→Tour de parole→Utterance). En pratique, on utilisera uniquement la structure « tour de parole » (*turn*), qui comporte les étiquettes nom et âge de l'enfant, locuteur et rôle du locuteur.

2.3 Qualitatif

Un corpus ne se résume pas qu'à des annotations. Comment le corpus a-t-il été constitué ? En quoi est-il intéressant pour votre problématique ? Quelles sont les limites de corpus pour votre problématique ? (nombre d'auteurs, conditions de création du corpus, situations naturelles/artificielles, qualité/exhaustivité des transcriptions…)

3 Créer un sous-corpus



Cette étape est optionnelle, mais peut être utile avec certains corpus et certaines problématiques. Par exemple, avec le corpus COLAJE, si vous voulez étudier uniquement les paroles des enfants.

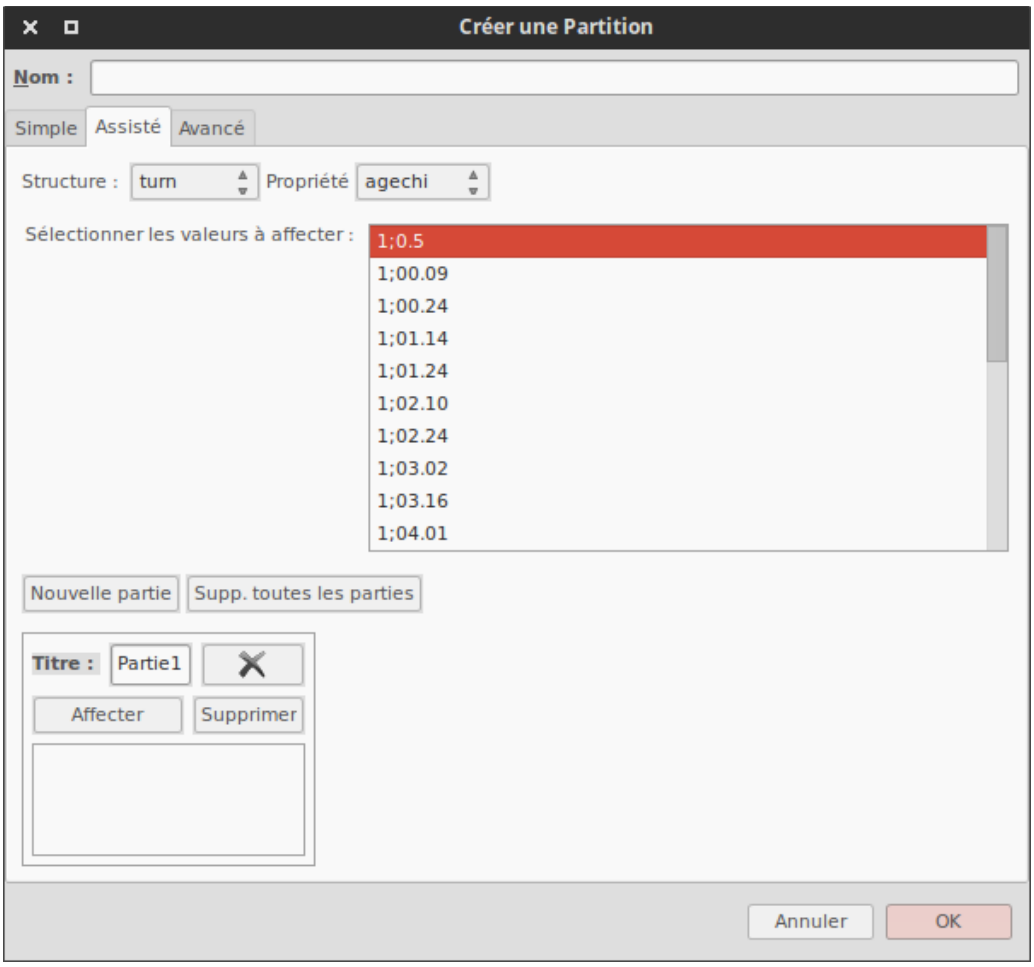
En général, pour créer un sous-corpus, le mode simple fait l'affaire (cf. image ci-contre). Sélectionnez la structure, l'étiquette et la/les valeur(s) qui vois intéressent.

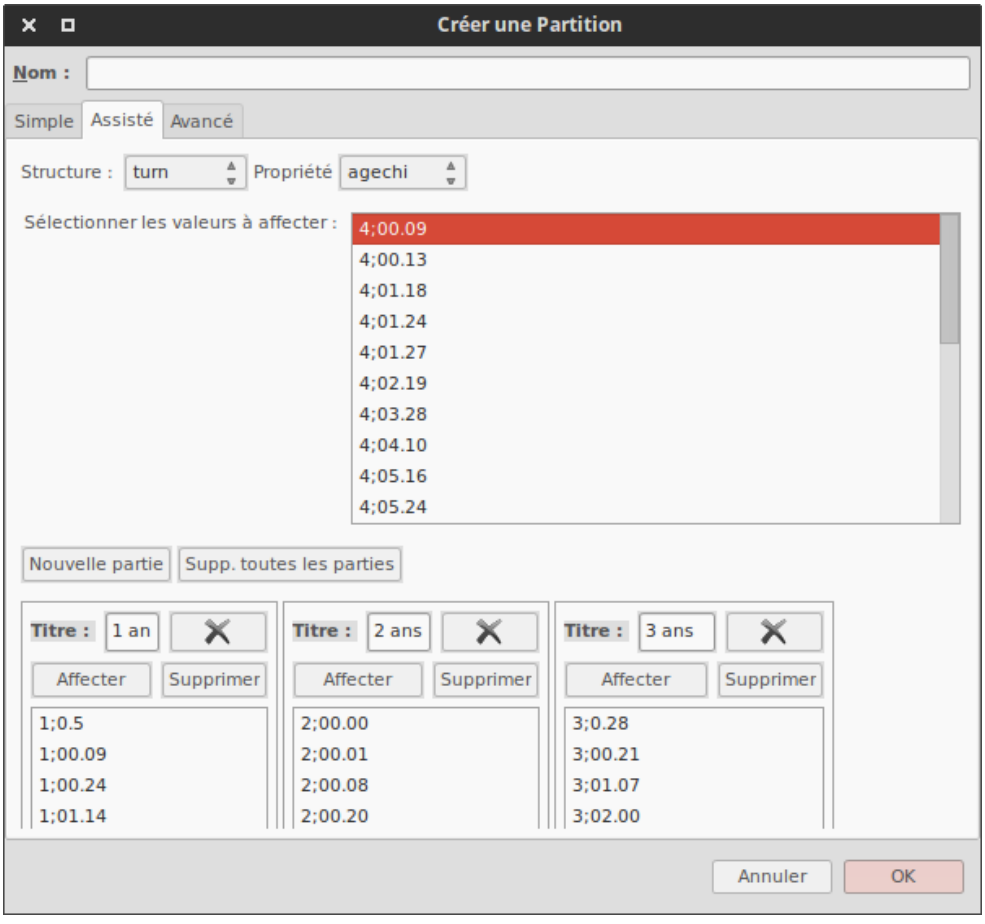
Attention ! La vue « Description » d'un sous-corpus n'indique que le nombre de mots du sous-corpus (qui est donc plus petit que le nombre de mots du corpus principal). Toutes les autres informations correspondent au corpus principal. De même, la vue « Édition » permet de parcourir le corpus principal, *pas* un sous-corpus.

Par contre, toutes les autres vues concernent le sous-corpus. Par conséquent, pour « parcourir » un sous-corpus (et ainsi vérifier qu'il est correct), il n'est pas possible d'utiliser la vue « Édition », mais on peut utiliser la vue « Concordances », en effectuant une concordance sur tous les mots : [word=".*"] (voir ci-dessous).

4 Créer une partition

Cette étape est utile dans la plupart des cas. Il vaut mieux utiliser le mode assisté (capture d'écran n°1, page suivante). Le principe est alors de créer des « boîtes » (capture d'écran n°2, page suivante) regroupant les valeurs qui vous intéressent (par exemple, l'âge de l'enfant en années de 1 à 3 ans).


Créer une Partition
Nom :
Simple
Assisté
Avancé
Structure :
turn
Propriété
agechi
Sélectionner les valeurs à affecter :
1;0.5
1;00.09
1;00.24
1;01.14
1;01.24
1;02.10
1;02.24
1;03.02
1;03.16
1;04.01
Nouvelle partie
Supp. toutes les parties
Titre :
Partie1
Affecter
Supprimer
Annuler
OK


Créer une Partition
Nom :
Simple
Assisté
Avancé
Structure :
turn
Propriété
agechi
Sélectionner les valeurs à affecter :
4;00.09
4;00.13
4;01.18
4;01.24
4;01.27
4;02.19
4;03.28
4;04.10
4;05.16
4;05.24
Nouvelle partie
Supp. toutes les parties
Titre :
1 an
Affecter
Supprimer
1;0.5
1;00.09
1;00.24
1;01.14
Titre :
2 ans
Affecter
Supprimer
2;00.00
2;00.01
2;00.08
2;00.20
Titre :
3 ans
Affecter
Supprimer
3;0.28
3;00.21
3;01.07
3;02.00
Annuler
OK

5 Concordances

Normalement, on ne peut faire des concordances que sur un corpus ou un sous-corpus, pas sur un corpus partitionné.

Mais à partir d'une partition, on peut créer un index sur tous les mots [word=".*"], puis sélectionner tous les mots de cet index (Ctrl+A), puis clic-droit et « Envoyer vers les concordances ».

6 Expressions régulières

N'importe quel mot : **.***

N'importe quel mot qui contient *blanc* : .*blanc.*

Le mot *blanc* exactement (il commence *et* finit par *blanc*) : blanc

N'importe quel mot se terminant par *-âtre* : .*âtre

N'importe quel mot commençant par *anti-* : anti.*

Le mot *blanc*, éventuellement suivi d'un *s* : blancs?

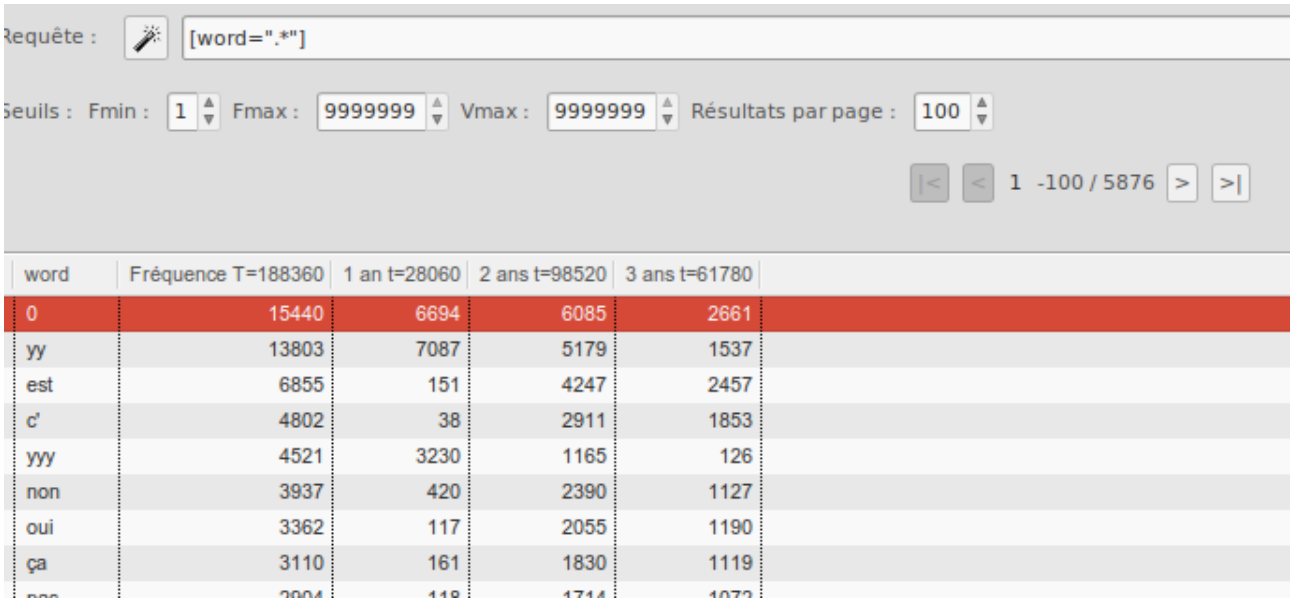
Le mot *blanc*, éventuellement fléchi : blanc**(**s|he|hes**)?** ou bien blanc**(**he**)?**s**?**

… et aussi *blancheur(s)* : blanc**(**he|heur**)**s**?**

7 Index

Les index sont particulièrement intéressants sur les corpus partitionnés. Par exemple, pour étudier l'évolution du lexique au fil du temps.

Attention ! Dans les exemples ci-dessous, la recherche se fait sur tous les mots, mais vous pouvez (et vous devriez !) être plus spécifiques. De même, les résultats des exemples sont toujours affichés en mots, mais souvent, c'est la vue en lemmes, voire parfois en parties du discours, qui sera la plus utile.



word	Fréquence T=188360	1 an t=28060	2 ans t=98520	3 ans t=61780
0	15440	6694	6085	2661
yy	13803	7087	5179	1537
est	6855	151	4247	2457
c'	4802	38	2911	1853
yyy	4521	3230	1165	126
non	3937	420	2390	1127
oui	3362	117	2055	1190
ça	3110	161	1830	1119

Attention, les fréquences indiquées dans cette vue (par exemple, 6694, 6085 et 2661) sont des fréquences absolues. Il faut **toujours** diviser par le nombre de mots dans la partition (dans l'exemple ci-dessus : 6694/28060, 6085/98520, 2661/61780).

8 Spécificités

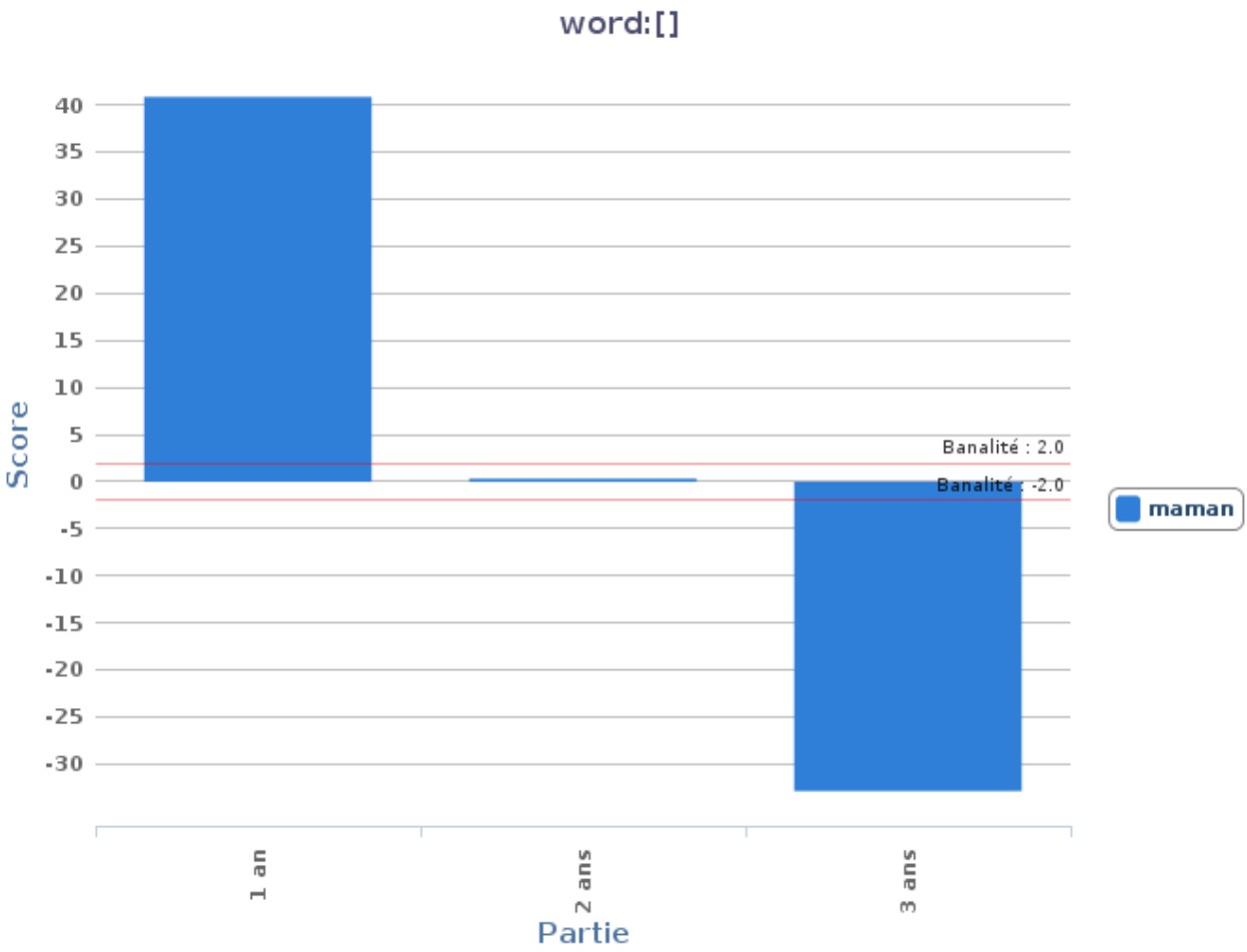
Si vous voulez savoir quels sont les mots (ou lemmes, ou parties du discours, etc.) les plus *spécifiques* pour une partition donnée (c'est à dire, qui apparaissent significativement plus souvent dans une partie que dans les autres), vous pouvez afficher la vue « Spécificités » d'un corpus partitionné. Cette vue se base sur les *spécificités de Lafon*, un calcul assez complexe (que TXM effectue pour nous).

L'exemple ci-dessous, trié sur le score de spécificité de la première année, indique donc que *yy*, *0*, *yyy*, *lait* et *maman* sont les mots les plus spécifiques de la première année, avec des scores de respectivement, 1000, 1000, 1000, 109 et 41. Au contraire, ces mots sont sous-spécifiques de la troisième année, avec des scores de -1000, -1000, -1000, -23 et -33.

Attention ! Les scores compris entre -2 et 2 sont considérés comme trop faibles pour être significatifs. Seuls les scores supérieurs à 2 ou inférieurs à -2 sont intéressants. Par contre, ces scores sont déjà des scores relatifs, il n'y a donc pas besoin de les diviser ; ils sont utilisables tels quels.

Unités	Fréquence T 186414	1 an t=27858	▼ score	2 ans t=97529	score	3 ans t=61027	score
yy	13803	7087	1 000,0	5179	-287,4	1537	-1 000,0
0	15440	6694	1 000,0	6085	-246,7	2661	-1 000,0
yyy	4521	3230	1 000,0	1165	-295,4	126	-1 000,0
lait	204	173	108,5	22	-35,5	9	-22,8
maman	1281	381	40,9	672	0,3	228	-32,9
ə	329	152	40,7	148	-2,3	29	-24,2
miam	91	68	36,4	23	-6,9	0	-15,7
assis	117	69	27,0	44	-3,0	4	-14,5
xx	2379	553	26,4	1045	-16,0	781	0,3
chat	156	78	24,1	56	-4,6	22	-7,1
boum	157	74	20,9	60	-3,6	23	-6,7
clé	58	40	19,6	11	-6,8	7	-3,6

On peut ensuite « Calculer le diagramme en bâtons » d'une ligne, par exemple « maman » :


word:[]
40
35
30
25
20
15
10
5
0
-5
-10
-15
-20
-25
-30
Score
Banalité : 2.0
Banalité : -2.0
maman
1 an
2 ans
3 ans
Partie