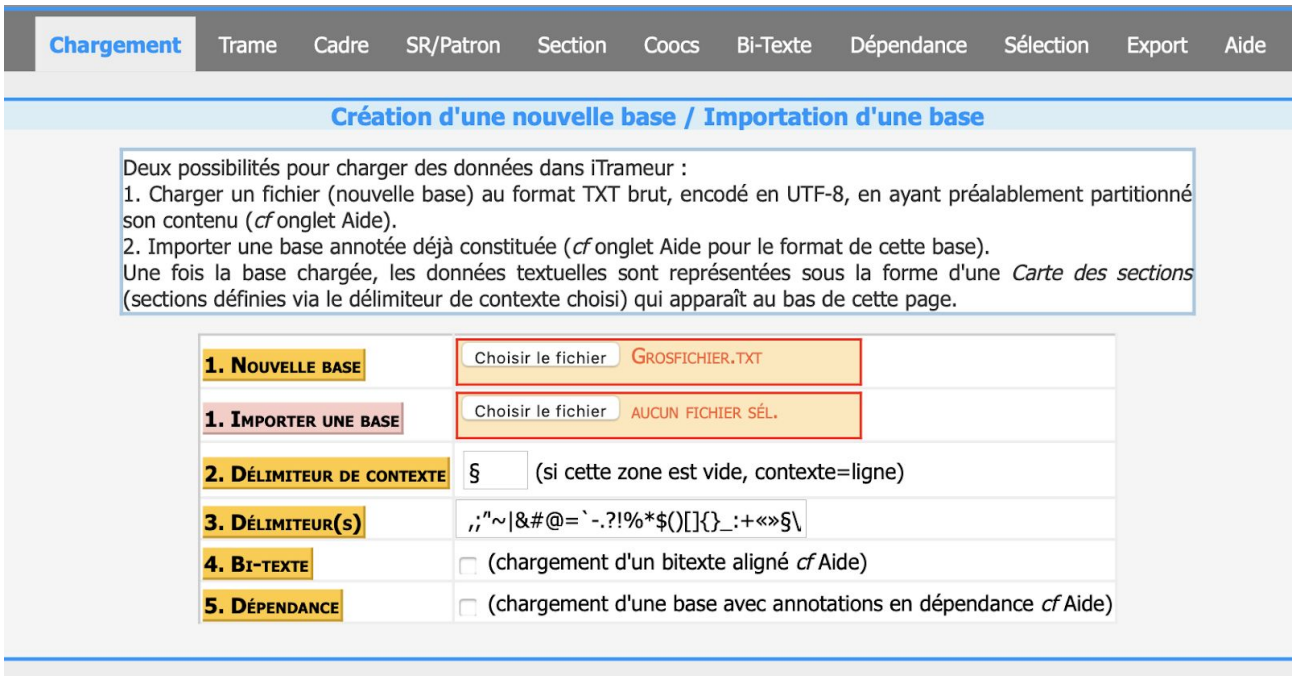
Le programme d’analyse par iTrameur

Chargement du corpus



figure(1)

Pour charger le corpus dans iTrameur, il nous faudrait de choisir un fichier en cliquant le premier bouton “choisir le fichier”. Comme on avait préparé ce corpus en amont nommé “Grosfichier” en combinant tous les textes après la segmentation en cas de textes chinois et aussi en ajoutant les balises. Donc, il nous suffit de choisir ce fichier en cliquant dessus.

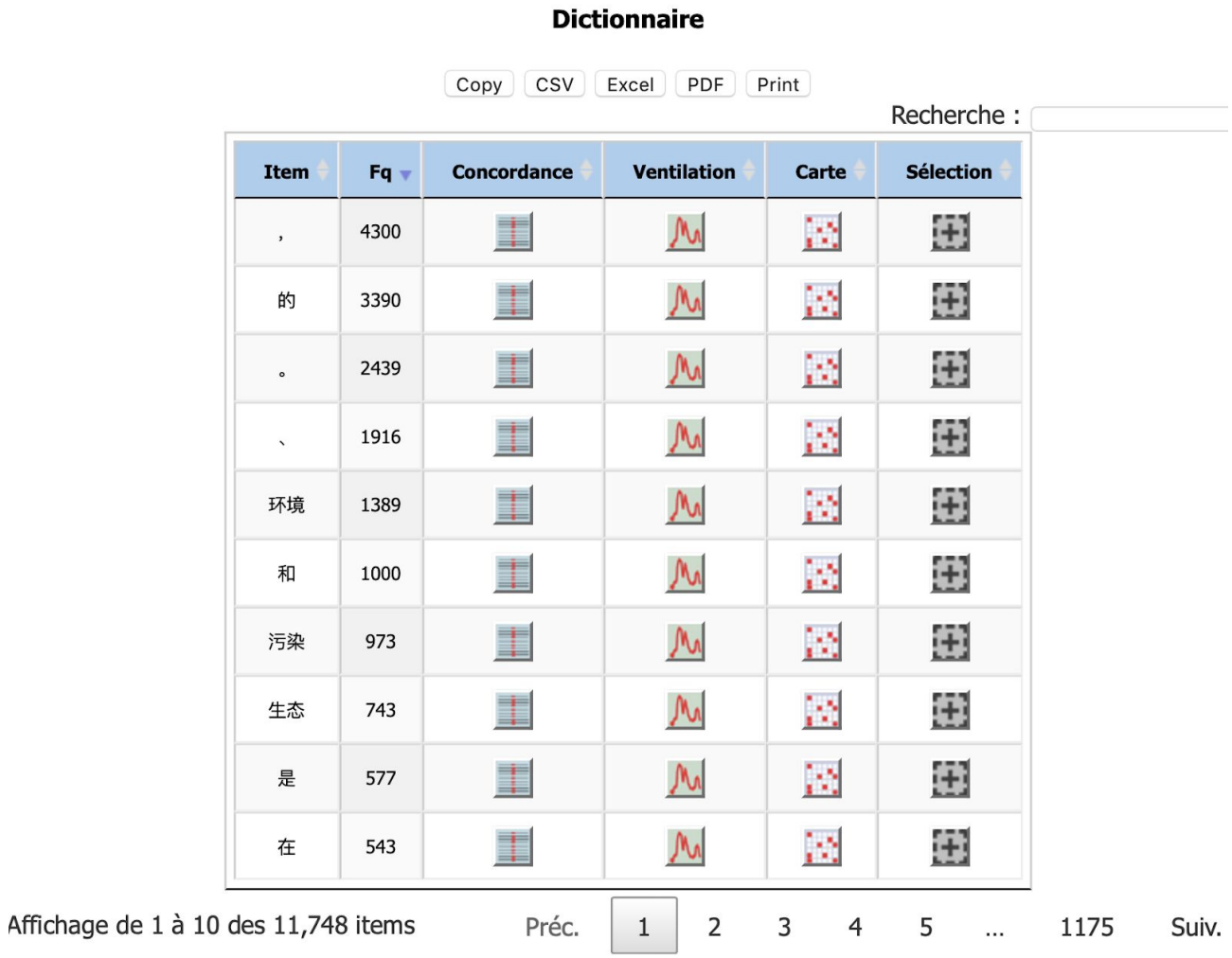
Cadre

Tout d'abord, nous vérifions notre corpus structuré avec les paires de balises ouvrante et fermantes pour chacune partie. Le bonton "cadre" permet de donner une représentation graphique de l'organisation des parties.

figure(2)

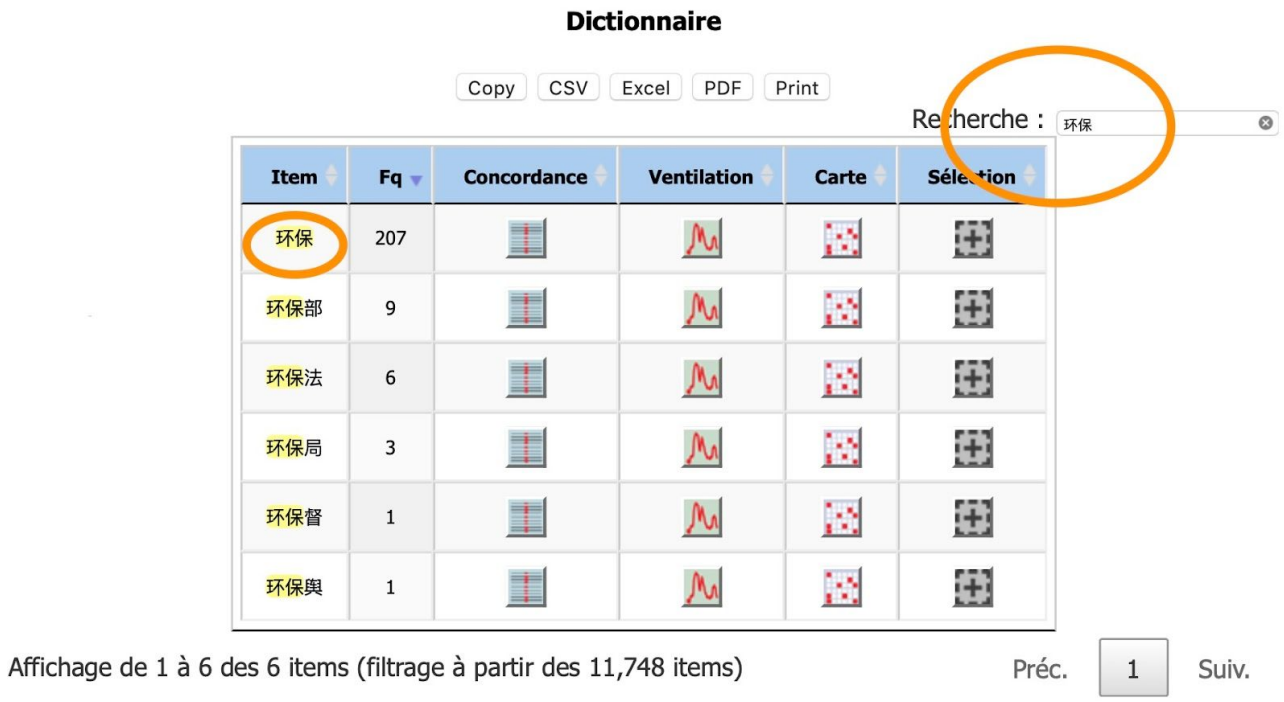
Dictionnaire

Cette fonction nous représente le contenu du dictionnaire du corpus, et aussi les fonctions permettant de montrer les fréquences, la ventilation, la concordance, la carte et la sélection qui sont tous dans la même ligne du tableau.



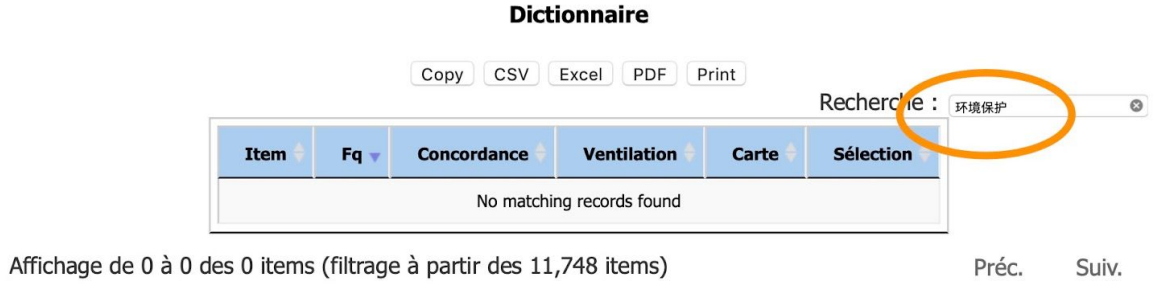
figure(3)

On voit que le mot “**的**” apparaît dans le corpus 3390 fois qui est le plus nombreux parmis les 11,784 mots. Ensuite, le groupe de mot “**环境**” qui signifie “l’environnement ” est le seconde le plus fréquent dans ce corpus. Alors, notre thématique est “la protection environnementale”. Donc, en l’occurrence, nous saisissons notre groupe de mots clés en mandarin “**环保**” qui veut dire “la protection environnementale”, voilà ce que nous obtenons:



figure(4)

On voit que ce cible “环保” (la protection environnementale) apparaît dans ce corpus 207 fois. Vu qu’il y a deux terms en mandarin qui désignent la même chose “la protection environnementale”, il vaut mieux de chercher et observer avec le deuxième term: “环境保护”. Le premier term “环保” est l’abréviation du deuxième “环境保护”.



figure(5)

Comme on a dit que l’abréviation “环保” (la protection environnementale) est employée 207 fois dans notre corpus, alors le term en toute lettre qui dit la même chose “环境保护” est apparu aucune fois. En fait, cela est à cause du gros fichier qu’on avait préparé en amont pour le corpus. En raison de la particularité du mandarin, nous somme obligé de faire la segmentation du texte en chinois dans le but de faire l’analyse. Alors, le processus de segmentation a fait diviser le groupe de mots “protection environnementale” en “protection” et “environnemental”, en l’occurrence, “环境保护”

qui est divisé en “环境” et “保护”. C’est la raison que la fréquence de “protection environnementale” (“环境保护”) est zéro. Pour vérifier si ce terme “environnement” est employé ensemble avec “protection”, on peut se servir de la fonction “Coocs”.

Cooccurrences

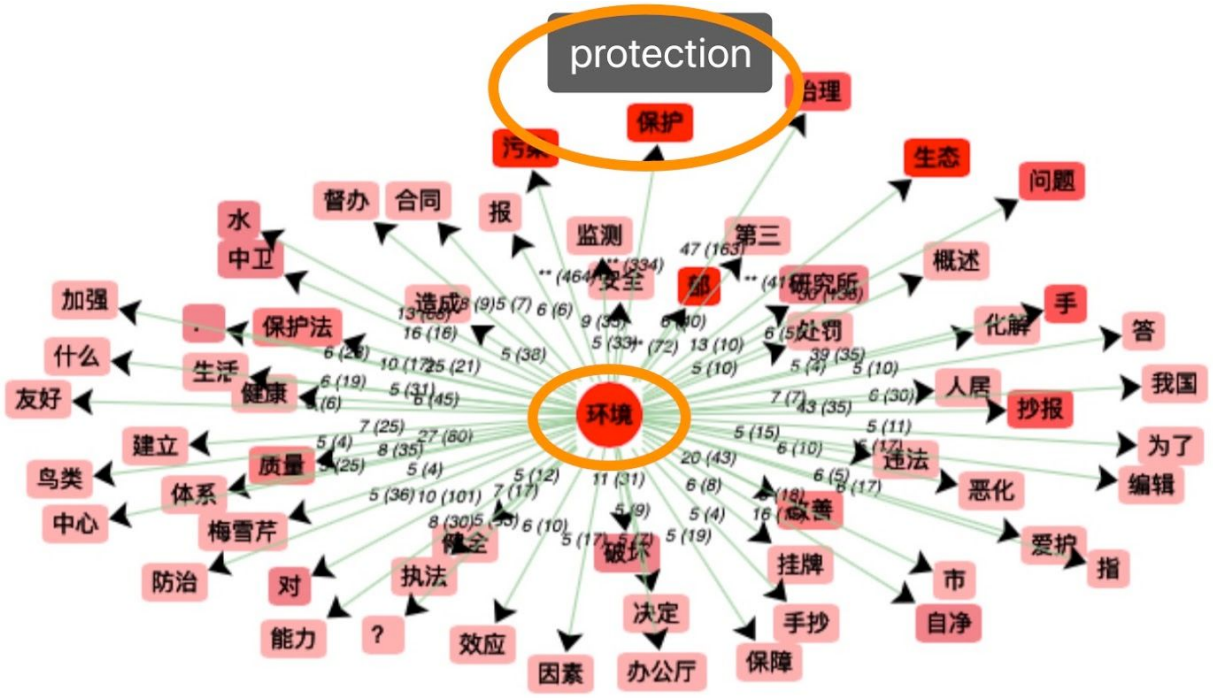
Cette fonction permet de calculer les cooccurrents du pôle source avec les contextes gauches et droites.

Cooc	FqCooc	CoFreq	IndSP
保护	423	334	**
生态	743	411	**
污染	973	464	**
部	65	72	**
治理	322	163	47
抄报	25	35	43
手	28	35	39
问题	320	136	30
质量	147	80	27
保护法	15	21	25
改善	63	43	20
自净	12	15	16

Couleur Noeud-Cooc : IndicSp<**5**<IndicSp<**10**<IndicSp<**20**<IndicSp<**30**<IndicSp<**50**<IndicSp
Arc Label : **POLE**–IndiceSpécif(Co-Freq)→**COOC**

figure(6)

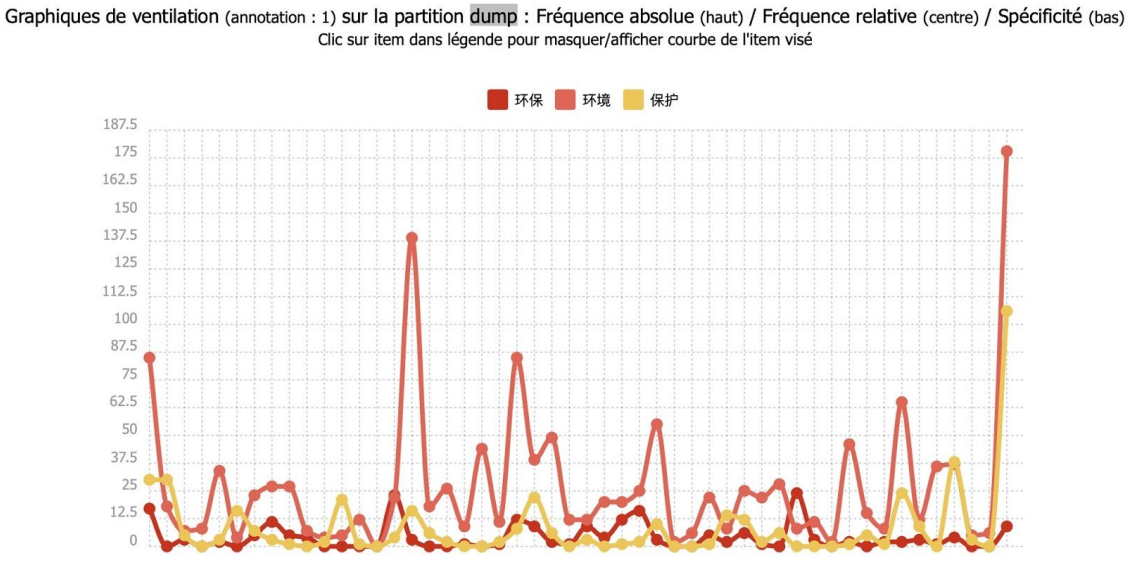
On voit que quand on saisit “environnement” dans le pôle source, ensuite on tape le bouton “Coocs”. Le terme le plus fréquent en cooccurrence avec “environnement” est “protection”.



figure(7)

Cadre - Ventilation

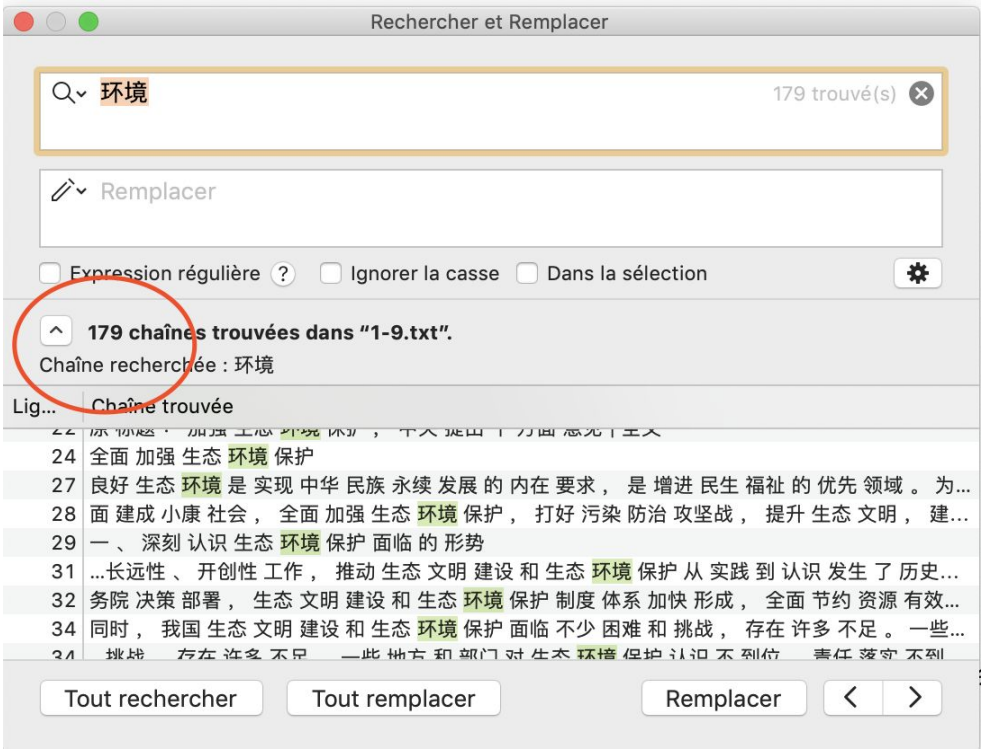
1, Fréquence absolue

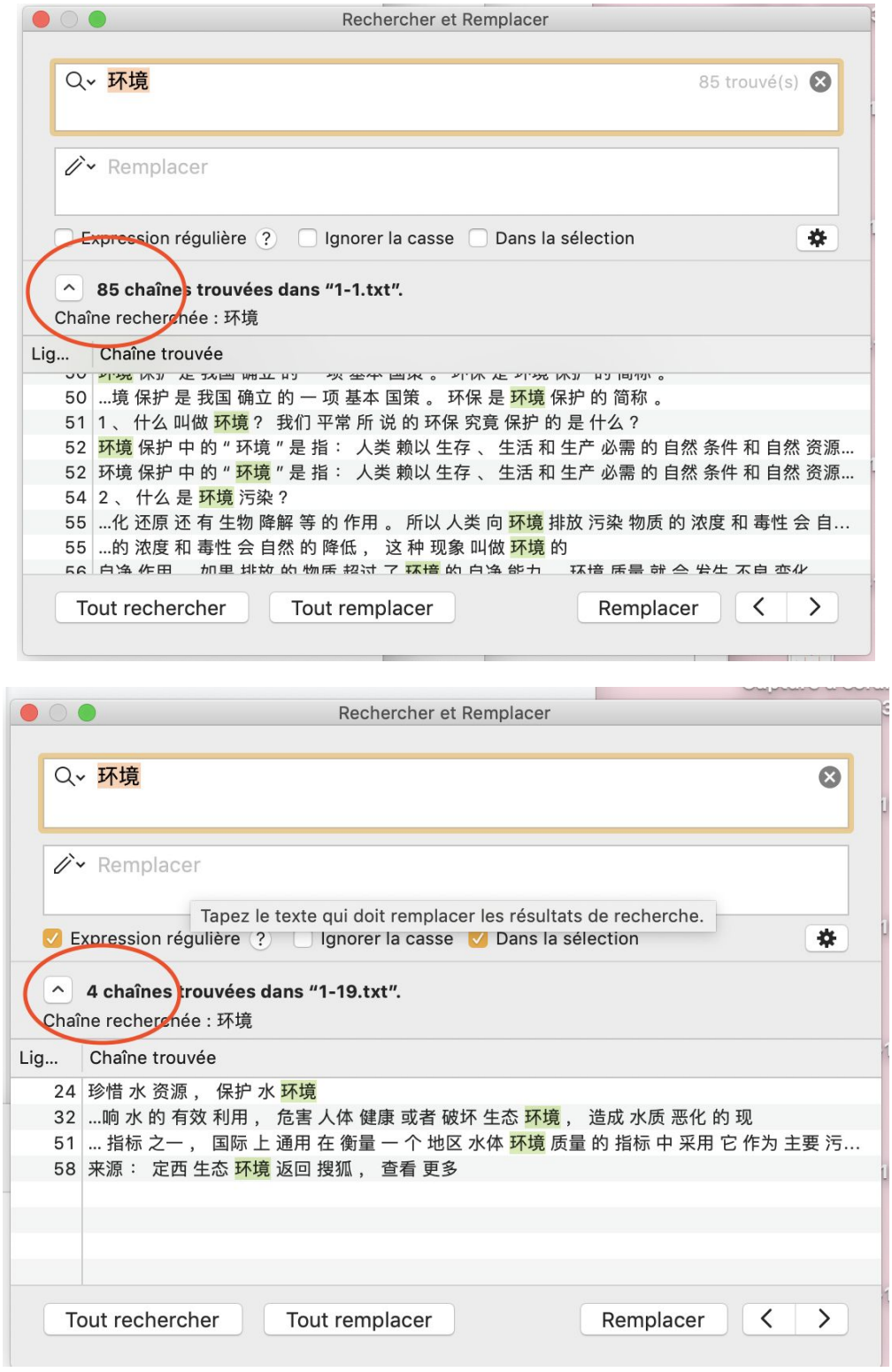


figure(8)

Le bouton de la ventilation nous permet d'apercevoir la répartition du mot en partie. La ventilation nous représente par les graphiques sur la fréquence absolue qui compte le nombre de fois que le pôle apparaît dans chaque partie, la fréquence relative qui illustre le nombre de fois que le pôle apparaît au total des mots dans la partie. En l'occurrence, le cible clé "环境" (environnement) apparaît dans des parties du corpus de façon instable par l'observation ces graphiques de ventilation. Par exemple, le cible clé "环境" représente 85 fois dans la première partie du corpus correspondant au texte "1-1.txt" par le graphique de fréquence absolue. Egalement, il existe des partie qui montre très peu de fréquence de termes, par exemple, le texte "1-19.txt" dont "环境" (environnement) apparaît juste 4 fois dans la partie. Alors on voit qu'il y a un pic qui représente le plus nombreu de fréquence sur le dernier text correspondant au texte numéro neuf "1-9.txt". Dans

ce texte, le cible clé “环境” représente 179 fois comme la fréquence absolue. Je vérifie les valeurs des fréquences absolues respectivement et je les montre en-dessous(figure(9)):

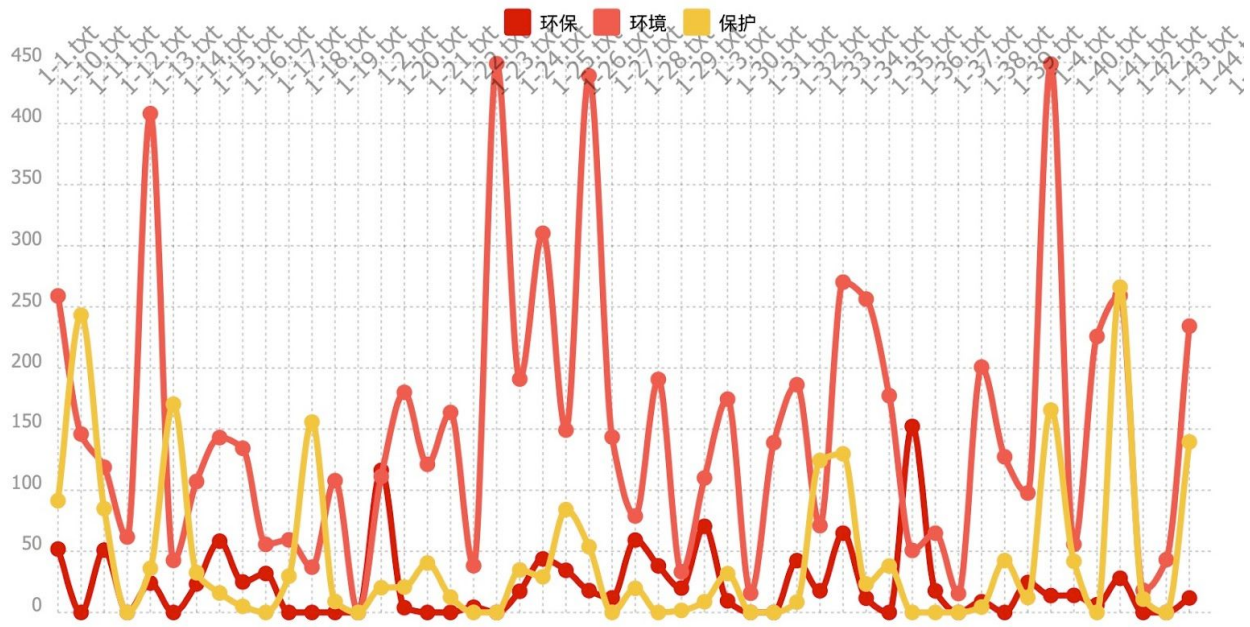




figure(9)

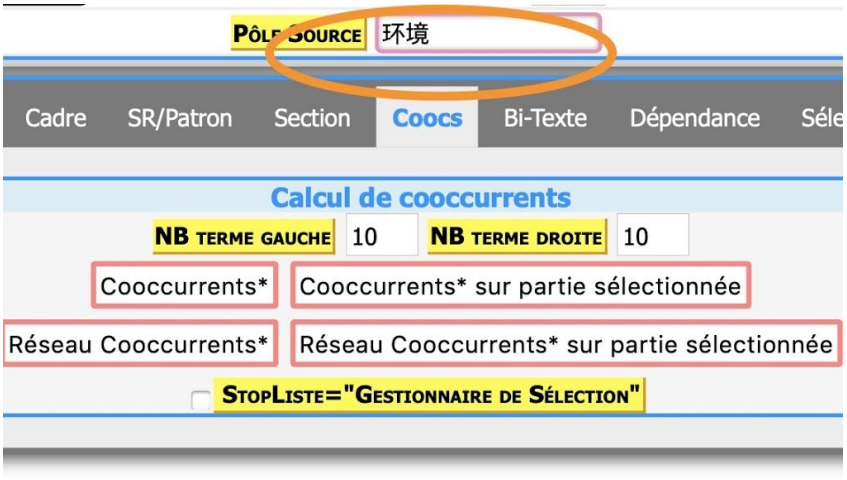
Enfin, même si ce sont tous les textes ayant le même thématique “la protection environnementale”, ils ont différents fréquences absolues du terme clé. On peut déduire que le nombre de fois d’apparition du terme clé n’est pas un facteur très important pour un thématique d’un texte. Ici, la thématique reste la même, alors le nombre de fois d’apparition du clé varie entre 2 fois et 179 fois.

2, Fréquence relative



figure(10)

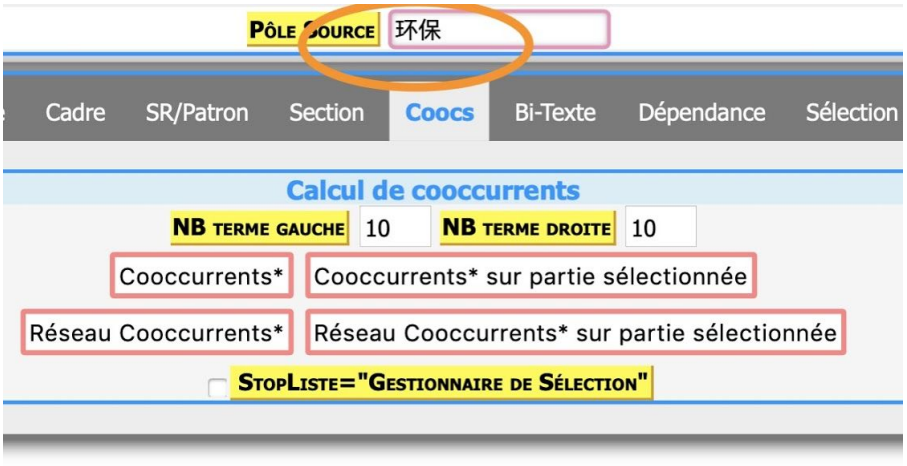
En ce qui concerne la fréquence relative, on voit que les fréquence du cible “环境”（l’environnement）est relativement beaucoup plus hautes que le cible “环保” (la protection environnementale) et “环境保护” (la protection environnementale). Ayant recours à la fonction de “Coocs”, on peut chercher la raison de cette haute fréquence relative de “环境”.



(clic sur mot : contextes)

Cooc	FqCooc	CoFreq	IndSP
保护	423	334	**
生态	743	311	**
污染	973	464	**
部	65	72	**
治理	322	163	47
抄报	25	35	43
手	28	35	39
问题	320	136	30
质量	147	80	27
保护法	15	21	25
改善	63	43	20

figure(11)



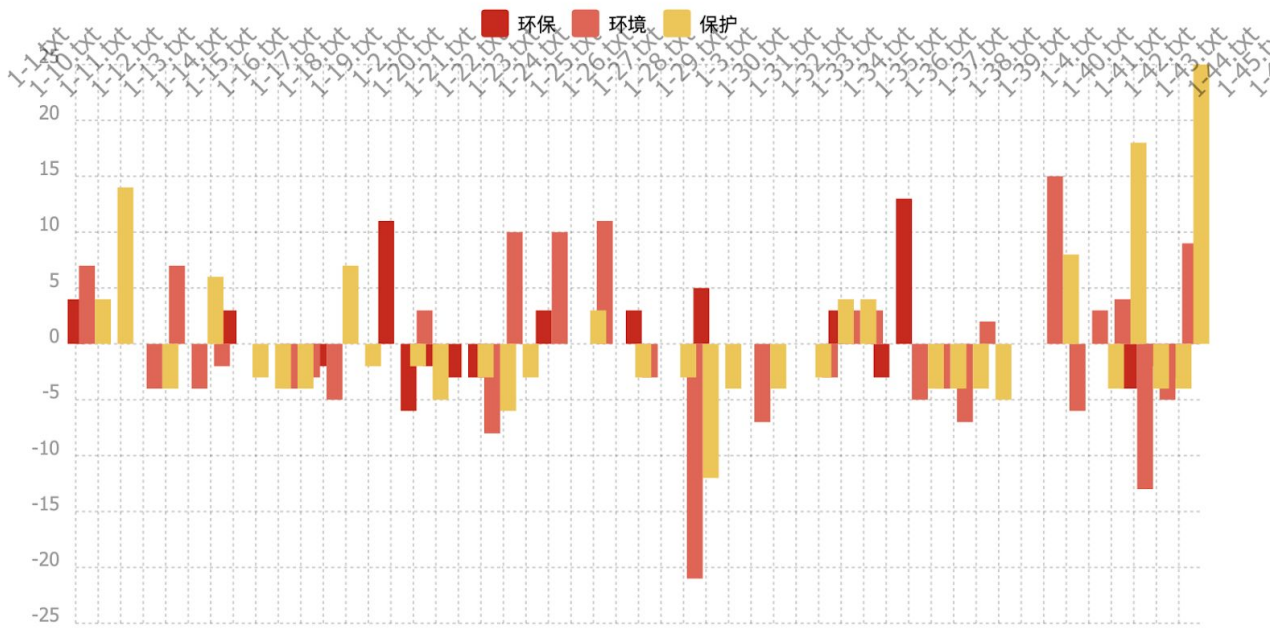
(clic sur mot : contextes)

Cooc	FqCooc	CoFreq	IndSP
产业	93	29	25
舆情	23	17	24
举报	82	23	19
督察	37	11	10
装备	30	10	10
节能	40	10	9
轮	13	7	9
扫描	11	6	8
中央	53	11	8
垂改	6	5	8
100	15	6	7

figure(12)

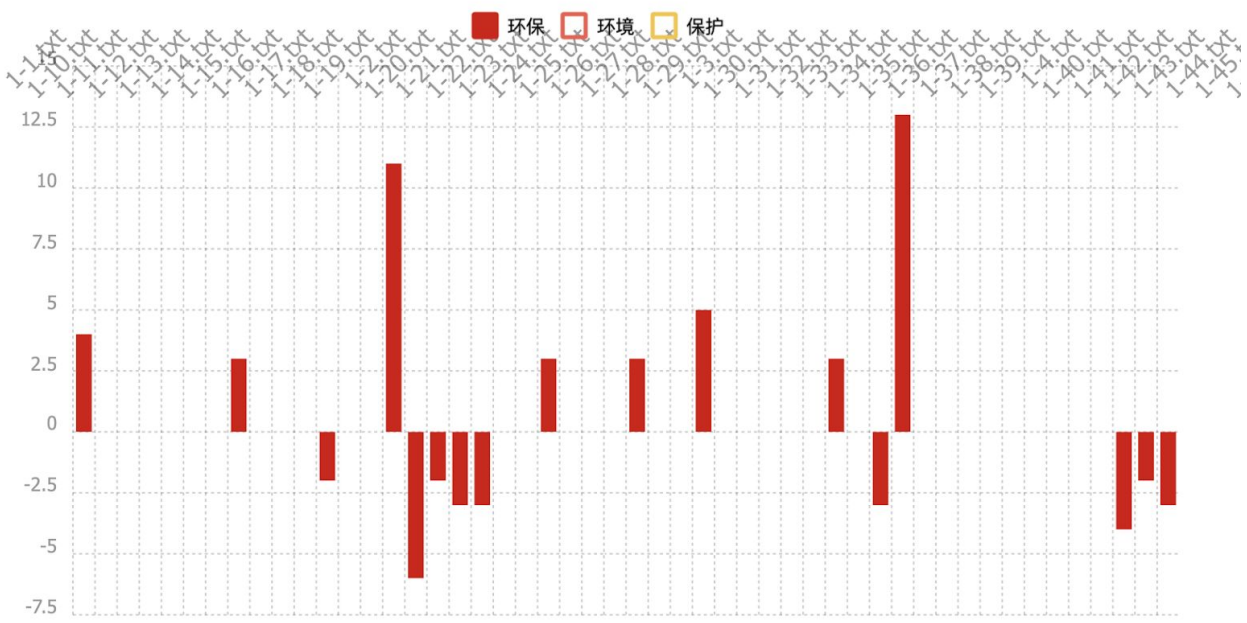
On voit que le cible “环境” est co-présenté très souvent avec d’autres termes. Par exemple, “la pollution de l'environnement”, “l’environnement écologique” etc. La fréquence de cooccurrence du terme “环境”(l’environnement) avec le terme “l’écologie” est 743 et celle avec le terme “la pollution” est 973. Alors on peut aussi observer les cooccurrences du cible “环保”(la protection environnementale), ce cible co-présente avec les termes comme “l’industrie”, “la situation” etc. Mais les fréquence de cooccurrence avec ces termes ne sont pas autant hautes que celles entre le cible “环境” et d’autres termes. C’est pour ça que les fréquences du cible “环境” （l’environnement）est relativement beaucoup plus hautes que les deux autres termes.

3, Spécificité



figure(13)

En ce qui concerne la spécificité, cette fonction sert à illustrer les comparaisons entre les parties. Avec la spécificité, nous nous attendons à avoir le nombre de fois équivalent que le pôle source apparaît pour chaque partie. S'il est sur-représenté, cela voudrait dire que le cible est particulièrement présent dans une partie par rapport aux autres parties; Par contre, s'il est en sous-représentation, celui-ci est particulièrement absent dans une partie comparé aux autres. En l'occurrence, on voit que ces trois termes apparaissent dans les textes de manière irrégulière. Si on veut juste observer les distributions de cible "环保" (la protection environnementale), on décoche les autres deux cases. Voilà ce que nous obtenons:



figure(14)

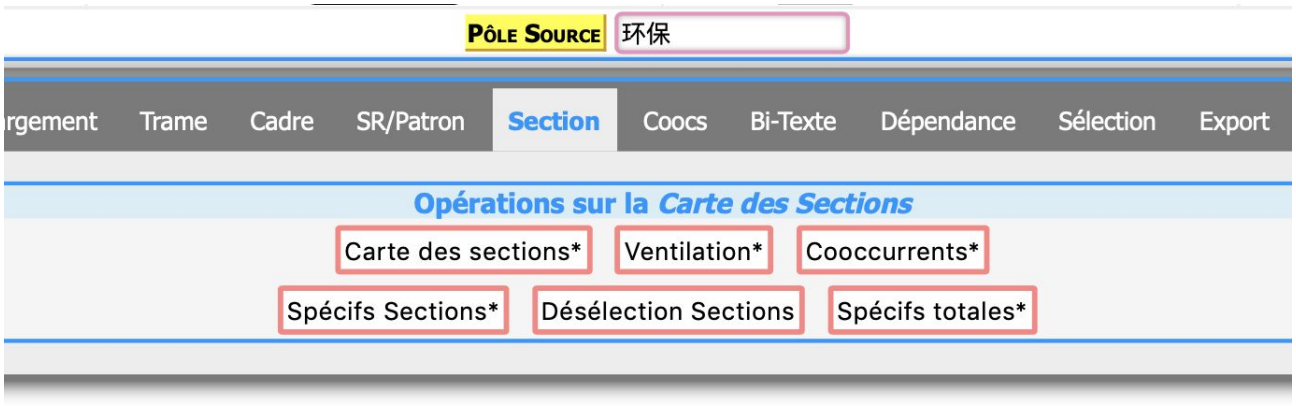
On constate que dans le texte numéro 15 et le texte numéro 38, le cible “环保” (la protection environnementale) sont particulièrement présent, donc sur-représenté dans ces deux parties par rapport aux autres. Egalement, il existe des parties dont le cible source est sous-représenté.

Barycentre temporel / Coefficient Von Neumann

Vu que ce corpus n’est pas récolté et structuré de façon chronologique, on ne va pas appliquer la fonction de Barycentre temporel et Coefficient Von Neumann. Parce que le bouton BT sert à constater à partir de quel moment ou une zone temporel, le cible est présent de la même quantité à gauche et à droite sur la ligne chronologique.

Sections -- Cooccurrents

On va regarder l’opération “Cooccurrents” sur la Carte des Sections.



(clic sur mot : contextes)

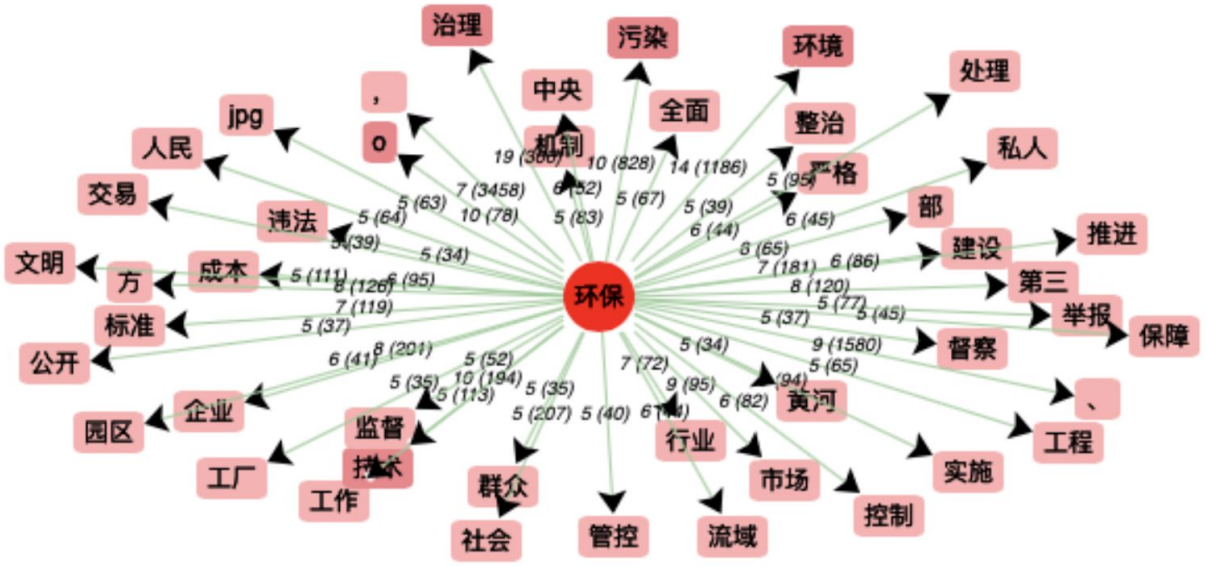
Cooc	FqCooc	CoFreq	IndSP
治理	322	306	19
环境	1389	1186	14
污染	973	828	10
技术	208	194	10
o	78	78	10
、	1916	1580	9
市场	97	95	9
第三	127	120	8
企业	220	201	8
部	65	65	8
，	4300	3458	7

Couleur Noeud-Cooc : IndicSp<**5**<IndicSp<**10**<IndicSp<**20**<IndicSp<**30**<IndicSp<**50**<IndicSp

Arc Label : **POLE**–IndiceSpécif(Co-Freq)→**COOC**

figure(15)

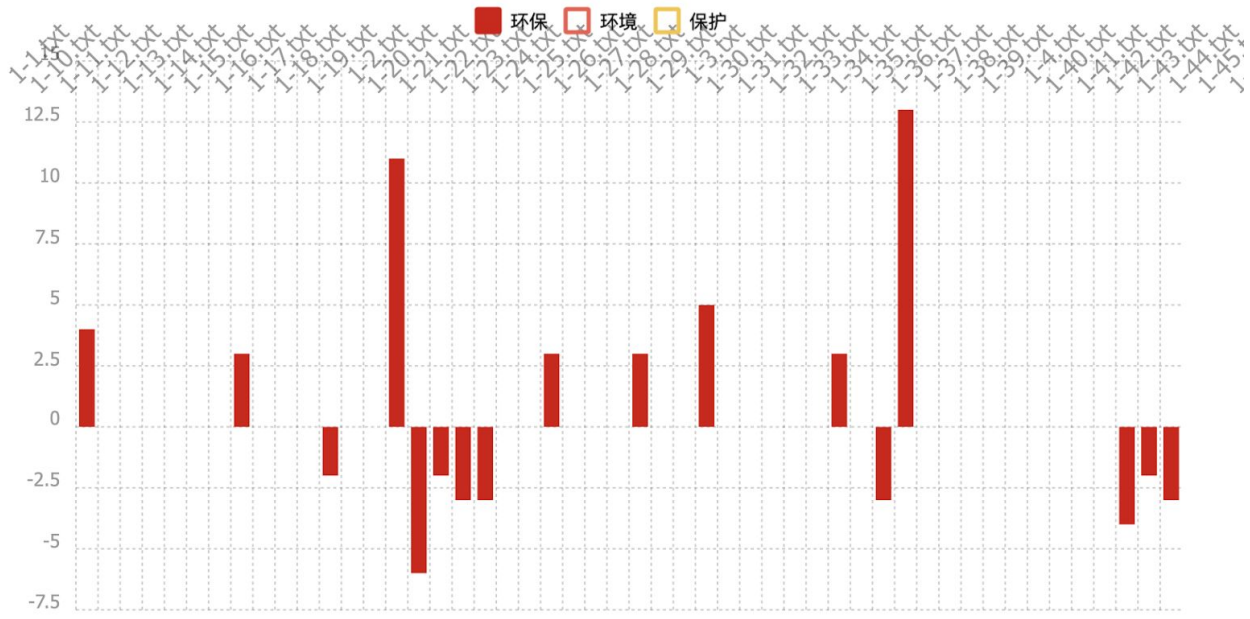
En saisissant le cible “环保” (la version raccourcie de “la protection environnementale”), on observe qu’il est co-présenté avec divers groupes de mots. Par exemple, “la technique”, “l’aménagement”, “la pollution”, “le marché”, “l’entreprise” etc. Grâce à cette fonction, nous obtenons deux représentations de cooccurrences du cible source. Premièrement, c’est le tableau contenant quatre colonnes qui sont le mot cooccurrent, la fréquence, la co-fréquence de deux mots ou deux groupes de mots et l’indice de spécificité. Deuxièmement, comme l’image montrée en-dessous, il illustre en une seule représentation tous les cooccurrents du cible source avec la co-fréquence et l’indice de spécificité.



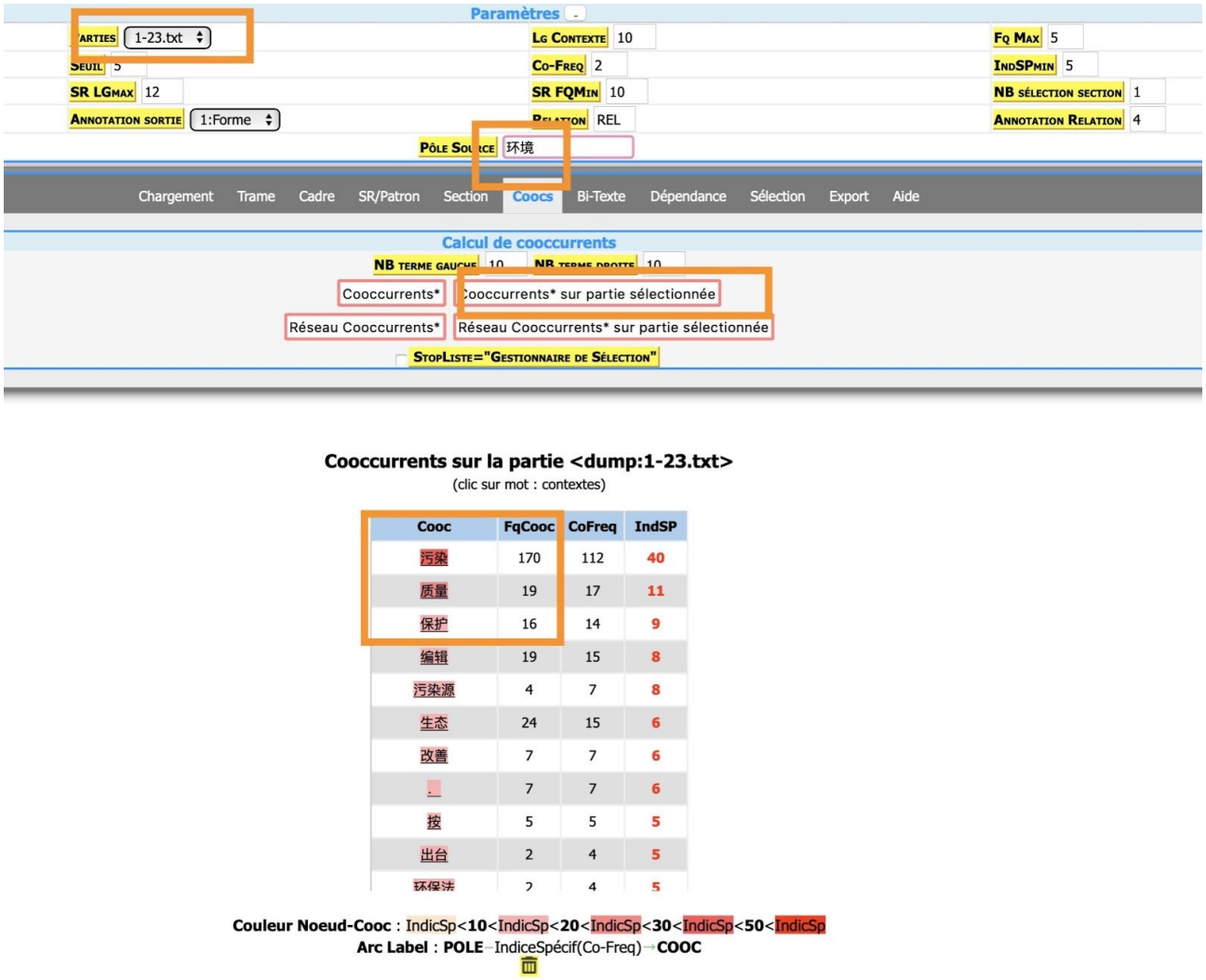
figure(16)

En conclusion, cet outil d'analyse textométrique de données iTrameur nous aide à explorer le corpus de façon intuitive, générale et précise. Grâce au outil iTrameur dont l'objectif est définir et compter des unités dans les partie de corpus et contraster entre ces parties afin de découvrir éventuellement des phénomènes textuelles. Par exemple, avec la fonction "Dictionnaire", nous pouvons regarder tous les mots avec leurs fréquences et d'autres valeurs textométriques du corpus. En l'occurrence, notre cible clé "环保" (la protection environnementale) apparaît 207 fois. En plus, il est co-présenté avec divers groupes des mots, comme "la technique", "l'aménagement". Par le bouton "Ventilation", on observe toutes les distributions du cible source dans chaque partie du corpus avec la fréquence absolue, la fréquence relative et la spécificité. En général, la répartition du cible "环保" (la protection environnementale) n'est pas équilibré dans toutes les parties du corpus. On peut considérer que ça fait partie de la rédaction textuelle en chinois: les fréquences d'un terme n'est pas forcément liés avec le sujet d'un texte. Par exemple, le sujet "la protection environnementale" peut être sur-représenté ou sous-représenté dans un texte, alors cela n'empêche pas qu'il est toujours le sujet de textes. Dans certains textes dont le cible "环保" (la protection environnementale) est sous-représenté, par exemple, par la figure dessous (figure 17), on constate que notre groupe de mots dans le texte 22 (1-22.txt) est sur-représenté et le texte (1-23.txt) dont le cible est sous-représenté. Pour découvrir ce qui se passe dans le texte 23 dont le sujet est "la protection environnementale" et en même temps, ce sujet n'apparaît pas autant de fois que les autres textes, on peut se servir de bouton 'Parties' dans le Paramètre et on choisit le texte '1-23.txt', ensuite, on tape le mot 'l'environnement' (环境) dans la case 'Pôle Source' pour regarder ces cooccurrences en cliquant le bouton 'Cooccurrnents sur partie sélectionnée' dans 'Calcul de cooccurrents' (figure 18). Donc, on constate que dans le texte 1-23, le mot le plus occurrent avec 'l'environnement' ou 'environnemental(e)' n'est pas 'la protection', mais c'est 'la pollution/contamination' qui apparaît 170 fois et 'la protection' représente juste 16 fois dans le texte. En fait, le contexte est autour de la pollution environnementale, alors le sujet est la protection environnementale. Ce genre de phénomène peut expliquer la sous-représentation du terme quand il

est sujet. Il existe aussi une autre explication qui est la vitalité et l'art es mots. On peut toujours trouver des termes similaires ou des synonymes du sujet au lieu d'utiliser le même terme dans tout le contexte. Cela peut aussi expliquer la sous-représentation du sujet dans un texte.



figure(17)



Cooccurrents sur la partie <dump:1-23.txt>

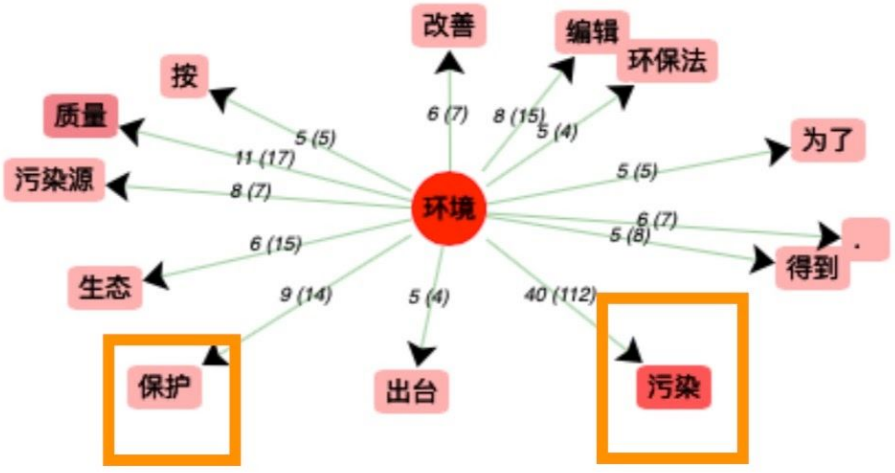
(clic sur mot : contextes)

Cooc	FqCooc	CoFreq	IndSP
污染	170	112	40
质量	19	17	11
保护	16	14	9
编辑	19	15	8
污染源	4	7	8
生态	24	15	6
改善	7	7	6
。	7	7	6
按	5	5	5
出台	2	4	5
环保法	2	4	5

Couleur Noeud-Cooc : IndicSp<**10**<IndicSp<**20**<IndicSp<**30**<IndicSp<**50**<IndicSp

Arc Label : **POLE**–IndiceSpécif(Co-Freq)→**COOC**

figure(18)



figure(19)