



Institut européen

des métiers de la **traduction** | IEMT

Université de Strasbourg

# Web, corpus, traduction : exploitations AntConc

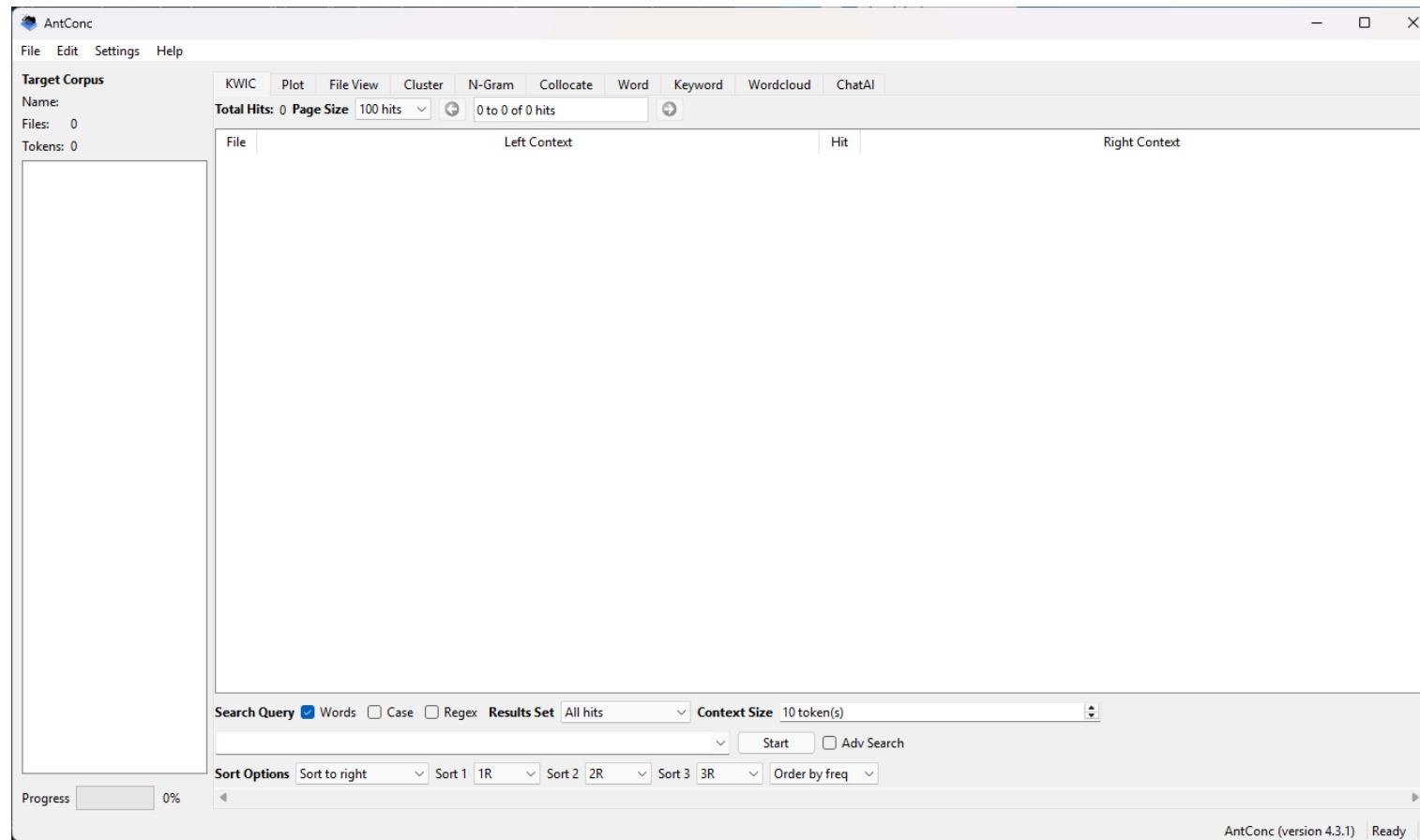
Enzo Doyen

2025 - M1

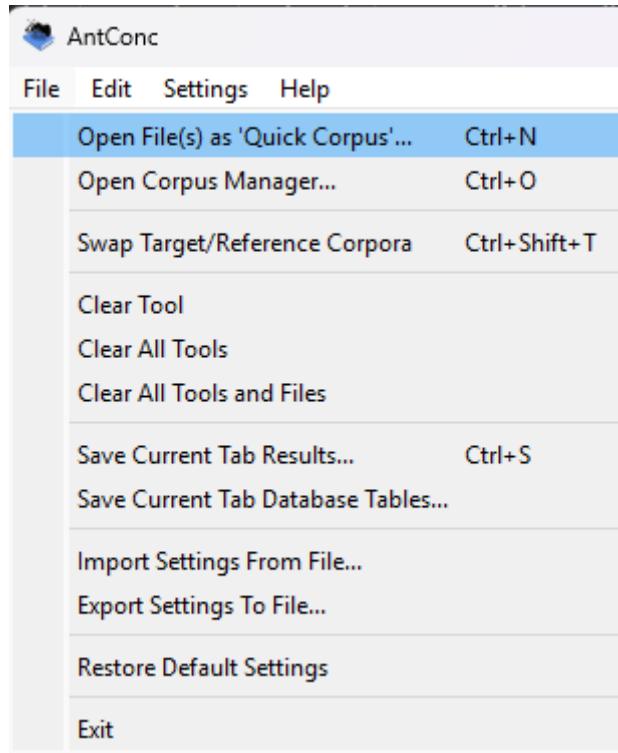
# Fonctionnalités principales d'AntConc

- *KWIC (Keyword in Context)* : permet de visualiser les occurrences d'un terme dans son contexte ;
- *File View* : affichage des fichiers chargés ;
- *Cluster* : regroupement des termes en fonction de leur proximité dans le contexte ;
- *N-Gram* : visualisation des séquences de mots (n-grammes) ;
- *Collocate* : visualisation des collocats d'un terme donné.

# Fenêtre principale



# Sélection d'un corpus



# KWIC (Keyword in Context)

AntConc

File Edit Settings Help

Target Corpus  
Name: temp  
Files: 1  
Tokens: 151362  
FR\_AssembleeNationale.txt

KWIC Plot File View Cluster N-Gram Collocate Word Keyword Wordcloud ChatAI

Total Hits: 138 Page Size 100 hits 1 to 100 of 138 hits

	File	Left Context	Hit	Right Context
1	FR_Assemblee...	fait [M. Marc Fesneau] L'évolution du budget de l'	Assemblée	ationale et du Sénat devrait permettre d'apaiser les
2	FR_Assemblee...	commissions permanentes des lois et des affaires sociales de l'	Assemblée	ionale et du Sénat, quant à la mise en
3	FR_Assemblee...	Parlement et des délégations aux droits des femmes de l'	Assemblée	nationale et du Sénat, nous le devons à la
4	FR_Assemblee...	conforme aux décisions prises par la commission commune à l'	Assemblée	nationale et au Sénat dont c'est la charge.
5	FR_Assemblee...	le moins, qu'un débat démocratique ait lieu à l'	Assemblée	nationale et au Sénat avant qu'on le déclenche.
6	FR_Assemblee...	M. André Schneider. Je constate que le Bureau de l'	Assemblée	nationale est constitué. Je le réunirai mercredi 6 octobre à
7	FR_Assemblee...	à l'Assemblée nationale. Une des deux déposées à l'	Assemblée	nationale est transpartisane et je l'ai sous les
8	FR_Assemblee...	à-dire d'une pratique morale de la politique, l'	Assemblée	nationale a étendu la compétence de son déontologue aux
9	FR_Assemblee...	hui pour examiner le tout dernier texte présenté devant l'	Assemblée	nationale au cours de cette quatorzième législature. Il s'
10	FR_Assemblee...	ont pris l'habitude de regarder les débats à l'	Assemblée	nationale c'est même devenu, le mercredi après-midi,
11	FR_Assemblee...	au travers de l'islamophobie et du nationalisme. À l'	Assemblée	nationale comme au Parlement européen, le Rassemblement national s'
12	FR_Assemblee...	a dit. M. Ruffin vous parle du contrôle par l'	Assemblée	nationale de l'évolution de la situation et des
13	FR_Assemblee...	vidéos s'exposeraient à des sanctions. La présidente de l'	Assemblée	nationale l'a rappelé à plusieurs reprises. Je souhaite
14	FR_Assemblee...	tenue des débats. Je souhaite que la présidence de l'	Assemblée	nationale lance une enquête, afin de savoir qui a
15	FR_Assemblee...	amendement, qui a été considéré par le président de l'	Assemblée	nationale lui-même comme contraire à l'article 70 alinéa 4,
16	FR_Assemblee...	du pays. J'ai envie, si le président de l'	Assemblée	nationale m'y autorise j'espère ne pas réveiller
17	FR_Assemblee...	sièges restant à pourvoir, je proclame vice-présidents de l'	Assemblée	nationale M. Marc Laffineur M. Maurice Leroy Mme Catherine

Search Query  Words  Case  Regex Results Set All hits Context Size 10 token(s)   Adv Search

assemblée

Sort Options Sort to right Sort 1 1R Sort 2 2R Sort 3 3R Order by freq

Progress 

100%

 Time taken (creating KWIC results): 0.1205 sec

# File View (affichage des fichiers chargés)

AntConc

File Edit Settings Help

Target Corpus  
Name: temp  
Files: 1  
Tokens: 151362  
FR\_AssembleeNationale.txt

KWIC Plot File View Cluster N-Gram Collocate Word Keyword Wordcloud ChatAI

File Hits 0 File Types 11765 File Tokens 151362 File Name FR\_AssembleeNationale.txt

[M. le président] La séance est ouverte. L'ordre du jour appelle la suite de la discussion de la seconde partie du projet de loi de finances pour 2022 Cet après-midi, l'Assemblée a commencé l'examen des crédits relatifs au conseil et au contrôle de l'Etat aux pouvoirs publics à la direction de l'action du Gouvernement et au budget annexe s'arrêtant aux porte-parole des groupes. La parole est à Mme Maud Petit.  
[Mme Maud Petit:] Les annexes au projet de loi de finances pour 2022 détaillent le montant des crédits par dotation de façon claire et précise. Elles contiennent toutes les informations requises afin que chacun puisse se forger un avis et décider de manière pertinente du vote des crédits à allouer. La mission montre que, pour l'exercice 2021, le fonctionnement de la présidence de la République a été moins affecté que l'année précédente par la crise sanitaire. Dès le printemps, des redéploiements opérés sur les crédits de déplacement ont conduit à abondier les crédits d'investissement, permettant d'engager de nouveaux projets de modernisation dès ici à la fin de cette année. Des leviers de performance ont aussi été identifiés et des résultats concrets ont d'ores et déjà été enregistrés. Par ailleurs, plusieurs projets contribuant à la préservation de l'environnement et à la transition écologique ont été conduits en vue d'influer sur les comportements, ce que le groupe Mouvement démocrate et démocrates apparentés saute. Concernant l'Assemblée nationale, le bilan est plus contrasté. En effet, alors que le budget pour 2021 s'inscrivait en baisse tant pour les dépenses de fonctionnement que pour celles d'investissement, le budget pour 2022 est marqué par une inflexion notable en raison des dépenses liées au renouvellement électoral. Malgré des ressources budgétaires propres en hausse, le résultat budgétaire serait déficitaire si la dotation restait inchangée. Pour notre groupe, il semble dès lors indispensable de concrétiser l'établissement d'une programmation budgétaire pluriannuelle. S'agissant du fonctionnement du Sénat, le total des dépenses exposées dans le projet de budget pour 2022 est en hausse. Les crédits d'investissement augmentent, atteignant un niveau particulièrement élevé. Notre groupe considère que, compte tenu de la diminution de l'activité du Sénat au cours des campagnes électorales de mars à juin 2022, des sources d'économies importantes pourraient être identifiées. Pour ce qui est du Conseil constitutionnel, les dépenses liées à ses membres sont en diminution. Cette baisse s'explique par l'absence de membres de droit siégeant actuellement en son sein. Toutefois, les autres dépenses sont en augmentation. Les dépenses de personnel sont en hausse, en raison de la professionnalisation du secrétariat général ainsi que des besoins en effectifs liés au déploiement de nouveaux projets. Les dépenses de fonctionnement courant progressent dans la même mesure. Le programme d'investissement pour l'année 2022 s'ordonne autour de la poursuite du plan d'économie d'énergie et de développement durable ainsi que de la nécessaire refonte des outils numériques. Le budget prévoit les dépenses pour l'élection présidentielle sur le modèle de celui mobilisé cinq ans plus tôt. Il tient compte des nouvelles charges qui pèsent sur le Conseil constitutionnel. Le budget annexe montre que la direction de l'information légale et administrative a adapté son organisation à la crise sanitaire afin de garantir la continuité de ses missions de service public pendant cette période. Elle poursuivra en 2022 la modernisation de ses activités afin d'améliorer le service rendu et de répondre aux besoins en constante évolution tant des citoyens que des entreprises. Le projet de budget pour 2022 devrait présenter un solde positif. Quant à la mission elle regroupe pour l'exercice 2022 les crédits et les emplois finançant l'activité des services directement rattachés au Premier ministre ainsi que ceux liés à la présidence française de l'Union européenne, ce qui impliquera une nette hausse du budget comme ce fut le cas lors de la précédente présidence en 2008.  
[M. Jean-Paul Lecocq:] Ce n'était pas une année d'élection présidentielle.  
[Mme Maud Petit:] La mission montre que de nouveaux moyens d'action au profit du Conseil économique, social et environnemental seront déclinés et développés au cours de la législature 2021-2026, conformément à la dernière réforme de l'institution. La nouvelle programmation triennale des contrôles des juridictions financières s'inscrit, quant à elle, dans la perspective de fonctions plus réactives et attachées aux préoccupations des citoyens, et sera marquée par l'affirmation de sa statut internationale grâce à l'exercice d'un mandat d'audit externe de l'Organisation des Nations unies à compter du 1er juillet 2022 pour six ans. Enfin, les juridictions administratives poursuivront la politique ambitieuse de modernisation de leur organisation et de leurs méthodes de travail. En ce sens, le renforcement des moyens alloués depuis plusieurs années a permis d'atteindre puis de dépasser l'objectif assigné à la juridiction administrative de ramener à un an le délai prévisible moyen de jugement. Le groupe Mouvement démocrate et démocrates apparentés donne son entière approbation aux crédits de ces missions.  
[Mme Isabelle Santiago:] Les missions budgétaires dont nous examinons les crédits répondent à des enjeux dont on ne peut nier l'importance puisqu'elles ont vocation à financer l'ensemble de notre structure institutionnelle, de l'Elysée aux assemblées parlementaires en passant par les services de Matignon, le Conseil constitutionnel, le Conseil d'Etat et nos autorités administratives indépendantes. Nous devons donc faire preuve d'une grande attention. S'agissant de la mission notons tout d'abord une stabilité des crédits, y compris pour l'Elysée qui voit son budget se stabiliser à 105,3 millions d'euros, après avoir augmenté depuis le début du quinquennat à raison de 6 millions d'euros chaque année. Les efforts de diminution entrepris sous le précédent quinquennat pour atteindre l'objectif des 100 millions d'euros n'ont pas été poursuivis. La stabilisation intervient en effet après une hausse. Nous regrettons, par ailleurs, le manque de précision de ce budget et déplorons que le contrôle exercé sur ce dernier par la Cour des comptes ne réponde pas à de nombreuses interrogations et observations à son sujet. C'est aussi un manque de précision que nous observons, comme dans les précédents projets de loi de finances, pour ce qui concerne les coûts des missions d'évaluation des politiques publiques, qu'elles soient menées par les commissions permanentes ou par le Comité d'évaluation et de contrôle des politiques publiques. Le budget alloué à cette instance, bras armé de l'Assemblée en matière d'évaluation, mériterait d'être détaillé afin que l'on puisse établir des comparaisons avec des institutions analogues à l'étranger. Si le CEC n'a pas à rougir de la qualité de ses travaux, sa mission est suffisamment essentielle pour que ses crédits soient sensiblement augmentés. Aujourd'hui, ils dépasseraient tout juste les 100 000 euros. La mission connaît une augmentation de 11,95 par rapport à la loi de finances initiale pour 2019. Le PLF pour 2022 traduit la poursuite des efforts en matière de sécurité informatique de l'Etat. L'Agence nationale de la sécurité des systèmes d'information bénéficia d'une augmentation de 50 ETDs équivalents temps plein travaillé après la création de 40 ETDs équivalents temps plein en 2021. En dépit de cette progression continue, on relève en matière de sécurité

Search Query  Words  Case  Regex  Hit Location 0

assemblée

Progress 100%

Start  Adv Search

Time taken (creating plot results): 0.1128 sec

# Cluster

KWIC Plot File View Cluster N-Gram Collocate Word Keyword Wordcloud ChatAI

Cluster Types 44 Cluster Tokens 138 Page Size 100 hits 1 to 44 of 44 hits

	Cluster	Rank	Freq	Range
1	assemblée nationale	1	67	1
2	assemblée a	2	11	1
3	assemblée générale	3	6	1
4	assemblée est	4	3	1
5	assemblée mme	4	3	1
6	assemblée qui	4	3	1
7	assemblée au	7	2	1
8	assemblée et	7	2	1
9	assemblée je	7	2	1
10	assemblée les	7	2	1
11	assemblée m	7	2	1
12	assemblée sur	7	2	1
13	assemblée à	7	2	1
14	assemblée c	14	1	1
15	assemblée ce	14	1	1
16	assemblée cela	14	1	1
17	assemblée comme	14	1	1
18	assemblée dans	14	1	1
19	assemblée de	14	1	1
20	assemblée doit	14	1	1

Search Query  Words  Case  Regex Cluster Size 2 Min. Freq 1 Min. Range 1  
assemblée Start  Adv Search

Sort by Frequency  Invert Order Search Term Position  On Left  On Right  On Left/Right

# Cluster : possibilité de définir la position du terme

The screenshot shows the AntConc search interface. At the top, there are search parameters: 'Search Query' (checked for 'Words'), 'Case' (unchecked), 'Regex' (unchecked), 'Cluster Size' set to 2, 'Min. Freq' set to 1, and 'Min. Range' set to 1. Below these are fields for the search term ('assemblée') and search buttons ('Start', 'Adv Search'). Underneath, there are sorting options ('Sort by Frequency', 'Invert Order') and the 'Search Term Position' section, which is highlighted with a red box. This section contains three radio button options: 'On Left' (selected), 'On Right', and 'On Left/Right'.

# N-Gram

	Type	Rank	Freq	Range	S1_TT	S1_Ent
1	la + de	1	556	1	0.383	0.904
2	I + de	2	512	1	0.309	0.82
3	le + de	3	488	1	0.375	0.883
4	ne + pas	4	472	1	0.472	0.899
5	n + pas	5	463	1	0.153	0.624
6	les + de	6	322	1	0.575	0.942
7	m + président	7	299	1	0.007	0.032
8	la + des	8	291	1	0.464	0.904
9	I + no	9	262	1	0.004	0.0
10	de + de	10	258	1	0.705	0.941
11	des + de	11	244	1	0.574	0.936
12	en + de	11	244	1	0.328	0.849
13	I + du	13	239	1	0.322	0.711
14	au + de	14	210	1	0.362	0.893
15	I + des	15	199	1	0.452	0.885
16	une + de	16	198	1	0.444	0.896
17	un + de	17	192	1	0.563	0.915
18	du + de	18	191	1	0.393	0.839
19	de + et	19	190	1	0.737	0.967
20	est + avis	20	187	1	0.011	0.048

Search Query  Words  Case  Regex N-Gram Size 3 Open Slots 1 Min. Freq 1 Min. Range 1

Sort by Frequency  Invert Order

# Collocate

Collocate Types 17 Collocate Tokens 344 Page Size 100 hits <input type="button"/> 1 to 17 of 17 hits <input type="button"/>								
	Collocate	Rank	FreqLR	FreqL	FreqR	Range	Likelihood	Effect
1	nationale	1	67	0	67	1	416.660	5.832
2	I	2	142	127	15	1	214.896	2.163
3	enceinte	3	20	20	0	1	150.588	6.777
4	parole	4	12	0	12	1	31.161	3.143
5	générale	5	6	0	6	1	26.132	4.504
6	dans	6	30	28	2	1	25.336	1.552
7	notre	7	16	15	1	1	25.066	2.265
8	fonctionnement	8	5	5	0	1	23.612	4.777
9	midi	9	4	4	0	1	23.442	5.607
10	sénat	10	6	1	5	1	23.022	4.114
11	vice	11	5	5	0	1	22.683	4.640
12	sein	12	5	4	1	1	21.068	4.399
13	cette	13	16	12	4	1	18.908	1.904
14	pas	14	1	0	1	1	18.105	-3.647
15	services	15	5	4	1	1	16.307	3.673
16	bras	16	2	1	1	1	15.033	6.777
17	commencé	16	2	0	2	1	15.033	6.777

Search Query  Words  Case  Regex Window Span From 5L  To 5R  Min. Freq 1  Min. Range 1   
assemblée  Start  Adv Search  
Sort by Likelihood  Invert Order

# Collocate

« You shall know a word by the company it keeps. »

J.R. Firth, 1957:11

**Collocats** : mots qui apparaissent fréquemment à proximité d'un mot donné dans un corpus.

# Collocate : formule log-likelihood

Quand on cherche des collocats dans un corpus, on veut savoir si deux mots apparaissent ensemble plus souvent que ce à quoi nous pourrions nous attendre par hasard → application de la formule *log-likelihood*.

# Collocate : formule log-likelihood

Quand on cherche des collocats dans un corpus, on veut savoir si deux mots apparaissent ensemble plus souvent que ce à quoi nous pourrions nous attendre par hasard → application de la formule *log-likelihood*.

On compare deux hypothèses :

- hypothèse nulle : les deux mots ne sont pas liés (leur cooccurrence est due au hasard, résultat non significatif) ;
- hypothèse alternative : les deux mots sont liés (leur cooccurrence n'est pas due au hasard, résultat statistiquement significatif).

Plus le score de *log-likelihood* est élevé, plus il est probable que les deux mots soient liés.

# Collocate : formule log-likelihood, valeurs

Soit :

- $N$  : nombre total de mots dans le corpus ;
- $O_{11}$  : nombre d'occurrences du terme A suivi du terme B ;
- $O_{12}$  : nombre d'occurrences du terme A suivi de n'importe quel terme autre que B ;
- $O_{21}$  : nombre d'occurrences du terme B suivi de n'importe quel terme autre que A ;
- $O_{22}$  : nombre d'occurrences où ni le terme A ni le terme B ne sont dans le contexte.

# Collocate : formule log-likelihood

## Tableau de contingence

	B	non-B	Total
A	$O_{11}$	$O_{12}$	$O_{11} + O_{12}$
non-A	$O_{21}$	$O_{22}$	$O_{21} + O_{22}$
Total	$O_{11} + O_{21}$	$O_{12} + O_{22}$	$N$

# Collocate : formule log-likelihood, hypothèse nulle

Calcul des valeurs attendues ( $E$ ) afin de voir à quel point les résultats obtenus diffèrent des valeurs observées ( $O$ ), et ainsi déterminer si la cooccurrence est significative (et n'est pas le fruit du hasard).

Sous l'hypothèse nulle d'indépendance, la probabilité que A co-occure avec B devrait correspondre au produit des probabilités individuelles de A et B, soit :

$$P(A, B) = P(A) \times P(B)$$

---

$$E_{ij} = \frac{(\text{total ligne}_i \times \text{total colonne}_j)}{N}$$

# Collocate : formule log-likelihood, hypothèse nulle

$$E_{11} = \frac{(O_{11} + O_{12}) \times (O_{11} + O_{21})}{N}$$

	B	non-B
A	$O_{11}$	$O_{12}$
non-A	$O_{21}$	$O_{22}$

# Collocate : formule log-likelihood, hypothèse nulle

$$E_{12} = \frac{(O_{11} + O_{12}) \times (O_{12} + O_{22})}{N}$$

	B	non-B
A	$O_{11}$	$O_{12}$
non-A	$O_{21}$	$O_{22}$

# Collocate : formule log-likelihood, hypothèse nulle

$$E_{21} = \frac{(O_{21} + O_{22}) \times (O_{11} + O_{21})}{N}$$

	B	non-B
A	$O_{11}$	$O_{12}$
non-A	$O_{21}$	$O_{22}$

# Collocate : formule log-likelihood, hypothèse nulle

$$E_{22} = \frac{(O_{21} + O_{22}) \times (O_{12} + O_{22})}{N}$$

	B	non-B
A	$O_{11}$	$O_{12}$
non-A	$O_{21}$	$O_{22}$

## Collocate : formule log-likelihood

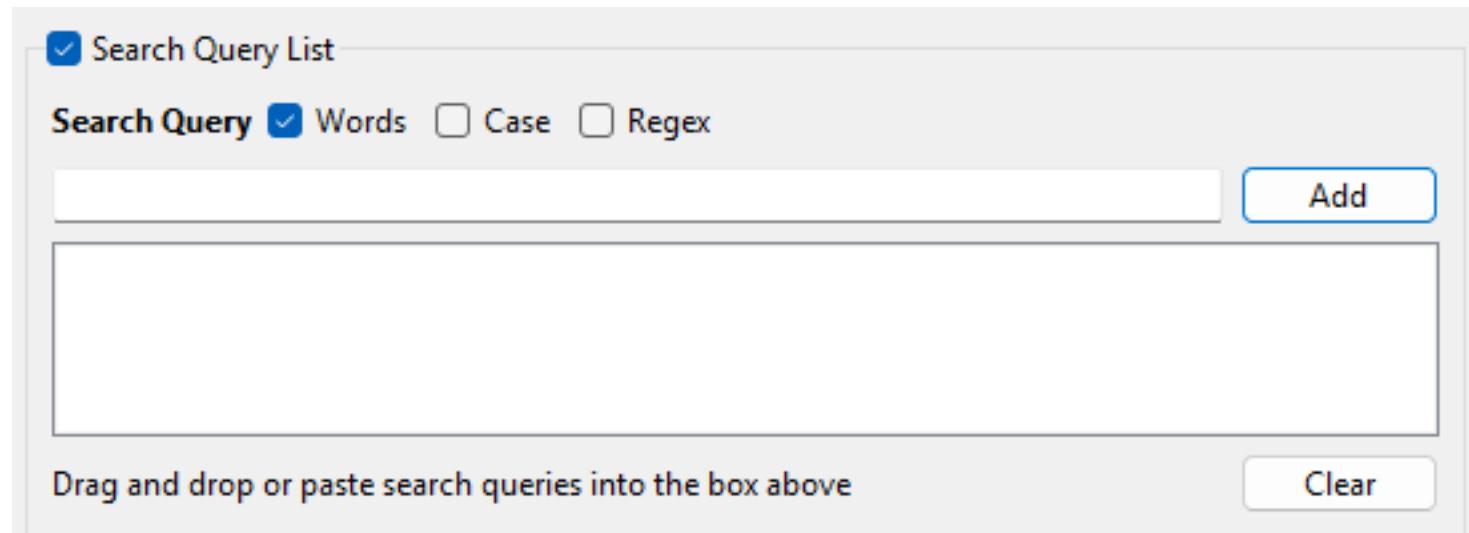
On peut ensuite utiliser nos valeurs attendues pour calculer le score de log-likelihood :

$$G^2 = 2 \sum \left( O_{ij} \log \frac{O_{ij}}{E_{ij}} \right)$$

Où  $i$  et  $j$  correspondent aux différentes combinaisons de termes : 11, 12, 21 et 22.

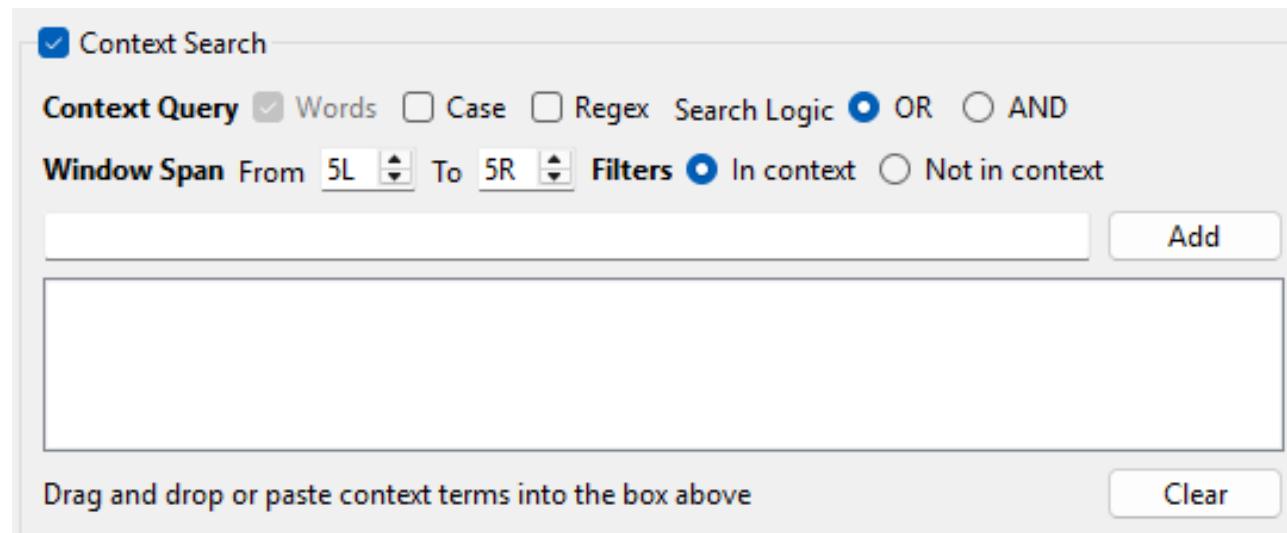
# Paramètres de recherche avancées

*Search Query List* : liste du ou des termes à rechercher



# Paramètres de recherche avancées

*Context Search* : liste du ou des termes qui doivent se trouver (ou non) dans le contexte gauche ou droit



# Paramètres de recherche avancées

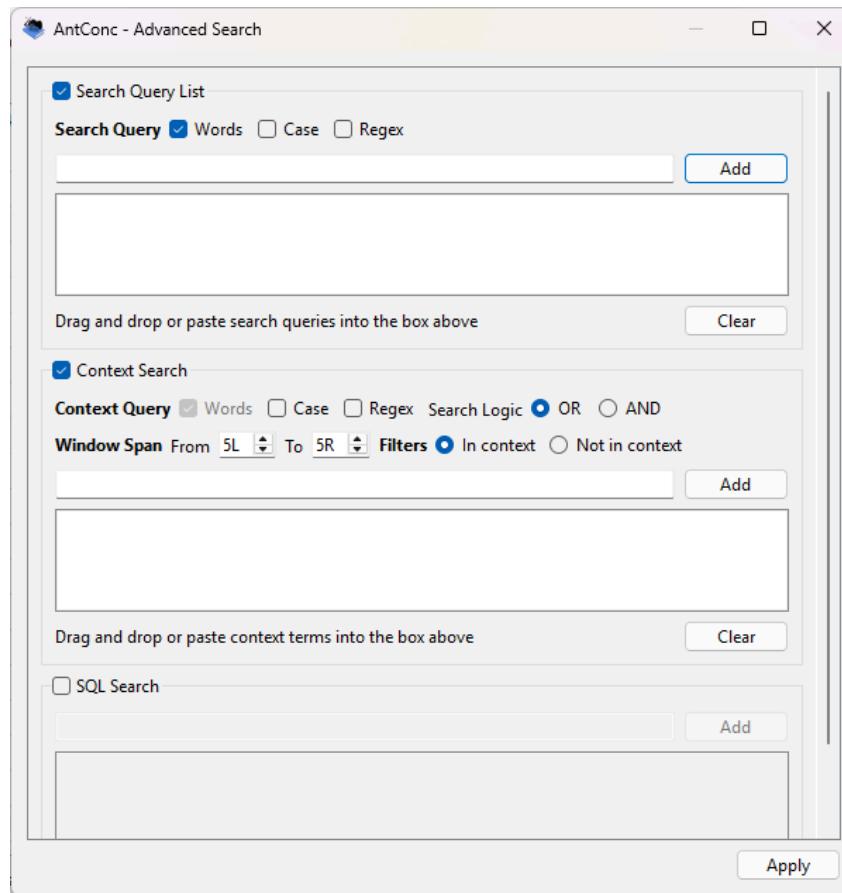
## Context Search

*Window Span* correspond à la **plage à utiliser pour l'inclusion/l'exclusion des termes.**

*Search Logic: OR* : utiliser l'un des termes de la liste.

*Search Logic: AND* : utiliser **tous** les termes de la liste.

# Recherche avancée



# Bonus : Wordcloud

