Mode d'emploi de l'interface guidée Scientext

Achille Falaise, Agnès Tutin

Avril 2016

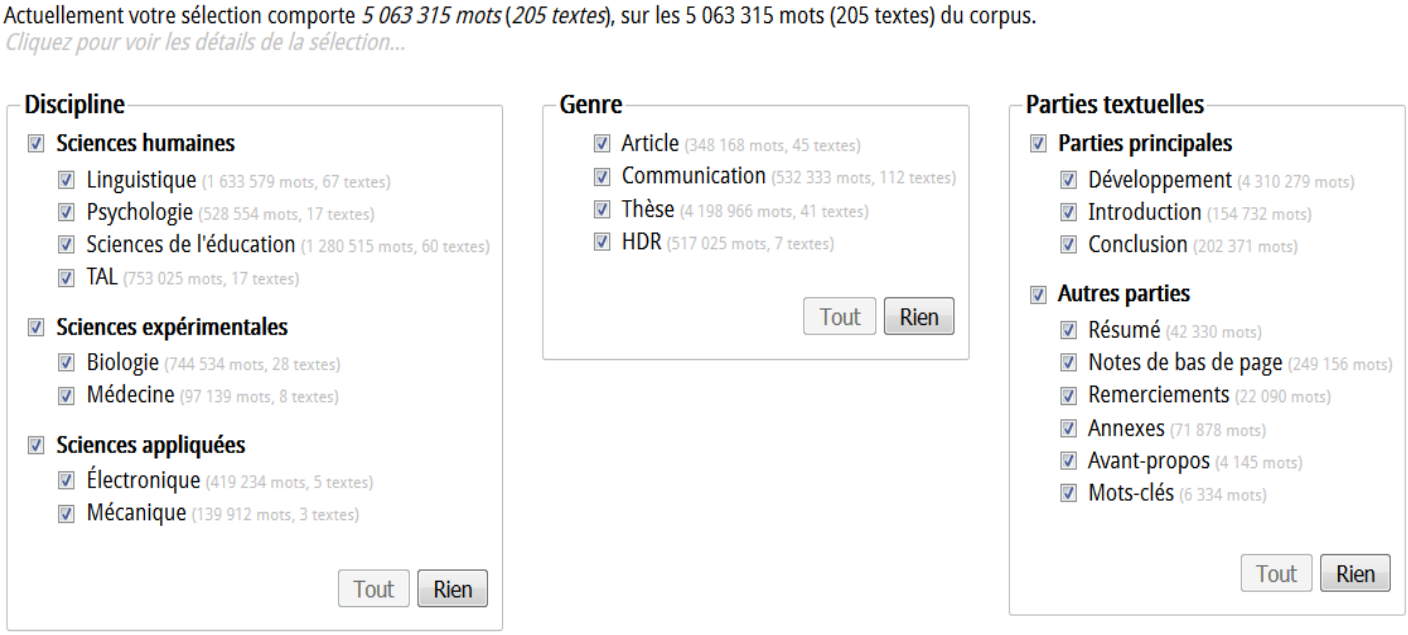
L'utilisation de l'interface simple et guidée se fait en trois temps :

1. Sélection du corpus
2. Recherche des expressions sélectionnées
3. Affichage et statistiques

Ces étapes sont détaillées dans les sections suivantes.

1 Étape 1 : Sélection du corpus

Une fois sur la page des textes, Sélectionner ensuite les critères pertinents : discipline (ou famille de disciplines), genre textuel, partie textuelle.



Aller ensuite sur Liste. La liste des textes sélectionnés s'affiche. On peut effectuer le tri par discipline, type de texte, titre, recueil et auteur. On pourra décocher les textes que l'on ne souhaite pas interroger.
Pour passer à la recherche dans les textes, on cliquera sur Recherche.

2 Recherche guidée dans les textes

Trois modes de recherche sont prévus pour explorer les textes :

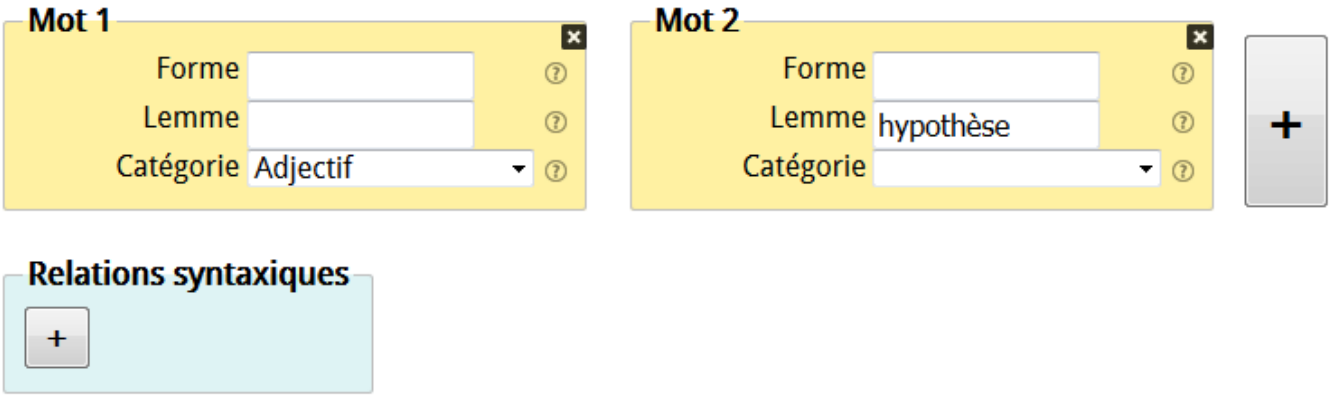
- un mode de recherche sémantique, avec des grammaires prédéfinies (pas

détaillé ici).
- un mode libre par recherche guidée.
- un mode de recherche avancé (qui ne sera pas détaillé ici) et qui utilise des expressions régulières.

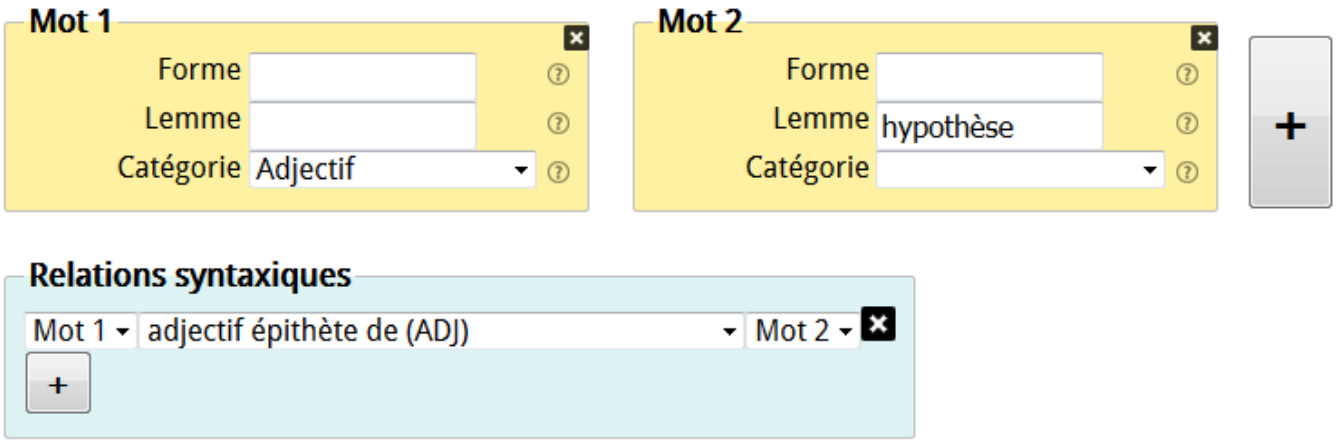
C'est ici le mode de recherche guidée qui est décrit.

2.1 Recherche libre

L'utilisateur utilise ici des menus, qui indiquent pour les mots : **la forme, la catégorie syntaxique** ou **le lemme**. Il est aussi possible d'utiliser des relations syntaxiques, grâce à une analyse syntaxique réalisée automatiquement à l'aide du logiciel Syntex (de Didier Bourigault). La recherche suivante permet de repérer les suites de mots où un adjectif apparait avant le lemme *hypothèse*, par exemple *première hypothèse*. En revanche, la recherche ne prendra pas en compte les cas où l'adjectif est postposé au nom ou les cas où l'adjectif n'est pas immédiatement antérieur au nom.



Pour proposer une requête plus adaptée, il faudra utiliser les relations syntaxiques (pour cela, cliquer sur « + » dans l'encadré « Relations syntaxiques ». On indique alors la relation syntaxique (de dépendance) entre les mots. **Attention, dans ce cas, il n'y a plus d'ordre linéaire !** La recherche suivante extrait les adjectifs épithètes qui accompagnent *hypothèse* indépendamment de leur position.

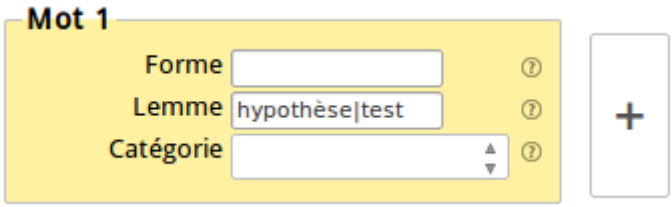


Les recherches qui utilisent les relations syntaxiques sont généralement plus intéressantes que celles qui n'utilisent que l'ordre linéaire, mais elles sont un peu plus complexes à maîtriser.

Pour les recherches, il est possible de paramétrer le nombre de résultats voulus. Par défaut, la taille est limitée à 1000 réponses.

2.2 Listes de mots

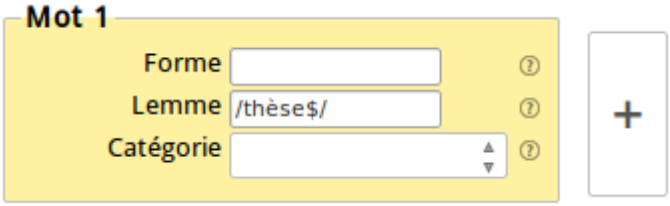
Pour chercher un mot parmi une liste, on utilise la caractère | (AltGr+6 sur un clavier PC). Ainsi, dans l'exemple ci-dessous, on recherchera toutes les occurrences du lemme *hypothèse* **ou** *test*.



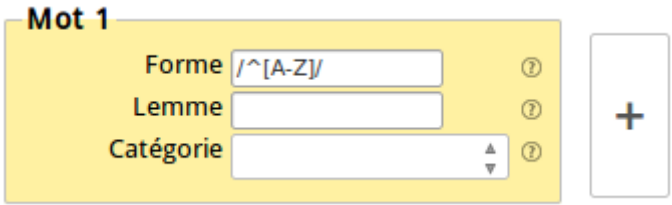
2.3 Expressions régulières

Pour aller plus loin, il est possible d'utiliser des expressions régulières. Par exemple :

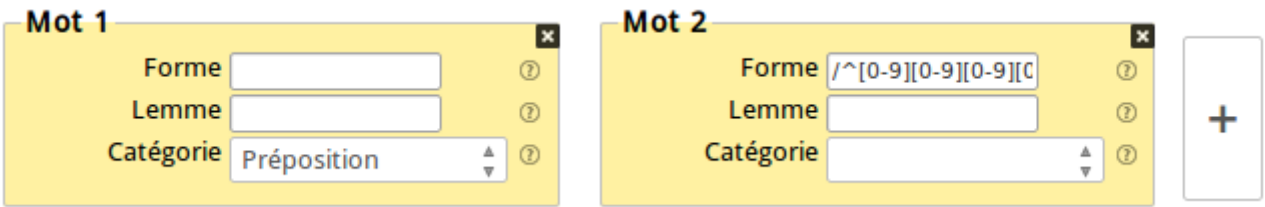
- */^hypo/* va rechercher tous les lemmes qui commencent par *hypo* : *hypothèse*, *hypothénuse*, etc.
- */thèse$/* va rechercher tous les lemmes qui se terminent par *thèse* : *thèse*, *synthèse*, etc.



- */^[A-Z]/* va rechercher toutes les formes qui commencent par une majuscule.



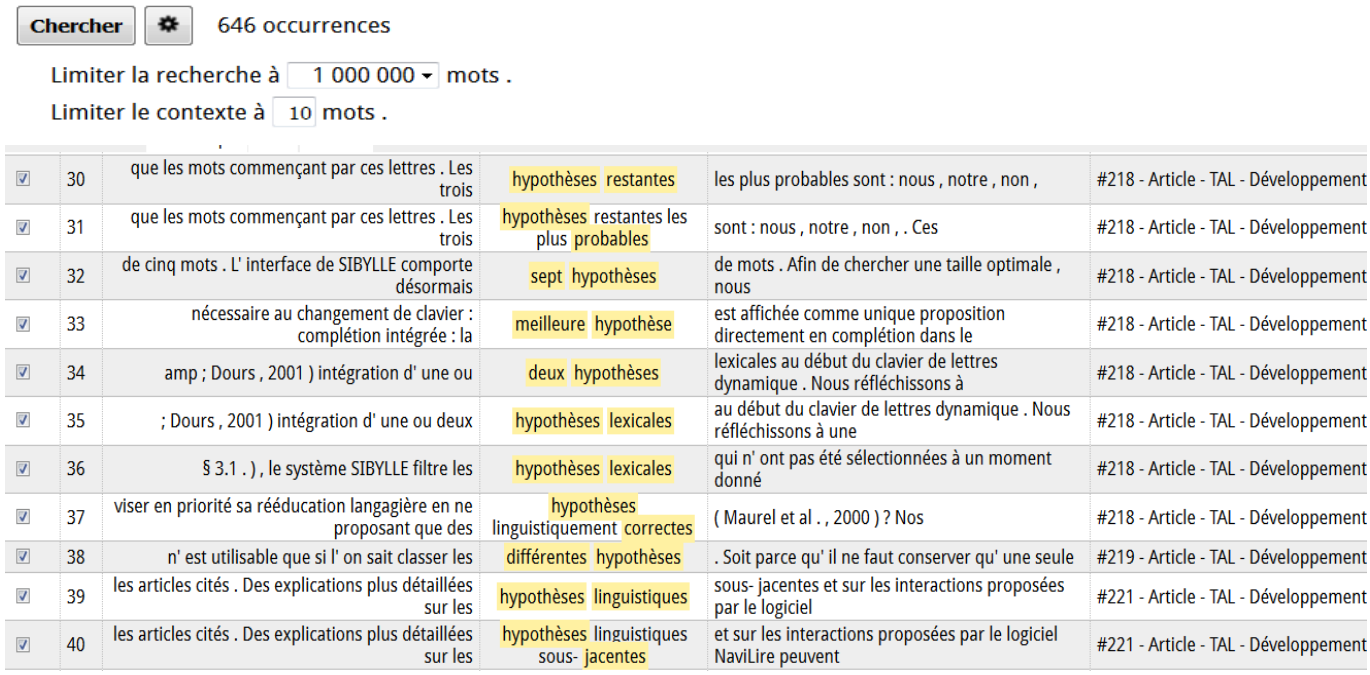
- /^[0-9][0-9][0-9][0-9]$/ va rechercher toutes les formes qui contiennent exactement quatre chiffres. Pratique pour rechercher des années ! Dans l'exemple ci-dessous, on recherche les prépositions qui introduisent une année.



3 Affichage des résultats et résultats

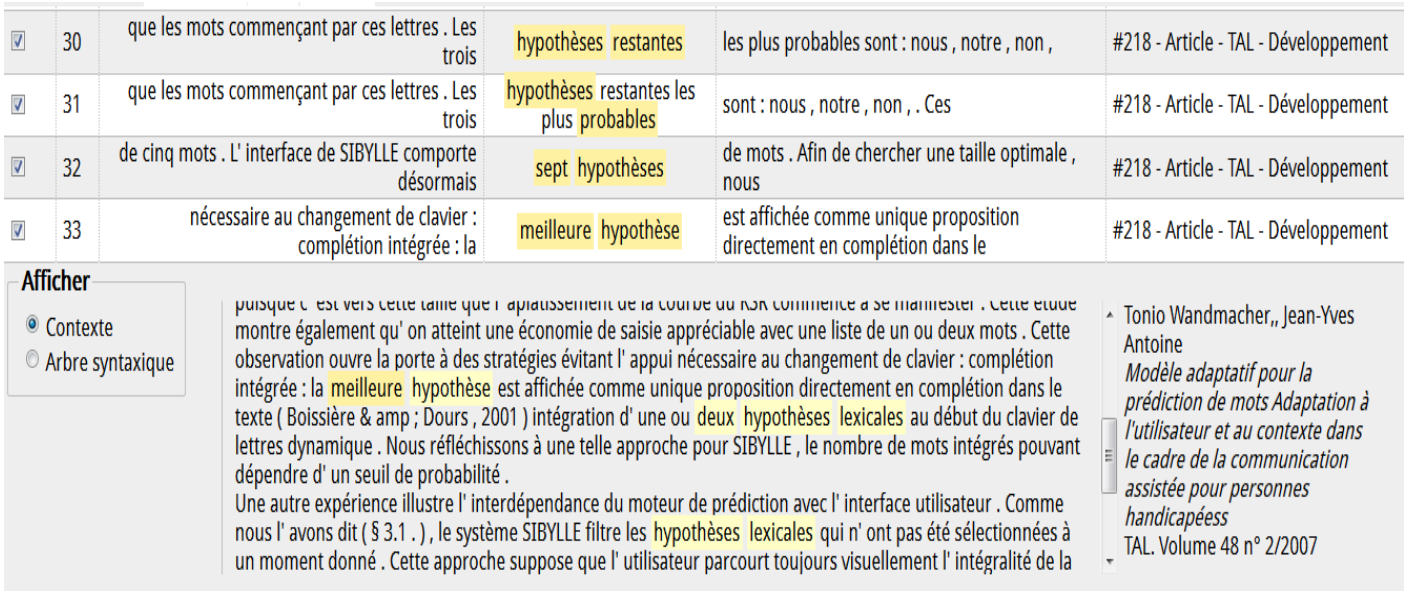
3.1 Affichage

Par défaut, l'affichage des résultats se fait dans une concordance KWIC dont la taille est paramétrable (par défaut, 10 mots avant, 10 mots après dans la même phrase), comme ci-dessous.



☑	30	que les mots commençant par ces lettres . Les trois	hypothèses restantes	les plus probables sont : nous , notre , non ,	#218 - Article - TAL - Développement
☑	31	que les mots commençant par ces lettres . Les trois	hypothèses restantes les plus probables	sont : nous , notre , non , . Ces	#218 - Article - TAL - Développement
☑	32	de cinq mots . L' interface de SIBYLLE comporte désormais	sept hypothèses	de mots . Afin de chercher une taille optimale , nous	#218 - Article - TAL - Développement
☑	33	nécessaire au changement de clavier : complétion intégrée : la	meilleure hypothèse	est affichée comme unique proposition directement en complétion dans le	#218 - Article - TAL - Développement
☑	34	amp ; Dours , 2001) intégration d' une ou	deux hypothèses	lexicales au début du clavier de lettres dynamique . Nous réfléchissons à	#218 - Article - TAL - Développement
☑	35	; Dours , 2001) intégration d' une ou deux	hypothèses lexicales	au début du clavier de lettres dynamique . Nous réfléchissons à une	#218 - Article - TAL - Développement
☑	36	§ 3.1 .) , le système SIBYLLE filtre les	hypothèses lexicales	qui n' ont pas été sélectionnées à un moment donné	#218 - Article - TAL - Développement
☑	37	viser en priorité sa rééducation langagière en ne proposant que des	hypothèses linguistiquement correctes	(Maurel et al . , 2000) ? Nos	#218 - Article - TAL - Développement
☑	38	n' est utilisable que si l' on sait classer les	différentes hypothèses	. Soit parce qu' il ne faut conserver qu' une seule	#219 - Article - TAL - Développement
☑	39	les articles cités . Des explications plus détaillées sur les	hypothèses linguistiques	sous- jacentes et sur les interactions proposées par le logiciel	#221 - Article - TAL - Développement
☑	40	les articles cités . Des explications plus détaillées sur les	hypothèses linguistiques sous- jacentes	et sur les interactions proposées par le logiciel NaviLire peuvent	#221 - Article - TAL - Développement

L'utilisateur peut avoir un contexte plus large en cliquant sur la concordance, comme ci-dessous.

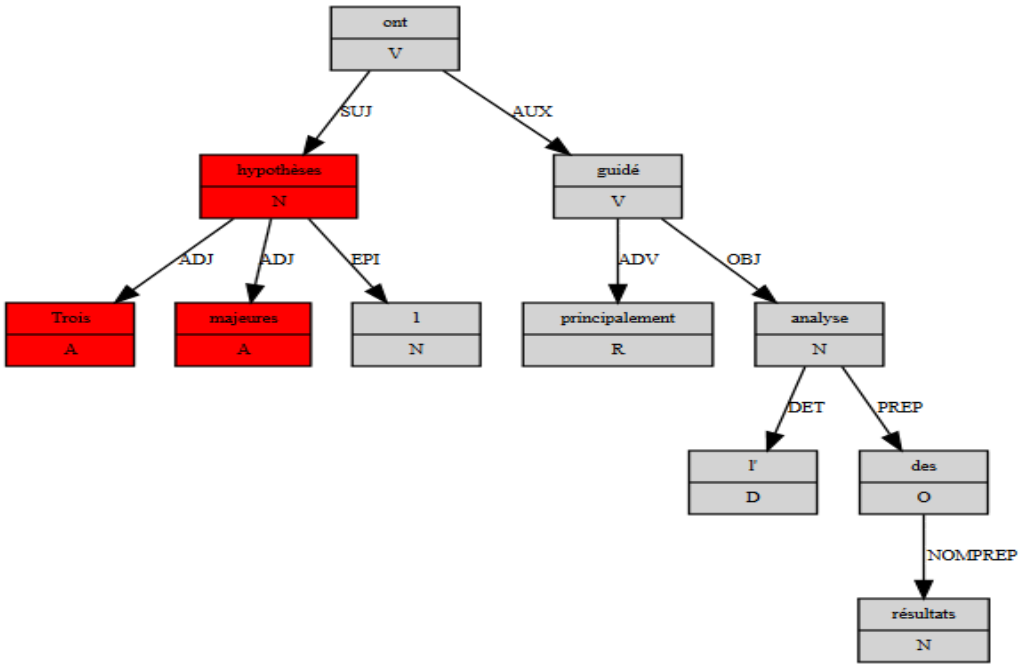


☑	30	que les mots commençant par ces lettres . Les trois	hypothèses restantes	les plus probables sont : nous , notre , non ,	#218 - Article - TAL - Développement
☑	31	que les mots commençant par ces lettres . Les trois	hypothèses restantes les plus probables	sont : nous , notre , non , . Ces	#218 - Article - TAL - Développement
☑	32	de cinq mots . L' interface de SIBYLLE comporte désormais	sept hypothèses	de mots . Afin de chercher une taille optimale , nous	#218 - Article - TAL - Développement
☑	33	nécessaire au changement de clavier : complétion intégrée : la	meilleure hypothèse	est affichée comme unique proposition directement en complétion dans le	#218 - Article - TAL - Développement

Il est également possible d'afficher l'analyse syntaxique de la phrase correspondant à la requête, en cochant arbre syntaxique. Par exemple, la figure ci-dessous indique l'analyse syntaxique pour la phrase

Trois hypothèses majeures 1 ont principalement guidé l'analyse des résultats :

:



3.2 Filtrage

Une case à cocher (cochée par défaut) se trouve à gauche de chaque ligne. Les lignes décochées ne seront pas comptabilisées dans les statistiques et n'apparaîtront pas dans les exportations. C'est utile pour filtrer les résultats qui ne sont pas pertinents.

3.3 Exportation

Il est possible d'extraire le résultat des concordances (voir les liens au bas de la fenêtre de résultats) aux formats :

- CSV : pour LibreOffice et OpenOffice.
- HTML : pratique pour imprimer (il vaut mieux utiliser le format paysage et dézoomer un peu).
- XLSX : pour Excel, fonctionne aussi avec LibreOffice et OpenOffice.

3.4 Statistiques

On peut avoir l'affichage de quelques statistiques simples. Pour cela, il faut cliquer sur statistiques ou sur Suite à partir des concordances.
On peut ainsi obtenir :

- la liste des lemmes correspondant à la requête, comme ci-dessous.

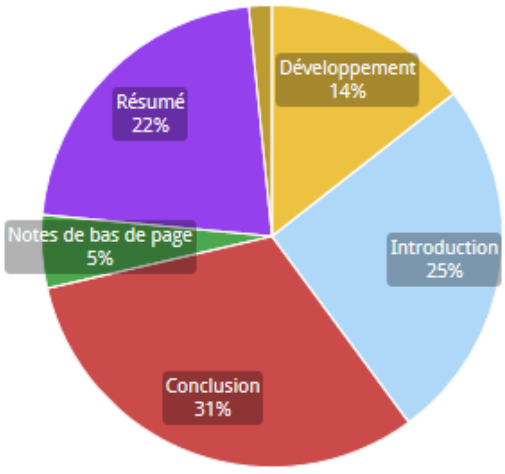
Lemme

Token	Nb occ	% occ	‰ corpus
deux hypothèse	65	10.0619	0.12837
premier hypothèse	39	6.0372	0.07702
général hypothèse	33	5.1084	0.06517
suivant hypothèse	25	3.87	0.04937
second hypothèse	24	3.7152	0.04740
tel hypothèse	20	3.096	0.03950
trois hypothèse	19	2.9412	0.03752
deuxième hypothèse	19	2.9412	0.03752
autre hypothèse	19	2.9412	0.03752
principal hypothèse	18	2.7864	0.03555
théorique hypothèse	16	2.4768	0.03160
alternatif hypothèse	15	2.322	0.02962
explicatif hypothèse	15	2.322	0.02962
nul hypothèse	13	2.0124	0.02567
différent hypothèse	13	2.0124	0.02567
interprétatif hypothèse	11	1.7028	0.02172
dernier hypothèse	11	1.7028	0.02172
nouvelle hypothèse	11	1.7028	0.02172

- la répartition des lemmes (fréquences absolues, relatives et schémas) selon :
  - la partie textuelle (Cf Schéma ci-dessous).
  - la discipline
  - le type de texte
  - le texte

Partie textuelle

Propriété	Nb occ		Nb T		Nb norm
Développement	537	/	4310279	=	0.00012459
Conclusion	55	/	202371	=	0.00027178
Introduction	34	/	154732	=	0.00021973
Notes de bas de page	11	/	249156	=	0.00004415
Résumé	8	/	42330	=	0.00018899
Annexes	1	/	71878	=	0.00001391



Il est également possible de rapatrier les résultats au format CSV pour une utilisation avec un tableur.

4 Historique

Enfin, l'onglet Historique permet de reprendre une requête effectuée auparavant.