

# 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.

Actuellement votre sélection comporte 5 063 315 mots (205 textes), sur les 5 063 315 mots (205 textes) du corpus.  
*Cliquez pour voir les détails de la sélection...*

Discipline	Genre	Parties textuelles
<input checked="" type="checkbox"/> <b>Sciences humaines</b> <ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Linguistique (1 633 579 mots, 67 textes)</li><li><input checked="" type="checkbox"/> Psychologie (528 554 mots, 17 textes)</li><li><input checked="" type="checkbox"/> Sciences de l'éducation (1 280 515 mots, 60 textes)</li><li><input checked="" type="checkbox"/> TAL (753 025 mots, 17 textes)</li></ul>	<input checked="" type="checkbox"/> Article (348 168 mots, 45 textes) <input checked="" type="checkbox"/> Communication (532 333 mots, 112 textes) <input checked="" type="checkbox"/> Thèse (4 198 966 mots, 41 textes) <input checked="" type="checkbox"/> HDR (517 025 mots, 7 textes)	<input checked="" type="checkbox"/> <b>Parties principales</b> <ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Développement (4 310 279 mots)</li><li><input checked="" type="checkbox"/> Introduction (154 732 mots)</li><li><input checked="" type="checkbox"/> Conclusion (202 371 mots)</li></ul>
<input checked="" type="checkbox"/> <b>Sciences expérimentales</b> <ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Biologie (744 534 mots, 28 textes)</li><li><input checked="" type="checkbox"/> Médecine (97 139 mots, 8 textes)</li></ul>	<input type="button" value="Tout"/> <input type="button" value="Rien"/>	<input checked="" type="checkbox"/> <b>Autres parties</b> <ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Résumé (42 330 mots)</li><li><input checked="" type="checkbox"/> Notes de bas de page (249 156 mots)</li><li><input checked="" type="checkbox"/> Remerciements (22 090 mots)</li><li><input checked="" type="checkbox"/> Annexes (71 878 mots)</li><li><input checked="" type="checkbox"/> Avant-propos (4 145 mots)</li><li><input checked="" type="checkbox"/> Mots-clés (6 334 mots)</li></ul>
<input checked="" type="checkbox"/> <b>Sciences appliquées</b> <ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Électronique (419 234 mots, 5 textes)</li><li><input checked="" type="checkbox"/> Mécanique (139 912 mots, 3 textes)</li></ul>		<input type="button" value="Tout"/> <input type="button" value="Rien"/>

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 apparaît 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.

Mot 1

Forme

Lemme

Catégorie Adjectif

Mot 2

Forme

Lemme hypothèse

Catégorie

+

Relations syntaxiques

+

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.

Mot 1

Forme

Lemme

Catégorie Adjectif

Mot 2

Forme

Lemme hypothèse

Catégorie

+

Relations syntaxiques

Mot 1 adjectif épithète de (ADJ) Mot 2

+

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*.

**Mot 1**

Forme	<input type="text"/>	?
Lemme	<input type="text" value="hypothèse test"/>	?
Catégorie	<input type="text"/>	?

+

## 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.

**Mot 1**

Forme	<input type="text"/>	?
Lemme	<input type="text" value="/thèse\$/"/>	?
Catégorie	<input type="text"/>	?

+

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

**Mot 1**

Forme	<input type="text" value="/^[A-Z]/"/>	?
Lemme	<input type="text"/>	?
Catégorie	<input type="text"/>	?

+

- `/^[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.

<b>Mot 1</b>	<input type="text"/>	?
Lemme	<input type="text"/>	?
Catégorie	<input type="text" value="Préposition"/>	?

<b>Mot 2</b>	<input type="text" value="/^[0-9][0-9][0-9][0-9]\$/"/>	?
Lemme	<input type="text"/>	?
Catégorie	<input type="text"/>	?

+

## 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.

646 occurrences

Limiter la recherche à  mots .

Limiter le contexte à  mots .

<input checked="" type="checkbox"/>	30	que les mots commençant par ces lettres . Les trois	hypothèses restantes	les plus probables sont : nous , notre , non ,	#218 - Article - TAL - Développement
<input checked="" type="checkbox"/>	31	que les mots commençant par ces lettres . Les trois	hypothèses restantes les plus probables	sont : nous , notre , non , . Ces	#218 - Article - TAL - Développement
<input checked="" type="checkbox"/>	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
<input checked="" type="checkbox"/>	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
<input checked="" type="checkbox"/>	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
<input checked="" type="checkbox"/>	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
<input checked="" type="checkbox"/>	36	§ 3.1 . ) , le système SIBYLLE filtre les	hypothèses lexicales	qui n' ont pas été sélectionnées à un moment donné	#218 - Article - TAL - Développement
<input checked="" type="checkbox"/>	37	viser en priorité sa rééducation langagière en ne proposant que des	hypothèses linguistiquement correctes	( Maurel et al . , 2000 ) ? Nos	#218 - Article - TAL - Développement
<input checked="" type="checkbox"/>	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
<input checked="" type="checkbox"/>	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
<input checked="" type="checkbox"/>	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.

<input checked="" type="checkbox"/>	30	que les mots commençant par ces lettres . Les trois	hypothèses restantes	les plus probables sont : nous , notre , non ,	#218 - Article - TAL - Développement
<input checked="" type="checkbox"/>	31	que les mots commençant par ces lettres . Les trois	hypothèses restantes les plus probables	sont : nous , notre , non , . Ces	#218 - Article - TAL - Développement
<input checked="" type="checkbox"/>	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
<input checked="" type="checkbox"/>	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

#### Afficher

- Contexte
- Arbre syntaxique

puisque c' est vers cette taille que l' aplatissement de la courbe du ROR commence à se manifester . Cette étude montre également qu' on atteint une économie de saisie appréciable avec une liste de un ou deux mots . Cette observation ouvre la porte à des stratégies évitant l' appui nécessaire au changement de clavier : complétion intégrée : la **meilleure hypothèse** est affichée comme unique proposition directement en complétion dans le texte ( Boissière & amp ; Dours , 2001 ) intégration d' une ou deux **hypothèses lexicales** au début du clavier de lettres dynamique . Nous réfléchissons à une telle approche pour SIBYLLE , le nombre de mots intégrés pouvant dépendre d' un seuil de probabilité .

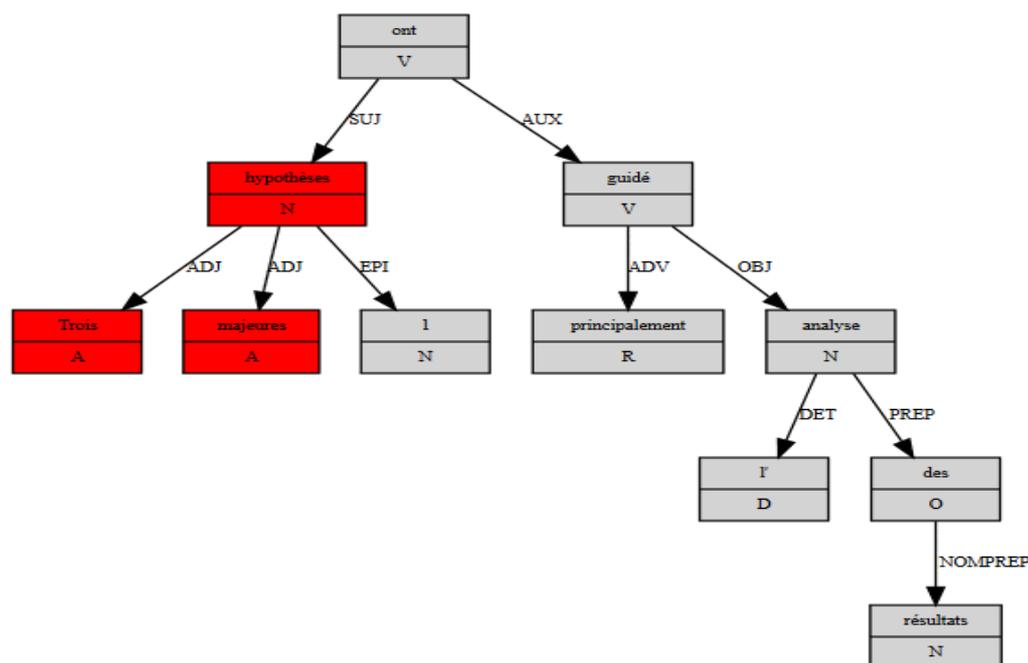
Une autre expérience illustre l' interdépendance du moteur de prédiction avec l' interface utilisateur . Comme nous l' avons dit ( § 3.1 . ) , le système SIBYLLE filtre les **hypothèses lexicales** qui n' ont pas été sélectionnées à un moment donné . Cette approche suppose que l' utilisateur parcourt toujours visuellement l' intégralité de la

▲ Tonio Wandmacher,, Jean-Yves Antoine  
*Modèle adaptatif pour la prédiction de mots Adaptation à l'utilisateur et au contexte dans le cadre de la communication assistée pour personnes handicapées*  
 TAL. Volume 48 n° 2/2007

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.

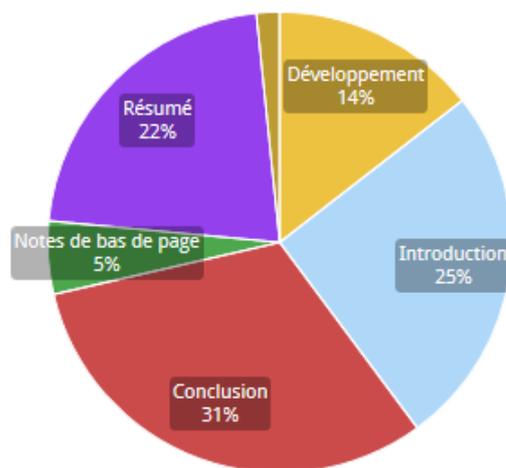
### Lemme

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

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