**1**

Travail de mémoire de recherche

**3**

Plan en 3 parties

**4**

Une bibliographie complètera

**6**

Haggan (2004) utilisation de phrases complètes est un trait majeur des titres de biologie

Plusieurs études (Haggan, 2004 ; Nagano 2015) constatent que les sciences dures et les sciences humaines forment deux blocs de disciplines qui se comportent de la même manière quelle que soit les propriétés étudiées : les sciences dures ont des titres plus longs, un taux de noms supérieur, et utilisent moins l’article défini *the* au début du titre que les sciences humaines.

**9**

Resserrement (narrowing)

**10**

Swales & Feak : Manuel d’écriture académique

Relation d’inclusion dans le 2e majeure : mineure

**11**

Le sujet n’est pas forcément un problème

**12**

Selon Huyghe (2015), *« les noms sont les items lexicaux privilégiés dans la réflexion générale sur la théorie sémantique et la structure du lexique »* et pour *« la construction du sens en contexte »*.

En explorant rapidement notre corpus de titres, nous avons remarqué que 84 % des segments après le double point avaient au moins deux noms.

**14**

Pas de différence fondamentale d’écriture entre les titres de types différents

**16**

Titres sans auteur, sans domaine

Nous écartons donc tous les titres en ayant plus (466) ainsi que ceux en ayant zéro (98 titres).

**17**

Stocké XML. Schéma XSD pour décrire le format.

**18**

Dans le cadre typologie reposant sur la fonction référentielle (Huyghe, 2015), on peut associer tous les noms de notre classe à des noms généraux = **un très faible contenu sémantique,** = pauvreté on ne sait pas vraiment à quoi fait référence une *approche* dans le monde réel.

permet en retour d’avoir une *« très large application référentielle »*  servir à dénoter énormément de référents, l’auteur parle de *« polyvalence référentielle »*

**19**

Une classe spécifique qu’on ne trouve qu’après le double point !

486 198 noms communs

224 400 avant double point soit 46 %,

261 798 après double point soit 54 %.

Donc 500 000 divisé par 2 en gros. 5000 = 1%

On peut aussi remarquer que la classe inversée, qui prendrait un taux de 70 % avant le double point et une fréquence minimum de 500 est réduite à un seul nom : *compte* qui compte 889 occurrences dont 71 % avant le double point.

**23**

Syntagme **immédiatement** après le double point

**25**

De … à … : nous y reviendrons

**33**

86 % pour SDV se répartissent entre 1 et 2

**38**

À quasi 50-50 entre « et » et « ou »

**40**

* Mythe ou réalité, Bilan et perspective
* Intérêt et limite, Théorie et application

approche phraséologique

Pour expliquer ces figements nous pouvons nous tourner vers la phraséologie, *« l’étude des séquences lexicales perçues comme préconstruites »* selon Legallois et Tutin (2013) sur lesquelles nous nous appuyons dans cette partie.

**41**

Reconstruire un syntagme à partir d’une liste d’étiquettes POS n’est pas une opération triviale

Une plage étroite de sable

Un tonneau plein de sable

Privilégier le modèle syntagmatique au modèle dépendanciel

**42**

Une amélioration conséquente à la recherche de structures et non plus de séquences linéaires et de s'affranchir de l’obligation d’avoir un ordre linéaire, une *« contrainte qui constitue une limite des approches classiques par segments répétés »* (Legallois & Tutin, 2013).

Patrons plus flexibles 🡺 analyse en dépendance, motif. Patron automatique ?

Basé sur structure et pas séquence

**44**

Legallois, D. & Tutin, A. (2013). Présentation : Vers une extension du domaine de la phraséologie. *Langages, 189*(1), 3-25.

Huyghe, R. (2015). Les typologies nominales : présentation. *Langue française, 185*, 5-27.

**Prévisions de questions**

**Why not CQL?**

Pour pouvoir avoir un contrôle fin sur le matching et faire des statistiques directement en Excel.

**Why not CONNELL / TEI P5 ou un autre format ?**

CONNELL : no XML

TEI P5 : pas adapté pour nos données

Notre format est XML 🡺 easy convertible > XSLT