Synthèse et questions

pour la réunion du 12 juin

Table des matières

[Introduction 1](#_Toc11081250)

[1ère partie : synthèse 2](#_Toc11081251)

[Le corpus de titres 2](#_Toc11081252)

[Segmentation des titres 2](#_Toc11081253)

[Catégorie du segment 2](#_Toc11081254)

[Catégorie du segment par domaine 3](#_Toc11081255)

[Études des racines 4](#_Toc11081256)

[Lemmes des racines 4](#_Toc11081257)

[Lemmes des racines par domaines 5](#_Toc11081258)

[Détermination des lemmes racines 7](#_Toc11081259)

[Complémentation des lemmes racines 7](#_Toc11081260)

[2ème partie : questions 8](#_Toc11081261)

[Questions 8](#_Toc11081262)

[Proposition d’échéancier 8](#_Toc11081263)

# Introduction

Ce document est une synthèse de notre travail pour en donner les éléments clés pour préparer la prochaine réunion de travail du mercredi 12 juin, 18h, à la maison de la recherche.

Nous avons réduit la part rédactionnelle pour n’y exposer que les questionnements successifs et leurs réponses découlant de notre problématique : **constructions centrales récurrentes dans les titres scientifiques**. « Construction » fait référence à la grammaire de construction, où niveaux lexical et syntaxique sont liés, pour apporter une sémantique qui est plus que la simple somme des composants de la construction. « Centrale » car nous nous intéressons à l’élément noyau du ou des syntagmes du titre, qui donne des informations sur la catégorie grammaticale du titre. « Récurrente » car nous voulons étudier les plus fréquentes, cette fréquence traduisant des habitudes et des pratiques, éventuellement variant selon les domaines scientifiques, que nous voulons mesurer et comprendre. Nous avons laissé dans le document les tableaux de résultats qui justifient l’avancée de notre cheminement scientifique, d’où son nombre de pages, mais nous pensons qu’il se lit rapidement en restant factuel. Un […] en bas d’un tableau indique que le tableau a été tronqué car le reste des lignes n’apportait pas d’informations significatives.

La première partie de ce document de synthèse se veut également représentatif d’un plan possible pour le document final, à discuter avec nos directeurs. Certaines parties non développées sont simplement signalées avec une flèche.

Dans la première partie, nous synthétisons notre démarche en rappelant les principaux apports. Dans la dernière partie, nous énumérons également des points à discuter pour cette réunion et rappelons l’échéancier proposé lors des séances de regroupement.

# 1ère partie : synthèse

## Le corpus de titres

### Segmentation des titres

339 687 titres en tout (avec auteurs, type de support, date, domaine scientifique) noté G.

171 890 constitué d’un segment (délimité par : ; . ? ! ) avec une racine unique (SG-1).  
 53 621 constitués de deux segments avec une racine dans le premier segment (SG-1:0).

30 554 constitués de deux segments avec une racine dans chacun (SG-1:1).

3 230 constitués de deux avec une racine dans le second segment (SG-0:1).

259 295 titres considérés dans mon corpus de travail (76 % du total)

Le corpus ne contient que des titres d’articles scientifiques ou de supports ayant des titres construits de manière similaire : chapitres d’ouvrages collectifs et communications dans des conférences. Tous les titres sont en français.

Nous entendons racine dans le sens de l’analyse syntaxique en dépendances : un mot uniquement régisseur.

Nous travaillons tout d’abord sur le sous-ensemble SG-1 qui est structurellement le plus simple. Nous voulons savoir la catégorie grammaticale du segment et donc du titre puisque celui-ci n’est composé que d’un seul segment.

### Catégorie du segment

Pour SG-1, on l’obtient en regardant la catégorie grammaticale de la racine du segment :

171 890 titres 100 % de SG-1

136 734 **NC** 76 % de SG-1

11 094 **NPP** 6 % de SG-1

8 186 **V** 4 % de SG-1

6 005 **P**  3 % de SG-1

5 135 **VINF** 3 % de SG-1

[…]

Si on regroupe les catégories nominales (NC et NPP), verbales (V, VIMP, VINF, VPP, VPR, VS) et prépositionnelles (P, P+D) :

171 890 titres 100 % de SG-1

147 828 **NOUN** 86 % de SG-1

15 749 **VERB**  9 % de SG-1

6 792 **PREP** 4 % de SG-1

Conclusion : 86 % des titres de SG-1 sont des groupes nominaux. Cela concorde avec la littérature.

### Catégorie du segment par domaine

On peut s’interroger sur la variation entre les domaines.

Notons que SG-1 est assez représentatif de l’ensemble du corpus des titres au niveau de la répartition par domaine :

**Dans G Dans SG-1 Position**

**Domaine Nombre % Nombre % Dans G Dans SG-1**

0.phys 35 538 10 % 24 688 14 %  **1 1**

1.shs.droit 30 517 9 % 16 679 10 % 3 2

1.shs.socio 32 228 9 % 13 739 8 % 2 3

0.sdv 24 638 7 % 13 150 8 % 6 4

NONE 25 115 7 % 12 585 7 %  **5 5**

0.info 16 932 5 % 11 435 7 % 8 6

1.shs.hist 25 764 8 % 11 434 7 % 4 7

1.shs.gestion 23 703 7 % 9 991 6 % 7 8

1.shs.ling 16 028 5 % 7 154 4 %  **9 9**

1.shs.archeo 13 745 4 % 6 704 4 % 11 10

1.shs.litt 14 938 4 % 6 356 4 % 10 11

En gras, les domaines ayant la même position dans G et SG-1.

Nous regardons la nature de la racine par domaine en pourcentage :

**Domaine POS 1 % POS 2 % POS 3 % Total**

0.phys NOUN 87 PREP 7 VERB 6 24 688

1.shs.droit NOUN 86 VERB 10 PREP 3 16 679

1.shs.socio NOUN 82 VERB 12 PREP 4 13 793

0.sdv NOUN 87 VERB 11 PREP 2 13 150

NONE NOUN 88 VERB 8 PREP 3 12 585

0.info NOUN 87 VERB 8 PREP 4 11 435

1.shs.hist NOUN 90 VERB 6 PREP 3 11 434

1.shs.gestion NOUN 82 VERB 14 PREP 3 9 991

1.shs.ling NOUN 83 VERB 10 PREP 5 7 154

1.shs.archeo NOUN 91 VERB 6 PREP 2 6 704

1.shs.litt NOUN 87 VERB 8 PREP 3 6 356

1.shs.edu NOUN 80 VERB 15 PREP 5 4 438

1.shs.phil NOUN 84 VERB 10 PREP 5 4 385

1.shs.scipo NOUN 83 VERB 12 PREP 4 3 998

1.shs.art NOUN 88 VERB 7 PREP 4 3 911

0.sde NOUN 85 VERB 11 PREP 3 3 494

1.shs.anthro NOUN 84 VERB 9 PREP 4 3 159

1.shs.infocom NOUN 82 VERB 11 PREP 6 3 073

0.math NOUN 86 VERB 7 PREP 6 2 240

1.shs.archi NOUN 82 VERB 12 PREP 5 2 119

0.sdu NOUN 88 VERB 7 PREP 3 1 944

0.chim NOUN 91 VERB 7 PREP 2 1 666

0.scco NOUN 84 VERB 11 PREP 5 1 509

1.shs.psy NOUN 84 VERB 11 PREP 4 1 286

1.shs.geo NOUN 85 VERB 11 PREP 3 509

0.qfin NOUN 84 VERB 14 PREP 2 196

1.shs.autre NOUN 90 VERB 8 PREP 2 48

Tous les domaines ont avant tout des groupes nominaux comme titres (de 80 à 91 %). Les catégories suivantes les plus fréquentes sont toujours les mêmes : verbes et prépositions, pour un syntagme verbal et un syntagme prépositionnel respectivement, sauf pour la physique ou les groupes prépositionnels sont légèrement plus fréquents que les groupes verbaux.

L’étude des racines nous a permis de déterminer la nature des titres. Mais les racines sont un sujet d’étude en elles-mêmes. Elles sont « au centre » du titre, de façon univoque lorsque celui-ci est constitué d’un seul segment, comme dans SG1.

## Études des racines

### Lemmes des racines

La grande majorité des titres étant des syntagmes nominaux, il convient de s’intéresser plus finement au nom noyau du syntagme, la racine. On s’interroge d’abord sur les lemmes les plus fréquents, nous prenons les 25 premiers sur les 15 692 lemmes différents, en attribuant un code couleur aux huit premiers :

**Pos Lemme Nombre % % cumulé**

01. étude 3855 2.61 % 2.61 %

02. analyse 1862 1.27 % 3.88 %

03. modélisation 1676 1.13 % 5.01 %

04. influence 1405 0.95 % 5.96 %

05. effet 1336 0.90 % 6.86 %

06. approche 1286 0.87 % 7.73 %

07. méthode 1058 0.72 % 8.45 %

08. modèle 1046 0.71 % 9.16 %

09. évolution 892 0.60 % 9.76 %

10. évaluation 884 0.60 % 10.36 %

11. caractérisation 837 0.57 % 10.93 %

12. mesure 834 0.56 % 11.49 %

13. droit 820 0.55 % 12.04 %

14. utilisation 805 0.54 % 12.58 %

15. rôle 749 0.51 % 13.09 %

16. recherche 703 0.48 % 13.57 %

17. impact 695 0.47 % 14.04 %

18. apport 664 0.45 % 14.49 %

19. contribution 632 0.43 % 14.92 %

20. construction 607 0.41 % 15.33 %

21. enjeu 588 0.40 % 15.73 %

22. développement 572 0.39 % 16.12 %

23. conception 554 0.37 % 16.49 %

24. représentation 543 0.37 % 16.86 %

25. application 538 0.36 % 17.22 %

Les 25 premiers noms les plus fréquents sont des noms abstraits appartenant au lexique scientifique. Nous également pouvons les analyser par domaine.

### Lemmes des racines par domaines

On fait cette même recherche par domaine (voir les résultats page suivante). On constate trois cas.

* Le premier cas (en couleurs) concerne les lemmes transdisciplinaires à la sémantique neutre vis-à-vis du domaine, comme « étude », qui se classe dans les cinq lemmes les plus fréquents dans 12 domaines sur 27, « analyse » (15/27), « modélisation » (6/27) que l’on peut rapprocher de « modèle », influence (3/27), effet (6/27) et approche (8/27).
* Le second cas (en gras) concerne les lemmes propres à un domaine particulier avec une sémantique univoque, comme « droit » pour le droit, « anthropologie » pour l’anthropologie, « philosophie » pour la philosophie, « architecture » pour l’architecture, « céramique » et « site » pour l’archéologie, « élections » pour les sciences politiques, « littérature » et « théâtre » pour la littérature, ou « vitrail », « musique » et « art » pour l’art. On ne retrouve pas ses lemmes de façon fréquente comme racines dans les autres domaines, ils sont exclusifs, et certains sont même directement le nom choisi pour représenter le domaine.
* Le troisième cas (en italique) concerne des lemmes non exclusifs à un domaine mais sans être autant transdisciplinaires que ceux du premier cas. Ils sont porteurs d’une sémantique en rapport avec un ou plusieurs domaines, citons par exemple « histoire » en histoire mais aussi en philosophie et en géographie, ou « ville » en sociologie et en architecture.

Notons quelques cas particuliers intéressants : « synthèse », le deuxième lemme le plus fréquent en chimie concerne la synthèse chimique d’éléments mais on peut le retrouver dans d’autres domaines dans une acceptation plus transdisciplinaire car sémantiquement plus large : la synthèse dans le sens d’un résumé. « politique » est le lemme le plus utilisé en sciences politiques qui étudient *la* politique, mais *une* politique peut désigner de façon beaucoup plus large tout ce qui a trait à une organisation en vue d’un but. Pour ces deux cas, qu’une acceptation relève du premier cas.

L’importance de la détermination pour « politique » nous amène à considérer de façon large la détermination des lemmes racines. Est-elle démonstrative, définie ou indéfinie ?

Les 5 lemmes les plus fréquents par domaine

**Domaine Lemme 1 % 1 Lemme 2 % 2 Lemme 3 % 3 Lemme 4 % 4 Lemme % 5 Titres**

0.phys étude 9.53 modélisation 3.96 mesure 2.36 analyse 2.18 influence 2.07 20987

1.shs.droit **droit** 4.58 chronique 2.26 responsabilité 1.48 protection 0.97 note 0.91 13664

0.sdv influence 4.90 étude 4.86 effet 4.69 utilisation 1.72 évolution 1.65 10725

NONE étude 2.87 modélisation 1.80 analyse 1.29 caractérisation 1.15 conception 0.79 10097

1.shs.socio *ville* 1.10 enjeu 1.01 évolution 0.96 dynamique 0.90 analyse 0.89 9908

0.info approche 3.10 modèle 2.71 modélisation 2.56 analyse 2.27 méthode 1.85 9665

1.shs.hist *histoire* 1.11 notice 0.54 inscription 0.47 question 0.46 femme 0.45 9515

1.shs.gestion analyse 1.89 impact 1.83 rôle 1.36 économie 1.18 évaluation 1.15 7305

1.shs.ling analyse 1.70 étude 1.61 construction 1.32 **langue** 1.22 représentation 0.95 5472

1.shs.archeo **céramique** 1.59 étude 1.55 **site** 1.35 apport 0.93 analyse 0.92 5348

1.shs.litt **littérature** 1.16 notice 0.82 **théâtre** 0.74 traduction 0.65 représentation 0.65 5266

1.shs.art **vitrail** 1.45 **musique** 1.36 **art** 1.27 notice 1.15 image 0.97 3304

1.shs.phil **philosophie** 1.35 *histoire* 1.16 science 1.13 critique 1.00 question 0.75 3194

1.shs.edu analyse 2.97 étude 1.79 effet 1.53 enseignement 1.34 évaluation 1.27 3069

1.shs.scipo politique 2.22 **élection** 1.26 enjeu 1.26 évolution 0.80 relation 0.80 3015

0.sde évaluation 2.55 brève 2.33 analyse 2.22 étude 1.99 modélisation 1.70 2708

1.shs.anthro **anthropologie** 1.04 corps 0.87 approche 0.75 étude 0.75 pratique 0.70 2412

1.shs.infocom analyse 1.71 communication 1.27 approche 1.23 pratique 1.10 représentation 0.88 2282

0.math estimation 4.46 modèle 2.62 analyse 2.57 modélisation 2.46 approche 2.06 1792

1.shs.archi **architecture** 2.57 *ville* 2.19 espace 1.75 projet 1.44 enjeu 1.06 1597

0.chim étude 7.41 synthèse 5.35 caractérisation 2.99 matériau 1.92 effet 1.92 1403

0.sdu étude 3.22 note 2.49 modélisation 2.42 apport 2.05 analyse 2.05 1365

0.scco étude 1.73 analyse 1.44 approche 1.44 évaluation 1.34 modèle 1.25 1043

1.shs.psy effet 2.82 étude 2.32 représentation 2.02 approche 1.71 analyse 1.51 992

1.shs.geo approche 2.06 *histoire* 1.80 évolution 1.54 politique 1.54 démographie 1.54 389

0.qfin déterminant 4.03 analyse 3.36 politique 3.36 effet 3.36 évaluation 2.01 149

1.shs.autre effet 12.20 influence 7.32 pratique 4.88 modèle 4.88 approche 4.88 41

* Suite à une légère correction, les chiffres varient de quelques dixièmes sans modifier l’ordre, tableau à mettre à jour.

### Détermination des lemmes racines

Nous nous restreignons, à l’intérieur de SG-1, à un sous-ensemble dont toutes les racines sont des noms communs, noté SG-1n. Nous avons analysé la détermination des racines dans SG-1n. Nous avons trouvé comme premiers résultats :

Racine sans détermination 87 512 59 %

Racine avec un déterminant 60 305 41 %

Racine avec deux déterminants 11 <1 %

Total 147 828 \_\_\_\_\_\_

La majorité des racines n’ont donc pas de déterminant. Parmi les déterminants des racines qui en ont, on trouve les formes suivantes :

**Déterminant Nombre % Cumul %**

La 15 023 24.90 24.90

Les 14 457 23.96 48.86

L' 10 908 18.08 66.94

Le 10 665 17.68 84.62

Une 2 785 4.62 89.24

Un 2 273 3.77 93.01

Des 1 238 2.05 95.06

Quelques 703 1.17 96.23

[…]

Pour les 60 305 racines ayant un déterminant, celui est défini à 85 %.

* Analyse transdisciplinaire
* Pluriel vs singulier
* Cas des racines avec deux déterminants (vérifier de visu que l’on peut les écarter)

En plus de la détermination, il est intéressant d’étudier la complémentation des racines.

### Complémentation des lemmes racines

Tout d’abord, nous étudions le nombre de compléments de la racine :

**Nombre de compléments Nombre %**

1 46620 31.54

0 43385 29.35

2 31747 21.48

3 15885 10.75

[…]

Nous observons que 29 % des racines n’ont pas de complément et 32 % en ont un seul. Nous regardons ensuite les prépositions introduisant le complément :

**Préposition Nombre %**

de 85 064 41.83

d' 25 336 12.46

dans 20 585 10.12

en 15 914 7.83

à 15 780 7.76

pour 11 060 5.44

sur 10 043 4.94

par 6 143 3.02

entre 2 815 1.38

chez 2 541 1.25

De/d’ compte pour 54 % des prépositions introduisant le complément.

* Analyse transdisciplinaire
* Cas des racines multi-complémentées
* Présence d’infinitif dans le complément (construction de + inf)

C’est ce dernier développement qui permettra éventuellement de relier nos racines aux constructions spécificationnelles de Legallois et Schmid. Si ce n’est pas le cas, nous nous rabattrons sur des schémas de la forme « X de Y » en détaillant les couples X et Y.

* Extension de cette réflexion aux sous-ensembles SG-1:0, SG-1:1, SG-0:1 avec des constructions sur plusieurs segments.
* La segmentation utilisant à la fois des marques de ponctuation segmentant des phrases et d’autres non, il est nécessaire d’introduire une distinction entre les segments correspondant à une phrase et les autres. Dans ces derniers, on pense notamment à SG-1:0 et SG-0:1, on rencontre des racines secondaires, elles sont régies uniquement par la racine se trouvant dans l’autre segment que celui de la racine secondaire.

Une fois SG-1 exploré, nous pouvons nous atteler aux autres sous-groupes de titres composés de titres ayant deux segments et entre une et deux racines.

# 2ème partie : questions

## Questions

1. Mon approche est-elle claire ? Partir d’une catégorisation des titres selon la segmentation pour trouver leur nature de syntagmes nominaux. Puis étudier plus avant la racine et les constructions/schémas dans lesquels elles s’inscrivent.
2. Mon approche va-t-elle dans la bonne direction ? Au final, je veux avoir une suite de construction (ou à défaut, des schémas), répartis selon les disciplines. Mon approche se distingue par l’utilisation de l’analyse syntaxique en dépendances pour construire ces constructions/schémas sous forme d’arbres.

## Proposition d’échéancier

**12 juin : réunion avec les directeurs**

L’objectif de cette réunion est de valider la délimitation de notre problématique et le cadre des réponses apportées. Il sera encore nécessaire de finaliser ces réponses, en émettant éventuellement de nouvelles hypothèses et en les vérifiant, en lançant des nouvelles requêtes sur notre corpus, en lisant d’autres articles ou en ajoutant un nouveau développement à nos outils, mais on veillera à rester dans le cadre arrêté par la réunion qui nous servira de fil conducteur pour la rédaction.

**Du 1er juin au 15 juillet (7 weekends) : rédaction du mémoire**

Plusieurs documents ont déjà été rédigés au cours de notre travail. Ils seront repris en partie pour constituer notre document final.

**15 juillet : rendu de la version initiale du document final**

*Du 15 juillet au 1er septembre, lecture par mes directeurs de la version initiale.*

**Du 1er au 15 septembre (3 weekends) : corrections et améliorations suite aux retours**

Suite aux retours de nos directeurs, une nouvelle version sera produite pour les prendre en compte.

**15 septembre : rendu de la version corrigée et améliorée du document final**

*Du 15 au 22 septembre, lecture par mes directeurs de la version corrigée et améliorée.*

**Du 23 au 30 septembre : soutenance**

éventuellement, une troisième version pourra être produite avant la soutenance si de nouveaux points à corriger ou à améliorer étaient soulevés par nos directeurs.