Projet de recherche : document du 1er semestre

Table des matières

[1. Problématique 1](#_Toc532807823)

[2. Actions effectuées 1](#_Toc532807824)

[3. Actions prévues 2](#_Toc532807825)

[3.1 Repérer les constructions 2](#_Toc532807826)

[3.2 Constituer un nouvel état de l’art 3](#_Toc532807827)

[4. Pistes et conseils 3](#_Toc532807828)

[5. Plan détaillé du document du 1er semestre 3](#_Toc532807829)

[Annexes 4](#_Toc532807830)

[A. Catégories du discours de Talismane 4](#_Toc532807831)

[B. Dépendances de Talismane 4](#_Toc532807832)

[C. Informations morphosyntaxiques de Talismane 4](#_Toc532807833)

[Bibliographie 4](#_Toc532807834)

# 1. Problématique

Notre travail part de l’utilisation des noms sous spécifiés, désormais NSS, et également appelés *shell nouns* par Schmid (2012), dans les titres d’article scientifiques en français.

Les NSS se caractérisent par un niveau de généricité très élevé et un contenu lexical très faible néanmoins non nul. Par exemple, le terme *problème* peut référencer un très grand nombre d’objet du monde réel et porte une sémantique d’obstacle et d’opposition.

Si on ne retrouve pas les deux patrons définitoires théorisés par Schmid (2012) dans les titres (déterminant + nom + that/wh-clause ou to-infinitive ou déterminant + nom + be + that/wh-clause ou to-infinitive), on retrouve néanmoins des occurrences de ces NSS.

Dans un titre au contenu si concis pourquoi utiliser un NSS et pour quels usages ? De plus, peut-on étendre ces réflexions à d’autres noms, non NSS, mais dotés de caractéristiques sémantiques proches, comme *outil*.

# 2. Actions effectuées

Le corpus des titres a été fourni par M. Tanguy et est tiré de l’archive ouverte HAL. Le socle technique d’interrogation du corpus est prêt, il a été créé en réutilisant certaines briques logicielles de l’an passé. Il a servi à mesurer le corpus dont les principales caractéristiques sont :

* 339 687 titres
* 4 9090 608 mots
* Une analyse complète des titres a été effectuée par Talismane qui fournit pour chaque mot :
  + Lemme, catégorie du discours, type de relation de dépendance et recteur, informations morphosyntaxiques additionnelles
* Les titres sont fournis avec les métadonnées suivantes :
  + Id, année, type du document, domaines, auteurs

Plusieurs articles sur les noms sous spécifiés, les titres et la littérature scientifique ont été lus et annotés, dont Hatier et al. (2016) et Schmid (2018).

# 3. Actions prévues

## 3.1 Repérer les constructions

Il s’agit de repérer des patrons d’usage des NSS à partir d’une **liste des NSS**. Certains NSS fournis dans l’ouvrage de Schmid (2012) peuvent être repris en les traduisant directement en français. Mme Ho-Dac et Mme Aleksandra Miletic travaillent également sur les NSS en français et donnaient un exposé lors d’une journée d’étude. Elles élaborent une liste que nous pourrions peut-être récupérer.

Il faut garder à l’esprit que certains termes, comme *outil*, *méthode*, *application*, s’ils ne sont pas des NSS, sont néanmoins très intéressants pour notre travail car ils possèdent eux aussi une grande capacité référentielle. Il faut réfléchir aux **critères d’inclusion** dans notre liste pour l’élargir à des non NSS. Dans le reste du document, on parlera de noms pivots (NP) et la **liste de noms pivots**.

Il faudra arbitrer sur la façon de présenter notre liste :

1. Est-ce une liste de NSS à laquelle on rajoute des membres, en justifiant ces rajouts en se fondant sur une analogie entre ces termes supplémentaires et les NSS ?
2. Ou doit-on établir a priori les contraintes d’appartenance à notre liste, contraintes respectées par les NSS et certains autres termes. La question de savoir si tous les NSS respectent ces contraintes se posera également.

Une fois une liste de noms pivots établie, il faut explorer le corpus pour faire émerger des patrons d’usage ou **constructions** utilisant ces mots pivots, d’abord par une exploration manuelle.

Nous cherchons également chercher à savoir si certains NP partagent des constructions. Pour trouver plus rapidement des constructions et savoir si nous avons des **constructions** **partagées** par plusieurs NP, on utilise l’algorithme itératif suivant :

* Établir une liste minimale de noms pivots nommée List(NP)
* Pour chaque terme N de List(NP) :
  + Repérer manuellement des constructions utilisant N en le projetant sur le corpus
  + Filtrer pour ne garder que les constructions fréquentes
  + Stocker ces constructions trouvées dans l’ensemble des constructions de N, Cons(N)
  + Pour chaque construction dans Cons(N)
    - Substituer N dans la construction par un joker.
    - Projeter la construction sur le corpus pour voir si d’autres termes l’utilisent
    - Pour chaque terme T ainsi identifié avec une fréquence suffisante :
      * Si T appartient déjà à List(NP), ajouter la construction à Cons(T) si elle n’y est pas
      * Si T n’appartient pas à List(NP), déterminer s’il respecte les critères d’inclusion et si oui, l’ajouter à List(NP) et initialiser Cons(T) avec cette construction

Il faudra déterminer **le seuil nécessaire** pour déterminer qu’une construction est fréquente.

Ces **constructions sont aussi bien syntaxiques que lexicales**. Quelques intuitions :

* *… : le problème de …*
* *ces X qui… / ces X de …*

Cette **substituabilité possible** des NP, entre *problème* et *cas* par exemple, devra être analysée finement.

Éventuellement, parmi ces constructions pourraient se trouver un **nouveau patron définitoire** pour les NSS, en plus de ceux théorisés par Schmid.

## 3.2 Constituer un nouvel état de l’art

Étape indispensable d’un projet de recherche, un état de l’art des questions soulevées doit être constitué à la fin du premier semestre. Il répertoriera les différents travaux déjà produits dans la littérature scientifique sur ces quatre questions :

* Les NSS
  + Définition
  + Usage
* Les noms à sémantisme faible et grande capacité référentielle
  + Critères de définition
  + Relations avec les NSS
* Les titres d’articles scientifiques
  + Certains articles sur la littérature scientifique sont aussi applicables, notamment ceux touchant au lexique employé.
* Les patrons lexico-syntaxiques

Cet état de l’art devra également traduire le **positionnement de mon travail** vis-à-vis de ces recherches.

Cet état de l’art s’insèrera dans un document de **15 à 20 pages** qui présentera également le sujet et un plan d’action pour le second semestre. Il servira d’évaluation pour le premier semestre.

Un plan détaillé de ce document est à envoyer aux codirecteurs avant les vacances de Noël pour valider les bases de ce travail. Nous en proposons une première version dans la partie 5 de ce document.

# 4. Pistes et conseils

Cette dernière partie évoque les conseils donnés et les pistes évoquées pour les étapes futures de mon travail :

* Il vaut mieux utiliser TXM plutôt qu’AntConc, ce dernier ne gère que les formes et pas les lemmes, ce qui est un sérieux défaut : ils comptent les formes fléchies *économique* et *économiques* du lemme ÉCONOMIQUEADJ de façon indépendante.
* Finir de lire les travaux de Dominique Legallois et Silvia Adler relatifs au NSS.
* Pour plus tard, il serait intéressant de construire la sémantique des constructions trouvées.

# 5. Plan détaillé du document du 1er semestre

**Introduction**

Problématique : pourquoi et comment utilise-t-on des termes avec un contenu sémantique faible dans l’espace très contraint que sont les titres d’articles scientifiques en français ?

1. **État de l’art**
   1. Sémantique faible et grande capacité référentielle

Sémantique propre faible car capacité référentielle trop grande ?

Cas des NSS

Autres termes

* 1. Les titres dans l’écriture scientifiques

Considérations lexicologiques

Considérations syntaxiques

Variations liées aux domaines

* 1. Patrons, motifs, schémas et constructions lexico-syntaxiques

1. **Données et outils**
   1. Présentation de HAL
   2. Présentation du corpus
   3. Présentation rapide des outils impliqués
   4. Présentation de la liste des noms pivots

Critères d’inclusions

1. **Analyses et résultats intermédiaires**
   1. Analyse lexicale des noms pivots

Présence dans les titres (1, 2, 3…)

* 1. Analyse syntaxique en dépendances
  2. Détection des constructions
  3. Substituabilité des termes et constructions communes

1. **Pistes et perspectives**

La sémantique derrière les constructions

La question d’une nouveau patron définitoire des NSS

**Conclusion  
Bibliographie**

# Annexes

## A. Catégories du discours de Talismane

ADJ, ADV, ADVWH, CC, CLO, CLR, CLS, CS, DET, DETWH, ET, I, NC, NPP, P, P+D, P+PRO, PONCT, PRO, PROREL, PROWH, V, VIMP, VINF, VPP, VPR, VS

## B. Dépendances de Talismane

\_, a\_obj, aff, arg, ato, ats, aux\_caus, aux\_pass, aux\_tps, comp, coord, de\_obj, dep, dep\_coord, det, mod, mod\_rel, obj, p\_obj, pont, prep, root, sub, suj

## C. Informations morphosyntaxiques de Talismane

g pour genre (m ou f), n pour nombre (s ou p), p pour personne (1, 2, 3, 12, 23, 123), poss pour nombre du possesseur (s ou p), t pour temps (pst pour présent, past pour passé, imp pour imparfait, fut pour futur, cond pour conditionnel).

# Remarques de Tanguy

Concernant votre compte-rendu, je voulais juste préciser (mais ça date un peu) qu'on avait surtout envisagé que vous étudierez les classes de noms "génériques" que l'on trouve à certaines positions dans les titres, en suivant la logique générale des NSS mais sans qu'il y ait nécessairement de recouvrement entre les deux ensembles. Les marqueurs que vous évoquez pour les NSS ont peu de chance d'être productifs sur vos données, alors que ceux qu'on avaient envisagés et testés à la volée l'étaient bien plus (sur le modèle de votre étude de M1 sur les noms de tête derrière les deux points).  
  
Concernant le plan détaillé, a priori rien à redire à ce stade de développement, sauf peut-être le fait qu'il faudrait éviter d'avoir une partie "état de l'art", et préférer une intégration plus fluide dans votre problématique.  
  
Avant la rédaction finale que nous évaluerons, vous pouvez nous envoyer un court texte de synthèse (type 1 page d'introduction) pour nous permettre de vérifier que nous sommes raccord sur la vision d'ensemble (hypothèses, objectifs et méthodes envisagées).

# Bibliographie

Hatier, S., Augustyn, M., Tran, T. T. H., Yan, R., Tutin, A., & Jacques, M. P. (2016). French cross-disciplinary scientific lexicon: extraction and linguistic analysis. In Proceedings of Euralex (pp. 355-366).

Schmid, H. J. (2012). *English abstract nouns as conceptual shells: From corpus to cognition* (Vol. 34). Walter de Gruyter.

Schmid, H. J. (2018). Shell nouns in English-a personal roundup. *Caplletra. Revista Internacional de Filologia*, (64), 109-128.