Approches outillées pour l'étude des noms sous-spécifiés ou noms capsules

Lydia-Mai Ho-Dac, Aleksandra Miletic, Marine Wauquier, Cécile Fabre

Approches outillées pour l'étude des noms sous-spécifiés ou noms capsules

Lydia-Mai Ho-Dac, Aleksandra Miletić, Marine Wauquier, Cécile Fabre

1. Les noms capsules : critères définitoires

Les noms qui nous intéressent dans cette étude sont des noms comme *fait*, *problème*, *objectif*, *idée*... qui sont souvent classés dans la catégorie des noms abstraits du fait de leur sémantisme sous-spécifié, générique, vague. Cette première caractéristique se double d'un fonctionnement spécifique en discours : ces noms ont tendance à apparaître dans des « constructions spécificationnelles » (Legallois 2006), à fonctionner comme des « noms à complément propositionnel » (Riegel 1996). Les exemples (1) et (2)[1] illustrent les deux constructions prototypiques auxquelles ces noms sont associés. Dans la première, le nom est lié à une infinitive (une complétive en *de*), dans le second à une conjonctive (une complétive en *que*).

> (1) [N + infinitive] […] Mon but n'est pas de jeter la pierre ou quoi que ce soit, je ne vois simplement pas ce qui te gêne dans [le *fait de qualifier* ainsi cet organisme]. (*WikiDisc*)
> (2) [N + conjonctive] […] [L'*idée que* cette population ne devait pas boire uniquement du vin lors des inhumations mais aussi dans la vie courante] a été avancée. (*WikiArt*)

Dans les exemples (1) et (2), le sens des noms *fait* et *idée* est spécifié (en quelque sorte « rempli ») par le contenu propositionnel des complétives. C'est cette propriété qui permet de définir ces noms comme des « coquilles conceptuelles » (*shell nouns*, Schmid 2000) ou « noms porteurs » (Huyghe 2018b), termes évoquant l'idée que ces noms vont se charger d'un contenu plus ou moins complexe en contexte. Dans les exemples précédents, le contenu de la complétive est encapsulé et étiqueté dans et par le nom, que nous choisissons finalement de qualifier, en reprenant une proposition de Richard Huyghe, de « nom capsule » (2018a).

Ce phénomène d'encapsulation ne se réduit pas au cas des noms à complément propositionnel. On voit dans l'exemple (3) qu'il peut englober, dans un empan de texte plus large, une série de faits énumérés (*prise de pouvoir illégale, absence de légitimité, etc.*) sous des labels (*raisons* et *aspects*) qui caractérisent sémantiquement les faits énumérés et peuvent également servir de référence discursive mobilisable dans la suite du texte.

> (3) Comme tu l'as probablement remarqué, l'introduction détaille de manière neutre et formelle [toutes les *raisons* associées au concept dictatorial] : prise de pouvoir illégale, absence de légitimité, pratiques à caractères arbitraires et sanguinaires ou encore aspect autoritaire de l'exercice des fonctions. Mieux, elle précise que les partisans de Pinochet ne nient pas [ces *aspects* des choses] mais le justifient par un intérêt supérieur. (*WikiDisc*)

Ce dispositif linguistique est défini par R. Huyghe, reprenant Schmid (2000 : 367-369), comme un procédé de « réification nominale » permettant de « conceptualiser » une portion de texte « comme un objet en soi » (Huyghe 2018b : 41-46). Nous appellerons ces constructions des constructions « encapsulantes » pour les distinguer des constructions « spécificationnelles » définies plus haut.

Comme l'illustre l'exemple (3), la réification nominale peut s'opérer vers l'avant, dans un mouvement de prospection (le contenu du nom *raisons* est spécifié *a posteriori* par l'énumération, cf. Francis 1994), ou vers l'arrière, dans un mouvement de reprise (Conte 1996), également appelé anaphore résumante (le contenu du nom *aspects* est chargé du contenu propositionnel véhiculé par l'ensemble de l'énumération). Cette capacité à encapsuler - annoncer ou résumer une portion de texte - donne à ces noms un rôle particulier dans l'organisation du discours en constituant un dispositif linguistique efficace pour « empaqueter » l'information (Chafe 1976). C'est en ce sens que Flowerdew et Forest (2015) proposent d'appeler ces noms *signalling nouns*, les définissant d'emblée comme des indices de cohésion plutôt que comme des éléments lexicaux simplement caractérisés par un sémantisme vague et un comportement syntaxique et discursif particulier.

En français, plusieurs études se sont intéressées à ces noms du point de vue de leur fonctionnement discursif : Legallois (2006) analyse la manière dont ces marqueurs participent aux structures d'enchaînement ; Roze et al. (2014) intègrent ces noms dans une méthode de détection automatique de « séquences organisationnelles, telles que problème-solution » ; Rebeyrolle et Péry-Woodley (2014) étudient leur rôle dans la signalisation des structures énumératives.

La variété terminologique que l'on constate est révélatrice de la diversité des points de vue sur ces noms. Deux points de vue se dégagent :

- un intérêt sémantique pour des noms au sémantisme vague : *unspecific nouns* (Winter 1992), *general nouns* (Halliday and Hasan 1976), noms super-ordonnés ou « sommitaux »[2] (Kleiber 2014), noms évaluatifs « en attente de valeur » (Blanche-Benveniste 1992) ;
- un intérêt lié au rôle discursif de certains noms présentant une capacité à « empaqueter » du contenu propositionnel (cf. Chafe 1976), à se charger référentiellement d'un contenu propositionnel : *anaphoric nouns* (Francis 1986),

[1] Tous les exemples présentés sont issus des deux corpus d'étude, WikiDisc et WikiArt, décrits en section 2.

[2] Ce terme est lié au fait que ces noms « soit occupent le sommet des hiérarchies, soit présentent une généralité et une abstractivité fonctionnelles très grandes. » (Kleiber 2014, note 4)

lexical signalling (Hoey 1979), *container nouns* (Vendler 1968), *shell nouns* (Schmid 2000), encapsulation (Conte 1996), noms-coquilles, noms-capsules (Huyghe 2018a, 2018b).

Cette différence de points de vue est intimement liée à la diversité des approches adoptées pour identifier ces noms capsules. Certaines études se situent dans une approche lexicale hors contexte qui tente de délimiter un lexique des noms capsules, en maintenant au second plan l'idée que certains mots porteraient en eux cette caractéristique discursive. D'autres travaux se situent dans une approche sémantique en contexte soutenant l'hypothèse que ce sont les situations syntaxiques et discursives qui dotent certains noms des propriétés du nom capsule. Nous explorons dans cet article les moyens d'étudier les noms en corpus selon cette double approche.

2. Notre approche : deux méthodes et deux corpus

L'objectif de ce travail est de faire émerger de l'analyse outillée de corpus la liste des noms présentant les propriétés des noms capsules. Nous mettons successivement en œuvre pour cela deux méthodes : la première (section 3) s'appuie sur le profil distributionnel des noms pour faire émerger une classe de noms sémantiquement homogène ; la deuxième (section 4) identifie les noms qui entrent dans les structures définitoires de la classe des noms capsules : les constructions spécificationnelles et encapsulantes. Ces deux points de vue sur le fonctionnement sémantique des mots requièrent des approches outillées différentes, qui relèvent de deux grandes familles de démarches pour l'analyse sémantique en linguistique de corpus : l'approche distributionnelle et l'approche par patrons syntaxiques. Nous montrons la contribution de chacune à l'objectif que nous nous sommes fixé. Nous faisons par ailleurs l'hypothèse que cette liste de noms ne peut pas être définie de façon totalement stable et *a priori*, mais qu'elle varie au contraire d'un genre de textes à l'autre, hypothèse que nous allons tester en menant l'expérience à partir de deux corpus différents.

Les deux corpus que nous utilisons sont issus de l'encyclopédie collaborative Wikipédia. Cette ressource offre l'avantage de fournir pour de nombreuses langues, dont le français, des données libres d'accès, de grande taille, couvrant un grand nombre de thématiques. Elle est limitée en termes de variété textuelle, mais donne néanmoins accès à deux genres de textes différents, correspondant à deux versants de l'écriture collaborative : les articles, qui constituent la face visible de Wikipédia, et les pages de discussion, sorte de « coulisses » de l'encyclopédie qui offrent un espace d'échanges pour réguler la démarche de rédaction participative (Cardon et Levrel 2009). Le tableau 1 donne un aperçu quantitatif des deux corpus d'étude : le corpus WikiArt est issu des articles contenus dans la version française de Wikipédia à la date du 1er octobre 2018 et fournit des textes principalement descriptifs et explicatifs ; le corpus WikiDisc est composé des textes argumentatifs issus des interactions entre les contributeurs qui participent à l'écriture des articles. WikiDisc contient toutes les interactions qui ont eu lieu dans les pages de discussion de la version française à la date du 20 octobre 2018.

corpus	**nombre de mots**	**nombre de noms communs**	**nombre de textes**	**genre**
wikiArt	925 515 503	946 903	2 256 568	descriptif et explicatif
wikiDisc	213 286 460	273 365	439 638	argumentatif

Tableau 1: Aperçu quantitatif des deux corpus d'étude

Les deux corpus offrent un contraste entre deux types d'écrits qui nous permet d'étudier la variation du vocabulaire des noms capsules selon le genre. Les deux méthodes que nous expérimentons requièrent un pré-traitement différent de ces corpus : comme nous allons le voir, l'approche distributionnelle peut être appliquée à un corpus brut, ou simplement lemmatisé. En revanche, l'approche par patrons, dans la mesure où elle s'intéresse à des indices morphosyntaxiques complexes et variés, nécessite une annotation morphosyntaxique et syntaxique. Ces aspects sont détaillés à l'occasion de la présentation de chaque méthode.

3. Approche sémantique des noms capsules

3.1. Sémantique distributionnelle automatique

L'approche distributionnelle du sens est définie par Harris (1954) comme une corrélation entre similarité sémantique et similarité distributionnelle : deux mots sont considérés comme proches sémantiquement s'ils partagent une large proportion de leurs contextes. Cette proximité est estimée empiriquement en recueillant les occurrences de ces mots dans des corpus et en comparant leurs environnements distributionnels. Cette hypothèse a suscité de nombreuses implémentations en traitement automatique des langues (Lenci 2018). Celles-ci consistent à construire des modèles distributionnels à partir de corpus, dans lesquels les mots sont représentés par des vecteurs de contextes. Le tableau 2 présente un exemple fabriqué et très simplifié d'une matrice à 3 lignes et 6 colonnes. Ici, chaque ligne correspond à un mot, chaque colonne à un contexte syntaxique, et chaque cellule à la fréquence absolue de la cooccurrence du mot et du contexte considéré. Dans cet espace matriciel, les contextes deviennent des dimensions et les fréquences absolues fournissent les coordonnées de chaque mot dans chaque dimension. Cette représentation permet de modéliser la distribution d'un mot sous la forme d'un vecteur défini par des coordonnées à 6 dimensions, ainsi *parcours* a comme localisation dans cet espace (33,5,92,146,9,25). Les modèles distributionnels réels sont calculés sur des corpus de très grande taille. Ils peuvent donc afficher des millions de dimensions (une pour chaque contexte).

	objet de *baliser*	sujet de *traverser*	complément du nom *étape*	complément du nom *incident*	modifié par *spirituel*	modifié par *libre*
parcours	33	5	92	146	9	25
itinéraire	10	6	18	0	38	0
voyage	0	0	137	0	13	0

Tableau 2 : Vecteurs simplifiés pour 3 mots cibles et 6 contextes

Une fois cette représentation calculée, la proximité sémantique entre mots est envisagée comme une proximité géométrique entre les vecteurs qui les représentent. Ces vecteurs peuvent être comparés en utilisant des mesures mathématiques classiques comme le cosinus, qui permet de calculer un score de similarité distributionnelle. On a donc accès à une mesure empirique graduelle de la proximité sémantique, qui ne préjuge pas de la nature de la relation sémantique impliquée, mais dont de multiples travaux ont montré la validité pour modéliser des phénomènes lexicaux variés, comme la synonymie, les préférences sélectionnelles, le changement sémantique (Baroni et Lenci 2010, Boleda 2019).

Ce principe a donné lieu à des propositions d'implémentation très variées. Les modalités de calcul de ces vecteurs ont considérablement évolué au cours des vingt dernières années (Fabre et Lenci 2015). De nombreux paramètres ont été considérés pour construire les vecteurs de contextes : la richesse de l'information contextuelle encodée (simples cooccurrences ou relations de dépendance syntaxique), la pondération des contextes (recours à des mesures d'association plutôt qu'à une fréquence), la taille de la fenêtre de contexte considérée, les seuils de fréquence définis pour filtrer les contextes, le nombre de dimensions prises en compte pour réduire les matrices et ainsi limiter la complexité des traitements, etc. (Baroni et Lenci 2010, Turney et Pantel 2010). Depuis peu, de nouveaux modèles, dits prédictifs, se sont imposés comme une alternative au décompte explicite des contextes. Basés sur les réseaux neuronaux, ces algorithmes apprennent à prédire automatiquement à partir de larges corpus les contextes d'un mot donné, et génèrent directement des matrices de dimension réduite plus faciles à manipuler (on parle de *word embeddings*). Ces modèles se sont imposés malgré l'opacité de la représentation produite, qui n'offre plus de prise sur les contextes originels.

De nombreux modèles sont aujourd'hui disponibles et leurs performances varient assez nettement selon les tâches considérées et les paramètres utilisés (Levy et al. 2015). Dans cette expérience, nous avons eu recours à l'outil Word2Vec (Mikolov et al. 2013) dont l'apport a été démontré sur une large gamme de tâches lexicales. Nous avons généré un modèle distributionnel par corpus, à partir d'une version lemmatisée, en fixant certains paramètres de l'algorithme comme la fréquence minimum des mots à prendre en compte (ici 50), la fenêtre de contexte considérée (ici 7 mots) et d'autres paramètres spécifiques aux réseaux de neurones : le nombre d'exemples négatifs (ici 5) et le nombre de dimensions du modèle (soit le nombre de colonnes dans la matrice, ici 300). Ces paramètres influant sur la nature du modèle généré (Pierrejean et Tanguy 2018), une étude plus systématique nécessiterait de répliquer l'expérience en les faisant varier afin de ne conserver que l'intersection des modèles résultants.

3.2. Construction de la classe des noms capsules par proximité distributionnelle

La capacité des modèles distributionnels à capter un large éventail de relations sémantiques a été maintes fois démontrée, en particulier dans des études consacrées au repérage de relations lexicales comme la synonymie ou l'hyperonymie (Baroni et Lenci 2010). Récemment, des travaux ont montré que la similarité distributionnelle recoupe la catégorisation des lexèmes en termes d'abstraction et de concrétude (Frassinelli et al. 2017). En d'autres termes, les voisins distributionnels des mots concrets sont eux-mêmes des mots concrets, et il en va de même pour les mots abstraits. Cette approche offre donc une piste intéressante pour l'extraction de noms abstraits à partir de corpus.

Afin de constituer une classe de noms abstraits susceptibles de présenter les propriétés des noms capsules, notre approche consiste à amorcer sa constitution à partir de quelques noms sous-spécifiés, considérés comme prototypiques. L'hypothèse est que nous pourrons alors faire émerger selon le critère de proximité distributionnelle des noms sémantiquement proches qui seront de bons candidats à la catégorie des noms capsules. En première approximation, nous avons choisi 7 noms régulièrement cités dans les travaux relatifs aux noms capsules, et dont nous estimons qu'ils figurent de façon incontestable dans la classe : *décision*, *fait*, *idée*, *possibilité*, *problème*, *question*, *raison*. Chacun de ces mots est représenté par un vecteur dans le modèle distributionnel. Un score de proximité entre vecteurs est fourni par Word2Vec, correspondant à la valeur du cosinus, qui varie de 0 (proximité nulle) à 1 (proximité parfaite). Cette valeur fournit une estimation de la proximité sémantique entre mots. À titre d'exemple, le mot *question* a pour plus proches voisins les mots *problématique* (0,70), *problème* (0,64), *interrogation* (0,6) dans le corpus WikiArt. Dans le corpus WikiDisc, les plus proches voisins de ce mot sont *quesiton* (graphie erronée), *problématique* et *pdd* (sigle pour *page de discussion*), avec des valeurs plus basses, respectivement 0,48, 0,45 et 0,43.

Pris isolément, chacun de ces 7 noms apporte des voisins distributionnels spécifiques, qui ne renseignent pas notre objectif de recherche d'une classe générique. Ainsi, le mot *possibilité* a très logiquement parmi ses premiers voisins des mots comme *permettre*, *possible*, ou *pouvoir*. Nous avons donc cherché à accéder à une sorte de représentation moyenne de ces mots en calculant un profil distributionnel partagé par l'ensemble des 7 noms amorces. La traduction mathématique de cette opération consiste à calculer le vecteur moyen de ces 7 vecteurs. Cette opération fournit un vecteur théorique, censé subsumer l'information contextuelle commune à ces vecteurs amorces. Nous suivons en cela une démarche déjà définie par Kintsch (2001) pour la prédication verbale, et qui a été appliquée dans le cadre d'une étude menée sur la suffixation (Wauquier et al. 2018). Le vecteur résultant fournit un point théorique dans l'espace vectoriel, autour duquel nous allons concentrer nos recherches.

À partir du vecteur moyen que nous avons calculé, nous produisons pour chacun des deux corpus la liste de ses plus proches voisins, en choisissant arbitrairement de retenir les 50 premiers. Le tableau 3 en montre un sous-ensemble, soit les 20 premiers voisins pour chaque corpus. Les mots figurant parmi les 7 mots amorces utilisés pour créer le vecteur moyen sont indiqués en gras. Il est normal de les trouver en bonne position, puisque l'information contextuelle les concernant a servi à construire ce vecteur prototypique. Dans le cas de WikiDisc, on notera néanmoins que le mot *fait* n'apparaît pas dans le tableau, car il figure seulement au 30e rang des voisins. Les surlignements signalent les noms qui sont absents de la liste des 50 premiers voisins générée à partir de l'autre corpus. Enfin, les astérisques mentionnent les noms répertoriés par l'étude de Legallois et Gréa (2006). Nous commentons ces deux caractéristiques plus loin.

Corpus WikiArt	Corpus WikiDisc
question	**raison***
problème*	**question**
fait*	**possibilité**
idée*	**idée***
problématique	*objection**
possibilité	difficulté*
motivation*	**problème***
raison*	**décision***
*conséquence**	*position**
argument*	hypothèse*
décision*	problématique
conclusion*	notion
nécessité	conclusion*
sujet	argument*
notion	considération
explication*	nécessité
solution*	volonté*
principe*	obligation*
proposition	option
situation	proposition

*Tableau 3 : Liste des 20 premiers voisins pour chaque corpus - *intersection avec Legallois et Gréa (2006)*

Afin de fournir une représentation visuelle plus complète des résultats, nous proposons sur la figure 1 un graphique rassemblant les 50 voisins les plus proches du vecteur moyen (signalé ici par le terme « barycentre ») obtenus pour le corpus WikiDisc.

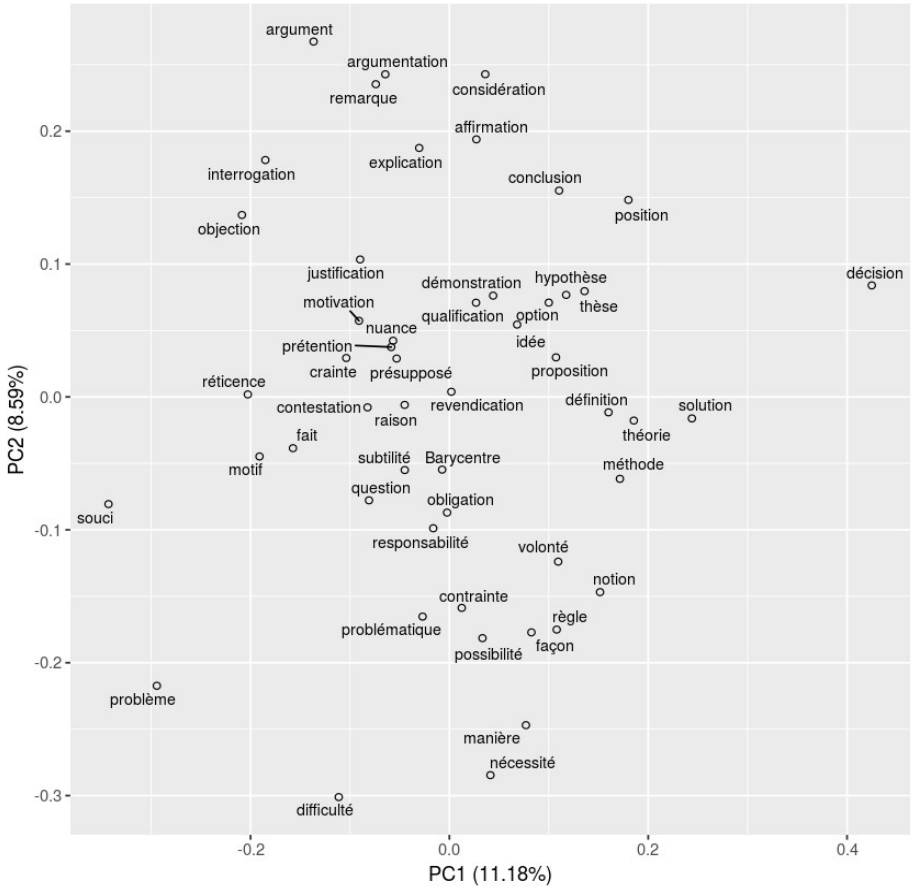


Figure 1 : Visualisation des 50 voisins du vecteur moyen pour le corpus WikiDisc

Notons que du fait de la projection des informations sur 2 dimensions uniquement (alors que nous avons un modèle à 300 dimensions), la visualisation ne donne qu'une approximation du résultat, ce qui produit par endroits une restitution erronée des valeurs initiales : les voisins les plus proches visuellement ne correspondent pas nécessairement aux scores de proximité les plus élevés. Ainsi, à gauche du point matérialisant le barycentre, *subtilité* est positionné plus près de celui-ci que *question* ; or, *subtilité* est situé au rang 37 des voisins, avec un score de proximité de 0,44, et *question* au rang 2, avec un score de 0,67.

3.3. Analyse des résultats

Plusieurs questions se posent à l'issue de cette première étape et guident l'observation des résultats : obtient-on, à partir de chaque corpus, une liste de noms abstraits ? S'agit-il de bons candidats pour la classe des noms capsules ? Observe-t-on une variation entre les 2 corpus ?

L'annotation des 100 mots issus des 2 listes de voisins distributionnels permet de répondre à la première question : tous les voisins sont des noms abstraits, à l'exception d'une anomalie : au 37e rang des voisins du vecteur moyen construit sur le corpus WikiArt figure la conjonction *donc*. Ce résultat démontre ainsi, comme nous en avions fait l'hypothèse, la capacité de cette méthode à extraire des noms qui partagent avec les noms amorces le trait abstrait. Cependant, rien ne nous permet d'affirmer que la proximité distributionnelle va préserver d'autres propriétés sémantiques de ces mots, et en particulier leur propension à être utilisés comme mots capsules. L'objectif de notre méthode est précisément d'utiliser un critère supplémentaire pour nous en assurer (section 4). Nous pouvons dès à présent entamer une première analyse en observant le degré d'intersection de notre liste avec une liste de noms sous-spécifiés établie par des travaux antérieurs. Nous avons choisi d'utiliser comme étalon une liste de noms proposée par Legallois et Gréa (2006) : il s'agit de 305 noms illustrant, sans visée exhaustive, les classes nominales qui apparaissent dans des constructions spécificationnelles au sein du corpus que ces auteurs ont analysé (corpus d'une année du *Monde*). On constate qu'une majorité des voisins distributionnels (58% pour WikiArt, 57,5% pour WikiDisc) apparaissent également dans la liste de Legallois et Gréa. C'est le cas des noms marqués d'un astérisque dans le tableau 3. L'observation des mots sans astérisque révèle qu'ils peuvent également être de bons candidats pour la catégorie visée. En effet, certains des voisins absents de la liste de référence peuvent rejoindre facilement une des classes nominales mises au jour par Legallois et Gréa. C'est le cas par exemple de *problématique* (classe PROBLÈME), *nécessité* (classe CONTRAINTE), ou *notion* (classe IDÉE). Par ailleurs, tous les noms passent le test consistant à les intégrer dans une construction encapsulante.

Si l'on s'intéresse dans un deuxième temps à la comparaison entre les deux corpus, la confrontation des deux listes constituées des 50 premiers voisins suggère des variations assez importantes, même si la majorité des noms sont partagés

(58% dans les deux cas). Il est intéressant de repérer l'orientation argumentative du corpus WikiDisc dans la liste des noms qui lui sont propres : *objection*, *position* (surlignés dans le tableau 3), *obligation*, *option*, *responsabilité*, *remarque*, *justification*, *réticence*, *qualification*, *prétention*, s'agissant des 10 premiers voisins absents de WikiArt. Le vocabulaire propre à WikiArt correspond à des noms abstraits plus divers : *conséquence*, *sujet* (surlignés dans le tableau 3), *principe*, *situation*, *réponse*, *discussion*, *affirmation*, *procédure*, *pratique*, *pertinence*.

Cette expérience pourrait être affinée de différentes manières. Outre la modification de certains paramètres dans le calcul initial du modèle distributionnel, il serait en particulier intéressant de reproduire cette méthode en étendant le nombre de mots amorces utilisés pour la construction du vecteur moyen. Ces premiers résultats nous permettent néanmoins de dresser un premier bilan : ils confirment la pertinence de l'approche distributionnelle pour extraire facilement un ensemble de noms ayant les propriétés caractéristiques des noms capsules. Mais ils montrent également une limite, souvent signalée, de l'approche distributionnelle. Celle-ci est fondée sur un critère très général et graduel de proximité distributionnelle. Si les résultats produits sont très intéressants lorsqu'on examine les premiers lexèmes candidats - à savoir les voisins distributionnels les plus proches du vecteur moyen que nous avons construit - ils se révèlent logiquement assez vite bruités lorsque l'on parcourt la suite de la liste. On rencontre alors des noms qui, bien qu'abstraits, ne semblent pas présenter le comportement attendu. C'est par exemple le cas des mots *pertinence* (rang 32 dans la liste des voisins de WikiArt) ou *réticence* (rang 29 dans la liste des voisins de WikiDisc). On peut donc considérer cette approche comme une première étape utile qu'il s'agit de combiner avec l'observation de critères complémentaires. Nous présentons dans la section suivante une deuxième direction de recherche, fondée sur une observation plus locale et contrôlée du comportement de ces noms dans les discours.

4. Approche fonctionnelle des noms capsules

Cette deuxième approche est plus classique pour l'étude des noms capsules. Elle repose sur l'une des propriétés évoquées plus haut, à savoir leur capacité à apparaître dans des structures syntaxiques spécifiques qui présentent un certain degré de figement et peuvent être considérés comme des constructions au sens de Goldberg (2006) ou de Legallois & François (2006). Deux types d'unités phraséologiques peuvent être considérées : les constructions spécificationnelles, par lesquelles le nom capsule incorpore le contenu de son complément propositionnel (exemples (1) et (2)) et les constructions encapsulantes, où le contenu incorporé par le nom est cette fois exprimé dans un segment de texte indépendant syntaxiquement (exemple (3)). Notre méthode consiste à identifier dans les corpus WikiArt et WikiDisc ces structures et, dans un deuxième temps, à recourir à des mesures pour filtrer, au sein de tous les noms ramenés, ceux qui présentent les propriétés les plus caractéristiques d'un emploi de nom capsule.

Une expérience de ce type a déjà été menée par Legallois (2008), qui analyse un corpus constitué de tous les articles publiés en 1995 dans le quotidien *Libération* et y identifie une liste de 361 termes définis comme « noms sous-spécifiés » car apparaissant fréquemment dans les constructions spécificationnelles modélisées par la formule donnée en (4), où NC signifie « nom commun » et (X | Y) indique une alternative entre la forme X ou Y.

(4) [Det NC (Ø | ce) être (que-clause | de-inf)]

L'idée (c') (est) que cette population ...
L'idée (c') (est) de qualifier ...

Selon le modèle de la grammaire des constructions dans lequel s'inscrit Legallois, ces constructions prototypiques des noms sous-spécifiés seraient en quelque sorte « stockées en mémoire, et donc convoquées lors de leur énonciation, et non produites *on line* à partir de règles combinatoires générales. » (Legallois et Gréa, 2006)

Roze et al. (2014) prolongent ce travail en effectuant une détection automatique de ces deux constructions spécificationnelles dans un grand corpus de presse, extrayant au total 1670 noms et adjectifs nominalisés. Cette liste s'avère bruitée, suggérant que le critère d'apparition dans une construction n'est en lui-même pas assez discriminant. À titre d'illustration, Roze et al. (2014) fournissent la liste des 20 noms les plus fréquemment associés à la construction spécificationnelle décrite en (4) :

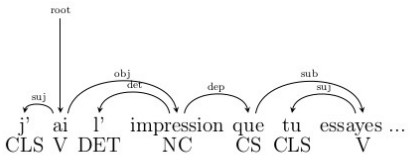
(5) *objectif, problème, but, question, idée, rôle, ambition, mission, essentiel, enjeu, intérêt, priorité, important, risque, difficulté, chose, souci, solution, mérite, vérité*

Afin de gérer ce bruit en filtrant les résultats bruts de l'extraction, différentes mesures ont été proposées, dont la *Dépendance* utilisée par Schmid (2000) et par la suite Legallois et Gréa (2006), qui consiste à calculer le pourcentage d'occurrences d'un NC qui apparaissent dans une construction spécificationnelle. Cette mesure permet de distinguer, d'une part, les NC qui montrent une fréquence fortement dépendante de la fréquence d'apparition des patrons retenus ; et, d'autre part, les NC qui montrent une fréquence élevée dans mais également en dehors des patrons. Seuls les premiers seraient alors à considérer comme des noms capsules. Nous avons choisi d'appliquer à nos données cette même méthode d'extraction associant patrons puis filtres en lui ajoutant la prise en compte des patrons associés aux constructions encapsulantes (voir section 4.2.2). Cela suppose que les corpus utilisés soient dotés d'une annotation syntaxique.

4.1. Analyse en dépendances syntaxiques

Les corpus WikiArt et WikiDisc sont tous les deux dotés d'une analyse syntaxique en dépendances (Tesnière 1959, Mel'čuk 1988). Dans ce type de représentation syntaxique, devenue un standard pour l'analyse syntaxique automatique

(Zeman et al. 2017, Nivre et al. 2016), la structure de la phrase se construit à partir des formes fléchies individuelles, reliées entre elles par des relations de dépendance (cf. exemple 6).
(6)



Chaque dépendance est orientée du gouverneur vers le dépendant, le gouverneur étant la forme qui légitime la présence du dépendant dans la phrase. Ainsi, dans l'exemple (6), la dépendance [obj] qui désigne un objet est orientée du verbe *avoir* vers le nom *impression*. Il en est de même pour la dépendance [dep] entre le nom *impression* et la conjonction *que*. Cet exemple illustre également deux contraintes que doivent respecter les arbres syntaxiques en dépendances : la complétude (tout mot de la phrase doit avoir un gouverneur) et l'unicité du gouverneur (tout mot ne peut avoir qu'un seul gouverneur).

L'annotation syntaxique de nos corpus a été effectuée à l'aide de l'analyseur automatique Talismane (Urieli, 2013). Cet outil est basé sur un algorithme d'apprentissage automatique entraîné sur le corpus annoté manuellement French Treebank (Candito et al. 2010a). En parcourant le corpus d'entraînement, l'algorithme apprend un modèle statistique qui permet, face à un nouveau corpus non annoté, de choisir l'analyse la plus probable pour chaque mot. Le modèle de Talismane utilisé dans ce travail atteint une exactitude de 86,9 % à 88,0 % quand il s'agit de déterminer les dépendances et les fonctions syntaxiques qu'elles portent, alors que son résultat se situe entre 89,5 % et 90,4 % si l'on se limite seulement à l'identification des dépendances sans se soucier des fonctions syntaxiques (Urieli, 2013, p. 154). Bien que ces résultats aient été comparables à l'état de l'art en analyse syntaxique automatique sur le français au moment de la parution de l'outil (Candito et al., 2010b), ils ne sont pas parfaits. Toute application basée sur cette annotation peut donc potentiellement être affectée par des erreurs d'analyse. Néanmoins, l'utilité d'une annotation syntaxique, même imparfaite, reste incontestable pour pouvoir mener des analyses systématiques sur de gros corpus.

4.2 Détection des patrons et identification des noms capsules

4.2.1. Projection des patrons syntaxiques

L'annotation en liens de dépendances décrite ci-dessus permet une détection à la fois plus précise et plus large des patrons qu'une simple recherche de séquences morpho-syntaxiques (ex : un nom suivi de la conjonction *que* sans certitude d'une quelconque relation entre eux), généralement pratiquée par les méthodes traditionnelles (Schmid 2000). Notre méthode permet ainsi (1) d'élargir la détection à des contextes où les éléments constitutifs du patron sont éloignés dans la phrase tout en étant reliés par des relations syntaxiques de longue distance (ainsi les cas de coordinations et de subordonnées) ; (2) de filtrer les contextes bruités ramenés par les patrons dans les cas où les éléments n'entretiennent pas de relation syntaxique.

Les constructions spécificationnelles étudiées dans les travaux de Legallois (2006), Legallois et Gréa (2006) et Roze et al. (2014) ont été retenues et formalisées en deux patrons se déclinant chacun en trois variantes (notées a, b, c) selon le degré d'intégration de la complétive au syntagme nominal :

- Spe_que : patron [Det NC[3] (Ø | ce) (Ø | être[4]) que-clause] où la complétive en *que* est reliée (a) par une relation de dépendance syntaxique au NC, (b) par une relation d'attribut du sujet au NC, ou (c) par une relation d'attribut du sujet au pronom *ce* dans une construction disloquée à gauche, avec ou sans présence de virgule ;
- Spe_de : patron [Det NC (Ø | ce) (Ø | être) de-inf] où la complétive en *de* est reliée (a) par une relation de dépendance syntaxique au NC, (b) par une relation d'attribut du sujet au NC, ou (c) par une relation d'attribut du sujet au pronom *ce* dans une construction disloquée à gauche, avec ou sans présence de virgule ;

Deux patrons de construction encapsulante viennent en complément :

- Enc_ponct : patron [NC_pluriel+:] ou [NC+(suivant)+:] qui correspond à la recherche de syntagmes nominaux suivis de près[5] par la ponctuation « : » et dont le nom est soit au pluriel, soit modifié par l'adjectif *suivant* ; ce patron correspond aux cas de *prospection* ;
- Enc_demo [ce+NC_sujet] où le NC est sujet et déterminé par un déterminant démonstratif.

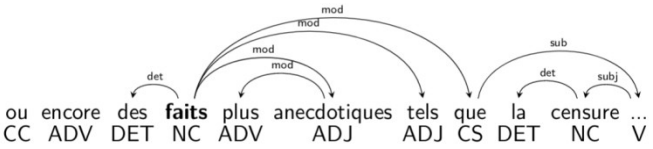
La détection de ces patrons repose sur l'analyse automatique des étiquettes attribuées par Talismane à chaque mot du corpus. Comme indiqué dans la section précédente, l'analyse automatique en dépendances peut être erronée et nécessiter une adaptation des patrons afin de réduire le bruit et le silence. L'exemple (7) montre un exemple de résultat bruité : une construction en *tel que* a été confondue avec la construction spécificationnelle Spe_que [Det *fait* que-clause]

[3] Seuls les noms déterminés ont été pris en compte (ce qui évite de considérer les contextes comme « Merci de modifier ... », très fréquents dans WikiDisc).

[4] La mention *être* désigne ici le lemme du verbe *être*, c'est-à-dire toutes ses formes fléchies, en excluant les cas où il est auxiliaire.

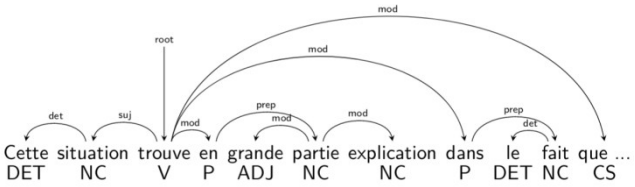
[5] « De près » signifie ici à moins de six éléments de distance – mots ou signe de ponctuation, sans possibilité de verbe intermédiaire.

(7)



Le silence concerne les constructions pertinentes mais qui, à cause d'une erreur d'analyse syntaxique, ne sont pas détectées, comme dans (8) où la complétive a été rattachée à tort au verbe (*Cette situation trouve que...*) plutôt qu'au NC *fait*.

(8)



Les résultats de la projection montrent que tous les patrons retenus sont largement présents dans les deux corpus. Presque trois millions de patrons ont été détectés (environ 70% dans WikiArt et 30% dans WikiDisc) et ont ramené 21 602 NC différents dans WikiArt et 10 742 NC différents dans WikiDisc. Le tableau 4 indique, pour chacun des quatre patrons, sa fréquence d'apparition (F_a) dans WikiArt (WA) et WikiDisc (WD), sa fréquence par 10 000 mots (F_r), le nombre de NC différents extraits (NC) et les NC les plus fréquemment identifiés, accompagnés pour les trois premiers et le dernier de leur fréquence dans le patron.

		F_a	F_r	NC	NC les plus fréquents (Freq.)
Spe_que	WD : WA :	108 058 249 222	5 3	6 470 13 410	WD : *fait (12194), problème (6138), article (4700), source, fois, point, idée, avis, raison, page, chose, argument, question (780)* WA : *fait* (7868), *problème* (3874), *fois* (3874*), taux, idée, but, raison, nombre, article, différence, objectif, résultat, nom, partie, point (931)*
Spe_de	WD : WA :	53 891 148 038	3 2	2 728 6 304	WD : *fait (5762), but (3420), façon (2435), question, idée, raison, moyen, manière, peine, volonté, objectif, intérêt (922)* WA : *but* (12843), *objectif* (*9862), fait (5355), idée, façon, moyen, manière, possibilité, volonté, mission, décision, droit, rôle (2116)*
Enc_ponct	WD : WA :	131 184 514 058	6 6	5 728 13 399	WD : *changement (20580), point (5324), source (4433), article, remarque, raison, chose, critère, phrase, question, page (1428)* WA : *enfant (9634), commune (6750), type (6173), partie, manière, catégorie, année, article, forme, domaine, propriété (2863)*
Enc_demo	WD : WA :	231 962 395 694	11 4	6 921 12 210	WD : *article (50528), page (7987), source (3354), information, phrase, genre, terme, section, liste, personne, point, paragraphe (1901)* WA : *module (35799), article (10900), espèce (3408), type, genre, nom, page, projet, terme, système, film, modèle, nombre, personne (1782)*

Tableau 4 : Résultats de la projection des patrons de noms capsules (F_a : fréquence absolue, F_r : fréquence relative sur 10 000 mots, NC : nombre de NC différents)

Si nous comparons tout d'abord les performances des différents types de constructions, nous observons que, de façon globale, les constructions encapsulantes sont plus fréquentes que les constructions spécificationnelles. Cependant, le nombre de NC différents ramenés est similaire. En d'autres termes, les deux types de construction montrent une forte différence en termes de diversité lexicale, avec une redondance beaucoup plus forte dans le cas des constructions encapsulantes. Par ailleurs, la liste des trois NC les plus fréquents montre peu de recouvrement entre les constructions. Si l'on reprend la classification proposée par Legallois et Gréa (2006), on observe que chaque construction semble attirée par des classes distinctes : Spe_que par la classe PROBLÈME et IDÉE, Spe_de par la classe OBJECTIF, Enc_demo et Enc_ponct par les noms sommitaux. En revanche, le recouvrement entre les deux corpus est assez fort.

Le tableau 5 complète le précédent en fournissant le détail des résultats concernant les variantes des constructions spécificationnelles.

	wikiDisc			wikiArt		
	freq	NC	NC les plus fréquents (freq)	freq	NC	NC les plus fréquents (freq)
Spe_que(a) *_que*	36 534	3 690	*fait* (27,9), *article* (2,5), *fois* (2,5)	95 943	9 337	*fait* (7,1), *fois* (3,4), *idée* (0,7)
Spe_que(b) *être_que*	64 080	5 110	*article* (5,8), *problème* (5,7), *fait* (3)	148 179	10 230	*problème* (1,7), *taux* (1,2), *but* (0,9)
Spe_que(c) *ce_être_que*	7 444	1 181	*problème* (31,2), *chose* (3,7), *différence* (3,2)	5 100	1 243	*problème* (22,7), *différence* (2,9), *chose* (2,6)
Spe_de(a) *_de*	37 534	2 488	*fait* (15,3), *façon* (5,9), *raison* (3,9)	98 926	6 070	*fait* (5,4), *façon* (3,8), *possibilité* (3,1)
Spe_de(b) *être_de*	14 995	830	*but* (20,5), *question* (10,8), *objectif* (5,8)	48 018	1 527	*but* (23,5), *objectif* (19,8), *mission* (4,5)
Spe_de(c) *ce_être_de*	1 362	310	*but* (9), *chose* (5,4), *problème* (5,1)	1 094	270	*idée* (7,1), *but* (6,8), *chose* (5,1)

Tableau 5 : Résultats de la projection des variantes des constructions spécificationnelles

Certains NC apparaissent fréquemment dans chacune des variantes d'un patron, comme l'illustrent les exemples suivants des variantes de Spe_que avec le NC *problème*.

(8) ***Le problème que*** *tu soulèves est le même que …* [wikiDisc]
(9) *Maintenant,* ***le problème*** *avec les souris nourries durant la nuit biologique* ***est que*** *le taux de leptine …* [wikiDisc]
(10) ***Le problème, c'est que****, contrairement à l'État, la notion de « pays" n'a …* [WikiArt]

4.2.2 Filtrer les noms capsules

Ces premiers résultats font émerger des NC qui ne sont pas tous des noms capsules. Parmi eux, nous retrouvons des NC très fréquents des corpus comme *article, source, page* ou *changement* pour WikiDisc et *article, commune, enfant* pour WikiArt[6], ainsi que le mot *fois* identifié précédemment. On retrouve l'idée que les constructions n'ont pas, en elles-mêmes, de pouvoir discriminant assez fort.

Un premier filtre tient compte de la variété (*breadth* selon Jones et al. 2007) des constructions dans lesquelles le NC apparaît. Nous décidons de sélectionner les NC qui apparaissent dans les quatre constructions retenues. Dans WikiDisc, 1 752 NC sur 10 742 sont retenus (soit 16 %), et 4 208 NC sur 21 602 dans WikiArt (soit 19,5 %). Ce filtre est complété par le score de *dépendance,* utilisé par Adler & Legallois (2018) reprenant Schmid (2000), qui mesure la propension d'un NC à apparaître dans un patron. Il s'agit d'un rapport entre la fréquence d'apparition d'un nom dans un patron et la fréquence d'apparition totale du nom dans le corpus. Sa valeur est donc de 100% si le NC apparaît exclusivement dans le patron considéré. Enfin, nous adoptons le même seuil de fréquence minimum des mots à prendre en compte (50, cf. section 3.1). Les listes suivantes présentent par ordre décroissant de dépendance, pour chaque construction, les 10 NC retenus par la combinaison de ces trois filtres. À titre d'indication, le score de dépendance du premier et du dixième NC est donné entre parenthèses.

Spe_que[WikiArt] : *désavantage* (4,8)*> inconvénient > particularité > probabilité > hypothèse >**fait**> taux >**problème**> avantage > supposition > originalité* (2)
Spe_que[WikiDisc] : *ennui* (25,3)*> idéal >**fait**>**problème**> probabilité > avantage > souci > inconvénient > bémol > différence (3)*
Spe_de[WikiArt] : *im**possibilité***(13,6)*> objectif > nécessité > manie > permis > volonté >**possibilité**> désir > fureur > refus > envie* (5,5)
Spe_de[WikiDisc] : *manie* (18,7)*> but > moyen > nécessité > im**possibilité**> volonté > peine > souhait > objectif >**possibilité** (6)*
Encaps_ponct[WikiArt] : *synonyme* (8,6)*> segment > critère > dérivé > prérequis > inconvénient > manière > enfant > variante > propriété > affluent* (3,2)
Encaps_ponct[WikiDisc] : *changement* (58,7)*> motif > remarque > patrouilleur > écueil > une > extincteur > bizarrerie > point > adresse (3,6)*
Encaps_demo[WikiArt] : *module* (37,6)*> ratio > affirmation > dernier > supposition > allégation > monsieur > chiffre > évènement > phénomène > hypothèse* (2,5)

[6] Les termes *commune* et *enfants* sont liés à la surreprésentation dans l'encyclopédie d'articles biographiques (« *de son mariage avec X sont issus les **enfants** suivants : »*) et de descriptions géographiques (*Nîmes est entourée des **communes** suivantes :*).

Encaps_demo[WikiDisc] : *genre* (6)> *sobriquet* > *évènement* > *assertion* > *manie* > *affirmation* > *incident* > *appellation* > *qualificatif* > *type* (4)

Un premier regard sur ces listes confirme l'efficacité de notre méthode de classement. Comme remarqué précédemment, les constructions encapsulantes semblent associées à une palette de noms assez différente de celle des constructions spécificationnelles et semblent être davantage influencées par le corpus. Les noms dépendant de constructions encapsulantes correspondent soit à des noms sommitaux (*genre*, *événement*, *phénomène*, *affirmation*) soit à des termes génériques propres au corpus (*patrouilleur*, *enfant, affluent*), complétant ainsi les noms dépendant de constructions spécificationnelles que l'on peut classer dans les catégories PROBLÈME, IDÉE et OBJECTIF.

Il est également intéressant de noter que la classe des noms amorces utilisés avec la méthode distributionnelle (en gras dans les listes) montre un fort degré de dépendance avec les seules constructions spécificationnelles, et ce dans les deux corpus. Il paraîtrait alors pertinent de recalculer un vecteur prototypique en prenant en compte des noms capsules dépendant de constructions encapsulantes. Un avant-goût de cette nouvelle liste de noms amorces est donnée dans les deux listes ci-dessous, résultant de la fusion de tous les patrons.

[WikiArt] : *im**possibilité*** (54,3) > *intention* >***possibilité***> *nécessité* > *module* > *opportunité* > *particularité* > *manie* > *souhait* > *permission* > *habitude* (28,6)
[WikiDisc] : *impression* (64,7)> *changement* > *manie* > *mérite* > *im**possibilité***> *nécessité* > *peine* >***possibilité***> *moyen* > *volonté* (32,9)

5. Conclusion

Nous avons présenté deux approches outillées pour l'extraction de noms sous-spécifiés, ou noms capsules, à partir de deux corpus de genre différent. La première a consisté à utiliser un modèle de sémantique distributionnelle pour constituer une liste de noms abstraits, au sémantisme général, dont le profil distributionnel se rapproche de noms capsules prototypiques utilisés comme amorces. La seconde nous a amené à nous focaliser sur la faculté qu'ont ces noms d'intégrer des structures spécificationnelles et encapsulantes. L'examen des résultats produits par ces deux méthodes a permis de montrer leurs apports et leurs limites. Chacune permet de dégager un ensemble de lexèmes candidats pertinents si l'on contraint suffisamment les seuils des mesures utilisées (score de proximité sémantique et indicateurs statistiques). Mais dans les deux cas, les limites de l'ensemble dégagé sont impossibles à poser de façon catégorique afin d'assurer des résultats optimaux en termes de précision et de rappel. Un examen manuel s'impose pour filtrer les propositions recueillies. Une alternative, qui constitue la prochaine étape de cette étude, consiste à croiser les deux approches pour constituer une liste de noms répondant simultanément à ces deux types de critères, afin de concilier le double point de vue sémantique et fonctionnel que nous venons de présenter. L'objectif sera de faire émerger des noms présentant à la fois un profil sémantique de nom abstrait général, et des propriétés syntaxico-discursives spécifiques.

Lydia-Mai Ho-Dac, Aleksandra Miletić, Marine Wauquier, Cécile Fabre
CLLE, Université de Toulouse et CNRS
hodac@univ-tlse2.fr, aleksandra.miletic@univ-tlse2.fr, marine.wauquier@univ-tlse2.fr, cfabre@univ-tlse2.fr

Bibliographie

ADLER, Silvia & LEGALLOIS, Dominique (2018), « Les noms sous-spécifiés dans le débat parlementaire : analyse fréquentielle et catégorisation modale », *Langue française*, 198(2), pp. 19-34.

BARONI , Marco & LENCI, Alessandro (2010), « Distributional memory : A general framework for corpus-based semantics », *Computational Linguistics*, 36(4), pp. 673-721.

BLANCHE-BENVENISTE, Claire (1992), « Sur un type de nom *évaluatif* portant sur des séquences verbales », *ITL-International Journal of Applied Linguistics*, 97(1), pp. 1-25.

BOLEDA, Gemma (2019). Distributional Semantics and Linguistic Theory. arXiv preprint arXiv : 1905.01896.

CANDITO, Marie, CRABBÉ, Benoît & DENIS, Pascal (2010a), « Statistical French dependency parsing: treebank conversion and first results », *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC2010), pp. 1840-1847, Valetta, Malta.

CANDITO, Marie, NIVRE, Joakim, DENIS, Pascal & ANGUIANO, Enrique Henestroza (2010b), « Benchmarking of statistical dependency parsers for French », *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (COLING2010), pp. 108-116, Beijing, China.

CARDON, Dominique, & LEVREL, Julien. (2009), « La vigilance participative. Une interprétation de la gouvernance de Wikipédia », *Réseaux*, 2, pp. 51-89.

CHAFE, Wallace (1976), « Givenness, contrastiveness, definiteness, subjects, topics and point of view », In Charles N. Li (eds.), *Subject and Topic*, New York, Academic Press, pp. 27-55.

CONTE, Maria-Elisabeth (1996), « Anaphoric encapsulation », *Belgian Journal of linguistics*, 10(1), pp. 1-10.
FABRE, Cécile. & LENCI, Alessandro (2015), « Distributional Semantics today », *Traitement Automatique des Langues* (*TAL*), 56(2), pp. 7-20.
FLOWERDEW, John & FOREST, Richard (2015), *Signalling Nouns in English : A Corpus-Based Discourse Approach*, Cambridge, Cambridge University Press.
FRANCIS, Gill (1986), « Anaphoric nouns », *Discourse Analysis Monographs*, 11, Birmingham, University of Birmingham Printing Section.
FRANCIS, Gill (1994). « Labelling discourse : an aspect of nominal-group lexical cohesion », In Coulthard, M. (ed.) *Advances in written text analysis*, Routledge, pp. 83-101.
FRASSINELLI, Diego, NAUMANN, Daniela, UTT, Jason & IM WALDE, Sabine Schulte (2017), « Contextual Characteristics of Concrete and Abstract Words, *IWCS 2017—12th International Conference on Computational Semantics—Short papers*, Montpellier, France.
GOLDBERG, Adele, (2006), *Constructions at work : the nature of generalization in language*, Oxford, Oxford University Press.
HALLIDAY, Michael & HASAN, Ruqaiya (1976), *Cohesion in English,* Longman, London.
HARRIS, Zellig (1954), « Distributional structure », *Word*, 10(2-3). pp. 146-162.
HOEY, Michael (1979), « Signalling in discourse », *Discourse Analysis Monographs*, 6, Birmingham, University of Birmingham Printing Section.
HUYGHE, Richard (2018a), « Noms généraux et noms sous-spécifiés : des relations à préciser », *Communication lors des journées d'étude S'Caladis : Les noms sous-spécifiés en français : du lexique au discours*, 15-16 novembre 2018, Université de Toulouse Jean Jaurès.
HUYGHE, Richard (2018b), « Généralité sémantique et portage propositionnel : le cas de *fait* », *Langue Française,* 198, pp. 35-50.
JONES, Steven, PARADIS, Carita., MURPHY, Lynne, & WILLNERS, Caroline (2007). Googling for 'opposites': A Web-based study of antonym canonicity. *Corpora*, *2*(2), pp. 129-155.
KINTSCH, Walter. (2001), « Predication », *Cognitive science*, 25, pp. 173-202.
KLEIBER, Georges (2014), « Lorsque l'opposition *massif / comptable* rencontre les noms superordonnés », *Travaux de linguistique*, 69(2), pp. 11-34.
LEGALLOIS, Dominique (2006), « Quand le texte signale sa structure : la fonction textuelle des noms sous-spécifiés », *Corela*, HS-5.
LEGALLOIS, Dominique. (2008), « Sur quelques caractéristiques des noms sous-spécifiés », *Scolia*, 23, pp. 109-127.
LEGALLOIS, Dominique, & FRANÇOIS, Jacques (2006), « Autour des grammaires de construction et de patterns », *Cahier du CRISCO*, 21.
LEGALLOIS, Dominique, & GREA, Philippe (2006), « L'objectif de cet article est de… Construction spécificationnelle et grammaire phraséologique », *Cahiers de praxématique*, 46, pp. 161-186.
LENCI, Alessandro (2018), Distributional models of word meaning, *Annual review of Linguistics*, 4, pp. 151–171.
LEVY, Omer, GOLDBERG, Yoav, & DAGAN, Ido (2015). « Improving distributional similarity with lessons learned from word embeddings ». Transactions of the Association for Computational Linguistics, 3, pp. 211-225.
MEL'ČUK, Igor (1988), Dependency syntax: Theory and practice, State University Press of New York, New York, USA.
MIKOLOV, Tomas, YIH, Wen-tau. T., & ZWEIG, Geoffrey (2013), « Linguistic regularities in continuous space word representations », In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pp. 746-751, Atlanta, Georgia, USA.
NIVRE, Joakim (2008), « *Algorithms for deterministic incremental dependency parsing* », Computational Linguistics, MIT Press, 34, pp. 513-553.
NIVRE, Joakim, DE MARNEFFE, Marie-Catherine., GINTER, Filip, GOLDBERG, Yoav, HAJIC, Jan, MANNING, Christopher, MCDONALD, Ryan, PETROV, Slav, PYYSALO, Sampo, SILVEIRA, Natalia, TSARFATY, Reut, & ZEMAN, Daniel (2016), Universal Dependencies v1: A multilingual treebank collection, *Proceedings of the 10th International Conference on Linguistic Resources and Evaluation (LREC2016)*, pp. 1659–1666, Portorož, Slovenia.
PIERREJEAN, Bénédicte & TANGUY, Ludovic. (2018). « Étude de la reproductibilité des *word embeddings* : repérage des zones stables et instables dans le lexique », In Actes de la conférence TALN, Rennes.
REBEYROLLE, Josette & PÉRY-WOODLEY, Marie-Paule (2014), « Énumération et structuration discursive », In *Actes du 4e Congrès Mondial de Linguistique Française*, pp. 3183-3196, Berlin, Allemagne.
RIEGEL, Martin (1996), « Les noms à compléments propositionnels : en quoi sont-ils plus abstraits que d'autres ? », In N. Flaux, Nelly, M. Glatigny & D. Samain (eds), *Les noms abstraits, Histoire et théories*, Villeneuve d'Ascq, Presses Universitaires du Septentrion, pp. 313-322.
ROZE, Charlotte, CHARNOIS, Thierry, LEGALLOIS, Dominique, FERRARI, Stéphane & SALLES, Mathilde (2014), « Identification des noms sous-spécifiés, signaux de l'organisation discursive », *Actes de la 21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, Marseille, France.
SCHMID, Hans-Jörg (2000), *English abstract nouns as conceptual shells : From corpus to cognition*, 34, Walter de Gruyter.
TESNIÈRE, Lucien (1959)*, Eléments de syntaxe structurale*, Paris, Klincksieck.
TURNEY Peter & PANTEL Patrick (2010), « From frequency to meaning : Vector space models of semantics », *Journal of artificial intelligence research*, 37(1), pp. 141-188.
URIELI, Assaf (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit* (Doctoral dissertation, Université Toulouse le Mirail-Toulouse II).

VENDLER, Zeno (1968). *Adjectives and Nominalizations,* The Hague, Mouton.
WAUQUIER, Marine, FABRE, Cécile & HATHOUT, Nabil. (2018). « Différenciation sémantique de dérivés morphologiques à l'aide de critères distributionnels ». *Congrès Mondial de Linguistique Française (CMLF)*, 46, EDP Sciences, Mons, Belgique.
WINTER, Eugene (1992), The notion of unspecific versus specific as one way of analysing the information of a fund-raising letter », In W.C. Mann & S.A. Thompson (eds.) *Discourse Descriptions: diverses Analyses of a Fund-Raising Letter*, Amsterdam, John Benjamins, pp. 131-170.
ZEMAN, Daniel, POPEL, Martin, STRAKA, Milan, HAJIC, Jan, NIVRE, Joakim, GINTER, Filip, LUOTOLAHTI, Juhani, PYYSALO, Sampo, PETROV, Slav, POTTHAST, Martin *et al*. (2017), CoNLL 2017 shared task : multilingual parsing from raw text to Universal Dependencies *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 1-19.

Résumé

Cet article s'intéresse au repérage en corpus des noms sous-spécifiés ou *noms capsules*. Ces noms se caractérisent par des propriétés à la fois sémantiques - ce sont des noms généraux, abstraits - et fonctionnelles - ils encapsulent sous une forme nominale un contenu propositionnel plus ou moins large. Nous adoptons deux approches outillées complémentaires, l'une s'intéressant à la proximité distributionnelle des mots de cette classe, l'autre à leur capacité à intégrer des patrons lexico-syntaxiques spécifiques. Nous montrons les apports et les limites de chacune des approches pour repérer cette classe de mots et en évaluer le comportement dans deux corpus contrastés : l'un constitué des articles de l'encyclopédie en ligne Wikipedia, l'autre des discussions associées à ces articles.

Abstract

This paper focuses on the detection of unspecific nouns, also called shell nouns, in corpora. These nouns are generally characterized by their semantic properties - they are general, abstract nouns - and at the same time by their ability, on a more functional level, to encapsulate a propositional content under a single nominal form. We adopt two complementary approaches to perform the extraction of shell nouns: we first use automatic distributional techniques to circumscribe the class on the basis of shared contexts in corpora, then we examine their ability to integrate specific lexico-syntactic patterns. We show the contributions and limits of each approach to identify this class of words and we evaluate their behaviour by comparing two corpora made up of Wikipedia articles and discussions.

Mots clés

noms sous-spécifiés, noms abstraits, linguistique outillée, patrons morpho-syntaxiques, modèles distributionnels

Keywords

shell nouns, abstract nouns, corpus-based methods, morpho-syntactic patterns, distributional models