# Identification des noms sous-spécifiés, signaux de l'organisation discursive

Charlotte Roze[1, 2] Thierry Charnois[3] Dominique Legallois[2] Stéphane Ferrari[1]
Mathilde Salles[2]
(1) GREYC, Université de Caen Basse-Normandie, Campus 2, 14000 Caen, France
(2) CRISCO, Université de Caen Basse-Normandie, Campus 1, 14000 Caen, France
(3) LIPN, Université Paris 13 Sorbonne Paris Cité, 93430 Villetaneuse, France
{charlotte.roze,dominique.legallois, stephane.ferrari, mathilde.salles}@unicaen.fr,
thierry.charnois@lipn.univ-paris13.fr

**Résumé.** Dans cet article, nous nous intéressons aux *noms sous-spécifiés*, qui forment une classe d'indices de l'organisation discursive. Ces indices ont été peu étudiés dans le cadre de l'analyse du discours et en traitement automatique des langues. L'objectif est d'effectuer une étude linguistique de leur participation à la structuration discursive, notamment lorsqu'ils interviennent dans des séquences organisationnelles fréquentes (e.g. le patron *Problème-Solution*). Dans cet article, nous présentons les différentes étapes mises en oeuvre pour identifier automatiquement ces noms en corpus. En premier lieu, nous détaillons la construction d'un lexique de noms sous-spécifiés pour le français à partir d'un corpus constitué de 7 années du journal *Le Monde*. Puis nous montrons comment utiliser des techniques fondées sur la fouille de données séquentielles pour acquérir de nouvelles constructions syntaxiques caractéristiques des emplois de noms sous-spécifiés. Enfin, nous présentons une méthode d'identification automatique des occurrences de noms sous-spécifiés et son évaluation.

**Abstract.** In this paper, we focus on *shell nouns*, a class of items involved in the signaling of discourse organisation. These signals have been little studied in Natural Language Processing and within discourse analysis theories. The main goal is to study their participation to discourse organisation, especially when they occur in Problem-Solution patterns. In this paper, we present the different steps involved in shell nouns identification of these nouns. First, we present the lexical acquisition of shell nouns from a large corpus. Second, we show how a method based on the extraction of sequential patterns (sequential data mining techniques) allows to discover new syntactic patterns specific to the use of shell nouns. Finally, we present a shell nouns identification system that we evaluate.

**Mots-clés :** noms sous-spécifiés, motifs séquentiels, structure discursive.

**Keywords:** shell nouns, sequential patterns, discourse structure.

## 1 Introduction

Le travail que nous présentons s'inscrit dans le cadre de l'analyse de l'organisation discursive et de l'étude des indices linguistiques de cette organisation. Parmi ces indices, une classe d'items a reçu jusqu'ici peu d'attention en Traitement Automatique des Langues et dans les théories d'analyse du discours telles que la RST (Mann et Thompson, 1988) ou la SDRT (Asher et Lascarides, 2003) : les *noms sous-spécifiés* (Legallois, 2008), appelés *"shell nouns"* (Schmid, 2000) ou *"signalling nouns"* (Flowerdew, 2003) dans les travaux sur l'anglais.

Les noms sous-spécifiés (désormais NSS) sont des noms comme *problème*, *idée* ou *objectif*, ayant non seulement la capacité de référer à des entités abstraites [1] décrites par une proposition syntaxique, une phrase, ou des unités discursives plus larges, mais aussi de leur attribuer un label, de caractériser leur contenu, ou leur fonction dans l'organisation du discours dans lequel ils apparaissent. Selon Schmid (2000), ils fonctionnent comme des « coquilles conceptuelles » : ils présentent une certaine incomplétude sémantique, qui est comblée par le contenu des entités auxquelles ils réfèrent. Cette incomplétude leur confère un statut proche de celui de prédicat. L'exemple (1) [2] présente une portion de discours

1. La notion d'entité abstraite a été introduite par (Asher, 1993).
2. Cet exemple est tiré du corpus *Le Monde*.

contenant trois occurrences de NSS : les noms *idée*, *objectif* et *résultat* (en gras). Ces noms réfèrent aux entités abstraites correspondant aux portions de textes entre accolades. Ici, *idée* et *objectif* étiquettent les unités auxquelles ils réfèrent comme comportant l'expression d'un but, et *résultat* l'expression d'une conséquence.

1. L'opérateur [SFR] se dit prêt à équiper, en 2003, 2000 sites capables d'accueillir des stations de base UMTS, ceci dans cinq villes françaises : Marseille, Lyon et Nice et les deux villes pilotes [...] Mais pas question d'étendre le futur réseau UMTS à l'ensemble du territoire. Pour SFR, cette technologie [...] sera réservée aux grandes agglomérations. L'**idée** est d'{offrir une continuité de services} grâce au réseau GPRS [...] SFR poursuit donc son investissement dans le GSM et le GPRS. L'**objectif** est d'{atteindre une couverture de 95 % du territoire fin 2005}, contre 84 % fin 2002. **Résultat**, {SFR accroît en 2003 ses investissements dans le réseau}.

Les NSS forment une classe fonctionnelle : les noms comme *problème* ou *objectif* n'ont pas la capacité de référer à des entités abstraites dans toutes leurs occurrences. Comme le souligne Schmid (2000), les noms comme *fact* ou *reason* ne sont pas des "shell nouns" grâce à une propriété qui leur est inhérente, ils deviennent des "shell nouns" dans certains de leurs emplois. De même, Legallois (2008) souligne que la notion de NSS s'applique à un type d'emploi nominal et non à une nature nominale. Plus précisément, il existe des constructions syntaxiques caractéristiques des emplois de NSS. Schmid (2000) décrit pour l'anglais un ensemble de patrons syntaxiques accueillant des "shell nouns", dont le patron *N be that–clause* (comme dans *the fact is that* par exemple). Pour le français, Legallois (2008) identifie les *constructions spécificationnelles* comme caractéristiques des emplois de NSS (voir section 2). Pour le français, Legallois et Gréa (2006) analysent les constructions spécificationnelles comme des dispositifs syntaxiques permettant la spécification du contenu indéterminé de ces noms. Ces constructions ont pour forme *Det N être* (*que–clause* | *de–inf*). Les noms comme *problème*, *solution* ou *objectif* possèdent donc la fonction de NSS dans certains de leurs emplois uniquement. Par exemple, en (2), l'occurrence de *problème* ne correspond pas à un emploi en tant que NSS.

2. Pour de nombreux économistes, le nouveau code du travail ne résoudra pas les **problèmes** de productivité et de compétitivité de l'économie portugaise.

Dans le cadre du Traitement Automatique des Langues, ces noms présentent évidemment un intérêt de recherche en ce qui concerne la résolution d'anaphores, et font récemment l'objet des travaux de Kolhatkar *et al.* (2013a,b), qui ont pour objectif principal d'améliorer, pour l'anglais, la résolution de leurs antécédents dans leurs emplois anaphoriques. En effet, ces noms peuvent référer à des unités abstraites de façon cataphorique, comme dans les emplois de *idée* et *objectif* dans l'exemple (1), mais aussi de façon anaphorique, comme l'emploi avec un déterminant démonstratif *ces résultats* en (3). Kolhatkar *et al.* s'intéressent à la tâche d'identification d'antécédents des emplois anaphoriques de "shell nouns", et se heurtent au manque de données annotées pour cette tâche. Kolhatkar *et al.* (2013a) se concentrent sur l'annotation manuelle des occurrences anaphoriques de ces noms par "crowdsourcing", afin de disposer de corpus d'apprentissage pour la tâche de résolution, et Kolhatkar *et al.* (2013b) utilisent comme données d'apprentissage les entités auxquelles réfèrent les "shell nouns" dans leurs emplois cataphoriques, la tâche d'identification étant bien plus aisée dans ce second cas, étant donné qu'elle peut au moins partiellement s'appuyer sur la syntaxe.

3. {Les IRM réalisées montrent une réactivation des zones du cortex moteur généralement dévolues à la main et au coude, sauf pour le troisième patient, dont l'accident était plus ancien.} [...] Comment expliquer ces **résultats** ? « Notre **hypothèse** est que {le fait de voir la main en mouvement réintroduit une cohérence dans le cerveau avec la représentation que le patient a de son corps} », avance Angela Sirigu.

Dans le présent travail, nous laissons de côté la question de l'identification des entités auxquelles les NSS réfèrent, et souhaitons nous intéresser à leur rôle dans l'organisation discursive, jusqu'ici peu étudié dans sa globalité — certains travaux, comme ceux de Vergez-Couret *et al.* (2011) concernant *pour deux raisons*, étudient le rôle discursif de cas particuliers d'emplois de NSS. Les NSS semblent pouvoir participer à l'organisation discursive, soit en spécifiant la fonction ou le contenu d'une unité au sein d'une unité discursive plus large (texte, paragraphe), soit en signalant un lien entre deux unités. Ils peuvent également constituer des marqueurs de frontières de segments textuels, ou des marqueurs de changement thématique (Schmid, 2000). Malgré ces propriétés, les NSS ont jusqu'ici suscité peu d'intérêt dans les recherches en analyse (linguistique ou automatique) du discours. L'objectif du travail présenté ici est d'effectuer une identification en corpus satisfaisante des occurrences de NSS en français, en vue d'une étude linguistique de leur participation à la structuration discursive, notamment lorsqu'ils interviennent dans des *séquences organisationnelles* fréquentes, c'est-à-dire des séquences du type *Problème–Solution* ("*Problem-Solution patterns*"), largement étudiées dans la littérature (Flowerdew, 2008). L'idée est d'identifier ces séquences organisationnelles fréquentes à l'aide des signaux que constituent les NSS (ou encore les connecteurs), et d'examiner les liens qu'entretiennent la présence de ces séquences organisationnelles avec la structure discursive, et plus particulièrement les relations de discours.

Cet article est organisé comme suit : à la section 2, nous présentons la construction d'un lexique de ces noms pour le français ; à la section 3, nous présentons une méthode d'identification des patrons syntaxiques caractéristiques des emplois

de NSS, qui s'appuie sur des techniques de fouille de données ; à la section 4, nous présentons une méthode d'identification automatique des NSS en corpus et son évaluation. Pour terminer, nous présentons les conclusions et perspectives de ce travail (section 5).

# 2 Construction d'un lexique

Dans cette section, nous présentons la construction/sélection d'un lexique de NSS pour le français. Cette construction est effectuée à partir d'un corpus constitué de 7 années d'articles du journal *Le Monde* (voir section 2.1 pour une description du corpus et des pré-traitements). Elle s'appuie sur l'extraction des occurrences de constructions syntaxiques identifiées dans la littérature comme étant caractéristiques des emplois de NSS, les constructions spécificationnelles (voir section 2.2). À la section 2.3, nous présentons les résultats de l'extraction, et la sélection du lexique, qui repose sur la fréquence d'apparition des noms dans les constructions spécificationnelles.

## 2.1 Corpus de travail et pré-traitements

Nous travaillons sur un corpus constitué de 7 années d'articles du journal *Le Monde* (de 2000 à 2006), ce qui correspond à 352 265 documents, 7 121 931 phrases et 165 097 356 tokens. Le corpus de départ est au format XML. À chaque document sont associées des métadonnées comme un secteur (une, société, france, débat, art, etc.), des catégories (critique, chiffre, mutation, chronique, important, opinion, etc.), un auteur, une date de parution, des mises à jour, etc. Au sein du texte contenu dans le document, les frontières de paragraphes sont identifiées [3].

Pour construire le lexique de NSS, nous nous appuyons sur l'identification de structures syntaxiques spécifiques, à savoir les constructions spécificationnelles (voir section suivante). Le pré-traitement du corpus comprend donc une phase d'analyse syntaxique. Nous utilisons l'analyseur syntaxique en dépendances Bonsaï (Candito *et al.*, 2010). Pour cela, nous convertissons les documents XML du corpus à un format pouvant être traité par l'analyseur en dépendances syntaxiques Bonsaï. Nous conservons un certain nombre d'informations pouvant être utiles pour nos expériences, comme les frontières de paragraphes, les métadonnées concernant le secteur et les catégories attribuées aux documents. Bonsaï prend en entrée du texte brut, opère une tokenisation (identification des composés), puis un étiquetage en parties du discours faisant appel au Melt Tagger (Denis et Sagot, 2009). L'analyse syntaxique proprement dite est effectuée par le MaltParser (Nivre *et al.*, 2006). Dans la phase de tokenisation du corpus, nous avons ajouté l'ensemble des connecteurs discursifs aux formes composées identifiées comme tokens avant l'analyse syntaxique, et également certaines formes composées auxquelles appartiennent des NSS comme *point*, mais ne correspondant pas (de façon autonome au moins) à des emplois comme NSS (*à ce point*, *point de vue*).

## 2.2 Méthode d'extraction

Comme nous l'avons vu précédemment, (Schmid, 2000) identifie comme caractéristiques des emplois de "shell nouns" en anglais les structures syntaxiques correspondant au patron : *Det N be* (*that–clause* | *wh–clause* | *to–inf*). Pour le français, les structures caractéristiques identifiées par (Legallois, 2008) sont les constructions dites spécificationnelles, qui couvrent des phrases copulatives dans lesquelles l'objet du verbe *être* est une complétive ou un infinitif. Ces constructions sont décrites par le patron suivant :

$$\textit{Det N } (\varnothing \mid \textit{ce}) \textit{ être } (\textit{que–clause} \mid \textit{de–inf}).$$

On trouve des exemples de constructions spécificationnelles dans les phrases en (4), avec une complétive, et (5), avec un infinitif. Parmi les constructions spécificationnelles, on trouve également des pseudo-clivées, comme en (6).

4. Le **risque** est que ce genre de comportement rappelle de bien mauvais souvenirs.
5. La **question** est de savoir ce que l'on prend comme élément de référence.
6. Le **problème**, c'est que les Occidentaux ne comprennent pas leur mentalité.

3. C'est une des raisons pour lesquelles nous avons choisi ce corpus. Nous verrons dans les perpectives que disposer de ce type de d'informations concernant l'organisation du texte pourra nous aider dans l'identification des séquences organisationnelles.

La notion de construction spécificationnelle s'appuie sur la classification de Higgins (1979) des phrases copulatives. Higgins identifie quatre types de phrases copulatives : prédicative (*cette voiture est rapide*), identificatrice (*la dame avec un chapeau, c'est Madame Dupont*), identité (*l'étoile du matin est l'étoile du soir*), spécificationnelle (*ce que je voudrais, c'est que tu gares la voiture*). Ce type de construction est également abordé par Apothéloz (2008), qui s'intéresse aux constructions spécificationnelles — pour lesquelles il emploie le terme de constructions identificatives — et plus particulièrement aux pseudo-clivées. Parmi ces constructions, Apothéloz relève notamment des cas dans lesquels le segment gauche est un adjectif nominalisé (*important*, *mieux*, *pire*), et des cas dans lesquels le segment gauche est un nom. Parmi les noms observés dans ces constructions, il observe des « lexèmes évaluatifs (*difficulté*, *problème*, *ennui*) ou des hyperonymes servant à construire un syntagme évaluatif (*une* ***chose*** *frappante*, *le* ***truc*** *sur lequel je ne suis pas d'accord*) » et des « lexèmes se rapportant à l'activité langagière, notamment dans ses aspects argumentatifs et explicatifs (cf. *remarque*, *hypothèse*, *preuve*, *raison*, *proposition*) ».

Nous reprenons ici l'idée de Legallois (2008), qui est de s'appuyer sur le repérage en corpus des constructions spécificationnelles pour identifier un lexique de NSS, en étendant la taille du corpus d'extraction. Nous rassemblons sous l'étiquette NSS des noms entrant dans d'autres constructions que les constructions spécificationnelles, mais nous appuyons sur celles-ci pour construire un lexique à partir duquel travailler. Nous effectuons un repérage des constructions spécificationnelles sur l'ensemble du corpus décrit dans à la section 2.1. Pour identifier les constructions, nous recherchons les contextes correspondant au schéma de la figure 1 dans les analyses en dépendances syntaxiques des phrases du corpus. Une fois ce contexte repéré dans une phrase du corpus, l'extracteur vérifie que plusieurs contraintes sur les dépendants de *être* sont respectées : il ne peuvent pas être des participes passés ou des adjectifs, il ne peuvent pas être des pronoms réflexifs, et ne peuvent pas non plus être des prépositions (*être en mesure de* analysé avec *en* dépendant de *être*, et *de* autre dépendant de *être*). En revanche, les adverbes sont admis comme dépendants de *être*.

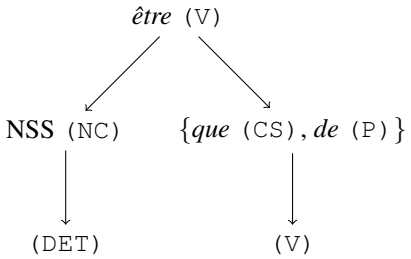


FIGURE 1 – Contextes recherchés dans les phrases du corpus

Nous donnons à la table 1 un fragment de phrase qui contient une occurrence des contextes recherchés dans le corpus.

| Id. | Forme | Lemme | Ét. POS | Informations morphologiques | Id. gouv. | Fonction synt. |
|---|---|---|---|---|---|---|
| 1 | Le | le | DET | g=m\|n=s\|s=def | 2 | det |
| 2 | problème | problème | NC | g=m\|n=s\|s=c | 4 | suj |
| 3 | aujourd'_hui | aujourd'hui | ADV | _ | 2 | mod |
| 4 | est | être | V | m=ind\|n=s\|p=3\|t=pst | 0 | root |
| 5 | que | que | CS | s=s | 4 | obj |
| 6 | les | le | DET | n=p\|s=def | 7 | det |
| 7 | hommes | homme | NC | g=m\|n=p\|s=c | 9 | suj |
| 8 | politiques | politique | ADJ | n=p\|s=qual | 7 | mod |
| 9 | tentent | tenter | V | m=ind\|n=p\|p=3\|t=pst | 5 | obj |

TABLE 1 – Fragment de phrase analysée par Bonsaï contenant une construction spécificationnelle

## 2.3 Résultats de l'extraction et sélection du lexique

Au total, 28 445 phrases contiennent la structure recherchée, soit 0,48 % des phrases du corpus. Nous identifions et comptons tous les noms trouvés dans ces contextes. Au total, 1 670 noms (ou adjectifs nominalisés) différents entrent

dans les contextes recherchés dans le corpus. Nous présentons dans le tableau 2 les noms les plus fréquemment employés dans les constructions syntaxiques recherchées. Pour chaque nom (ou adjectif nominalisé), nous renseignons le nombre d'occurrences, et le pourcentage sur le nombre total d'occurrences du patron syntaxique.

| Lemme | Nb. d'occurrences | % | Lemme | Nb. d'occurrences | % |
|---|---|---|---|---|---|
| objectif | 3635 | 12.78 | intérêt | 365 | 1.28 |
| problème | 1620 | 5.7 | priorité | 324 | 1.14 |
| but | 1597 | 5.61 | important | 317 | 1.11 |
| question | 1293 | 4.55 | risque | 309 | 1.09 |
| idée | 1176 | 4.13 | difficulté | 288 | 1.01 |
| rôle | 554 | 1.95 | chose | 285 | 1.0 |
| ambition | 475 | 1.67 | souci | 275 | 0.97 |
| mission | 435 | 1.53 | solution | 249 | 0.88 |
| essentiel | 392 | 1.38 | mérite | 248 | 0.87 |
| enjeu | 382 | 1.34 | vérité | 245 | 0.86 |

TABLE 2 – Les 20 noms apparaissant le plus fréquemment dans les contextes recherchés

Au sein des constructions spécificationnelles identifiées, les NSS peuvent être modifiés par un adjectif, comme en (7), par un syntagme prépositionnel, comme en (8), ou une proposition relative, comme en (9). Il peuvent avoir un déterminant possessif, comme en (10). On trouve également dans les constructions spécificationnelles des adjectifs nominalisés comme *essentiel* ou *important*. On retrouve les hyperonymes mentionnés par Apothéloz, comme *chose* ou *côté*, qui fonctionnent généralement avec un modifieur, comme en (11).

7. En fait, le principal **problème** de cette présentation est que « les membres de l'ONU à qui elle s'adressait n'en ont rien à faire », commente le New Republic sur son site Internet.
8. L'**objectif** de l'entraîneur, c'est de donner au joueur de l'autonomie.
9. La **leçon** que j'en tire, c'est qu'il faut éviter ces passages à vide, qu'il faut de la constance.
10. Notre **position** est que nous n'aurions pas dû être en Irak, en premier lieu.
11. Le **côté** positif de cette motion de censure, c'est de ressouder la majorité autour du premier ministre.

Pour les expériences présentées dans la suite de cet article, nous conservons la portion du lexique pour laquelle le nombre d'occurrences de chaque nom dans les constructions spécificationnelles dépasse 0,1 % du total d'occurrences des constructions spécificationnelles en corpus (28 445). Le lexique ainsi sélectionné contient 122 noms, les moins fréquents du lexique retenu étant *possibilité* et *option*, avec chacun 29 occurrences. Parmi l'ensemble des noms ayant été identifiés dans les constructions spécificationnelles, 1 546 sont donc exclus du lexique retenu. Le lexique sélectionné est volontairement restreint pour éviter tout bruit lié au lexique. Les noms dont les fréquences sont les plus basses correspondent généralement à des erreurs d'analyse syntaxique, pour des phrases dont la structure est moins fréquente — voir par exemple *ljubljanais* en (12).

12. D'ailleurs, le dernier chic culinaire **ljubljanais** est de mettre sur la carte des restaurants un « simple poisson avec son filet d'huile d'olive » après une tapenade – qu'on accompagne d'un tokay de Goriska Brda, cette région proche de l'Italie qui produit de bons crus.

## 3 Fouille de données pour la découverte de contextes syntaxiques spécifiques

Comme nous le verrons à la section 4, nous souhaitons identifier les occurrences de NSS dans le corpus décrit à la section 2.1, afin d'extraire des séquences organisationnelles fréquentes. A priori, nous n'avons aucun moyen de distinguer dans le corpus les emplois comme NSS des items de notre lexique de leurs autres emplois sans procéder à une annotation manuelle — excepté dans les cas où ils entrent dans une construction spécificationnelle. Or, l'identification de séquences organisationnelles fréquentes suppose de travailler sur un corpus conséquent, pour lequel une annotation complètement manuelle des emplois de NSS n'est pas envisageable. Nous verrons à la section 4 que considérer toutes les occurrences des noms du lexique comme des occurrences de NSS n'est pas une solution satisfaisante en ce qui concerne le bruit, et que se limiter aux seules constructions spécificationnelles n'est pas satisfaisant non plus en ce qui concerne le silence. Nous avons donc opté pour une solution intermédiaire d'identification des NSS, pour laquelle nous ne disposions que d'un lexique d'items ayant la capacité d'être des NSS, et d'un corpus conséquent analysé pour les dépendances

syntaxiques. Cependant, nous savons que les NSS sont susceptibles d'apparaître dans des constructions syntaxiques particulières, dans lesquelles d'autres noms ne peuvent pas entrer. Pour l'anglais, des études sur corpus des emplois de NSS ont déjà été effectuées par Schmid (2000) et Flowerdew (2003), et recensent des contextes syntaxiques autres que le patron *Det N be* (*that–clause* | *wh–clause* | *to–inf*), comme le patron *Det N* (*that–clause* | *wh–clause* | *to–inf*) (Schmid, 2000) ou le patron *Det N be Det N*, dans lequel le second nom est un déverbal (Flowerdew, 2003). Pour le français, les constructions spécifiques aux NSS n'ont pas été inventoriées. On suppose qu'il existe en français d'autres constructions que la construction spécificationnelle [4] dans lesquelles on trouve des emplois de NSS. L'objectif du travail présenté dans cette section est précisément **d'identifier ces constructions**.

Nous partons de l'hypothèse que les noms communs n'appartenant pas au lexique des NSS ne peuvent pas entrer dans les constructions spécifiques aux emplois de NSS. L'idée qui sous-tend cette approche est que les propriétés discursives ou sémantiques des emplois de NSS sont corrélées à des traits syntaxiques. Pour identifier des constructions syntaxiques caractéristiques des emplois de NSS, nous utilisons une méthode permettant d'identifier des patrons linguistiques fréquents en corpus, héritée des techniques de fouille de données, et proposée par Béchet *et al.* (2012) : l'extraction de motifs séquentiels. La *fouille de données séquentielles*, introduite par Agrawal et Srikant (1995), est une technique qui permet de découvrir de nouvelles connaissances sous forme de régularités dans des bases de données, en tenant compte de l'ordre temporel entre les données. Suivant une méthodologie proche de celle de Quiniou *et al.* (2012) concernant la stylistique, nous nous intéressons aux particularités des contextes syntaxiques des NSS, via le calcul de *motifs séquentiels émergents*. Ces motifs émergents sont identifiés à partir de motifs séquentiels impliquant des noms retenus dans le lexique de NSS (étiquetés `NSS`), et de motifs impliquant des noms communs (étiquetés `NC`) — et ne contenant aucun des noms retenus dans le lexique. L'avantage de cette méthode de fouille est d'être symbolique. Les motifs extraits, et découverts automatiquement, sont interprétables par un humain. Ils se prêtent donc facilement à une analyse linguistique.

À la section 3.1, nous présentons les principes généraux du calcul de motifs séquentiels et de l'identification de motifs émergents. À la section 3.2, nous décrivons l'extraction de motifs émergents pour les NSS, et donnons les résultats quantitatifs de cette extraction. À la section 3.3, nous décrivons les principaux patrons syntaxiques caractéristiques des emplois de NSS, identifiés par l'analyse des motifs émergents.

## 3.1 Les motifs séquentiels et les motifs émergents

Nous définissons dans cette section les notions impliquées dans le calcul de motifs séquentiels et des motifs émergents, en les illustrant à travers des exemples similaires aux données que nous traitons ici. Néanmoins, il faut noter que la méthode de fouille de données séquentielles décrite ici trouve de très diverses applications, comme par exemple l'extraction de chaînes d'ADN ou de séquences de protéines.

Un *itemset* $I$ est un ensemble de littéraux appelés *items*, représenté par $I = (i_1 \dots i_n)$. Par exemple, (*problème* $NC$) est un itemset contenant un item correspondant à un lemme et un item correspondant à une étiquette POS. Une *séquence* $S$ est une liste ordonnée d'itemsets, représentée par $S = \langle I_1 \dots I_m \rangle$. Notons que de nombreuses applications ne nécessitent qu'un seul item dans leurs itemsets. Ces séquences sont appelées des *séquences d'items* et sont représentées par $S = \langle i_1 \dots i_n \rangle$, où $i_1 \dots i_n$ sont des items. Dans la suite de cet article, les deux types de séquences seront considérés : les séquences d'items (lemmes) et les séquences d'itemsets (lemmes et étiquettes POS). Une séquence $S = \langle I_1 \dots I_n \rangle$ est *contenue* dans une séquence $S' = \langle I'_1 \dots I'_m \rangle$ s'il existe des entiers $1 \leq j_1 < \dots < j_n \leq m$ tels que $I_1 \subseteq I'_{j_1}$, ..., $I_n \subseteq I'_{j_n}$. La séquence $S$ est appelée *sous-séquence* de $S'$, ce qui est noté $S \preceq S'$. Par exemple, la séquence $\langle$(`DET`) (`NC`) (*être* `V`)$\rangle$ est contenue dans la séquence $\langle$(*le* `DET`) (*solution* `NC`) (*être* `V`) (*de* `P`)$\rangle$. Une base de séquences $B$ est un ensemble de tuples $(S_{id}, S)$, où $S_{id}$ est un identifiant de séquence et $S$ une séquence. La table 3 représente ainsi une base de trois séquences. Dans notre cas, une séquence correspond à la représentation d'une phrase du corpus. Les itemsets peuvent contenir des lemmes ou des étiquettes POS.

| Identifiant | Séquence |
|---|---|
| 1 | $\langle$(*le* `DET`) (*solution* `NC`) (*être* `V`)(*de* `P`) (*partir* `VINF`) (. `PONCT`)$\rangle$ |
| 2 | $\langle$(*le* `DET`) (*problème* `NC`) (: `PONCT`)(*trouver* `VINF`) (*un* `DET`) (*financement* `NC`) (. `PONCT`)$\rangle$ |
| 3 | $\langle$(*le* `DET`) (*solution* `NC`) (: `PONCT`) (*partir* `VINF`) (. `PONCT`)$\rangle$ |

TABLE 3 – Une base de séquences

4. (*Det N* (∅ | *ce*) *être* (*que–clause* | *de–inf*) décrite en section 2.2.

Un tuple $(S'_{id}, S')$ *contient* une séquence $S$ si $S \preceq S'$. Le *support* d'une séquence $S$ dans une base de séquences $B$, noté *sup*$(S)$, est le nombre de tuples contenant $S$ dans la base. Par exemple, dans la table 3, le support de la séquence $(\langle(\textit{le}\ \texttt{DET})\ (\textit{solution}\ \texttt{NC})\rangle)$ est 2. Le *support relatif* peut aussi être utilisé, comme défini par l'équation suivante.

$$sup(S) = \frac{|\{(S'_{id}, S') \mid (S'_{id}, S') \in B \wedge (S \preceq S')\}|}{|B|}$$

Un *motif fréquent* est une séquence dont le support est supérieur ou égal à un seuil fixé : *minsup*. Les algorithmes de fouille de motifs séquentiels extraient tous les motifs fréquents apparaissant dans une base de séquences. L'ensemble des motifs fréquents pouvant être très grand, il existe une représentation condensée permettant d'éliminer les redondances sans perte d'information : les *motifs clos* (Yan *et al.*, 2003). Un motif fréquent $S$ est clos s'il n'existe aucun motif fréquent $S'$ tel que $S \preceq S'$ et *sup*$(S) =$ *sup*$(S')$. Certaines contraintes peuvent être définies pour diriger l'extraction de motifs selon les besoins de l'utilisateur et éliminer des motifs non pertinents (Dong et Pei, 2007), comme la contrainte de fréquence (en donnant une valeur au support minimal) ou la contrainte *gap* : un motif avec une contrainte *gap* égale à $[x, y]$, noté $P_{[x,y]}$, est un motif dont chaque couple d'itemsets est séparé par au moins $x - 1$ itemsets et au plus $y - 1$ itemsets.

Les motifs émergents sont des motifs dont le support augmente de manière significative d'un ensemble de données à un autre (Dong et Li, 1999). Les motifs émergents sont ainsi des motifs dont le taux de croissance ("growth rate"), c'est-à-dire le rapport des supports dans deux ensembles de données, est supérieur à un seuil fixé $\rho$. Un motif $P$ d'un ensemble de données $D_1$ est alors un *motif émergent*, par rapport à un autre ensemble de données $D_2$, si $TC(P) \geq \rho$, avec $\rho > 1$ et $TC(P)$ défini par l'équation suivante.

$$TC(P) = \begin{cases} \infty & \text{si } sup_{D_2}(P) = 0 \\ \frac{sup_{D_1}(P)}{sup_{D_2}(P)} & \text{sinon} \end{cases}$$

## 3.2 Extraction de motifs émergents pour les noms sous-spécifiés

Pour calculer des motifs émergents relatifs aux noms sous-spécifiés, nous constituons tout d'abord une base de séquences $B_{NSS}$, dans laquelle chaque séquence correspond à une phrase. Notons que sur chaque séquence nous ajoutons deux itemsets particuliers – (`INIT`) et (`END`) – qui symbolisent respectivement les débuts et fins de phrase. Pour constituer la base $B_{NSS}$, on extrait du corpus présenté à la section 2.1 toutes les phrases contenant au moins un nom du lexique de NSS d'une part, en remplaçant leur étiquette `NC` (nom commun) par l'étiquette `NSS` (nom sous-spécifié), ou leur étiquette `ADJ` (adjectif) par l'étiquette `ASS` (adjectif nominalisé sous-spécifié). Au total, 1 881 526 phrases du corpus contiennent au moins une occurrence d'un nom du lexique retenu (soit environ 26 % des phrases), donc $|B_{NSS}|$ vaut 1 881 526. L'ensemble de ces phrases contient 2 035 850 occurrences étiquetées comme `NSS`, et 400 099 occurrences étiquetées comme `ASS`.

Pour limiter le nombre de motifs émergents à analyser, et pour extraire des motifs émergents plus génériques, nous ne conservons dans la base de séquences que l'étiquette POS (on ne tient pas compte des lemmes), excepté pour les mots appartenant aux classes fermées (`DET`, `P`, `P+D`, `CS`, `CC`, `PROREL`), ainsi que pour les auxiliaires *être* et *avoir*. Cela nous permet d'éliminer des motifs émergents trop spécifiques, se rapportant essentiellement à un élément du lexique de NSS. En effet, nous avons mené une première extraction de motifs émergents dont les séquences contiennent des itemsets avec lemmes et étiquettes POS. L'observation des motifs émergents les plus fréquents montre qu'on extrait par cette méthode beaucoup de motifs spécifiques à un NSS en particulier. Parmi les motifs ainsi extraits, on trouve des motifs positifs, comme *[poser le NSS de]* (par exemple avec *question* ou *problème*) ou *[NSS consister à]* (par exemple avec *solution*), correspondant à des occurrences de noms sous-spécifiés, et des motifs négatifs, ne correspondant pas à des occurrences de noms sous-spécifiés. Parmi les motifs émergents les plus fréquents, on trouve par exemple le motif *projet de NC* (pour *projet de loi*) qui ne peut pas être considéré comme une occurrence du NSS *projet*. C'est également le cas pour des motifs comme *faire l'objet de*, *marché du travail*, *contrat de travail*, *point de vente*, *être sur le point de*, *mettre au point*, etc. La prise en compte de ces motifs émergents plus spécifiques peut être utile pour raffiner l'identification des NSS, mais dans un premier temps, nous voulons identifier des motifs émergents plus génériques.

Nous effectuons le choix du support *minsup* pour l'extraction des motifs dans la base $B_{NSS}$ en fixant la contrainte suivante : les motifs extraits doivent être présents dans au moins 0,1 % du nombre d'occurrences de l'étiquette `NSS`. Ce nombre d'occurrence étant de 2 035 850, le support choisi est de 2036. Pour réduire les redondances dans les motifs extraits, nous utilisons l'extraction de motifs clos. Nous fixons la contrainte *gap* à $[1, 1]$, c'est-à-dire que les items ou itemsets des motifs extraits sont séparés par 0 items ou itemsets. La longueur minimale des motifs est fixée à 2, et la longueur maximale à 7.

Après l'extraction des motifs effectuée sur la base de séquences $B_{NSS}$, nous retenons deux ensembles de motifs :

- l'ensemble $D_{NSS}$, qui rassemble les motifs calculés à partir de $B_{NSS}$ et contenant une étiquette NSS, comme le motif (*le* DET) (NC) (*être* V) (*que* CS) ;
- l'ensemble $D_{NC}$, qui rassemble les motifs calculés à partir de $B_{NSS}$ contenant des étiquettes NC mais pas d'étiquette NSS, comme (*le* DET) (NC) (V) (ADJ).

Le tableau 4 indique le nombre de motifs clos fréquents (en valeur absolue et en pourcentage) en fonction de leur longueur et pour ces 2 ensembles.

| Ensemble de motifs | Longueur 2 | Longueur 3 | Longueur 4 | Longueur 5 | Longueur 6 | Longueur 7 | Total |
|---|---|---|---|---|---|---|---|
| $D_{NSS}$ | 79 (5.71 %) | 419 (30.3 %) | 527 (38.1 %) | 290 (20.97 %) | 65 (4.7 %) | 3 (0.22 %) | 1 383 (100 %) |
| $D_{NC}$ | 191 (3.19 %) | 1 110 (18.53 %) | 2 246 (37.5 %) | 1 658 (27.68 %) | 629 (10.5 %) | 156 (2.6 %) | 5 990 (100 %) |

TABLE 4 – Nombre de motifs fréquents et clos extraits

| Motifs émergents | Longueur 2 | Longueur 3 | Longueur 4 | Longueur 5 | Longueur 6 | Longueur 7 | Total |
|---|---|---|---|---|---|---|---|
| $D_{NSS}/D_{NC}$ | 1 (0.87 %) | 23 (20 %) | 46 (40 %) | 34 (29.57 %) | 11 (9.57 %) | 0 (0 %) | 115 (100 %) |
| $D_{NSS}/D_{NC}$ (avec $TC = \infty$) | 1 (1.39 %) | 11 (15.28 %) | 25 (34.72 %) | 26 (36.11 %) | 9 (12.5 %) | 0.0 (0 %) | 72 (100 %) |

TABLE 5 – Nombre de motifs émergents

L'ensemble de motifs dans lequel nous voulons identifier des motifs émergents est $D_{NSS}$, c'est-à-dire les motifs contenant des étiquettes NSS. Nous avons procédé à l'extraction de motifs émergents de $D_{NSS}$ par rapport à $D_{NC}$. Pour cela, on considère chaque motif appartenant à $D_{NSS}$, puis on calcule son taux de croissance en recherchant le même motif dans $D_{NC}$ ; après avoir substitué à l'étiquette NSS l'étiquette NC. Le tableau 5 donne le nombre de motifs émergents de $D_{NSS}$ par rapport à $D_{NC}$, tous taux de croissance confondus puis uniquement avec les motifs ayant un taux de croissance infini.

On remarque que le calcul des émergents produit un nombre de motifs relativement faible (187) qui permet une analyse manuelle. D'autre part, parmi ces motifs émergents, 72 ont un taux de croissance infini : ce sont des patrons syntaxiques caractéristiques des NSS puisque ces constructions n'apparaissent pas avec un NC. Notons que ce sont surtout les motifs de longueur 3 à 5 qui semblent les plus pertinents. Le tableau 6 montre l'ensemble des motifs émergents de longueur 3.

Parmi l'ensemble des motifs émergents, on retrouve la construction spécificationnelle. Par exemple, dans les motifs de longueur 3, on la retrouve en (e) dans le tableau 6. On la retrouve également dans les motifs de longueur 5, sous la forme (DET *le*) (NSS) (V *être*) (P *de*) (VINF). Mais de plus, l'un des résultats intéressants est que la méthode fondée sur les motifs émergents a permis de découvrir de nouvelles constructions spécifiques aux NSS. La section suivante présente des observations en corpus à partir de ces patrons caractéristiques découverts.

## 3.3 Patrons identifiés à partir des motifs émergents

Dans cette section, nous présentons les principaux patrons caractéristiques des emplois de NSS que nous avons identifiés grâce à l'extraction des motifs émergents et l'exploration manuelle du corpus guidée par les motifs. Les patrons identifiés correspondent à des emplois cataphoriques de NSS.

**Le patron (V *avoir*) (P *pour*) (NSS) (P *de*)** Ce patron a été identifié à partir de motifs émergents dont le taux de croissance est égal à $\infty$ et dont le support est parmi les plus élevés. Il couvre par exemple les motifs $(a)$ et $(b)$ du tableau 6, (P *pour*) (NSS) (P *de*) et (V *avoir*) (P *pour*) (NSS), dont on trouve respectivement une occurrence dans les exemples (13) et (14) [5], tirés du corpus *Le Monde*. Il correspond également à des motifs émergents plus longs, comme le motif de longueur 5 : (V *avoir*) (P *pour*) (NSS) (P *de*) (VINF).

13. Le gouvernement s'est fixé *pour* ***objectif*** *de* {parvenir à 1,6 % du PIB en 2008 et à 2 % en 2010}.

5. Dans les exemples présentés dans cette section, nous notons les NSS en gras, les patrons présentés en italiques, et les entités auxquelles réfèrent les NSS entre accolades.

| Motif | Taux de croissance | Support absolu | Support relatif |
|---|---|---|---|
| $(a)$ (P *pour*) (NSS) (P *de*) | $\infty$ | 7 305 | 0.0039 |
| $(b)$ (V *avoir*) (P *pour*) (NSS) | $\infty$ | 4 770 | 0.0025 |
| $(c)$ (NSS) (P *de*) (CLO *cla*) | $\infty$ | 3 293 | 0.0018 |
| $(d)$ (NSS) (PONCT :) (VINF) | $\infty$ | 3 268 | 0.0017 |
| $(e)$ (INIT) (NSS) (PONCT :) | $\infty$ | 3 164 | 0.0017 |
| $(f)$ (V *avoir*) (NSS) (P *à*) | $\infty$ | 3 139 | 0.0017 |
| $(g)$ (VPP *avoir*) (DET *le*) (NSS) | $\infty$ | 2 868 | 0.0015 |
| $(h)$ (NSS) (V *être*) (CS *que*) | $\infty$ | 2 654 | 0.0014 |
| $(i)$ (V *avoir*) (DET *aucun*) (NSS) | $\infty$ | 2 626 | 0.0014 |
| $(j)$ (DET *aucun*) (NSS) (P *de*) | $\infty$ | 2 614 | 0.0014 |
| $(k)$ (NSS) (P *de*) (VINF *avoir*) | $\infty$ | 2 373 | 0.0013 |
| $(l)$ (NSS) (P *de*) (VINF *être*) | 3.05 | 6 858 | 0.0036 |
| $(m)$ (DET *le*) (NSS) (CS *que*) | 2.43 | 17 055 | 0.0091 |
| $(n)$ (NSS) (V *être*) (P *de*) | 2.25 | 11 442 | 0.0061 |
| $(o)$ (V *avoir*) (DET *le*) (NSS) | 2.02 | 11 512 | 0.0061 |
| $(p)$ (NSS) (P *de*) (VINF) | 2.01 | 78 905 | 0.0419 |
| $(q)$ (P *pour*) (DET *de*) (NSS) | 1.82 | 6 316 | 0.0034 |
| $(r)$ (DET *le*) (NSS) (PROREL *dont*) | 1.65 | 5 894 | 0.0031 |
| $(s)$ (NSS) (P *de*) (CLR *se*) | 1.6 | 6 768 | 0.0036 |
| $(t)$ (V *avoir*) (DET *de*) (NSS) | 1.49 | 3 503 | 0.0019 |
| $(u)$ (V *être*) (DET *son*) (NSS) | 1.19 | 2 584 | 0.0014 |
| $(v)$ (V *être*) (ADV) (NSS) | 1.16 | 3 468 | 0.0018 |
| $(w)$ (DET *un*) (NSS) (PONCT :) | 1.08 | 3 646 | 0.0019 |

TABLE 6 – Exemple de motifs émergents de longueur 3 de $D_{NSS}$ par rapport à $D_{NC}$

14. Ces deux procédures *ont pour* ***résultat*** *de* {déplacer insensiblement le centre d'intérêt, tenu dans la première version par l'éblouissante héroïne, vers le personnage effacé et hypocondriaque de son mari}...

**Le patron** (**NSS**) (**PONCT** :) Toujours parmi les motifs émergents dont le taux de croissance est égal à $\infty$ et dont le support est parmi les plus élevés, on trouve des motifs comme ceux étiquetés $(d)$ et $(e)$ dans le tableau 6. Dans le second motif, INIT symbolise le début de phrase. On trouve respectivement des occurrences de ces motifs dans les exemples (15) et (16).

15. Leur ***objectif*** *:* {*être* identifiés comme des « adultes disponibles » avec lesquels on peut parler de tout et de rien}.
16. ***Conclusion*** *:* {à ce jour, la monnaie unique n'a guère enrayé le malaise économique européen et l'on ne peut manquer de s'interroger sur son éventuelle responsabilité dans les difficultés économiques actuelles de la zone euro}.

**Les patrons** (**NSS**) (**P** ***de***) (**VINF**) **et** (**NSS**) (**CS** ***que***) Parmi les motifs émergents dont le taux de croissance n'est pas égal à $\infty$, on trouve des motifs correspondant aux cas dans lesquels les NSS sont suivis d'un infinitif ou d'une complétive, comme les motifs $(k)$, $(l)$, $(m)$ et $(p)$. Ces motifs semblent caractériser des emplois sous-spécifiés de noms, comme le montrent les exemples en (17) et (18).

17. C'est un site commercial qui a été lancé en juin 2002, à New York, avec l'***objectif*** *de* {*mettre* les gens en relation les uns avec les autres autour d'un sujet d'intérêt commun}.
18. *L'**idée*** {*que* les inspecteurs puissent renifler les armes et les documents qui s'y rapportent sans l'aide des autorités irakiennes} est absurde.

# 4 Identification des occurrences de noms sous-spécifiés

Dans cette section, nous présentons la méthode utilisée pour l'identification des occurrences de noms sous-spécifiés dans le corpus de travail. Cette identification s'appuie à la fois sur le lexique sélectionné à la section 2 et les patrons identifiés à la section précédente. Nous avons procédé à une évaluation manuelle de l'identification des occurrences de noms sous-spécifiés, que nous présentons également.

**Contextes recherchés lors de l'identification** Lors de l'identification des noms sous-spécifiés, seules les occurrences de noms appartenant au lexique sélectionné à la section 2 sont considérées par le système. Chaque occurrence rencontrée est étiquetée comme nom sous-spécifié (NSS) lorsque son contexte d'apparition correspond à un des 6 patrons présentés ci-dessous (et dans le tableau 8), soit comme nom commun (NC) dans le cas contraire. Parmi les contextes menant à un étiquetage comme NSS, on retrouve les patrons syntaxiques décrits à la section 3.3, que nous noterons ici NSS_etre_que_de (pour la construction spécificationnelle), pour_NSS_de, NSS_punct (pour les cas dans lesquels le nom est suivi de deux points) et NSS_que_de (pour les noms prédicatifs suivis d'une complétive ou d'une infinitive). Au cours des différentes étapes du travail et de l'exploration manuelle du corpus, nous avons également identifié des contextes d'occurrences anaphoriques de NSS, qui n'ont pas été identifiées parmi les motifs émergents, et que nous avons intégrés aux contextes déclenchant l'étiquetage comme NSS. Ces contextes sont désignés par les patrons : dem_NSS, pour les occurrences dans lesquelles le NSS a un déterminant démonstratif, comme dans *ce problème* ; root_NSS, pour les occurrences dans lesquelles le NSS est la tête d'une phrase averbale, et dans lesquelles il peut être modifié par une relative, comme en (19).

19. Une **perspective** qui incite les français à accélérer leur internationalisation.

Pour chaque patron, le repérage peut être effectué de façon surfacique (c'est-à-dire en s'appuyant uniquement sur la séquences de lemmes et d'étiquettes POS, et en n'autorisant aucune distance entre les différents éléments du patron) ou en s'appuyant sur les dépendances syntaxiques, comme nous l'avons présenté à la section 2.2 pour l'identification des constructions spécificationnelles. Cela nous permet notamment de prendre en compte les cas dans lesquels le NSS est modifié par un adjectif, un groupe prépositionnel ou une relative, ce qui n'est pas permis par le repérage surfacique.

**Évaluation de l'identification** L'évaluation porte sur les occurrences de noms (ou adjectifs) du lexique de NSS, étiquetées préalablement par le système présenté précédemment [6]. L'annotation a été effectuée par un annotateur sur 38 documents du corpus, dans lesquels 600 occurrences de noms du lexique ont été repérées par le système, et 120 ont été étiquetées comme NSS (20 % du total des occurrences). Pour chaque occurrence, l'annotation a consisté à vérifier que l'étiquette attribuée était correcte. Pour les cas dans lesquels le rôle de nom sous-spécifié est incertain, l'occurrence a été annotée comme unknown [7].

Nous présentons dans le tableau 7 la répartition des occurrences en fonction de l'étiquetage du système et de l'annotation manuelle. Parmi les 600 occurrences de noms du lexique, 150 ont été annotées manuellement comme NSS (soit 25 %). Le rappel du système est de 0,61. En mettant de côté les occurrences étiquetées comme NSS et annotées comme unknown, la précision de l'identification des NSS effectuée par le système est de 0,83 et le F-score de 0,71. En considérant les occurrences étiquetées comme NSS et annotées comme unknown comme des faux positifs, la précision est de 0,77 et le F-score de 0,68. Une des conclusions que l'on peut tirer de cette évaluation, c'est que considérer toutes les occurrences des noms du lexique comme des occurrences de NSS est susceptible de bruiter considérablement les données extraites. Si notre système avait étiqueté toutes les occurrences de noms du lexique retenu comme NSS, la précision aurait été de 0,29, le rappel de 1, et le F-score de 0,45.

| Occurrences | Annotées NSS | Annotées NC | Annotées unknown |
|---|---|---|---|
| Étiquetées NSS | 92 | 18 | 10 |
| Étiquetées NC | 58 | 355 | 67 |

TABLE 7 – Répartition des occurrences évaluées, en fonction de l'étiquetage du système et de l'annotation manuelle

Dans le tableau 8, nous présentons le nombre d'occurrences et le pourcentage d'apparition des différents patrons sur l'ensemble des 120 occurrences étiquetées comme NSS par le système. Nous présentons également, pour chaque patron, la répartition des annotations et la précision de l'étiquetage du système (en ne comptant comme correctes que les occurrences annotées NSS). Ces résultats montrent que l'intégration d'autres patrons que les constructions spécificationnelles est nécessaire à une extraction moins silencieuse. En effet, on peut comparer les résultats de l'identification à ceux d'une identification qui s'appuierait uniquement sur la présence des constructions spécificationnelles. Nous observons que seulement 10 occurrences de NSS ont été extraites grâce à la présence du patron NSS_etre_que_de. Si l'on avait choisi de considérer comme NSS uniquement ces occurrences, le rappel du système aurait été de 0,07, la précision de 1, et le F-score de 0,13.

6. Nous n'évaluons pas la couverture du lexique sélectionné. Comme nous l'avons déjà dit, le lexique sélectionné est volontairement restreint pour interdire le bruit qui pourrait être lié au lexique.

7. Parmi les occurrences annotées unknown, on trouve essentiellement des cas dans lesquels le NSS réfère bien à une entité abstraite, mais dans lesquels cette entité est nominalisée, comme dans *« D'abord, un sentiment de fragilisation professionnelle. »*, et pour lesquels une étude plus approfondie des NSS nous permettrait de trancher.

| Patron | Nb. d'occ. | Annotées NSS | Annotées NC | Annotées unknown | Précision |
|---|---|---|---|---|---|
| NSS_que_de | 40 (43,5 %) | 31 | 7 | 2 | 0,78 |
| dem_NSS | 35 (38 %) | 29 | 2 | 4 | 0,83 |
| NSS_punct | 17 (18,5 %) | 11 | 4 | 2 | 0,65 |
| root_NSS | 16 (17,4 %) | 9 | 5 | 2 | 0,57 |
| NSS_etre_que_de | 10 (10,9 %) | 10 | 0 | 0 | 1 |
| pour_NSS_de | 0 (0 %) | 0 | 0 | 0 | – |

TABLE 8 – Proportion des différents patrons dans l'ensemble des occurrences étiquetées comme NSS et résultats de l'annotation

Pour les 58 occurrences NSS qui n'ont pas été repérés par le système, nous avons annoté le contexte d'apparition. Pour 18 d'entre eux, le contexte correspond à un patron (NSS_etre_que_de, NSS_punct ou NSS_que_de) qui n'a pas été correctement identifié, du fait d'un cas particulier dans l'analyse syntaxique. On trouve également une vingtaine d'occurrences anaphoriques et cataphoriques avec déterminant défini ou indéfini, comme dans *« La question est posée. »* (anaphorique) ou *« Revenons sur les faits. »* (cataphorique).

## 5 Conclusion et perspectives

Dans cet article, nous nous sommes intéressés à une classe d'items peu étudiée dans le cadre du Traitement Automatique des Langues et dans le cadre de l'analyse du discours : les noms sous-spécifiés. Nous avons essentiellement détaillé les étapes d'une identification satisfaisante des occurrences de ces noms en corpus, à savoir : l'acquisition d'un lexique d'items pouvant être des NSS dans certains de leurs emplois ; l'identification de constructions syntaxiques dans lesquelles ces items sont susceptibles d'avoir un emploi en tant que NSS. Cette identification s'appuie sur une méthode de fouille de données : l'extraction de motifs séquentiels émergents. Ces motifs émergents, constituant des patrons syntaxiques, peuvent ensuite être analysés et validés manuellement. Ces premières étapes nous ont permis (notamment par l'intégration des patrons syntaxiques identifiés à l'aide des motifs émergents) d'obtenir un système d'identification automatique des NSS en corpus, que nous avons évalué sur 600 occurrences de noms du lexique. Le résultat de cette évaluation est encourageant, puisque la précision de l'identification se situe aux alentours de 0,8, le rappel est de 0,61, et le F-score aux alentours de 0,7.

Les NSS, et plus précisément les séquences de NSS, sont des signaux potentiels de ce que nous appelons des *séquences organisationnelles*, telles que *problème–solution*. Notre objectif est maintenant d'utiliser notre système d'identification des NSS pour extraire des séquences organisationnelles fréquentes en corpus, afin d'effectuer une analyse linguistique de ces séquences, d'étudier les interactions entre ces séquences et les relations de discours, de clarifier le rôle des NSS dans l'organisation discursive, ainsi que leur statut au sein des indices de la structure discursive. L'identification automatique des séquences organisationnelles, qui peuvent a priori elles-mêmes être des signaux de relations de cohérence, pourrait permettre d'améliorer certaines tâches liées à l'analyse automatique du discours, l'extraction d'information, etc. L'extraction des séquences fréquentes devra tenir compte d'informations que nous avons jusqu'ici ignorées : les informations sur la position des noms au sein du document, au sein du paragraphe, etc. Les connecteurs discursifs seront également intégrés aux séquences organisationnelles.

## Remerciements

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme Investissements d'Avenir portant la référence ANR-10-LABX-0083 et du projet Hybride ANR-11-BS02-002.

## Références

AGRAWAL, R. et SRIKANT, R. (1995). Mining Sequential Patterns. *In Proceedings of the Eleventh International Conference on Data Engineering*, ICDE '95, pages 3–14, Washington, DC, USA. IEEE Computer Society.

APOTHÉLOZ, D. (2008). À l'interface du système linguistique et du discours : l'exemple des constructions identificatives (e.g. pseudo-clivées). *Discours, diachronie, stylistique du français. Études en hommage à Bernard Combettes*, pages 75–92.

ASHER, N. (1993). *Reference to Abstract Objects in Discourse*. Kluwer.

ASHER, N. et LASCARIDES, A. (2003). *Logics of Conversation*. Cambridge University Press.

BÉCHET, N., CELLIER, P., CHARNOIS, T. et CRÉMILLEUX, B. (2012). Discovering linguistic patterns using sequence mining. *In Proceedings of Springer LNCS, 13th International Conference on Intelligent Text Processing and Computational Linguistics – CICLing'2012*, volume 1, pages 154–165.

CANDITO, M.-H., JOAKIM, N., DENIS, P. et HENESTROZA ANGUIANO, E. (2010). Benchmarking of Statistical Dependency Parsers for French. *In Proceedings of COLING'2010*, Beijing, China.

DENIS, P. et SAGOT, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. *In PACLIC 2009*, Hong-Kong, China.

DONG, G. et LI, J. (1999). Efficient Mining of Emerging Patterns : Discovering Trends and Differences. *In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, pages 43–52, New York, NY, USA. ACM.

DONG, G. et PEI, J. (2007). *Sequence Data Mining*, volume 33 de *Advances in Database Systems*. Kluwer.

FLOWERDEW, J. (2003). Signalling Nouns in Discourse. *English for Specific Purposes*, 22:329–346.

FLOWERDEW, L. (2008). *Corpus-based Analyses of the Problem-Solution Pattern*. Studies in Corpus Linguistics 29. John Benjamins, Philadelphia.

HIGGINS, F. R. (1979). *The pseudo-cleft construction in English*. Garland, New York.

KOLHATKAR, V., ZINSMEISTER, H. et HIRST, G. (2013a). Annotating Anaphoric Shell Nouns with their Antecedents. *In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 112–121, Sofia, Bulgaria.

KOLHATKAR, V., ZINSMEISTER, H. et HIRST, G. (2013b). Interpreting Anaphoric Shell Nouns using Antecedents of Cataphoric Shell Nouns as Training Data. *In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 300–310.

LEGALLOIS, D. (2006). Quand le texte signale sa structure : la fonction textuelle d'une certaine catégorie nominale. *Corela*.

LEGALLOIS, D. (2008). Sur quelques caractéristiques des noms sous-spécifiés. *Scolia*, 23:109–127.

LEGALLOIS, D. et GRÉA, P. (2006). L'objectif de cet article est de... construction spécificationnelle et grammaire phraséologique. *In* LECOLLE, M. et LEROY, S., éditeurs : *Changement linguistique et phénomènes de fixation : figement, lexicalisation, catachrèse*, volume 46 de *Cahiers de praxématique*, pages 161–184. Montpellier : Publications Montpellier 3.

MANN, W. et THOMPSON, S. (1988). Rhetorical structure theory : Towards a functional theory of text organization. *Text*, 8:243–281.

NIVRE, J., HALL, J. et NILSSON, J. (2006). MaltParser : A Data-Driven Parser-Generator for Dependency Parsing. *In Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*, pages 2216–2219, Genoa, Italy.

QUINIOU, S., CELLIER, P., CHARNOIS, T. et LEGALLOIS, D. (2012). Fouille de données pour la stylistique : cas des motifs séquentiels émergents. *In Actes des Journées Internationales d'Analyse Statistique des Données Textuelles (JADT'12)*, Liège, Belgique.

SCHMID, H.-J. (2000). English Abstract Nouns As Conceptual Shells : From Corpus to Cognition. *Topics in English Linguistics*, 34.

SWALES, J. (1981). *Aspects of Article Introductions*. Birmingham : University of Aston.

UPTON, T. A. et COHEN, M. A. (2009). An approach to corpus-based discourse analysis : The move analysis as example. *Discourse Studies*, 11(5):585–605.

VERGEZ-COURET, M., BRAS, M., PRÉVOT, L., VIEU, L. et ATALLAH, C. (2011). Discourse contribution of enumerative structures involving 'pour deux raisons' (regular paper). *In* ASHER, N. et DANLOS, L., éditeurs : *Constraints in Discourse (CID), Agay, France*. INRIA.

YAN, X., HAN, J. et AFSHAR, R. (2003). CloSpan : Mining Closed Sequential Patterns in Large Datasets. *In SDM*, pages 166–177.