Rapport n°3 – novembre 1998

Repérage de variantes dérivationnelles de termes

Cécile Fabre*

Repérage de variantes dérivationnelles de termes

* Equipe de Recherche en Syntaxe et Sémantique (UMR 5610 - CNRS), Université de Toulouse-Le Mirail; email: cfabre@univ-tlse2.fr

Cécile Fabre
ERSS, Université Toulouse-Le-Mirail
5 allées A. Machado, 31058 Toulouse Cedex
cfabre@univ-tlse2.fr

Résumé : Notre objectif est d'identifier un cas particulier de variation terminologique qui, à un terme nominal (ex : *augmentation de production*, *méthode d'estimation*) fait correspondre une séquence composée d'un élément verbal dérivé d'un des noms du terme de départ (ex : *augmente la production*, *estimés par une méthode*). Cette étude a pour objet l'acquisition automatique de terminologie et se base sur les résultats du logiciel FASTER, dédié au repérage de variantes terminologiques. Nous réalisons une étude linguistique des variantes extraites par FASTER, et proposons une réécriture des règles utilisées pour le repérage de ces variantes : nous montrons que des règles fondées sur des critères morpho-syntaxiques sont insuffisantes et doivent être enrichies pour contrôler la validité des variantes sur le plan sémantique.

1. Objectif : repérer la variation

Savoir repérer et mettre en relation sous des formulations distinctes l'expression d'une même "idée" est un objectif partagé par diverses applications du Traitement automatique des Langues (TAL). Il s'agit de déterminer qu'on a affaire à une même information, exprimée à travers des réalisations linguistiques diversifiées. Les types de paraphrase envisagés et la nature des informations linguistiques prises en compte sont multiples : dans le domaine de la recherche d'informations, pour augmenter les chances d'appariement entre une requête et les textes de la base à explorer, on cherche à mettre en relation des séquences textuelles qui partagent des concepts sémantiquement voisins, en exploitant des relations de proximité sémantique (principalement l'hyperonymie et la synonymie) ; on cherche parfois à prendre en compte la distribution argumentale des mots, ce qui permet par exemple de repérer un rapport de paraphrase entre des syntagmes comme *evaluating computer performance*, *the performance of information retrieval systems can be evaluated*, et le composé *performance evaluation* [Gay et Croft 89, Pugeault et al. 94, Fabre 96]. En acquisition de terminologie, qui est le domaine d'application que nous visons dans un premier temps, les termes d'un domaine subissent des variations linguistiques importantes et doivent être repérés sous leurs différents modes d'apparition. FASTER [Jacquemin 97] est un logiciel dédié au repérage de variantes terminologiques de nature morpho-syntaxique. Par le biais de métarègles, il repère par exemple une relation de variation entre les groupes *développement du squelette* et *développement squelettique* (dérivation morphologique), *fixation de l'azote* et *fixation biologique de l'azote* (variation syntaxique de type insertion). Nous cherchons ici à prolonger les possibilités de FASTER en nous intéressant à un type de variation basé sur une relation de dérivation intercatégorielle (`nom` $\rightarrow$ `verbe`), qui s'établit entre des groupes nominaux (*fixation de l'azote*, *porteur du gène*) et verbaux (*fixer l'azote*, *portant le gène*). Les critères morpho-syntaxiques utilisés par le logiciel s'avèrent insuffisants pour faire la part entre des variantes sémantiquement correctes (*porteur du gène* / *portant le gène*) et incorrectes (ex : *méthode d'utilisation* / *utilise une méthode*). Notre but est de proposer des critères pour expliciter et formaliser ce jugement de correction et filtrer les variantes valides, tout en respectant un des prérequis de FASTER, à savoir l'utilisation d'indices formels minimaux, requérant peu de connaissances lexicales, et donc peu coûteux en termes de codage informatique.

Il s'agit d'étendre les patrons de variation que FASTER est capable de repérer, mais aussi d'étudier par ce biais les possibilités d'acquisition de terminologie de nature verbale. Nous proposons donc une méthode pour accéder à des unités verbales, sans toutefois nous prononcer dans le cadre de cet article sur leur statut terminologique. Les systèmes d'acquisition terminologique se sont en effet principalement concentrés sur la recherche de termes nominaux (c'est le cas d'Acabit [Daille 95], ou de Lexter [Bourigault 94]), défendant l'idée selon laquelle la dénomination et la construction de taxonomies est réalisée essentiellement à l'aide de noms et de leurs expansions. Dans Lexter, les formes verbales sont ainsi considérées comme des marques de frontière pour délimiter les candidats termes. Dans un deuxième temps, il est certain qu'on ne peut plus exclure ainsi de la description terminologique les relations verbales, qui contribuent à structurer la terminologie d'un domaine. Si

l'on considère le groupe *analyse de croissance* comme un terme, il est indispensable de savoir qu'il peut se retrouver sous une forme verbale dans le groupe *analyser la croissance* ; de même pour *calcul de coefficient* / *coefficient a été calculé, variabilité de rendement* / *rendements varient*, etc. Plutôt que d'envisager l'acquisition autonome de groupes verbaux, nous nous appuyons sur un ensemble de termes nominaux déjà recensés pour l'étendre du côté du vocabulaire verbal en utilisant des principes de variation dérivationnelle. Cette approche ne permet pas de couvrir tous les syntagmes verbaux d'un domaine spécialisé parce que bon nombre d'entre eux ne sont pas liés à une forme nominale. Cette restriction n'affaiblit cependant pas la méthode car même si l'on disposait d'un extracteur de syntagmes verbaux, la possibilité de regrouper les syntagmes verbaux correspondant à la même forme nominale serait un complément indispensable. En effet, tant en acquisition de connaissances qu'en recherche d'informations, il est nécessaire de regrouper les formes sémantiquement apparentées afin de remédier à l'éparpillement des données.

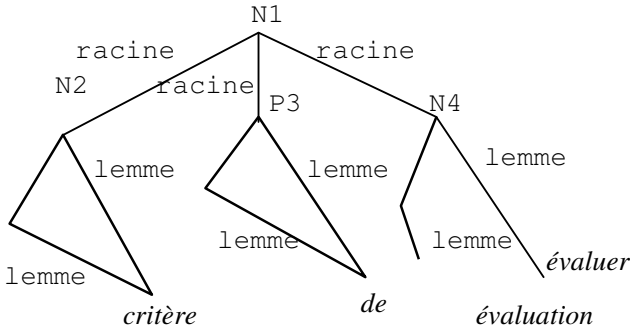
Après une description des principes de fonctionnement de FASTER et en particulier des métarègles pour l'acquisition de variantes de termes, nous étudions les couples terme / variante extraits par le logiciel selon des critères morpho-syntaxiques. Nous montrons ensuite que la validité d'une variante doit être évaluée en termes sémantiques, dans la mesure où la préservation de la relation argumentale est requise pour que l'on puisse parler de variante valide. Cela nous amène enfin à proposer une réécriture des métarègles, en intégrant de nouveaux traits dans la description des termes et de leurs variantes.

2. L'extraction de variantes nomino-verbales par FASTER

FASTER est un analyseur syntaxique reposant sur une grammaire d'unification. La description grammaticale des termes est faite de deux parties : un squelette hors contexte donnant la structure syntaxique du terme et un ensemble d'(in)équations décrivant les informations attachées aux feuilles lexicales[1]. Ainsi, la règle présentée à la page suivante[2] décrit le terme *critère d'évaluation*.

Chaque feuille est garnie d'un lemme identifié par le trait `<lemme>`. Avec la catégorie (implicitement donnée par l'étiquette de la feuille), le lemme définit de façon unique chaque feuille lexicale d'un terme. Afin de pouvoir travailler sur les liens morphologiques, les feuilles lexicales sont également garnies d'un trait `<racine lemme>` qui donne le lemme de la racine. Dans le terme précédent, seul *évaluation* n'est pas sa propre racine. Sa racine est le verbe *évaluer*.

Règle N1 → N2 P3 N4:
<N2 lemme> = *critère*
<N2 racine lemme> = *critère*
<N2 lemme> = *de*
<N2 racine lemme> = *de*
<N3 lemme> = *évaluation*
<N2 racine lemme> = *évaluer*.



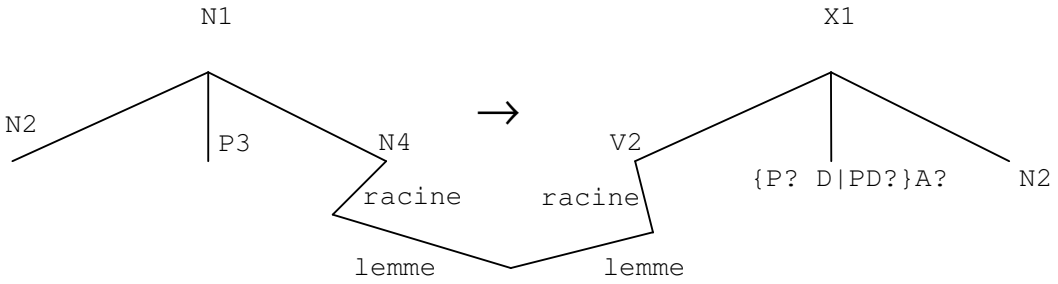
[1] Les abréviations employées doivent se lire : `N` pour `Nom`, `V` pour `Verbe`, `A` pour `Adjectif`, Av pour `Adverbe`, `P` pour `Préposition`, `D` pour déterminant, et `C` pour `Conjonction`.

[2] Cette métarègle est appelée règle $N_xPN_y \rightarrow V_xN_y$ dans la suite de l'article : cela signifie qu'elle relie un terme de structure `NPN` et une variante de structure `VN` (les éléments optionnels ne sont pas reproduits dans le nom de la règle) ; la coindexation indique l'égalité des lemmes lorsqu'il s'agit de deux noms (N_y et N_y en partie gauche et droite de la règle) , et la relation de dérivation entre un nom et un verbe (N_x et N_y de part et d'autre).

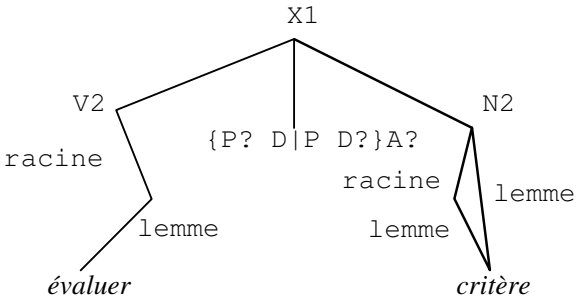
Aux règles ayant cette structure correspond un ensemble de métarègles dont le rôle est de produire de nouvelles règles décrivant les variantes. Par exemple, la métarègle suivante transforme un terme de structure `Nom Préposition Nom` en un syntagme verbal (transformation nomino-verbale, notée `NtoV`) dont la tête est morphologiquement liée à celle du syntagme prépositionnel. Le schéma de la règle initiale est donné par la partie gauche de la métarègle, le schéma de la variation par la partie droite.

```
NtoV(N1 → N2 P3 N4) = X1 → V2 <{P? D | P D?} A?> N2 :
      <V2 racine lemme> = <N4 racine lemme>
```

En voici une illustration graphique :



Les règles transformées sont obtenues en unifiant la partie gauche de la métarègle avec le terme initial (*critère d'évaluation* dans l'exemple précédent). Les traits présents dans le nom `N2` sont transmis dans la partie droite de la métarègle en raison de la présence du symbole `N2` dans cette partie. Des traits de `N4`, seul le trait donnant le lemme de la racine est transmis en partie droite. Après unification de la partie gauche avec le terme *critère d'évaluation* et recopie de la partie droite, on obtient la variante suivante :



Cette variante peut se commenter ainsi : elle repère une structure syntaxique `V P? D A? N` ou `V P D? A? N`, où le verbe a pour racine *évaluer* et où le nom est *critère*. Le `?` marque l'optionalité du constituant qu'il suit, et la barre verticale `|` l'alternative. En particulier, cette règle transformée reconnaît la séquence *évalué selon certains critères* comme une variante correcte *de critère d'évaluation*. C'est le deuxième patron syntaxique qui est choisi avec, comme verbe de racine *évaluer*, le verbe *évaluer* lui-même, *selon* comme préposition et *certains* comme article. L'adjectif antéposé optionnel est absent.

Cette recherche d'informations minimales pour le filtrage des variantes n'est possible que parce qu'elles sont liées à des termes de structure nominale par des transformations morpho-syntaxiques particulières. Les deux mots pleins constituant la variante verbale sont donc préalablement trouvés (à des variantes morphologiques près) en association dans la construction d'un terme complexe, donc dans une relation de dépendance lexicale. La transformation syntaxique est choisie en espérant que la dépendance existant dans le terme nominal est conservée dans la forme verbale. Ce travail est fait dans l'esprit de la linguistique harrissienne qui recherche les transformations préservant le contenu informationnel des documents et les ramène à des formes

noyaux. [Harris et al. 90] considèrent par exemple qu'en français, les formes *les plasmocytes sont producteurs d'anticorps, les plasmocytes produisent des anticorps*, *la production plasmocytaire d'anticorps*, transmettent une même information. Tout le problème de notre approche consiste donc à définir des liens de transformations syntaxiques entre une forme nominale et une forme verbale en espérant que les deux formes pourront être effectivement rapprochées du point de vue informationnel.

3. Étude préalable des données

En prenant en compte le phénomène de la variation dérivationnelle, nous élargissons la portée des règles de découverte de la variation terminologique. Les variantes que nous allons étudier sont couvertes par plusieurs catégories de métarègles, que nous allons présenter. Toutes relèvent du cas où les deux groupes (le terme et sa variante) sont liés par une relation de dérivation nomino-verbale. Le nom tête ou le nom modifieur du terme de départ est morphologiquement apparenté à la forme verbale de la variante. Par exemple :

injection de la solution (terme) → *injecter une solution* (variante)

L'expérience menée consiste à définir les contraintes qui permettent de considérer la séquence comme une variante sémantiquement valide du terme nominal attesté.

3.1. Description de l'expérience menée

Notre corpus est constitué de 5883 associations terme-variante, obtenues par application de règles de variation basées sur des critères morphosyntaxiques. Ces variantes ont été extraites d'une collection de résumés d'articles scientifiques dans le domaine de l'agriculture et mis à disposition par l'INIST dans le cadre du projet ILIAD (GIS Sciences de la cognition). Ces données, contenant 1,2 million de mots, ont été préalablement étiquetées au moyen de l'analyseur morphologique et du désambiguïsateur du français développés au Laboratoire Bell, Lucent Technologies [Tzoukermann et Radev 96]. Les patrons syntaxiques ayant permis de récolter ces associations terme / variante ont été mis au point par une double démarche modélique et empirique. L'aspect introspectif de la démarche consiste à représenter les associations groupe nominal / groupe verbal telles qu'elles sont décrites dans les études linguistiques comme [Harris et al. 90] ou [Mel'≠uk 84]. La mise au point expérimentale consiste à enrichir les transformations conçues sur les modèles linguistiques par l'observation des occurrences trouvées en corpus. Afin de découvrir ou d'enrichir des patrons syntaxiques, nous collectons des variantes très lâches correspondant à des cooccurrences dans une fenêtre textuelle. Certaines de ces occurrences correspondent à des variantes correctes qui ne sont pas couvertes par les patrons créés *a priori*. La démarche expérimentale consiste alors à enrichir les patrons existants ou à créer de nouvelles transformations afin de décrire ces variations non prises en compte.

Les résultats reportés dans cet article sont basés sur l'étude de 1050 associations. Le terme est un groupe nominal de structure `Nom Préposition Nom` (les unités sont lemmatisées). La variante est un groupe contenant une forme verbale, dans lequel on retrouve les deux mots pleins du terme, soit directement, soit sous une forme dérivée : dans ce dernier cas, le verbe de la variante est morphologiquement lié à un nom du terme. Voici trois exemples mêlant des variantes correctes et incorrectes.

terme	variante	nom de la métarègle
(1) *détermination de le structure*	*déterminer la structure*	$N_xPDN_y \rightarrow V_xN_y$
(2) *développement de le méthode*	*méthode de mesure développe*	$N_xPDN_y \rightarrow N_yV_x$
(3) *méthode de utilisation*	*méthodes de mesure indépendantes utilisées*	$N_xPN_y \rightarrow N_xV_y$

Par exemple, les métarègles illustrées en (1) et (3) produisent des groupes qui maintiennent l'ordre des deux mots, alors que la métarègle en (2) les inverse. Dans le cas de (1), la tête du groupe est le verbe (on obtient un groupe verbal), alors que la forme verbale est en position finale dans les

deux autres cas (on obtient un verbe et son sujet en (2), un groupe nominal en (3)). Sur ces exemples, on constate que les règles ne permettent pas toujours de repérer des variantes correctes du terme : si la séquence *déterminer la structure* peut être considérée comme une variante correcte du terme *détermination de la structure*, les deux autres variantes introduisent des notions distinctes du terme de départ. On peut dire dans ces deux cas que la variation morpho-syntaxique n'a pas permis de repérer une variation sémantique. Notre objectif est donc de contrôler l'application de ces règles pour obtenir des séquences associées aux termes de départ, et pour caractériser la relation sémantique qui est induite. Cela revient à éliminer le bruit produit par des règles morpho-syntaxiques trop permissives, pour leur substituer des règles valides sémantiquement.

3.2. Les métarègles utilisées

Nous donnons ici l'ensemble des règles de transformation `nom → verbe`, notées `NtoV`, qui ont permis de générer les couples terme / variante étudiés. Les catégories non optionnelles sont notées en gras pour faciliter la lecture des métarègles.

3.2.1. **règle $N_xPDN_y \rightarrow V_xN_y$**

```
NtoV( N1 -> N2 P3 Dd4 N4 ) = X1 -> V2 <Av? {P? D | P} A?> N4:
                               <V2 der ref> = <N2 ref>
```

Le terme nominal est constitué d'un nom tête suivi d'un groupe prépositionnel pourvu d'un déterminant défini (Dd4). Le verbe de la variante est dérivé du nom tête, il est suivi d'un nom, (N4), identique au deuxième nom du terme de départ (adverbe, préposition, déterminant et adjectif sont facultatifs, mais on trouve obligatoirement soit une préposition, soit un déterminant). Voici quelques exemples de séquences instanciant cette métarègle :

caractérisation de l'activité / caractérise l'activité (`V2 D N4`)
mutation du gène / mutées dans les gènes (`V2 P D N4`)
utilisation de l'eau / utilisent généralement l'eau (`V2 Av D N4`)

3.2.2. **règle $N_xPN_y \rightarrow V_xN_y$**

```
NtoV( N1 -> N2 P3 N4 ) = X1 -> V2 <Av? {P? D | P} A?> N4:
                           <V2 der ref> = <N2 ref>
```

L'absence de déterminant défini dans le terme initial est le seul élément qui distingue cette mérarègle de la précédente.

Ex : *amélioration de rendement / améliore le rendement* (`V2 D N4`)
analyse en chromatographie / analysant par chromatographie (`V2 P N4`)

3.2.3. **règle $N_xPDN_y \rightarrow N_yV_x$**

```
NtoV( N1 -> N2 P3 Dd4 N5 ) = X1 -> N5 <A?
                               <P D? A? N <A <C A>?>?>?
                               <C D? Av? A? N A?>? V? V? Av?>V2:
            <V2 der ref> = <N2 ref>
```

Le terme nominal est constitué d'un nom tête suivi d'un groupe prépositionnel avec déterminant défini. La tête de la variante équivaut au nom expansion du terme de départ (`N5`) ; le verbe de la variante (`V2`) est dérivé du nom tête (`N2`). La règle accepte différents éléments optionnels.

Ex : *application de l'analyse / analyse est appliquée* (`N5 V2`)
utilisation de l'analyse / analyses rapides utilisant (`N5 A V2`)
calcul de le coefficient / coefficient instantané de mortalité naturelle a été calculé

(N5 A P N A V2)

3.2.4. règle $N_xPN_y \rightarrow N_yV_x$

```
NtoV( N1 -> N2 P3 N4 ) = X1 -> N4 <A?
                            <P D? A? N <A <C A>?>?>?
                            <C D? Av? A? N A?>? V? V? Av?>V2:
                            <V2 der ref> = <N2 ref>
```

Là encore, cette métarègle est identique à la première, à la présence du déterminant près dans le terme d'origine.

Ex : *apport d'engrais / engrais apporté* (N4 V2)
détermination de sensibilité / sensibilités des glossines saines déterminent (N4 P N A V2)

3.2.5. règle $N_xPN_y \rightarrow V_yN_x$

```
NtoV( N1 -> N2 P3 N4 ) = X1 -> V2 <{P? D | P D?} A?> N2:
                                <V2 der ref> = <N4 ref>
```

Le terme nominal est constitué d'un nom tête suivi d'un groupe prépositionnel sans déterminant. Le verbe de la variante est dérivé du nom expansion (N4), et suivi du nom tête du terme de départ (N2).
Ex : *processus d'induction / induit des processus* (V2 D N2)
critère d'évaluation / évalués selon les critères (V2 P D N2)

3.2.6. règle $N_xPN_y \rightarrow N_xPV_y$

```
NtoV( N1 -> N2 P3 N4) = X1 -> N2 < <Av? A <C Av ? A>? >?  V? P> V5:
                            <V5 der ref> = <N4 ref>
```

Le terme est constitué d'un nom tête suivi d'un groupe prépositionnel sans déterminant. Le verbe de la variante est dérivé du nom expansion (N4). Il est précédé du nom tête du terme de départ (N2) et d'une préposition (P).

Ex : *méthode de caractérisation / méthode pour caractériser* (N2 P V5)
type d'utilisation / type "bleu" et "emmental" en utilisant (N2 A C A P V5)

3.2.7. règle $N_xPN_y \rightarrow N_xV_y$

```
NtoV( N1 -> N2 P3 N4) = X1 -> N2 <<A? {V? | P D?} <Av? A>? Av?>> V5:
           <V5 der ref> = <N4 ref>
```

Par rapport à la métarègle précédente, la préposition est optionnelle dans la variation.

Ex : *plante de traitement / plantes traitées* (N2 V5)
système de production / systèmes enzymatiques produisant (N2 A V5)

3.3. Description linguistique des séquences extraites

Les séquences nominales concernées sont des séquences *Nom Prép Nom.* Le nom tête du syntagme prépositionnel peut être précédé d'un article. Dans 92% des cas, la préposition est *de* (sinon, *en*). Nous nous concentrons donc principalement sur le problème de la sémantique des séquences *Nom de Nom* (avec ou sans déterminant), dont le nom tête ou expansion est morphologiquement apparenté à un verbe. Ce nom peut être un nom déverbal processif (désignant une action, un processus), comme dans :

(1a) *colmatage de membrane*
(1b) *migration de cellule, infiltration d'eau*

En (1a), on a affaire à un déverbal processif objectif[3], complété par un groupe nominal (GN) objet (= *la membrane est colmatée*), alors que l'exemple (1b) correspond à un déverbal processif subjectif, complété par un GN sujet (= *la cellule migre*, *l'eau s'infiltre*). On trouve également des noms exprimant le résultat de l'action exprimée par le verbe (*déverbaux résultatifs*). Il peut s'agit d'un résultat concret :

(2a) *échantillon*, *conserve*, ...

ou abstrait :

(2b) *influence*, *carence*, *limite*, ...

On a souvent affaire dans cette catégorie à une conversion nomino-verbale par suppression du suffixe verbal, appelée dérivation inverse[4].

Le nom peut également être un *déverbal agentif* (le nom désigne l'agent de l'action exprimée par le verbe), comme dans :

(3) *porteur du gène*, *fixateur d'azote*

On trouve enfin des noms dérivés d'un adjectif. Dans ce cas, le lien avec le verbe est indirect et passe par une forme adjectivale intermédiaire.

(4) *variabilité*, *applicabilité, stabilité, solubilité*

Les exemples de notre corpus relèvent massivement de la catégorie (1a). Les suffixes servant à la construction de ces déverbaux processifs sont principalement :

-tion : *localisation*, *induction*, *inhibition* ...
-ment : *développement*, *éclairement*, ...
-age : *clonage*, *séchage* ,...
Ø : *analyse*, *mesure*, *recherche*, ...

Remarque : Nous ne discutons que les associations résultant d'une analyse morphologique correcte. Par conséquent, les quelques erreurs liées à la surgénération du module d'analyse morphologique (ex : *type de fromage* engendre *fromage type*, où *type* est identifié comme verbe dans la variante, entretenant une relation de dérivation avec le nom tête) ne sont pas prises en compte. On observe plusieurs cas d'erreur (qui concernent la suffixation zéro), principalement lorsque le nom n'est pas dérivé du verbe au sens où il apparaît dans le corpus (*conditions de production* / *conditionnent la production* : *condition* est lié à *conditionner* dans le sens "être la condition de", mais c'est *conditionnement* qui serait la forme dérivée correcte ici) ou lorsque le nom appartient à une autre famille lexicale, même si éventuellement une parenté morphologique a pu exister en diachronie (*présent* / *présentent*; *part*, *partie* / *partir*; *comportement* / *comportant*).

Le problème de la sémantique des séquences *N prép N*, et plus spécifiquement des séquences *N de N*, a été en particulier étudié par [Bartning 87] et [Fabre 96]. Ces études permettent de déterminer les configurations sémantiques que l'on peut trouver dans ce contexte, c'est-à-dire le type de relation qui peut exister entre le nom tête et son expansion.

(a) nom tête déverbal objectif

Le deuxième nom est l'argument objet du nom tête.
localisation du gène → action de *localiser*(objet : *gène*)

(b) nom tête déverbal subjectif

Le deuxième nom est l'argument sujet du nom tête.
migration de cellule → action de *migrer*(sujet : *cellule*)

(c) nom tête déverbal agentif

Le nom tête est l'agent de l'action, le deuxième nom est l'objet.

---

[3] La terminologie est empruntée à [Bartning 87].

[4] Le verbe est considéré comme un dénominal. Ainsi, *échantillon* a donné *échantillonner* et *influence* a donné *influencer* alors que *localiser* a donné *localisation*. On peut d'ailleurs trouver la série *échantillon*, *échantillonner*, *échantillonnage* ou *conserve*, *conserver*, *conservation.*

porteur du gène → agent de l'action de *porter*(objet : *gène*)
(d) nom expansion déverbal processif
Le nom tête peut être l'agent de l'action ou désigner un argument non essentiel (lieu, instrument, but) :
région de production → lieu de l'action de *produire*
mécanisme de contrôle → instrument de l'action de *contrôler*

Cette définition des différentes interprétations possibles pour une séquence *N prép (dét) N* va nous permettre de déterminer les variantes verbales acceptables.

4. Caractériser la variation sémantique

À partir de ces quelques données linguistiques, nous allons tenter de distinguer les variantes valides des variantes invalides, en partant de l'observation des classes de variation obtenues par l'application de chaque règle.

4.1. Les différents types de variation du point de vue sémantique

Rappelons que nous n'émettons pas d'hypothèse sur le statut terminologique des séquences extraites par transformation des termes de départ. Cependant, cette première analyse des différentes catégories de variantes nous amène à tenter de construire une typologie de la variation du point de vue linguistique. On trouve aux extrêmes les *variantes de type paraphrastique*, qui manifestent la plus grande proximité entre les deux syntagmes :
pilotage d'irrigation *piloter les irrigations*
et les *variantes impropres* :
utilisation de plante *utilisés sur jeunes plantes*
Dans ce cas, les deux séquences partagent bien sûr les mêmes mots simples (on parle dans les deux cas de *plante* et d'*utilisation*) mais la relation entre ces deux mots est différente (la plante est dans le premier cas l'objet, dans le second le support de l'utilisation, l'objet étant exprimé en dehors du syntagme).

Entre les deux, on rencontre des variantes qui semblent correspondre à des exemplaires plus spécifiques du concept de départ :

utilisation de spectrométrie	*spectrométrie d image pourrait être utilisée*
utilisation de chromatographie	*chromatographie liquide est utilisée*
application de la méthode	*méthode de sorption oscillante appliquée*

L'ajout d'un adverbe, d'un adjectif ou d'un complément du nom a pour effet de créer une *variante hyponyme*.

La variation peut éventuellement introduire un antonyme ou un concept complémentaire du terme, en particulier par le biais du jeu sur les prépositions :
plantation en sol *plantés sur sol*
Sur cet exemple, l'antonymie n'est pas avérée (elle doit être vérifiée en fonction du domaine concerné) mais elle est plausible. Ce dernier point suggère l'intérêt qu'il peut y avoir à récupérer des variantes sémantiques qui ne sont pas des équivalents du terme.
On a également affaire à des variantes qui introduisent un *déplacement de focalisation* :
air de séchage *séchées à l'air*
Les positions tête et modifieur sont échangées. Il s'agit d'une même réalité vue sous deux angles différents (le moyen d'action dans un cas, l'action et l'objet sur lequel elle porte dans l'autre).

Enfin, le terme associé peut présenter des variantes plus subtiles du terme de départ, qu'il est difficile de rapporter à des relations lexicales, mais où l'on peut voir plutôt des différences d'actualisation : changement de nombre (*repérage de plante / repérer les plantes*), changement de détermination (*exploitation de forêt / exploiter cette forêt*).
Insistons sur le fait qu'il est souvent difficile de décider si deux séquences peuvent être ramenées à un même contenu informationnel sans tenir compte du contexte ni faire intervenir des connaissances

propres au domaine. Par exemple, la règle $N_x PN_y \rightarrow N_x V_y$ produit quatre instances de la variante N2 P D N V5, V5 étant une forme de participe passé. S'agit-il de variantes correctes ?

densité de inoculation *densité de la souche parasite préalablement inoculée*
étude de consommation *étude de la puissance consommée*
méthode de dosage *méthode des ajouts dosés*
qualité de production *qualité du bois produit*

On se rend compte qu'un schéma identique produit des variantes de qualité très variable. Si l'on considère que *qualité du bois produit* peut constituer une variante hyponyme de *qualité de production*, il est par contre fort peu probable que l'on puisse rapprocher *densité d'inoculation* et *densité de la souche préalablement inoculée* (il s'agit de deux mesures différentes de la densité)[5]. Dans ce cas (qui concerne cependant un schéma de variation très peu productif), il paraît difficile de pouvoir généraliser ces observations pour aboutir à une règle capable de réaliser un filtrage adéquat.

4.2. Un critère de répartition : la préservation de la structure argumentale

Notre objectif est de formaliser la distinction entre variations paraphrastiques et variations incorrectes. À la lumière des observations que nous venons de mener, il semble que l'on puisse faire passer cette frontière entre deux catégories d'associations terme / variante : celles qui préservent le rapport argumental entre les concepts simples et celles qui le perturbent. Ce critère permet de différencier les variantes incorrectes et les variantes valides, sachant que cette deuxième catégorie est hétérogène et doit être affinée dans un deuxième temps en tenant compte des nuances que nous venons d'apporter. Il s'agit donc du paramètre principal de classement que nous allons réaliser. Dans ce qui suit, nous proposons une représentation prédicative du syntagme, dans laquelle le nom déverbal (noté N_v) ou le verbe sont représentés munis d'une structure argumentale remplie par les arguments indicés 0 (argument sujet ou externe[6]), 1 (1er argument ou argument interne 1), 2 (2ème argument ou argument interne 2) ou circ (argument non essentiel, circonstanciel).

Les cas de préservation de la structure argumentale sont au nombre de trois :
(1) Le nom argument est un argument interne du nom tête du terme et du verbe de la variante.
$N_V(arg_1) \rightarrow V(arg_1)$ (premier argument = objet)
évaluation de l'efficacité / évaluer l'efficacité
application de l'analyse / analyse est appliquée
$N_V(arg_2) \rightarrow V(arg_2)$ (deuxième argument)
application à l'étude / appliquées à l'étude
(2) Le nom argument est l'argument externe (sujet) du nom tête du terme et du verbe de la variante.
$N_V(arg_0) \rightarrow V(arg_0)$
variation de la production / production varie
fonctionnement de l'enzyme / enzyme pouvait fonctionner
(3) Le nom argument est un argument non thématique jouant le même rôle sémantique (circonstanciel) par rapport au nom tête du terme et au verbe de la variante.
$N_V(arg_{circ}) \rightarrow V(arg_{circ})$
traitement à la chaleur / traitée à la chaleur (moyen)
acclimatation en serre / acclimatées en serre (locatif)

Les deux premières catégories s'enrichissent d'un cas particulier : dans le terme, la relation argumentale peut être sous-spécifiée. La variante permet de décider si l'agent est ou non explicité :
$N_V(arg_0|arg_1) \rightarrow V(arg_1)$

---

[5] Merci à Éric Gaussier pour ces remarques sur ce point.
[6] Les notions d'argument thématique (externe ou interne) et non thématique, que l'on peut faire correspondre aux notions traditionnelles d'argument essentiel (sujet ou complément) et non essentiel, sont empruntées à la terminologie générativiste (cf. en particulier [Grimshaw 90]).

variation du temps / varier le temps
augmentation de concentration / augmente la concentration

$N_V(arg_0|arg_1) \rightarrow V(arg_0)$
augmentation de l'intensité / intensité augmente
arrêt de croissance / croissance s'arrête

Les cas de distribution argumentale non respectée constituent les configurations complémentaires de celles que nous venons de présenter (l'astérisque marque le fait que les deux séquences ne sont pas dans un rapport de variation valide).

$N_V(arg_1) \ *\rightarrow V(arg_0)$
détermination de facteur / facteurs déterminent

$N_V(arg_1) \ *\rightarrow V(arg_2)$
application de l'analyse / appliqué à l'analyse

$N_V(arg_1) \ *\rightarrow V(arg_{circ})$
réalisation du modèle / réalisée sur modèle (manière)
utilisation de l'analyse / utilisée pour l'analyse (but)
introduction du gène / introduit dans le gène (locatif)

$N(arg_{nonthém}) \ *\rightarrow V(arg_0)$
taux d'augmentation / taux augmente

$N(arg_{nonthém}) \ *\rightarrow V(arg_1)$
méthode d'utilisation / utilise une méthode

Ces différents patrons sémantiques s'illustrent dans les métarègles que nous avons étudiées. Nous allons donc tenter de ne retenir que les configurations sémantiquement valides dans l'ensemble des variations morpho-syntaxiques produites.

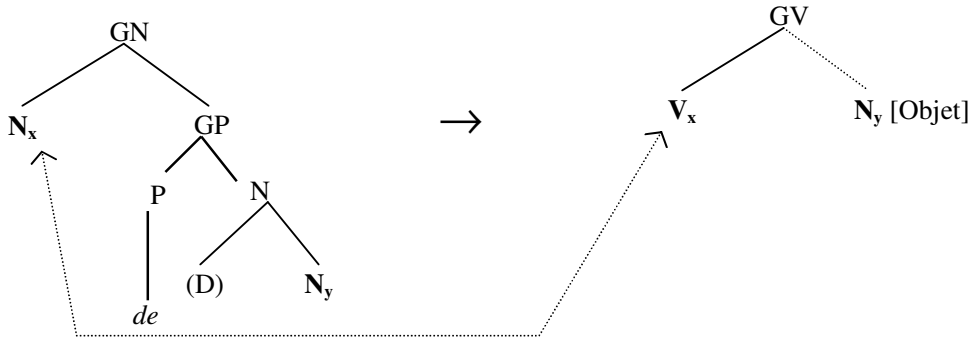
4.3. Relecture des données selon le critère de la structure argumentale

Nous proposons une nouvelle lecture, sémantique cette fois, des termes et de leurs variantes. Du point de vue de l'illustration sous forme graphique, nous substituons à la représentation morpho-syntaxique une représentation sémantique, simplifiée, pour illustrer les cas de variation valide.

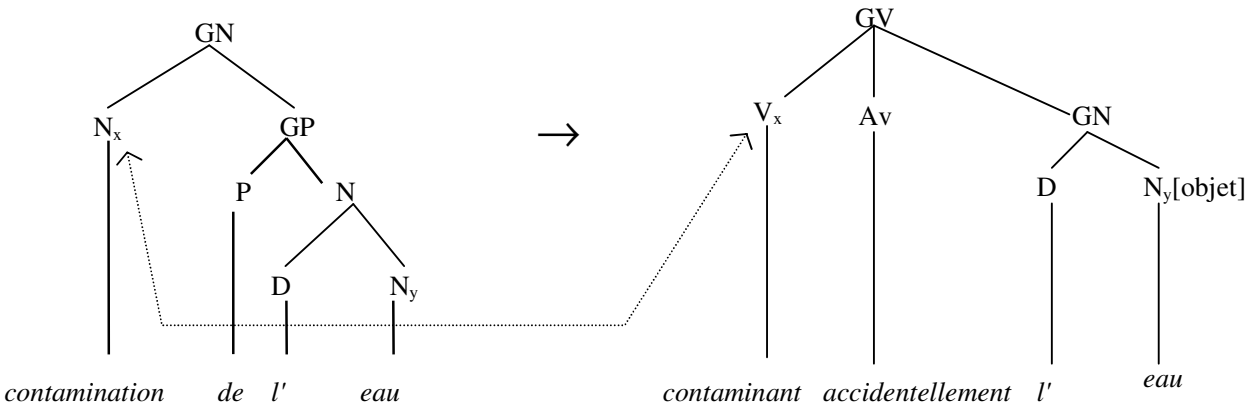
♦ **Métarègles : $N_xPDN_y \rightarrow V_xN_y$ et $N_xPDN_y \rightarrow V_xN_y$**

Il faut distinguer deux cas de variation valide : (a) lorsque la préposition dans le terme initial est *de*, le nom de la variante doit être l'argument objet du verbe. Dans le cas contraire, (b) seules les variations comportant la même préposition que le terme initial sont valides.

(a) La préposition du terme initial est *de*



Dans la représentation graphique, l'arbre de la variante est simplifié, puisque nous ne représentons que les éléments obligatoires. D'autres éléments (adjectifs, déterminants, etc.) peuvent être présents dans les différentes instanciation de la règle ainsi schématisée. La branche en pointillée indique que le nom est situé dans un certain rapport de dépendance au verbe (ici la relation objet) au sein du groupe verbal, mais que des insertions peuvent se produire. Voici une instance possible de ce schéma :

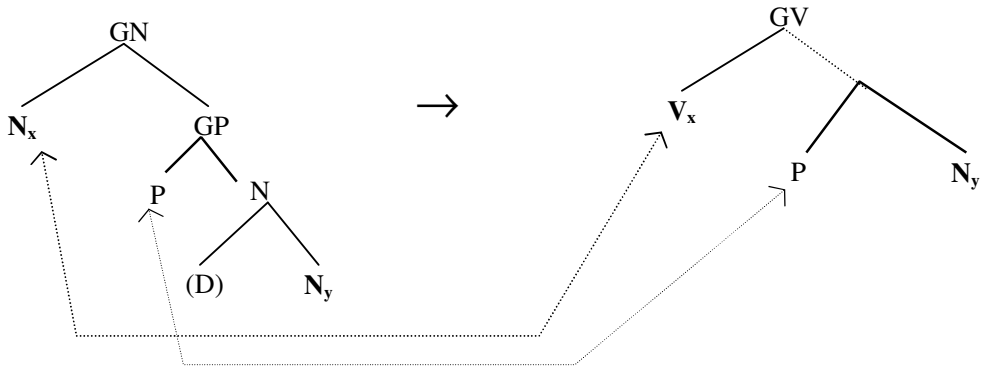


Dans ces deux métarègles, qui se distinguent uniquement par la présence ou non d'un déterminant dans le terme, on trouve de nombreux cas de paraphrase :

amélioration de l'efficacité *améliore l'efficacité*
amélioration de rendement *améliore le rendement*

Dans ce cas, la séquence *verbe* + *objet* répond à la séquence *déverbal* (N_v) + *objet*.
L'insertion d'un adverbe (*contamination de l'eau* / *contaminant accidentellement l'eau*) ou d'un adjectif (*utilisation de système* / *utilisant un nouveau système*) produit des variantes plus spécifiques du terme.

(b) La préposition du terme initial n'est pas *de*



Dans ce cas, l'identité des prépositions dans le terme et la variante garantit la stabilité de la relation argumentale, qu'il s'agisse d'un argument interne prépositionnel (non objet) ou d'un argument circonstanciel :

application à l'étude *appliquée à l'étude*
utilisation en industrie *utilisés en industrie*

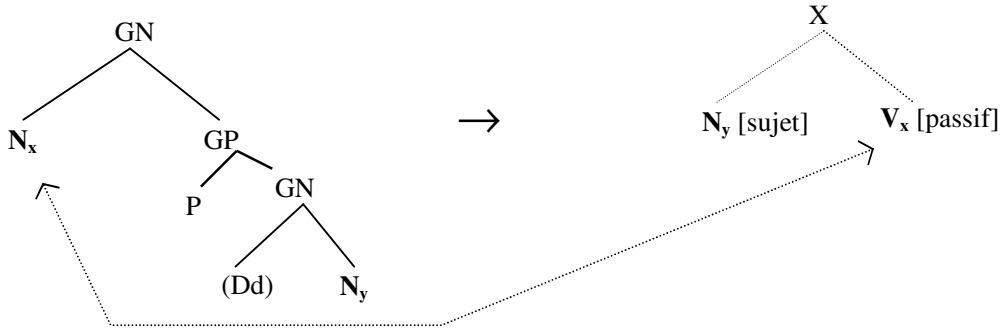
À côté de ces deux schémas, on rencontre des variations incorrectes dûes à une relation argumentale différente dans la séquence verbale :

introduction du gène *introduit dans le gène*

réalisation du modèle *réalisée sur modèle*
détermination de nombre *déterminé probablement par un grand nombre*

La présence d'une préposition dans la variante est à l'évidence un indice essentiel pour déterminer que la configuration argumentale n'est pas maintenue.

♦ **Métarègles : $N_xPDN_y \rightarrow N_yV_x$ et $N_xPDN_y \rightarrow N_yV_x$**

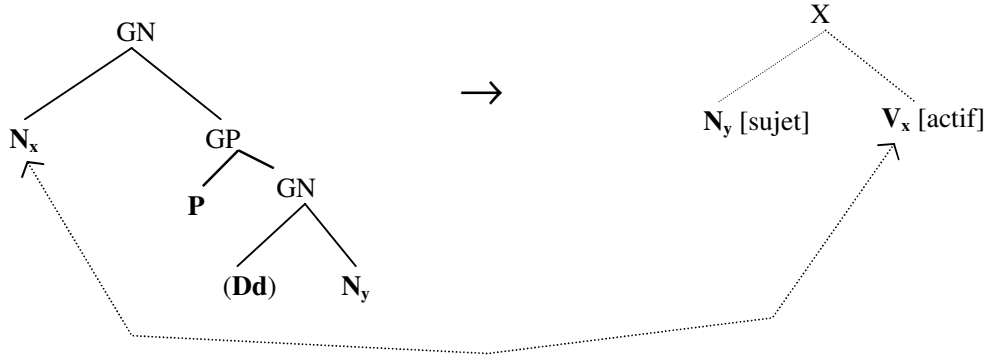


Dans ces deux catégories (différenciées par le déterminant), le verbe est placé en fin de séquence. Dans le cas d'un nom tête processif objet, les variantes correctes correspondent à une forme verbale passive.

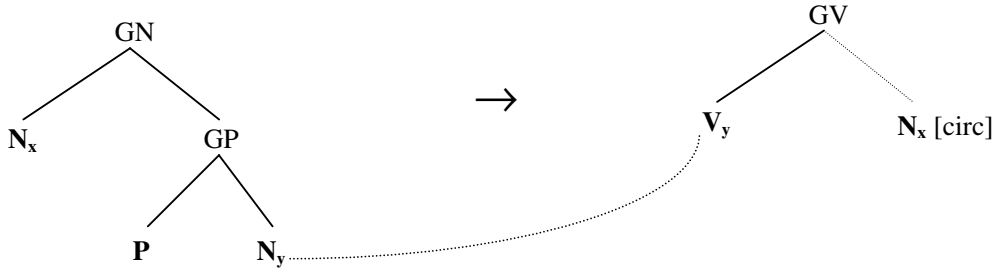
calcul de le coefficient *coefficient instantané de mortalité naturelle a été calculé*
alimentation de larve *larves de Diprion pini alimentée*

Lorsque le nom tête est un déverbal subjectif, on doit à l'inverse trouver une forme active :

augmentation de concentration *concentration en azote augmente*



♦ **Métarègle $N_xPN_y \rightarrow V_yN_x$**



Les variantes correctes du terme correspondent à une relation circonstancielle entre le verbe et son argument (moyen, manière) :

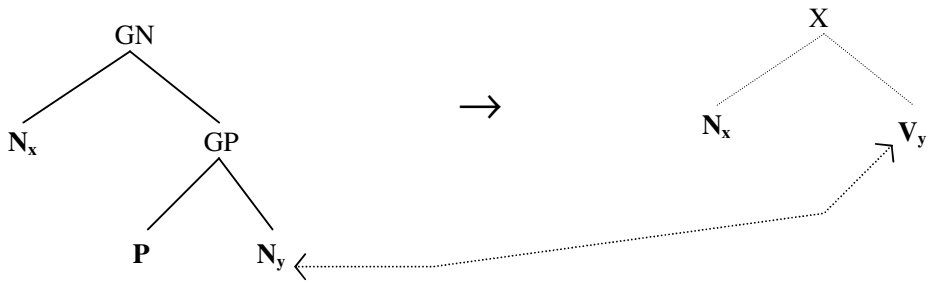
condition de test *testées dans les conditions*
méthode d'estimation *estimés par une méthode*

A l'inverse, une relation objet telle que :

modèle de développement *développer un modèle*

produit une association incorrecte.

♦ **Métarègles $N_xPN_y \rightarrow N_xPV_y$ et $N_xPN_y \rightarrow N_xV_y$**



Il est difficile de préciser ce schéma, car les cas de variation valide forment un ensemble à la fois très hétérogène et très restreint. Le déverbal est expansion de la séquence nominale. Ces deux catégories se distinguent par la présence ou non d'une préposition dans la variante. Ces cas de variation donnent lieu à des associations incorrectes, dans environ 3/4 des cas pour la première règle et 9/10ème des cas pour la deuxième :

type de utilisation *type "bleu" et "emmental" en utilisant*
méthode de utilisation *méthode utilisée*

Il est parfois délicat de décider si une variante est ou non correcte. Par exemple :

taux d'augmentation *taux de remplissage augmente*

Les variations correctes sont par exemple :

capacité de accumulation *capacité à accumuler*
méthode de détermination *méthodes utilisées pour déterminer*
modèle de simulation *modèle simulant*

5. Redéfinition des métarègles

Notre objectif est de traduire les résultats de cette analyse linguistique en termes de contraintes sur les règles de variation. Ces contraintes ne doivent pas donner lieu à une description lexicale des données linguistiques trop coûteuse d'un point de vue informatique.

5.1. Définition de traits supplémentaires

En principe, pour identifier quel schéma syntactico-sémantique est instancié par une variation, il faut vérifier un ensemble de paramètres, qui sont la satisfaction des contraintes argumentales (nombre d'arguments), le respect des restrictions sélectionnelles (type des arguments), la prise en compte de la sémantique de la préposition (apte à rendre compte de telle ou telle valeur argumentale). Dans la mesure où nous ne pouvons pas nous baser sur un lexique assez riche pour réaliser ces vérifications, nous proposons des règles reposant sur une description lexicale minimale. Les éléments-clé pour la description de la structure argumentale sont :

- la préposition
- la valence du verbe (son nombre d'arguments)
- la voix du verbe (passif ou actif).

Les contraintes supplémentaires introduites dans les règles sont de la forme :

– `<atome lemme> = chaîne_lemme` : le mot de catégorie `atome` doit avoir la valeur `chaîne_lemme`.

Ex : **<P3 lemme> = `'de'`** signifie que la préposition doit être *de*.

– `<atome classe> = symbole_classe` **:** le mot de catégorie `atome` doit prendre une des valeurs possibles de la classe mentionnée.

Ex : `<V2 voix> = actif` signifie que le verbe `V2` doit être à la voix active.

– `<atome classe> ≠ symbole_classe` **:** le mot de catégorie `atome` ne doit pas appartenir à la classe mentionnée.

Ex : **< V2 voix> ≠** `actif` signifie que le verbe `V2` ne doit pas être à la voix active.

Les classes utilisées sont les suivantes :

(a) pour les verbes :

- `voix` : elle peut prendre la valeur `actif` ou `passif`

- `forme` : elle donne la forme du verbe. Ses valeurs peuvent être `ppé`, `ppr`, `inf`, `conj` : (participe passé, participe présent, infinitif, forme conjuguée).

- `valence` : elle indique la valence du verbe (son nombre d'arguments). Ses valeurs possibles sont `intrans`, `trans`, `ditrans`, `erg`. Les intransitifs n'admettent qu'un argument (le sujet), les transitifs deux (sujet et objet) et les ditransitifs trois (sujet, objet et objet indirect). Les ergatifs sont une catégorie de verbes qui peuvent être transitifs ou intransitifs et qui acceptent le même GN comme objet des propositions transitives et comme sujet des propositions intransitives. Ce sont donc des verbes qui peuvent décrire une action du point de vue de l'agent ou du point de vue du patient : *la température augmente / on augmente la température*. Dans le corpus étudié, cela correspond à l'alternance : changement naturel, changement provoqué artificiellement. On trouve donc principalement dans cette catégorie des verbes de changement : *augmenter*, *fermenter*, *muter*, *baisser*, *saturer*, *cristalliser*, etc. La forme intransitive est plus souvent une forme pronominale : *dégrader*, *arrêter*, *régénérer*, *transformer*, *améliorer*, *dénaturer*, *développer*, *immobiliser*, *infecter*, *réguler*, etc.

(b) pour les noms :

- `dév`. Elle indique qu'on a affaire à un nom déverbal. Ses valeurs peuvent être `agentif` ou `processif`.

(3) pour toutes les catégories variables :

- `acc` : permet d'indiquer le partage des traits d'accord.

5.2. Réécriture des métarègles

♦ **métarègle $N_xPDN_y \rightarrow V_xN_y$**

règle de départ :

```
NtoV( N1 -> N2 P3 Dd4 N4 ) = X1 -> V2 <Av? {P? D | P} A?> N4:
                               <V2 der ref> = <N2 ref>
```

nouvelles règles :

règle $N_xPDN_y \rightarrow V_xN_y$ (type *dépassement du plafond / dépasser le plafond*)

```
NtoV( N1 -> N2 P3 Dd4 N4 ) = X1 -> V2 A? D A? N4:
                        <V2 der ref> = <N2 ref>
                        <P3 lemme> = 'de'
                        <N2 dev> = processif
                        <V2 voix> = actif
```

La préposition du terme est *de*, le nom tête est processif.

La variante ne comporte pas de préposition, la forme verbale doit être active (exclusion du participe passé).

règle $N_xPDN_y \rightarrow V_xPN_y$ (type *exposition à le lumière* / *exposées à la lumière*)

```
NtoV( N1 -> N2 P3 Dd4 N4 ) = X1 -> V2 Av? D A? P3 N4:
                        <V2 der ref> = <N2 ref>
                        <N2 dev> = processif
                        <P3 lemme> ≠ 'de'
```

La préposition du terme et celle de la variante doivent être identiques.

♦ **métarègle $N_xPN_y \rightarrow V_xN_y$**

règle de départ :

```
NtoV( N1 -> N2 P3 N4 ) = X1 -> V2 <Av? {P? D | P} A?> N4:
                                <V2 der ref> = <N2 ref>
```

nouvelles règles :

règle $N_xPN_y \rightarrow V_xDN_y$ (type *caractérisation de paramètre* / *caractériser les paramètres*)

```
NtoV( N1 -> N2 P3 N4 ) = X1 -> V2 <Av? D A?> N4:
                           <V2 der ref> = <N2 ref>
                           <N2 dev> = processif
                           <P3 lemme> = 'de'
                           <V2 voix> = actif
```

La tête du terme est un processif, la préposition est *de*.
La variante n'a pas de préposition, la forme verbale est à l'actif.

règle $N_xPN_y \rightarrow V_xPN_y$ (type *élevage en laboratoire* / *élevés en laboratoire*)

```
NtoV( N1 -> N2 P3 N4 ) = X1 -> V2 <Adv? P3 Dét Adj?> N4:
                        <V2 der ref> = <N2 ref>
                        <N2 dev> = processif
```

La préposition du terme et celle de la variante doivent être identiques.

♦ **métarègle $N_xPDN_y \rightarrow N_yV_x$**

règle de départ :

```
NtoV( N1 -> N2 P3 Dd4 N5 ) = X1 -> N5 <(A?
                            <P D? A? N <A <C A>?>?>?
                            <C D? Av? A? N A?>? V? V? Av?>V2:
           <V2 der ref> = <N2 ref>
```

nouvelles règles :

règle $N_xPDN_y \rightarrow N_yV_{x[passif]}$ (type *application de l'analyse* / *analyse est appliquée*)

```
NtoV( N1 -> N2 P3 Dd4 N5 ) = X1 -> N5 <(Adj?
                 <P D? A? N <A <C A>?>?>?
                 <C D? Av? A? N A?>? V? V? Av?>V2 :
           <V2 der ref> = <N2 ref>
           <P3 lemme> = 'de'
           <N2 dev> = processif
           <V2 voix> = passif
           <V2 valence> ≠ intransitif
```

La tête du terme est un déverbal processif. La contrainte d'intransitivité est un moyen de contrôler qu'on a affaire à un processif objectif (c'est une condition nécessaire mais non suffisante). Le verbe de la variante est un passif (*être ppé* ou *ppé* seul)

règle $N_xPDN_y \rightarrow N_yV_{x[actif]}$ (type *augmentation de l'intensité / intensité augmente*)

```
NtoV( N1 -> N2 P3 Dd4 N5 ) = X1 -> N5 <(A?
                 <P D? A? N <A <C A>?>?>?
                 <C D? Av? A? N A?>? V? V? Av?>V2:
                 <V2 der ref> = <N2 ref>
                 <P3 lemme> = 'de'
                 <N2 dev> = processif
                 <V2 voix> = actif
                 <V2 valence> = ergatif | intransitif
```

La tête du terme est un processif. Le verbe de la variante est à l'actif. Il doit s'agir d'un ergatif ou d'un intransitif, si l'on veut que le déverbal soit de type subjectif (sinon, on trouve par exemple la variante incorrecte : *utilisation de l'expérience* / *expérience utilisant*).

♦ **métarègle $N_xPN_y \rightarrow N_yV_x$**

règle de départ :

```
NtoV( N1 -> N2 P3 N4 ) = X1 -> N4 <(Adj?
                            <Prép Dét? Adj? N <Adj <C Adj>?>?>?
                            <C Dét? Adv? Adj? N Adj?>? V? V? Adv?>V2:
                                  <V2 der ref> = <N2 ref>
```

nouvelles règles :

règle $N_xPN_y \rightarrow N_yV_{x[passif]}$ (type *analyse de condition / conditions de culture sont analysées*)

```
NtoV( N1 -> N2 P3 N4 ) = X1 -> N4 <(Adj?
                 <P D? A? N <A <C A>?>?>?
                 <C D? Av? A? N A?>? V? V? Av?>V2:
                                  <V2 der ref> = <N2 ref>
                                  <N4 acc> = <V2 acc>
                                  <P3 lemme> = 'de'
                                  <N2 dev> = processif
                                  <V2 voix> = passif
                                  <V2 valence> ≠ intransitif
```

La tête du terme est un nom processif, la préposition est *de*. La forme verbale de la variante doit être à la voix passive.

règle $N_xPN_y \rightarrow N_yV_{x[actif]}$(type *augmentation de concentration / concentrations en glucose et fructose augmentent*)

```
NtoV( N1 -> N2 P3 N4 ) = X1 -> N4 <(A? <P D? A? N <A <C A>?>?>?
                 <C D? Av? A? N A?>? V? V? Av?> V2:
                                  <V2 der ref> = <N2 ref>
                            <P3 lemme> = 'de'
                            <N2 dev> = processif
                            <V2 voix> = actif
                            <V2 valence> = ergatif | intransitif
```

La tête du terme est un nom processif, la préposition est *de*. La forme verbale de la variante doit être à la voix active, il s'agit d'un ergatif ou d'un intransitif.

♦ **métarègle $N_xPN_y \rightarrow V_yN_x$**

règle de départ :

```
NtoV( N1 -> N2 P3 N4) = X1 -> N4 <A? <P? D | P D?} A?> N2:
                <V2 der ref> = <N2 ref>
```

nouvelle règle :
règle $N_xPN_y \rightarrow V_yPN_x$ (type *critère d'évaluation / évalués selon les critères*)

```
NtoV( N1 -> N2 P3 N4) = X1 -> N4 <A? P D? A?> N2:
                <V2 der ref> = <N2 ref>
```

La présence d'une préposition est rendue obligatoire pour éliminer les relations objet (type *processus d'induction / induit des processus*).

♦ **métarègle $N_xPN_y \rightarrow N_xPV_y$**

règle de départ :

```
NtoV( N1 -> N2 P3 N4) = X1 -> N2 < <Av? A <C Av ? A>? >?  V? P> V5:
                <V5 der ref> = <N4 ref>
```

Cette métarègle est très peu productive. On ne peut formuler que des règles très lexicalisées, qui couvrent moins d'une dizaine de cas.

nouvelles règles :
règle $N_xPN_y \rightarrow N_xPV_y$ (type *capacité de production / capacité à produire*)

```
NtoV( N1 -> N2 P3 N4) =
          X1 -> N2 < <Av? A <C Av ? A>? >?  V? P3> V5:
                <V5 der ref> = <N4 ref>
                <N2 lemme> = 'capacité' | 'possibilité'
                <N4 dev> = processif
                <P3 lemme> = 'à' | 'de'
                <V5 forme> = 'inf'
```

règle $N_xPN_y \rightarrow N_xAPV_y$ (type *méthodes de détermination / méthodes utilisées pour déterminer*)

```
NtoV( N1 -> N2 P3 N4 ) = X1 ->
N4 <A P> V2:
                <V2 der ref> = <N2 ref>
                <P3 lemme> = 'à' | 'de' | 'pour'
                <N4 dev> = processif
                <A lemme> = 'susceptible' |'permettant' | 'utilisé'
                <V5 forme> = 'inf'
```

Les variantes possibles sont de la forme : adj (de,pour) Vinf :
conditions de transformation / conditions nécessaires pour transformer
méthodes de caractérisation / méthodes susceptibles de caractériser
méthode de détermination / méthode permettant de déterminer
On pourrait également accepter *N qui conj, qui sont susceptibles de vinf*, etc :

```
NtoV( N1 -> N2 P3 N4 ) = X1 ->
     N4 C (.*) V2:
     <V2 der ref> = <N2 ref>
     <P3 lemme> = 'à' | 'de' | 'pour'
     <N4 dev> = processif
     <C lemme> = 'qui'
```

♦ **métarègle $N_xPN_y \rightarrow N_xV_y$**

règle de départ :

```
NtoV( N1 -> N2 P3 N4) = X1 -> N2 < <Adj? {V? | Prép Dét?} <Adv? Adj>?
Adv?> > V5:
<V5 der ref> = <N4 ref>
```

Nous n'avons retenu qu'un cas de variation valide, le plus productif :

nouvelle règle :

règle $N_xPN_y \rightarrow N_xV_y$ (type *modèle de simulation / modèle simulant ou système de production / systèmes enzymatiques produisant*)

```
NtoV( N1 -> N2 P3 N4) = X1 -> N2 < <A? {V? | P D? <Av? A>? Av?> V5:
      <V5 der ref> = <N4 ref>
      <V5 forme> = ppr
```

La forme verbale de la variante doit être un participe présent.

6. Conclusion

L'étude linguistique des cas de variation dérivationnelle a permis d'aboutir à la réécriture des métarègles pour contrôler le repérage de variantes terminologiques valides sur le plan sémantique. Ce travail montre comment les résultats d'une analyse linguistique peuvent être intégrés à un système informatique qui n'est pas conçu pour supporter des descriptions lexicales trop lourdes : cela suppose la définition de traits et de contraintes minimaux, dont il reste à vérifier l'efficacité sur de nouvelles données. Ce sera la prochaine étape de ce travail.

Plusieurs prolongements de cette étude sont envisagés, sur le plan terminologique, linguistique et informatique.

Du point de vue terminologique, en explicitant le lien entre formes nominales et verbales, ce travail entame une réflexion plus générale sur l'acquisition de données terminologiques de nature verbale. Sans nous confronter d'emblée à toute la complexité de la reconnaissance et du filtrage des syntagmes verbaux, nous avons montré que le regroupement d'expressions linguistiques de contenu semblable met au jour des séquences que les analyseurs terminologiques ont jusqu'à présent exclues, mais qui s'avèrent tout aussi pertinentes que les séquences nominales, puisqu'elles leur sont équivalentes du point de vue informationnel. Il reste à discuter du statut de ces séquences : s'agit-il de termes ? Peut-on envisager de façon autonome le repérage de séquences terminologiques de nature verbale, et selon quels critères ?

Du point de vue linguistique, les résultats recueillis pour cette étude sur un corpus technique fournissent des données "réelles" qui permettent d'éprouver la généralité de certaines modélisations (ici, concernant la sémantique des *N prép N*, ou la relation entre un nom déverbal et le verbe morphologiquement associé). Ce travail pourra sur le même principe se prolonger par l'étude d'autres données produites par FASTER, qui impliquent cette fois-ci la sémantique de l'adjectif. C'est le cas des couples terme / variante suivants, que nous avons écartés de ce travail car ils nécessitent une étude linguistique spécifique :

augmentation significative / augmenter de façon significative
production industrielle / produits dans les silos industriels
recherche intensive / rechercher une utilisation intensive

Plus généralement, ce travail nous a permis d'aborder concrètement le problème de la variation et de l'équivalence sémantique, mais ces notions doivent encore être précisées, ainsi que les moyens de caractériser la relation entre un terme et sa variante. Nous avons esquissé une typologie de la variation qui a montré que la distance entre le terme et sa variante, du point de vue linguistique et informationnel, fluctue très largement. En cherchant dans un premier temps à exclure les variantes incorrectes, nous ne nous sommes pas confrontée à la description des différents degrés d'équivalence

que nous avons pu relever dans les variantes retenues, et à l'influence importante du contexte et du domaine. Cette étude repose sur l'hypothèse que des critères formels, purement linguistiques, permettent de faire la part entre ce qui est équivalent du point de vue terminologique et ce qui ne l'est pas. La phase de test des règles ainsi reformulées fournira sans doute des éléments de réponse.

Sur le plan de l'application informatique, il s'agit tout d'abord d'intégrer les nouvelles métarègles dans FASTER et de les tester pour estimer leur pouvoir de filtrage. En outre, dans la mesure où ces règles comportent de nouveaux traits, il est nécessaire de concevoir des méthodes qui permettront de les acquérir automatiquement. C'est le cas de la valence des verbes, du type des noms (déverbal processif ou agentif), de la voix et de la forme des verbes. Des indices de surface, comme la terminaison des noms et des verbes, seront exploités dans ce but. Mais l'information concernant la structure argumentale des verbes est plus délicate à acquérir : doit-on considérer par exemple que la catégorie des ergatifs forme un ensemble restreint facile à lister, et qui reste stable quel que soit le domaine concerné ? Les travaux existants sur la syntaxe des verbes (en particulier dans le cadre du lexique-grammaire) fourniront certainement des informations précieuses, mais il sera nécessaire de prendre en compte les variations propres aux différents corpus techniques, ce qui nous conduirait à privilégier l'acquisition automatique sur le listage.

7. Remerciements

Ce travail a bénéficié des suggestions et des précisions apportées par Christian Jacquemin, en particulier sur les aspects techniques du fonctionnement de FASTER. Je le remercie vivement pour cette aide, pour ses relectures successives, et plus globalement, pour m'avoir confié les résultats de FASTER.
Christian Jacquemin se joint à moi pour remercier Jean Royauté et Xavier Polanco de l'INIST, ainsi que Yannick Toussaint du LORIA, coordinateur du projet ILIAD, pour avoir permis ces expérimentations. Évelyne Tzoukermann du laboratoire Bell est également remerciée pour avoir mis à disposition l'étiqueteur du français et les données lexicales qui ont servi au traitement préalable du corpus.
Les termes nominaux dont les variantes sont recherchées dans le corpus ont été extraits au moyen du logiciel Acabit. Merci donc à Béatrice Daille de l'IRIN d'avoir effectué cette extraction sur les textes fournis par l'INIST.
Je remercie enfin Éric Gaussier (Rank Xerox) et Pascale Sébillot (IRISA) pour leurs remarques sur cet article.

8. Références

Bartning, I. (1987). L'interprétation des syntagmes binominaux en 'de' en français contemporain. *Cahiers de grammaire*, vol.12, 1-64.

Bourigault, D. (1994). *LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes*. Thèse en mathématiques, informatique appliquée aux sciences de l'homme, École des Hautes Études en Sciences Sociales, Paris.

Daille, B. (1995). Repérage et extraction de terminologie par une approche mixte statistique et linguistique, *TAL*, 36, 1-2, 1995, 101-118.

Fabre, C. (1996). *Interprétation automatique des séquences binominales en anglais et en français. Application à la recherche d'informations*, thèse de doctorat en informatique, Université de Rennes I.

Gay, L.S. et Croft, W.B. (1990). Interpreting nominal compounds for information retrieval, *Information Processing and Management*, n°26, vol.1, 21-38.

Grimshaw, J. (1990). *Argument Structure*, MIT Press, Cambridge.

Harris, Z., Gottfried M., Ryckman T., Mattick Jr P., Daladier A, Harris T., Harris S., (1989). *The Form of Information in Science*, Analysis of Immunology Sublanguage, Kluwer Academic Publisher, Dordrecht, 1989.

Jacquemin, C. and Tzoukermann, E., (1997). NLP for term variant extraction: A synergy of morphology, lexicon, and syntax. In T. Strzalkowski, editor*, Natural Language Processing and Information Retrieval*. Kluwer, Boston, MA.

Jacquemin, C. (1997) *Variation terminologique : reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*, mémoire d'habilitation à diriger des recherches, IRIN, Université de Nantes.

Mel'ćuk, I. (1984). *Un nouveau type de dictionnaire : le dictionnaire explicatif et combinatoire du français contemporain*, Presses de l'Université de Montréal, Québec.

Pugeault, F., Saint-Dizier, P. et Monteil, M.-G., (1994). "Knowledge extraction from texts: a method for extracting predicate-argument structures from texts", In *Actes, 15th International Conference on Computational Linguistics*, Kyoto, Japon.

Tzoukermann, E. et Radev D.R., (1996). "Using word class for Part-of-speech disambiguation", in *Actes, Coling - SIGDAT Workshop*, Copenhagen, Danemark.