Guide pour l'annotation des noms déverbaux

Équipe Nomage
STL, Université de Lille 3

28 mai 2009

Table des matières

1 Cadre général

1.1 Finalité du projet

Le projet Nomage vise la description et la modélisation des nominalisations en français, plus précisément des noms dérivés de verbes (ex. *construction*, dérivé de *construire*). Le projet s'intéresse notamment aux propriétés aspectuelles de ces noms, c'est-à-dire à l'aspect plus ou moins dynamique des situations qu'ils dénotent (ex. *manifestation vs. croyance*), à leur aspect plus ou moins borné (ex. *construction vs. jardinage*) ou encore à leur aspect plus ou moins duratif (ex. *attente vs. découverte*). La principale question théorique qui sous-tend ce projet est celle de savoir dans quelle mesure les noms héritent des propriétés aspectuelles des verbes dont ils sont dérivés.
La description et la modélisation des noms déverbaux va reposer, dans le cadre du projet Nomage, sur l'annotation de leurs occurrences en corpus. Cette étape préliminaire d'annotation des nominalisations fait l'objet de ce guide.

1.2 Étude des nominalisations en corpus

Le corpus utilisé pour l'annotation des nominalisations est le French TreeBank, corpus du français développé il y a une dizaine d'années au laboratoire LLF (CNRS & Université Paris 7) sous la direction d'Anne Abeillé. Le French TreeBank regroupe des articles du journal *Le Monde* parus entre 1989 et 1995 (environ 1 million de mots) et propose trois principaux niveaux d'annotation linguistiques :

- annotation morpho-syntaxique (adjectif, nom commun, préposition, etc.)
- annotation en constituants (groupe nominal, groupe prépositionnel, etc.)
- annotation en fonctions (sujet, objet, etc.)

Pour l'heure, environ 11 800 nominalisations ont été extraites de la sous-partie du corpus annotée en fonctions (soit à peu près la moitié du French TreeBank). Un certain nombre de ces nominalisations ne seront toutefois pas analysées dans le cadre du projet Nomage, pour l'une des trois raisons suivantes :

- la nominalisation est un nom dérivé d'un adjectif (par ex. *indulgence*, dérivé de *indulgent*) or nous n'étudions ici que les noms dérivés de verbes ;
- le nom est lie à un verbe mais le sens de la dérivation n'apparaît pas clairement. Par exemple, on ne sait pas si *voyager* est dérivé de *voyage* ou si c'est l'inverse ;
- il ne s'agit pas d'une nominalisation. Les nominalisations sont extraites automatiquement en fonction d'un suffixe (par exemple *-tion*, *-age*, etc.) et d'une longueur minimale de caractères précédant ce suffixe. Cette heuristique nous conduit imman-

quablement à la sélection de candidats qui ne sont pas des nominalisations (par ex. *sarcophage*).

Une "stop-list" a été dressée pour exclure automatiquement la plupart des cas mentionnés ci-dessus. Les annotateurs sont toutefois invités à vérifier que les noms qu'ils ont à traiter sont bien des nominalisations dérivées de verbes.

2 Procédure d'annotation

2.1 Annotation des propriétés observées

La première étape de l'annotation consiste à indiquer quelles sont les propriétés morphologiques et syntaxiques du nom candidat. Ces propriétés étant pour la plupart observables, nous les regroupons sous le terme de "propriétés observées".

2.1.1 Présentation du candidat

Les propriétés suivantes sont renseignées automatiquement et ne sont pas modifiables :

- l'attribut **wordForm** indique le nom candidat tel qu'il apparaît dans le corpus (ex. wordForm="gouvernements") ;
- l'attribut **isLemma** indique le lemme du nom candidat (ex. pour wordForm="gouvernements", isLemma ="gouvernement") ;
- l'attribut **morphoCue** donne l'indice morphologique qui a conduit à la sélection du nom candidat (ex. pour wordForm="intégration", morphoCue="tion").
- l'attribut **hasMorpho** indique certaines propriétés morphologiques (nombre et genre) associées au candidat (ex. pour wordForm="intégration", hasMorpho="fs").

2.1.2 Origine morphologique du candidat

La propriété suivante concerne l'origine morphologique du nom candidat. Si le nom possède **en synchronie** un verbe morphologiquement apparenté, les annotateurs saisissent ce verbe comme valeur du trait **isDerivedFromVerb**. Ils doivent saisir la valeur "null" dans le cas contraire, comme illustré ci-dessous :

- (bombardement) **isDerivedFromVerb**="bombarder"
- (allocation) **isDerivedFromVerb**="allouer"
- (indulgence) **isDerivedFromVerb**="null"
- (sarcophage) **isDerivedFromVerb**="null"

La réponse à la question de savoir si le candidat est dérivé d'un verbe doit être immédiate : si aucun verbe ne vient spontanément à l'esprit, la valeur "null" doit être saisie. L'annotation de la fiche du candidat s'arrête alors ici. En pratique, un verbe de la liste *Verbaction*[1] sera en général proposé, les annoteurs n'ayant qu'à vérifier que c'est bien le nom qui est dérivé du verbe et non l'inverse.

2.1.3 Position syntaxique du candidat

La seconde propriété observée concerne la position syntaxique du candidat à l'intérieur du groupe nominal (GN) dans lequel il figure. Le nom peut en effet être la tête du GN, comme illustré en (1.a), ou dépendant de la tête, comme illustré en (1.b). Disons, pour faire très simple, que la tête est le nom qui figure le plus à gauche dans le GN (indiqué dans un des champs d'une fiche d'annotation, voir section 2.3). Dans le premier cas, l'attribut **isSyntHead** recevra la valeur "yes", dans l'autre la valeur "no".

(1) a. *Neues Deutschland, l'organe du parti, a ainsi publié cette semaine une contribution réclamant [la constitution d'un véritable parti social-démocrate]*$_{\boldsymbol{GN}}$.

b. *Grand Metropolitan poursuit [sa politique d'acquisition de marques]*$_{\boldsymbol{GN}}$ *dans le domaine des vins et spiritueux.*

Les candidats qui ne sont pas en position de tête dans le GN ne doivent pas être traités par les annotateurs car leur position syntaxique rend difficile l'utilisation des tests[2].

2.2 Annotation des propriétés inférées

La seconde partie de l'annotation consiste à appliquer une série de tests qui nous indiqueront certaines propriétés sémantiques du candidat, notamment ses propriétés aspectuelles. Le sens n'étant pas observable, nous parlerons ici de propriétés inférées (à partir du résultat des tests).
Chacun des tests présentés ci-dessous est illustré de deux exemples, un premier exemple où le test marche (l'annotateur doit alors indiquer **yes**), un second exemple où le test ne marche pas (**no**). Dans chaque cas, le nom candidat est souligné (avec son éventuel déterminant) tandis que la modification correspondant à l'application du test est indiquée en gras.

[1] Verbaction est un lexique de noms d'actions morphologiquement apparentés à des verbes. Il a été développé par le CLLE-ERSS de l'Université de Toulouse Le Mirail.
[2] Ces noms seront toutefois étudiés ultérieurement.

T1_Plusieurs Ce test consiste à remplacer le déterminant du nom par "plusieurs" ou à ajouter cette séquence, si le nom n'a pas de déterminant.

- *Une allocation sera versée en décembre.* → ***Plusieurs** allocations seront versées en décembre.* (**yes**)
- *Tout au plus des petites choses à changer sur l'intégration.* → **Tout au plus des petites choses à changer sur **plusieurs** intégrations.* (**no**)

*****

T2_Avoir lieu Dans ce test, le nom est modifié au moyen d'une relative de la forme "qui AVOIR lieu + complément de temps".

- *La décentralisation des universités apparaît comme l'un des grands débats de l'année.* → *La décentralisation des universités **qui a lieu en ce moment** apparaît comme l'un des grands débats de l'année.* (**yes**)
- *Entre gens qui ont des vraies convictions, il peut y avoir convergences.* → **Entre gens qui ont des vraies convictions, il peut y avoir convergences **qui ont lieu en ce moment**.* (**no**)

Remarque sur le verbe de la relative Le verbe de la relative (indiqué en lettres capitales) doit être conjugué à un temps (passé, présent, futur, etc.) et à un mode (conditionnel, indicatif) qui conviennent dans le contexte de la phrase.

Remarque sur le complément de temps Ce complément peut être de l'une des formes suivantes :

- il peut s'agir d'un nom (*lundi*), d'un groupe nominal (*le 23 février, ce mois-ci*) ou d'un adverbe (*hier, demain*).
- il peut s'agir d'un groupe prépositionnel (*à 8h, en février dernier, au printemps, en 1987...*). On peut également, **au besoin**, recourir à un complément de durée (*pendant deux jours*) ou à un complément de fréquence (*chaque samedi*).

Remarque sur le type de la relative Les propositions relatives peuvent être de deux types selon la façon dont elles modifient le nom auxquelles elles s'appliquent : **spécificatives**, comme illustré en (2.a) ou **explicatives**, comme illustrées en (2.b). Cette différence se marque dans la phrase par l'utilisation ou non de virgules encadrant la relative.

(2) a. *Les randonneurs **qui sont fatigués** pourront faire une pause.*

b. *Les randonneurs**, qui sont fatigués,** pourront faire une pause.*

Les annotateurs doivent, chaque fois que cela est possible, utiliser une relative spécificative. Si le contexte rend cela impossible ou bizarre, il est possible d'essayer

d'appliquer le test avec une relative explicative.

Remarque sur la position de la relative La relative se place, lorsque cela est possible, le plus à droite possible du nom candidat. Toutefois, il arrive fréquemment que la relative ne puisse être placée qu'après le(s) complément(s) prépositionnel(s) du nom, comme c'est le cas ci-dessus avec *la décentralisation des universités.*

Notons que si le nom candidat est déjà modifié par une relative, il convient de coordonner cette relative avec la relative du test au moyen d'un *ou*, d'un *et* ou d'un *mais.*

- *C'est une manifestation sans ambition spatiale excessive qui s'est installée dans les salles restantes. → C'est une manifestation sans ambition spatiale excessive* ***qui a lieu le 4 mars*** *et qui s'est installée dans les salles restantes* (**yes**)

*****

T3_Éprouver/ressentir Dans ce test, le nom est modifié au moyen d'une relative de la forme "que x EPROUVER/RESSENTIR (+ complément de temps)".

- *L'admiration, sinon la confiance, se sont émoussées. → L'admiration, sinon la confiance,* ***qu'on éprouvait depuis longtemps*** *se sont émoussées.* (**yes**)
- *Cette revendication vient d'être rappelée par l'association nationale des élus locaux. → *Cette revendication,* ***qu'on éprouve depuis longtemps,*** *vient d'être rappelée par l'association nationale des élus locaux.* (**no**)

Remarque sur le verbe de la relative Cf. T2

Remarque sur le complément de temps Cf. T2

Remarque sur le type de la relative Cf. T2

Remarque sur la position de la relative Cf. T2

Remarque sur l'application du test Le test s'applique si l'une et/ou l'autre des relatives (celle avec *éprouver* ou celle avec *ressentir*) s'applique. Il ne s'applique pas si aucune des deux relatives n'est possible.

*****

T4_Un peu de Ce test consiste à remplacer le déterminant du nom par la séquence "un peu de" ou à ajouter cette séquence, si le nom n'a pas de déterminant.

- *Mais, en dénonçant avec un certain agacement la "french mafia", nos confrères étrangers ne rendent-ils pas indirectement hommage à une nouvelle forme d'influence française ? → Mais, en dénonçant avec* ***un peu d'****agacement la "french mafia", nos confrères étrangers ne rendent-ils pas indirectement hommage à une nouvelle forme d'influence française ?* (**yes**)
- *Une réunion du courant Socialisme et République samedi soir tranchait la question :*

l' hostilité à "l'inconstance politique" de Mr Dray l' emportait. → **Une réunion du courant Socialisme et République samedi soir tranchait* ***un peu de*** *question : l' hostilité à "l'inconstance politique" de Mr Dray l' emportait.* (**no**)

Remarque sur les noms au pluriel Les noms au pluriel doivent pouvoir être mis au singulier :

- *Un peu de convergences* : *un peu de convergence* (**yes**)
- *Un peu de logements* : **un peu de logement* (**no**)

Remarque sur un éventuel changement de sens Il se peut la transformation change le sens du candidat, comme illustré dans la phrase ci-dessous. On considérera dans ces cas-là que le test s'applique, sachant que l'analyse post-annotation reviendra sur cette différence sémantique.

- *L'élue de l'Essonne ne cachait pas que l'expérience la tentait.* → *L'élue de l'Esonne ne cachait pas qu'****un peu d'****expérience la tentait.* (**yes**)

*****

T5_Durer x temps Dans ce test, le nom est modifié au moyen d'une relative de la forme "qui DURER x temps".

- *La décentralisation des universités apparaît comme l'un des grands débats de l'année.* → *La décentralisation des universités****, qui durera deux ans,*** *apparaît comme l'un des grands débats de l'année.*
- *Monsieur Jospin a insisté sur la part que doivent prendre les collectivités locales dans les décisions et les investissements.* → **Monsieur Jospin a insisté sur la part que doivent prendre les collectivités locales dans les décisions* ***qui dureront trois jours*** *et les investissements.* (**no**)

Remarque sur le complément de durée Dans le complément de durée, *x* doit correspondre à un chiffre, *temps* à minute, heure, jour, semaine, mois ou année.

Remarque sur le verbe de la relative Cf. T2

Remarque sur le type de la relative Cf. T2

Remarque sur la position de la relative Cf. T2

*****

T6_Se trouver Dans ce test, le nom est modifié au moyen d'une relative de la forme "qui SE TROUVER (+ complément de lieu)".

- *Le gouvernement a décidé d'accorder la maîtrise d'ouvrage aux collectivités locales, pour les constructions universitaires.* → *Le gouvernement a décidé d'accorder la maîtrise d'ouvrage aux collectivités locales, pour les constructions universitaires*

qui se trouveront en banlieue. (**yes**)

- *Mais de leur côté, les collectivités demandent que cette participation s'accompagne d'une extension de leurs compétences à l'enseignement supérieur.* → **Mais de leur côté, les collectivités demandent que cette participation**, **qui se trouvera en banlieue,** *s'accompagne d'une extension de leurs compétences à l'enseignement supérieur.* (**no**)

Remarque sur le type de la relative Cf. T2
Remarque sur la position de la relative Cf. T2

*****

T7_Effectuer/procéder Dans ce test, le nom est modifié au moyen d'une relative de la forme "que x ÉFFECTUER (+ complément de temps)" et/ou d'une relative de la forme "auquel on PROCÉDER (+ complément de temps)".

- *Ça c'est colossal, parce qu'enfin, jusqu'à l'annexion, les pays baltes et scandinaves, question niveau de vie, c'était du pareil au même* → *Ça c'est colossal, parce qu'enfin, jusqu'à l'annexion* ***à laquelle on a procédé en 2000****, les pays baltes et scandinaves, question niveau de vie, c'était du pareil au même* (**yes**)
- *Entre gens qui ont des vraies convictions, il peut y avoir convergences.* → **Entre gens qui ont des vraies convictions, il peut y avoir convergences**, ***auxquelles on procédera bientôt.*** (**no**)

Remarque sur le verbe de la relative Cf. T2
Remarque sur le complément de temps Cf. T2
Remarque sur le type de la relative Cf. T2
Remarque sur la position de la relative Cf. T2
Remarque sur l'application du test Cf. T3

*****

T8_État de Ce test consiste à placer la séquence "état de" juste à la gauche du candidat, c'est-à-dire entre le candidat et son déterminant.

- *Sa vie et la diversité de son talent en auront fait une sorte de voyageur "professionnel" dans une Europe en pleine effervescence.* → *Sa vie et la diversité de son talent en auront fait une sorte de voyageur "professionnel" dans une Europe en plein* ***état d'****effervescence.* (**yes**)
- *Il y avait un moyen simple de prouver cette intention.* → **Il y avait un moyen simple de prouver cet* ***état d'****intention.* (**no**)

*****

T9_Se dérouler Dans ce test, le nom est modifié au moyen d'une relative de la forme

"qui SE DÉROULER (+ complément de temps)".

- *Chaque année ou presque, La Royal Academy of Art consacre une de ses expositions majeures à l'architecture.* → *Chaque année ou presque, La Royal Academy of Art consacre une de ses expositions majeures**, qui se déroule en général au printemps,** à l'architecture.* (**yes**)
- *Cette revendication vient d'être rappelée par l'association nationale des élus locaux.* → **Cette revendication**, qui se déroule aujourd'hui,** *vient d'être rappelée par l'association nationale des élus locaux.* (**no**)

Remarque sur le verbe de la relative Cf. T2
Remarque sur le complément de temps Cf. T2
Remarque sur le type de la relative Cf. T2
Remarque sur la position de la relative Cf. T2

*****

T10_Cardinal Le test consiste à remplacer le déterminant (qu'il soit défini ou indéfini) par un cardinal (par exemple, *trois*, *trente*, *deux cents*, etc.). Il est également possible de placer le cardinal entre le candidat et son déterminant, lorsque celui-ci est défini ou encore d'ajouter le cardinal en position de déterminant lorsque le nom n'a pas déjà un déterminant.

- *Quelles sont les possibilités et les intentions des différents acteurs ?* → *Quelles sont les possibilités et les intentions des **trois** différents acteurs ?* (**yes**)
- *Mais, de leur côté, les collectivités demandent que cette participation s'accompagne d'une extension de leurs compétences à l'enseignement supérieur.* → **Mais, de leur côté, les collectivités demandent que cette participation s'accompagne de **trois** extensions de leurs compétences à l'enseignement supérieur.* (**no**)

2.3 Exemples d'annotation de candidats

Nous présentons ci-dessous deux exemples de fiches annotées (figures 1 et 2). Chaque fiche débute par une phrase extraite du French Treebank dans laquelle figure le candidat à annoter (indiqué en gras) et par l'indication du GN dans lequel figure ce candidat. Puis viennent la partie consacrée aux propriétés observées (les propriétés déjà renseignées sont ici soulignées) et la partie consacrée aux propriétés inférées. Notons que les réponses apportées ici ne reflètent l'intuition que d'une personne et peuvent donc diverger de l'intuition d'une autre personne.

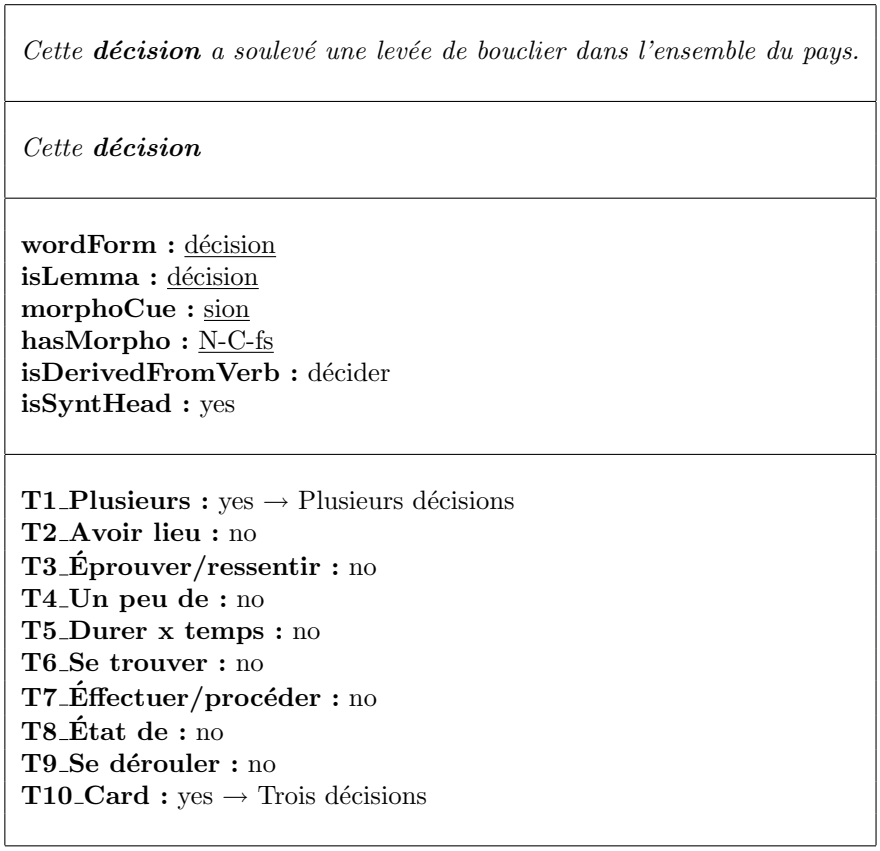

FIG. 1 – Annotation d'une occurrence du nom *décision*

*Ça, c'est colossal, parce qu'enfin, jusqu'à l'**annexion**, les pays baltes et scandinaves,question niveau de vie, c'était du pareil au même.*
*l'**annexion***
wordForm : annexion **isLemma :** annexion **morphoCue :** xion **hasMorpho :** N-C-fs **isDerivedFromVerb :** annexer **isSyntHead :** yes
T1_Plusieurs : no **T2_Avoir lieu :** yes → l'annexion qui a eu lieu l'année dernière **T3_Éprouver/ressentir :** no **T4_Un peu de :** no **T5_Durer x temps :** yes → l'annexion, qui a duré 3 ans **T6_Se trouver :** no **T7_Éffectuer/procéder :** yes → l'annexion à laquelle ont procédé les russes **T8_État de :** no **T9_Se dérouler :** no **T10_Card :** no

FIG. 2 – Annotation d'une occurrence du nom *annexion*

3 Environnement pour l'annotation

3.1 Base de donnée Open Office

L'annotation des nominalisations extraites du French Treebank se fait à l'aide d'un formulaire de base de données Open Office. Téléchargez la dernière version du logiciel (3.1.0) ou faites une mise à jour si vous disposez d'une version antérieure d'Open-Office.
Pour accéder au formulaire, téléchargez la base de donnée disponible sur le site Nomage. Une fois la base ouverte, cliquez sur le bouton "Formulaires" (à gauche), puis double-cliquer sur "NOMAGE_TOUT" (au centre). L'ouverture et la fermeture de "NOMAGE_TOUT" peuvent être longues, la base contenant presque 12 000 enregistrements.

3.2 Procédure de sauvegarde des données

Il est vivement recommandé d'effectuer une sauvegarde de la base de travail à chaque fin de session d'annotation, voire deux fois par session. La procédure de sauvegarde est la suivante :

1. Créer un dossier à votre nom sur le serveur nomage
2. A chaque session d'annotation, créer un sous-dossier ayant pour nom la date du jour et y faire une copie de la base de travail. L'important est de toujours garder les deux ou trois versions précédentes de votre base de travail, les versions plus anciennes pouvant être au fur et à mesure effacées pour ne pas encombrer l'espace.

On aura donc, sur le serveur Nomage, la hiérarchie de dossiers suivante :

→ `NomAnno` contenant
  → `06_Juin` contenant le fichier `base.odb`
  → `07_Juin` contenant le fichier `base.odb`
  → etc.

Cette procédure de sauvegarde atypique est due au fait que la base de travail ne peut être enregistrée sous un autre nom que son nom d'origine sans perdre l'ensemble des données. On se sert donc de noms de dossier pour identifier les différentes versions de la base annotée.