Repérage automatique de génériques dans les définitions terminographiques

Selja SEPPÄLÄ
Laboratoire de terminologie/TIM/ETI – Université de Genève
Bd du Pont-d'Arve 40, 1211 Genève 4
selja.seppala@eti.unige.ch

Résumé. Cet article présente une procédure de repérage et de balisage de l'élément générique de la définition terminographique exploitant les caractéristiques formelles du sous-langage définitoire. La procédure, qui comporte quatre étapes, constitue l'une des sous-tâches d'un analyseur (semi-)automatique de la structure conceptuelle des définitions terminographiques, destiné à faciliter l'annotation d'un corpus en vue de l'étude de régularités dans cette structure. La tâche décrite consiste à mettre au point un système d'annotation automatique basé sur le repérage d'indices morphosyntaxiques, sans recourir à d'autres ressources linguistiques informatisées.

Abstract. This article presents a procedure to locate and tag the generic elements of terminographic definitions, taking advantage of the formal characteristics of the definition sublanguage. This four step procedure is part of a larger (semi-)automatic parser of the conceptual structure of terminographic definitions, intended to ease the tagging of a corpus for studying conceptual regularities in definition structure. The method involves the development of an automatic tagging system, based on the identification of morphosyntactic boundary markers, which does not require the use of additional linguistic resources.

Mots-clés : définition terminographique, annotation automatique, repérage de frontière, indices morphosyntaxiques, sous-langage.

Keywords : terminographic definition, automatic tagging, boundary location, morphosyntactic markers, sublanguage.

1 Introduction

La présente communication s'inscrit dans le cadre d'une recherche en terminologie portant sur l'analyse de la structure conceptuelle de définitions terminographiques à travers l'étude d'un corpus de définitions annoté (Seppälä, 2004, 2005). L'un des objets de notre travail est la conception d'un analyseur (semi-)automatique de cette structure conceptuelle, qui facilite l'annotation du corpus en automatisant le plus grand nombre de tâches. Nous présentons ici l'une des étapes de l'annotation du corpus, à savoir la mise en place d'une procédure de

repérage et de balisage automatique des génériques (GEN) des définitions en compréhension, le générique étant l'élément qui rattache le concept défini à un concept plus général (souvent le *genre prochain*). La tâche décrite dans cet article est basée sur un traitement qui exploite les caractéristiques du sous-langage des définitions (Barnbrook, 2002). L'implémentation, en vue d'un balisage automatique des GEN des définitions, est faite à l'aide d'un programme Perl où les règles de repérage sont exprimées sous la forme d'expressions régulières.

Après une brève présentation de la définition terminographique, de sa structure générale et de l'élément générique (section 2), nous passons à l'identification du problème, afin de mieux cerner la tâche à accomplir (section 3). Nous abordons ensuite la description de l'expérience à proprement parler. Dans cette partie plus empirique, nous présentons tout d'abord les corpus utilisés (section 4), puis la méthode qui a été mise en place (section 5), pour terminer par une évaluation des performances de la procédure (section 6). En conclusion, nous évoquons les limites de cette méthode et quelques perspectives futures.

2 La définition terminographique

En terminographie, les définitions servent à décrire et distinguer les concepts d'un domaine (ou système conceptuel) spécialisé, plus ou moins clos, où les concepts entretiennent des relations généralement hiérarchiques. Elles consistent, dans la très grande majorité des cas, en une définition en compréhension[1]. Une définition en compréhension est composée, comme l'illustre l'exemple ci-dessous, d'une seule phrase où le GEN est suivi d'un ou de plusieurs spécifiques (SPE). Le SPE a pour fonction de préciser la portée du GEN en énumérant les traits ou propriétés du concept défini, tout en distinguant ce dernier des autres concepts du domaine auquel il appartient.

Ex : *peptidyl-transférase* = $_{GEN}$Enzyme $_{SPE1}$située dans le ribosome, $_{SPE2}$qui lie successivement un acide aminé supplémentaire à la chaîne polypeptidique en cours d'élongation.

La forme de ce type de définitions est relativement contrainte : elle se compose d'une seule phrase, où le GEN précède généralement le(s) SPE. Parfois, le GEN peut être précédé d'un adjectif – considéré ici comme SPE –, mais cela reste assez rare en français. À l'exception de ces cas de figure facilement identifiables, une définition respectant les conventions de rédaction terminographiques généralement admises (rappelées dans Seppälä, 2004, 2006) ne devrait pas être précédée d'un SPE en français.

Selon notre schéma d'analyse[2], chaque définition est segmentée en GEN et en SPE ; ces éléments sont ensuite annotés avec des étiquettes conceptuelles. La segmentation et l'annotation sont réalisées en XML, le but étant d'automatiser ces opérations au maximum. L'exemple suivant montre une version annotée de la fiche terminologique du concept *peptidyl-transférase*.

---

[1] 97 % d'un corpus étudié lors d'une précédente expérience (Seppälä, 2004, 2005), le reste étant des définitions en extension. Étant donné notre tâche, nous excluons de cette étude ces définitions, qui se composent d'une énumération d'espèces isonymes et n'ont donc pas de générique.

[2] Plusieurs niveaux d'analyse sont possibles selon la granularité voulue (Barnbrook, 2002). Pour plus de détails sur celui que nous avons adopté, voir (Seppälä, 2004).

Ex :
```
<FICHE langue="FR">
<NI>25</NI>
<CM>biosynthèse des protéines</CM>
<VE>peptidyl-transférase</VE>
<DF><GEN relation_VE="GENRE"
classe_conceptuelle="NATUREL">Enzyme </GEN>
<SPE relation_GEN="SPATIAL">située dans le ribosome,</SPE>
<SPE relation_GEN="FONCTION">qui lie successivement un acide
aminé supplémentaire à la chaîne polypeptidique en cours
d'élongation.</SPE></DF>
</FICHE>
```
[3]

L'élément générique de la définition, également appelé *incluant*, a pour fonction de rattacher le concept défini à un concept plus général. Sa forme présente un certain nombre de variations, néanmoins limitées et régulières. Le plus souvent il est constitué d'un seul mot[4]. Lorsque le GEN n'est pas un mot unique, il peut prendre la forme d'un terme complexe formé de deux mots ou plus (*Acide nucléique* ou *Mesures ou ouvrages*), d'un syntagme nominal (*Triplet de nucléotides*) ou encore d'un *faux incluant*[5] commençant par un marqueur relationnel (*Partie de…* ou *Ensemble de…*), dont la liste est relativement restreinte et qui permet d'annoter le GEN avec la relation conceptuelle (PARTIE, TOUT, GENRE ou GENRE PROCHAIN) qui l'unit au défini. Dans ce cas, c'est l'ensemble « marqueur + mot ou syntagme » qui constitue le GEN. Dans les deux cas, le GEN peut comporter une autre entrée du domaine (environ 50 % du corpus d'entraînement). En effet, les concepts définis en terminographie s'insèrent généralement dans un système conceptuel hiérarchisé, où le concept superordonné peut servir de générique à la définition du concept subordonné. Ceci se traduit, linguistiquement parlant, par la reprise dans le GEN, de l'un des termes ou synonymes (l'une des vedettes) désignant ce concept, comme le montre l'exemple suivant, où la vedette (*acide nucléique*) du concept superordonné devient le GEN de la définition du concept subordonné (*acide ribonucléique*).

Ex : ***acide nucléique*** = *Molécule constituée d'un enchaînement de nucléotides disposé le long d'un brin ou deux.*
acide ribonucléique = ***Acide nucléique*** *formé d'un seul brin et participant à toutes les étapes de la synthèse des protéines*.

Dans le cas des définitions d'adjectifs, peu fréquentes, c'est l'expression du type *Se dit de…* qui est considérée comme étant le GEN de la définition. Ainsi le GEN présente des régularités marquées qui contribuent à faire du discours définitoire un sous-langage spécifique (Barnbrook, 2002).

---

[3] NI = numéro d'identification unique ; CM = domaine ; VE = vedette (terme ou synonyme) ; DF = définition ; GEN = générique ; SPE = spécifique ; relation_VE = relation conceptuelle entre le GEN et la VE ; relation_GEN = relation conceptuelle entre le SPE et le GEN.

[4] Nous entendons par mot, une chaîne de caractères suivie d'un espace.

[5] On trouve le *faux incluant* dans cinq types de situation : lorsque la chose est définie par ses parties ; lorsqu'il y a définition de la chose transformée ; lorsque la chose est définie par sa cause ou sa conséquence ; lorsque l'incluant marque le rapport de la chose à l'unité ; et lorsqu'il y a faux incluant d'existence (Rey-Debove, 1971).

3 Identification de la tâche

Notre tâche est de repérer le générique de la définition et de le marquer, de part et d'autre, avec les balises XML suivantes : <GEN relation_VE="…">xxxxx</GEN>. Le contenu du GEN n'étant pas en cause, la question revient par conséquent à déterminer la frontière qui sépare le GEN du ou des SPE adjacent(s), selon qu'il est ou non précédé de cet élément. Dans la grande majorité des cas, le GEN se trouve en effet en début de définition terminographique ; il correspond même souvent au premier mot de la phrase (env. 70 % des GEN de notre corpus d'entraînement ; 97 % du corpus d'évaluation). On sait, par ailleurs, qu'un GEN est toujours suivi d'un SPE (en terminologie). Suivant le même constat que (Barnbrook, 2002), nous pouvons dire que la principale difficulté consiste donc à repérer sa limite à droite, plus exactement la frontière qui le sépare du premier SPE. Nous proposons d'exploiter les régularités du discours définitoire en identifiant ces frontières (début et fin de GEN) à l'aide de marqueurs morphosyntaxiques, lesquels peuvent ensuite être traduits en règles de repérage sous la forme d'expressions régulières.

La constitution d'une liste de marqueurs morphosyntaxiques nécessite une étude détaillée des principales caractéristiques des GEN et des SPE antéposés, ainsi que du début (limite à gauche) des SPE qui viennent immédiatement après le GEN. Pour ce faire, nous avons eu recours, d'une part, à la littérature (Iris, et al., 1988, Iso, 2000, L'homme, 2003, Rebeyrolle, 2000, etc.) et, d'autre part, à un corpus préalablement annoté à la main (voir section 4), afin d'y repérer les régularités susceptibles de servir d'indices de repérage. L'identification des marqueurs peut être réalisée manuellement et/ou automatiquement, par des méthodes d'apprentissage automatique. Dans la mesure où le discours définitoire constitue un sous-langage relativement contraint, présentant peu de variations, notamment au niveau des éléments morphosyntaxiques marquant la frontière entre les éléments de la définition (Barnbrook, 2002), nous avons opté pour un repérage manuel assisté d'un concordancier. L'identification d'indices peut également faire intervenir des traitements préalables (Vossen, et al., 1989), comme un étiquetage morphosyntaxique ou une lemmatisation ; la grande régularité du discours définitoire nous a là encore permis de nous en tenir à une méthode simple, basée sur les seuls éléments de surface, c'est-à-dire sur des chaînes de caractères.

Une fois les marqueurs identifiés et traduits en règles de repérage, il convient de mettre au point une procédure d'étiquetage qui n'applique qu'une seule règle par définition, afin d'éviter les erreurs d'annotation. Deux paramètres sont pris en compte :

- la hiérarchisation des étapes et des règles d'analyse de la plus spécifique à la plus générale, de façon à ce qu'elles entrent le moins possible en conflit entre elles, et
- la recherche systématique des vedettes (VE) de la base terminologique en début de définition, afin de pouvoir les inclure le cas échéant dans les GEN (étape appelée « GEN=VE »).

Selon ces constats, le repérage des SPE antéposés doit avoir préséance sur l'application des règles de repérage des GEN comprenant une VE, laquelle se doit de précéder la recherche de la fin du GEN lorsqu'on ne connaît pas son contenu, c'est-à-dire tous les autres cas. Pour terminer, il y a lieu d'évaluer la performance de ce système et d'envisager des pistes d'amélioration.

4 Présentation des corpus

Les expériences réalisées dans le cadre de cette étude portent sur deux corpus distincts : un corpus d'entraînement et un corpus d'évaluation. Le premier a servi à identifier les indices de repérage des GEN et les marqueurs pour chaque relation conceptuelle, ainsi qu'à tester, affiner et hiérarchiser les règles de repérage, de façon à obtenir la meilleure performance possible. Il s'agit d'un échantillon de 490 définitions en compréhension tirées de la *Banque de terminologie du canton de Berne* (Lingua-PC), où chaque définition a été préalablement segmentée et annotée à la main en GEN et en SPE[6]. Le second corpus a été utilisé pour tester la performance générale de la procédure d'annotation proposée. Il s'agit d'un ensemble de 92 définitions extraites d'un glossaire sur la *Terminologie de la biosynthèse des protéines chez les cellules eucaryotes* (Bourjault, 2005). Les deux corpus respectent les conventions de rédaction terminographiques et remplissant les critères de bonne formation.

5 Description du système

Dans cette partie, nous présentons l'architecture générale du système et une synthèse de ses principales étapes, en nous concentrant plus spécifiquement sur les difficultés qu'elles posent. Comme nous l'avons déjà souligné, les différentes étapes du traitement doivent être ordonnées pour éviter les conflits de règles. Nous distinguons quatre types de tâches à réaliser dans l'ordre suivant : repérage des SPE antéposés et marquage des balises de début de GEN (section 5.1) ; repérage d'éventuelles vedettes dans les génériques (GEN=VE) (section 5.2) ; repérage des fin de GEN (section 5.3) ; et finalement, marquage du premier mot des définitions non traitées au cours des étapes précédentes (section 5.4).

5.1 Repérage des spécifiques antéposés et marquage du début du GEN

Le repérage d'éventuels spécifiques antéposés et l'insertion des balises de début de GEN doivent précéder les autres étapes du traitement, au risque d'engendrer des problèmes de repérage des GEN. Une vedette utilisée en tant que générique pourrait, par exemple, ne pas être repérée, simplement parce qu'elle est précédée d'un SPE. Cette tâche consiste à assigner aux SPE antéposés les balises `<SPE>xxxxx</SPE><GEN>`. La dernière balise, qui est également ajoutée au début de toutes les autres définitions, marque le début du GEN et est nécessaire pour que les règles suivantes puissent être uniformément appliquées à l'ensemble des définitions. Les SPE antéposés peuvent être de deux types – adjectifs et indication de domaine –, dont les caractéristiques formelles susceptibles de servir d'indice à leur repérage sont aisément identifiables et peu nombreuses.

1. Adjectifs : La nature non indexicale, objective et factuelle des définitions terminographiques implique qu'en français, la liste des adjectifs antéposables pouvant servir de marqueur de SPE antéposé est relativement réduite. Le corpus d'entraînement, par exemple, ne fait état que d'adjectifs numéraux ordinaux (*première*, *deuxième*, etc.) et de quelques adjectifs qualificatifs (*grand*, *petit*, etc.) ou de leurs superlatifs. Formellement parlant, ces adjectifs donnent lieu à deux types de SPE antéposés :

[6] Pour une description plus détaillée du corpus et de l'annotation conceptuelle, voir (Seppälä, 2004, 2005).

- Le premier consiste en un adjectif qui peut être précédé d'un superlatif absolu et qui est immédiatement suivi du GEN.
  Ex : <SPE> (superlatif absolu)[7] + adj. </SPE> + <GEN>
  ⇒ *très grand [vaisseau]* ou *première [phase]*

- Le second consiste en un adjectif, superlatif relatif ou numéral ordinal, précédé d'un article défini et éventuellement d'un comparatif, et suivi d'un article indéfini et éventuellement d'un nombre.
  Ex : <SPE> art. défini + (comparatif) + adj. + art. indéfini + (nombre) </SPE> + <GEN>
  ⇒ *le plus grand des [singes]* ou *la première des trois [étapes]*

2. Indication de domaine : Dans les bases de données terminologiques, l'indication du domaine et des sous-domaines se fait dans un champ d'indexation propre. Or, ce type d'indication apparaît parfois en début de définition, souvent pour restreindre la portée du concept défini à un sous-domaine plus spécifique. Une définition bien formée ne devrait pas inclure ce type d'élément, mais un analyseur automatique de définitions doit néanmoins prévoir ces cas de figure. La forme de cette indication est, elle aussi, très contrainte : elle commence par une préposition, suivie d'un ou de plusieurs mots, et se termine par une virgule.
   Ex : <SPE> En | Dans | Sur + mot (s) quelconque(s) + virgule </SPE> + <GEN>
   ⇒ *En droit constitutionnel,…* ou *Sur une locomotive,…*

5.2 Repérage des génériques incluant une vedette

La deuxième étape consiste à vérifier si le GEN reprend l'un des termes du domaine. Pour ce faire, toutes les vedettes de la base sont extraites dans une liste (18080 termes dans le corpus d'entraînement), qui doit être triée par ordre alphabétique, du terme le plus long au plus court, afin d'éviter que certains éléments des termes complexes ne soient exclus du GEN. L'exemple suivant montre la séquence à respecter pour les termes complexes dérivés d'une même base lexicale : *traitement annuel déterminant* → *traitement annuel* → *traitement*. Le programme vérifie ainsi pour chaque VE si elle apparaît en début de définition ou à la suite d'une expression relationnelle. Si c'est le cas, l'élément correspondant est marqué comme GEN et se voit attribuer la relation GENRE PROCHAIN ou celle qui correspond au marqueur de relation qui le précède ; sinon, la définition est remise dans le circuit pour être traitée par l'une des règles présentées dans la section suivante. Le développement de cette tâche présente deux types de difficultés.

1. Deux VE pour un même GEN : Il arrive que le programme repère une VE plus longue alors que la plus courte existe aussi dans la liste et que c'est celle-là qu'il aurait convenu de marquer : la liste contient par exemple les trois VE *crédit de paiement*, *crédit d'engagement* et *crédit*. Si la définition commence par *Crédit de paiement ou d'engagement…*, le programme repère tout d'abord *crédit de paiement* et le marque comme GEN. Or, il serait dans ce cas plus adéquat, soit de ne retenir que *crédit* et de considérer les deux éléments suivants comme un SPE précisant la nature de ce crédit, soit de marquer les deux comme composant un seul GEN.

[7] Les parenthèses marquent les éléments optionnels ; le « | » sépare différentes variantes ; et les crochets indiquent le GEN.

2. Traitement des pluriels : Les termes en vedette sont généralement sous la forme canonique, donc au singulier. La liste des VE prise en entrée du programme comporte donc principalement cette forme. Dans un GEN, ces termes peuvent en revanche apparaître au pluriel, ce qui n'a pas d'incidence lorsque, pour des questions d'usage (certains termes ne s'utilisent qu'au pluriel), ils figurent déjà au pluriel dans la base. C'est en revanche plus problématique lorsqu'il faut repérer la forme plurielle d'une VE au singulier, en particulier dans les cas de pluriels irréguliers, mais surtout dans ceux des mots composés. Notons toutefois que le taux d'erreurs dues à ce type de cas devrait rester très faible et qu'il peut être aisément réduit dans la mesure où, dans des définitions bien formées, seuls les termes qui suivent une expression relationnelle tendent à apparaître au pluriel, et que ces cas sont pris en charge par les règles plus générales de la troisième étape.
   Lors du développement, nous avons tout de même constaté que le fait de prévoir le repérage de VE suivies de la marque du pluriel la plus générale, à savoir un « s », permet d'améliorer la précision (VE = *dépense d'investissement* → GEN = *dépense d'investissements*). Seule réserve, mais plutôt relative au fond : il arrive que le GEN marqué et le pluriel de la VE reconnue dans la liste soient homographes, auquel cas ce n'est pas le véritable terme de la base qui est repéré, mais un autre terme au singulier (par exemple, VE : sg. *fond* → pl. *fonds* et GEN : *fonds*, mais ce dernier n'est pas le pluriel de *fond*, il s'agit d'un autre terme au singulier). Cependant, l'élément ainsi étiqueté reste un vrai GEN et le résultat est considéré comme acceptable, étant donné que la tâche est ici de trouver la frontière des GEN.

5.3 Repérage des fins de générique

La troisième étape consiste à soumettre toutes les définitions traitées lors de la première étape, mais ignorées lors de la deuxième, aux règles de repérage des marqueurs de frontière de fin de GEN. Nous avons vu que les définitions terminographiques présentent beaucoup de régularités, notamment au niveau des frontières autour desquelles s'articulent les GEN et les SPE. D'après (Barnbrook, 2002), un analyseur peut identifier ces limites en utilisant une combinaison de trois éléments : une règle générale identifiant les participes présents et passés réguliers ; une liste de moins de 100 mots comportant des membres des classes fermées (*selon*, *qui*, *dans*, etc.) et des participes passés irréguliers ; ainsi qu'une liste d'exclusion comprenant les mots susceptibles d'être mal traités par la règle de repérage générale. Bien que l'analyse que nous envisageons ne soit pas tout à fait la même, la présence des mêmes types de régularités a néanmoins été vérifiée sur notre corpus d'entraînement. Suivant ce constat, nous avons établi des règles correspondant à ces différents cas de figure, en prenant en compte toutes les réalisations possibles d'un marqueur. La marque du participe présent en *–ant* doit par exemple être déclinée en *–ante*, *–ants* ou *–antes*.

Les règles visant à contraindre le repérage de la fin du GEN exploitent donc les caractéristiques des contextes droit et gauche du GEN, et gauche du SPE qui le suit, et tiennent compte du nombre de mots apparaissant généralement entre les deux. Nous distinguons deux types de règles : des règles de fin spécifiques et des règles de fin générales.

1. Les règles spécifiques concernent les cas de faux incluants et exploitent à la fois les expressions relationnelles en début de GEN (*Type de…*, *Ensemble de…*, *Partie de…*, etc.), dont le nombre est restreint et qui permettent d'associer une relation spécifique au GEN (GENRE, TOUT ou PARTIE), et la ponctuation à la fin du GEN. (Dans les cas de GEN=VE, ce sont les VE qui constituent la limite à droite du GEN.)

2. Les règles générales concernent tous les autres cas et se basent sur trois types de marqueurs :

- sur la terminaison du premier mot après le GEN ($_{fin}MOT_{GEN+1}$), sachant qu'il s'agit généralement d'un participe présent ou passé ;
- sur le premier mot qui suit le GEN (MOT_{GEN+1}). Dans ce cas, les règles (lexicales) intègrent un certain nombre de mots entiers, tels que des conjonctions (*que*, *quand*, etc.), des pronoms relatifs (*qui*, *tel*, etc.) ou des prépositions (*par*, *dans*, etc.).
- Soit enfin sur la ponctuation qui suit le GEN, à savoir la virgule ou la parenthèse.

La principale difficulté liée au développement de ces règles est de les hiérarchiser en sorte qu'une règle plus générale ne marque pas des GEN qui devraient être traités par une règle plus spécifique. L'exemple suivant illustre un conflit de règles où l'étiquetage est faux si la règle lexicale plaçant le GEN avant *par*, précède la règle qui place la frontière avant un mot qui se termine par *–é*. Si on inverse l'ordre, le résultat est juste.

FAUX : *Gain brut réalisé</GEN> par…*
VRAI : *Gain brut</GEN> réalisé par…*

Une autre difficulté est liée à la préposition *de*. Ce marqueur apparaît aussi bien dans les mots composés à l'intérieur du GEN, qu'après un GEN (en début de SPE). La solution adoptée jusqu'à présent est de considérer que la préposition marque le début d'un SPE et que le GEN est donc composé d'un seul mot. Cette solution est en effet la plus probable étant donné que près de 70 % des GEN du corpus d'entraînement et 60 % du corpus d'évaluation sont des mots simples. Il semblerait aussi que les cas ambigus soient souvent mieux traités par les autres règles générales de type $_{fin}MOT_{GEN+1}$. La règle du marqueur *de* doit donc être classée parmi les dernières.

Le classement des règles est réalisé manuellement, en fonction de la fréquence d'application et de la performance de chaque règle prise individuellement, ainsi que des observations faites lors des tests et des ajustements successifs. L'ordre de classement général est le suivant :

1. Les règles spécifiques sont appliquées en premier (y compris dans l'étape GEN=VE), étant donné qu'elles sont contraintes par le début du GEN. Comme elles n'entrent pas en conflit entre elles, leur ordre interne n'a pas d'importance.
2. Suivent les règles lexicales permettant de repérer des mots MOT_{GEN+1} (*auquel*, *dont*, *lors*, *où*, etc.) qui précédent les participes.
3. Viennent ensuite les règles $_{fin}MOT_{GEN+1}$, qui précisent la fin du premier mot après le GEN. Celles-ci non plus n'entrent pas en conflit entre elles, dans la mesure où les formes des patrons recherchés sont bien distinctes. Il est donc possible des les placer sans ordre spécifique dans un même bloc précédant
4. une seconde règle lexicale qui comporte des prépositions (*dans*, *entre*, *par*, *pour*, *sur*) ne pouvant être placées avant les règles recherchant des participes.
5. Viennent finalement les règles les plus « bruyantes », qui tendent à provoquer beaucoup d'erreurs lorsqu'elles sont placées avant les autres (préposition *de* et *ponctuation*).

Dans tous les cas, il conviendrait de tester et d'ajuster l'ensemble des règles sur un corpus plus large, et de les classer ensuite statistiquement de façon à obtenir la meilleure performance possible.

5.4 Étiquetage des définitions non traitées

La quatrième et dernière étape consiste à soumettre au même traitement toutes les définitions dont les GEN n'auraient pas encore été balisés. Il s'agit de leur appliquer la règle la plus générale (règle par défaut) qui marque systématiquement le premier mot de la définition comme étant le GEN. L'application de cette seule règle à la troisième étape (une fois les GEN=VE exclus et en ignorant les autres règles) montre que 72,7 % des définitions du corpus d'entraînement auraient été étiquetées convenablement. L'ensemble des règles de repérage de fin de GEN (celles de la 3e étape) visent donc à couvrir les 27,3 % de cas d'échec de la règle par défaut (celle qui marque systématiquement le 1er mot). Ce constat réduit en fait considérablement le nombre d'exemples du corpus d'entraînement permettant d'apprécier toute la gamme des variations possibles de la forme du GEN, mais cela n'empêche en rien l'étude des marqueurs post-GEN qui indiquent la frontière à droite de l'élément.

6 Évaluation de la procédure

L'évaluation réalisée sur le second corpus montre que la performance générale de ce système est relativement bonne : 88 % des GEN sont correctement annotés, ce qui constitue une très nette amélioration par rapport à une annotation par défaut du 1er mot (60 %) de chaque définition. Cette amélioration de la performance est due à la prise en compte des VE (57 % des GEN) et des expressions relationnelles (8 %), ainsi que, dans une moindre mesure, au repérage des SPE antéposés (quelque 3 %). La plupart des erreurs (7/11 erreurs) sont dues aux règles générales, dont la performance pourrait être améliorée en les affinant davantage ou en les hiérarchisant différemment. Un apprentissage automatique de ces règles pourrait également être envisagé. Les autres erreurs (4/11) sont dues au non repérage d'un faux incluant, qui s'explique généralement par l'absence d'un des éléments clés du marqueur, comme les mots soulignés dans ces GEN : *Branche de la biologie* ou *Constituant d'une cellule*. Ces erreurs peuvent être corrigées en insérant le mot en question dans la règle, au risque cependant de la rendre plus ambigüe, notamment si elle est appliquée à un autre domaine. Ce serait sans doute le cas du mot *région* du GEN *Région de l'ADN*. Pour voir dans quelle mesure les marqueurs sont spécifiques à un domaine et vérifier ces résultats, il conviendrait de poursuivre l'évaluation sur un plus grand nombre de données, provenant de domaines variés.

7 Conclusion

Dans cet article, nous avons présenté un procédé de repérage et d'étiquetage des éléments génériques des définitions terminographiques, qui exploite les régularités formelles du sous-langage définitoire, afin de créer des règles de repérage les plus précises possibles de la frontière entre le générique et le(s) spécifique(s). Il ne nécessite aucun prétraitement morphosyntaxique ni recours à des ressources de TALN spécifiques, mais exige que les différentes étapes du traitement et les règles, susceptibles d'entrer en conflit, soient hiérarchisées. Après une présentation des quatre étapes de traitement des définitions et des difficultés inhérentes à leur application, nous avons montré que le système permet d'obtenir de bonnes performances globales. Cette performance est principalement obtenue grâce aux règles qui repèrent les spécifiques antéposés, les termes du domaine et les expressions relationnelles. Les résultats devraient cependant être consolidés sur des corpus plus grands, couvrant différents domaines, et des améliorations pourraient être étudiées en ayant recours à

des procédés statistiques ou d'apprentissage automatique. Finalement, si l'architecture de ce système semble transposable à d'autres langues, l'exploitation des variations morphosyntaxiques suppose toutefois que des règles de repérage spécifiques soient créées pour chaque langue, ce qui restreint considérablement leur réutilisabilité. Des techniques d'analyse indépendantes des langues basées sur l'apprentissage automatique des règles devraient permettre de pallier cette limite. Nous envisageons à l'avenir d'explorer ces pistes statistiques afin d'optimiser les performances du système. Nous étudierons également la possibilité d'étendre ce type de procédé à l'annotation des spécifiques, l'objectif final étant de disposer d'un analyseur qui facilite l'annotation de corpus de définitions terminographiques, en vue d'en étudier les éventuelles régularités dans la structure conceptuelle. Ce type d'analyseur pourrait finalement s'avérer intéressant pour enrichir les options de recherche dans les bases de données terminologiques.

Références

BARNBROOK, G. (2002). *Defining Language : A local grammar of definition sentences*. Amsterdam, Philadelphia: John Benjamins.

BOURJAULT, A. (2005). *Terminologie de la biosynthèse des protéines chez les cellules eucaryotes : anglais-français*. Université de Genève, École de traduction et d'interprétation.

IRIS, M. A., et al. (1988). Problems of the part-whole relation. *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*, 261-287.

ISO (2000). *Travaux terminologiques : principes et méthodes (ISO 704)*. Genève: ISO.

L'HOMME, M.-C. (2003). Indices de relations conceptuelles dans les définitions terminologiques. Application au domaine de l'informatique. Actes de *I Jornada Internacional sobre la Investigación en Terminología y Conocimiento Especializado*, 44-50.

LINGUA-PC. *Banque de terminologie du canton de Berne*. Chancellerie d'État du Canton de Berne.

REBEYROLLE, J. (2000). Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes. Actes d'*IC'2000*.

REY-DEBOVE, J. (1971). *Étude linguistique et sémiotique des dictionnaires français contemporains*. The Hague, Paris: Mouton.

SEPPÄLÄ, S. (2004). *Composition et formalisation conceptuelles de la définition terminographique*. Université de Genève, École de traduction et d'interprétation.

SEPPÄLÄ, S. (2005). Structure des définitions terminographiques : une étude préliminaire. Actes de *Terminologie et Intelligence Artificielle, TIA'05*, 19-29.

SEPPÄLÄ, S. (2006). Semi-Automatic Checking of Terminographic Definitions. Actes de *TermEval Workshop - LREC 2006*, 22-27.

VOSSEN, P., et al. (1989). Meaning and structure in dictionary definitions. *Computational Lexicography for Natural Language Processing*, 171-192.