

Structure des définitions terminographiques : une étude préliminaire

Selja Seppälä

Terminologie, Département de traitement informatique multilingue
École de traduction et d'interprétation, Université de Genève
40, bd du Pont-d'Arve, CH-1211 Genève 4
selja.seppala@eti.unige.ch
<http://www.unige.ch/eti/personnel/seppala.htm>

Résumé

Les recherches vont bon train dans le domaine de l'extraction d'informations définitoires dans du texte « libre », mais qu'en est-il de leur exploitation ultérieure, notamment dans le domaine de la terminographie ? Afin que des outils d'aide à la terminographie efficaces puissent voir le jour, il convient de disposer de tous les éléments formels indispensables à leur mise en œuvre. Ainsi, l'automatisation ne serait-ce que partielle des tâches de rédaction de définitions terminographiques, ou encore de contrôle de leur composition, nécessite une compréhension claire et précise des définitions de ce type. Or, la terminologie semble présenter un certain nombre de lacunes, tant théoriques que méthodologiques, concernant la structure interne des définitions. Pour tenter de pallier ce problème et, plus concrètement, de répondre à un certain nombre de questions relatives à la composition sémantique des définitions terminographiques, nous avons procédé à une étude empirique exploratoire, dont nous présentons ici les résultats. Ce travail sur corpus vise, d'une part, à dégager des régularités structurelles formalisables en vue, notamment, de leur exploitation informatique et, d'autre part, à voir dans quelle mesure l'étude des définitions peut être automatisée. Car automatiser la segmentation et l'étiquetage des éléments définitoires, et, à terme, l'extraction de mots-clés qui y sont contenus, c'est non seulement contribuer au développement d'outils d'aide à la terminographie, mais aussi rendre les définitions terminographiques existantes exploitables, tant à des fins terminologiques que pour des outils de TALN.

Mots-clés : définition terminographique, structure définitoire, classes conceptuelles, relations conceptuelles, théorie terminologique, patrons définitoires, annotation de corpus

1. Introduction

Nombreuses sont les études portant sur le repérage automatique d'éléments définitoires dans du texte « libre » à des fins diverses [Auger, 1997; Malaisé, *et al.*, 2004; Rebeyrolle, 2000a; 2000b, etc.], et notamment en vue d'enrichir les ressources terminologiques. Peu de recherches se penchent cependant sur la structure des définitions elles-mêmes. Celle-ci est pourtant d'un grand intérêt pour la terminologie et, plus particulièrement, pour l'activité terminographique. En effet, l'une des tâches principales du terminologue consiste à fournir des informations sur les concepts propres à un domaine et donc à les définir. Pour ce faire, il procède au dépouillement de corpus adéquats, dont il extrait les informations pertinentes, qu'il combine ensuite dans une définition. Que cette tâche se fasse manuellement ou

(semi-)automatiquement – et à plus forte raison dans le second cas de figure –, le terminologue est souvent confronté à une question bien précise : qu'est-ce qu'une « bonne » définition ?

La terminologie, qui régit les fondements de la rédaction de fiches terminographiques, pose les jalons d'une définition-type, sans pour autant en détailler la composition interne. La rédaction de ce type de définitions suit ainsi certaines règles dont les principes plus ou moins implicites sont néanmoins communément admis par la plupart des terminographes.

Or, à vouloir automatiser ne serait-ce que partiellement la rédaction ou le contrôle des définitions, il est essentiel d'en comprendre la structure interne et de la formaliser en vue de son exploitation informatique. Il semblerait, à l'issue d'un large inventaire de la littérature terminologique en matière de définitions [Seppälä, 2004], que cette discipline présente toutefois certaines lacunes d'ordre théorique et méthodologique en ce qui concerne leur composition sémantique. Et pourtant, les terminographes rédigent tout de même de « bonnes » définitions, jugées utiles¹, et ce, de manière intuitive, sans suivre de modèles trop contraignants.

Les fonctions mêmes de la définition et l'existence d'une poignée de contraintes définitoires nous amènent à postuler l'hypothèse suivante : bien que le terminographe jouisse d'une liberté de rédaction relativement grande, la structure formelle de la définition n'est pas totalement libre. Cette dernière se construit intuitivement à l'intérieur de certaines contraintes de rédaction. Il doit donc exister des structures définitoires implicites que l'on doit pouvoir découvrir et qui devraient nous permettre de définir des règles d'écriture et/ou de contrôle des définitions terminographiques les plus naturelles possibles, ou pour le moins nous fournir des informations utiles à l'automatisation de l'annotation.

Pour mesurer la validité de ce postulat, nous nous proposons d'étudier la structure interne de la définition au travers d'une analyse de corpus de définitions terminographiques. Plus précisément, nous tenterons de répondre aux questions suivantes : Quels types de traits définitoires semblent pertinents et pourraient être considérés comme nécessaires dans la définition terminographique ? Combien d'éléments spécifiques semblent suffisants pour définir ? Dans quel ordre convient-il de les placer ? De quoi dépendent ces traits et leur ordre (du domaine, de la nature du terme, etc.) ?

Nous présentons ici les résultats d'une recherche préliminaire en la matière, ainsi qu'un certain nombre de travaux futurs. Avant de nous pencher sur l'étude empirique à proprement parler et sur ses résultats, nous passerons brièvement en revue les caractères de la définition terminographique, ainsi que les conventions de rédaction généralement admises en terminographie. Ce tour de la question permettra de mieux comprendre ce qui guide les terminographes dans leur travail de rédaction de définitions, ainsi que de nous rendre compte des limites des connaissances actuelles en matière de structure définitoire.

Nous présenterons ensuite notre travail empirique. Dans un premier temps, nous évoquerons les caractéristiques du corpus étudié. Puis, nous exposerons rapidement la méthodologie suivie et les descripteurs utilisés pour marquer la structure des définitions, ainsi que les outils d'analyse employés. Pour terminer, nous proposerons une synthèse des analyses effectuées et des résultats observés. En conclusion, nous fournirons des éléments de réponse aux questions posées et nous esquisserons quelques pistes futures.

¹ Même si elles ne respectent pas toujours les principes théoriques, tel que celui de substituabilité au terme.

2. La définition terminographique

2.1. Caractères de la définition terminographique

Qu'est-ce qu'une définition terminographique ? Il s'agit d'une proposition formulée en langue, à propos d'un concept désigné par un terme dans un domaine donné, qui peut avoir diverses fonctions : décrire, expliquer, expliciter et/ou délimiter un concept ; distinguer des concepts les uns des autres ; reconnaître le défini ; attester l'existence d'un concept ; fixer un concept ; faire le lien entre unité linguistique, concept et référent ; structurer ou refléter un système conceptuel ; établir l'équivalence et la synonymie entre unités linguistiques ; remplir une fonction didactique et/ou normalisatrice. Les deux fonctions essentielles sont cependant, comme le précise la norme ISO 704 [2000], d'« identifier le *concept* et le différencier nettement des autres *concepts*. »

Cette proposition, qui a le statut d'équivalent périphrastique du défini et qui est en principe substituable au terme, se compose dans la plupart des cas de deux parties : un élément générique (GEN), ou incluant, rattachant le défini à un concept plus général, et un ou plusieurs éléments spécifiques (SPE) – également appelés caractères ou traits – qui précisent la portée conceptuelle du générique et qui permettent, notamment, de distinguer des concepts les uns des autres. On parle alors de définition en intension ou en compréhension.

Ex : *dépotoir à boues* = ^{GEN}*Chambre* ^{SPE1}*avec un sac à boues* ^{SPE2}*qui reçoit les eaux usées, retient les matières solides, et restitue les eaux au réseau de canalisations.*

Parfois, la définition terminographique prend la forme d'une définition en extension, auquel cas elle se compose d'une énumération d'espèces isonymes. Mais ces cas de figure sont assez rares dans la pratique terminographique.

Ex : *bateau* = ^{EXT1}*Véhicule servant à la navigation sur l'eau,* ^{EXT2}*corps flottant destiné au déplacement sur l'eau* ^{EXT3}*ou engin flottant.*

Dans certains cas, l'incluant ne correspond pas à une classe conceptuelle plus générale, mais à une expression relationnelle (comme *partie* ou *tout*). Dans ces génériques, nommés « faux incluant », un marqueur relationnel (comme *ensemble de...*) rattache un concept donné au défini, et c'est alors l'ensemble « marqueur+concept » qui est modifié par un ou plusieurs spécifiques. Ces derniers entretiennent avec les génériques, quelle que soit leur nature, un certain type de relation, en fonction de laquelle il est possible de classer les différents types de spécifiques. Ainsi Sager [1990] précise, par exemple, que c'est souvent la définition fonctionnelle – dont au moins un spécifique dénote la fonction du générique – qui prévaut en terminologie. Mais il reste difficile de se faire une idée de la proportion que représentent d'autres types de spécifiques (donc de relations) utilisés pour définir.

S'agissant du nombre de spécifiques que comporte une définition terminographique, de Bessé [1996] considère que les définitions ne devraient pas en compter plus de cinq, mais c'est là la seule attestation aussi précise que nous ayons pu trouver relativement à cette question, les autres auteurs restant généralement plus vagues. Quant à leur ordre à l'intérieur de la définition, il semblerait qu'il ne fasse pas l'objet d'un grand nombre de recherches. De là à savoir ce qui pourrait éventuellement influencer cet ordre...

2.2. Conventions de rédaction communément admises

À ces éléments définitoires de la définition terminographique, il convient d'ajouter des précisions sur les règles de rédaction conventionnellement suivies par les praticiens de la terminologie. [Voir notamment Auger et Rousseau, 1988; de Bessé, 1996; Dubuc, 1978; ISO

704, 2000; Rey, 1992; Rondeau, 1984] Ces conventions trouvent leur explication dans les principes théoriques de la définition [Seppälä, 2004], mais nous n'entrerons pas ici dans ces détails.

En ce qui concerne son contenu, la définition terminographique doit définir des concepts (et non des unités linguistiques) à un moment donné ; elle doit être spécifique au domaine ou au sous-domaine traité, sans pour autant contenir d'indication de domaine ; elle ne doit pas contenir le défini ; elle devrait contenir les caractères essentiels du concept ; elle ne devrait pas être négative, à moins qu'il ne s'agisse d'un concept négatif ; elle suppose la neutralité du point de vue.

Pour ce qui est de sa forme, la définition terminographique est, en principe, constituée de deux parties : un générique et un ou plusieurs spécifiques ; elle devrait être concise ; elle devrait définir la forme nominale de la dénomination, et de ce fait commencer par un nom ; elle consiste en une seule phrase, qui dans la tradition francophone commence par une majuscule et se termine par un point ; elle devrait éviter autant que possible tous les signes de ponctuation, à l'exception de la virgule. Autant de facteurs susceptibles d'imposer une contrainte sur la structure interne de la définition.

Dans l'ensemble, les apports théoriques et méthodologiques relatifs à la composition de la définition restent donc relativement vagues. Si la littérature souligne l'importance d'y inclure les caractères essentiels d'un concept, c'est-à-dire nécessaires et suffisants pour le comprendre et/ou le distinguer d'autres concepts, elle ne dit en revanche pas grand chose sur la nature, le nombre ou l'ordre de ces SPE à l'intérieur de la définition. Aussi, nous proposons-nous entre autres d'apporter des éléments de réponse à ces questions en recherchant, à l'intérieur de définitions terminographiques réelles (produites par des terminologues professionnels), des constantes structurelles au niveau des éléments définitoires en général et des éléments spécifiques en particulier.

3. Étude empirique

Pour pouvoir étudier la structure de définitions de type terminographique, notre corpus se devait d'être composé de définitions « bien formées » et issues de domaines variés, de façon à ne pas introduire de biais dans l'expérience. C'est pourquoi il est composé de 500 définitions terminographiques tirées de la banque de terminologie d'une administration cantonale suisse (*LINGUA-PC : Banque de terminologie du canton de Berne*) et couvrant un total de 28 domaines. Ces définitions, rédigées en français par des terminologues avertis, respectent – nous l'avons vérifié [Seppälä, 2002] – les conventions de rédaction communément admises, énumérées au point 2.2.

Afin d'observer d'éventuelles régularités dans la composition interne des définitions, il est nécessaire d'établir des comparaisons entre des définitions issues de domaines très différents. Pour pouvoir comparer ce qui est comparable, nous avons fait abstraction du contenu des définitions en les découpant, manuellement², en éléments génériques (*GEN*) et spécifiques (*SPE*), et en annotant (toujours manuellement) les segments ainsi obtenus avec des classes conceptuelles – susceptibles d'expliquer d'éventuelles régularités structurelles – pour les *GEN* et des relations conceptuelles pour les *SPE*.

² Il ne s'agit ici que d'une étude préliminaire visant à évaluer l'opportunité de poursuivre ces recherches à plus grande échelle. L'annotation manuelle réalisée dans le cadre de ce travail pourra servir de base à l'apprentissage automatique des tâches de segmentation et d'annotation des définitions.

Structure des définitions terminographiques

Les classes et les relations conceptuelles utilisées sont le fruit d'un travail à la fois de synthèse théorique – qui a permis l'adoption de jeux d'étiquettes de départ – et empirique – qui a permis de les affiner et de les définir au fur et à mesure de l'étiquetage, afin de les adapter à nos besoins. S'agissant des classes conceptuelles, nous sommes partie des onze classes supérieures proposées pour les noms³ dans le réseau lexical *WordNet*⁴, pour aboutir à un ensemble de 14 classes permettant de rattacher les GEN à des classes ontologiques : ABSTRACTION, ACTE, ACTIVITE, ANIME, ARTIFICIEL, ENTITE, ESPACE, ETAT, GROUPE, INANIME, NATUREL, PHENOMENE, TEMPS. Concernant les relations conceptuelles (au total 22) destinées à annoter les SPE, nous nous sommes inspirée des classifications proposées tant par les terminologues [notamment Gouadec, 1990; Kageura, 1997; Sager, 1990], que dans *WordNet* [2001] ou Sanfilippo, *et al.* [1999] : AGENT, BENEFICIAIRE, DESTINATAIRE, INSTRUMENT, OBJET_VISE, PATIENT; CAUSE, CONDITION, CONSEQUENCE, FONCTION, UTILITE; CONTENU, EXTENSION, GENRE, PARTIE, TOUT; DOMAINE, PROPR_ABSTRAITE, PROPR_METRIQUE, PROPR_PHYSIQUE, SPATIAL, TEMPOREL. Sans entrer dans de plus amples détails concernant les autres types d'étiquettes, ajoutons toutefois que nous avons également marqué, lorsque cela semblait pertinent, les indices (JONCTEURS) susceptibles de permettre un repérage et/ou la classification automatique des spécifiques.

Pour l'étiquetage du corpus, nous avons utilisé le langage XML, qui est extrêmement souple et modulable, et qui permet l'emploi de balises personnalisées. Le contenu de ces balises est défini dans une DTD séparée. Des feuilles de style adaptées aux différentes analyses ont permis de transformer le texte balisé en XML en un ensemble de données comportant les informations à observer, par exemple les étiquettes GEN et SPE, ou SPE et SPE, ou encore les étiquettes SPE ou JONCT avec le texte auquel elles sont associées (voir les exemples ci-dessous). Les listes ainsi obtenues sont ensuite triées de sorte à effectuer différents calculs de fréquence ou, dans le cas des JONCT, à permettre des rapprochements entre marqueurs, puis des généralisations sous forme de patrons morphosyntaxiques.

Exemple de fiche terminologique avant annotation :

```
NI 0000146
CM AD6
VE division de la police des districts
DF Division du corps de police qui se charge, dans les districts, des
tâches relevant de la gendarmerie et de la police judiciaire.
```

Exemple de fiche terminologique après annotation :

```
<FICHE langue="FR">
  <NI>0000146</NI>
  <CM>AD6</CM>
  <VE>division de la police des districts</VE>
  <DF>
    <GEN relation_VE="GENRE" classe_conceptuelle="GROUPE">Division du
    corps de police</GEN>
    <SPE relation_GEN="FONCTION" voir_SPE="DEBUT"><JONCT>qui se
    charge</JONCT>,</SPE>
    <SPE relation_GEN="SPATIAL">dans les districts,</SPE>
    <SPE relation_GEN="FONCTION" voir_SPE="FIN">des tâches relevant de
    la gendarmerie et de la police judiciaire.</SPE>
  </DF>
</FICHE>
```

³ Cette limitation aux *noms*, par opposition aux *verbes* ou aux *adjectifs*, est justifiée dans un travail terminologique du fait que la plupart des concepts définis se présentent sous la forme nominale.

⁴ L'avantage de ces classes est qu'elles ont été largement utilisées dans diverses applications en TALN, qu'elles semblent couvrir la plupart des cas de figure, et que les distinctions qu'elles opèrent entre les différents types conceptuels ne sont ni trop générales ni trop spécifiques.

Exemples de transformations et d'analyses appliquées à la définition ci-dessus :

Listes des patrons de GEN+SPE permettant de calculer leur fréquence :

{GROUPE}[FONCTION_début] [SPATIAL] [FONCTION_fin]

Listes des patrons de SPE+SPE permettant de calculer leur fréquence :

[FONCTION_début] [SPATIAL] [FONCTION_fin]

Listes des SPE+texte associé permettant de calculer le nombre de mots par SPE :

[SPATIAL]dans les districts ⇒ SPATIAL = 3 mots/SPE

Listes des JONCT(texte) afin de dégager des patrons de marqueurs morphosyntaxiques :

JONCT(qui se charge) ⇒ qui+["se charger"_présent]

4. Analyses effectuées et résultats

On distingue principalement trois types d'analyse : le premier concerne plutôt des aspects généraux de la définition terminographique, tels que le nombre de différents GEN et SPE, et leur répartition, le nombre de SPE par définition, la relation du GEN au terme défini (GENRE/PARTIE/TOUT/VIDE), ou la répartition par domaines. Le deuxième est davantage axé sur ses aspects structurels et plus particulièrement sur les patrons définitoires (GEN+SPE ; SPE+SPE), et sur le comportement des éléments définitoires pris individuellement. Et pour terminer, nous avons effectué quelques analyses d'ordre plutôt linguistique (nombre de mots par définition et par SPE ; JONCTEURS). Nous ne présentons ici que les résultats les plus significatifs. Ceux-ci ne sont pour le moment que de simples constatations de certains comportements plus réguliers. Il s'agirait à présent d'étudier ces phénomènes de plus près pour tenter de les expliquer.

4.1. Aspects généraux de la définition terminographique

S'agissant des aspects généraux de la définition, cette étude apporte une confirmation empirique de certains points théoriques rencontrés dans la littérature.

Elle vérifie l'assertion de de Bessé [1996] selon laquelle une définition (terminographique) ne comporte pas plus de cinq spécifiques. En effet, les définitions étudiées ne comportent jamais plus de cinq SPE et, dans pratiquement 95 % des cas, leur nombre varie de un à trois, ce qui porte le nombre moyen de SPE par GEN à deux.

nb. de spe par déf.	fréq.	%
1 spe	161	32.2
2 spe	201	40.2
3 spe	112	22.4
4 spe	19	3.8
5 spe	7	1.4
total	500	100%

S'agissant du mode définitoire, notre échantillon de définitions confirme que les définitions en compréhension priment en terminographie [Depecker, 2002]. En effet, la très grande majorité (91,4 %) des GEN entretient une relation de GENRE par rapport au terme, le reste (8,6 %) étant composé pour près de la moitié de la relation PARTIE (4 %), ainsi que de l'étiquette VIDE (3 %) qui indique une définition en extension ; seule une infime part des GEN étudiés dénote la relation TOUT (1,4 %) ou un FAUX⁵ générique (0,2 %).

De même, elle permet de nous rendre compte que les définitions fonctionnelles sont les plus répandues [Rey, 1992; Sager, 1990] : FONCTION = 22,2 % des SPE ; PROPR_ABSTRAITE = 15,6 % des SPE ; la part des autres SPE est inférieure à 7,4 %.

⁵ Il ne s'agit pas ici d'un faux GEN, tel que nous l'avons défini plus haut, mais d'une locution introduisant la définition d'un adjectif ; il n'est de ce fait pas considéré comme étant vraiment un GEN.

4.2. Aspects structurels de la définition terminographique

4.2.1 Patrons définitoires

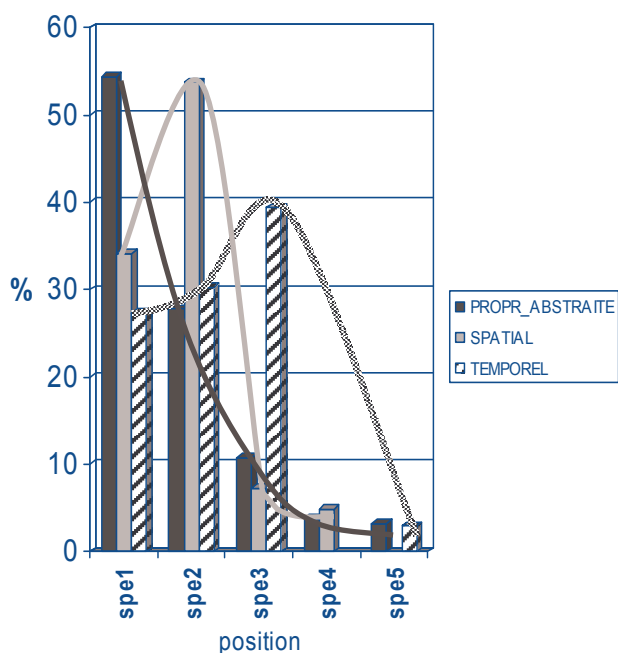
Toutes les définitions qui ont un GEN ont nécessairement au moins un SPE. C'est là une différence marquante par rapport à la lexicographie où l'on trouve davantage de définitions par simple synonymie, donc avec un simple GEN (cas qui n'apparaît jamais ici). Par ailleurs, nous pouvons affirmer qu'en français, peu de SPE apparaissent avant le GEN.

Par rapport à l'ensemble du corpus, les régularités au niveau des patrons définitoires en général ne sont pas très marquées, seul 8,6 % des définitions correspondent à des patrons se répétant plus de 10 fois (fréquences de 18, 14 et 11). En revanche, la comparaison des patrons à différents niveaux (patrons GEN+1SPE, GEN+2SPE, etc. ou GEN+1er SPE, GEN+2e SPE, etc. selon que l'on tient compte ou non du nombre total de SPE), ainsi que l'observation des 15 patrons les plus fréquents, montrent qu'il convient de se concentrer sur l'étude du comportement du GEN+1er SPE qui le suit (voire des deux premiers SPE qui le suivent), car c'est là que l'on trouve le plus de régularités. Ces résultats montrent également qu'il est plus intéressant d'étudier le comportement des éléments définitoires séparément.

4.2.2 Comportement des éléments définitoires

L'observation du comportement des éléments définitoires pris isolément a, en effet, permis de mettre en lumière un certain nombre de régularités ou pour le moins de comportements privilégiés. Ainsi, il apparaît que certains GEN privilégient certains types de SPE. Par exemple, un GEN de type ABSTRACTION (*Principe selon lequel...*) semble plutôt être défini à l'aide des SPE de type CONTENU, FONCTION, PROPR_ABSTRAITE ou OBJET_VISE. De même, le SPE de type FONCTION ne semble pas être utilisé pour définir des concepts de type PHENOMENE ou NATUREL alors qu'il est le plus fréquent.

Nous relevons également un certain nombre de régularités et de tendances dans le comportement des SPE. Nous observons par exemple, dans la succession des SPE, que FONCTION et PROPR_ABSTRAITE forment un tandem où l'un suit généralement l'autre, et vice versa. De même, la succession CONTENU + INSTRUMENT ou CONDITION est également récurrente.



Ex : inspection d'admission = Inspection

PR ABST	officielle	FONCT	qui a pour but
d'établir si un bateau est conforme			
aux prescriptions relatives à la			
construction et s'il est équipé			
conformément aux prescriptions.			

Par ailleurs, il est extrêmement intéressant de noter que tous les types de SPE n'ont pas le même comportement selon leur position ; leur distribution à l'intérieur de la définition varie selon leur type. Ainsi trois tendances se dégagent (voir le graphique ci-contre) : les SPE apparaissant plutôt en début de définition (FONCTION, PROPR_ABSTRAITE, OBJET_VISE, CONTENU, etc.), les SPE apparaissant le plus souvent en deuxième position (SPATIAL,

INSTRUMENT, CONDITION, AGENT, etc.) et ceux qui figurent de préférence en troisième position (TEMPOREL, UTILITE, CONSEQUENCE).

4.3. Analyse d'ordre linguistique

Dans la perspective d'une automatisation de ce type d'annotation, nous avons tenté de voir dans quelle mesure les éléments textuels étaient susceptibles de servir au repérage automatique des SPE, mais aussi à la délimitation du GEN. Les résultats nous montrent que la possibilité de dégager des patrons lexico-syntaxiques est très variable. En observant le cas des quatre SPE les plus fréquents, nous constatons que le total des SPE avec *joncteur non pertinent*⁶ et *aucun joncteur* s'élève respectivement à 34,8 %, 28,4 % et à 20 % pour *PROPR_ABSTRAITE*, *OBJET_VISE* et *AGENT*, mais qu'il n'est que de 9,3 % dans le cas de *FONCTION*, ce qui laisse à penser que ce dernier sera sans doute plus facilement repéré par une machine que les trois autres. Il n'en reste pas moins que, même dans les trois autres cas, l'automatisation pourrait être envisagée, sachant qu'au moins 65 % des SPE comportent un joncteur "explicite". Un problème majeur risque cependant de se poser : le "surrepérage" d'éléments (phénomène de bruit), du fait que certains marqueurs de SPE sont ambigus (par exemple, *pour* et *dont*+*[NP]*+*[vb]* s'appliquent à la fois à *FONCTION* et à *PROPR_ABSTRAITE*); et du fait que ces joncteurs sont susceptibles de s'appliquer non seulement à des SPE, mais aussi à des éléments à l'intérieur du SPE qui ne sont pas rattachés au GEN. Malgré tout, ce résultat est suffisamment encourageant pour que l'on tente une expérience d'automatisation du repérage au moins de ces quelques relations.

5. Conclusion

À bon nombre d'égards, notre recherche s'apparente à une étude similaire réalisée par Sager et L'Homme [1994], à cette différence que leurs conclusions ne visaient pas, nous semble-t-il, l'observation de régularités structurelles, mais plutôt une description méthodologique de ce type d'analyse. De fait, le découpage et le métalangage que nous proposons pour l'étude des définitions correspondent, à quelques exceptions près, aux leurs. La plupart des attributs de nos descripteurs sont les mêmes (marquage du domaine, de la classe conceptuelle du GEN, de la relation du GEN au terme, ou du type de SPE); les quelques différences apparaissent surtout au niveau de leurs valeurs, notamment des classes conceptuelles utilisées ou des types de relations terme-incluant considérées.

Afin de nous rendre compte de l'opportunité de poursuivre nos recherches, voyons dans quelle mesure nos résultats permettent de répondre aux questions posées au départ.

Quels types de traits semblent pertinents et pourraient être considérés comme nécessaires dans la définition terminographique ?

Tout d'abord, notre analyse confirme que toutes les définitions en compréhension, à quelques exceptions près, ont un générique et au moins un spécifique. Ensuite, s'agissant de la nature des relations, nous avons observé que les éléments *FONCTION* et *PROPR_ABSTRAITE*

OBJET_VISE		
indices formels	nb.	%
["de"]+([det])+[nom]	36	53.7
joncteur non pertinent	12	17.9
aucun joncteur	7	10.4
<i>pour</i>	5	7.5
<i>sur</i>	3	4.5
["applicable"]+["à"]	2	3.0
["concerner"]	2	3.0
total	67	100%

AGENT		
indices formels	nb.	%
[vb participe]+ <i>par</i>	38	50.7
joncteur non pertinent	15	20.0
[devoir]+[à]	13	17.3
<i>par</i> + <i>[det]</i> + <i>[nom]</i>	5	6.7
<i>que</i> + <i>[det]</i> + <i>[nom]</i> + <i>[vb]</i>	4	5.3
total	75	100%

⁶ Lorsque ni la fréquence, ni le sémantisme du joncteur ou de son patron ne semblent utiles au repérage du SPE.

Structure des définitions terminographiques

semblent être les spécifiques les plus utilisés, bien qu'ils ne soient pas « nécessaires » à proprement parler. Mais ces propos sont à nuancer, car on ne peut considérer ces deux catégories comme les seules pertinentes. La pertinence des SPE semble, en effet, dépendre du type de générique que l'on définit. Aussi, un générique de type ACTIVITE sera-t-il principalement défini à l'aide du spécifique OBJET_VISE, de même que ABSTRACTION le sera plutôt par un trait précisant son CONTENU, pour ne prendre que ces exemples.

Combien d'éléments spécifiques semblent suffisants pour définir ?

D'après les résultats de nos analyses, il semblerait que le nombre de spécifiques suffisant soit souvent de un et, très souvent, de deux, mais en tout cas rarement plus de trois et jamais plus de cinq. Si l'on s'en tient à la moyenne des spécifiques par définition, le nombre « idéal » qui semble se dégager de notre corpus serait de deux.

Dans quel ordre convient-il de les placer ?

Tout comme il n'y a pas de traits nécessaires, il n'y a pas non plus d'ordre absolu qui s'applique à toutes les définitions. Là encore, il y a lieu d'observer les tendances fortes qui se dégagent du corpus : le GEN apparaît presque toujours en premier ; les SPE occupent souvent des positions privilégiées selon leur type ; la succession des GEN+SPE et SPE+SPE présentent des régularités intéressantes, notamment une probabilité plus grande de trouver les combinaisons de SPE suivantes : AGENT + FONCTION OU CONTENU, CONTENU + CONDITION OU INSTRUMENT, CAUSE + PROPR_ABSTRAITE, ou encore BENEFICIAIRE + CONTENU. Ces régularités peuvent s'avérer intéressantes lors de l'automatisation de l'étiquetage.

De quoi dépendent ces traits et leur ordre ?

Comme nous l'évoquions en réponse à la première question, il semblerait que le type de classe conceptuelle à laquelle est rattaché le GEN ait une certaine influence pour le moins sur la nature des SPE, si ce n'est aussi sur leur ordre. Il y aurait donc lieu d'approfondir la question des relations entre les différents types de GEN et de SPE, ainsi qu'entre les SPE et les domaines. Ainsi, si notre étude ne nous autorise pas encore à répondre à cette question, elle nous permet néanmoins de proposer quelques modèles de définitions « bien formées », par exemple :

GEN [GROUPE/ANIME/ENTITE/ESPACE] + SPE [FONCTION] + SPE [PROPR_ABSTRAITE] .

école de ski = ^{GROUPE}Établissement_{FONCT} qui enseigne le ski par groupes de plus de quatre personnes, _{PR_ABS} qui doit obtenir pour ce faire une autorisation de la Division du tourisme.

À défaut de nous fournir des modèles prêts à l'emploi, ni même des résultats catégoriques sur les définitions en général, ou des réponses tranchées à nos questions, cette analyse aura au moins permis de vérifier certaines affirmations rencontrées dans les ouvrages terminologiques et qui sont, la plupart du temps, avancées sans justification empirique. Elle aura également permis de mettre à jour certaines tendances très intéressantes, surtout, au niveau du comportement des éléments définitoires, qui peuvent s'avérer utiles pour l'automatisation de l'annotation et qui indiquent que cette étude mérite d'être poursuivie. Il serait, en effet, intéressant de diversifier encore les sources des définitions et d'augmenter la taille du corpus pour vérifier les résultats et les tendances observées, et passer ensuite à leur interprétation. Mais un élargissement du corpus devrait s'accompagner au moins d'une semi-automatisation des opérations, piste que nous entendons explorer plus en détail dans nos futurs travaux. Ce type de développement nous conduira notamment à nous pencher davantage sur les méthodes d'acquisition et d'extraction de relations conceptuelles [Auger, 1997; Claveau, *et al.*, 2001;

Malaisé, *et al.*, 2004; Marshman, *et al.*, 2002; Rebeyrolle, 2000a; 2000b, etc.], et donc à comparer nos modèles avec les indices de repérage utilisés dans d'autres cas de figure. Ce sera l'occasion de tester l'applicabilité de ces patrons à l'identification des composants de la définition.

La présente étude ouvre également bon nombre d'autres perspectives. Pourquoi, en effet, ne pas l'étendre à d'autres champs de la fiche terminologique ? Par exemple à la note, ce qui permettrait de mieux préciser la nature d'une « bonne » définition, et de voir où s'arrête la définition et où commence la note. Il pourrait également s'avérer utile d'analyser des corpus de définitions spécialisées dans des domaines donnés, afin de voir si les régularités sont plus grandes, ou de types différents. De même, il pourrait être intéressant d'étudier des corpus multilingues, des définitions non terminographiques (lexicographiques ou encyclopédiques, par exemple) ou des définitions destinées à d'autres publics cibles (comme les enfants), pour voir si l'on observe des différences de structuration. Et pour tirer pleinement parti des éventuelles méthodes automatiques d'analyse ainsi développées, il serait également envisageable de pousser les recherches concernant les différents éléments spécifiques, afin d'en extraire les informations pertinentes en vue de leur exploitation informatique.

Notre objectif est en fin de compte double. D'une part, il s'agit de sonder les définitions terminographiques pour tenter d'en dégager la composition sémantique interne et répondre ainsi à certaines interrogations concrètes que soulève leur rédaction. D'autre part, nous cherchons, autant que possible, à formaliser ce type de définitions en vue de leur exploitation informatique, qu'il s'agisse de les générer ou d'en contrôler la composition (semi-)automatiquement. Ce travail se veut donc une modeste contribution tant à la construction d'une théorie plus aboutie, qu'à l'élaboration d'outils d'aide à la terminographie efficaces.

Remerciements

Nous tenons à remercier Pierrette Bouillon, Bruno de Bessé et Donatella Pulitano, nos relecteurs, pour leur disponibilité, la pertinence de leurs remarques et leurs encouragements. Nous sommes également très reconnaissante au professeur de Bessé pour ses enseignements en terminologie et pour l'intérêt qu'il porte à ces travaux.

Références

WordNet 1.7.1 Copyright ©, by Princeton University. All rights reserved, <http://www.cogsci.princeton.edu/~wn/man1.7.1/uniqbeg.7WN.html#sect0> (3 janvier 2005).

AUGER A. (1997), *Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles*, Université de Neuchâtel, Neuchâtel, 224 p.

AUGER P. et ROUSSEAU L.-J. (1988), *Méthodologie de la recherche terminologique*, Office de la langue française, Québec, 80 p.

CLAVEAU V., SÉBILLOT P., BOUILLON P. et FABRE C. (2001), "Acquérir des éléments du lexique génératif : quels résultats et à quels coûts?" in *Traitement automatique des langues*, vol. 42, n° 3, pp. 729-753.

DE BESSÉ B. (1996), "Chapitre 2.3.: La définition", in *Notes de cours*, non publié, pp. 68-87.

DEPECKER L. (2002), *Entre signe et concept : Éléments de terminologie générale*, Presses Sorbonne nouvelle, Paris, 198 p.

DUBUC R. (1978), *Manuel pratique de terminologie*, Linguatex, CILF, Montréal, Paris, 98 p.

Structure des définitions terminographiques

- GOUADEC D. (1990), *Terminologie: Constitution des données*, AFNOR, Paris, XV, 218 p.
- ISO 704 (2000), *Travaux terminologiques : principes et méthodes (ISO 704)*, 2e éd., ISO, Genève, 41 p.
- KAGEURA K. (1997), "On Intra-Term Relations of Complex Terms in the Description of Term Formation Patterns", in *Mélanges de Linguistique Offerts à R. Kocourek*, Les Presses d'ALFA, Halifax, pp. 105-111.
- MALAISÉ V., ZWEIGENBAUM P. et BACHIMONT B. (2004), "Extraction d'informations sémantiques pour l'aide à la construction d'ontologies différentielles", In *Actes Journées d'étude Terminologie, Ontologie et Représentation des Connaissances*, Lyon.
- MARSHMAN E., MORGAN T. et MEYER I. (2002), "French patterns for expressing concept relations", in *Terminology*, vol. 8, n° 1, pp. 1-29.
- REBEYROLLE J. (2000a), *Forme et fonction de la définition en discours*, Université Toulouse II-Le Mirail, Toulouse, 227 p.
- REBEYROLLE J. (2000b), "Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes", In *IC'2000 - Actes de la conférence Journées francophones d'Ingénierie des Connaissances, 10-12 mai 2000*, Toulouse.
- REY A. (1992), *La terminologie : noms et notions*, 2e édition corrigée, "Que sais-je ?" n° 1780, Presses Universitaires de France, Paris, 128 p.
- RONDEAU G. (1984), *Introduction à la terminologie*, 2e éd., Gaëtan Morin, Québec, XLV, 238 p.
- SAGER J. (1990), *A practical course in terminology processing*, John Benjamins, Amsterdam, Philadelphia, 254 p.
- SAGER J.C. et L'HOMME M.-C. (1994), "A model for the definition of concepts : rules for analytical definitions in terminological databases", in *Terminology*, vol. 1, n° 2, pp. 351-373.
- SANFILIPPO A., CALZOLARI N., GAIZAUSKAS R., SAINT-DIZIER P., *et al.* (1999), *EAGLES LE3-4244, Preliminary Recommendations on Lexical Semantic Encoding, Final Report*, 289 p.
- SEPPÄLÄ S. (2002), *La définition terminologique: Analyse d'un corpus*, Travail de séminaire non publié, Université de Genève, 22 p.
- SEPPÄLÄ S. (2004), *Composition et formalisation conceptuelles de la définition terminographique*, Université de Genève, École de traduction et d'interprétation, Genève, 200 p.