RECITAL 2010, Montréal, 19–23 juillet 2010

Automatiser la rédaction de définitions terminographiques : questions et traitements

Selja Seppälä
TIM/ISSCO, ETI, Université de Genève, Suisse
seppala2@etu.unige.ch

Résumé. Dans cet article, nous présentons une analyse manuelle de corpus de *contextes conceptuels* afin (i) de voir dans quelle mesure les méthodes de TALN existantes sont *en principe* adéquates pour automatiser la rédaction de définitions terminographiques, et (ii) de dégager des question précises dont la résolution permettrait d'automatiser davantage la production de définitions. Le but est de contribuer à la réflexion sur les enjeux de l'automatisation de cette tâche, en procédant à une série d'analyses qui nous mènent, étape par étape, à examiner l'adéquation des méthodes d'extraction de définitions et de contextes plus larges au travail terminographique de rédaction des définitions. De ces analyses émergent des questions précises relatives à la *pertinence* des informations extraites et à leur *sélection*. Des propositions de solutions et leurs implications pour le TALN sont examinées.

Abstract. A manual corpus analysis of *conceptual contexts* is presented in order (i) to indicatively evaluate to what extent NLP methods can *in principle* be used to automate the production of terminographic definitions, and (ii) to identify central questions to be answered if one wants to further automate the task. The objective is to contribute to reflection on the challenges faced by the automation of this task. Through a series of analyses, the adequacy of extraction methods for *defining* or *knowledge-rich contexts* is examined in the light of the terminographic activity of definition writing. Precise questions emerge from these analyses relating to the *relevance* and the *selection* of the extracted information. Some solutions are proposed and their implications to NLP reviewed.

Mots-clés : Terminologie, définitions terminographiques, sélection des traits, pertinence des traits, extraction de définitions, contextes conceptuels, traitement automatique des définitions.

Keywords: Terminology, terminographic definitions, feature selection, feature relevance, definition extraction, conceptual contexts, definition processing.

1 Introduction

En terminologie, la rédaction de définitions implique, traditionnellement, le dépouillement *manuel* de vastes corpus textuels spécialisés. Afin de faciliter la tâche, le terminologue a souvent recours au TALN pour automatiser autant que possible la production de ressources terminologiques (dictionnaires, bases de données, etc.). S'agissant de la rédaction de définitions, au moins deux méthodes de TALN peuvent être intégrées à la méthodologie terminographique : les méthodes d'acquisition automatique de définitions « canoniques », par exemple (Cartier, 2004; Malaisé *et al.*, 2004; Rebeyrolle & Tanguy, 2000; Rebeyrolle,

2000, pour le français) et (Meyer, 2001; Pearson, 1998, pour l'anglais), et celles d'extraction d'informations ou de relations conceptuelles (par exemple, Marshman *et al.*, 2002). Cependant, dans la perspective du terminologue traditionnel, il est légitime de se demander (i) dans quelle mesure l'application de ces méthodes de TALN permet, *en principe*, de réaliser un travail définitoire équivalent au travail manuel et (ii) ce qui pourrait être envisagé pour parvenir à une plus grande automatisation de la tâche. Dans cet article, nous tentons de répondre à ces questions à titre indicatif. Pour ce faire, nous analysons manuellement un corpus de *contextes conceptuels* tirés de corpus textuels spécialisés. L'adéquation des deux méthodes citées plus haut est évaluée sur la base (i) du contenu du corpus de contextes et (ii) par comparaison avec des définitions rédigées manuellement suivant la méthodologie terminographique enseignée et pratiquée par la plupart des terminologues dans un cadre professionnel. Cet examen vise ainsi l'adéquation des méthodes de TALN à la méthodologie de rédaction des définitions terminographiques, et non l'inverse. Il ne s'agit donc pas d'évaluer les performances de ces méthodes en soi, mais de voir quelle peut être leur couverture *théorique* lorsqu'elles sont appliquées à la terminographie. Nous proposons ensuite de mener une réflexion sur la nature des informations en corpus et leur relation aux définitions. Ces réflexions nous amèneront à identifier des questions centrales auxquelles la terminologie doit répondre pour parvenir à une plus grande automatisation de la production de définitions. Ces considérations peuvent également être utiles pour le TALN en matière de génération automatique de définitions.

Après une brève présentation du corpus (section 2), nous proposons, premièrement (section 3), une analyse quantitative (mais non significative) de la proportion de définitions canoniques figurant dans un corpus de dépouillement terminographique et en comparons les résultats aux définitions issues d'un travail terminographique classique (*manuel*). Un décalage apparent nous amène, deuxièmement (section 4), à considérer les contextes formant le corpus à la lumière de la méthodologie terminographique que l'on souhaite automatiser, à savoir la production de définitions. Ceci implique la prise en compte d'une définition de la définition qui corresponde à cette pratique : un énoncé construit à partir de *traits* extraits de corpus. Une nouvelle analyse du corpus en termes de traits nous permet de soulever la question de la pertinence des informations, ce qui nous conduit à un réexamen des traits selon leur degré de pertinence (section 5). Cet examen permet d'identifier l'une des problématiques centrales du traitement (en l'occurrence, la génération) automatique des définitions en terminologie : la sélection des traits à inclure dans une définition. Cette problématique recouvre différentes questions qui doivent être résolues pour toute (semi-)automatisation du travail terminographique. Finalement (section 6), nous proposons quelques pistes de solutions possibles, ainsi que l'examen d'une solution empirique à la question de la sélection d'informations définitoires.

2 Présentation du corpus de contextes conceptuels

Les résultats indicatifs présentés dans ce travail proviennent de l'analyse de 127 *contextes conceptuels* (voir paragraphe suivant) relatifs à 56 concepts tirés de cinq domaines distincts : 21 concepts du domaine des *catastrophes écologiques*, 18 du *traitement des déchets*, 4 de la *météorologie*, 5 de la *lutte antidopage* et 8 du *stress*. Les corpus spécialisés dont sont tirés les contextes regroupent des textes rédigés par des experts des différents domaines. Ces textes comportent la terminologie des domaines concernés et apportent des informations sur les concepts du domaine, ainsi que sur l'usage des termes. Les sources ont donc été sélectionnées pour leur fiabilité et pour la richesse de leur contenu informatif [1]. Ces données proviennent

1. Sur les critères de sélection de textes à inclure dans un corpus de dépouillement terminographique, voir notamment (Pearson, 1998).

de travaux de dépouillement effectués dans le cadre de cours de méthodologie terminographique dispensés à l'École de traduction et d'interprétation de l'Université de Genève ; la nature du corpus peut de ce fait sembler quelque peu artificielle. Cependant, aussi bien la méthodologie de constitution des corpus, que celles d'extraction manuelle des données et de rédaction des définitions sont les mêmes que celles qui sont pratiquées par la plupart des terminologues dans un cadre professionnel. Soulignons également qu'au vu du nombre restreint de concepts examinés, la présente étude ne vise ni à l'exhaustivité ni à la représentativité, son but est principalement de voir dans quelle mesure les méthodes de TALN sont *en principe* adéquates pour l'automatisation de la rédaction de définitions (sachant que les chiffres indiquent seulement une tendance à vérifier sur d'autres données [2]).

Un *contexte conceptuel* correspond à un extrait de texte tiré d'un ouvrage spécialisé qui comporte une ou plusieurs unité(s) d'information conceptuelle (*trait(s)*, voir section 4) relative(s) au concept à définir, y compris parfois des définitions canoniques. Comme le précise Meyer (2001, 282, 299), ce type de contexte a une triple fonction en terminographie : fournir des définitions, fournir des points de départ pour la rédaction de définitions, et enrichir les connaissances du terminologue relativement au domaine traité. La notion de contexte ici considérée est cependant plus large et moins formelle que celle qui est proposée notamment par (Meyer, 2001; Pearson, 1998; Rebeyrolle, 2000), puisqu'elle ne suppose pas forcément la présence du terme désignant le concept, même si, de fait, il y figure souvent. L'extraction des contextes a été réalisée manuellement après identification de toutes les informations sur le concept susceptibles ou non d'être inclues dans une définition. Le repérage des ces informations étant basé sur le contenu-même des contextes, on peut considérer qu'il n'y a que peu de *silence* dans les informations extraites. Chaque concept du corpus est, par ailleurs, accompagné d'une définition rédigée manuellement [3] suivant la méthodologie traditionnelle. La structure générale des données pour chaque concept du corpus est la suivante [4] :
`<concept>[<contexte(s)>[<trait(s)>|<déf_en_contexte>]][<déf>].`

Voici l'exemple du concept ***gravité*** dans le domaine des *catastrophes géologiques* :

Ctxt$_1$ En faisant référence à un [*espace du danger dont*] [*les deux dimensions*] sont la gravité et la probabilité, [*les grandes catastrophes ont* une faible fréquence mais *une gravité importante.*]

Ctxt$_2$ [*Un danger peut être représenté selon*] [*deux paramètres qui sont la gravité*] et la probabilité. C'est en agissant sur ces deux axes que l'on pourra diminuer le nombre et [*l'ampleur des catastrophes*] qui ne [*sont* que *réalisation du danger.*]

Déf. *Paramètre représentant l'une des deux dimensions du danger, qui permet de mesurer l'importance d'une catastrophe.*

Le tableau 1 récapitule le nombre de contextes par concept, d'occurrences de traits [5] par concept et d'occurrences de traits par contexte. Si le nombre de contextes par concept est en moyenne de 2,3, il importe de noter pour la suite que le nombre d'occurrences le plus fréquent, c'est-à-dire le mode, est de 1 contexte par concept (28/56 concepts, soit 50 %). Au total, 75 % des concepts dans le corpus étudié font l'objet d'au maximum 2 contextes. Les concepts faisant l'objet de 3 contextes ou plus sont par conséquent, ici, exceptionnels.

2. Les données sont actuellement annotées par un autre terminologue en vue d'une évaluation interannotateurs.

3. Les définitions existantes pour chaque concept ont été, pour la plupart, rédigées par nous-même dans le cadre des cours de terminographie ; il y aurait donc lieu d'effectuer le même type d'étude sur des définitions rédigées par un tiers.

4. Nous n'entrons pas, ici, dans les détails du schéma d'annotation XML utilisé pour l'analyse des données.

5. Nous distinguons plus loin le nombre d'occurrences de traits (499) des types de traits (380).

	contextes/concept	occ. de traits/concept	traits/contexte
totaux	127/56	499/56	499/127
moyenne	2,3	8,9	3,9
mode	1 (50 %)	3 (16 %)	1 (24 %)

TABLE 1 – Contextes et traits par concept et par contexte

3 Définitions canoniques dans le corpus

Pour voir dans quelle mesure l'extraction automatique de définitions à structure canonique peut être utilisée pour l'automatisation de la production de définitions, nous avons identifié toutes les définitions canoniques du corpus. Nous distinguons les définitions *entières* des définitions *partielles* [6]. Une définition est considérée comme entière lorsqu'elle correspond au patron général *générique+spécifique* et qu'elle peut être extraite suivant les méthodes proposées notamment par Rebeyrolle (2000) ou Cartier (2004), pour le français. En ce sens, elle peut être (ou a effectivement été, dans les définitions basées sur ces données) reprise telle quelle du corpus de dépouillement, moyennant des modifications mineures de mise en forme pour satisfaire les conventions de rédaction dictionnairiques. Elle est considérée comme partielle lorsqu'il manque un ou plusieurs traits — en particulier le générique, mais aussi un spécifique, si on la compare à la définition qui a été rédigée à partir des contextes disponibles. Parmi les 127 contextes analysés, seul 21

	nb. concepts	%	nb. contextes	%
définitions (entières ou partielles)	16	29	21	16,5
définitions entières	11	19,6	13	10,2
définitions partielles	5	8,9	8	6,3

TABLE 2 – Définitions dans le corpus

(16,5 %) comportent une définition entière ou partielle. Dans la mesure où un même concept peut faire l'objet de plusieurs contextes comportant une définition [7], il convient plutôt de rapporter le total des définitions « canoniques » au nombre de concepts dans le corpus. Ainsi considéré, le total est de 16 définitions pour 56 concepts (28,6 %), parmi lesquelles 11 définitions (19,6 %) sont entières et 5 (8,9 %), partielles. Ces chiffres étant relativement réduits, il semble raisonnable de dire que les corpus (pour le moins ceux qui ont été utilisés dans ces travaux, mais ces chiffres semblent confirmer les observations de Meyer (2001, 284)) ne comportent que peu de définitions à part entière.

Du point de vue de l'automatisation de la production de définitions, ces résultats impliquent que même si un système avait extrait toutes les définitions du corpus, moins d'un tiers des concepts seraient au final accompagnés d'une définition en bonne et due forme. Or, chacun des 56 concepts du corpus s'est vu attribuer une définition sur la base du même corpus étudié. Donc, en principe, le recours aux seules méthodes d'extraction automatique de définitions canoniques est insuffisant. Ce décalage s'explique par le fait que la méthodologie terminographique de rédaction de définitions ne repose pas exclusivement sur le repérage de définitions canoniques, mais aussi sur celle d'informations inclues dans des contextes conceptuels plus larges. Ces informations ou traits permettent ensuite de construire une définition canonique, d'où l'intérêt de recourir à l'extraction de contextes conceptuels plus larges. En effet, comme le souligne Meyer (2001, 284), même s'ils ne comportent pas de définitions à part entière, les contextes extraits (semi-)automatiquement n'en sont pas moins utiles pour le terminologue. La question serait alors plutôt de savoir comment

6. Voir aussi les *définitions formelles* et *semi-formelles* chez Pearson (1998).
7. Par exemple, des contextes tirés de sources distinctes.

passer à une (semi-)automatisation de l'étape suivante : la construction de définitions. Il convient, pour ce faire, de prendre en compte la méthodologie traditionnelle, afin d'identifier les questions qui se posent et d'en examiner les conséquences sur l'automatisation de la production de définitions.

4 Traits dans le corpus

Partant de la méthodologie terminographique de rédaction des définitions, nous proposons la définition de *définition* suivante : *une définition est un énoncé composé d'un ensemble de traits tirés d'un corpus spécialisé*. Elle comporte généralement un trait générique et un ou plusieurs trait(s) spécifique(s). Un *trait* correspond à une unité d'information textuelle contenue dans un contexte. Il ne s'agit pas d'une unité syntaxique mais conceptuelle. Ainsi, une phrase à l'intérieur d'un contexte peut être subdivisée en plusieurs traits, dans la mesure où ceux-ci peuvent chacun être caractérisés par un type de classe ou de relation conceptuelle [8] donné et être autonomes par rapport aux autres traits. Ces traits sont repérés dans les corpus de dépouillement. Une fois les traits pertinents identifiés, le terminologue procède à une synthèse des informations sélectionnées pour chaque concept en les intégrant dans une phrase unique. Les définitions ainsi obtenues figurent dans les ressources (dictionnaires ou bases de données) terminographiques. Les définitions associées aux concepts du corpus analysé ont été rédigées suivant cette méthodologie, ce qui explique le décalage observé dans la section précédente.

Sur la base de cette définition, nous avons manuellement annoté le corpus de contextes conceptuels de façon à faire ressortir tous les traits relatifs aux 56 concepts des domaines considérés, qu'ils soient ou non susceptibles d'être utilisés pour définir. Dans cette section, nous faisons une analyse de la répartition des traits dans le corpus, afin de quantifier la proportion de traits susceptibles d'être utiles pour définir un concept. Un examen du corpus sous forme de traits devrait permettre d'identifier la problématique sous-tendant la rédaction de définitions.

Le corpus compte un total de 380 types de traits (ou 499 occurrences de traits si l'on compte les traits qui apparaissent deux fois ou plus). Dans la plupart des analyses, le décompte final du nombre de traits ne tient pas compte des répétitions, car méthodologiquement parlant, ce sont les *types* d'information susceptibles d'entrer dans une définition qui comptent et non leurs différents modes d'expression.

On constate, là encore, un décalage entre la moyenne et le mode, respectivement de 8,9 et 3 traits par concept et de 3,9 et 1 trait(s) par contexte. Concernant les occurrences de traits par concept et par contexte, on constate un grand nombre de configurations distinctes avec des fréquence allant de 1 à 54 traits par concept et de 1 à 25 traits par contexte, où les fréquences les plus élevées sont de 3 à 7 traits par concept (51,8 %) et de 1 à 5 traits par contexte (27,5 %). On sait, par ailleurs, à travers des études de corpus de définitions terminographiques (Seppälä, 2005), que le nombre de traits par définition le plus fréquent (le mode) est de 3 (40,2 % de 500 cas), générique et spécifiques confondus. Ces chiffres montrent que le corpus recèle donc bien plus d'informations sous forme de traits qu'il n'est nécessaire pour définir un concept. Ce constat amène une première conclusion : tous les traits figurant dans le corpus ne semblent pas forcément utiles pour définir un concept. Il y a donc lieu d'analyser les contextes plus en détail pour en extraire tous les traits (les informations) relatifs à chaque concept et voir dans quelle mesure ils sont susceptibles d'être inclus dans une définition. Se pose donc la question de la *pertinence* des traits.

8. Par exemple, les classes `ARTEFACT`, `ENTITE ABSTRAITE`, `etc.`, ou les relations `FONCTION`, `PARTIE`, `CAUSE`, `CONSEQUENCE`, `etc.`. Il est à noter que ce corpus-ci n'est pas annoté avec ces classes et ces relations.

5 Classification des traits en fonction de leur pertinence

Afin de voir si tous les types de traits dans le corpus sont susceptibles d'être inclus dans une définition, nous faisons appel à la notion intuitive de pertinence des traits [9]. Nous proposons donc un cadre descriptif permettant de catégoriser les traits figurant dans le corpus selon qu'ils sont susceptibles ou non de figurer dans une entrée (fiche) terminologique (*traits latents vs traits saillants*), et selon qu'ils sont potentiellement pertinentes ou non pour définir le concept considéré (*traits saillants vs traits potentiellement pertinents*). Les niveaux considérés sont les suivants :

Trait latent Un trait latent (L) correspond à une information qui n'est en principe pas exprimée dans une entrée terminologique [10], par exemple, le trait « [...] ils ressemblent aux mouvements d'un liquide dans une marmite placée sur une source chaude. » associé au concept *courant de convection* dans le domaine des *catastrophes écologiques*.

Trait saillant Un trait saillant (S) correspond à une information qui est en principe exprimée dans tout autre champ d'une entrée terminologique que celui de la définition, par exemple, le trait « [...] représenter 40 % du volume des ordures et 15 % de leur poids. » associé au concept *emballage* dans le domaine du *traitement des déchets*.

Trait potentiellement pertinent Un trait potentiellement pertinent (PP) correspond à une information qui est susceptible d'être exprimée dans le champ définition d'une entrée terminologique, par exemple, le trait « [...] établis pour faire face à des risques naturels ou technologiques ne faisant pas l'objet de plans particuliers. » associé au concept *plan de secours spécialisé* dans le domaine des *catastrophes écologiques*.

Les traits figurant dans le corpus de contextes conceptuels ont été (section 5.1) annotés à l'aide de ces descripteurs, ainsi qu'avec des marques indiquant leur éventuel rôle dans la définition (le type d'élément définitoire : *générique*, *spécifique* ou *extension*). Ces annotations ont permis deux types d'analyses : (i) déterminer la proportion de traits PP, S et L dans le corpus, et (ii) voir la répartition des traits selon l'éventuel type d'information définitoire auquel ils pourraient correspondre. Par ailleurs, les définitions proposées pour chaque concept ont été (section 5.2) décomposées en traits, que nous appelons *pertinents* (P) [11]. Chaque occurrence de trait pertinent a ensuite été recherchée dans le corpus afin de voir son niveau de pertinence (PP, S ou L) originel. Cette analyse permet (i) de déterminer la proportion de traits en corpus finalement pertinente pour définir, et (ii) de voir si tous les traits P correspondent à des traits PP en corpus.

5.1 Niveaux de pertinence des traits dans le corpus

Dans la suite de l'analyse, nous ne considérons que les *types de traits* de façon à avoir une idée plus précise de la répartition effective des traits selon (i) leur degré de pertinence (L, S et PP), (ii) le type d'élément auquel un trait pourrait correspondre dans une définition (gen, spe ou ext), et (iii) leur nature L, S ou PP lorsqu'ils sont pertinents (P), c'est-à-dire lorsqu'ils figurent dans la définition finale proposée pour chaque concept.

9. Le choix des traits inclus dans la définition est, en effet, bien souvent fait intuitivement, sur la base des informations contenues dans les textes, des recommandations d'experts et des connaissances d'arrière-plan du terminologue.

10. Même si cette information peut être utile dans un dictionnaire à visée vulgarisatrice, par exemple, elle n'est pas significative pour l'expert dans la connaissance de son domaine.

11. Nous choisissons de les considérer comme pertinents, dans la mesure où ils ont été délibérément sélectionnés pour définir les concepts considérés.

L'analyse de la répartition des types de traits en fonction de leur degré de pertinence montre que plus de la moitié (240/380, 64 %) sont PP, que 125/380 (33 %) sont S et que la part des types de traits L est très petite (13/380, 3 %) [12]. Ce résultat implique, à nouveau, que les informations contenues dans le corpus ne sont pas toutes forcément susceptibles d'être utiles pour définir, puisque le corpus comporte des traits qui ne sont *a priori* pas pertinents pour définir. Se pose alors la problématique de la sélection des traits.

S'agissant de la répartition des types de traits selon les éléments définitoires auxquels ils sont susceptibles de correspondre dans une définition, on constate que la plupart correspondent soit à des spécifiques (44,2 %) soit à aucun élément définitoire (35,3 %) puisqu'il s'agit de traits L ou S. Une infime proportion de types de traits (3,4 %) correspond à des traits extensionnels ; de ce fait, ils pourraient être compris dans une définition de ce type ou dans une définition mixte mêlant traits spécifiques et extensionnels. Il est également à noter que les contextes comportent davantage de génériques (17,1 %) qu'il n'en faudrait proportionnellement au nombre de concepts considérés : 65 génériques potentiels pour 56 concepts. Cette disproportion implique à nouveau que, dans plusieurs cas, un choix doit être fait entre différents génériques disponibles pour un même concept.

5.2 Niveaux de pertinence des traits dans les définitions

L'analyse de la répartition des traits pertinents montre que, sur les 380 types de traits du corpus, 180 sont P (47,4 %) [13], c'est-à-dire qu'ils ont été sélectionnés pour définir un concept. Presque tous ces traits P sont issus de l'ensemble de traits PP (170/180, soit 94 %). Ceci semble tout à fait normal, puisque les traits PP sont justement ceux qui ont été considérés comme susceptibles d'entrer dans une définition. On notera cependant que la part des types de traits PP du corpus qui sont également P est de 70,2 % (170 PP=P sur 242 PP). Parmi les traits P, 10 sur 180 (6 %) correspondent à des traits qui ont été catégorisés comme S. Si l'on observe ces traits-là de plus près, on constate qu'il s'agit pour la plupart de traits extensionnels (6/10 traits). Les autres correspondent à des traits S sélectionnés comme pertinents pour différentes raisons que nous n'avons pas la place de détailler ici. Il est néanmoins à noter que ces derniers auraient tout aussi bien pu être mis dans un champ *note* séparé. Par conséquent, si l'on fait abstraction de ces traits S=P, on peut considérer qu'une définition ne comporte que des traits PP, ce qui confirme la nécessité de distinguer entre traits PP et traits S et L, en vue de la sélection des premiers.

Cette analyse des traits, en corpus et dans les définitions, en fonction de leur degré de pertinence permet de mettre en lumière une problématique générale de la rédaction de définitions : *la sélection des traits*. Elle permet, en particulier, d'identifier trois questions précises à résoudre pour toute (semi-)automatisation de cette tâche, et donc pour le traitement automatique des définitions, mais aussi plus généralement dans une théorie des définitions. La première question, issue des constats qu'un tiers des types traits dans le corpus sont S ou L, et que les traits P de la définition sont PP en corpus, est de savoir comment distinguer les traits PP des traits S et L. La deuxième question est de savoir comment sélectionner les traits P parmi les traits PP. Elle émerge des constats que les traits P correspondent à quelques exceptions près aux traits PP dans

12. Ces résultats sont néanmoins à nuancer, car il se peut que la part des traits S et L soit, ici, plus réduite que ce que l'on peut trouver généralement dans les corpus de dépouillement, y compris ceux dont ont été extraits les contextes analysés. De fait, les données dont nous disposions étaient destinées à des cours sur la méthodologie de rédaction de définitions ; il se peut dès lors que certains contextes qui ne comportaient que des traits S ou L n'aient pas été retenus par les enseignants. Il est cependant intéressant de noter que, même en tenant compte de ce biais, les contextes comportent une proportion importante de traits S. Cet éventuel biais ne nuit donc en rien à notre propos, mais il y aurait néanmoins lieu d'effectuer le même type d'analyses sur d'autres données pour voir si l'on obtient des résultats comparables.

13. Rapporté au nombre total d'occurrences de traits, le nombre de traits P est de 283 sur 499, soit 56,7 % des traits.

le corpus, et qu'environ 30 % des types de traits PP ne sont pas effectivement P. La troisième question, qui découle du fait que le corpus comporte plus de génériques qu'il n'y a de concepts, est de savoir comment sélectionner le générique adéquat.

Si le terminologue peut souvent faire confiance à ses intuitions, à son expérience et à l'aide d'experts du domaine pour catégoriser (implicitement) les traits en fonction du niveau de pertinence et faire ces différents choix [14], ce n'est pas le cas d'un programme informatique. L'adoption de ce cadre descriptif inspiré de la pratique terminographique a donc des conséquences pour le traitement automatique, puisqu'il exige que la notion de niveau de pertinence soit définie plus formellement. L'absence de ce type de critères a notamment des répercussions sur l'évaluation des systèmes d'extraction, comme le souligne Meyer (2001, 298). Il y aurait, par conséquent, lieu de déterminer des critères de pertinence des traits qui permettent d'automatiser, au moins partiellement, la sélection de traits à inclure dans une définition.

6 Examen d'un critère de sélection empirique

Dans cette section, nous nous intéressons plus particulièrement à la première et à la deuxième question, à savoir comment distinguer les traits PP des traits S et L, et comment sélectionner les traits P parmi les traits PP. L'une des solutions proposée par Meyer (2001) est de cibler l'extraction non pas sur les définitions à proprement parler ou les contextes en général, mais sur les traits en fonction du type de relation qu'ils expriment. Cette méthode, qui permet de disposer des traits pouvant être combinés pour former une définition, implique à son tour deux difficultés : (i) mettre au point des algorithmes d'extraction capables de catégoriser les informations en fonction du type de relation conceptuelle exprimé ; et, plus fondamentalement, (ii) savoir, ensuite, quels sont les types de relations pertinents pour définir [15]. Cette dernière question se pose d'ailleurs aussi au terminologue qui compose la définition manuellement. Une solution à ce problème serait d'étudier la nature conceptuelle (en termes de classes et de relations) des traits composant des définitions existantes, afin de voir si elles présentent des régularités, par exemple, en fonction du type de classe conceptuelle définie ou du domaine [16]. Il est à noter que ce type d'entreprise pose, lui aussi, des défis en TALN, notamment en ce qui concerne la segmentation des définitions en traits et leur annotation conceptuelle. Une autre solution, mais qui ne permettrait que de distinguer entre traits S et PP (ce qui ne répond pas à la question des traits pertinents), serait de rechercher des marqueurs linguistiques caractéristiques des traits S, notamment la présence de noms propres, de valeurs numériques, ou d'autres types d'informations non généralisantes (voir notamment les propositions de Pearson (1998) et de Meyer (2001), tels que les modaux). Ce type de solution implique, là encore, la mise au point d'algorithmes tels que ceux qui sont utilisés pour l'extraction d'entités nommées, par exemple. Les deux types de solutions proposées impliquent, en outre, un minimum de théorie des définitions, si ce n'est des concepts, qui rende compte de la nature des concepts et de leurs relations, ou de la nature discursive des propositions pouvant être énoncées à propos des objets et de leur statut définitoire.

Pour tenter de pallier les problèmes évoqués ci-dessus, nous proposons d'examiner une hypothèse de pertinence des traits purement empirique : *la répétition d'un trait en corpus est un signe de pertinence*. Il s'agit, ici, de repérer et d'analyser dans le corpus les traits qui figurent deux fois ou plus dans un ou

14. Même s'il serait souhaitable qu'il dispose d'autres critères pour ce faire, et donc d'une théorie de la définition qui rende compte de ces critères.

15. Au sujet des relations, voir également la discussion de Meyer (2001, 297–298), sections 4.2 et 4.3.

16. Des travaux en ce sens ont été initiés dans Seppälä (2005) et font actuellement l'objet de notre thèse de doctorat.

plusieurs contexte(s) relatif(s) à un même concept. Là encore, il y a lieu de distinguer les occurrences des types de traits qui se répètent (respectivement 185/499 traits et 66/380 types). La moyenne de répétitions d'un trait est de 2,76 traits/type, le nombre d'occurrences le plus fréquent (le mode) étant de 2 traits/type et concerne 61 % des types. Ensemble, les types se répétant deux et trois fois représentent 84 % du total des types de traits. Pour 48 types sur 66 (73 %), la répétition a lieu dans des contextes distincts (*contextes multiples*) ; seul 18 types sur 66 (27 %) ont leurs différentes occurrences à l'intérieur d'un même contexte (*contexte unique*).

En ce qui concerne l'analyse du niveau de pertinence des types de traits à occurrences multiples, on constate que la très grande majorité de ces types de traits (61/66, soit 92 %) sont des traits PP ; il n'y a aucun trait L et seulement 8 % de traits S (5/66 types de traits). On constate, par ailleurs, que 54/66 types (82 %) sont aussi pertinents (P) [17], ce qui tend à confirmer notre hypothèse sur la pertinence des traits qui se répètent. Parmi ces derniers, on ne trouve qu'un seul trait S ; les 53 restants sont tous des traits PP. La répétition pourrait donc être un bon indice pour discriminer les traits PP des traits S et L. En outre, seul 3 % des types de traits à occurrences multiples (soit 2/66 types) ne correspondent à aucun élément de la structure définitoire (contre 35,5 % pour l'ensemble des types de traits du corpus). Ces résultats tendent à montrer que les types de traits qui se répètent ont potentiellement leur place dans une définition. En revanche, la proportion de génériques (27 %, contre 17,1 %), de spécifiques (65 %, contre 44,2 %) et de traits extensionnels (5 %, contre 3,4 %) est similaire.

Ramenés à l'ensemble des traits du corpus, les types de traits à occurrences multiples représentent 17,4 % des types de traits du corpus (66/380), soit 37,1 % des occurrences de traits (185/499). Ils représentent, par ailleurs, un quart des traits PP et seulement 4 % des traits S. En termes d'éléments de la structure définitoire, ils représentent plus ou moins un quart des génériques, des spécifiques et des extensions, mais seulement 1,5 % des traits qui ne sont attribuables à aucun élément définitoire (soit 2/134). Le résultat le plus intéressant reste, cependant, que les types de traits à occurrences multiples qui sont effectivement P ne représentent que 30 % de l'ensemble des traits P figurant dans les définitions [18], ce qui oblige à relativiser les résultats considérés uniquement par rapport aux types de traits à occurrences multiples dans le corpus. Ce chiffre tend également à infirmer l'hypothèse testée.

Ainsi, la répétition des traits semble, à première vue, être un bon indice empirique pour décider de leur pertinence, et ce dès deux occurrences du même trait (configuration de loin la plus fréquente, qui concerne 61 % des types de traits qui se répètent [19]). Cette conclusion est néanmoins à nuancer au vu d'un certain nombre de limites susceptibles de remettre en cause la fiabilité de cette méthode pour déterminer la pertinence d'un trait, et surtout son implémentation : les types de traits à occurrences multiples pertinents ne représentent que 30 % de l'ensemble des traits pertinents ; la plupart du temps (73 %), les différentes occurrences d'un même type de trait figurent dans des contextes distincts, or l'analyse générale du corpus de contextes (section 3) révèle que la moitié des concepts ne fait l'objet que d'un seul contexte, il peut dès lors s'avérer difficile d'utiliser cette méthode de façon systématique, notamment si l'on souhaite l'automatiser ; la même information peut être formulée de façons si différentes d'une occurrence à l'autre qu'il peut parfois être difficile de dire s'il s'agit vraiment du même type de trait — là encore, il peut s'avérer difficile d'utiliser cette méthode de façon systématique — ; l'information répétée peut être implicite et nécessiter une intervention humaine, voire celle d'un expert, pour déterminer s'il s'agit bien de la même information ou non ; la méthode, purement empirique, manque de pouvoir explicatif.

17. Ramené au nombre d'occurrences de traits qui se répètent, les traits P représentent 85 % du total, soit 157/185 traits.
18. Cela implique, à l'inverse, que 70 % des traits P ne se répètent pas dans le corpus.
19. À savoir 40 types de traits P se répétant deux fois sur un total de 66 types de traits à occurrences multiples.

7 Conclusion

Dans cet article, nous avons présenté une analyse manuelle de corpus de contextes conceptuels afin (i) de voir dans quelle mesure les méthodes de TALN existantes sont *en principe* adéquates pour automatiser la rédaction de définitions, et (ii) de dégager des question précises dont la résolution permettrait d'automatiser davantage la production de définitions. Grâce à une première analyse de contextes conceptuels, nous avons montré que la part des définitions « canoniques » dans les corpus est relativement réduite, et donc que la seule extraction automatique de définitions est en principe insuffisante. Ce constat nous a conduit à prendre en compte la méthodologie terminographique de rédaction de définitions, afin d'examiner le corpus suivant un cadre descriptif correspondant : une analyse sous forme de traits. À l'issue de cette analyse, nous avons montré que toutes les informations contenues dans un corpus ne sont pas forcément utiles pour définir un concept. Pour voir dans quelle mesure les différents traits figurant dans le corpus sont susceptibles d'entrer dans une définition, nous avons posé un cadre descriptif permettant d'annoter les traits en fonction de leur degré de pertinence. Nous avons ensuite procédé à l'analyse des données ainsi annotées pour les comparer avec les traits pertinents composant les définitions existantes de chaque concept. Les résultats nous ont permis de dégager de façon empirique trois questions précises liées à la problématique centrale de la sélection des traits. Une fois les questions identifiées, nous avons considéré la question de leur résolution, c'est-à-dire de l'identification de critères de sélection des traits pertinents. Nous avons proposé quelques pistes et évoqué leurs implications pour le TALN, puis examiné une hypothèse purement empirique : *la répétition d'un trait dans le corpus est un signe de pertinence*. Cette dernière étude tend à montrer que l'automatisation de la sélection des traits sur la seule base de la fréquence est limitée. La difficulté pourrait en partie être due au fait qu'il s'agit de questions liées au contenu conceptuel ou sémantique des traits, donc difficilement abordables sans théorie. Les résultats de cette analyse nous permettent de conclure à l'intérêt qu'il y aurait à considérer des solutions théoriques à la problématique de la sélection des traits, pour ensuite étudier la possibilité de les implémenter de façon (semi-)automatique.

Références

CARTIER E. (2004). *Repérage automatique des expressions définitoires*. PhD thesis.

MALAISÉ V., ZWEIGENBAUM P. & 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. & MEYER I. (2002). French patterns for expressing concept relations. *Terminology*, **8**(1), 1–29.

MEYER I. (2001). Extracting knowledge-rich contexts for terminography. In *Recent Advances in Computational Terminology*. John Benjamins.

PEARSON J. (1998). *Terms in Context*. Studies in corpus linguistics 1.

REBEYROLLE J. (2000). Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes. *Actes IC'2000*, p. 105–114.

REBEYROLLE J. & TANGUY L. (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de grammaire*, **25**, 153–174.

SEPPÄLÄ S. (2005). Structure des définitions terminographiques : une étude préliminaire. In *Actes TIA'05*, p. 19–29.