Identification de Contextes Riches en Connaissances en corpus comparable

Firas Hmida

▶ **To cite this version:**

Firas Hmida. Identification de Contextes Riches en Connaissances en corpus comparable. RECITAL, Jul 2014, Marseille, France. hal-01142531

HAL Id: hal-01142531
https://hal.science/hal-01142531

Submitted on 15 Apr 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Identification de Contextes Riches en Connaissances en corpus comparable

Firas Hmida
LINA UMR 6241,Université de Nantes
firas.hmida@univ-nantes.fr

Résumé. Dans les études s'intéressant à la traduction assistée par ordinateur (TAO), l'un des objectifs consiste à repérer les passages qui focalisent l'attention des traducteurs humains. Ces passages sont considérés comme étant des contextes « riches en connaissances », car ils aident à la traduction. Certains contextes textuels ne donnent qu'une simple attestation d'un terme recherché, même s'ils le renferment. Ils ne fournissent pas d'informations qui permettent de saisir sa signification. D'autres, en revanche, contiennent des fragments de définitions, mentionnent une variante terminologique ou utilisent d'autres notions facilitant la compréhension du terme. Ce travail s'intéresse aux « contextes définitoires » qui sont utiles à l'acquisition de connaissances à partir de textes, en particulier dans la perspective de traduction terminologique assistée par ordinateur à partir de corpus comparables. En effet, l'association d'un exemple à un terme permet d'en appréhender le sens exact. Nous proposons, tout d'abord, trois hypothèses définissant la notion d'exemple définitoire. Ensuite nous évaluons sa validité grâce une méthode s'appuyant sur les Contextes Riches en Connaissances (CRC) ainsi que les relations hiérarchiques reliant les termes entre eux.

Abstract. Some contexts provide only a simple explanation of a given term even if they contain it. However, others contain fragments of definitions, mention a terminological variant or use other concepts to make it easy the term understanding. In this work we focus on « definitory contexts » that would be valuable to a human for knowledge acquisition from texts, mainly in order to assist in terminological translation from comparable corpora. Indeed, provide the term with an example, makes it possible to understand its exact meaning. First, we specify three hypothesis defining the concept of a defnitory example. Then we evaluate its validity through a method based on the knowledge-Rich Contexts (KRCs) and hierarchical relationships between terms.

Mots-clés : Contextes Riches en Connaissances, CRC, identification de définitions, identification d'exemples, énoncé définitoire, terminologie, traduction terminologique.

Keywords: Knowledge-Rich Contexts, KRCs, mining definitions, minging examples, terminology, terminological translation.

1 Introduction

Ces dernières années, de nombreux travaux se sont tournés vers l'exploitation des corpus comparables. Ces corpus sont définis par Bowker & Pearson (2002) comme étant : des corpus contenant des documents multilingues qui ne sont pas des traductions mais qui partagent certaines caractéristiques telles que la période et le thème. Les principaux travaux en extraction de terminologies bilingues à partir de ces corpus se basent sur l'hypothèse qu'un mot et sa traduction apparaissent souvent dans les mêmes environnements lexicaux (Firth, 1957). Les approches standard (Fung & McKeown, 1997; Rapp, 1999) et par similarité inter-langue (Déjean & Gaussier, 2002; Daille & Morin, 2005), dédiés à l'extraction de lexiques bilingues à partir de corpus comparables, reposent plus particulièrement sur ce principe. En effet, elles permettent, à partir d'un terme à traduire (dans une langue source) d'obtenir une liste ordonnée de traductions candidates (dans une langue cible). Ces traductions sont souvent obtenues en comparant le contexte traduit, en langue cible, du terme source avec l'ensemble des contextes des termes de la langue cible. Les traductions candidates se présentent sous la forme d'une liste plate qui ne fournit pas d'information contextuelle structurée permettant de saisir le contexte d'utilisation du terme visé. Par exemple la méthode standard propose *axillary dissection*, *axillary node dissection* où *dissection* comme traductions candidates (en Anglais) correspondant au terme *curage axillaire* (en Français).
En termes de performance, ces approches de traduction semblent avoir atteint leurs limites, et les recherches les plus récentes se focalisent plutôt sur l'évaluation de ces approches (Laroche & Langlais, 2010). Si l'accès à la terminologie bilingue s'avère indispensable au processus de traduction, le potentiel des méthodes de traduction adoptées doit être amélioré au moyen d'une contextualisation pertinente des termes. En effet, il faut être capable d'appréhender le sens exact

d'un terme et de l'employer correctement. Par exemple, le terme *clavier* désigne dans le domaine musical un instrument de musique, tandis qu'il correspond en informatique à un périphérique d'entrée.
Nous étudions, ici, la possibilité d'associer à un terme à traduire dans une langue source (ou à un terme traduit dans une langue cible) un contexte permettant d'en appréhender le sens exact, dans le but d'aider à sa traduction. Pour cela, nous nous appuyons sur les Contexte Riche en Connaissances (CRC) qui sont introduits par Meyer (2001), dans le but d'améliorer la traduction terminologique assistée par ordinateur.
Ce travail s'inscrit dans le cadre du projet CRISTAL [1] ayant pour objectif de développer une technologie d'extraction de CRC permettant de produire de nouveaux dictionnaires. Ces derniers sont censés pouvoir lister pour chaque terme (ou sa traduction), les CRC et illustrer les connaissances qu'ils contiennent. Par conséquent, ce type de contextes auront pour effet d'attirer l'attention de l'utilisateur (éventuellement non expert en terminologie) sur des phénomènes linguistiques qu'il ne soupçonne pas, et de réduire le temps d'accès à l'information pertinente.
Après un rappel de la terminologie utilisée et de l'état de l'art en identification de contextes riches en connaissances (section 2 et 3), nous présentons notre contribution sur le repérage et l'exploitation de ces énoncés (section 4). Ensuite, nous discutons les résultats obtenus pour conclure dans la (section 5) sur des perspectives à ce travail.

2 Définitions

2.1 Contexte Riche en Connaissances (CRC)

Meyer (2001) fut la première à proposer l'appellation Contextes Riches en Connaissances (CRC), pour désigner les contextes qui permettent de repérer, grâce à des éléments lexico-syntaxiques, des relations (souvent lexicales ou lexico-syntaxiques) entre plusieurs termes. Il s'agit de portions de textes qui contiennent i) des termes d'un domaine spécialisé et ii) des marqueurs explicitant des relations entre ces termes.
Par exemple, la phrase *Les graisses dans le sang sont essentiellement le cholestérol et les triglycérides* est définie comme un contexte riche en connaissances pour le terme *graisse dans le sang*. Dans cette phrase, la présence du marqueur *être_essentiellement* explicite une relation hiérarchique entre les termes *graisse dans le sang*, *cholestérol* et *triglycéride*.
Par ailleurs, les CRC sont intéressants pour l'acquisition de relations sémantiques entre les termes et pour construire des définitions. Dans l'exemple précédent, il existe une relation de définition (par dénotation) entre *graisse dans le sang*, *cholestérol* et *triglycéride*.

2.2 Marqueur de relation

Dans la littérature, les CRC sont souvent identifiés grâce à des marqueurs de relations. Ces derniers servent à repérer et classer finement les relations terminologiques dans un corpus spécialisé. Il s'agit d'un ensemble de mots, d'expressions ou de symboles révélant de façon récurrente une relation terminologique. Par exemple, la structure *telle que* est un marqueur de relation qui exprime une relation d'hyperonymie pouvant relier deux termes comme dans la phrase qui suit : *une hormone telle que l'insuline*... où *hormone* et *insuline* sont deux termes.

2.3 Patron de connaissances (PC)

Les marqueurs de relations sont habituellement utilisés afin d'identifier les CRC. Ils sont le plus souvent modélisés et mis en œuvre grâce aux patrons de connaissances. Ces derniers sont l'une des principales stratégies utilisées dans le but d'isoler les CRC, et ainsi d'écarter les contextes jugés moins utiles. Meyer (2001) et Pearson (1998) ont montré l'intérêt des marqueurs linguistiques indiquant des relations sémantiques entre des termes pour exploiter, dans un but lexicographique, les corpus spécialisés. Cette démarche permet à un terminologue de pointer vers le sous-ensemble des phrases susceptibles de véhiculer les informations souhaitées (Barrière, 2004).
Un patron de connaissances est une expression régulière, formée de mots, de catégories grammaticales ou sémantiques et de symboles, visant à identifier des fragments de texte explicitant des formes lexicales et des catégories grammaticales. Par exemple, dans la phrase *X est un type de Y* (X et Y étant deux termes différents), la structure *est un type de* est un patron de connaissances modélisant le marqueur *être_un_type_de*. Nous appelons PC définitoire, un PC s'articulant autour d'un marqueur de définition.

1. www.projet-cristal.org

Synthèse : Les contextes riches en connaissances sont des contextes contenant un patron de connaissances associé à un marqueur de relation. Reprenons l'exemple précédent :

CRC : *les graisses dans le sang sont essentiellement le cholestérol et les triglycérides*.

PC : *X sont essentiellement Y et Z* ; X, Y et Z sont trois termes.

Marqueur de relation : *être_essentiellement*.

2.4 Terme secondaire

Selon Saggion (2004), un terme secondaire [2] peut être un nom, un verbe ou un adjectif qui cooccurre souvent avec un terme donné dans des définitions provenant de ressources externes, comme le Web par exemple. Saggion (2004) considère les termes secondaires comme un indice pour identifier les définitions. L'intuition derrière l'apparition de cette notion revient à l'observation suivante : cherchant les définitions du mot *Goth* parmi 217 phrases contenant toutes ce mot, Saggion (2004) a remarqué que le mot *subculture* apparaît régulièrement dans les définitions du mot *Goth* du Web. En examinant les 217 phrases de départ, il s'est avéré que seulement 5 d'entre elles étaient des définitions contenant toutes le mot *subculture*. Citons à titre d'exemple la phrase *The goth is a contemporary subculture found in many countries*. À travers cette observation, nous pouvons noter qu'un terme à traduire et son terme secondaire apparaissent souvent dans les définitions et rarement dans les autres contextes (non définitoires).

3 État de l'art

De nombreux travaux se sont intéressés à l'identification automatique de définitions dans différents domaines : terminologie (Gangemi *et al.*, 2003; Velardi *et al.*, 2013), lexicographie (Saggion, 2004), etc. Dans ce travail, nous abordons plutôt l'identification des définitions dans la perspective de l'aide à la traduction terminologique. Nous étudierons, tout d'abord, les approches à base de PC, puis les approches supervisées et semi-supervisées. Dans ces travaux les limites séparant les définitions des CRC n'étaient pas bien déterminées puisqu'ils ont souvent été exploités dans le but d'identifier les définitions.

3.1 Approches à base de PC

Les méthodes basées sur les patrons de connaissances ont été adoptées dans plusieurs travaux de la littérature. Auger (1997), par exemple, a consacré son travail à repérer des définitions avec des PC lexicaux. Quant à Rebeyrolle (2000), elle a utilisé des PC lexico-syntaxiques et a également tenu compte des contraintes liées à la ponctuation. Rebeyrolle (2000) a évalué sa méthode en considérant un corpus étiqueté manuellement par les définitions, comme référence. Elle a obtenu une précision comprise entre 17,95 % et 79,19 % et un rappel entre 94,75 % et 100 % selon les PC utilisés afin de repérer les définitions. Muresan & Klavans (2002) ont proposé leur outil DEFINDER également basé sur des PC (ex. *is defined as, is called*) et des marqueurs de relation comme *()* et - -. Cet outil permet tout d'abord de sélectionner des définitions candidates à partir d'articles médicaux disponibles sur le Web. Ensuite, les définitions complexes sont filtrées par un analyseur grammatical et les sorties de cet outil (DEFINDER) sont comparées à un ensemble de textes étalon. L'évaluation a donné 86,95 % de précision et 75,47 % de rappel.
Malaisé *et al.* (2004) se sont appuyés sur les travaux de Auger (1997) et de Rebeyrolle (2000) pour définir une liste de marqueurs adaptés à leurs corpus. Ils ont davantage pris en considération des marqueurs liés à la ponctuation. Malaisé *et al.* (2004) ont tenté de repérer les définitions pour en extraire ultérieurement les relations entre les termes afin d'aider à la construction d'ontologie. L'évaluation de ce point a soulevé le problème d'avoir une précision faible quand il s'agit des marqueurs linguistiques de reformulation plutôt que des marqueurs lexicaux métalinguistiques [3]. Cette remarque a également été mentionnée par Rebeyrolle (2000).

Saggion (2004) a proposé un système reposant sur l'utilisation des PC et les termes secondaires afin d'identifier les passages définitoires pour ensuite extraire des définitions. Pour cela, il disposait en amont d'une liste de 69 PC. Il a présupposé

2. « *Terms that co-occur with the definiendum (outside the target collection) in definition-bearing passages seem to play an important role for the identification of definitions in the target collection [..] Our methode considers nouns, verbs and adjective as candidate secondary terms.* »(Saggion, 2004)

3. renferment un lexique métalinguistique

que cette liste pourrait servir à identifier les passages et sélectionner les définitions indépendamment des corpus. Parmi ces 69 PC, 36 sont destinés aux questions générales (Qu'est ce que X ? ; X est à définir), et 33 pour traiter des questions spécifiques (Qui est X ? ; X est à définir). Saggion (2004) a sollicité WordNet, Britannica et le Web (ressources externes) afin de déterminer les termes secondaires. Dans WordNet seulement sont considérés les hyperonymes du mot X (à définir) ou les mots les plus fréquents dans son contexte. Tandis que dans Britannica, les termes secondaires ne sont extraits que si les phrases contenant une référence explicite de X. Dans le cas des mots venant d'autres pages Web, la phrase doit contenir un PC pour tenir compte des termes secondaires qu'elle contient. Pour identifier les passages contenant des définitions, Saggion (2004) introduit ses requêtes enrichies par les termes secondaires comme des entrées. Ensuite une phrase est retenue comme une définition si elle contient soit un PC, soit le mot à définir avec au moins trois termes secondaires. Lors de TREC 2003, le système de Saggion (2004) a obtenu un score de 0,236 (meilleur 0,555, moyen 0,192) en termes de F-mesure. Les définitions ont été évaluées par rapport à des définitions de référence.
Barrière (2004) considère les PC comme un outil "clé" permettant de repérer les CRC. Elle a organisé les PC présents dans les définitions du dictionnaire numérique American Heritage First Dictionary (AHFD) selon leurs types et la relation sémantique qu'elles expriment. Les PC sont classés en trois grandes catégories : **statiques** donnant des contextes qui ne sont pas liés à des événements, **dynamiques** contenant les relations causales et temporelles, et **événementielles** introduisant des événements (intrinsèques/extrinsèques). Ensuite, elle a analysé la généralité/spécificité des PC par rapport aux domaines et aux relations sémantiques exprimées par ces PC dans le domaine de la plongée sous-marine (1 million de mots). Barrière (2004) a remarqué que les relations sémantiques d'hyperonymie et méronymie, par exemple, sont exprimées de la même façon dans le corpus de la plongée sous-marine que dans le dictionnaire AHFD. Cependant, il existe d'autres relations telles que *risk prevention* qui sont exprimées avec de nouveaux PC. Celles-ci sont considérées comme spécifiques au domaine de la plongée sous-marine : les PC explicitant ces relations n'apparaissent que dans le corpus du domaine étudié. Afin de repérer les énoncés définitoires, les travaux présentés dans cette section se sont basés sur des marqueurs, habituellement lexico-syntaxiques, signalant des CRC. Si Barrière (2004) et Rebeyrolle (2000) ont étudié les structures linguistiques exprimant la définition dans les textes, d'autres travaux comme celui de Saggion (2004) ont choisi d'exploiter les termes secondaires comme étant un indice de définition. Cette notion de termes secondaires étant similaire à celle de la collocation (limitée aux énoncés définitoires), elle permet d'identifier le contexte définitoire « typique ».

3.2 Approches supervisées

Les méthodes utilisant les patrons de connaissances ont mis en évidence plusieurs difficultés. En effet, les définitions peuvent être exprimées de manières différentes. Ceci rend difficile l'obtention d'un ensemble de PC permettant d'identifier toutes les définitions. C'est pour cette raison que plusieurs recherches se sont orientées vers des méthodes moins dépendantes des PC. Fahmi & Bouma (2006), par exemple, ont proposé une méthode reposant sur l'apprentissage supervisé afin de repérer les définitions dans un corpus allemand du domaine médical. Ce corpus est constitué des pages de Wikipedia. Ils ont commencé par extraire toutes les phrases contenant le marqueur *to be* afin d'obtenir des définitions candidates. Parmi ces phrases, les définitions sont isolées manuellement. Ensuite, Fahmi & Bouma (2006) ont déduit les traits permettant de distinguer les définitions des autres phrases. Il s'agit de la position de la phrase, la distribution des mots et des bigrammes, ainsi que des traits syntaxiques comme le type du déterminant et la position du sujet dans la phrase. Ils ont alors intégré ces traits dans trois systèmes d'apprentissage différents : Bayésien naïf, SVM (Support Vector Machine) et MaxEnt (Maximum Entropy). Les résultats obtenus varient de 77 % à 92.3 % (le meilleur fourni par MaxEnt) en termes de précision. Quelques années plus tard, Westerhout (2009), inspirée par Fahmi & Bouma (2006), a proposé une méthode hybride dans laquelle elle a eu recours à l'apprentissage supervisé et aux patrons de connaissances. Ainsi, elle a exploité à la fois des informations linguistiques (telles que la l'étiquette grammaticale) et des informations structurelles. L'auteur a ajouté le type des noms et la structure du document aux traits de son système. Les meilleurs résultats (F-mesure=0.63) proviennent du patron *is a*.

3.3 Approches semi-supervisées

Peu de travaux, comme celui de Kilgarriff *et al.* (2008), se sont penchés sur l'identification des exemples. Kilgarriff *et al.* (2008) présente GDEX, un outil qui propose aux lexicographes plusieurs exemples permettant de définir un mot donné. Il s'est référé à Atkins & Rundell (2008) pour qualifier un bon exemple comme étant : lisible et informatif. Afin de concrétiser ces critères, il a proposé des traits privilégiés (positifs) tels que la longueur de phrase, la présence de la collocation souhaitée dans la clause principale de la phrase. Il a également considéré la présence de pronoms et d'anaphores comme des traits pénalisant (négatifs). Pour ce faire, il a effectué des jeux de test basées sur des comparaisons avec son corpus

d'apprentissage considéré comme un corpus de référence. D'après les résultats, la longueur de la phrase et la fréquence des mots sont les traits qui influencent principalement le choix des exemples. Toujours dans le but d'aider les lexicographes, Didakowski *et al.* (2012) se sont également intéressés à l'extraction des exemples. Même si leur approche était similaire à celle de Kilgarriff *et al.* (2008), les traits dans Didakowski *et al.* (2012) ont été exploités afin d'associer des scores aux phrases.
Pour faire face au problème de portabilité des PC, Reiplinger *et al.* (2012) ont proposé une méthode semi-supervisée. Ils ont appliqué des couples de termes liés par des relations sémantiques afin d'extraire des définitions à partir des articles ACL Anthology Reference Corpus (ACL ARC) (Bird *et al.*, 2008). À partir d'un ensemble limité de paires de terme-définition et des PC définis auparavant, leur système a acquis de nouvelles paires terme-définition ainsi que de nouveaux PC. Les résultats obtenus montrent que cette technique peut être appliquée pour extraire des définitions. Quant à Navigli & Velardi (2010), ils considèrent comme définition toute phrase pouvant être associée à un automate également associé à une définition et généré en amont. Cet automate correspond à un patron de connaissances s'intéressant plutôt à la structure de la phrase. Navigli & Velardi (2010) disposent d'un corpus de définitions extraites de Wikipedia et étiquetées manuellement lui permettant d'en déduire des modèles de définitions sous forme d'automates (finis déterministes).
La plupart des méthodes abordent le repérage automatique des définitions avec des structures linguistiques définies auparavant (excepté Kilgarriff *et al.* (2008) et Didakowski *et al.* (2012)). On distingue principalement deux types d'approches : linguistiques et informatiques. Barrière (2004), par exemple, s'est donnée explicitement pour objectif la description de patrons de connaissances signalant des énoncés riches en informations sémantiques. D'autres méthodes ont essayé de trouver un compromis entre les deux, c'est-à-dire soit en traduisant la structure linguistique en modèle générique (comme Navigli & Velardi (2010)), soit en utilisant des méthodes informatiques en parallèle avec des méthodes linguistiques, telles que les PC.

Nous poursuivons notre travail sur le même principe qui est d'exploiter les PC et la présence d'autres termes comme un indice de définition. Nous proposons tout d'abord de profiter des relations hiérarchiques reliant les termes entre eux, puis ensuite de sélectionner les définitions à l'aide des marqueurs de définitions.

4 Détection de CRC

Rappelons que notre objectif consiste à exploiter les corpus comparables dans le but d'aider à la traduction terminologique. Pour cela nous proposons une méthode permettant d'illustrer un terme à traduire dans une langue source (ou un terme traduit dans une langue cible) pour aider à sa traduction. En effet, l'association d'un exemple à un terme permet d'en appréhender le sens exact.

4.1 Notion d'exemple définitoire (ED)

Nous postulons tout d'abord que le contexte d'apparition d'un terme est un exemple candidat. Cependant, les contextes d'apparition d'un terme, qui peuvent être nombreux, ne sont pas tous des exemples pertinents. Voici par exemple 3 contextes du terme *diabète de type 2* :

(a) *En ce qui concerne le risque que l'enfant ait lui-même un diabète de type 2 à la cinquantaine ou fasse un diabète gestationnel s'il s'agit d'une fille, on a pendant longtemps estimé qu'il dépendait essentiellement de la transmission par la mère d'un capital génétique favorisant le diabète de type 2 (même risque dans ce cas que si le père de l'enfant à naître a un diabète de type 2) mais des études récentes semblent en faveur d'un rôle possible du niveau de glycémie pendant la grossesse lorsque le diabète n'est pas maîtrisé.*

(b) *Dans le monde, 150 millions de personnes souffrent de diabète, dont 90 % de diabète de type 2.*

(c) *Le diabète de type 2 est un diabète qui s'accompagne souvent d'un excès de poids.*

Nous allons proposer des caractéristiques permettant de considérer un contexte comme un exemple pertinent.

4.1.1 Aspect définitoire

Hypothèse 1 : *"un exemple est illustré par une définition explicite"*

Le contexte doit donner des renseignements sur le terme visé afin d'en appréhender le sens exact. Lehmann & Martin-Berthet (1998) distinguent trois types de contexte :

1. Les contextes définitoires sont des fragments textuels servant à enrichir des définitions canoniques [4] du terme.
   Par exemple *le diabète gestationnel est un diabète qui se développe pendant la grossesse* est une définition canonique représentant un contexte définitoire pour le terme *diabète gestationnel*.
2. Les contextes encyclopédiques donnent une information complémentaire sur le terme.
   La phrase *le diabète gestationnel concerne entre 1 et 4 % des grossesses* ne fournit pas une définition mais donne plutôt une information supplémentaire sur le terme concerné.
3. Les contextes linguistiques caractérisent le comportement du terme dans le discours notamment par rapport aux collocations.
   Par exemple *une glycémie, mesurée à jeun, avec une valeur normale, ne permet pas d'exclure le diabète gestationnel,* permet d'associer *glycémie* au terme *diabète gestationnel* dans un discours spécialisé.

Debora Farji-Haguet [5] (traductrice chargée de cours en traduction technique et en terminologie à l'Université de Paris 7) affirme que les traducteurs s'intéresseront plus aux définitions qu'aux autres types de contexte afin de différencier deux termes du même domaine tels que *diabète de type 1* et *diabète de type 2*.
Dans ce travail, nous nous intéressons aux aspects définitoires que peuvent révéler les contextes. D'autre part, les énoncés définitoires peuvent être repérés automatiquement dans le texte grâce à des structures signalant des segments de définitions (Rebeyrolle & Tanguy, 2000). Par exemple, parmi les contextes (a), (b) et (c), seulement (c) correspond à une structure définitoire exprimée par le verbe *est* explicitant la définition. Notre objectif vise les contextes contenant des indices de définitions.

4.1.2 Niveau phrastique

Hypothèse 2 : *"un exemple est limité à une phrase"*

Plusieurs travaux traitant des contextes tels que Meyer (2001) et Martínez *et al.* (2009) ont choisi de travailler sur des unités textuelles plus petites que les phrases. Ces contextes sous-phrastiques risquent de ne pas donner assez d'informations sur le terme en question. En outre, dans les paragraphes le risque est plus élevé d'extraire des informations imprécises concernant ce terme. Afin d'éviter cet écueil nous avons fait le choix de travailler uniquement sur des phrases entières comme étant un compromis, et ainsi considérer le contexte (a) comme un contexte non-pertinent.

4.1.3 Présence d'hyperonyme

Hypothèse 3 : *"un exemple contient le terme à illustrer et son hyperonyme"*

Selon Lehmann & Martin-Berthet (1998), un terme est souvent défini par recours à son hyperonyme. En effet, définir un terme revient à déterminer d'abord sa classe générale, puis le spécifier par rapport aux autres termes appartenant à celle-ci. On rencontre deux types de définitions : la définition intensionnelle (non intentionnelle) qui décrit le terme par son hyperonyme (habituellement dans les définitions canoniques)(ex. *un diabète gestationnel est un diabète*), contrairement à la définition par extension qui donne l'hyponyme du terme (ex. *le diabète contient le diabète de type 1, diabète de type 2...*) (Lehmann & Martin-Berthet, 1998). Dans notre travail nous nous intéressons aux contextes contenant un hyperonyme du terme en question étant donné que bon nombre de travaux considèrent la relation d'hyperonymie comme la plus fréquemment utilisée pour le définir (Green *et al.*, 2002). Même si le contexte (b) respecte cette contrainte, *diabète*, qui est l'hyperonyme de *diabète du type 2*), n'explicite pas la définition de ce dernier.

4.1.4 Synthèse

À partir des trois hypothèses précédentes, nous proposons de définir la notion d'exemple définitoire (ED) considéré comme un contexte pertinent permettant de préciser le sens du terme concerné comme étant **une phrase contenant un hyperonyme définissant le terme visé**. Par conséquent, parmi les exemples précédents ((a), (b) et (c)) seulement (c) sera

4. La définition canonique consiste à désigner d'abord le genre (la classe générale), dont relève le référent du terme à définir, puis à spécifier les différences qui le séparent des autres espèces appartenant au même genre (Lehmann & Martin-Berthet, 1998).
5. http://hosting.eila.univ-paris-diderot.fr/~juilliar/sitetermino/cours/cours_total_deb_john_2003.htm

retenu comme étant un exemple définitoire associé au terme *diabète du type 2*.

Néanmoins, les définitions du même terme peuvent varier d'un dictionnaire à l'autre. D'une part en raison du choix de l'hyperonyme et d'autre part parce que parfois l'hyperonyme direct peut être ambigu. La problématique est donc la suivante : quel hyperonyme faut-il choisir afin de mieux définir le terme ?

4.2 Identification d'exemples définitoires

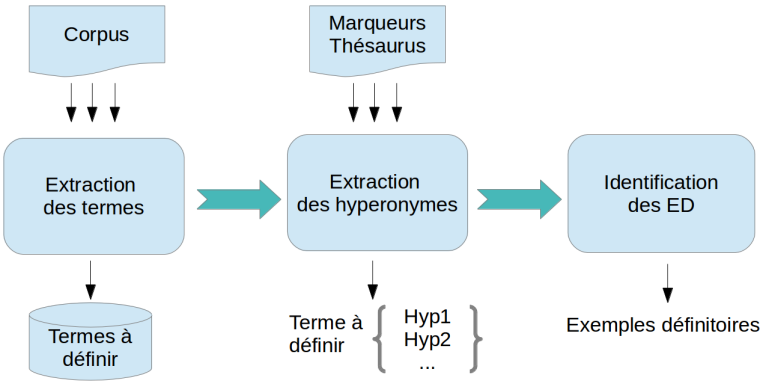

FIGURE 1 – Méthode d'identification des exemples définitoires.

Afin d'évaluer la validité des hypothèses précédentes et vérifier si notre définition de l'exemple définitoire est très restrictive, nous avons suivi la méthodologie illustrée par la figure 1. Cette méthodologie se décompose en 3 étapes :

1. **Extraction terminologique :** Dans le but d'associer à chaque terme du corpus ses exemples, nous avons d'abord extrait automatiquement la terminologie. Ensuite, nous avons filtré les termes les moins fréquents du corpus. Dans la suite, nous désignons par T la liste finale des termes filtrés.
2. **Extraction des hyperonymes :** En exploitant les relations hiérarchiques reliant les termes entre eux, nous déterminons, pour chaque terme appartenant à T, tous les hyperonymes présents dans le corpus étudié.
3. **Sélection des exemples :** Les phrases contenant un terme avec un de ses hyperonymes obtenus pendant l'étape précédente, seront considérées comme des exemples définitoires candidats. Ensuite, ces exemples candidats sont évalués manuellement dans le but d'identifier ceux qui sont des exemples définitoires.

4.3 Évaluation

4.3.1 Expérimentations

Nous avons appliqué la méthodologie décrite dans la section précédente sur un corpus français spécialisé. Il s'agit d'un corpus relatif à la thématique « diabète et alimentation » composé d'articles scientifiques contenant 206 460 mots (Goeuriot, 2009), dans lequel des définitions de 137 termes (simples et composés) ont été manuellement annotées (Nakao, 2010).

1. **Extraction terminologique :** Nous avons tout d'abord utilisé TermSuite [6] afin d'extraire respectivement les termes simples et les termes composés. Ensuite, nous avons considéré seulement les termes qui apparaissent plus de 10 fois dans le corpus pour les termes simples, et plus de 5 fois pour les termes composés. Ce choix est supposé maximiser la probabilité d'avoir des définitions parmi les phrases où occurre le terme. La terminologie obtenue contient, après ce filtrage, 677 termes composés et 809 termes simples.
2. **Extraction des hyperonymes :** Afin de pouvoir comparer les ED avec les définitions de référence (annotées dans le corpus), nous nous sommes intéressé seulement aux termes dont la définition est manuellement identifiée par Nakao (2010). Deux stratégies ont été adoptées de manière indépendante pour identifier les hyperonymes des termes

6. https ://logiciels.lina.univ-nantes.fr/redmine/projects/termsuite

définis : l'exploitation des marqueurs d'hyperonymie et l'utilisation d'un thésaurus.
À défaut de disposer de toutes les relations hiérarchiques reliant les termes entre eux, nous avons eu recours au thésaurus MeSH [7] afin de déterminer pour chaque terme extrait et présent dans le MeSH ses différents hyperonymes (jusqu'au quatrième niveau supérieur si possible). Par exemple si on considère la fonction *Hyperonyme (cholestérol alimentaire)* indiquant l'hyperonyme direct du terme *cholestérol alimentaire*, on obtient :

Hyperonyme (cholestérol alimentaire) = cholestérol : hyperonyme de niveau 1
Hyperonyme (cholestérol) = stérol : hyperonyme de niveau 2
Hyperonyme (stérol) = lipides : hyperonyme de niveau 3

D'autre part, le thésaurus permet éventuellement d'enrichir la terminologie avec les termes qui sont présents dans le corpus mais qui n'ont pas été extraits par l'outil d'extraction, comme *stérol*. Après avoir projeté la terminologie filtrée sur le thésaurus MeSH, seulement 18 termes simples et 10 termes composés sont retenus.
Concernant les marqueurs d'hyperonymie, nous avons utilisé une liste de 33 marqueurs de relation issus de la thèse de Séguéla (2001), dont 22 sont présentés dans la table 1. Contrairement aux marqueurs utilisés dans l'état de l'art, ceux que nous avons appliqués intègrent le terme et son hyperonyme. Ceci s'explique par le fait que nous nous intéressons à associer à un terme son ED (considéré comme CRC), tandis que les travaux de la bibliographie identifient les CRC en ne prenant pas en compte la présence d'un terme précis.

X_EST_UN_Y	ON_VERBE_DEFINITOIRE_X_UN_Y
X_ETRE_UNE_SORTE_DE_Y	X_EST_UN_Y_TRES
X_EST_LE_Y_LE_PLUS	X_ETRE_LE_PLUS_DE_TOUS_LES_Y
X_ET_AUTRES_Y	X_ET_D_AUTRES_Y
Y_ET_ADVERBE_DE_SPECIFICATION_X	X_VIRGULE_LE_Y_LE_PLUS
X_VIRGULE_LE_PLUS_ADJ_DES_Y	LE_PLUS_ADJ_DES_Y_VIRGULE_SOIT_X
Y_VIRGULE_ADVERBE_DE_SPECIFICATION_X	Y_VIRGULE_X_ADVERBE_DE_SPECIFICATION
Y_VIRGULE_ADVERBE_INCLUSION_X	Y_VIRGULE_ADVERBE_EXCLUSION_X
Y_PARENTHESES_X	INCLUSION_Y_VIRGULE_X
Y_PARMI_LESQUELS_X	Y_EXEMPLIFICATION_X
Y_ENUMERATION_X	Y_DEUX_POINTS_X

TABLE 1 – Exemples de marqueurs d'hyperonymie utilisés (X étant le terme et Y son hyperonyme)

4.3.2 Résultats

Terme	ex. sans hyp		avec hyp1		avec hyp2		avec hyp3		avec hyp4	
	ED	occ.	ED	occ.	ED	occ.	ED	occ.	ED	occ.
diabète de type1	2	32	4	45	0	0	0	0	0	0
diabète de type 2	2	46	6	69	0	0	0	1	0	0
diabète gestationnel	2	37	3	16	0	0	0	0	0	0
cholestérol alimentaire	0	5	6	6	0	0	0	0	0	0
ration calorique	0	14	0	1	0	0	0	0	0	0
excès de poids	6	49	0	1	0	0	0	0	0	0
régime alimentaire	0	14	0	0	0	0	0	0	0	0
perte de poids	0	16	0	0	0	0	0	0	0	0
prise de poids	0	31	0	0	0	0	0	0	0	0
activité physique	-	-	0	0	0	0	0	0	0	0

TABLE 2 – Résultats d'identification des exemples définitoires : hyperonymes extraits à partir du MeSH : cas des termes composés

La première colonne des tables 2 et 4 contient les termes définis dans le corpus étudié, qui sont obtenus après la projection sur le thésaurus. Suite à l'utilisation des marqueurs d'hyperonymie, les termes définis, et qui sont retenus, sont illustrés par

7. http ://mesh.inserm.fr/mesh/

Terme	nbr. d'hyperonymes	nbr. d'ED
diabète de type 1	1	1
diabète de type 2	4	1
diabète non insulino-dépendant	1	1
autosurveillance glycémique	1	0
diabète gestationnel	1	0
united kingdom prospective diabetes study	1	0
diabetes control and complications trial research group	1	0

TABLE 3 – Résultats d'identification des exemples définitoires : hyperonymes extraits en utilisant les marqueurs de relation : cas des termes composés

la première colonne des tables 3 et 5. La seconde colonne (ex. sans hyp) des tables 2 et 4 représente le nombre d'exemples définitoires (comme défini en section 4.1) identifiés parmi l'ensemble des phrases contenant le terme indépendamment de ses hyperonymes i (i étant le niveau d'hyperonymie). Nous avons considéré un exemple définitoire comme valide, s'il a été annoté dans le corpus comme une définition associée au terme en question. Par exemple, le terme *diabète de type 1* apparaît dans 32 phrases dont seulement 2 sont des exemples définitoires sans présence de ses hyperonymes. Ces deux exemples définitoires sont marqués dans le corpus étudié comme des définitions du terme *diabète de type 1*. Ce dernier co-occure également avec son hyperonyme direct (i.e *diabète*) dans 4 exemples définitoires parmi 45 phrases (colonne 3 de la table 2). Par contre le couple (*diabète de type 1, Hyperonyme 3(diabète de type 1)*) n'apparaît pas dans le corpus.

Terme	ex. sans hyp		avec hyp1		avec hyp2		avec hyp3		avec hyp4	
	ED	occ.	ED	occ.	ex.	occ.	ED	occ.	ED	occ.
glycémie	4	622	1	42	0	0	0	0	0	0
agpi	2	31	0	8	0	3	0	0	0	0
dnid	0	8	0	4	0	0	0	0	0	0
saccharose	2	23	0	0	0	0	0	0	0	0
amidon	0	23	0	0	0	0	0	0	0	0
pancréas	1	76	0	0	0	0	0	0	0	0
diabète	4	649	0	0	0	3	0	0	0	0
artère	0	89	0	0	0	0	0	0	0	0
athérosclérose	0	19	0	0	0	0	0	0	0	0
insuline	6	498	0	0	0	42	0	0	0	0
hyperglycémie	0	65	0	0	0	0	0	0	0	0
acétonurie	1	10	0	0	0	0	0	0	0	0
cétonurie	1	16	0	0	0	0	0	0	0	0
hypoglycémie	0	92	0	0	0	0	0	0	0	0
glycosurie	0	9	0	0	0	0	0	0	0	0
fructosamine	1	5	0	0	0	0	0	0	0	0
cholestérol	0	107	0	0	0	0	0	0	0	0
glucose	0	103	0	0	0	3	0	0	0	0

TABLE 4 – Résultats d'identification des exemples définitoires : hyperonyme extrait à partir du MeSH : cas des termes simples

La deuxième et la troisième colonne des tables 3 et 5 contiennent le nombre d'hyperonymes associés à chaque terme, en fonction des patrons utilisés. Par exemple, dans le cas du terme *diabète de type 2*, 4 hyperonymes sont trouvés et un seul exemple définitoire a été identifié.
Rappelons que nous sommes partis de l'hypothèse qu'un exemple définitoire est une phrase contenant à la fois le terme et son hyperonyme explicitant une définition. Dans un premier temps, nous nous sommes basés seulement sur la présence du terme et son hyperonyme afin d'identifier un ED, comme présenté dans les résultats précédents. Dans un second temps, nous avons exploité, en plus des hyperonymes, les marqueurs de relation qui explicitent la définition, afin de repérer les ED. En effet, la première colonne des tables 6 et 7 présentent respectivement les termes simples et composés définis qui sont retenus après avoir utilisé les marqueurs d'hyperonymie. La troisième colonne de ces tables contient le nombre d'occurrences d'un terme avec son hyperonyme (colonne 2). Parmi ces occurrences, celle qui contient un

Terme	nbr. d'hyperonymes	nbr. d'ED
diabète	4	3
insuline	4	3
glucose	2	0

TABLE 5 – Résultats d'identification des exemples définitoires : hyperonymes extraits en utilisant les marqueurs : cas des termes simples

marqueur explicitant la définition, est considérée comme étant un ED. En ce qui concerne les termes dont les hyperonymes sont proposés par le MeSH, seulement 2 ED ont été identifiés dans le cas des termes *diabète de type 1* et *diabète de type 2*. Ces résultats ont été manuellement validés. La table 2 montre que, d'une part, s'il existe des ED dans le corpus, l'

Terme	hyperonyme	nbr. d'occ	nbr. d'ED
diabète	cause	48	2
diabète	facteur	116	4
diabète	facteur de risque vasculaire	5	2
diabète	maladie	100	1
insuline	facteur	26	1
insuline	hormone	21	2
insuline	moment	7	0
insuline	patient	52	0
glucose	facteur	11	0
glucose	phénomène	5	0

TABLE 6 – Résultats d'identification des ED : hyperonyme (extrait avec des marqueurs d'hyperonymie) + Marqueurs de relation explicitant la définition : cas des termes simples

Terme	hyperonyme	nbr. d'occ	nbr. d'ED
united kingdom prospective diabetes study	u.k.p.d.s	2	2
diabetes control and complications trial research group	d.c.c.t	5	2
diabète de type 2	argument	9	0
diabète de type 2	critère	6	0
diabète de type 2	diabète	69	2
diabète de type 2	maladie	5	3
diabète de type 1	diabète	45	2
diabète non insulino-indépendant	maladie métabolique	1	1
Autosurveillance glycémique	outil	1	1
diabète gestationnel	risque	10	2

TABLE 7 – Résultats d'identification des ED : hyperonyme (extrait avec des marqueurs d'hyperonymie) + Marqueurs de relation explicitant la définition : cas des termes composés

hyperonyme de niveau 1 s'avère plus performant que les autres hyperonymes proposés par le MeSH. D'autre part, les résultats montrent que les définitions contenant cet hyperonyme direct sont majoritaires par rapport aux définitions sans hyperonymes.
Par ailleurs les colonnes (avec. hyp) des tables 2 et 4 (ainsi que la table 5) montrent que les hyperonymes sont rares dans les contextes des termes en question. Ceci explique le nombre limité des termes présents dans les tables 3 et 5. Ainsi, nous déduisons que l'hypothèse 3 est restrictive pour identifier les exemples définitoires. Concernant l'hypothèse 2, la segmentation du corpus peut influencer les résultats. Par exemple, une phrase mal-segmentée ne peut pas être considérée comme un exemple définitoire. De plus, dans le cas d'une définition contenant deux phrases, où la première contient le terme visé et la deuxième contient son hyperonyme, l'hypothèse 2 présente également une contrainte supplémentaire empêchant de retenir une des deux phrases comme un exemple définitoire. Par exemple, aucune de ces deux phrases *ces schémas insuliniques sont également appelés basal-bolus. Bolus étant un mot latin signifiant action de jeter, coup de dé, coup de filet...* ne peut être considérée comme un exemple définitoire du terme *basal-bolus*.

5 Conclusion et perspectives

Dans ce travail nous nous sommes intéressés à l'identification des exemples définitoires dans un corpus comparable comme étant des CRC. Nous avons tout d'abord proposé trois hypothèses définissant cette notion. Nous avons considéré la relation d'hyperonymie comme un indicateur de définitions. Ensuite, nous avons proposé une méthode permettant de sélectionner ces exemples définitoires. Les résultats obtenus ont montré que les hypothèses proposées sont valides mais contraignantes étant donné que les hyperonymes sont peu fréquents dans les contextes des termes visés.
En ce qui concerne l'évaluation, les deux stratégies que nous avons adoptées pendant les premières expériences ne semblent pas être appropriées. D'une part, car la comparai- son des CRC candidats avec des définitions, qui sont souvent rares dans un petit corpus, ne permet pas d'avoir des résultats significatifs. Un CRC a été considéré comme valide, si et seulement s'il déclenchait une définition qui était préalablement annotée. Autrement dit, il s'agissait d'évaluer la richesse des CRC en termes de « définitoires », tandis qu'ils pouvaient apparaître dans des phrases autres que les définitions. D'autre part, parce que faire valider manuellement les CRC est un exercice sans doute très coûteux. Dans un premier temps, nous proposons d'exploiter séparément les hypothèses présentées, afin de rendre moins restrictives la définition de l'exemple définitoire. Dans un second temps, nous utiliserons d'autres relations telles que la causalité et la méronymie, comme indice de CRC.

Remerciement

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-12CORD-0020.

Références

ATKINS B. S. & RUNDELL M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

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

BARRIÈRE C. (2004). Knowledge-rich contexts discovery. *Proceedings of the Seventeenth Canadian Conference on Artificial Intelligence (AI'2004)*.

F. BENARMARA, N. HATOUT, P. MULLER & S. OZDOWSKA, Eds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

B. BIGI, Ed. (2014). *Actes de TALN 2014 (Traitement automatique des langues naturelles)*, Marseille. ATALA, LPL.

BIRD S., DALE R., DORR B. J., GIBSON B. R., JOSEPH M., KAN M.-Y., LEE D., POWLEY B., RADEV D. R. & TAN Y. F. (2008). The acl anthology reference corpus : A reference dataset for bibliographic research in computational linguistics. In *LREC* : European Language Resources Association.

BOWKER L. & PEARSON J. (2002). *Working with specialized language : a practical guide to using corpora*. Routledge.

DAILLE B. & MORIN E. (2005). French-english terminology extraction from comparable corpora. In *Natural Language Processing–IJCNLP 2005*, p. 707–718. Springer.

DÉJEAN H. & GAUSSIER E. (2002). Une nouvelle approchea l'extraction de lexiques bilinguesa partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, p. 1–22.

DIDAKOWSKI J., LEMNITZER L. & GEYKEN A. (2012). Automatic example sentence extraction for a contemporary german dictionary. In *Proceedings EURALEX*, p. 343–349.

FAHMI I. & BOUMA G. (2006). Learning to identify definitions using syntactic features. In *Proceedings of the EACL 2006 workshop on Learning Structured Information in Natural Language Applications*, p. 64–71.

FIRTH J. R. (1957). A synopsis of linguistic theory 1930-55. *The Philological Society*, **1952-59**, 1–32.

FUNG P. & MCKEOWN K. (1997). Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, p. 192–202.

GANGEMI A., NAVIGLI R. & VELARDI P. (2003). The ontowordnet project : Extension and axiomatization of conceptual relations in wordnet. In R. MEERSMAN, Z. TARI & D. C. SCHMIDT, Eds., *CoopIS/DOA/ODBASE*, volume 2888 of *Lecture Notes in Computer Science*, p. 820–838 : Springer.

GOEURIOT L. (2009). *Découverte et caractérisation des corpus comparables spécialisés*. PhD thesis, Université de Nantes.

GREEN R., BEAN C. & MYAENG S. (2002). *The Semantics of Relationships : An Interdisciplinary Perspective*. Information science and knowledge management. Kluwer Academic Publishers.

KILGARRIFF A., HUSÁK M., MCADAM K., RUNDELL M. & RYCHLÝ P. (2008). Gdex : Automatically finding good dictionary examples in a corpus. In *Proc. Euralex*.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benarmara *et al.*, 2007), p. 101–110.

LAROCHE A. & LANGLAIS P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd international conference on computational linguistics*, p. 617–625 : Association for Computational Linguistics.

LEHMANN A. & MARTIN-BERTHET F. (1998). *Introduction à la lexicologie : sémantique et morphologie*. Collection Lettres supérieures. Ed. Dunod.

MALAISÉ V., ZWEIGENBAUM P. & BACHIMONT B. (2004). Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie. *Actes de TALN*, p. 269–278.

MARTÍNEZ R., MARTÍNEZ G., DE LINGÜÍSTICA APLICADA U. P. F. I. U. & BACH C. (2009). *Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios*. Sèrie tesis. Universitat Pompeu Fabra.

MEYER I. (2001). Extracting knowledge-rich contexts for terminography - A conceptual and methodological framework. In *Recent Advances in Computational Terminology*, p. 279–302.

MURESAN S. & KLAVANS J. (2002). A method for automatically building and evaluating dictionary resources. In *LREC* : European Language Resources Association.

NAKAO Y. (2010). *Analyse contrastive français-japonais du discours en langue de spécialité-modalité et définition phrastique*. PhD thesis, Université de Nantes.

NAVIGLI R. & VELARDI P. (2010). Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 1318–1327, Uppsala, Sweden : Association for Computational Linguistics.

PEARSON J. (1998). *Terms in context*, volume 1. John Benjamins Publishing.

RAPP R. (1999). Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, p. 519–526 : Association for Computational Linguistics.

REBEYROLLE J. (2000). *Forme et fonction de la définition en discours*. PhD thesis, Toulouse 2.

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

REIPLINGER M., SCHÄFER U. & WOLSKA M. (2012). Extracting glossary sentences from scholarly articles : A comparative evaluation of pattern bootstrapping and deep analysis. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, ACL '12, p. 55–65, Stroudsburg, PA, USA : Association for Computational Linguistics.

SAGGION H. (2004). Identifying definitions in text collections for question answering. In *LREC* : European Language Resources Association.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benarmara *et al.*, 2007), p. 401–410.

VELARDI P., FARALLI S. & NAVIGLI R. (2013). Ontolearn reloaded : A graph-based algorithm for taxonomy induction. *Computational Linguistics*, **39**(3), 665–707.

WESTERHOUT E. (2009). Extraction of definitions using grammar-enhanced machine learning. In A. LASCARIDES, C. GARDENT & J. NIVRE, Eds., *EACL (Student Research Workshop)*, p. 88–96 : The Association for Computer Linguistics.