Compréhension ciblée de texte par reconnaissance de relations

Anne-Laure Ligozat

Université Paris-Sud

Mémoire d'Habilitation à Diriger des Recherches

Compréhension ciblée de textes par reconnaissance de relations

ou Comment établir des relations simples avec ses voisins

Anne-Laure Ligozat

HDR soutenue le 9 décembre 2016

Jury :

- Patrice Bellot, Professeur en informatique, Aix-Marseille Université, LSIS
- Béatrice Daille, Professeur en informatique à l'Université de Nantes, LINA
- Cédrick Fairon, Professeur à l'Université Catholique de Louvain, Belgique, directeur du laboratoire CENTAL
- Chantal Reynaud, Professeur en informatique à l'Université Paris-Sud, LRI
- Isabelle Tellier, Professeur à l'Université Paris 3, LaTTiCe
- Brigitte Grau, Professeur en informatique à l'ENSIIE, LIMSI

Table des matières

Chapitre 1

Introduction

La disponibilité massive de documents textuels sous forme électronique et la création de ressources informatiques et linguistiques (outils d'analyse, bases de connaissance...) ont considérablement facilité les possibilités d'analyse et de fouille de tels documents. Des tâches de compréhension ciblée des textes, telles que de recherche ou d'extraction d'information précise, bénéficient de nouvelles possibilités d'analyse.

Mes travaux de recherche sont centrés sur ces processus de compréhension ciblée des textes, qui cherchent à extraire une représentation partielle ou simplifiée de ces textes pour une application donnée. L'objectif applicatif est d'améliorer l'accès à une information précise, soit pour un utilisateur humain, soit pour un traitement ultérieur des données extraites.

Ce type de processus amène à traiter deux types de relations entre énoncés (entendu comme un court extrait de texte [1]) : des relations entre termes d'un énoncé, c'est-à-dire des relations syntagmatiques entre deux occurrences de concepts dans un même énoncé, et des relations de possible substitution entre termes, c'est-à-dire des relations paradigmatiques entre deux occurrences de concepts. Prenons un exemple pour illustrer ces deux types de relations.

Le premier type de relation pourra entrer en jeu par exemple pour extraire automatiquement la réponse à une question : dans l'exemple de la figure 1.1, la reconnaissance de la relation de type *affiliation* entre «Rosa Parks» et la réponse attendue dans la question, puis entre « Rosa Parks » et «NAACP» dans la phrase réponse, permettra d'extraire la réponse.

Le second type de relation pourra notamment être utilisé pour simplifier la phrase réponse, afin par exemple de produire la phrase «Rosa Parks était membre de la NAACP» de l'exemple de la figure 1.2 : dans ce cas, la relation de proximité sémantique entre «s'était investie dans ses activités militantes au sein de» et «était membre de» sera recherchée.

1. Nous parlons d'énoncé plutôt que de phrase, bien qu'en pratique ces énoncés soient de l'ordre de la phrase, car ce terme est plus général, et permet également d'inclure des tours de dialogue par exemple

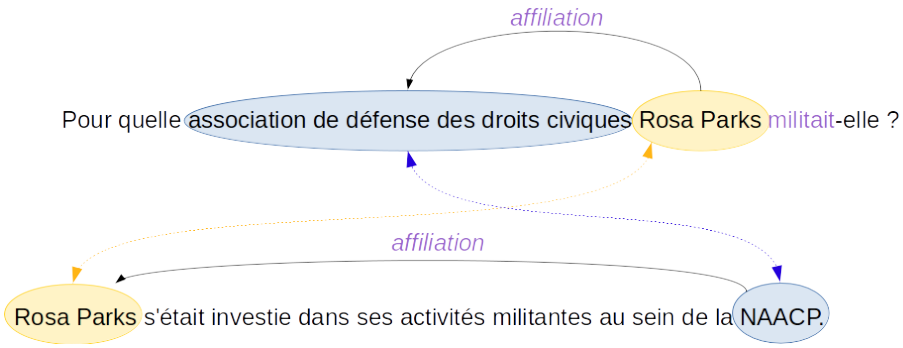


FIGURE 1.1 – Relation sémantique (*affiliation*) entre termes de l'énoncé

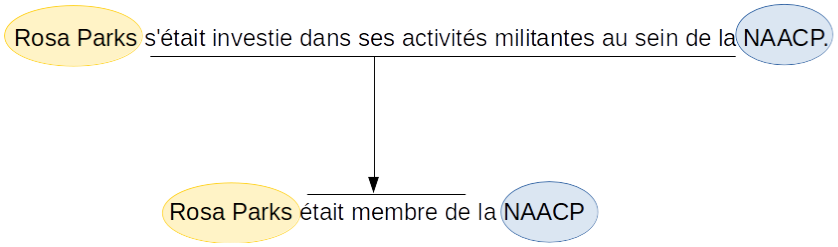


FIGURE 1.2 – Relation de simplification entre deux énoncés

En réalité, dans le premier exemple, plusieurs types de relations sémantiques entrent en jeu :

- la relation, non explicite dans les énoncés, d'instanciation entre «association de défense des droits civiques» et «NAACP», entre la question et la phrase réponse ;
- la relation d'identité entre les deux entités «Rosa Parks» ;
- la relation d'affiliation entre «Rosa Parks» et «NAACP». Cette relation est explicite, au sens où elle possède des marqueurs dans les énoncés. Afin de reconnaître cette relation, il est possible de s'appuyer sur ses expressions dans les énoncés, «militait pour» et «s'était investie au sein de», et de rapprocher chaque expression de la relation *affiliation* ; une autre possibilité consiste à reconaître que «militait pour» et «s'était investie au sein de» sont des paraphrases, et donc qu'il existe une relation sémantique entre les deux énoncés.

Ces problèmes peuvent donc être considérés sous l'angle de la détection de relations de proximité sémantique. La différence essentielle entre relations intra et inter-énoncés est que pour les premières, la reconnaissance peut s'appuyer sur des marqueurs explicites de la relation.

Les définitions des types de relation pouvant varier légèrement selon les auteurs, nous présentons maintenant les notions auxquelles nous ferons référence

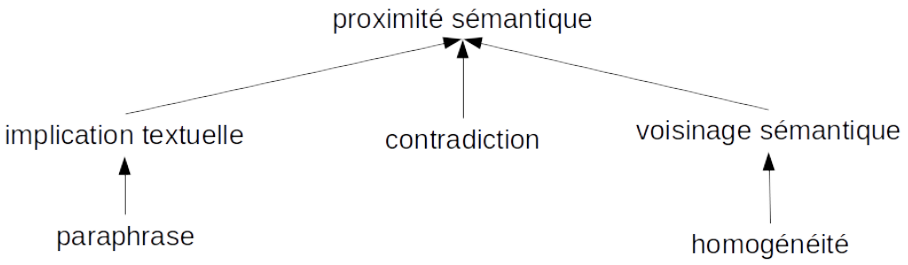


FIGURE 1.3 – Relations de proximité sémantique entre énoncés

dans ce manuscrit (voir figure 1.3), en nous appuyant sur [Dagan *et al.*, 2013, MacCartney, 2009, Pavlick *et al.*, 2015, Agirre *et al.*, 2016].

En traitement automatique des langues, de nombreux problèmes peuvent être posés comme des problèmes d'implication textuelle, donc nous commencerons par définir ce terme. Il existe une relation d'implication textuelle (*textual entailment*) entre un texte T et une hypothèse H (T implique H) si et seulement si un humain lisant H en conclurait que H est très probablement vraie [Dagan *et al.*, 2006, Mihalcea *et al.*, 2006, Dagan *et al.*, 2013]. L'implication textuelle correspond donc à une notion d'inférence sur des textes. Certains cas d'implication textuelle correspondent à la formulation de nouvelles informations en utilisant des connaissances du monde : ainsi, « L'alliance CDU-CSU de la chancelière allemande Angela Merkel connaît une forte érosion de ses intentions de vote selon un sondage publié mardi. » [2] implique « Angela Merkel est une citoyenne allemande ». D'autres cas correspondent à des généralisations ou des équivalences. Le cas de généralisation correspond à des couples d'énoncés comportant un énoncé plus général et un énoncé plus spécifique, et peut se décliner sur le plan syntaxique ou sémantique [Pais, 2013].

Le cas d'équivalence, c'est-à-dire d'implication textuelle bidirectionnelle, correspond à la notion de *paraphrase* interprétée de façon stricte. En réalité, la notion de paraphrase comprend souvent également certains cas de généralisation (« Ikea commercialise un nouveau modèle de fauteuil » peut être considéré comme une paraphrase de « Ikea commercialise un nouveau meuble »).

Lorsque les énoncés ne sont pas en relation d'implication textuelle, ils peuvent être *contradictoires* : l'hypothèse H contredit le texte T si un lecteur humain dirait que H a très peu de chances d'être vraie étant donné T [Dagan *et al.*, 2013]. « Yahoo a annoncé aujourd'hui le rachat d'Overture pour 1,63 milliard de dollars. » contredit par exemple l'hypothèse « Yahoo a vendu Overture ».

Lorsqu'il existe une relation entre les énoncés, qui ne correspond pas à l'un des cas précédents, par exemple une relation thématique, on parle généralement de *voisinage sémantique* (*semantic relatedness*) [3]. Ainsi, « La femme joue du

2. Ouest France du 20/10/2016
3. Certains auteurs emploient aussi le terme de similarité textuelle (*Semantic Textual Si-*

violon » et « La jeune fille aime écouter la guitare » [4] peuvent être considérées comme des phrases voisines car portant sur le même thème bien que non équivalentes.

Nous introduirons dans le chapitre 4 la notion d'*homogénéité* sémantique entre énoncés, que nous considérons comme un cas particulier de voisinage sémantique.

Enfin, deux énoncés peuvent évidemment être sans rapport l'un avec l'autre.

Le degré de proximité entre énoncés, allant de l'équivalence à l'indépendance, peut être évalué de façon discrète (ce qui était le cas dans certaines campagnes d'évaluation d'implication textuelle par exemple), ou de façon continue, par un score de proximité.

Dans mes travaux, j'ai exploré plusieurs domaines nécessitant la reconnaissance de relations inter ou intra énoncés, qui peuvent être considérés sous l'angle de la similarité textuelle (voir tableau 1.1).

Ainsi, la reconnaissance de relations entre entités d'un énoncé peut être vue comme un problème d'implication textuelle : il s'agit de montrer que l'énoncé considéré implique une mention de la relation à classifier. Les méthodes d'extraction de relations sont cependant fondées en général sur la reconnaissance de la proximité sémantique entre l'énoncé à catégoriser et un ensemble d'exemples de mentions de la relation considérée.

En simplification textuelle, l'énoncé original implique l'énoncé simplifié, et la génération de cet énoncé simplifié nécessite de prendre en compte divers niveaux linguistiques afin d'améliorer la lisibilité.

Les systèmes de réponse à des questions (QR dans le tableau 1.1) s'appuient sur de l'extraction de relations lorsque les ressources dans lesquelles les réponses sont recherchées sont des bases de connaissance, ou plutôt de la reconnaissance de paraphrase ou d'implication textuelle lorsque les ressources sont des corpus textuels.

Je présenterai également les travaux que nous avons menés en sélection de distracteurs pour des QCM, que nous avons posé comme un problème de reconnaissance de voisinage sémantiques, et plus précisément d'homogénéité.

J'indique enfin dans le tableau 1.1 le problème de la sélection de réponses pour un système de dialogue non supervisé, que je n'aborde pas dans ce manuscrit, mais sur lequel j'ai commencé à travailler récemment, et pour lequel nous avons mis en œuvre des méthodes fondées sur un voisinage sémantique entre l'initiative de l'utilisateur et les initiatives d'un corpus de dialogues.

Dans chaque domaine, je me suis intéressée aux questions suivantes : quel type de représentation des textes permet de capter au mieux la relation considérée ; quel type de connaissances sont utiles pour cette reconnaissance ; et quel type de méthode est adaptée à chaque type de relation ? J'ai généralement abordé ces questions en commençant par constituer et analyser des corpus pour la tâche considérée, avant de modéliser le problème, puis de proposer une méthode pour tester les hypothèses de recherche.

milarity), qui me semble cependant porter à confusion.

4. Exemple de la tâche 1 de la campagne SemEval-2016 : Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation

	implication textuelle	contradiction	voisinage
relations intra-énoncés	extraction de relations QR bases de connaissances		
relations inter-énoncés	simplification QR textes	sélection de distracteurs	*sélection de réponses*

TABLE 1.1 – Positionnement des travaux présentés en fonction du type de proximité entre énoncés

Je décris dans ce mémoire les travaux qui sont au centre de mes recherches. Le chapitre 2 présente mes travaux en extraction de relations, en terminant par les systèmes de réponse à des questions, qui peuvent être envisagés comme proches de l'extraction de relations, ou proches de la reconnaissance d'implication textuelle. Le chapitre 3 aborde le domaine de la simplification de textes, et le chapitre 4 la recherche de voisins sémantiques dans un cadre pédagogique.

Bien que ces travaux se trouvent regroupés dans mon manuscrit, ils sont tous (ou presque) le fruit de nombreuses collaborations, que je citerai au fur et à mesure. J'ai eu la chance de travailler avec des collègues qui m'ont énormément appris et ouvert de nouvelles perspectives, et d'encadrer des stagiaires et doctorants à qui je dois une partie de ce manuscrit.

Chapitre 2

Extraction de relations

La première partie de ce document est consacrée à mes travaux en extraction de relations textuelles, pour l'extraction d'information ou la recherche d'information précise.

2.1 Définitions

Nous considérons les textes comme des potentielles sources d'information structurée, c'est-à-dire contenant des relations entre entités, qui peuvent être interrogées ou fouillées. Les relations visées sont donc des relations sémantiques, qui structurent l'information des textes. L'objectif sera soit d'extraire des informations ciblées d'un ensemble de textes, soit d'extraire toutes les relations d'un extrait de texte, dans le cadre de l'extraction d'information, ou de la recherche d'information précise.

Nous commençons par définir plusieurs concepts [Grau *et al.*, 2015].

Une *entité* e_n représente un objet du monde, avec un identifiant unique, par exemple son *uri* dans une base de connaissances. Elle est liée à un type, ou une classe, et est référée dans les textes par différentes séquences de mots, correspondant à des *mentions*. Les entités nommées reconnues dans un texte, par exemple des personnes ou des lieux, sont des mentions d'entités auxquelles on a associé un type. Plus généralement, les entités auxquelles nous ferons référence ici peuvent être des entités nommées, mais aussi des termes comme des noms de maladies ou de médicaments.

Une *relation binaire* ou *classe de relation* R associe des entités d'un domaine $D1$ à des entités d'un domaine $D2$ (voir figure 2.1). Ces relations sont appelés *propriétés* dans le cadre des ontologies. Un domaine correspond à une classe d'entités, ou à un ensemble de classes d'entités. Une relation peut être orientée ou symétrique, si l'ordre des arguments n'a pas d'importance.

Les relations peuvent également être *n-aires* si elles associent plus de deux entités entre elles. Les relations n-aires sont parfois appelées *événements*. Selon

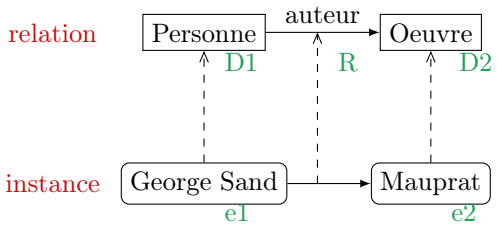


FIGURE 2.1 – Relation : définitions

les recommandations du W3C [1], une relation n-aire doit être représentée par un concept, auquel seront rattachés les autres éléments par des propriétés. Pour simplifier, nous donnons dans la suite des définitions et exemples pour les relations binaires, mais dans nos travaux, nous nous sommes intéressée aux deux types de relations.

Une *instance* de relation associe une entité $e1$ à une entité $e2$ par une relation R. Une *mention* de relation est la réalisation phrastique d'une instance de relation dans un énoncé, reliant les mentions d'entités.

Par exemple *auteur_de(Personne, Œuvre)* est une relation, *auteur_de(George_Sand, Mauprat)* est une instance de la relation *auteur* et «Mauprat, que Sand écrivit entre 1835 et 1837, est bien un roman capital dans son œuvre.» est une mention de la relation, et par extension «écrivit». La mention de la relation est ici explicite, puisqu'elle se manifeste via le verbe «écrire», mais elle peut être implicite, au sens où l'énoncé donne peu d'indication sur la relation entre les entités, comme par exemple dans la phrase «George Sand crée dans Mauprat un personnage sans équivalent dans l'univers pourtant divers des personnages romanesques» ou encore dans «Le grand art de George Sand, qui se vérifie dans Mauprat comme dans Les Maîtres sonneurs et d'autres romans, est de ne pas laisser son message politique nuire à l'intérêt du récit.»

La tâche d'*extraction de relations* consiste, étant donné deux entités (ou plus) dans un énoncé (soit données en entrée, soit extraites précédemment), à déterminer si ces entités sont reliées par une relation et à identifier cette relation le cas échéant. Les entités à relier sont généralement des instances de concepts (comme dans la figure 2.1).

Les classes de relations à identifier sont généralement issues d'un ensemble donné, pouvant aller de quelques classes à plusieurs milliers de classes, lorsque ces relations sont celles d'une base de connaissances en domaine ouvert.

L'extraction de ces relations nécessite de définir une représentation des infor-

1. https://www.w3.org/TR/swbp-n-aryRelations/

mations recherchées ainsi que des corpus annotés en fonction de cette représentation. Le processus d'extraction peut ensuite s'appuyer sur différents niveaux de connaissances linguistiques pour les relations.

2.2 Représentation des relations

Définir une représentation des relations suppose de choisir à la fois le format des relations (triplets, prise en compte des relations n-aires, langage de représentation des données...) et le jeu de relations utilisé

Nous présentons quelques exemples d'ensembles de relations afin de situer nos modèles de représentation. En domaine général et dans le cas d'un ensemble restreint de relations, deux jeux de données font généralement référence : celui d'ACE 2005 [2] et celui de SemEval 2010 tâche 8 [3]. Le tableau 2.1 présente les relations ACE 2005 ; le tableau 2.2 celles de SemEval.

Les relations d'ACE sont des relations entre deux entités au sein d'une phrase, entité étant pris au sens large puisqu'une entité peut être une entité nommée, mais aussi un nom ou groupe nominal, ou un pronom faisant référence à ou décrivant une entité. Les relations entre ces entités peuvent faire référence à des relations entre concepts, ou à des relations instanciées. Certaines relations sont symétriques. Une relation possède en outre les attributs suivants : type, sous-type (voir tableau 2.1), modalité (relation possible - *asserted* - ou fait établi, mais pas de négation de la relation) et temps (passé, présent, futur, ou non spécifié).

En plus de ces relations, cette tâche définit également des événements, qui peuvent être considérés comme des relations dans d'autres référentiels : ainsi, ACE définit un événement *BE-BORN*, ayant comme arguments une personne, une date et un lieu, qui sera décomposé en deux relations binaires dans la base de connaissances Wikidata (relations *place of birth* et *date of birth* de la personne).

Les relations de SemEval sont des relations sémantiques entre groupes nominaux comportant un nom commun comme tête. Les relations ne se recoupent pas entre elles. Seules les relations du «monde réel» sont annotées, ce qui exclut donc les relations possibles, supposées ou avec négation.

Les ensembles de relations issus du web sémantique sont beaucoup plus importants : l'ontologie de DBpedia par exemple contient plus d'un millier de types de relations entre concepts (1 650 propriétés en 2013, [Lehmann *et al.*, 2013]), tandis que Wikidata en contient environ 2 500 en 2016 [4].

Les choix effectués pour créer ces ensembles de relations peuvent être différents, certains choisissant par exemple de représenter une relation et sa négation avec une même relation, accompagnée d'un attribut de négation, d'autres sous forme de deux relations différentes, d'autres encore de ne représenter que les relations vérifiées.

2. Guide d'annotation et guide d'évaluation
3. Site web de la tâche
4. Propriétés Wikidata

Type	Sous-type	Exemple
artifact (ART)	User-Owner-Inventor-Manufacturer	*My*$_1$ *house*$_2$ is in West Philadelphia
Gen-affiliation (GEN-AFF)	Citizen-Resident-Religion-Ethnicity	Some *Missouri voters*
	Org-Location	its *Beijing branch*
Org-affiliation (ORG-AFF)	Employment	the *CEO* of *Microsoft*
	Founder	*Coursera* co-founder *Daphne Koller*
	Ownership	*Dallas Cowboys*$_2$ *owner*$_1$
	Student-Alum	*Card*$_1$ graduated from the *University of South Carolina*$_2$
	Sports-Affiliation	*Zidane* led *France* to the European title this year
	Investor-Shareholder	In 1992, the *Motorola Company* invested 120 million US dollars in *Tianjin*...
	Membership	a *member* of the Supreme *Court*
part-whole (PART-WHOLE)	Geographical	*Moscow*, *Russia*
	Subsidiary	*Microsoft*'s accounting *department*
person-social* (PER-SOC)	Business	*his lawyer*
	Family	*his wife*
	Lasting-Personal	*his* friendship with some right-wing *mayors*
physical* (PHYS)	Located	*He* was campaigning in his home *state* of Tennessee
	Near	a *town* some 50 miles south of *Salzburg* in the central Austrian Alps

TABLE 2.1 – Relations ACE (les relations accompagnées de * sont les relations symétriques, et les têtes des arguments sont en italique souligné)

Le terme de relation sémantique dans la littérature regroupe par ailleurs des relations de diverses natures : relations hiérarchiques, spatiales, lexicales ; relations entre entités nommées ou entre tout type d'entité ; relations contextuelles (entre deux instances de concept dans un contexte textuel donné) ou absolues (relations lexicales par exemple)... Le type de relations recherchées et de contraintes utilisées dépendra en partie des types d'application envisagés :

Type	Exemple
Cause-Effect (CE)	The *news* brought about a *commotion* in the office.
Instrument-Agency (IA)	*Carpenters* build many things from *wood* and other materials, like buildings and boats.
Product-Producer (PP)	The *government* built 10,000 new *homes*.
Content-Container (CC)	I emptied the *wine bottle* into my glass and toasted my friends.
Entity-Origin (EO)	One basic trick involves a spectator choosing a *card* from the *deck* and returning it.
Entity-Destination (ED)	He sent his *painting* to an *exhibition*.
Component-Whole (CW)	Feel free to download the first *chapter* of the *book* (PDF - 78 kb) as free sample.
Member-Collection (MC)	A person who is serving on a *jury* is known as *juror*.
Message-Topic (CT)	Mr Cameron asked a *question* about tougher *sentences* for people carrying knives.

TABLE 2.2 – Relations tâche 8 SemEval 2010 (les arguments de la relation sont en italique et sont toujours dans l'ordre e1 puis e2)

selon qu'il s'agit par exemple de peupler une base de connaissance entre entités ou une base de données de résultats médicaux, il ne s'agira pas du même type de relations, ni du même type de méthodes. Le type de relations dépend de l'objectif visé : pour la construction d'une base de connaissance, les relations avérées seront recherchées, tandis que dans un cadre d'extraction d'information, des relations négatives pourront être pertinentes.

Nous présentons maintenant deux exemples de représentation de relations que nous avons définis, le premier concernant des relations binaires, le second des relations n-aires.

2.2.1 Exemple de formalisation de relations binaires : informations médicales

Le projet Cabernet (ANR 2013-2017) vise à produire une analyse fine du contenu des textes du domaine biomédical, et en particulier des textes cliniques. L'objectif du point de vue médical est de disposer de représentations structurées des informations contenues dans les différents documents médicaux, qui seront exploitables de façon automatique. Dans le cadre de ce projet, nous avons mis en place un schéma d'annotation des textes permettant d'extraire les informations médicales pertinentes pour un traitement ultérieur (analyse de données, recherche d'information...). Ce travail a été réalisé avec Aurélie Névéol, Cyril Grouin, Louise Deléger, Thierry Hamon et Leonardo Campillos.

L'objectif était de créer un schéma d'annotation à large couverture, afin de pouvoir annoter tout type de document clinique. Ce schéma se compose de

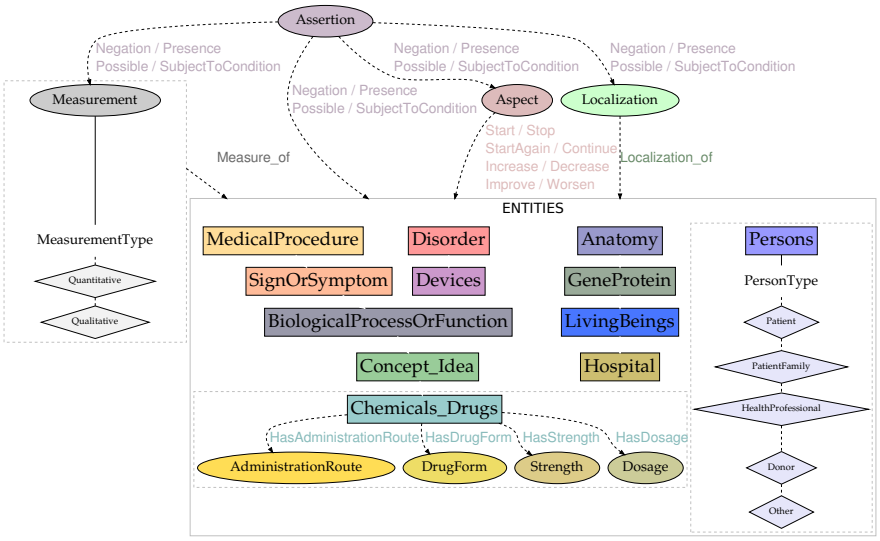


FIGURE 2.2 – Entités (rectangles) et attributs (losanges pour les classes fermées, ellipses pour les attributs avec ancres textuelles) du schéma d'annotation

trois types d'annotations : des entités, des attributs et des relations. Les entités correspondent en partie à des concepts UMLS [5] (Anatomy, Chemicals & Drugs, Living Beings...), auxquels nous avons ajouté trois catégories particulièrement importantes pour le domaine clinique : Sign or Symptom, Persons et Hospital. Les attributs donnent des informations supplémentaires sur une entité : présence ou absence, aspect... La description complète des entités et attributs est donnée dans [Deléger *et al.*, 2014] et une vue synthétique est donnée figure 2.2.

L'ensemble des relations a été créé en utilisant en partie le réseau sémantique de l'UMLS, mais en tirant également parti des schémas d'annotation cliniques existants [Roberts *et al.*, 2009, Savova *et al.*, 2012, Albright *et al.*, 2013, Uzuner *et al.*, 2011]. Les contraintes que nous nous étions données étaient les suivantes :

— se rapprocher autant que possible des modélisations standard, notamment pour le domaine médical celle de l'UMLS. Pour certains choix de modélisation ou d'annotation, nous avons fait appel à des experts médicaux, afin de déterminer quelle modélisation serait la plus utile, ou quelle annotation leur paraissait la plus pertinente. Nous avons également cherché à être cohérents vis-à-vis de TimeML [6] pour les aspects temporels ;

— unifier les schémas existants, notamment ceux de [Ogren *et al.*, 2008],

5. Unified Medical Language System, méta-thésaurus du domaine médical, https://www.nlm.nih.gov/research/umls/

6. Langage pour l'annotation d'expressions temporelles dans les textes, http://www.timeml.org/

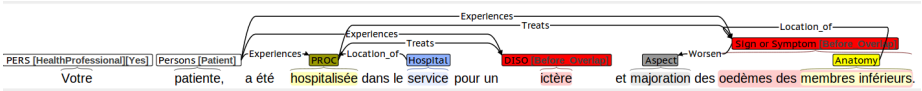


FIGURE 2.3 – Exemple d'annotation d'une phrase en entités et relations

CLEF [Roberts *et al.*, 2009], i2b2 [Uzuner *et al.*, 2010, Uzuner *et al.*, 2011, Sun *et al.*, 2013], IxA-Med-GS [Oronoz *et al.*, 2015], THYME [Styler IV *et al.*, 2014], SHARP, MiPACQ, et ShARe/CLEF eHealth ;

— choisir des relations les plus génériques possibles, qui s'appliquent à un maximum de catégories d'entités ;

— choisir un jeu de relations suffisamment restreint pour être utilisé facilement dans une tâche d'annotation.

Le jeu de relations résultant est donné dans les tableaux 2.3, 2.4 et 2.5, et un exemple de phrase annotée en entités et relations est présenté dans la figure 2.3.

On peut remarquer que certaines relations ne sont pas spécifiques au domaine médical : relations spatiales et temporelles notamment, mais aussi *Causes* par exemple, qui pourraient s'appliquer en domaine général. Certaines relations sont en revanche spécifiques au domaine médical, comme *Reveals* ou *Treats*. Contrairement à d'autres schémas, le jeu de relations n'inclut pas de relations négatives : si un compte rendu médical indique par exemple qu'un examen n'a pas montré de signe d'un problème médical, la relation *reveals* sera utilisée pour l'annotation, avec un attribut de négation sur l'entité du problème médical.

Ces relations ont été utilisées pour l'annotation d'un corpus de 500 comptes rendus cliniques en français, provenant de différents hôpitaux français et concernant différentes spécialités médicales. Un premier système d'annotation des entités a été développé, et dans la suite du projet Cabernet, un système d'extraction de relations sera développé.

2.2.2 Exemple de formalisation de relations n-aires : résultats expérimentaux

Nous avons jusqu'ici discuté principalement des relations binaires, mais nous présentons maintenant un exemple de formalisation de relation n-aire. Le cadre d'application était la modélisation de résultats expérimentaux en physiologie rénale : il existe en effet dans ce domaine une base de données, *Quantitative Kidney DataBase*, ou QKDB [Dzodic *et al.*, 2004], regroupant des résultats expérimentaux quantitatifs utiles pour la modélisation, à partir de leurs publications dans les articles de physiologie rénale et qui a été créée dans le contexte du projet Physiome international [Hunter *et al.*, 2010].

Ce travail a fait partie du travail de thèse d'Anne-Lyse Minard [Minard, 2012] et a été réalisé avec Anne-Lyse Minard, Brigitte Grau et Stephen Randall Thomas.

Nous appelons *résultat expérimental* un résultat quantitatif obtenu suite

Relation	Définition
Affects	Produces a direct effect on a process or function
Causes	Brings about a condition or an effect. Implied here is that an agent, such as for example, a pharmacologic substance or an organism, has brought about the effect. This includes induces, effects, evokes, and etiology
Complicates	Causes to become more severe or complex or results in adverse effects
Conducted	When a test is conducted to investigate a Disorder and the outcome is unknown/does not result in a diagnosis
Experiences	When a Living Being (e.g. patient) is affected by a Disorder, Sign or Symptom ; when a Living Being (e.g. patient) is subjected to a Medical Procedure
Interacts_with	Acts, functions, or operates together with
Measure_of	The quantitative or qualitative result of a medical procedure such as lab test or physical examination
Performs	A person conducts a procedure
Physically_related_to	Related by virtue of some physical attribute or characteristic
Prevents	Stops, hinders of eliminates an action or condition
Reveals	When a test is conducted and the outcome is known/leads to a diagnosis
Treats	Applies a remedy with the object of effecting a cure or managing a condition
Used_for	When a device is used, e.g. to conduct a treatment or to administer a drug

TABLE 2.3 – Schéma d'annotation pour les relations cliniques

Relation	Définition
HasAdministrationRoute	links a medication to its administration route
HasDosage	links a medication to its dosage
HasStrength	links a medication to its strength
HasDrugForm	links a medication to its form

TABLE 2.4 – Schéma d'annotation pour les relations concernant les prescriptions

Relation	**Définition**
Relations spatiales	
Localization_of	The spatial or relative localization of an entity
Location_of	The position, site, or region of an entity or the site of a process
Relations temporelles	
Before	An event precedes another event/temporal expression
Begins_on	The event starts on an event or temporal expression
During	The temporal span of an event is completely contained within the span of another event or temporal expression
Ends_on	The event finishes on an event or temporal expression
Overlap	An event happens almost at the same time, but not exactly, as another event/temporal expression
Simultaneous	An event happens at exactly the same time as another event/temporal expression
Relations aspectuelles	
Continue	Shows the continuation of an event
Decrease	A lowering value (e.g. of dose)
Improve	An improvement (e.g. in condition)
Increase	A rising value (e.g. of dose)
Recurrence_ StartAgain	Indicates that an event begins occurring again
Start	Indicates the initiation of an event
Stop	Indicates the ending of an event
Worsen	A negative change (e.g. in health)
Assertions	
Negation	An event is negated
Possible	An event may occur
Presence	An event occurs
SubjectToCondition	An event may occur on condition that another event occurs

TABLE 2.5 – Schéma d'annotation pour les relations spatiales, temporelles, aspectuelles et modalités

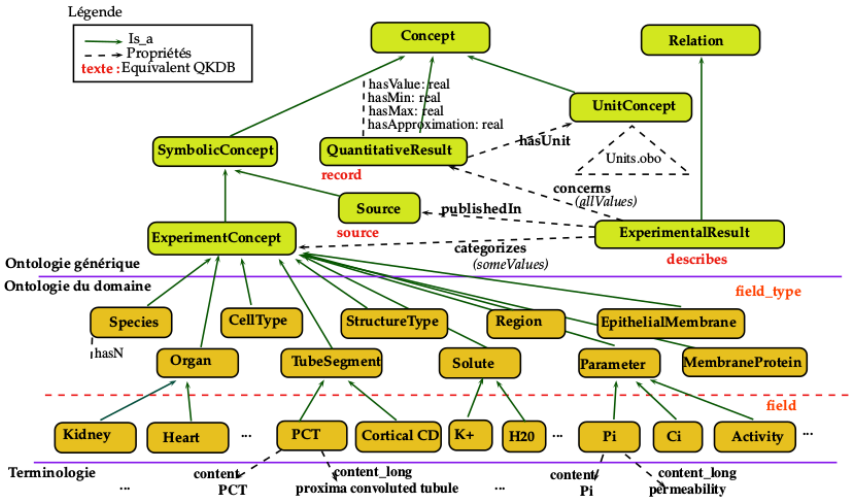


FIGURE 2.4 – Ressource termino-ontologique

à une expérience, contextualisé par les informations décrivant l'expérience. L'exemple 2.2.1, extrait d'un article scientifique, comprend une mention d'un résultat expérimental.

Exemple 2.2.1. Apical membrane P_f averaged (in cm/s) **9.37 +- 0.77 e-4** (n=5) at 20°C

Cet exemple indique que l'expérience consistait à mesurer la perméabilité (P_f) de la membrane apicale, à une température de 20°C sur 5 individus (l'espèce n'est pas précisée dans la phrase). Le résultat de cette expérience est exprimé en cm/s. L'ensemble de ces informations sera appelé résultat expérimental.

La formalisation que nous avons proposée est fondée sur une ressource termino-ontologique composée d'une ontologie générique, représentant des concepts généraux indépendants du domaine, d'une ontologie de domaine, et d'une terminologie, mettant en relation les concepts de l'ontologie avec leurs expressions dans la langue.

Un résultat expérimental est défini par un résultat quantitatif et les différents descripteurs de l'expérimentation qui ont permis de l'obtenir, et peut donc être vu comme une relation n-aire ; il sera donc représenté par un concept principal, relié à plusieurs autres concepts par des propriétés. Contrairement au travail [Touhami *et al.*, 2011] sur l'extraction de relations n-aires en microbiologie, nous souhaitions modéliser un résultat de façon indépendante du domaine, et traitons ainsi de la même façon tous les descripteurs d'un résultat. La figure 2.4 présente la ressource termino-ontologique définie.

Un résultat expérimental est représenté par un concept *ExperimentalResult*. Il est relié à sa valeur quantitative représentée par un concept *QualitativeResult*,

à ses descripteurs, correspondant à des *ExperimentConcept* et à la publication dont il est extrait *Source*. Ce niveau générique de l'ontologie décrit un résultat d'expérimentation, quelque soit le domaine concerné. Il est complété d'une ontologie du domaine, qui va décrit les concepts du domaine en deux niveaux : un premier niveau définit les catégories des concepts, le niveau des feuilles décrivant les instances des concepts. L'application visée relevant dans notre cas du domaine de la physiologie rénale, un concept peut par exemple être *Organ*, qui peut s'instancier en *Kidney*.

Enfin, la partie terminologique associe à chaque feuille un ensemble de termes qui correspondent à des mentions possibles du concept dans les textes, dans lesquelles sont distinguées un terme préféré (choisi par les experts du domaine lors de la construction de la terminologie), et les variantes de ce terme.

Cette première étape de formalisation est nécessaire pour que les informations puissent être extraites des textes, et être ensuite exploitées de façon cohérente. Si la formalisation n'est pas pré-existante, cette étape nécessite généralement la rédaction d'un guide d'annotation, l'annotation de textes et la consultation d'experts du domaine si les informations à extraire sont celles d'un domaine de spécialité. Les méthodes d'extraction d'information se fonderont alors sur cette représentation.

2.3 Méthodes d'extraction de relations

L'extraction de relations a fait l'objet de nombreux travaux, en domaine ouvert comme en domaine de spécialité, depuis les conférences MUC (*Message Understanding Conference*) organisées par la DARPA [7]. À MUC-7 en 1998 était introduite une tâche d'extraction de relations entre entités (relations *employee_ of,product_ of et location_ of*, qui a ensuite donné lieu à de nombreux travaux en extraction de relations sémantiques. Les conférences ACE du NIST ont ensuite pris le relais entre 1999 et 2008, suivies des conférences TAC (*Text Analysis Conference* [8]) depuis 2008, avec notamment la tâche KBP (*Knowledge Base Population*). Cette dernière tâche a pour objectif de remplir une base de connaissances à partir de textes, ce qui consiste à extraire des entités et leurs relations entre elles pour remplir les champs des entités (*slots*).

L'extraction de relations est généralement posée comme un problème de classification, la forme la plus simple étant une classification binaire : étant données deux entités et une relation, il s'agit de savoir si cette relation est vraie entre ces deux entités. La classification est cependant généralement multiclasses : il s'agit alors, étant donné un couple d'entités et un ensemble de relations, de déterminer quelle relation relie les entités, si elles sont en relation (voir figure 2.5). L'extraction de relations est parfois décomposée en deux étapes : la détection d'une relation, qui vise à déterminer si les deux entités sont en relation, et l'identifi-

7. *Defense Advanced Research Projects Agency*
8. http://www.nist.gov/tac/about/index.html

La carrière de *Cecilia Bartoli* débute en *1985*, à Rome : elle a dix-neuf ans et incarne la pétulante Rosina du « Barbier de Séville ».

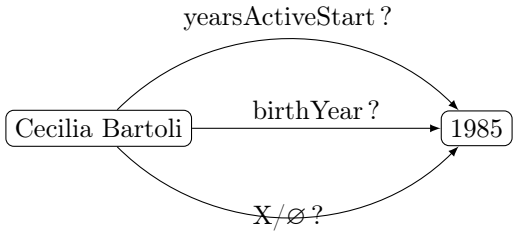


FIGURE 2.5 – Extraction de relations : définition

cation ou classification de la relation, qui rattache la relation trouvée à un jeu de relations donné.

Extraire des relations nécessite de définir plusieurs points :

— le point d'ancrage de la méthode : jeu de relations dans un cadre complètement supervisé, ou textes dans le cadre de l'extraction d'information ouverte ;
— la représentation des textes en entrée : texte brut, analysé morpho-syntaxiquement, syntaxiquement ;
— la mesure de similarité utilisée pour rapprocher les exemples à classifier ;
— l'empan considéré pour extraire les relations, le niveau le plus fréquent étant celui de la phrase ;
— la représentation du contexte local aux entités.

Dans le cas d'un jeu de relations restreint, les méthodes se sont appuyées sur un apprentissage supervisé exploitant des corpus annotés et des connaissances linguistiques de différents niveaux (syntaxiques et sémantiques). Les travaux correspondant à ces approches peuvent prendre en entrée une représentation vectorielle des textes (*feature-based methods*) correspondant à un certain nombre d'attributs [Kambhatla, 2004, Boschee *et al.*, 2005, Zhou *et al.*, 2005], ou une représentation syntaxique utilisant généralement des similarités de noyaux (*kernel-based methods*) [Roth et Yih, 2002, Zelenko *et al.*, 2003, Culotta et Sorensen, 2004].

Ce type d'approche nécessite cependant un grand nombre d'exemples pour chaque relation, ce qui peut être difficile à obtenir, notamment lorsque l'ensemble de relations est de taille importante, typiquement dans le cadre du web sémantique. Une autre difficulté des approches supervisées est que le jeu d'apprentissage est souvent très déséquilibré car le nombre d'exemples qui ne sont pas en relation est bien plus important que le nombre d'exemples qui sont en relation. En outre, l'adaptation à un nouveau domaine ou un nouveau type de corpus nécessite de réentraîner un modèle. Enfin, ces approches sont très sensibles aux performances des outils d'analyse.

Plus récemment, les méthodes supervisées se sont tournées vers les ré-

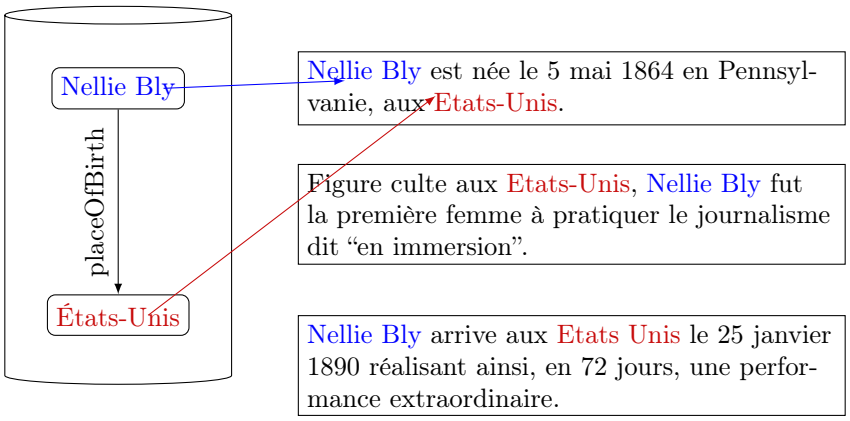


FIGURE 2.6 – Principe de la supervision distante (relations dbpedia)

seaux de neurones profonds [Zeng *et al.*, 2014, Nguyen et Grishman, 2015] afin de contourner plusieurs limitations des approches précédentes, et notamment la dépendance à des pré-traitements (étiquetage morpho-syntaxique et analyse syntaxique), dont les performances influent sur la classification, et qui perdent en performance lors d'un changement de domaine.

Afin de contourner la contrainte du nombre d'exemples annotés, problématique dans le cadre du web sémantique, de nombreux travaux se sont tournés vers de la supervision distante [Mintz *et al.*, 2009, Wu et Weld, 2007, Niu *et al.*, 2012], dont le principe consiste à annoter des textes en utilisant les relations présentes dans des bases de connaissance, donc de façon bruitée (voir figure 2.6). L'hypothèse sous-jacente est que toute phrase qui contient les entités participant à une relation connue, est susceptible d'exprimer cette relation. Ainsi, la base de données Wikidata par exemple contient la relation *notable work* entre *Virginia Woolf* et *Mrs Dalloway*, et il est donc possible d'annoter un corpus en exemples (vraisemblablement) positifs de la relation *notable work* lorsque les instances des deux entités sont repérées. Un autre avantage de ce type d'approche est son indépendance à un corpus, puisque tout type de texte peut en théorie être annoté ainsi. [Riedel *et al.*, 2010] ont cependant montré que la précision est dégradée si le corpus utilisé n'est pas directement lié à la base de connaissances. Comme le montre la figure 2.6, les phrases comprenant deux entités en relation ne correspondent pas toujours réellement à des mentions de cette relation.

[Riedel *et al.*, 2010] ont donc reformulé le problème de la supervision distante comme un problème d'apprentissage multi-instances (*multi-instance learning*), en relâchant les contraintes sur les instances de phrases trouvées, c'est-à-dire en supposant qu'au moins une des phrases contient une mention de la relation. Leur modèle améliore nettement la précision d'un modèle standard de supervision distante. Un inconvénient de ce modèle est qu'il ne permet pas de gérer les cas où des relations se recouvrent, comme *CEO* et *Founded* [Hoffmann *et al.*, 2011].

[Hoffmann *et al.*, 2011] étendent avec leur outil MULTIR l'approche précédente, en prenant en compte le fait que les phrases contenant deux entités données peuvent exprimer plusieurs relations possibles entre les deux entités, ou pas de relation. [Zhang *et al.*, 2012] ont montré qu'une grande taille du corpus permettait d'obtenir de meilleurs résultats, et le système DeepDive [Niu *et al.*, 2012] a été fondé sur ces travaux. Les travaux récents utilisant la supervision distante cherchent également à se passer des pré-traitements linguistiques en utilisant des réseaux de neurones profonds [Zeng *et al.*, 2015, Lin *et al.*, 2016]. Les performances de tels systèmes sont néanmoins limitées par la qualité des données d'apprentissage, et [Liu *et al.*, 2016] ont montré que l'ajout d'exemples annotés obtenus par myriadisation (*crowdsourcing*) améliorait les performances de plusieurs algorithmes de supervision distante, lorsque l'annotation est faite avec certaines contraintes (selon le protocole proposé dans l'article), pour la tâche TAC-KBP Slot Filling.

Un autre type d'approche, appelée extraction d'information ouverte (*Open Information Extraction)* cherche à extraire des relations en se fondant sur leur expression dans les textes plutôt que sur un ensemble prédéfini, partant ainsi des textes comme source première de connaissance, plutôt que de bases structurées [Banko *et al.*, 2007, Wu et Weld, 2010, Angeli *et al.*, 2015, Xu *et al.*, 2015b] (voir figure 2.7). [Banko *et al.*, 2007] ont introduit ce terme avec leur système TEXTRUNNER dont le principe est le suivant : un premier classifieur est entraîné sur un corpus analysé syntaxiquement, en s'appuyant sur des heuristiques pour extraire des relations, afin de déterminer si un tuple candidat a des chances d'être relié par une relation ; ensuite des couples d'entités sont annotés sur un grand corpus, et les mots compris entre les entités sont extraits puis simplifiés pour normalisation (par exemple les adverbes non nécessaires sont supprimés) ; ces relations sont enfin évaluées en fonction de leur nombre d'occurrences. [Wu et Weld, 2010] utilisent quant à eux les *infobox* de Wikipédia pour apprendre un système d'extraction de relations générique c'est-à-dire indépendant des relations exprimées par les infobox de Wikipédia. Les relations ainsi obtenues ne sont cependant pas normalisées et sont par conséquent difficiles à rattacher à un schéma de base de connaissance donné. Certains travaux se sont donc attachés à rapprocher les relations des bases de connaissances des relations obtenues par extraction d'information ouverte [Riedel *et al.*, 2013, Angeli *et al.*, 2015, Verga *et al.*, 2016] : [Angeli *et al.*, 2015] rapprochent les relations au schéma de KBP en cherchant des relations co-occurrant dans un corpus, tandis que [Riedel *et al.*, 2013, Verga *et al.*, 2016] cherchent à reconnaître des implications entre relations, afin d'inférer de nouvelles instances de relations.

Les méthodes supervisées restent néanmoins toujours pertinentes lorsque les relations considérées sont contextuelles, c'est-à-dire qu'il n'est pas possible d'exploiter la redondance textuelle au moment de l'apprentissage. C'est le type de méthode que nous avons appliqué dans nos travaux d'extraction de relation, que nous présentons dans les sections suivantes.

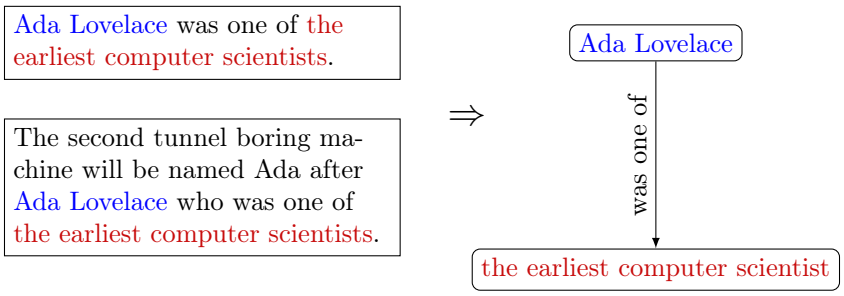


FIGURE 2.7 – Principe de base de l'extraction d'information ouverte (exemples de http://openie.allenai.org/)

2.3.1 Extraction de relations binaires

Dans le cadre de la thèse d'Anne-Lyse Minard [Minard, 2012], nous nous sommes intéressées à l'extraction de relations binaires dans un contexte biomédical. La tâche d'extraction de relations est posée comme une tâche de classification de couples d'entités, supposés connus : étant donné un couple d'entités (e_1, e_2), il s'agit de déterminer si ces entités sont reliées par l'une des relations données, et laquelle le cas échéant (les cas de *non-relation* peuvent correspondre à deux entités qui ne sont pas en relation, ou qui sont reliées par une autre relation que celles de la classification donnée).

Anne-Lyse a mis en place un système d'extraction de relations appelé REMED (*Relation Extraction in bio-Medical Domain)*, qui a été exploité pour étudier plusieurs questions : quels types d'information sont les plus utiles à l'extraction des relations, et en particulier quelles représentations syntaxiques ; et les méthodes de simplification de phrases peuvent-elles améliorer les performances de l'extraction ?

Ces études ont été menées sur plusieurs corpus de relations. Le cadre d'application principal était celui de l'évaluation internationale i2b2 en 2010, dont l'une des tâches consistait à repérer des relations entre concepts dans des comptes rendus médicaux [Uzuner *et al.*, 2011]. Ces concepts étaient soit des *problèmes médicaux* (maladies, symptômes...), soit des *traitements* (médicaments, interventions...), soit des *tests* (procédures, mesures...) et les relations reliaient deux concepts : ainsi, la relation TrIP (*Treatment improves problem*) indiquait par exemple que le traitement administré au patient a amélioré son problème. La liste des relations est présentée dans le tableau 2.6. Les corpus étaient constitués de comptes rendus hospitaliers provenant de plusieurs centres médicaux aux États-Unis.

Dans un premier temps, nous avons utilisé un classifieur SVM et une représentation sous forme de vecteurs d'attributs de nos données. Les étapes du système sont les suivantes :

— prétraitements (notamment analyse morpho-syntaxique) des fichiers d'entrée et annotation des concepts ;

Traitement-Problème	
TrIP	Le traitement améliore le problème
TrWP	Le traitement aggrave le problème
TrCP	Le traitement cause le problème
TrAP	Le traitement est administré en raison du problème
TrNAP	Le traitement n'est pas administré en raison du problème
Test-Problème	
TeRP	Le test révèle le problème
TeCP	Le test est conduit en raison du problème
Problème-Problème	
PIP	Un problème en indique un autre

TABLE 2.6 – Relations i2b2 2010

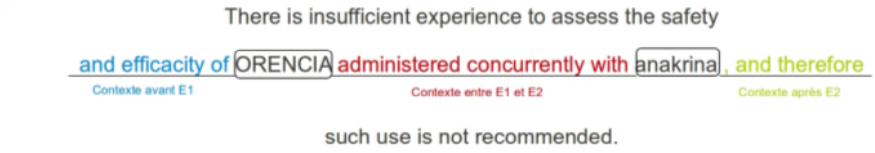


FIGURE 2.8 – Contexte local d'une relation

— traitement des coordinations et extraction des attributs ;
— classification.

Nous avons observé que dans le corpus i2b2, de nombreuses phrases contenaient des énumérations de concepts médicaux, ce qui pouvait éloigner les entités en relation et poser problème pour l'extraction des relations. Nous avons donc ajouté un prétraitement pour supprimer les entités coordonnées avec une des deux entités candidates à la relation, via un système de règles.

Afin d'extraire les attributs nécessaires à la classification, nous avons séparé le contexte des entités en trois, suivant [Zhou *et al.*, 2005, Roberts *et al.*, 2008] : mots avant la première entité, mots après la seconde (fenêtre de 3 mots), et mots entre les deux entités (voir figure 2.8). Les attributs utilisés sont de quatre ordres : attributs de surface (position des entités, ordre, distance...), lexicaux (mots et lemmes des entités et du contexte), morpho-syntaxiques (catégories morpho-syntaxiques, présence d'une préposition ou de ponctuation) et sémantiques (type UMLS, types de concepts, classes des verbes du contexte...). Des attributs donnent également des informations sur la présence d'une coordination préalable. Une description détaillée du système peut être trouvée dans la thèse d'Anne-Lyse [Minard, 2012] et dans [Minard *et al.*, 2011c].

Les différentes étapes et attributs ont été évalués. Ainsi, le module de gestion de la coordination permet d'améliorer légèrement la précision du système, lorsque les attributs de coordination sont ajoutés, tandis que les attributs morpho-syntaxiques permettent d'améliorer son rappel. Il est intéressant

de noter que les différentes classes d'attributs n'ont pas la même incidence sur les résultats en fonction de la relation considérée. Une méthode de sélection des attributs a donc été utilisée, en évaluant le caractère discriminant de chaque type d'attribut pour chaque classe.

Le système REMED a été classé 3e lors de l'évaluation i2b2, avec une f-mesure de 0,707 sur le corpus d'évaluation pour l'ensemble des relations, le meilleur système ayant obtenu 0,736 de f-mesure. L'un des inconvénients principaux de REMED est qu'il n'arrive pas à classifier les relations présentant peu d'exemples sur le corpus d'entraînement (notamment la relation TrWP), ce qui a été compensé par l'utilisation de patrons lors de l'évaluation [Grouin *et al.*, 2010a, Minard *et al.*, 2011a].

En outre, la représentation des phrases contenant les relations par des traits de surface ne permet pas de bien capturer des relations entre termes distants. Nous avons par conséquent étudié l'apport de traits syntaxiques pour la reconnaissance des relations [Minard *et al.*, 2011b], en utilisant les représentations sous forme d'arbre de constituants, et d'arbre de dépendances. L'ajout de ces informations sous forme vectorielle n'améliore pas les performances. En revanche, l'utilisation des arbres syntaxiques et de mesures de similarité adaptées (*tree kernels*) permet d'améliorer les performances du système REMED, et en particulier l'association des arbres complets, des sous-arbres minimaux et des attributs. L'étude des résultats permet de confirmer que les relations supplémentaires détectées correspondent bien à des phrases dans lesquelles les entités sont éloignées. REMED a également été testé sur deux autres corpus, avec des résultats au niveau de l'état de l'art : le corpus DDI, dans lequel il faut détecter si deux médicaments sont en interaction ou non [Minard *et al.*, 2011d] ; et les corpus PPI, où il s'agit d'identifier les paires de protéines ou gènes qui interagissent. Pour ces corpus, l'utilisation d'informations syntaxiques sous forme vectorielle améliore les performances du système, contrairement aux résultats obtenus sur le corpus i2b2. Ceci peut probablement s'expliquer par le fait que les corpus DDI et PPI sont constitués de résumés d'articles scientifiques, sur lesquels l'analyse syntaxique est de meilleure qualité que sur des comptes rendus médicaux, dans lesquels les phrases ne sont pas toujours bien construites et comportent de nombreuses énumérations et abréviations. L'extraction des relations sous cette forme est donc très sensible à la qualité des prétraitements.

La prise en compte de la structure syntaxique n'étant pas suffisante pour résoudre le problème de la variabilité d'expression des relations, nous avons également étudié si des méthodes de simplification de phrases permettait d'améliorer la classification, notamment pour les relations peu représentées dans le corpus. Quatre méthodes de simplification ont été testées (voir [Minard *et al.*, 2012] pour une présentation détaillée des méthodes et de leurs résultats) :

— la première méthode repose sur l'annotation d'entités telles que les informations posologiques pour normalisation, la suppression de segments de phrases supposés inutiles pour la reconnaissance de la relation, comme par exemple les mots précédant la première entité et séparés par une virgule et la suppression des entités en coordination avec une entité considérée ;

— la seconde méthode utilise un outil existant, bioSimplify [Jonnalagadda et Gonzalez, 2010], avec une phase de post-traitement, l'outil n'ayant pas été développé pour la tâche d'extraction de relations ;
— la troisième méthode est fondée sur des règles de suppression appliquées sur les arbres de constituants ;
— la dernière méthode est fondée sur un apprentissage CRF annotant les informations considérées comme inutiles pour l'extraction de relations.

Ces méthodes ne permettent pas d'améliorer les performances de REMED, mais les relations correctement classées diffèrent avec et sans simplification, et la combinaison des prédictions du système avec et sans simplification permet d'améliorer la f-mesure de 1,5%.

Le système d'extraction de relations que nous avons développé obtient donc de bonnes performances globales. Cependant, les systèmes même actuels nécessitent un grand nombre d'exemples pour apprendre à reconnaître une relation, ce qui est problématique dans le cadre médical, où les corpus sont rares. Les relations considérées sont en outre assez restreintes, et il sera intéressant de voir comment se comportent les méthodes sur un ensemble de relations plus complet tel que celui défini dans le projet Cabernet.

2.3.2 Extraction de relations n-aires

Toujours dans le cadre de la thèse d'Anne-Lyse Minard, nous avons mis en place une méthode d'extraction de résultats expérimentaux. Les informations extraites dans ce cadre sont des relations n-aires, également appelées événements, par exemple entre le résultat et son unité, la précision, l'organe concerné, le soluté éventuel (voir section 2.2.2 pour la représentation des résultats). L'extraction d'événements peut être traitée comme une tâche de classification de termes comprenant notamment le déclencheur de l'événement, puis de mise en relation de ces termes [Ahn, 2006, Buyko *et al.*, 2009, Riedel *et al.*, 2009, Grouin *et al.*, 2010b, Björne *et al.*, 2010, Patrick et Li, 2010, McClosky *et al.*, 2011, Huang *et al.*, 2016], ou inversement de détection de relations binaires, puis de reconstruction de l'événement complet à partir des relations [McDonald *et al.*, 2005].

Dans le cadre considéré, et contrairement à la plupart des tâches existantes en extraction d'événement, les informations à extraire ne sont pas reliées à la mention de l'événement, mais à l'expression d'un résultat quantitatif. En outre, ces informations sont majoritairement exprimées par des termes du domaine, et non des entités nommées.

Une des originalités de notre approche est la prise en compte de tous les types de résultats expérimentaux, recherchés dans l'article complet, y compris dans les tableaux, et non limités à une phrase (contrairement à [Corney *et al.*, 2004, Garten et Altman, 2009] notamment).

La méthode proposée se décompose en trois étapes :
— détection d'une valeur numérique (exemples donnés figure 2.9), qui joue le rôle de déclencheur et constituera l'élément pivot du résultat. Cette

détection s'apparente à une reconnaissance d'entité numérique. Deux méthodes ont été évaluées : une extraction à base de règles, expressions régulières fondées sur la présence d'un nombre et d'une précision ; et une annotation par apprentissage supervisé, fondée sur les caractéristiques du nombre et de son contexte. D'autres attributs sont également reconnus à cette étape : le nombre d'animaux concernés par l'expérience, la précision du résultat, et l'unité ;

— reconnaissance de descripteurs d'expérience. Il s'agit alors de faire le lien entre les concepts de l'ontologie et les termes de l'article. Pour cela, la terminologie est utilisée, ainsi que des variantes flexionnelles, dérivationnelles et syntaxiques ;

— mise en relation des attributs ou descripteurs et de la valeur numérique (figure 2.10). Cette mise en relation prend en compte la distance entre la valeur numérique du résultat et les attributs ou descripteurs, ainsi que des critères de fréquence.

The urinary Ca2+ concentration of the knockout mice reached values as high as 20 mM, compared with 6 mM for TRPV5+/+ littermates.

Apical membrane Pf averaged (in cm/s) 9.37 +- 0.77 e-4 (n = 5) at 20°C, and two values obtained at 37°C were 33.7 and 33.2 e-4 cm/s.

Parameter	Group 1 (n = 7)	Group 2 (n = 5)	Groupe 3 (n = 5)
Body weight (g)	24.8 +- 0.47	23.5 +- 0.48	26.1 +- 0.8

FIGURE 2.9 – Exemples de valeurs numériques à extraire

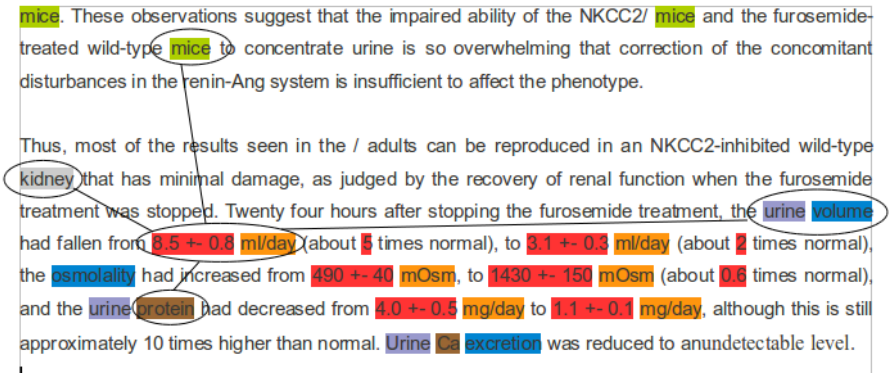


FIGURE 2.10 – Exemples de mise en relation des attributs ou descripteurs et de la valeur numérique

La méthode s'appuyant sur la terminologie associée à l'ontologie, il était

nécessaire de disposer d'un ensemble de termes le plus complet possible. Pour cela, nous avons utilisé un certain nombre de ressources spécifiques au domaine, comme une liste d'unités de base et composées (issue de Gene Ontology). Nous avons également appliqué une méthode d'acquisition de termes à partir de corpus [Minard *et al.*, 2010] afin de compléter la terminologie.

Le modèle termino-ontologique est ensuite représenté par la base de données QKDB qui permet de stocker les instances trouvées dans les articles.

Le système complet d'extraction des résultats expérimentaux a été évalué de façon intrinsèque et extrinsèque.

L'évaluation intrinsèque a été effectuée sur un corpus de test de 15 articles, ce qui représente 855 résultats expérimentaux. Une partie de cette évaluation est présentée dans le tableau 2.7.

Le système final utilise la détection des valeurs numériques par règles dont le rappel est de 1 (f-mesure de 0,77) : nous avons privilégié la méthode ayant le meilleur rappel, afin que dans l'outil final, les utilisateurs n'oublient pas de résultat, suivant notamment les résultats de [Alex *et al.*, 2008] qui ont montré dans leurs expériences de curation que les curateurs préféraient une annotation la plus complète possible. La méthode apprentissage présente quant à elle un rappel de 0,93 (f-mesure de 0,85) : seuls certains types de résultats prototypiques sont reconnus, le nombre d'exemples étant probablement trop faible pour le reste des cas.

L'extraction des résultats expérimentaux globale obtient une f-mesure de 0,61, avec un rappel de 0,75 et une précision de 0,51. Si l'on évalue l'extraction et mise en relation des attributs et descripteurs uniquement pour les résultats expérimentaux correctement extraits, la f-mesure est de 0,78.

Ces résultats sont au niveau de l'état de l'art, si l'on se réfère à la tâche similaire de i2b2 3009 portant sur des termes médicaux. La précision reste bonne, ce qui signifie que le nombre de résultats proposés ne constituera pas un bruit trop important pour les curateurs.

	Rappel	Précision	F-mesure
Évaluation générale	0,75	0,51	0,61
Résultats quantitatifs	1,00	0,63	0,77
Attributs et descripteurs pour les résultats quantitatifs corrects	0,75	0,81	0,78

TABLE 2.7 – Évaluation de l'extraction des résultats expérimentaux

Nous avons également évalué l'apport du système d'extraction d'information pour la curation de la base QKDB, en intégrant le système d'extraction dans une interface web (voir figure 2.11). Cette évaluation a été menée auprès de 5 experts, qui ont annoté chacun trois articles dans trois conditions différentes : à partir de l'annotation de référence, avec les annotations de notre système d'extraction, et manuellement. Les experts ont passé en moyenne 105s par résultat avec notre système contre 174s manuellement, et ont annoté presque deux fois plus de résultats. L'assistant facilite donc bien la tâche d'annotation, et assure

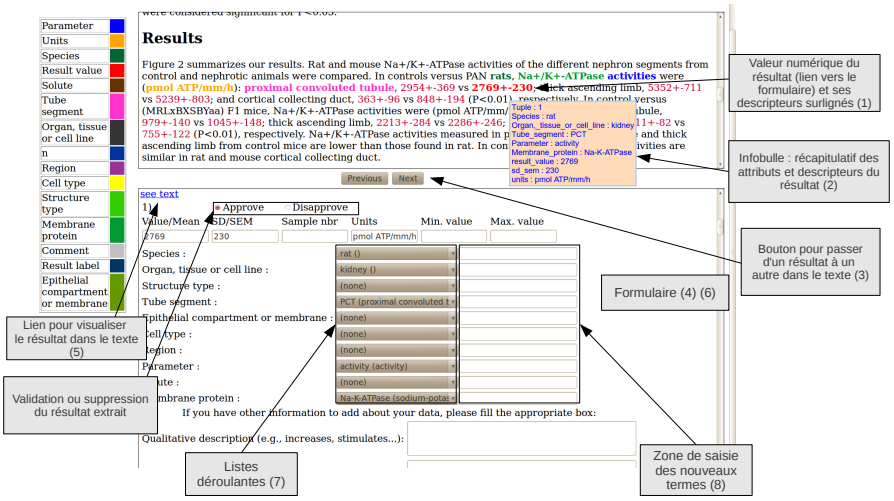


FIGURE 2.11 – Interface test de curation de QKDB

une meilleure qualité des annotations.

Nous avons ainsi mis en œuvre une méthode d'extraction de relations n-aires, par complétion de la terminologie associée à une ontologie, puis utilisation de règles fondées sur la reconnaissance des termes de la Ressource Termino-Ontologique, et leur contexte, fréquence et position.

Ce système est pour l'instant spécifique au domaine de la physiologie rénale, mais sa construction a été pensée comme générique, et il serait intéressant de voir comment il s'adapte à d'autres domaines.

D'un point de vue de l'extraction de relations, la méthode utilisée ici est largement fondée sur la reconnaissance des termes de la ressource et leurs variantes flexionnelles, dérivationnelles et syntaxiques. Chaque résultat est extrait de façon indépendante des autres, mais l'extraction conjointe des résultats pourrait certainement permettre d'augmenter les performances de la méthode.

Après avoir abordé l'extraction de relations en tant que telle, pour des tâches de compréhension ciblée de textes, nous nous intéressons maintenant à son rôle dans une application de recherche d'information précise, la réponse à des questions.

2.4 Réponse à des questions

Nous présentons dans cette section nos travaux en cours sur les systèmes de réponse à des questions, que nous abordons ici comme un cas d'application de l'extraction de relations.

La présentation de la problématique de cette section reprend en partie l'article [Grau *et al.*, 2015] écrit avec Martin Gleize et Brigitte Grau.

La recherche d'information couvre un vaste éventail de tâches, dans le but de

répondre à des besoins utilisateurs différents. Lorsqu'il s'agit de rechercher des documents, ce besoin est généralement exprimé par une liste de mots-clés, tandis que pour rechercher une information précise, concernant un fait ou une entité, ce besoin peut être formulé simplement par une question en langage naturel, comme par exemple « Dans quelle ville est né l'assassin de Martin Luther King ? ». Les systèmes de réponse à des questions, ou systèmes de questions-réponses (QR) cherchent à répondre à ce type de besoin d'information précise, dans des sources de connaissances qui peuvent être structurées (bases de connaissances) ou textuelles (corpus de documents textuels).

La nature des informations recherchées conduit à développer des processus de recherche différents.

La recherche de réponses dans des textes est fondée sur un appariement ou une similarité entre la question et les extraits de texte sélectionnés constituant des passages réponses. Les informations extraites de la question permettent de formuler des contraintes pour cet appariement ou d'établir des critères de similarité.

La recherche de réponses dans une base de connaissances suppose la reconnaissance et l'identification des entités et des relations qui les lient, pour construire une requête en un langage formel et interroger la base. L'ensemble des relations, celui du schéma de la base, est alors connu. Les systèmes exploitent, comparent ou transforment des représentations sémantiques structurées sous forme de graphes d'entités et de relations pour produire une requête dans un langage formel.

On voit ainsi apparaître deux domaines de recherche à chaque extrémité, qui sont la recherche d'information et l'interrogation de bases de connaissances du Web sémantique, les deux s'appuyant sur une analyse de la langue. Néanmoins, le problème peut être posé sous un même point de vue, conduisant à une formalisation plus unifiée : les textes peuvent être vus comme des bases de connaissances, dans la mesure où ils contiennent des relations entre éléments d'information, qui peuvent être obtenues par analyse syntaxique et sémantique. Cela est d'autant plus justifié que les systèmes de QR sur du texte qui s'appuient sur une représentation structurée des phrases sont souvent plus performants.

Je me suis intéressée aux systèmes de questions-réponses sur le texte depuis ma thèse de doctorat, et j'ai contribué à ce domaine depuis concernant notamment l'évaluation de type boîte transparente [El Ayari *et al.*, 2009], l'analyse des questions [El Ayari *et al.*, 2009, Ligozat, 2013], la validation de réponses [Grappy *et al.*, 2011], et l'étude de l'apport des informations morphologiques [Ligozat *et al.*, 2012a, Tribout *et al.*, 2012, Ligozat *et al.*, 2012b]. Plus récemment, la thèse de Romain Beaumont, que je co-encadre avec Brigitte Grau, étudie les possibilités d'exploiter les deux types de ressources. Cette voie de recherche établit un pont entre les travaux en extraction de relations et ceux sur la reconnaissance de paraphrase ou plus généralement d'implication textuelle, dans la lignée de [Angeli *et al.*, 2015] par exemple. Nous avons proposé dans [Grau *et al.*, 2015] une mise en parallèle des deux tâches, visant à définir un formalisme de représentation des informations applicable dans les deux tâches et permettant la recherche hybride de réponses. Une question peut en effet être

vue comme un ensemble de mentions de relation ($m_i r_j$) entre mentions d'entités ($m_i e_j$). La réponse recherchée est une entité particulière indiquée par ? et certaines relations ou entités peuvent ne pas être explicites (x) (voir figure 2.12).

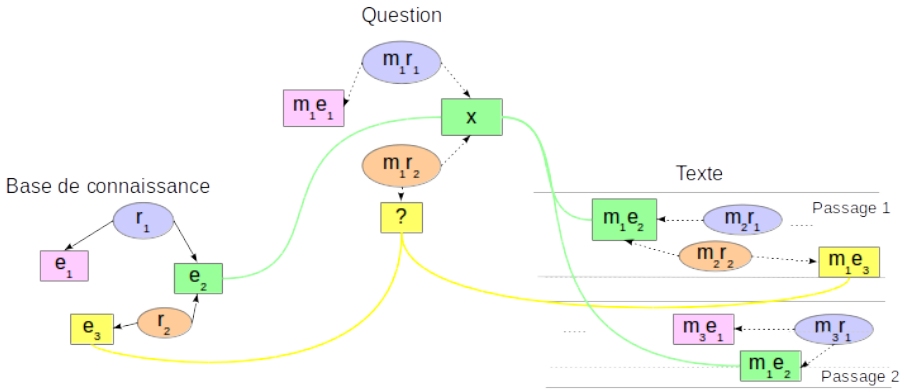


FIGURE 2.12 – Recherche de réponses dans une base de connaissances ou dans des textes

Les systèmes de QR suivent généralement l'architecture présentée figure 2.13.

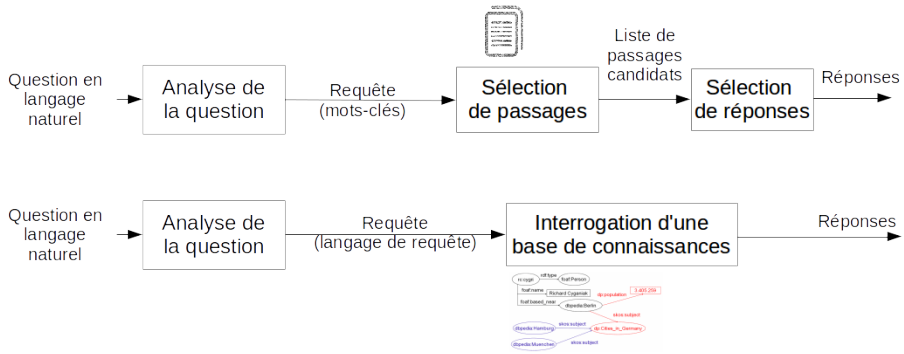


FIGURE 2.13 – Architecture de recherche de réponses à partir de questions en langage naturel

Quelle que soit la ressource interrogée, la première étape consiste à analyser la question en langage naturel afin d'en tirer toutes les informations exploitées par les processus de recherche de passages et de sélection de réponses.

Ensuite, l'interrogation de textes pose le problème de comparer une représentation comportant des mentions de relation et entités de la question avec différentes mentions de ces mêmes entités et relations, présentes dans les passages pertinents. Puisque l'on ne cherche pas forcément à identifier les relations sémantiques présentes dans les énoncés, mais à mettre en relation deux énoncés et donc reconnaître des paraphrases, la définition de ce qui est mention de

relation et entité n'est pas assujettie à une représentation des connaissances explicite. De plus, la même information est souvent donnée sous différentes formes, ce qui ne rend pas toujours nécessaire de reconnaître toutes les variantes d'une même information pour trouver une réponse car généralement on ne recherche pas toutes ses occurrences.

L'interrogation d'une base de connaissances pose le problème de la transformation d'une structure comportant des mentions de relation et d'entité de la question vers une structure, *e.g.* un graphe, formée de l'instanciation partielle des relations de la base. Le problème des variations lexicales entre labels associés aux entités et relations de la base et les termes employés par l'utilisateur se pose alors, puisque ce dernier n'est pas guidé par la connaissance du schéma de la base. Se pose également le problème de la résolution d'ambiguïtés sémantiques, car un même terme peut faire référence à différents éléments sémantiques. Cependant, la représentation formelle des relations fournit des contraintes sémantiques pour reconnaître les arguments d'une relation, et permet de concevoir un processus d'identification des instances de relation plus informé.

Illustrons ces processus avec la question suivante : « Who is the daughter of Bill Clinton married to ? » La réponse peut être trouvée dans les passages suivants :

> On 31 July 2010, 30-year-old *Chelsea Clinton* (the daughter of former U.S. president Bill Clinton and Secretary of State Hillary Clinton), who had recently received a master's degree from Columbia University's Joseph L. Mailman School of Public Health, married 32-year-old *Marc Mezvinsky*, an investment banker.

Dans ce premier passage, les mots de la question se retrouvent à l'identique, mais en revanche ils sont séparés par deux incises, ce qui rend la structure syntaxique nécessaire pour les relier.

> Famously protective of her private life, the 32-year-old daughter of Bill and Hillary Clinton has long stayed silent on her relationship with husband *Marc Mezvinsky*.

Dans ce second passage, la relation de mariage est exprimée cette fois par une variation sémantique, le nom « husband », à relier au mot « married » de la question, et le nom de Bill Clinton est associé à celui de sa femme, ce qui rend son identification plus complexe. En outre, cet exemple montre qu'il n'est pas nécessaire d'identifier la fille de Bill Clinton pour trouver la réponse.

La réponse peut aussi être trouvée dans DBpedia par la requête indiquée dans la figure 2.14. Afin de former cette requête, il faut repérer l'existence d'une entité implicite correspondant à Chelsea Clinton (?a dans la requête), relier l'entité Bill Clinton à la ressource correspondante, et identifier les relations « spouse », correspondant à « married », et « child » correspondant à « daughter ». Il faut également pouvoir relier ces différentes ressources entre elles pour former les relations, puis la requête complète.

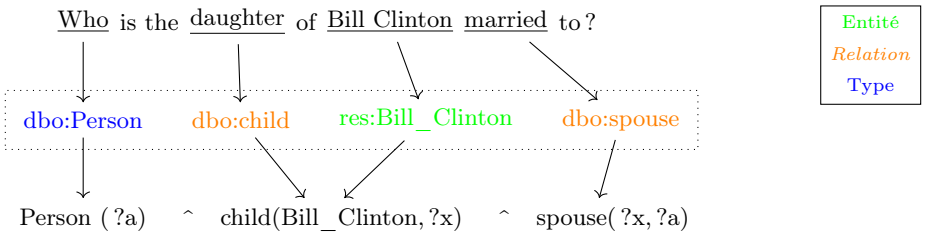


FIGURE 2.14 – Analyse de la question et construction de la requête

2.4.1 Méthodes de réponses à des questions

La recherche de réponses dans des textes procède à une sélection de passages dans les documents interrogés afin de restreindre l'espace de recherche, en mettant en œuvre des méthodes de recherche d'information. Un passage de texte est pertinent s'il contient l'information donnée dans la question et la réponse courte attendue. Généralement, cette information n'est pas exprimée dans les mêmes termes que ceux de la question et différents types de variations linguistiques sont à traiter entre la question et les passages. Le problème de sélection d'une réponse peut être modélisé par une implication textuelle de la forme $T => H$, dont nous rappelons que la signification est la suivante [Glickman, 2006] : un lecteur humain lisant T peut raisonnablement en déduire H. Dans le cas de la recherche de réponses, le texte T est un passage sélectionné, et l'hypothèse H correspond à une formulation déclarative de la question dans laquelle la place de la réponse attendue est marquée et éventuellement instanciée par la réponse que l'on cherche à valider. La recherche de réponses dans des textes s'est principalement focalisée sur la sélection de passages ces dernières années, laissant donc de côté l'extraction de la réponse (à l'exception par exemple de [Yao *et al.*, 2013b] qui utilisent des CRF pour étiqueter la réponse). La sélection de passages est généralement traitée comme un problème de reconnaissance d'implication textuelle, et abordée avec différents types d'algorithmes d'alignement [Yao *et al.*, 2013a, Yih *et al.*, 2013, Yao *et al.*, 2013b].

Les travaux récents côté bases de connaissance sont fondés sur la reconnaissance de relations, utilisant des méthodes proches de celles d'extraction de relations présentées précédemment. Deux types d'approches sont possibles : les approches du type analyse sémantique [Berant *et al.*, 2013], qui visent à apprendre une représentation sémantique de la question, et nécessitent généralement un grand nombre d'exemples annotés ; et les approches s'appuyant sur l'extraction d'information. Dans ces dernières, la question est alors principalement considérée comme une relation entre deux entités : une entité sur laquelle porte la question, et l'entité réponse. Selon les systèmes, cette modélisation peut être un peu étendue : la relation permettant de trouver la réponse peut être une relation n-aire [Yih *et al.*, 2015] ; ou la question peut comporter deux relations portant sur la même entité [Xu *et al.*, 2016]. Par rapport

à la représentation générale présentée précédemment, l'analyse de la question se focalise sur une entité principale, qui se rapproche de notre définition du focus d'une question [El Ayari *et al.*, 2009, El Ayari *et al.*, 2010]. Les autres informations de la question peuvent être considérées comme des contraintes supplémentaires [Yih *et al.*, 2015]. Certains systèmes (comme [Fader *et al.*, 2013]) s'appuient sur des bases de connaissances construites par extraction d'information ouverte, ce qui les rapproche de travaux sur le texte.

Ce type d'approche ne permet pas de traiter des questions complexes, dans lesquelles plusieurs relations seraient réellement nécessaires pour répondre, telles que « Who is the mother of the father of Prince William ? ».

Une autre limitation des systèmes interrogeant les bases de connaissances est que la base de connaissances peut ne pas contenir l'ensemble des informations requises par la question, nécessitant ainsi une recherche hybride.

2.4.2 Intégrer ressources textuelles et structurées

Certains travaux ont cherché à exploiter ces deux types de ressources.

Une première possibilité d'hybridation des deux types de systèmes consiste à rechercher des réponses dans des textes et dans des bases de connaissances en parallèle, puis à fusionner les réponses obtenues dans chacune des ressources [Hildebrandt *et al.*, 2004, Cucerzan et Agichtein, 2005]. [Clarke *et al.*, 2002] commencent par rechercher une réponse dans des données structurées, puis, si aucune réponse n'y est trouvée, recherchent des informations dans des textes. La recherche de réponses se fait cependant dans une ressource d'un seul type.

[Katz *et al.*, 2005] décomposent les questions et ressources, structurées ou semi-structurées, afin d'obtenir une représentation sémantique commune, leur permettant de rechercher la réponse dans les deux types de ressources. Cependant, la robustesse de cette méthode n'est pas discutée, et le système n'est pas évalué.

Plus récemment, [Yahya *et al.*, 2013] ont mis en œuvre des techniques d'extension et de relaxation de requêtes afin de rechercher des informations dans des descriptions textuelles associées aux triplets : la première technique identifie les termes de la question qui ne sont pas reliés à un triplet dans la requête générée, et les ajoute comme mots-clés ; la deuxième identifie les triplets ou ensembles de triplets de la requête qui n'ont pas de correspondance dans la base, et les termes correspondant aux triplets menant à une absence de résultat sont ajoutés comme mots-clés ; enfin, la troisième technique, utilisée en dernier recours, consiste à considérer une simple requête pour vérifier le type de réponse attendu trouvé par l'analyse de la question, et ajouter les autres termes comme mots-clés associés. Les réponses ainsi obtenues sont ensuite réordonnées en fonction des entités saillantes et des mots-clés pertinents. L'utilisation des descriptions textuelles permet de passer d'un MRR de 0,53 à 0,72 sur le jeu de questions factuelles de QALD2, et d'obtenir une f-mesure de 0,68 (contre 0,74 pour le meilleur système participant). [Usbeck et Ngomo, 2015] ont une approche assez voisine, fondée sur la génération de requêtes hybrides : si aucune ressource n'est

associée à un nœud du graphe de la question, ce nœud est ajouté à la requête sous forme textuelle.

[Park *et al.*, 2015] au contraire, recherchent d'abord des informations correspondant à la décomposition d'une question dans des textes annotés en entités de la base de connaissances, et complètent la recherche textuelle par des requêtes SPARQL si la recherche textuelle est infructueuse. Ce système obtient une bonne précision mais traite peu de questions sur le jeu de données hybrides de QALD5. De même, [Xu *et al.*, 2016] ont un système permettant de rechercher la ou les réponses dans une base de connaissances, et valident les réponses candidates en s'appuyant sur les pages Wikipédia.

Le système Watson qui a participé à Jeopardy ! [Chu-Carroll *et al.*, 2012] effectue également une recherche hybride, mais en sélectionnant les relations à rechercher dans les bases de connaissances plutôt que comme stratégie de secours. Il utilise des bases de connaissances, soit existantes comme DBPedia, soit propres au système : la ressource PRISMATIC a ainsi été construite à partir de relations syntaxiques ou sémantiques extraites de textes. Les relations contenues dans ces bases sont exploitées de deux façons : la recherche de réponses lorsque l'une des relations Freebase les plus fréquentes est détectée dans la question, et la vérification du type de la réponse.

Cependant, chacun de ces systèmes garde une ressource principale, et fait appel à l'autre type de ressource uniquement pour la validation, ou en cas d'échec, ce qui n'en fait pas réellement des systèmes hybrides.

De fait, les deux types de ressources ne contiennent pas les mêmes types d'information et dans [Grau *et al.*, 2015] nous avons examiné différentes caractéristiques des questions sous l'angle d'une résolution hybride. Les questions sont des questions factuelles dans les deux types de ressources, mais toutes ne peuvent profiter des informations provenant d'une base de connaissances. Cependant, lorsque la question mentionne une entité, une hybridation sera possible. Nous avons donc annoté en entités nommées un corpus de questions sur des textes comportant 9 227 questions. De manière à évaluer le taux d'erreur, nous avons vérifié l'annotation sur un échantillon de 200 questions. 59 % des questions comportent au moins une entité et entrent donc de fait dans les cas d'hybridation. Celle-ci peut fournir des informations intermédiaires nécessaires à la résolution de la question ou fournir la réponse, et cela selon les types d'information cherchés :

- la réponse est une caractéristique de l'entité (relation directe) : « Qui est le mari de la fille de Bill Clinton ? » La réponse peut être cherchée dans l'une ou l'autre source ;
- la réponse porte sur un événement, soit un rôle ou le nom de l'événement : « Qui est l'assassin de Martin Luther King ? » La réponse proviendra de textes, et d'autant plus si l'événement fait intervenir des entités non connues ;
- une combinaison des deux, par exemple une relation directe d'une entité ayant un rôle dans un événement (composition de relations) « Où est né l'assassin de Martin Luther King ? » On pourra effectuer une recherche hybride ;

— la réponse est une instance de concept ou un concept : « Quel animal pond des œufs bleus ? » La réponse viendra du texte, avec par exemple « Les Collonca sont sans queue et pondent des œufs bleus. », et la vérification que les Collonca sont des animaux pourra être opérée sur l'une ou l'autre source ;
— la relation avec la réponse est contextuelle (opinion ou liée à un événement par exemple comme dans « Quel pays a acheté du pétrole à l'Irak durant l'embargo ? ») Cette dernière ne pourra être trouvée que dans le texte ;
— une définition : « Qu'est-ce qu'un atome ? » La réponse est dans le texte ;
— un résultat provenant d'un opérateur d'agrégation (comparaison, classement, comptage) : « Donnez les dix plus grandes entreprises françaises. » Les réponses peuvent provenir de textes, dès lors que l'information cherchée est explicite, mais elles seront plus aisément déduites d'une base de connaissances.

Un intérêt supplémentaire à hybrider vient du fait que la langue et le schéma de la base ne font pas état du même niveau de granularité. Un mot peut faire référence à un chemin dans une base de connaissances et rendre la recherche dans cette base très complexe, comme pour la question « Donnez le nom d'un cosmonaute. » où « cosmonaute » doit être apparié à « astronaute de nationalité russe ». Inversement, le texte peut ajouter des informations, *i.e.* des contraintes, permettant de choisir la bonne entité parmi les possibles, par exemple « Quelle chanson des Rolling Stones parle de la fin d'une histoire d'amour ? » (réponse : *Angie*).

Des systèmes hybrides exploitant à la fois les bases de connaissances et les textes seraient donc plus performants.

2.5 Discussion

Nous avons créé des représentations adaptées pour deux tâches d'extraction de relation, des relations binaires dans le domaine médical [9], et des relations n-aires dans le domaine biologique. Dans les deux cas, nous avons essayé de créer une représentation la plus générique possible, mais nous n'avons pas encore eu d'opportunité de les utiliser dans d'autres domaines que ceux étudiés initialement, et j'espère que des projets futurs le permettront.

Nous avons également mis en œuvre deux méthodes d'extraction de relations : le système Remed, qui nous a permis d'étudier l'apport de différents types de connaissances, et notamment des informations syntaxiques ; et l'outil de curation de QKDB, pour lequel nous avons étudié l'acquisition de termes pour la complétion de la terminologie sur laquelle est fondée la méthode.

Concernant la réponse à des questions, la recherche d'informations hybride, c'est-à-dire dans des bases de connaissances et des textes, est une piste qui a peu été explorée : les systèmes hybrides actuels s'appuient principalement sur une source alors que la coopération des deux types de sources devrait être

9. Le schéma d'annotation et la méthodologie de constitution du corpus annoté ont fait l'objet de deux articles qui ont été soumis dans des revues.

exploitée. La thèse de Romain devrait aboutir à une représentation des questions permettant une réelle recherche hybride.

Chapitre 3

Simplification textuelle

3.1 Motivations

L'objectif de la simplification de textes est de réécrire un texte en entrée afin qu'il soit plus facile à lire. La problématique de la simplification de textes cherche à répondre aux difficultés rencontrées par certains lecteurs lors de la lecture de textes, qu'il s'agisse d'apprenants de langue maternelle (enfants [De Belder et Moens, 2010], lecteurs avec un faible niveau d'alphabétisation [Watanabe *et al.*, 2009], lecteurs non experts [Elhadad et Sutaria, 2007]), d'apprenants de langue seconde ([Petersen et Ostendorf, 2007]) ou de personnes avec un trouble ou une pathologie liée au langage, comme les dyslexiques [Rello *et al.*, 2013b], les aphasiques [Carroll *et al.*, 1999, Canning *et al.*, 2000, Max, 2006], ou les sourds [Inui *et al.*, 2003] (voir [Feng, 2008] ou [Siddharthan, 2014] pour des explications détaillées sur les difficultés rencontrées par chaque type de lecteur). Nous laisserons ici de côté les applications au domaine médical car elles sont un peu en-dehors de notre étude, mais il est à noter que le même type de problématique se retrouve dans le cadre de l'accès à des textes médicaux (publications, comptes rendus hospitaliers...) pour le grand public (voir par exemple [Severance et Cohen, 2015] ou [Bouamor *et al.*, 2016] pour une présentation du problème de l'accès à l'information médicale pour les patients ou le grand public).

Il peut ainsi être utile de pouvoir évaluer *a priori* la difficulté d'un texte, et si besoin de le simplifier. Il est néanmoins nécessaire de tenir compte du fait que dans une situation d'apprentissage, des textes trop simples risquent de limiter l'apprentissage, tandis que des textes trop complexes risquent de conduire à une mauvaise compréhension du contenu et à une perte de motivation. Cependant, plusieurs travaux ont montré que l'utilisation de textes simplifiés pouvait rendre les textes plus faciles à lire, sans perte d'information : [Rello *et al.*, 2013a] notamment ont montré que l'utilisation de mots plus fréquents améliore la compréhension des documents par des lecteurs dyslexiques. [Ziegler *et al.*, 2015] ont

également mené des tests de lecture auprès de dix enfants dyslexiques en leur présentant des textes standards ou manuellement simplifiés, et ont montré que la lecture est plus rapide et la compréhension meilleure lorsque les enfants ont lu les textes simplifiés.

Signalons enfin que la simplification a également été employée comme prétraitement afin d'améliorer les performances d'autres applications de traitement automatique des langues, comme l'analyse syntaxique [Chandrasekar *et al.*, 1996], l'extraction de relations [Miwa *et al.*, 2010] ou l'étiquetage en rôles sémantiques [Vickrey et Koller, 2008].

3.2 Définitions

Les notions de lisibilité, complexité et difficulté d'un texte ont fait l'objet de nombreux travaux, notamment en psychologie cognitive et en pédagogie. La notion de complexité linguistique recouvre des acceptations vastes (voir par exemple [Blache, 2011] pour une discussion du terme de complexité), mais nous nous intéresserons ici uniquement à la complexité de lecture de textes pour un groupe de lecteurs donné, soit ce que [Blache, 2011] appelle la complexité relative ou difficulté linguistique.

La lisibilité peut être définie comme l'évaluation de cette complexité, c'est-à-dire de la difficulté de lecture d'un texte pour une classe d'individus. Le lecteur pourra se rapporter à la thèse de Thomas François [François, 2012] pour une présentation détaillée du point du vue psycholinguistique sur le processus de lecture en langues première et seconde, ainsi que des travaux sur la lisibilité, des premières formules à l'utilisation de la linguistique computationnelle (nous n'aborderons pas ici l'ensemble des travaux en lisibilité, mais ferons éventuellement référence à certains travaux lorsque nécessaire).

La tâche de simplification de textes consiste, à partir d'un texte donné en entrée, à produire une version plus simple de ce texte, c'est-à-dire présentant une meilleure lisibilité pour un groupe de lecteurs donné, tout en préservant au maximum son sens et son contenu informationnel. Les tâches d'évaluation de la lisibilité et de simplification sont donc liées, et les travaux sur la lisibilité peuvent être utilisés en simplification, notamment pour déterminer des critères de simplification, et pour l'évaluation des méthodes de simplification.

Un exemple de texte simplifié est présenté ci-dessous.

> Comptant parmi les plus grands compositeurs du XIXe siècle, tant par l'importance et l'étendue de son répertoire que par sa qualité, son nom se rattache surtout à l'opéra, l'un des plus populaires étant — encore de nos jours — Il barbiere di Siviglia (d'après Le Barbier de Séville de Beaumarchais).
>
> C'est un compositeur du XIXème siècle. Il est notamment connu pour avoir composé l'opéra Le Barbier de Séville, adapté de la pièce de Beaumarchais.

(exemple - un peu raccourci - issu des pages Vikidia/Wikipédia sur Gioachino Rossini, consultées le 13 juin 2016)

Remarquons que cet exemple comprend plusieurs types de simplification : la phrase du texte d'origine est divisée en deux, certaines informations pouvant être considérées comme secondaires sont supprimées (les raisons de l'importance de Rossini), certaines expressions sont paraphrasées pour employer des termes plus simples («son nom se rattache surtout à» devient «Il est notamment connu pour»), le titre d'origine du Barbier de Séville n'est pas donné...

3.3 Modélisation de la simplification de textes

Remarquons tout d'abord que bien que le terme de «simplification textuelle» soit largement utilisé, il est en réalité souvent un peu abusif, car la simplification est généralement appliquée aux phrases du texte et non pas à l'ensemble du texte. Appliquée au niveau des phrases donc, la simplification textuelle est généralement vue comme un processus comprenant trois opérations : découpage, d'une phrase en plusieurs phrases plus courtes ; suppression ou réordonnancement de mots, constituants ou phrases ; et substitution d'un mot ou d'une expression par un équivalent plus simple [Zhu *et al.*, 2010, Narayan et Gardent, 2014].

Afin d'étudier plus précisément les phénomènes impliqués dans le processus de simplification, nous avons souhaité nous appuyer sur un corpus de textes parallèles afin d'établir une typologie précise.

3.3.1 Corpus

Bien que la simplification automatique de textes ait fait l'objet de nombreux travaux, il existe peu d'études des types de simplification fondées sur corpus. La simplification est généralement vue comme un processus impliquant des phénomènes lexicaux, et des phénomènes syntaxiques [Carroll *et al.*, 1999, Inui *et al.*, 2003, De Belder et Moens, 2010] ; peu de travaux ont étudié les aspects discursifs [Siddharthan, 2006].

Afin d'établir une typologie des phénomènes concernés pour le français, nous avons constitué un corpus de textes parallèles en français. Ce travail, ainsi que le système de simplification en résultant, ont été menés en collaboration avec Laetitia Brouwers, Thomas François et Delphine Bernhard, et sont présentés dans [Brouwers *et al.*, 2014] et [Brouwers *et al.*, 2012].

Corpus existants

Les travaux récents en simplification se sont beaucoup appuyés sur des corpus parallèles, et notamment celui composé des versions en anglais et en anglais simplifié de Wikipédia, constitué par [Zhu *et al.*, 2010]. La version en «Simple English» de Wikipédia est destinée aux enfants et adultes qui apprennent l'anglais, et les consignes de rédaction de ses articles indiquent de préférer les mots

simples et des phrases courtes. Ce corpus contient plus de 65 000 articles parallèles, reliés entre eux dans Wikipédia comme correspondant au même article dans deux langues différentes. Sa taille et sa disponibilité en ont fait le jeu de données de référence des travaux de simplification pour l'anglais.

Ce corpus a récemment fait l'objet de discussions par [Xu *et al.*, 2015a], qui ont montré qu'il était à utiliser avec précaution car relativement bruité. [Coster et Kauchak, 2011b] avaient déjà montré que les articles parallèles des deux versions ne contiennent que peu de phrases alignables, les articles en anglais simplifié n'étant généralement pas une traduction directe des articles en anglais standard. En outre, dans les phrases alignables, beaucoup sont identiques (27% de leur résultat d'alignement). [Xu *et al.*, 2015a] ont appronfondi cette étude par une annotation manuelle de 200 alignements de phrases choisies aléatoirement. Ils ont montré que seuls 50% des paires de phrases correspondent en réalité à une simplification réelle (17% ne constituent pas une réelle paire parallèle, et dans les 33% des cas restants, la phrase de Wikipédia en Simple English n'est pas plus simple que l'originale). En outre, les 50% de «réelle simplification» comprennent par exemple des phrases avec seulement un mot de moins que l'originale, mais pas réellement plus simples.

Une autre difficulté de la constitution de corpus pour la simplification automatique vient du fait de considérer ce processus comme indépendant du public visé. Or en réalité, les difficultés de lecture des textes sont dépendantes du type de public visé, et varient même au sein d'un même groupe de lecteurs, des apprenants d'une langue par exemple, en fonction de leur niveau dans la langue. La constitution d'un corpus parallèle nécessite donc toujours de spécifier quel est le public visé, et les systèmes appris un corpus particulier ne sont adaptés qu'au public visé par le corpus.

Afin de répondre au problème du bruit présent dans le corpus Wikipédia anglais, [Xu *et al.*, 2015a] proposent l'utilisation du corpus Newsela, constitué de 1 130 textes journalistiques simplifiés par des éditeurs professionnels pour quatre niveaux scolaires différents. Leur analyse de corpus montre par exemple que la taille du vocabulaire diminue fortement entre les deux versions analysées du corpus Newsela (50% de réduction), alors qu'elle ne varie que de 18% dans le corpus Wikipédia. Les caractéristiques linguistiques du corpus, ainsi que la présence de quatre versions correspondant aux niveaux dans la langue des lecteurs visés, en feraient un corpus plus adapté pour les systèmes de simplification automatique.

Un corpus similaire avait également été créé pour l'espagnol dans le cadre du projet Simplext [Bott *et al.*, 2012] comprenant 200 articles de journaux en espagnol, ainsi que leur version simplifiée par des éditeurs professionnels.

Corpus pour le français

Concernant le français, il n'existe pas pour le moment de corpus constitué par des éditeurs professionnels, bien qu'un tel corpus soit en cours de création (travaux de Núria Gala) adapté à des enfants dyslexiques. Nous avons donc créé

un corpus en utilisant des ressources existantes [1].

Notre corpus est composé de deux types de textes : des textes informatifs, issus des encyclopédies Wikipédia et Vikidia ; et des textes narratifs, contes de Perrault, Maupassant et Daudet, accompagnés de leur version simplifiée à destination des apprenants du français.

Afin de créer le premier sous-corpus, appelé corpus *Wiki* dans la suite, nous avons collecté les articles de Vikidia et ceux de même titre dans Wikipédia, grâce à l'API MediaWiki. L'outil WikiExtractor [2] a ensuite été appliqué aux articles, afin de supprimer la syntaxe wiki et de ne conserver que le contenu textuel principal des articles. Le corpus ainsi constitué contient 13 638 textes (7 460 de Vikidia et 6 178 de Wikipédia, certains articles de Vikidia n'ayant pas d'équivalent côté Wikipédia).

Nous avons ensuite utilisé l'algorithme d'alignement monolingue décrit dans [Nelken et Shieber, 2006] afin d'extraire des couples de phrases exprimant les mêmes informations. Cet algorithme est simplement fondé sur une mesure cosinus entre deux vecteurs de mots pondérés par leur *tf.idf* représentant les phrases. Un exemple de paire de phrases parallèles est présenté ci-dessous.

> (Wikipédia) Ses parents s'y opposant, il choisit de suivre des études de lettres, qui l'amèneront finalement à la profession d'enseignant.
> (Vikidia) Mais ses parents s'y opposent et il devient enseignant après des études de lettres.

Les alignements produits sont accompagnés d'un score de confiance. Afin de mener une étude de corpus, nous avons sélectionné 20 articles, ce qui représente 72 phrases de Wikipédia et 80 phrases de Vikidia.

Le second sous-corpus, appelé corpus *Contes* dans la suite, contient 16 textes narratifs, en l'occurrence des contes, de Perrault, Maupassant et Daudet. Nous avons utilisé des contes car leur version simplifiée est plus proche de l'originale que dans le cas de romans plus longs, ce qui facilite l'alignement. Les versions simplifiées des contes proviennent de deux collections destinées aux apprenants du français, et leurs niveaux de difficulté vont de A1 à B1 sur l'échelle CERL (Cadre européen commun de référence pour les langues - Apprendre, Enseigner, Évaluer). Nous avons procédé à un alignement manuel de 3 textes par deux annotateurs, ce qui correspond à 83 phrases dans les textes originaux, et 98 dans les versions simplifiées, et constitue un corpus d'une taille comparable à celui des encyclopédies en ligne.

Étude du corpus français

Nous avons souhaité étudier ces corpus d'un point de vue lexical, afin de vérifier l'hypothèse que la simplification de textes implique bien également une simplification au niveau du lexique. Cette étude a fait l'objet du stage d'Anaïs

1. Le travail de constitution du corpus encyclopédique a fait l'objet d'un stage de M1 par Antoine Sylvain en 2011, co-encadré par Delphine Bernhard et moi-même.
2. https ://github.com/attardi/wikiextractor

	contes originaux	contes simplifiés	Wikipédia	Vikidia
moyenne	2 745	1 839	2 594	156

TABLE 3.1 – Longueur des textes en nombre de mots

Tack en 2014, et a été prolongée avec Anaïs, Thomas François et Cédrick Fairon dans les publications [Tack *et al.*, 2016b] et [Tack *et al.*, 2016a].

Afin de comparer la complexité lexicale des textes originaux et simplifiés, nous avons utilisé deux ressources lexicales graduées, renseignant sur le niveau d'apprentissage d'apprenants du français pour un terme donné : Manulex et FLELex. Notre objectif était de vérifier la complexité lexicale comparée des deux versions du corpus, ainsi que d'évaluer l'utilisation de ces ressources lexicales pour cette vérification.

Le lexique Manulex [Lété *et al.*, 2004] a été créé à partir de manuels scolaires destinés à des enfants francophones et donne la fréquence des mots et des lemmes dans trois niveaux des manuels : CP, CE1 et cycle 3 (CE2 à CM2). La version constituée des formes fléchies contient 48 900 entrées, tandis que la version constituée des lemmes en comprend 23 900. Dans cette étude, nous avons utilisé les lemmes et la fréquence observée simple.

Contrairement à Manulex, FLELex [François *et al.*, 2014] a été établi à partir de manuels de français langue étrangère, répartis selon les six niveaux du Cadre européen commun de référence pour les langues (CECRL), à savoir les niveaux A1, A2, B1, B2, C1 et C2. Ce lexique reporte les fréquences de chaque mot (avec son étiquette morpho-syntaxique) pour chacun des niveaux. Deux versions existent également, établies avec un étiqueteur CRF spécifique et le TreeTagger. Nous avons utilisé la première version, qui contient plus de lemmes, incorpore des expressions multi-mots et est plus précise. Suivant [Gala *et al.*, 2014], nous avons dans la suite considéré que le niveau d'un mot était le niveau le plus bas dans lequel il apparaît.

Pour cette étude, nous avons utilisé le corpus *Contes* complet et une sélection du corpus *Wiki* [3]. Dans un premier temps, nous avons constaté qu'il y avait une différence très nette entre la longueur des textes originaux et simplifiés et que cet écart se marquait avant tout dans le corpus Wiki (tableau 3.1). Dans la suite, nous avons donc normalisé les fréquences pour chaque version des deux sous-corpus afin de réduire l'impact de la différence de longueur.

Nous avons ensuite annoté les corpus avec les entrées de Manulex et Flelex afin d'étudier la répartition des différents niveaux lexicaux dans les corpus.

Le tableau 3.2 présente les fréquences des niveaux Manulex.

Dans toutes les versions, le niveau le plus fréquent est le niveau G1, ce qui est logique car les mots les plus fréquents sont les mots grammaticaux, présents dès le niveau G1. En revanche, lorsque l'on considère les fréquences relatives au sein des trois niveaux, on constate que le niveau le plus bas est plus fréquent dans les versions simplifiées, tandis que les deux autres niveaux sont plus fréquents

3. Voir [Tack, 2014] et [Tack *et al.*, 2016a] pour plus de détails sur la sélection et l'annotation des corpus.

Version		G1	G2	G3-5
contes originaux	% dans version	94,6	3,1	2,3
	% dans niveau	49,1	70,5	76,7
contes simplifiés	% dans version	98,0	1,3	0,7
	% dans niveau	50,9	29,5	23,3
Wikipédia	% dans version	84,1	7,8	8,1
	% dans niveau	49,3	52,7	55,1
Vikidia	% dans version	86,4	7,1	6,6
	% dans niveau	50,7	47,3	44,9

TABLE 3.2 – Fréquences des niveaux Manulex dans les corpus Contes et Wiki

Version		A	B	C
contes originaux	% dans version	96,1	3,5	0,4
	% dans niveau	49,3	77,8	100,0
contes simplifiés	% dans version	99,0	1,0	0,0
	% dans niveau	50,7	22,2	0,0
Wikipédia	% dans version	92,8	5,8	1,4
	% dans niveau	49,4	57,4	63,6
Vikidia	% dans version	94,9	4,3	0,8
	% dans niveau	50,6	42,6	36,4

TABLE 3.3 – Fréquences des niveaux FLELex dans les corpus Contes et Wiki

dans les versions originales.

Concernant l'annotation par les niveaux de FLELex, les six niveaux ont été regroupés en trois niveaux (A1-A2, B1-B2, C1-C2) afin de pouvoir appliquer des tests statistiques, les fréquences des trois niveaux les plus élevés étant souvent trop faibles pour pouvoir les appliquer sinon. Le tableau 3.3 présente les résultats obtenus.

Ces résultats semblent à nouveau confirmer notre hypothèse : les corpus originaux comptent plus de mots qui relèvent du niveau intermédiaire que les corpus simplifiés, ces derniers ne comprenant que peu ou pas de mots du niveau avancé.

Enfin, nous avons également étudié les fréquences d'apparition pour chaque catégorie grammaticale pour les mots lexicaux (adjectifs, adverbes, noms, verbes). Les résultats (détaillés dans [Tack, 2014]) montrent que les mots lexicaux et en particulier les noms et verbes sont simplifiés entre les versions originales et les versions simplifiées.

L'analyse de la complexité lexicale du corpus confirme donc notre hypothèse que la simplification de textes mène également à une simplification du lexique. Ces observations nous donnent également des caractéristiques générales de notre corpus sur le plan lexical.

3.3.2 Typologie de simplification

Nous avons également observé manuellement les phrases alignées de notre corpus afin d'en dégager les différents phénomènes intervenant dans la simplification de texte. Si les consignes de rédaction de textes simples ou de simplification de textes sont nombreuses, nous souhaitions établir une typologie précise des phénomènes occurrant en corpus. L'analyse des phrases alignées de notre corpus a permis d'établir une typologie fondée sur les niveaux linguistiques lexical, discursif et syntaxique. Cette typologie est résumée dans le tableau 3.4.

Lexical	**Discursif**	**Syntaxique**
Traduction	Réorganisation	Temps
Anaphore	Ajout	Modification
Définition et paraphrase	Suppression	Regroupement
Synonyme ou hyperonyme	Cohérence and cohésion	Suppression
	Personnalisation	Découpage

TABLE 3.4 – Typologie des simplifications

Concernant les aspects lexicaux, les phénomènes observés relèvent de quatre types de substitution différents :

— des termes jugés difficiles peuvent être remplacés par un synonyme ou un hyperonyme ;
— certaines expressions anaphoriques, considérées comme plus simples ou plus explicites, sont préférées à l'anaphore originale. Ainsi, des anaphores nominales sont régulièrement utilisées à la place d'anaphores pronominales, notamment dans les contes ;
— des termes jugés difficiles peuvent également être remplacés par une définition ou une paraphrase explicative ;
— dans le cas où le texte original contient des mots dans une langue étrangère, ces mots sont généralement traduits dans la version simplifiée.

Sur le plan discursif, les auteurs des textes simplifiés s'attachent à rendre l'organisation de l'information la plus claire et concise possible.

— à cet effet, des clauses peuvent être échangées afin que la présentation de l'information soit plus lisible ;
— des informations d'importance secondaire peuvent être supprimées ou au contraire, des exemples ou explications peuvent être ajoutés ;
— une attention particulière est portée à la cohérence et cohésion du texte : ainsi, les auteurs explicitent souvent les relations entres les phrases ;
— enfin, les structures personnelles sont souvent modifiées.

Sur le plan syntaxique, les types de modification suivants sont observés :

— les temps utilisés dans les textes simplifiés sont plus communs et moins littéraires que dans les textes originaux ; ainsi, le présent et le passé composé sont préférés au passé simple, à l'imparfait et au plus-que-parfait ;
— des informations secondaires ou redondantes peuvent être supprimées des phrases, notamment des clauses adverbiales ou subordonnées ;

— lorsque des structures complexes ne sont pas supprimées, elles sont généralement déplacées pour faciliter la compréhension ; c'est souvent le cas par exemple pour des phrases négatives, des structures impersonnelles, du discours indirect et des propositions subordonnées ;
— les auteurs choisissent parfois de diviser des phrases trop longues, ou au contraire de regrouper des phrases proches (souvent après les avoir simplifiées).

Cette typologie est proche par exemple de celle de [Medero et Ostendorf, 2011], qui proposent trois catégories de phénomènes syntaxiques (division, suppression et extension) et de celle de [Zhu *et al.*, 2010], qui identifient des phénomènes de division, suppression, réorganisation et substitution, mais les affine et fait apparaître d'autres phénomènes (tel que la traduction des mots étrangers).

Les trois niveaux ne sont bien entendu pas indépendants les uns des autres, puisque la substitution d'un mot par une paraphrase peut par exemple entraîner un changement syntaxique.

3.4 Méthodes

Le processus de simplification peut être considéré comme un processus de reconnaissance de paraphrase dirigée ou d'implication textuelle, avec une contrainte de lisibilité sur le texte simplifié. La relation de simplification est donc asymétrique, contrairement à la paraphrase, mais se rapproche ainsi de l'implication textuelle puisque le texte original implique le texte simplifié. Certains travaux sur la simplification se rapprochent donc de travaux sur l'acquisition ou la reconnaissance de paraphrases, notamment pour les aspects lexicaux [Yatskar *et al.*, 2010, Biran *et al.*, 2011]. [Yatskar *et al.*, 2010] et [Pavlick et Nenkova, 2015] présentent la dimension de simplicité comme une dimension de style, au sens où il s'agit de présenter la même information différemment pour s'adapter à un public visé. La base de paraphrases PPDB a d'ailleurs été annotée automatiquement avec une dimension de complexité [Pavlick *et al.*, 2015].

La simplification textuelle dans son ensemble peut également être posée comme un problème de traduction automatique monolingue, et plusieurs travaux ont exploité des méthodes issues de la traduction, utilisant des modèles n-grams [Coster et Kauchak, 2011a, Wubben *et al.*, 2012], des arbres syntaxiques [Zhu *et al.*, 2010] ou des représentations sémantiques [Narayan et Gardent, 2014]. Des modèles spécifiques à la tâche de simplification sont néanmoins nécessaires pour tenir compte des phénomènes de suppression et de découpage, peu fréquents en traduction bilingue [Narayan et Gardent, 2014]. [Zhu *et al.*, 2010] est le premier article à avoir posé le problème de la simplification comme un problème de traduction monolingue, les auteurs exploitant le corpus parallèle PWKP issu des Wikipédia en anglais et anglais simple. Leur méthode est fondée sur les opérations de découpage, suppression, réordonnancement et substitution s'ap-

pliquant aux arbres de constituants des phrases. [Coster et Kauchak, 2011a, Wubben *et al.*, 2012, Narayan et Gardent, 2014] utilisent également des modèles de traduction automatique statistique en prenant en compte les particularités de la problématique de la simplification, notamment la suppression de mots pour [Coster et Kauchak, 2011a], le découpage [Narayan et Gardent, 2014] et la dissimilarité pour [Wubben *et al.*, 2012].

Enfin, la simplification peut également être rapprochée de la compression de textes [Coster et Kauchak, 2011b], utilisée notamment en résumé automatique, qui correspond à prendre en compte en particulier les opérations de suppressions difficiles à intégrer dans les modèles précédents, mais ne prend pas en compte les reformulations.

Deux sous-tâches peuvent être dégagées, qui ne sont pas nécessairement traitées à part dans les systèmes, mais ont fait l'objet d'études et d'évaluations spécifiques : la simplification lexicale, qui vise à remplacer certains mots considérés comme trop complexes par des équivalents plus simples, et la simplification syntaxique, où seuls les aspects syntaxiques sont pris en compte.

3.4.1 Aspects lexicaux de la simplification

Certains travaux se sont concentrés uniquement sur les aspects lexicaux de la simplification, c'est-à-dire à l'opération de substitution évoquée précédemment, qui s'applique généralement à une sous-partie très restreinte de la phrase, que nous appellerons «mot cible» dans la suite, mais qui peut en réalité être constitué d'un terme multi-mots, d'une expression ou d'un syntagme.

En 2012, la tâche de substitution lexicale de la campagne d'évaluation SemEval 2012 [Specia *et al.*, 2012] a permis une certaine normalisation de la tâche, ainsi que la mise à disposition d'un corpus pour la simplification lexicale. L'objectif était le suivant : étant donné un contexte (une phrase), un mot cible donné dans cette phrase, et un ensemble de mots équivalents au mot cible en contexte, appelés substituts (ensemble issu de la campagne de désambiguisation lexicale de SemEval 2007), il s'agissait de classer les mots équivalents par ordre décroissant de simplicité. L'exemple ci-dessous présente un contexte, avec le mot cible qu'il convient (potentiellement) de simplifier, ainsi que la liste des substituts, triés par ordre décroissant de simplicité.

> **Contexte** : The US ranks 37th in a World Health Organization *examination* of the world's health care systems.
> **Mot cible** : *examination*
> **Substituts** (par ordre décroissant de simplicité) : *study, investigation, examination, inspection, scrutiny*

Le corpus d'environ 200 mots cibles présentés chacun dans 10 contextes différents avait été annoté par des locuteurs non natifs de l'anglais. Un inconvénient de ce corpus, qui montre aussi la difficulté de la tâche, est que l'accord interannotateur était bas (κ [4] de 0,39 environ) ; les annotations de référence ont par

4. adapté à la comparaison paire à paire pour une tâche de traduction automatique [Callison-Burch *et al.*, 2011]

conséquent été générées en effectuant une moyenne des annotations. Cette évaluation a permis de rendre compte de la difficulté de la tâche, et de justifier de s'intéresser en particulier à ces aspects lexicaux puisque les systèmes participants [5] n'ont pas dépassé la baseline fondée sur la fréquence des mots dans le corpus Google Web IT Corpus (κ de 0,47).

[Paetzold et Specia, 2016] ont quant à eux repris les corpus LSeval [De Belder et Moens, 2012] (fondé sur le même corpus de SemEval 2007 que le précédent) et LexMTurk [Horn *et al.*, 2014] (fondé sur des phrases alignées des Wikipédia en anglais et anglais simple puis annotation via Mechanical turk), mais les ont fait annoter à nouveau par 400 locuteurs non natifs de l'anglais afin qu'ils indiquent s'ils considéraient les mots cibles de ces corpus comme complexes. Le corpus final, NNSeval, ne contient que les 239 mots cibles jugés comme complexes, et les substituts jugés comme non complexes, ce qui fournit un corpus d'évaluation un peu différent de celui de SemEval.

Prise dans son ensemble, la tâche de simplification lexicale comporte les processus suivants : identification des mots complexes (*complex word identification*), génération des substituts (*substitution generation*), sélection des substituts en tenant compte du contexte (*substitution selection*), et ordonnancement des substituts (*substitution ranking*) [Paetzold et Specia, 2015, Paetzold et Specia, 2016]. L'identification des mots complexes sera discutée plus loin mais dans la plupart des travaux en simplification lexicale, elle est supposée donnée par le corpus en entrée [Paetzold et Specia, 2016] ou issue de l'alignement d'un corpus parallèle [Biran *et al.*, 2011, Horn *et al.*, 2014].

Concernant la génération des substituts, l'approche classique consiste à utiliser des synonymes, hyperonymes ou paraphrases issues de bases lexicales [Carroll *et al.*, 1998, De Belder et Moens, 2010]. Les substituts peuvent également être extraits de corpus parallèles [Biran *et al.*, 2011, Yatskar *et al.*, 2010] ou de représentations vectorielles des mots [Paetzold et Specia, 2016]. [Yatskar *et al.*, 2010] ont ainsi acquis des simplifications lexicales («collaborate» → «work together») à partir des révisions des pages Wikipédia rédigées en anglais simplifié. Les paires extraites par ces modèles ont une bonne précision dans les premiers rangs, mais une faible couverture, et ne tiennent pas compte du contexte.

La sélection de substituts contextuellement valables se rapproche d'une tâche de désambiguisation, et est généralement traitée soit directement comme telle [De Belder et Moens, 2010], soit comme une comparaison des contextes d'apprentissage et du mot cible [Biran *et al.*, 2011] soit intégrée à l'ordonnancement des substituts [Horn *et al.*, 2014].

Le coeur de la simplification lexicale est cependant l'ordonnancement des substituts, qui est généralement considérée comme une tâche d'apprentissage d'ordonnancement, avec différents classifieurs (SVM et boundary ranker notamment). La complexité lexicale de chaque substitut est alors estimée en utilisant les mêmes types de critères qu'en lisibilité computationnelle, car il s'agit de

5. Nous mettons de côté le système de [Jauhar et Specia, 2012] qui obtenu des résultats supérieurs à la baseline, mais a en partie été développé par les organisateurs de l'évaluation.

mesurer le même phénomène, bien que l'objectif final soit ici de comparer la complexité de deux mots, et non d'estimer la lisibilité ou la classe de lisibilité d'un mot unique.

Ces critères peuvent être de plusieurs ordres [Gala *et al.*, 2014, Pavlick et Callison-Burch, 2016] :

— critères (ortho)graphiques : longueur du mot en nombre de lettres, de phonèmes ou de syllabes, complexité orthographique (nombre et fréquence des voisins orthographiques, cohérence phonème-graphie, présence de graphèmes complexes) structure syllabique, n-grams de caractères ;

— critères morphologiques : nombre et fréquences d'affixes, composition, taille de la famille morphologique, étiquette morpho-syntaxique (en particulier dans le cas d'expressions) ;

— critères sémantiques : polysémie, représentation vectorielle (*word embeddings*) ;

— critères statistiques : fréquence dans divers lexiques et corpus, notamment supposés contenir des termes simples, fréquences comparées (Simple Wikipedia vs. Wikipedia).

Le dernier type de critère est en réalité un peu différent puisque l'on peut supposer qu'il tient compte des critères des autres types, mais les informations statistiques ne peuvent être utilisées seules car elles sont en partie spécifiques à un corpus et ne permettent pas de généraliser suffisamment, même en présence d'un grand corpus. [Gala *et al.*, 2014] montrent en effet (sur le français) que la fréquence dans un corpus de sous-titres de films (Lexique3) est la variable la plus corrélée au niveau de complexité lexicale (estimé avec les données de Manulex et FLELex), mais que l'ajout des autres variables permet néanmoins un gain de performance. [Pavlick et Callison-Burch, 2016] montrent également que l'utilisation de divers critères apporte un gain de performance non négligeable (5 points de précision supplémentaires par rapport au modèle n'utilisant que les word embeddings) pour une tâche de simplification lexicale.

La fréquence des mots reste cependant le critère principal utilisé pour la simplification lexicale, fréquence obtenue via une base lexicale, comme la base psycholinguistique MRC[6] ou calculée par une analyse en corpus. L'hypothèse sous-jacente est qu'un mot est d'autant plus simple qu'il apparaît fréquemment en corpus. Parmi les premiers travaux, on peut citer ceux de [Carroll *et al.*, 1998], qui comprennent un module de simplification lexicale interrogeant la base de données Oxford Psycholinguistic Database afin de classer les synonymes WordNet selon leur fréquence de Kucera-Francis. Plus récemment, [De Belder et Moens, 2010] utilisent la fréquence du substitut en corpus. Cette fréquence peut être étendue à la fréquence du n-gram comprenant le substitut et son contexte proche [Jauhar et Specia, 2012] ou issue d'un modèle de langue [Paetzold et Specia, 2015] appris sur un large corpus, typiquement

6. La base MRC Psycholinguistic Database fournit différentes indications psycholinguistiques sur une entrée lexicale, telle que son âge d'acquisition, son caractère concret, sa fréquence en corpus de Kucera-Francis etc., http://www.psych.rl.ac.uk/. Elle constitue également la base de l'Oxford Psycholinguistic Database.

Google 1T. Nous avons montré dans [Ligozat *et al.*, 2013] qu'une fenêtre de 4 mots à droite et à gauche du mot cible permettait d'obtenir les meilleurs résultats avec un modèle n-gram sur les données de SemEval2012.

Concernant le corpus de référence, [Paetzold et Specia, 2016] ont montré que l'utilisation d'un corpus de sous-titres de films permet d'obtenir des résultats comparables à ceux obtenus avec Google 1T mais avec un corpus beaucoup plus petit. Nous avons également montré dans [Ligozat *et al.*, 2013] que l'utilisation de la Simple English Wikipédia pour calculer les fréquences des substituts donnait des résultats équivalents à l'utilisation des fréquences de Google 1T, ce qui signifie que l'utilisation d'un corpus adapté (langue plus simple) permet d'obtenir des résultats équivalents à l'utilisation d'un corpus plus grand. En présence de corpus parallèles, il est également possible de prendre en compte la probabilité de l'alignement entre le mot cible et le substitut [Horn *et al.*, 2014].

Citons quelques résultats récents pour l'ordonnancement des substituts : [Paetzold et Specia, 2016] testent leur modèle sur le corpus de SemEval 2012, dont le meilleur résultat est un TRank [7] de 0,627 et utilisent un modèle de langue 5-gram sur un corpus de sous-titres de films et séries pour enfants SubIMDB. Le meilleur système à SemEval obtenait un TRank de 0,602.

Les modèles présentés ci-dessus sont cependant des modèles généraux de simplification lexicale, qui ne tiennent pas compte des profils des lecteurs. Dans [Tack *et al.*, 2016b], nous avons cherché à inverser le problème de la complexité lexicale en nous focalisant sur les apprenants plutôt que sur les mots. Nous avons ainsi mené des expériences de prédiction de la compétence lexicale d'apprenants de français langue étrangère (FLE). En effet, les prédictions classiques de lisibilité ne prennent en compte que le niveau de maîtrise de l'apprenant, et ne rendent pas compte des différences possibles entre apprenants du même niveau. Ensuite, puisque les prédictions sont dérivées d'une connaissance symbolique, a priori, de la compétence lexicale, ces modèles classiques n'intègrent pas de données pour modéliser a posteriori la compétence lexicale d'un apprenant. Enfin, étant donné que les prédictions des modèles restent constantes, les modèles prédictifs résultant ne sont pas adaptatifs et ne rendent pas compte de la nature incrémentale de l'acquisition du vocabulaire L2. Nous avons donc défini trois modèles de la connaissance lexicale : un premier modèle, appelé modèle expert, estime la compétence lexicale moyenne des apprenants d'un niveau donné à partir de la ressource FLELex ; le deuxième modèle, modèle du niveau, se fonde sur la moyenne des annotations faites par les apprenants du même niveau ; le troisième, modèle personnalisé, repose uniquement sur l'annotation de la connaissance lexicale faite par un apprenant individuel. Les deux derniers modèles utilisent un ensemble de traits lexicaux inspiré des travaux de [Gala *et al.*, 2014].

Ce travail a permis d'observer un gain d'exactitude significatif des modèles personnalisés par rapport à un modèle de référence. Nous avons également pu noter que certaines erreurs de prédiction étaient bien dépendantes du profil de

7. TRank au rang i : proportion d'instances pour lesquelles un candidat de rang de référence $r \leq i$ est classé en premier

l'apprenant, et notamment de sa langue maternelle : les modèles de prédiction de la complexité lexicale ne permettent pas encore de prendre en compte l'effet du transfert des connaissances à partir de la langue maternelle (L1) sur la compétence lexicale d'un apprenant. Une dernière observation est que les modèles étant fondés sur les lemmes, ils ne sont pas capables de distinguer la difficulté de certaines formes telles que les formes fléchies d'un même lemme verbal. Or, certains apprenants ne connaissaient par exemple pas la forme au passé simple du verbe *être*, alors que l'indicatif présent avait été annoté comme connu.

La simplification lexicale a donc connu d'importants progrès ces dernières années, mais les résultats d'annotation et d'ordonnancement montrent que la tâche est complexe. En outre, les modèles de simplification sont actuellement adaptés uniquement au niveau supposé de l'apprenant dans la langue, alors que nos résultats dans [Tack *et al.*, 2016b] montrent que des modèles adaptés plus finement au profil de l'apprenant aboutissent à une meilleure annotation de la complexité lexicale.

3.4.2 Aspects syntaxiques

Les aspects syntaxiques de la simplification ont déjà été évoqués lors de la présentation de systèmes de simplification complets, mais certains travaux se sont focalisés sur cet aspect. Ces systèmes sont généralement fondés sur des règles de réécriture d'arbres de constituants [Canning *et al.*, 2000, De Belder et Moens, 2010] ou de dépendances [Bott *et al.*, 2012, Drndarević *et al.*, 2013, Siddharthan, 2011]. [Siddharthan, 2006] se fondent sur les chunks mais annotent ensuite des informations de discours afin de préserver la cohésion du texte lors de la simplification. Dans tous les cas, les opérations de transformation correspondent à des opérations de découpage, suppression, réordonnancement, ou substitution, ce qui correspond bien à notre typologie. Une caractéristique importante des systèmes est le type d'information qu'ils prennent en compte : les modifications syntaxiques peuvent avoir des répercussions sur les plans morphologiques, sémantiques et discursifs notamment (nous laissons de côté le plan purement lexical, qui est plutôt pris en compte dans les systèmes complets évoqués en début de section). Ainsi, [Siddharthan, 2011] enrichit ses règles de modifications morphologiques, par exemple pour le passage de la voix active à la voix passive. [Siddharthan, 2006] effectue des transformations mettant en jeu la cohésion du texte (prise en compte des anaphores, ajout de connecteurs) pour conserver ou expliciter les relations discursives du texte.

Dans [Brouwers *et al.*, 2012, Brouwers *et al.*, 2014], nous avons utilisé notre typologie pour implémenter un système de simplification syntaxique pour le français. Afin de disposer d'un système adapté au ressources du français, et facilement adaptable à différents publics, nous avons choisi de créer un système s'appuyant sur une grammaire de réécriture des phrases, utilisant des arbres syntaxiques enrichis d'informations morphologiques comme entrée.

Notre processus de simplification est en deux étapes : dans un premier temps, l'ensemble des simplifications possibles est généré, puis la simplification respec-

tant au mieux les critères de lisibilité choisis est sélectionnée.

L'étape de surgénération s'appuie sur un ensemble de règles (19) fondées sur des critères morphosyntaxiques et syntaxiques. Une première étape consiste ainsi à annoter les textes de notre corpus avec les outils MELT [Denis et Sagot, 2009] et Bonsai [Candito *et al.*, 2010], et produit des arbres syntaxiques, sur lesquels les règles syntaxiques vont pouvoir s'appuyer.

Les règles définies sont de trois types : suppression (12 règles), modification (3 règles) et division (4 règles). Par rapport à notre typologie, les phénomènes de regroupement et de modification de temps ne sont donc pas traités. Les phénomènes de regroupement sont peu fréquents dans notre corpus, et difficiles à traiter sans conflit avec les règles de division et de suppression. Les modifications de temps posent également problème car elles impliquent des changements au niveau du texte, puisque l'emploi des temps au fil du texte doit être cohérent, et une erreur de modification du temps risque fort d'altérer la cohérence du texte et de dégrader sa lisibilité.

Afin d'appliquer l'ensemble des règles, les structures candidates à la simplification sont détectées en utilisant des expressions régulières sur les arbres syntaxiques, grâce à l'outil Tregex [Levy et Andrew, 2006]. Ensuite, les règles sont appliquées via un jeu d'opérations les implémentant dans l'outil Tsurgeon.

Les différentes règles sont appliquées récursivement à chaque phrase jusqu'à ce que toutes les alternatives possibles aient été générées. La plupart du temps, il y aura donc plusieurs variantes simplifiées pour une phrase donnée. L'étape suivante consiste à sélectionner une variante parmi l'ensemble des simplifications générées.

Nous avons choisi d'utiliser des critères de lisibilité pour effectuer cette sélection, qui sont combinés par une approche de programmation linéaire en nombres entiers (ILP) [Gillick et Favre, 2009]. Les critères choisis sont simples, et consistent à vérifier que la complexité lexicale est également améliorée, en plus de la complexité syntaxique. Nous avons utilisé quatre critères pour sélectionner une variante : la longueur de la phrase en terme de nombre de mots (h_w), la longueur moyenne des mots en termes de caractères (h_s), la familiarité du vocabulaire (h_a) et la présence de mots-clés (h_c). La familiarité des mots est estimée à l'aide de la liste de Catach [Catach, 1985], qui contient environ 3 000 mots considérés comme étant les plus fréquents du français. Les mots-clés ont été définis ici comme tout terme apparaissant plus d'une fois dans le texte.

Ces quatre critères ont été combinés grâce à un module ILP [8].

La simplification syntaxique créée des changements substantiels dans les phrases, et il est donc important de vérifier que l'application d'une règle ne génère pas d'erreurs qui rendraient les phrases produites incompréhensibles ou non grammaticales. Une évaluation manuelle de notre système sur ce point a été menée, en sélectionnant un ensemble de textes qui n'avait pas été utilisé pour l'analyse typologique précédente, soit 9 nouveaux articles de Wikipédia (202 phrases) et deux autres contes de Perrault (176 phrases). Dans cette éva-

8. Module ILP fondé sur `glpk` et disponible à l'adresse http://www.gnu.org/software/glpk/

luation, toutes les phrases simplifiées générées sont prises en compte et non pas simplement celles sélectionnées par l'ILP. Les résultats sont présentés dans le tableau 3.5. Deux types d'erreurs sont distinguées : celles provenant du prétraitement morphosyntaxique et syntaxique, et les erreurs de notre système de simplification.

Parmi les 202 phrases sélectionnées pour l'évaluation dans le corpus informatif, 113 (56%) ont été simplifiées, générant 333 variantes. Notre analyse d'erreur manuelle a montré que 71 de ces phrases (21%) contiennent des erreurs, parmi lesquelles 89% proviennent du prétraitement, tandis que les erreurs de simplification ne représentent que 11% des erreurs.

<table>
<tr><th colspan="4">Corpus informatif</th></tr>
<tr><th>nb. phr.</th><th>% correctes</th><th>% erreurs prétrait.</th><th colspan="2">% erreurs de simplification</th></tr>
<tr><td rowspan="2">333</td><td rowspan="2">262 (78,7 %)</td><td rowspan="2">63 (18,9%)</td><td colspan="2">8 (2,4 %)</td></tr>
<tr><td>syntaxe : 6 (1,8%)</td><td>sémantique : 2 (0,6%)</td></tr>
<tr><th colspan="4">Corpus narratif</th></tr>
<tr><th>nb. phr.</th><th>% correctes</th><th>% erreurs prétrait.</th><th colspan="2">% erreurs de simplification</th></tr>
<tr><td rowspan="2">369</td><td rowspan="2">292 (79,1 %)</td><td rowspan="2">39 (10,6%)</td><td colspan="2">38 (10,3 %)</td></tr>
<tr><td>syntaxe : 20 (5,4%)</td><td>sémantique : 18 (4,9%)</td></tr>
</table>

TABLE 3.5 – Performance du système de simplification sur les deux corpus

Les scores obtenus sur le corpus narratif sont légèrement inférieurs : parmi les 369 phrases générées à partir des 154 phrases originales, 77 (21%) présentent des erreurs, mais seules 51% sont dues à des erreurs de prétraitement, tandis que 49% dont dues à l'application de nos règles. Ces erreurs sont cependant dues à 2 ou 3 règles en particulier, qui pourraient être revues. La différence entre les deux corpus s'explique notamment par la présence importante de discours indirect dans les contes, qui est difficilement dintingable des clauses subordonnées simplifiables.

Globalement, ces résultats semblent proches de ceux des systèmes similaires développés pour l'anglais, mais cependant la méthodologie d'évaluation est généralement trop éloignée pour permettre une comparaison directe. [Siddharthan, 2006] ont fait évaluer leur système par trois juges, qui ont considéré qu'environ 80% des phrases simplifiées étaient grammaticales, et 87% préservaient le sens original. Ces résultats sont du même ordre que les nôtres. [Drndarević *et al.*, 2013] ont également fait évaluer les sorties de leur système par des évaluateurs humains qui ont estimé que 60% des phrases étaient grammaticales, et que 70% conservaient le sens. Ces scores sont un peu inférieurs aux nôtres, mais leur système utilise également des règles lexicales, et les erreurs incluent donc des erreurs à la fois grammaticales et lexicales.

Cette évaluation montre que les règles développées sont plus adaptées aux textes informatifs, et que les règles de suppression sont particulièrement difficiles à utiliser, notamment en présence de discours indirect.

3.5 Discussion

Le domaine de la simplification textuelle est ainsi très actif actuellement, avec des travaux sur différentes dimensions de la lisibilité et de la simplification. Quelques corpus de grande taille et ressources sont désormais disponibles : notamment pour l'anglais le corpus de Newsela et PPDB 2.0. Le corpus de Newsela est particulièrement intéressant car il comporte différents niveaux de simplification, ce qui est peu géré par les systèmes actuels.

Le parallèle avec la traduction automatique permet l'utilisation de méthodes devenues standard, mais comme dans ce domaine, se pose le problème de l'évaluation des systèmes. L'évaluation de la qualité de la simplification a fait l'objet d'un atelier et d'une tâche à LREC en 2016 [Štajner *et al.*, 2016]. Les résultats de la tâche d'évaluation de la qualité des simplifications ont montré que, si les métriques issues de l'évaluation de la traduction automatique sont efficaces pour mesurer la grammaticalité et la préservation du sens, elles ne permettent pas de mesurer la simplicité des textes produits. L'ajout de critères d'estimation de la qualité des traductions [9] améliore cependant les performances.

Un autre point qui nous semble intéressant est la comparaison des informations statistiques et des informations linguistiques, pour mieux caractériser les critères de complexité linguistique les plus pertinents, en allant dans la direction par exemple de [Panchenko, 2016] : relier les informations linguistiques à celles de fréquence (fréquences en corpus simples ou plongements lexicaux) donnerait une meilleure compréhension des résultats, et de mieux guider le processus de simplification, qu'il soit manuel ou automatique.

9. Évaluation des traductions sans traductions de référence [Specia *et al.*, 2010]

Chapitre 4

Recherche de voisins sémantiques

Nous nous sommes jusqu'ici intéressée à des variations entre énoncés proches de la similarité (paraphrase, implication textuelle). Dans cette section, nous nous intéressons cette fois à un type de variation un peu différent que nous avons appelé *homogénéité sémantique*. Les travaux présentés ici sont les résultats de la thèse de Van-Minh Pho [Pho, 2015], co-encadrée par Brigitte Grau, Yolaine Bourda et moi-même, et du stage de Thibault André en 2013 [André, 2013].

4.1 Cadre applicatif

La motivation de ce travail est issue d'un ensemble de problématiques provenant d'un cadre d'application pédagogique : correction automatique de réponses, et surtout ici validation et génération automatique de Questionnaires à Choix Multiples (QCM). En effet, les QCM sont largement utilisés dans de nombreux contextes d'apprentissage et d'évaluation. Les principales raisons en sont que leur évaluation peut être automatisée et que leur pertinence, ainsi que leur objectivité dans l'évaluation des compétences de l'apprenant, ont été prouvées [Haladyna *et al.*, 2002]. Cependant, la rédaction de QCM est coûteuse en temps, et la qualité des QCM est cruciale si l'on veut s'assurer que les résultats des apprenants correspondent à leurs compétences. Des défauts de conception peuvent engendrer un biais dans l'évaluation des apprenants : si les QCM donnent des indices sur la réponse correcte, ou sont trop compliqués, ils ne permettent pas d'évaluer correctement le niveau de connaissance des apprenants. Pour remédier à ces problèmes, des études de psychologie de l'éducation ont été dédiées à la conception de QCM, dont ont découlé différentes consignes de rédaction. Cependant, les enseignants peuvent ne pas avoir connaissance de l'existence de ces consignes ou rencontrer des difficultés pour les appliquer [Tarrant et Ware, 2008]. Ainsi, des outils d'évaluation automatique de la qualité de QCM pourraient assister les enseignants dans leur travail de production

d'exercices pédagogiques ou de tests. Avec le développement des Environnements Informatiques pour l'Apprentissage Humain (EIAH) sous diverses formes comme les Environnements Numériques de Travail, ou les MOOC (Massive Open Online Courses), de nombreux enseignants fournissent des supports de cours, des exercices pédagogiques et des tests sous forme électronique, et il est donc possible d'envisager d'intégrer des outils permettant de faciliter la tâche des enseignants. Nous nous sommes par conséquent intéressés à la validation automatique de QCM.

4.2 Définitions

Commençons par quelques définitions. Un *Questionnaire à Choix Multiples* ou QCM est un ensemble de *questions* ou *items*, chacune d'entre elles étant composée de deux parties :

- l'*amorce* ou consigne donnée à l'apprenant, généralement sous forme interrogative, mais parfois aussi consigne ou texte à trous ;
- les *options*, choix possibles de réponses à une amorce, et incluant
  - la *réponse* (option correcte)
  - et un ou plusieurs *distracteurs* (options incorrectes).

Nous considérons en outre ici que nous disposons d'un *document* permettant de répondre au QCM.

L'exemple 4.2.1 présente ces différentes composantes d'un item.

Exemple 4.2.1.

Dans quelle ville le service de vélo en libre-service s'appelle-t-il Vélo'v ? ← amorce

- Valence ← distracteur
- *Lyon* ← réponse
- Caen ← distracteur
- Paris ← distracteur

Amour et liberté. Le nom imaginé à Lyon, Vélo'v, est un croisement entre les mots vélo et love. Le terme vélo, il est vrai, est magique. C'est l'anagramme de «love». Il est lové dans développement, marque le début de «évolution». Paris, avec le Velib', créé en 2007, préfère invoquer la liberté, tout comme le Libélo, à Valence ou Cristolib, à Créteil. ← document

Dans la thèse de Van-Minh Pho, nous nous sommes intéressés en particulier à l'évaluation de la qualité des distracteurs, c'est-à-dire leur capacité à évaluer l'assimilation de la notion évaluée par les apprenants. Dans les travaux de psychologie de l'éducation, la qualité d'un QCM est mesurée a posteriori, à partir des scores obtenus par des apprenants, et selon deux critères principaux : la fiabilité, c'est-à-dire la reproductibilité d'un test, et la validité, c'est-à-dire sa capacité à mesurer ce qu'il est censé mesurer, critère qui ne sont bien entendu pas indépendants. La validité peut se décomposer en trois paramètres [Considine *et al.*, 2005] : la validité du contenu, qui mesure si les items sont représentatifs des connaissances et des compétences évaluées ; la validité

apparente, c'est-à-dire la clarté, lisibilité des items pour les apprenants ; et la validité conceptuelle, qui évalue la capacité d'un test à mesurer ce qu'il prétend mesurer. Les inconvénients de ce type d'évaluations sont qu'elles ne mesurent la qualité d'un test qu'a posteriori, et ne permettent donc pas de corriger les items avant de faire passer le test ; en outre, elles sont difficilement reproductibles puisqu'il n'est pas possible de faire passer le même test plusieurs fois aux mêmes apprenants ; enfin, ces évaluations sont lourdes car elles nécessitent de mobiliser un grand nombre d'apprenants.

Afin de générer des QCM de qualité a priori, des consignes de rédaction ont été proposées [Bernard et Fontaine, 1982, Burton *et al.*, 1991, Haladyna et Downing, 1989, Haladyna *et al.*, 2002]. La taxonomie de [Haladyna *et al.*, 2002], la plus couramment utilisée, comporte 5 catégories de consignes concernant : le contenu de l'item, le format de l'item, le style de l'item, la rédaction de l'amorce, et la rédaction des options (voir la thèse de Van-Minh [Pho, 2015] pour une présentation détaillée). Certaines consignes sont directement implémentables (et d'ailleurs souvent implémentées dans les interfaces pédagogiques numériques courantes), par exemple le fait de faire varier l'emplacement de la réponse en fonction du nombre d'options. Dans nos travaux, nous nous sommes focalisés sur les consignes concernant la qualité des distracteurs, et nous sommes plus particulièrement intéressés aux consignes qui sont moins facilement implémentables mais qui pourraient être prises en compte avec des outils de TAL, comme par exemple « Rendre plausibles tous les distracteurs ».

Plusieurs analyses de corpus ont montré qu'environ la moitié des items utilisés pour des examens d'apprenants comportent au moins une violation de consigne [Downing, 2005, Tarrant et Ware, 2008]. Concernant les distracteurs, les consignes les moins respectées concernent leur homogénéité, dont nous discuterons plus loin. [Downing, 2005] a également montré que les items ne respectant pas les consignes rendent le test plus difficile.

Certains travaux se sont par conséquent intéressés à la sélection de distracteurs pour générer des QCM, en se fondant sur des mesures de similarité ou de voisinage. Le travail de [Mitkov *et al.*, 2009] est le plus proche du nôtre. Leur méthode est fondée sur l'utilisation de WordNet et Wikipédia, en fonction de la taille de la réponse. Les items engendrés concernent des cours de linguistique en langue anglaise. Une première étape consiste à obtenir une liste de distracteurs potentiels assez large, parmi lesquels seront sélectionnés des distracteurs en fonction de leur similarité à la réponse. Si la réponse est un mot, les distracteurs potentiels sont ses coordonnées dans l'arborescence de WordNet, c'est-à-dire les concepts partageant un ancêtre direct commun avec la réponse. Si la réponse est constituée de plusieurs mots, les distracteurs potentiels sont les syntagmes nominaux extraits des titres des pages Wikipédia dont la tête est identique à celle de la réponse. Ensuite, plusieurs stratégies de sélection ont été testées pour sélectionner automatiquement les distracteurs à partir des distracteurs potentiels, en appliquant différentes mesures de voisinage sémantique fondées sur WordNet, une mesure de voisinage distributionnel, une mesure de similarité phonétique, et une combinaison (union) des mesures. Les items ainsi engendrés ont été évalués

de façon extrinsèque, en se fondant sur les résultats d'apprenants et en calculant des mesures psychométriques. L'évaluation de ces items a montré qu'aucune de ces mesures seule n'est réellement meilleure que les autres, mais la mesure de Lin fondée sur WordNet, et la combinaison des mesures ont généralement les meilleurs résultats. L'une des limites de cette approche est de ne permettre la sélection de distracteurs que pour des noms et des syntagmes nominaux. En outre, l'utilisation des mesures de façon indépendante limite la robustesse de cette approche.

Le travail de [Karamanis *et al.*, 2006] repose sur le même type de méthode, et a pour objectif la génération de QCM médicaux en anglais, à partir d'un document textuel et du méta-thésaurus UMLS. Les distracteurs potentiels sont des termes de même classe sémantique que la réponse dans l'UMLS. Pour chacun de ces distracteurs potentiels, un score de voisinage distributionnel est calculé à partir d'un corpus de référence, et les distracteurs sélectionnés sont ceux avec le meilleur score. L'évaluation des items engendrés a été effectuée par des experts du domaine qui ont jugé les distracteurs ainsi sélectionnés. Environ la moitié des items engendrés automatiquement ont été invalidés par les experts.

[Papasalouros *et al.*, 2008] et [Cubric et Tosic, 2011] s'appuient quant à eux sur une ontologie de domaine, et des règles de sélection des distracteurs en fonction de leur place dans l'ontologie : par exemple, un distracteur peut être une instance d'un concept coordonné au concept dont la réponse est une instance. Cette méthode nécessite de disposer d'une ontologie de domaine, limite les types de réponse traitées, et le type de distracteurs possibles. [Foulonneau, 2011] a étudié la possibilité d'utiliser des ressources du web sémantique à la place d'une ontologie et montre les intérêts et limites de cette approche.

Enfin, de nombreux travaux se sont intéressés à la génération de QCM pour l'apprentissage des langues, pour des amorces de type texte à trou ou de vocabulaire (comme par exemple « Complétez la phrase «La personne suivante attendait qu'ils.....la porte.» : 1. aient franchi, 2. franchissent, 3. eurent franchi, 4. eussent franchi ») [Brown *et al.*, 2005, Liu *et al.*, 2005, Sumita *et al.*, 2005, Aldabe *et al.*, 2006, Lin *et al.*, 2007, Lee et Seneff, 2007, Sung *et al.*, 2007, Goto *et al.*, 2010]. La problématique est alors un peu différente car il s'agit de tester la connaissance de la langue, et non la compréhension d'un texte ; par conséquent, les distracteurs sont généralement sélectionnés en fonction de leur forme (différente conjugaison d'un verbe, termes de la même famille morphologique), ou de leur coocurrence avec la réponse, et éventuellement sur des critères sémantiques simples (synonymes ou antonymes dans WordNet) en cas de question de vocabulaire.

Dans nos travaux, nous avons cherché à traiter des QCM de compréhension de textes ou de test de connaissances. En outre, plutôt que d'appliquer des méthodes de sélection de distracteurs en supposant a priori qu'elles fourniront des distracteurs valables, nous avons souhaité comparer différentes méthodes de sélection, et définir un mode d'évaluation intrinsèque de ces méthodes qui permette d'effectuer cette comparaison.

Remarquons que dans tous ces travaux, les distracteurs sont sélectionnés en fonction de leur relation à la réponse. Nous faisons dans la suite la même

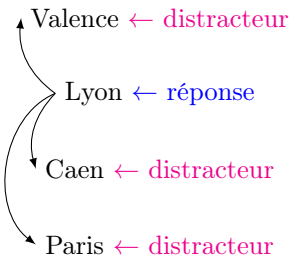


FIGURE 4.1 – Homogénéité entre distracteurs et réponse

hypothèse simplificatrice (figure 4.1) : nous considérerons que les distracteurs doivent être homogènes à la réponse, au lieu de considérer le cas plus général d'options (distracteurs plus réponses) homogènes entre elles.

En réalité, les distracteurs ne sont pas toujours homogènes avec la réponse, comme le montre l'exemple 4.2.2 : les deux options indiquant des bonnes réponses sont proches, et les deux distracteurs sont homogènes entre eux, mais pas avec les réponses.

Exemple 4.2.2.

Que peut-il advenir aux propos d'un internaute publiés sur un forum ? [1]

A - Ils pourront figurer dans les résultats d'un moteur de recherche.

B - Ils pourront rester accessibles en ligne pendant plusieurs années.

C - L'internaute pourra les modifier à tout moment.

D - Les autres utilisateurs du forum pourront les modifier sans son autorisation.

4.3 Homogénéité des options

Bien que les travaux antérieurs abordent la recherche de distracteurs comme un problème de mesure de similarité sémantique, il s'agit en réalité d'une notion un peu différente, car les distracteurs doivent être différents de la réponse, et non pas réellement similaires, et leur contenu doit être de même niveau, ce que traduisent les consignes suivantes :

— « Rendre les options indépendantes les unes des autres : le sens de l'une ne doit pas être inclus dans le sens de l'autre » ;

— « S'assurer que seulement une option correspond à la réponse » ;

— « Rendre plausibles tous les distracteurs » ;

— et « Rendre la formulation des options homogène en contenu et en structure grammaticale ».

La traduction de ces consignes en notions utilisables en TAL a été commencée dans le cadre du stage de Thibault André, et a été prolongée dans la thèse de Van-Minh Pho.

1. Exemple c2i

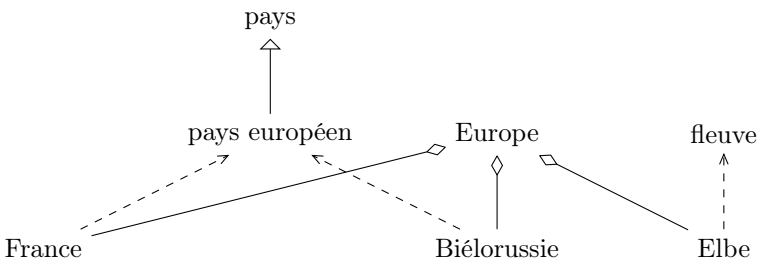


FIGURE 4.2 – Caractérisation sémantique de paires de nœuds

Nous avons traduit cette relation syntaxique et sémantique par la notion d'*homogénéité*, qui se décline en deux aspects : une *homogénéité* syntaxique et une *homogénéité* sémantique.

Du point de vue syntaxique, l'homogénéité signifie que les structures syntaxiques des options sont proches les unes des autres. Dans l'exemple 4.3.1, les options sont ainsi constituées d'un groupe verbal avec un verbe au gérondif puis du même groupe nominal, et d'un groupe prépositionnel.

Exemple 4.3.1.

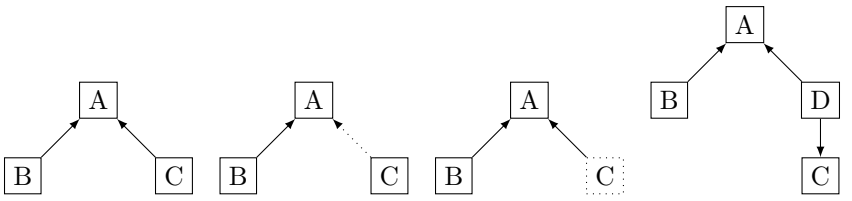
Comment peut-on signifier dans une requête qu'un mot-clé doit être exclu des réponses ? [2]

A - En faisant précéder le mot-clé du symbole — (moins).
B - En faisant précéder le mot-clé du symbole + (plus).
C - En mettant le mot-clé entre guillemets.
D - En mettant le mot-clé entre crochets.

Du point de vue sémantique, l'homogénéité signifie que les options doivent être sémantiquement voisines et de même niveau, sans être trop similaires. Nous donnons la définition de plusieurs notions utiles à la définition de l'homogénéité sémantique, en faisant référence à une organisation des connaissances sous la forme d'un graphe hiérarchique tel que présenté dans la figure 4.2 et contenant des concepts typés et des relations sémantiques. Nous considérons pour l'instant que les options sont limitées à des concepts, mais discuterons la généralisation à tout type d'option plus loin.

La définition du *voisinage sémantique* est la suivante : «Le voisinage sémantique indique dans quelle mesure deux concepts sont sémantiquement distants dans un réseau ou une taxonomie en utilisant toutes les relations entre eux (c'est-à-dire des relations d'hyponymie/hyperonymie [3], d'antonymie [4], de méronymie [5] et toutes sortes de relations fonctionnelles incluant *is-made-of*, *is-an-attribute-of*, etc.)» [Ponzetto et Strube, 2007]. Le voisinage sémantique est établi entre deux termes lorsqu'il existe un chemin entre les concepts auxquels ils se réfèrent, et le

2. exemple c2i
3. Deux concepts dont le premier a un sens plus spécifique/général que le second.
4. Deux concepts dont les sens sont opposés.
5. Deux concepts dont le premier est une partie ou un membre du second.



(a) Idéalement (b) Lien manquant (c) Concept manquant (d) Différence de granularité

FIGURE 4.3 – Différents cas problématiques

degré de voisinage est dépendant de la longueur du chemin. Dans la figure 4.2, tous les concepts peuvent être considérés comme sémantiquement voisins.

Nous définissons la *similarité sémantique* comme un cas particulier de voisinage sémantique : deux termes sont similaires s'ils partagent le même sens (c'est-à-dire qu'ils sont des synonymes) ou un sens partiel, c'est-à-dire que les concepts auxquels ils se réfèrent sont liés par une chaîne ascendante ou descendante de relations *is-a* ou de méronymie, tels que «Biélorussie» et «Europe» dans la figure 4.2.

Nous définissons la *spécificité sémantique* comme un cas particulier de voisinage sémantique entre deux termes dont les concepts auxquels ils se réfèrent partagent un ancêtre commun direct, comme «France» et «Biélorussie» dans la figure 4.2.

Nous proposons enfin la définition de l'*homogénéité sémantique* comme étant un cas particulier de *voisinage sémantique* qui considère toutes les relations entre les concepts comparés mais exclut la notion de *similarité sémantique* : deux options ne peuvent être similaires. Enfin, une meilleure homogénéité est atteinte si la *spécificité sémantique* est respectée.

Dans une représentation structurée, deux termes seront homogènes s'ils sont, idéalement, coordonnées dans l'arborescence, c'est-à-dire qu'ils partagent un même ancêtre commun direct (exemple 4.3(a)). C'est d'ailleurs l'un des critères utilisés par certains travaux sur la sélection de distracteurs [Mitkov *et al.*, 2009, Papasalouros *et al.*, 2008, Cubric et Tosic, 2011]. Cependant, il n'existe pas toujours de représentation structurée du domaine considéré, et même si elle existe, elle n'est pas nécessairement complète, ou du niveau de granularité souhaité (exemples 4.3(b), 4.3(c) et 4.3(d)). Il sera donc nécessaire d'approximer cette relation d'homogénéité en utilisant des mesures de proximité entre termes fondées sur corpus.

En outre, les options ne correspondent généralement pas à n'importe quels concepts coordonnés : les distracteurs devant être des réponses plausibles, ils sont des concepts proches sémantiquement de la réponse, selon le contexte fourni par le document associé à l'item. Des mesures de voisinage sémantique appliquées aux termes du document permettraient donc de sélectionner des distracteurs plus pertinents.

Dans nos travaux, nous avons donc choisi d'utiliser deux types de mesures

de voisinage pour valider les distracteurs : des mesures fondées sur des représentations structurées, permettant d'estimer la distance entre deux concepts en fonction des relations entre ces concepts, et des mesures fondées sur corpus, qui indiquent à quel point deux concepts sont voisins sémantiquement. Chaque mesure sera appliquée à la réponse et à l'ensemble formé des distracteurs et des exemples négatifs, afin de tenter de classer les distracteurs dans les premiers rangs. Le choix des mesures a été fait en deux temps. Nous avons d'abord choisi des mesures représentatives des différentes approches de calcul du voisinage sémantique. Puis nous avons étudié la distribution des scores de proximité des distracteurs afin de sélectionner les mesures qui semblaient les plus pertinentes. Deux types d'approches sont classiquement distingués en calcul de voisinage sémantique : les approches fondées sur des ressources fournissant des relations sémantiques explicites ; et les approches distributionnelles, qui estiment la proximité sémantique de deux mots en fonction de la similarité de leurs contextes. Nous avons ainsi sélectionné plusieurs mesures exploitant les relations sémantiques de WordNet (qui sont par ailleurs utilisées dans les travaux de sélection de distracteurs, ce qui nous permet de comparer notre modèle à ces mesures) ; à l'opposé une mesure fondée sur les distributions en corpus ; et enfin deux mesures qui exploitent des relations non typées (liens Wikipédia, gloses de WordNet).

Afin de valider les hypothèses d'homogénéité syntaxique et sémantique en corpus, puis la possibilité d'utiliser des méthodes de TAL pour les calculer, nous avons constitué des corpus de QCM, l'un en anglais et l'autre en français.

Corpus

Le corpus anglais comprend des QCM provenant :

— d'évaluations de systèmes de compréhension automatique de textes, QA4MRE [Peñas *et al.*, 2013] 2011, 2012 et 2013 ;

— de plusieurs sites d'apprentissage de la langue anglaise, ayant pour objectifs de tester soit la compréhension de la langue, soit les connaissances sur un sujet donné. Nos critères de sélection étaient que nous cherchions des QCM de compréhension, avec un document associé.

Ce corpus contient au total 196 items et 741 options (soit environ 4 options par item).

Le corpus français rassemble des QCM de compréhension de la langue française, et contient 556 items et 1572 options.

Tous ces QCM comprennent des documents associés aux items, sur lesquels sont posées les questions.

Afin d'étudier la validité des hypothèses d'homogénéité en corpus, nous avons mené une analyse de corpus, dans un premier temps manuelle, puis automatique. L'analyse manuelle a été faite dans le cadre du stage de Thibault [André, 2013].

Afin de tester l'homogénéité syntaxique, nous avons défini plusieurs catégories d'annotation des distracteurs selon leur proximité syntaxique avec la réponse :

— syntaxe identique, lorsque les groupes syntaxiques de plus bas niveau sont de même type, comme dans l'exemple 4.3.2 ;

— syntaxe partiellement identique, lorsqu'un groupe seulement est de type différent (exemple 4.3.3) ;
— syntaxe globalement identique, lorsque le groupe syntaxique de plus haut niveau est identique (exemple 4.3.4) ;
— syntaxe différente lorsqu'aucun des cas précédent n'est satisfait (exemple 4.3.5).

Exemple 4.3.2.
distracteur : **NP**(The number) **PP**(of tortoises) **VP**(began) **PP**(to decrease)

réponse : **NP**(The number) **PP**(of tortoises) **VP**(began) **PP**(to grow)

Exemple 4.3.3.
distracteur : **NP**(it) **VP**(resists) **NP**(diseases)

réponse : **NP**(it) **VP**(is not) **AP**(profitable)

Exemple 4.3.4.
distracteur : **NP**(The total figures)

réponse : **NP**(**NP**(The number) **PP**(of units) **AP**(recorded))

Exemple 4.3.5.
distracteur : **ADVP**(First) **ADVP**(there) **VP**(was) **NP**(a tie) , **ADVP**(then) **NP**(Chestnut) **VP**(won)

réponse : **NP**(Joey Chestnut)

Les résultats de l'annotation manuelle sont résumés dans le tableau 4.1.

	Nombre
Syntaxe identique	40%
Syntaxe partiellement identique	19%
Syntaxe globalement identique	29%
Syntaxe différente	12%

TABLE 4.1 – Répartition des distracteurs selon leur homogénéité syntaxique avec la réponse

Ce tableau montre qu'environ 40% des distracteurs partagent une syntaxe commune avec la réponse. Ces distracteurs sont principalement des entités nommées, mais quelques phrases et clauses appartiennent à cette catégorie. La moitié des distracteurs annotés comme ayant une syntaxe partiellement identique à celle de la réponse sont des listes ou correspondent à des réponses qui sont des listes, et ne diffèrent que par le nombre d'éléments. La plupart des distracteurs ayant une syntaxe globalement identique à celle de la réponse sont des clauses, ce qui était attendu puisque leur structure syntaxique a moins de chances d'être strictement la même que celle de la réponse. L'homogénéité syntaxique semble

donc bien respectée pour une grande partie des distracteurs, si l'on prend en compte les légères variations évoquées.

Afin de tester l'homogénéité sémantique, nous avons également effectué une annotation sémantique des options. Si les options doivent être homogènes à la réponse, elles devraient être de même type sémantique que la réponse. Nous avons donc utilisé comme critère d'homogénéité sémantique le type des options, comparé d'une part au type sémantique attendu selon l'amorce, et d'autre part au type sémantique de la réponse.

Le type sémantique attendu selon l'amorce peut être un *type spécifique* (par exemple l'amorce « Quel président a eu le plus d'enfants ? » attend une réponse correspondant à un nom de président) ; un *type d'entité nommée* comme un nom de personne, de lieu ou d'organisation (la question « Qui a inventé le téléphone ? » attend un nom de personne en réponse) ; ou enfin un *rôle sémantique* (« Pourquoi les patients en Afrique n'ont-ils pas accès aux médicaments rétroviraux ? »). Une annotation manuelle a été effectuée, en prenant en compte les catégories suivantes :

— type conforme : le type de l'option est conforme au type attendu. Pour les types spécifiques et entités nommées, cela signifie que les deux types sont identiques. Pour les rôles sémantiques (causes, définitions...), l'option est considérée comme conforme si elle constitue un argument possible pour ce rôle ;
— type non conforme : le type de l'option est différent du type attendu ;
— conformité inconnue : options pour lesquelles il n'est pas possible d'identifier le type attendu dans l'amorce ou pour l'option.

Le tableau 4.2 indique les résultats de cette annotation.

	Pourcentage
Type conforme	76%
Type non conforme	4%
Conformité inconnue	20%

TABLE 4.2 – Conformité des options avec le type attendu de l'amorce

Ces résultats montrent qu'environ trois quarts des options correspondent au type attendu par l'amorce. Les 20% de conformité inconnus correspondent aux situations dans lesquelles déterminer le type attendu est impossible, notamment des items de type questions à trous (« If you "out do" someone, you ... ? »).

Concernant la conformité avec le type de la réponse, l'annotation est fondée sur la taxonomie des entités nommées du système de questions-réponses QALC [Ferret *et al.*, 2000], qui présente l'avantage d'être assez générale et de regrouper entités nommées et numériques. Les catégories d'annotation sont les suivantes :

— type d'entité nommée identique : distracteurs ayant le même type d'entité nommée que la réponse ;
— type d'entité nommée différent : distracteurs n'ayant pas le même type d'entité nommée que la réponse ;

— distracteur non entité nommée : distracteurs qui ne sont pas des entités nommées.

Le tableau 4.3 indique les résultats de cette annotation.

	Pourcentage
Type identique	21%
Type différent	4%
Non entité nommée	75%

TABLE 4.3 – Conformité des options avec le type de la réponse

Environ trois quarts des distracteurs ne sont pas des entités nommées de type QALC ; parmi les distracteurs entité nommée, la plupart sont du même type que la réponse. Les cas restants sont des cas où le distracteur est d'un type hyponyme ou hyperonyme de celui de la réponse, ou bien des cas où la réponse n'est pas une entité nommée, contrairement au distracteur (par exemple le distracteur « In New York » et la réponse « A proper geographical term »).

Cette étude préalable de corpus nous a également permis de constater que l'homogénéité syntaxique et sémantique est plus simple à observer et calculer pour les options « courtes » (comme par exemple les entités nommées) que pour les réponses plus longues. Les réponses longues respectent généralement l'homogénéité syntaxique, mais l'homogénéité sémantique peut être partielle (exemple 4.3.6) ou se décomposer en plusieurs parties (exemple 4.3.7). Pour cette raison, nous avons choisi dans un premier temps de nous limiter aux items dont la réponse est une réponse courte, c'est-à-dire une entité nommée ou un groupe nominal.

Exemple 4.3.6.
— Réponse : To show how simple, traditional cultures can have complicated grammar structures
— Distracteur : To demonstrate how difficult it is to learn the Cherokee language

Exemple 4.3.7.
— Réponse : Installing a handle.
— Distracteur : Using a vise.

Modèle

Pour évaluer automatiquement la qualité des distracteurs, nous avons choisi un modèle d'ordonnancement fondé sur des critères d'homogénéité sémantique entre les candidats et la réponse. L'architecture globale est présentée dans la figure 4.4.

Afin d'apprendre un modèle de validation des distracteurs, nous avons créé des exemples que nous considérerons comme négatifs, sélectionnés en fonction d'un critère d'homogénéité syntaxico-sémantique. Ces exemples négatifs sont d'abord extraits selon deux méthodes distinctes : la première consiste à les

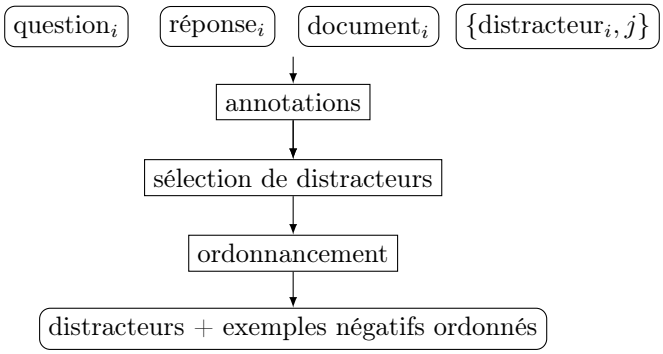


FIGURE 4.4 – Architecture présentant les étapes du modèle

extraire du document de référence de l'item ; la seconde consiste à sélectionner les options des autres items du corpus évalué. Un premier filtrage consiste à ne garder que les exemples de type syntaxique ou entité nommée proche de celui de la réponse. Un second filtrage vise à supprimer les exemples similaires à une option.

Exemple 4.3.8.
Quel est le pays d'origine de Nelson Mandela ?
— Réponse : Afrique du Sud
— Distracteur : Namibie
— Distracteur : Mali
— Distracteur : Tanzanie
— Exemple négatif : Hong Kong
— Exemple négatif : Zambie
— Exemple négatif : Johannesburg

Pour chaque item, nous disposons donc de l'amorce, la réponse, les distracteurs, le texte associé, et un ensemble d'exemples négatifs (voir exemple 4.3.8). Le nombre d'exemples pour chaque corpus est indiqué dans le tableau 4.4. Notons que parmi les exemples considérés comme négatifs, certains pourraient constituer des distracteurs pertinents (comme par exemple « Zambie » dans l'exemple 4.3.8). Pour tenir compte de cela, nous ne modélisons pas notre problème comme un problème de classification binaire (quoiqu'il serait possible de le faire, en considérant que nos exemples négatifs sont bruités), mais comme un problème d'ordonnancement, c'est-à-dire que nous cherchons à ce que les distracteurs soient classés dans les premiers rangs.

Pour pouvoir extraire les exemples négatifs et calculer les différents critères d'homogénéité, il est nécessaire de disposer d'informations syntaxiques et sémantiques, ce qui est effectué par différents outils : analyse syntaxique avec le Stanford Parser [Chen et Manning, 2014] ; annotation en entités nommées avec le Stanford Named Entity Recognizer [Finkel *et al.*, 2005], complétée par des listes de déclencheurs ; annotation des entités DBpedia par DBpedia

	réponse entité nommée	réponse non EN
Distracteurs	299	545
Exemples négatifs (extraits du document)	7 583	42 396
Exemples négatifs (extraits des autres items)	27 390	69 558

TABLE 4.4 – Nombre de distracteurs et d'exemples négatifs

Spotlight [Daiber *et al.*, 2013], également étendue de façon ad-hoc ; annotation en synsets de WordNet de la tête syntaxique. Les annotations des options et exemples négatifs sont tirées de leurs mentions dans le texte (plus fiable), ou par défaut de leur mention isolée.

Comme indiqué précédemment, nous avons souhaité associer deux types de mesures, fondées sur des représentations structurées (WordNet, DBpedia...), et sur corpus. Nous avons fondé notre choix de mesures sur les travaux en sélection de distracteurs (afin de déterminer les performances des critères utilisés dans ces travaux), et sur les travaux en reconnaissance de voisinage sémantique.

Un premier type de mesure très simple consiste à approximer la spécificité en utilisant une annotation en types sémantiques. Le premier critère consiste donc en une comparaison simple du type d'entité nommée (critère *meme_type_EN* dans la suite). Cependant, les types d'entité nommée sont très généraux, et ne permettent souvent pas de faire la différence entre diverses catégories de lieu par exemple (ville, montagne...).

Un second critère sera donc les types sémantiques issus de DBpedia, qui sont plus précis que les types entités nommées (par exemple « Harvard » est de type *Private_university*, alors que son type entité nommée serait *Organization*), et présentent également l'avantage d'être organisés hiérarchiquement, ce qui permet de calculer une mesure plus précise qu'une simple égalité des types. Pour comparer les types DBpedia, nous avons appliqué la mesure de [Wu et Palmer, 1994] sur les types DBpédia des termes (*type_dbpedia* dans la suite). L'annotation en types DBpédia couvre environ 80% des distracteurs et 40 à 80% des exemples négatifs (selon la méthode de sélection) correspondant à des réponses entités nommées ; pour les réponses non entités nommées, la couverture descend à 16% des distracteurs, et 13 à 19% des exemples négatifs. Cette annotation sera donc une information a priori importante, mais n'a pas une couverture complète et n'est pas suffisamment sélective.

Le second type de mesure utilise les relations sémantiques entre concepts de la ressource WordNet, qui permettra a priori de couvrir des cas différents de DBpedia, c'est-à-dire ceux qui ne correspondent pas à des entités nommées. Nous avons utilisé les quatre mesures sélectionnées par [Mitkov *et al.*, 2009] dans leur travail de sélection automatique de distracteurs, qui représentent des aspects complémentaires du voisinage sémantique : la mesure de recoupement étendu

de gloses qui évalue la proximité textuelle des gloses des synsets ; la mesure de [Leacock et Chodorow, 1998] fondée sur le plus court chemin entre les synsets donc sur les relations explicites entre les concepts ; et enfin les mesures de [Jiang et Conrath, 1997] et [Lin, 1997], fondées sur le contenu informationnel (avec le corpus par défaut de WordNet), qui combinent les connaissances sémantiques de ressources structurées avec les distributions des concepts dans un grand corpus. Les termes pouvant être polysémiques, ils sont associés à plusieurs synsets, et nous conservons la mesure maximale entre deux termes. L'analyse de la couverture de ces mesures indique que les termes correspondant à des réponses entités nommées sont moins bien couverts que les autres, et que la couverture des mesures utilisant le corpus est moins grande que celle des mesures fondées sur les synsets et leurs relations uniquement.

Le troisième type de mesure est fondé sur les liens en corpus des deux termes. Le premier critère de ce type est la comparaison des liens des pages Wikipedia, fondé sur l'outil Wikipedia Miner [Milne et Witten, 2013]. Le voisinage de deux concepts est fondé sur la similarité des liens des pages associées. Enfin, le second critère de ce type exploite les distributions des termes dans les pages Wikipédia, grâce à l'outil ESAlib [6]. D'un point de vue couverture, la couverture de Wikipédia sur les entités nommées est bonne, moins sur les non entités nommées. L'ESA s'apuyant sur la présence des termes dans les pages Wikipédia, sa couverture est très importante (environ 90% des termes).

L'ordonnancement des candidats intégrant les différents critères d'homogénéité définis précédemment est effectué avec l'outil SVMRank dans les résultats présentés ici, mais d'autres outils d'ordonnancement ont également été testés dans le cadre du stage d'Emilie Piérot en 2015. L'objectif est que pour tout distracteur d et tout exemple négatif e, le rang de d soit inférieur au rang de e.

Les métriques d'évaluation principales sont le rappel et la précision, qui sont définis de la façon suivante, pour un item i comprenant n_i distracteurs :

$$Rappel_i = \frac{\#distracteurs_i \text{ de rang} \leq n_i}{n_i}$$

$$Precision_i = \frac{\#distracteurs_i \text{ de rang} \leq n_i}{\#candidats_i \text{ de rang} \leq n_i}$$

La précision d'un ensemble d'items est simplement la moyenne des précisions pour chaque item, de même pour le rappel. Ces mesures permettent de vérifier que les distracteurs sont bien classés dans les premiers rangs. Nous calculons également la MAP (*Mean Average Precision*), pour un ensemble de I items :

$$MAP = \frac{\sum_{i=1}^{I} Pmoy_i}{I}$$

basée sur la précision moyenne d'un item :

$$Pmoy_i = \frac{\sum_{d=1}^{n} \frac{rel(d)}{abs(d)}}{|\text{Di}|}$$

6. http://ticcky.github.io/esalib

où d est un distracteur, $rel(d)$ son rang relatif parmi les distracteurs de l'item i, $abs(d)$ son rang parmi l'ensemble des distracteurs et exemples négatifs et Di l'ensemble des distracteurs de l'item i.

Nous avons créé quatre modèles d'apprentissage, un par type de réponse (entité nommée ou non) et par type de filtrage des exemples négatifs (extraits du document ou issus des autres items). Les modèles sont évalués par validation croisée en 7 sous-ensembles. Les résultats de ces modèles, ainsi que de chaque critère individuel, sont présentés dans les tableaux 4.5 et 4.6.

	Réponses entités nommées				Réponses non entités nommées			
	R	P	F	MAP	R	P	F	MAP
type_EN	0,83	0,13	0,23	0,85				
type_dbpedia	0,68	0,30	0,41	0,73	0,93	0,07	0,13	0,92
reg	0,68	0,22	0,34	0,73	0,44	0,14	0,21	0,46
lch	0,70	0,21	0,33	0,75	0,45	0,13	0,20	0,48
jcn	0,80	0,16	0,26	0,82	0,54	0,12	0,19	0,57
lin	0,81	0,17	0,28	0,83	0,55	0,11	0,18	0,57
liens Wikipédia	0,39	0,27	0,32	0,48	0,64	0,13	0,22	0,66
ESA	0,27	0,20	0,23	0,36	0,28	0,17	0,21	0,32
Modèle	0,42	0,40	**0,41**	**0,49**	0,22	0,22	**0,22**	**0,27**

TABLE 4.5 – Résultats du modèle d'ordonnancement dans le cas où les distracteurs sont extraits des documents

	Réponses entités nommées				Réponses non entités nommées			
	R	P	F	MAP	R	P	F	MAP
type_EN	0,83	0,03	0,05	0,84				
type_dbpedia	0,51	0,08	0,13	0,53	0,87	0,04	0,07	0,87
reg	0,60	0,12	0,20	0,62	0,42	0,11	0,18	0,43
lch	0,65	0,09	0,16	0,67	0,45	0,10	0,17	0,46
jcn	0,74	0,06	0,11	0,75	0,53	0,10	0,16	0,54
lin	0,74	0,07	0,12	0,76	0,53	0,09	0,16	0,55
liens Wikipédia	0,35	0,25	0,29	0,39	0,58	0,13	0,21	0,60
ESA	0,28	0,21	0,24	0,33	0,30	0,19	0,23	0,33
Modèle	0,43	0,42	**0,43**	**0,42**	0,21	0,18	**0,19**	**0,27**

TABLE 4.6 – Résultats du modèle d'ordonnancement dans le cas où les distracteurs sont ceux des autres items

Dans le cas où les exemples négatifs sont extraits du document, le modèle d'ordonnancement obtient un meilleur équilibre entre rappel et précision que les critères individuels, quel que soit le corpus. Il obtient notamment de meilleures performances que les mesures fondées sur WordNet utilisées par [Mitkov *et al.*, 2009].

Certains critères ont un meilleur rappel que le modèle global, mais ont une faible précision ou couverture. Ainsi, les mesures fondées sur les types ont un rappel élevé et une large couverture, mais de nombreux candidats se voient attribuer le même rang. Ce type de mesure sera donc intéressant pour filtrer les exemples de type sémantique différent de celui de la réponse, mais n'est pas un critère suffisant.

Les mesures fondées sur WordNet ont une faible couverture pour les entités nommées, mais des résultats du même ordre que les mesures fondées sur les types. Sur les autres réponses, la couverture est meilleure que pour les entités nommées, mais les performances moins bonnes, notamment du fait de l'ambiguïté des réponses.

Les mesures fondées sur corpus sont moins performantes que celles fondées sur des connaissances structurées pour les entités nommées, ce qui est attendu car il n'y a pas d'indication sur les relations sémantiques explicites, mais sur les autres réponses, elles obtiennent des résultats de même ordre, et sont plus performantes que les mesures fondées sur WordNet.

Le modèle est globalement moins performant pour les réponses non entités nommées, ce qui était attendu puisque les distracteurs sont alors associés à moins d'informations sémantiques, notamment sur leur type.

Les résultats dans le cas où les exemples négatifs sont issus des autres items vont dans le même sens, mais avec des performances moindres généralement.

4.4 Discussion

D'un point du vue du traitement automatique des langues, nous avons montré que la combinaison de méthodes complémentaires permet d'améliorer la reconnaissance de l'homogénéité sémantique. La reconnaissance des exemples similaires à la réponse pourrait cependant être enrichie. Ces travaux devront également être étendus pour traiter des types de réponse plus complexes, c'est-à-dire non limités à un groupe nominal ou entité nommée. Nous souhaiterions également pouvoir utiliser des ressources sémantiques de domaine, pour appliquer le système à des domaines de spécialité.

D'un point de vue pédagogique, nous avons mis en place un système de validation de distracteurs avec des résultats suffisamment fiables pour être intégré dans un environnement numérique de travail. Une interface web a d'ailleurs été développée pour ce système, que nous souhaitons intégrer dans Moodle. Le système a enfin commencé à être adapté au français, mais l'évaluation dans cette langue n'est pas encore complète.

Chapitre 5

Conclusion et perspectives

5.1 Conclusion

J'ai présenté dans ce manuscrit plusieurs axes de recherche que j'ai abordés comme des problèmes de proximité textuelle entre énoncés, pour reconnaître des relations sémantiques intra ou inter-énoncés. Les informations syntaxiques offrent un premier niveau de représentation permettant de s'éloigner de la formulation initiale des textes et d'ajouter un premier niveau de relation dans les énoncés. Les connaissances lexicales et sémantiques issues de corpus ou de bases de connaissances ajoutent des relations implicites entre énoncés, qui permettent d'évaluer leur proximité et de typer leurs relations.

Un point qui me semble essentiel est l'importance de disposer de ressources de qualité, correspondant au problème considéré, et permettant des évaluations reproductibles. Ainsi, lors de la constitution du schéma d'annotation de Cabernet et du corpus associé, nous avons constamment vérifié la cohérence des annotations afin de nous assurer de la qualité du schéma et du guide d'annotation associé. Lors de la thèse de Van-Minh Pho, nous avons également réfléchi à la mise en place d'un protocole d'évaluation de ses travaux qui n'implique pas nécessairement de passer par des tests auprès d'apprenants, afin de pouvoir reproduire facilement l'évaluation des différentes approches et comparer leurs performances.

Enfin, la notion de proximité sémantique entre énoncés est au cœur de nombreux domaines du TAL, mais n'est pas nécessairement explicite dans les travaux associés. Si la question de la proximité sémantique entre mots a été largement étudiée et est très actuelle du fait de la généralisation de l'utilisation de plongements lexicaux, les travaux sur la proximité sémantique entre énoncés sont plus épars (bases de paraphrases, tâches de reconnaissance d'implication textuelle ou de similarité textuelle...) bien que certains rapprochements soient faits, par exemple entre évaluation de la traduction automatique et implication textuelle depuis plusieurs années. Rendre explicite les rapprochements (et les différences) entre les différentes tâches du TAL permettrait une meilleure exploitation des

méthodes et ressources.

5.2 Perspectives

5.2.1 Généricité des modèles

L'un des enjeux du TAL est la généricité des modèles créés, que ce soit les représentations des connaissances, ou les méthodes : un «simple» changement de type de corpus peut rendre les modèles inadaptés, et il est difficile d'établir des modèles qui se généralisent à travers plusieurs domaines voire dans des langues différentes. Dans mes travaux, j'ai essayé de créer des représentations les plus génériques possibles ; les représentations définies pour les relations cliniques et les résultats expérimentaux s'inscrivant dans cette démarche, j'envisage de les valider en les utilisant dans des domaines différents de ceux traités jusqu'ici dans des projets à venir. De même, le corpus créé et annoté dans le cadre du projet Cabernet étant désormais finalisé, il sera possible de tester les outils développés pour l'extraction de relations en anglais et pour un jeu restreint de relations, avec un schéma plus complexe et une nouvelle langue. L'existence de corpus comparables en anglais et français permet en effet des transferts d'une langue à une autre, et il sera intéressant de pouvoir utiliser les deux langues pour confronter les connaissances acquises. La problématique du transfert entre langues est également au centre du projet Restaure, qui vise à fournir des ressources informatiques et des outils de traitement automatique pour trois langues régionales de France : alsacien, occitan et picard.

5.2.2 Ressources structurées et ressources textuelles

Le TAL établit de plus en plus de liens avec les ressources du web sémantique. Cependant la mise en commun des ressources structurées et textuelles reste difficile, comme le montrent les travaux en questions-réponses : la recherche d'informations hybride, c'est-à-dire dans des bases de connaissances et des textes, est une piste qui a peu été explorée : les systèmes hybrides actuels s'appuient principalement sur une source alors que la coopération des deux types de sources devrait être exploitée. La thèse de Romain Beaumont devrait aboutir à une représentation des questions permettant une réelle recherche hybride. Les rapprochements entre représentations, comme les travaux présentés dans la section 2.3 qui visent à rapprocher les relations issues des bases de connaissances et celles issues des textes, me semblent particulièrement intéressants. Cela est d'autant plus important que les bases de connaissances ne couvrent pas tous les types de relations, et les méthodes qui s'appuient sur ces ressources structurées ne sont pas adaptées à tout type de problème.

5.2.3 Interactions entre recherche en TAL et applications

De façon plus générale, et bien que ces aspects soient peu abordés dans ce manuscrit, qui se concentre sur les problématiques du TAL, le côté applicatif

répondant à des enjeux sociétaux (culturels, éducatifs, biomédicaux) est une motivation importante pour mon travail, qui se traduit notamment dans mes participations à des projets, par exemple avec le projet Restaure sur le traitement automatique des langues pour les langues régionales de France ou le projet Cabernet visant l'analyse automatique de dossiers électroniques patient.

En plus d'être une motivation importante, les tâches de TAL actuelles sont très guidées par les données, alors que les expériences que nous avons menées en simplification lexicales montrent que des méthodes s'adaptant à l'utilisateur sont plus performantes que celles fondées sur des connaissances a priori. Les méthodes et leur évaluation ne doivent donc pas se déconnecter des utilisateurs finaux. L'intégration de problématiques issues d'autres domaines apporte par ailleurs des questions de recherche originales pour le TAL. Ainsi, les applications pédagogiques ont des particularités qui en font des problèmes intéressants pour le TAL, comme nous avons pu le voir dans la thèse de Van-Minh Pho, et inversement les avancées du TAL permettent de se poser de nouvelles questions pédagogiques (correction automatique, génération d'exercices...) dès lors que l'on dispose d'outils performants.

Bibliographie

[Agirre *et al.*, 2016] Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R. et Wiebe, J. (2016). Semeval-2016 task 1 : Semantic textual similarity, monolingual and cross-lingual evaluation. *In Proc. of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, USA, June. Association for Computational Linguistics.*

[Ahn, 2006] Ahn, D. (2006). The stages of event extraction. *In Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, ARTE'06, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Albright *et al.*, 2013] Albright, D., Lanfranchi, A., Fredriksen, A., Styler, W., Warner, C., Hwang, J., Choi, J., Dligach, D., Nielsen, R., Martin, J., Ward, W., Palmer, M. et Savova, G. (2013). Towards comprehensive syntactic and semantic annotations of the clinical narrative. *J Am Med Inform Assoc*, 20(5):922–30.

[Aldabe *et al.*, 2006] Aldabe, I., De Lacalle, M. L., Maritxalar, M., Martinez, E. et Uria, L. (2006). Arikiturri : an automatic question generator based on corpora and nlp techniques. *In International Conference on Intelligent Tutoring Systems*, pages 584–594. Springer.

[Alex *et al.*, 2008] Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E., Matthews, M., Roebuck, S., Tobin, R. et Wang., X. (2008). Assisted Curation : does Text Mining Really Help ? *In Proceedings the Pacific Symposium on Biocomputing.*

[André, 2013] André, T. (2013). Génération automatique de distracteurs dans le cadre de QCM. Rapport technique, LIMSI.

[Angeli *et al.*, 2015] Angeli, G., Premkumar, M. J. et Manning, C. D. (2015). Leveraging linguistic structure for open domain information extraction. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing.*

[Banko *et al.*, 2007] Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M. et Etzioni, O. (2007). Open Information Extraction from the Web. *In IJCAI*, volume 7, pages 2670–2676.

[Berant *et al.*, 2013] Berant, J., Chou, A., Frostig, R. et Liang, P. (2013). Semantic Parsing on Freebase from Question-Answer Pairs. *In EMNLP*, pages 1533–1544.

[Bernard et Fontaine, 1982] Bernard, H. et Fontaine, F. (1982). *Les Questions à choix multiple : guide pratique pour la rédaction, l'analyse et la correction*. Université de Montréal, Service pédagogique,.

[Biran *et al.*, 2011] Biran, O., Brody, S. et Elhadad, N. (2011). Putting it simply : A context-aware approach to lexical simplification. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : Short Papers - Volume 2*, HLT '11, pages 496–501, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Björne *et al.*, 2010] Björne, J., Ginter, F., Pyysalo, S., Tsujii, J. et Salakoski, T. (2010). Complex event extraction at PubMed scale. *Bioinformatics*, 26:i382–i390.

[Blache, 2011] Blache, P. (2011). A computational model for linguistic complexity. *Biology, Computation and Linguistics. New Interdisciplinary Paradigms*, pages 155–167.

[Boschee *et al.*, 2005] Boschee, E., Weischedel, R. et Zamanian, A. (2005). Automatic information extraction. *In Proceedings of the International Conference on Intelligence Analysis*, volume 71. Citeseer.

[Bott *et al.*, 2012] Bott, S., Saggion, H. et Mille, S. (2012). Text Simplification Tools for Spanish. *In LREC*, pages 1665–1671.

[Bouamor *et al.*, 2016] Bouamor, D., Campillos-Llanos, L., Ligozat, A.-L., Rosset, S. et Pierre, Z. (2016). Transfer-Based Learning-to-Rank Assessment of Medical Term Technicality. *In LREC*.

[Brouwers *et al.*, 2012] Brouwers, L., Bernhard, D., Ligozat, A.-L. et François, T. (2012). Simplification syntaxique de phrases pour le français. *In Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2012)*, page 14p, Grenoble, France.

[Brouwers *et al.*, 2014] Brouwers, L., Bernhard, D., Ligozat, A.-L. et François, T. (2014). Syntactic sentence simplification for French. *In The 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2014)*.

[Brown *et al.*, 2005] Brown, J. C., Frishkoff, G. A. et Eskenazi, M. (2005). Automatic question generation for vocabulary assessment. *In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 819–826. Association for Computational Linguistics.

[Burton *et al.*, 1991] Burton, S. J., Sudweeks, R. R., Merrill, P. F. et Wood, B. (1991). How to prepare better multiple-choice test items : Guidelines for university faculty. *Brigham Young University Testing Services and the Department of Instructional Science*.

[Buyko *et al.*, 2009] Buyko, E., Faessler, E., Wermter, J. et Hahn, U. (2009). Event extraction from trimmed dependency graphs. *In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing : Shared Task*, BioNLP '09, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Callison-Burch *et al.*, 2011] Callison-Burch, C., Koehn, P., Monz, C. et Zaidan, O. F. (2011). Findings of the 2011 workshop on statistical machine translation. *In Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64. Association for Computational Linguistics.

[Candito *et al.*, 2010] Candito, M., Crabbé, B. et Denis, P. (2010). Statistical french dependency parsing : treebank conversion and first results. *In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1840–1847.

[Canning *et al.*, 2000] Canning, Y., Tait, J., Archibald, J. et Crawley, R. (2000). Cohesive Generation of Syntactically Simplified Newspaper Text. *In Text, Speech and Dialogue : Third International Workshop, TSD 2000*, pages 145–150, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Carroll *et al.*, 1998] Carroll, J., Minnen, G., Canning, Y., Devlin, S. et Tait, J. (1998). Practical simplification of English newspaper text to assist aphasic readers. *In Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

[Carroll *et al.*, 1999] Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S. et Tait, J. (1999). Simplifying Text for Language-Impaired Readers. *In Proceedings of EACL*, pages 269–270.

[Catach, 1985] Catach, N. (1985). *Les listes orthographiques de base du français*. Nathan, Paris.

[Chandrasekar *et al.*, 1996] Chandrasekar, R., Doran, C. et Srinivas, B. (1996). Motivations and methods for text simplification. *In Proceedings of the 16th conference on Computational linguistics*, pages 1041–1044.

[Chen et Manning, 2014] Chen, D. et Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 1, pages 740–750.

[Chu-Carroll *et al.*, 2012] Chu-Carroll, J., Fan, J., Boguraev, B., Carmel, D., Sheinwald, D. et Welty, C. (2012). Finding needles in the haystack : Search and candidate generation. *IBM Journal of Research and Development*, 56(3.4):6–1.

[Clarke *et al.*, 2002] Clarke, C. L., Cormack, G. V., Kemkes, G., Laszlo, M., Lynam, T. R., Terra, E. L. et Tilker, P. L. (2002). Statistical Selection of Exact Answers (MultiText Experiments for TREC 2002). *In TREC*.

[Considine *et al.*, 2005] Considine, J., Botti, M. et Thomas, S. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*, 12(1):19–24.

[Corney *et al.*, 2004] Corney, D., Buxton, B., Langdon, W. et Jones, D. (2004). BioRAT : extracting biological information from full-length papers. *Bioinformatics*, 20(17):3206.

[Coster et Kauchak, 2011a] Coster, W. et Kauchak, D. (2011a). Learning to Simplify Sentences Using Wikipedia. *In Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 1–9.

[Coster et Kauchak, 2011b] Coster, W. et Kauchak, D. (2011b). Simple English Wikipedia : a new text simplification task. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers-Volume 2*, pages 665–669. Association for Computational Linguistics.

[Cubric et Tosic, 2011] Cubric, M. et Tosic, M. (2011). Towards automatic generation of e-assessment using semantic web technologies. *International Journal of e-Assessment*.

[Cucerzan et Agichtein, 2005] Cucerzan, S. et Agichtein, E. (2005). Factoid Question Answering over Unstructured and Structured Web Content. *In TREC*, volume 72, page 90.

[Culotta et Sorensen, 2004] Culotta, A. et Sorensen, J. (2004). Dependency tree kernels for relation extraction. *In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 423. Association for Computational Linguistics.

[Dagan *et al.*, 2006] Dagan, I., Glickman, O. et Magnini, B. (2006). The pascal recognising textual entailment challenge. *In Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

[Dagan *et al.*, 2013] Dagan, I., Roth, D., Sammons, M. et Zanzotto, F. M. (2013). Recognizing textual entailment : Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

[Daiber *et al.*, 2013] Daiber, J., Jakob, M., Hokamp, C. et Mendes, P. N. (2013). Improving Efficiency and Accuracy in Multilingual Entity Extraction. *In Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*.

[De Belder et Moens, 2010] De Belder, J. et Moens, M.-F. (2010). Text Simplification for Children. *In Proceedings of the Workshop on Accessible Search Systems*.

[De Belder et Moens, 2012] De Belder, J. et Moens, M.-F. (2012). A dataset for the evaluation of lexical simplification. *In Computational Linguistics and Intelligent Text Processing*, pages 426–437. Springer.

[Deléger *et al.*, 2014] Deléger, L., Ligozat, A., Grouin, C., Zweigenbaum, P. et Névéol, A. (2014). Annotation of specialized corpora using a comprehensive entity and relation scheme. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 1267–1274.

[Denis et Sagot, 2009] Denis, P. et Sagot, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. *In Proceedings of PACLIC*.

[Downing, 2005] Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students : the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, 10(2):133–143.

[Drndarević *et al.*, 2013] Drndarević, B., Štajner, S., Bott, S., Bautista, S. et Saggion, H. (2013). Automatic Text Simplification in Spanish : A Comparative Evaluation of Complementing Modules. *In Computational Linguistics and Intelligent Text Processing*, pages 488–500. Springer.

[Dzodic *et al.*, 2004] Dzodic, V., Hervy, S., Fritsch, D., Khalfallah, H., Thereau, M. et Thomas, S. R. (2004). Web-based tools for quantitative renal physiology. *Cellular and Molecular Biology*, 50(7):795–800.

[El Ayari *et al.*, 2009] El Ayari, S., Grau, B. et Ligozat, A.-L. (2009). REVISE, un outil d'évaluation précise des systèmes questions-réponses. *In CORIA*, pages 385–396.

[El Ayari *et al.*, 2010] El Ayari, S., Grau, B. et Ligozat, A.-L. (2010). Fine-grained Linguistic Evaluation of Question Answering Systems. *In LREC*.

[Elhadad et Sutaria, 2007] Elhadad, N. et Sutaria, K. (2007). Mining a lexicon of technical terms and lay equivalents. *In Proceedings of the Workshop on BioNLP 2007 : Biological, Translational, and Clinical Language Processing*, pages 49–56. Association for Computational Linguistics.

[Fader *et al.*, 2013] Fader, A., Zettlemoyer, L. et Etzioni, O. (2013). Paraphrase-driven learning for open question answering. *In Association for Computational Linguistics (ACL)*.

[Feng, 2008] Feng, L. (2008). Text Simplification : A Survey. Technical report, City University of New York.

[Ferret *et al.*, 2000] Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., Jacquemin, C., Masson, N. et Lecuyer, P. (2000). Qalc–the question-answering system of limsi-cnrs. *In TREC*.

[Finkel *et al.*, 2005] Finkel, J. R., Grenager, T. et Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

[Foulonneau, 2011] Foulonneau, M. (2011). Generating educational assessment items from linked open data : the case of dbpedia. *In Extended Semantic Web Conference*, pages 16–27. Springer.

[François *et al.*, 2014] François, T., Gala, N., Watrin, P. et Fairon, C. (2014). FLELex : a graded lexical resource for French foreign learners. *In Proceedings of the International conference on Language Resources and Evaluation*, LREC '14, pages 3766–3773.

[François, 2012] François, T. (2012). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère.* Thèse de doctorat, Université Catholique de Louvain.

[Gala *et al.*, 2014] Gala, N., François, T., Bernhard, D. et Fairon, C. (2014). Un modèle pour prédire la complexité lexicale et graduer les mots. *In Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles*, TALN '14, pages 91–102.

[Garten et Altman, 2009] Garten, Y. et Altman, R. (2009). Pharmspresso : a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC bioinformatics*, 10(Suppl 2):S6.

[Gillick et Favre, 2009] Gillick, D. et Favre, B. (2009). A scalable global model for summarization. *In Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, ILP '09, pages 10–18.

[Glickman, 2006] Glickman, O. (2006). *Applied textual entailment.* Thèse de doctorat, Bar Ilan University.

[Goto *et al.*, 2010] Goto, T., Kojiri, T., Watanabe, T., Iwata, T. et Yamada, T. (2010). Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management & E-Learning : An International Journal (KM&EL)*, 2(3):210–224.

[Grappy *et al.*, 2011] Grappy, A., Grau, B., Falco, M.-H., Ligozat, A.-L., Robba, I. et Vilnat, A. (2011). Selecting answers to questions from web documents by a robust validation process. *In Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 55–62. IEEE Computer Society.

[Grau *et al.*, 2015] Grau, B., Ligozat, A.-L. et Gleize, M. (2015). Recherche d'information précise dans des sources d'information structurées et non structurées : défis, approches et hybridation. *Traitement Automatique des Langues.*

[Grouin *et al.*, 2010a] Grouin, C., Abacha, A. B., Bernhard, D., Cartoni, B., Deléger, L., Grau, B., Ligozat, A.-L., Minard, A.-L., Rosset, S. et Zweigenbaum, P. (2010a). CARAMBA : Concept, assertion, and relation annotation using machine-learning based approaches. *In i2b2 Medication Extraction Challenge Workshop.*

[Grouin *et al.*, 2010b] Grouin, C., Deléger, L. et Zweigenbaum, P. (2010b). Extracting medical information from narrative patient records : the case of medication-related information. *JAMIA*, 17:555–558.

[Haladyna et Downing, 1989] Haladyna, T. M. et Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied measurement in education*, 2(1):37–50.

[Haladyna *et al.*, 2002] Haladyna, T. M., Downing, S. M. et Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3):309–333.

[Hildebrandt *et al.*, 2004] Hildebrandt, W., Katz, B. et Lin, J. J. (2004). Answering definition questions using multiple knowledge sources. *In HLT-NAACL*, pages 49–56.

[Hoffmann *et al.*, 2011] Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L. et Weld, D. S. (2011). Knowledge-based weak supervision for information extraction of overlapping relations. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, pages 541–550. Association for Computational Linguistics.

[Horn *et al.*, 2014] Horn, C., Manduca, C. et Kauchak, D. (2014). Learning a Lexical Simplifier Using Wikipedia. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 458–463, Baltimore, Maryland, USA.

[Huang *et al.*, 2016] Huang, L., Taylor Cassidy, X. F., Ji, H., Voss, C. R., Han, J. et Sil, A. (2016). Liberal event extraction and event schema induction. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

[Hunter *et al.*, 2010] Hunter, P., Coveney, P. V., Bono, B. d., Diaz, V., Fenner, J., Frangi, A. F., Harris, P., Hose, R., Kohl, P., Lawford, P., McCormack, K., Mendes, M., Omholt, S., Quarteroni, A., Skår, J., Tegner, J., Thomas, S. R., Tollis, I., Tsamardinos, I., Beek, J. H. G. M. v. et Viceconti, M. (2010). A vision and strategy for the virtual physiological human in 2010 and beyond. *Phil. Trans. R. Soc. A*, 368:2595–2614.

[Inui *et al.*, 2003] Inui, K., Fujita, A., Takahashi, T., Iida, R. et Iwakura, T. (2003). Text simplification for reading assistance : a project note. *In Proceedings of the second international workshop on Paraphrasing*, pages 9–16.

[Jauhar et Specia, 2012] Jauhar, S. K. et Specia, L. (2012). Uow-shef : Simplex–lexical simplicity ranking based on contextual and psycholinguistic features. *In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 477–481. Association for Computational Linguistics.

[Jiang et Conrath, 1997] Jiang, J. J. et Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *In Proceedings of the 10th Research on Computational Linguistics International Conference (ROCLING/IJCLCLP)*, pages 19–33.

[Jonnalagadda et Gonzalez, 2010] Jonnalagadda, S. et Gonzalez, G. (2010). BioSimplify : an open source sentence simplification engine to improve recall in automatic biomedical information extraction. *In AMIA Annual Symposium Proceedings*.

[Kambhatla, 2004] Kambhatla, N. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. *In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 22. Association for Computational Linguistics.

[Karamanis *et al.*, 2006] Karamanis, N., Ha, L. A. et Mitkov, R. (2006). Generating multiple-choice test items from medical text : A pilot study. *In Proceedings of the fourth international natural language generation conference*, pages 111–113. Association for Computational Linguistics.

[Katz *et al.*, 2005] Katz, B., Borchardt, G. et Felshin, S. (2005). Syntactic and semantic decomposition strategies for question answering from multiple resources. *In Proceedings of the AAAI 2005 workshop on inference for textual question answering*, pages 35–41.

[Leacock et Chodorow, 1998] Leacock, C. et Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification. *WordNet : An electronic lexical database*, 49(2):265–283.

[Lee et Seneff, 2007] Lee, J. et Seneff, S. (2007). Automatic generation of cloze items for prepositions. *In INTERSPEECH*, pages 2173–2176.

[Lehmann *et al.*, 2013] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S. et Bizer, C. (2013). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 1(5).

[Lété *et al.*, 2004] Lété, B., Sprenger-Charolles, L. et Colé, P. (2004). Manulex : A grade-level lexical database from French elementary-school readers. *Behavior Research Methods, Instruments and Computers*, 36:156–166.

[Levy et Andrew, 2006] Levy, R. et Andrew, G. (2006). Tregex and Tsurgeon : tools for querying and manipulating tree data structures. *In Proceedings of LREC*, pages 2231–2234.

[Ligozat, 2013] Ligozat, A.-L. (2013). Question classification transfer. *In Proceedings of the Association for Computational Linguistics (ACL shorts papers)*.

[Ligozat *et al.*, 2012a] Ligozat, A.-L., Grau, B. et Tribout, D. (2012a). Morphological resources for precise information retrieval. *In TSD*.

[Ligozat *et al.*, 2013] Ligozat, A.-L., Grouin, C., Garcia-Fernandez, A. et Bernhard, D. (2013). Approches à base de fréquences pour la simplification lexicale. *TALN-RÉCITAL 2013*, page 493.

[Ligozat *et al.*, 2012b] Ligozat, A.-L., Tribout, D. et Grau, B. (2012b). Intérêt des ressources morphologiques pour la recherche d'information précise. *In Conférence en Recherche d'Information et Applications (CORIA 2012)*, pages 1–13.

[Lin, 1997] Lin, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 64–71. Association for Computational Linguistics.

[Lin *et al.*, 2016] Lin, Y., Shen, S., Liu, Z., Luan, H. et Sun, M. (2016). Neural relation extraction with selective attention over instances. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

[Lin *et al.*, 2007] Lin, Y.-C., Sung, L.-C. et Chen, M. C. (2007). An automatic multiple-choice question generation scheme for english adjective understanding. *In Workshop on Modeling, Management and Generation of Problems/Questions in eLearning, the 15th International Conference on Computers in Education (ICCE 2007)*, pages 137–142.

[Liu *et al.*, 2016] Liu, A., Soderland, S., Bragg, J., Lin, C. H., Ling, X. et Weld, D. S. (2016). Effective Crowd Annotation for Relation Extraction. *In Proceedings of NAACL-HLT 2016*.

[Liu *et al.*, 2005] Liu, C.-L., Wang, C.-H., Gao, Z.-M. et Huang, S.-M. (2005). Applications of lexical information for algorithmically composing multiple-choice cloze items. *In Proceedings of the second workshop on Building Educational Applications Using NLP*, pages 1–8. Association for Computational Linguistics.

[MacCartney, 2009] MacCartney, B. (2009). *Natural language inference*. Thèse de doctorat, Stanford University.

[Max, 2006] Max, A. (2006). *Writing for Language-Impaired Readers*, pages 567–570. Springer Berlin Heidelberg, Berlin, Heidelberg.

[McClosky *et al.*, 2011] McClosky, D., Surdeanu, M. et Manning, C. D. (2011). Event extraction as dependency parsing. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, pages 1626–1635. Association for Computational Linguistics.

[McDonald *et al.*, 2005] McDonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y. et White, P. (2005). Simple algorithms for complex relation extraction with applications to biomedical IE. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 491–498, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Medero et Ostendorf, 2011] Medero, J. et Ostendorf, M. (2011). Identifying Targets for Syntactic Simplification. *In Proceedings of the SLaTE 2011 workshop*.

[Mihalcea *et al.*, 2006] Mihalcea, R., Corley, C. et Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *In AAAI*, volume 6, pages 775–780.

[Milne et Witten, 2013] Milne, D. et Witten, I. H. (2013). An open-source toolkit for mining wikipedia. *Artificial Intelligence*, 194:222–239.

[Minard, 2012] Minard, A.-L. (2012). *Extraction de relations en domaine de spécialité*. Thèse de doctorat, Paris-Sud.

[Minard *et al.*, 2010] Minard, A.-L., Grau, B. et Ligozat, A.-L. (2010). Extraction de résultats expérimentaux d'articles scientifiques pour le peuplement d'une base de données. *In Journées internationales d'analyse statistique des données textuelles (JADT)*.

[Minard *et al.*, 2011a] Minard, A.-L., Ligozat, A.-L., Ben Abacha, A., Bernhard, D., Cartoni, B., Deléger, L., Grau, B., Rosset, S., Zwei-

genbaum, P. et Grouin, C. (2011a). Hybrid methods for improving information access in clinical documents : concept, assertion, and relation identification. *Journal of the American Medical Information Association*, 18(5):588–593.

[Minard *et al.*, 2011b] Minard, A.-L., Ligozat, A.-L. et Grau, B. (2011b). Apport de la syntaxe pour l'extraction de relations en domaine médical. *In Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, page 11p, Montpellier, France.

[Minard *et al.*, 2011c] Minard, A.-L., Ligozat, A.-L. et Grau, B. (2011c). Multi-class svm for relation extraction from clinical reports. *In Recent Advances in Natural Language Processing (RANLP)*, pages 604–609, Hissar, Bulgaria.

[Minard *et al.*, 2012] Minard, A.-L., Ligozat, A.-L. et Grau, B. (2012). Simplification de phrases pour l'extraction de relations. *In Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2012)*, page 14p, Grenoble, France.

[Minard *et al.*, 2011d] Minard, A.-L., Makour, L., Ligozat, A.-L. et Grau, B. (2011d). Feature selection for drug-drug interaction detection using machine-learning based approaches. *In DDIExtraction 2011. First Challenge Task : Drug-Drug Interaction Extraction - SEPLN 2011 satellite workshop*, pages 43–50, Huelva, Spain.

[Mintz *et al.*, 2009] Mintz, M., Bills, S., Snow, R. et Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

[Mitkov *et al.*, 2009] Mitkov, R., Ha, L. A., Varga, A. et Rello, L. (2009). Semantic similarity of distractors in multiple-choice tests : extrinsic evaluation. *In Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 49–56. Association for Computational Linguistics.

[Miwa *et al.*, 2010] Miwa, M., Sætre, R., Miyao, Y. et Tsujii, J. (2010). Entity-focused Sentence Simplification for Relation Extraction. *In Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 788–796, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Narayan et Gardent, 2014] Narayan, S. et Gardent, C. (2014). Hybrid Simplification using Deep Semantics and Machine Translation. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 435–445.

[Nelken et Shieber, 2006] Nelken, R. et Shieber, S. (2006). Towards robust context-sensitive sentence alignment for monolingual corpora. *In Proceedings of EACL*, pages 161–168.

[Nguyen et Grishman, 2015] Nguyen, T. H. et Grishman, R. (2015). Relation extraction : Perspective from convolutional neural networks. *In Proceedings of NAACL-HLT*, pages 39–48.

[Niu *et al.*, 2012] Niu, F., Zhang, C., Ré, C. et Shavlik, J. W. (2012). Deepdive : Web-scale knowledge-base construction using statistical learning and inference. *VLDS*, 12:25–28.

[Ogren *et al.*, 2008] Ogren, P., Savova, G. et Chute, C. (2008). Constructing evaluation corpora for automated clinical named entity recognition. *In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

[Oronoz *et al.*, 2015] Oronoz, M., Gojenola, K., Pérez, A., de Ilarraza, A. D. et Casillas, A. (2015). On the creation of a clinical gold standard corpus in spanish : Mining adverse drug reactions. *Journal of biomedical informatics*, 56:318–332.

[Paetzold et Specia, 2015] Paetzold, G. H. et Specia, L. (2015). LEXenstein : A Framework for Lexical Simplification. *In Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 85–90, Beijing, China.

[Paetzold et Specia, 2016] Paetzold, G. H. et Specia, L. (2016). Unsupervised Lexical Simplification for Non-Native Speakers. *In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*.

[Pais, 2013] Pais, S. (2013). *Asymmetric Distributional Similarity Measures to Recognize Textual Entailment by Generality*. Thèse de doctorat, Ecole Nationale Supérieure des Mines de Paris.

[Panchenko, 2016] Panchenko, A. (2016). Best of Both Worlds : Making Word Sense Embeddings Interpretable. *In LREC*.

[Papasalouros *et al.*, 2008] Papasalouros, A., Kanaris, K. et Kotis, K. (2008). Automatic generation of multiple choice questions from domain ontologies. *In IADIS e-Learning*, pages 427–434.

[Park *et al.*, 2015] Park, S., Kwon, S., Kim, B. et Lee, G. G. (2015). Isoft at qald-5 : Hybrid question answering system over linked data and text data. *In Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*.

[Patrick et Li, 2010] Patrick, J. et Li, M. (2010). High accuracy information extraction of medication information from clinical notes : 2009 i2b2 medication extraction challenge. *JAMIA*, 17:524–527.

[Pavlick et Callison-Burch, 2016] Pavlick, E. et Callison-Burch, C. (2016). Simple PPDB : A Paraphrase Database for Simplification. *In ACL*.

[Pavlick et Nenkova, 2015] Pavlick, E. et Nenkova, A. (2015). Inducing Lexical Style Properties for Paraphrase and Genre Differentiation. *In Proceedings of the 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL*, pages 218–224.

[Pavlick *et al.*, 2015] Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B. et Callison-Burch, C. (2015). Ppdb 2.0 : Better

paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. *In ACL (short papers)*.

[Peñas *et al.*, 2013] Peñas, A., Miyao, Y., Forner, P. et Kando, N. (2013). Overview of qa4mre 2013 entrance exams task. *In CLEF (Online Working Notes/Labs/Workshop)*, pages 1–6.

[Petersen et Ostendorf, 2007] Petersen, S. E. et Ostendorf, M. (2007). Text Simplification for Language Learners : A Corpus Analysis. *In Proceedings of SLaTE2007*, pages 69–72.

[Pho, 2015] Pho, V.-M. (2015). *Génération automatique de questionnaires à choix multiples pédadogiques : évaluation de l'homogénéité des options*. Thèse de doctorat, Université Paris-Sud.

[Ponzetto et Strube, 2007] Ponzetto, S. P. et Strube, M. (2007). Knowledge derived from wikipedia for computing semantic relatedness. *J. Artif. Intell. Res.(JAIR)*, 30:181–212.

[Rello *et al.*, 2013a] Rello, L., Baeza-Yates, R., Dempere-Marco, L. et Saggion, H. (2013a). Frequent words improve readability and short words improve understandability for people with dyslexia. *In Human-Computer Interaction–INTERACT 2013*, pages 203–219. Springer.

[Rello *et al.*, 2013b] Rello, L., Bayarri, C., Gòrriz, A., Baeza-Yates, R., Gupta, S., Kanvinde, G., Saggion, H., Bott, S., Carlini, R. et Topac, V. (2013b). DysWebxia 2.0 ! : more accessible text for people with dyslexia. *In Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, page 25. ACM.

[Riedel *et al.*, 2009] Riedel, S., Chun, H.-W., Takagi, T. et Tsujii, J. (2009). A markov logic approach to bio-molecular event extraction. *In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing : Shared Task*, pages 41–49. Association for Computational Linguistics.

[Riedel *et al.*, 2010] Riedel, S., Yao, L. et McCallum, A. (2010). Modeling Relations and Their Mentions without Labeled Text. *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

[Riedel *et al.*, 2013] Riedel, S., Yao, L., McCallum, A. et Marlin, B. M. (2013). Relation extraction with matrix factorization and universal schemas. *In Proceedings of NAACL-HLT 2013*.

[Roberts *et al.*, 2008] Roberts, A., Gaizauskas, R. et Hepple, M. (2008). Extracting clinical relationships from patient narratives. *In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 10–18. Association for Computational Linguistics.

[Roberts *et al.*, 2009] Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I. et Setzer, A. (2009). Building a semantically annotated corpus of clinical texts. *J Biomed Semantics*, 42:950–966.

[Roth et Yih, 2002] Roth, D. et Yih, W.-t. (2002). Probabilistic reasoning for entity & relation recognition. *In Proceedings of the 19th international*

conference on Computational linguistics-Volume 1, pages 1–7. Association for Computational Linguistics. based on shallow parsing but belief network and non kernel.

[Savova *et al.*, 2012] Savova, G., Styler, W., Albright, D., Palmer, M., Harris, D., Zaramba, G., Haug, P., Clark, C., Wu, S. et Ihrke, D. (2012). SHARP template annotations : Guidelines. Rapport technique, Mayo Clinic.

[Severance et Cohen, 2015] Severance, S. et Cohen, K. B. (2015). Measuring the readability of medical research journal abstracts. *In Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), ACL-IJCNLP 2015*, page 127.

[Siddharthan, 2006] Siddharthan, A. (2006). Syntactic Simplification and Text Cohesion. *Research on Language & Computation*, 4(1):77–109.

[Siddharthan, 2011] Siddharthan, A. (2011). Text simplification using typed dependencies : A comparison of the ro- bustness of different generation strategies. *In Proceedings of the 13th European Workshop on Natural Language Generation*.

[Siddharthan, 2014] Siddharthan, A. (2014). A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

[Specia *et al.*, 2012] Specia, L., Jauhar, S. K. et Mihalcea, R. (2012). Semeval-2012 task 1 : English lexical simplification. *In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 347–355. Association for Computational Linguistics.

[Specia *et al.*, 2010] Specia, L., Raj, D. et Turchi, M. (2010). Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.

[Štajner *et al.*, 2016] Štajner, S., Popovic, M., Saggion, H., Specia, L. et Fishel, M. (2016). Shared Task on Quality Assessment for Text Simplification. *In Proceedings of the LREC Workshop on Quality Assessment for Text Simplification (QATS), Portoroz, Slovenia*, pages 22–31.

[Styler IV *et al.*, 2014] Styler IV, W., Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P., Erickson, B., Miller, T., Lin, C., Savova, G. et Pustejovsky, J. (2014). Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

[Sumita *et al.*, 2005] Sumita, E., Sugaya, F. et Yamamoto, S. (2005). Measuring non-native speakers' proficiency of english by using a test with automatically-generated fill-in-the-blank questions. *In Proceedings of the second workshop on Building Educational Applications Using NLP*, pages 61–68. Association for Computational Linguistics.

[Sun *et al.*, 2013] Sun, W., Rumshisky, A. et Uzuner, O. (2013). Evaluating temporal relations in clinical text : 2012 i2b2 challenge. *J. of Amer. Med. Inform. Assoc.*, 20(5):806–813.

[Sung *et al.*, 2007] Sung, L.-C., Lin, Y.-C. et Chen, M. C. (2007). The design of automatic quiz generation for ubiquitous english e-learning system. *In Technology Enhanced Learning Conference (TELearn 2007), Jhongli, Taiwan*, pages 161–168.

[Tack, 2014] Tack, A. (2014). Annotation lexicale adaptée à différents niveaux utilisateurs. Stage de master 2, Université Catholique de Louvain, LIMSI, CNRS.

[Tack *et al.*, 2016a] Tack, A., François, T., Ligozat, A.-L. et Fairon, C. (2016a). Evaluating Lexical Simplification and Vocabulary Knowledge for Learners of French : Possibilities of Using the FLELex Resource. *In The 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

[Tack *et al.*, 2016b] Tack, A., François, T., Ligozat, A.-L. et Fairon, C. (2016b). Modèles adaptatifs pour prédire automatiquement la compétence lexicale d'un apprenant de français langue étrangère. *In 23ème Conférence sur le Traitement Automatique des Langues Naturelles (JEP-TALN-RECITAL 2016)*.

[Tarrant et Ware, 2008] Tarrant, M. et Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, 42(2):198–206.

[Touhami *et al.*, 2011] Touhami, R., Buche, P., Dibie-Barthélemy, J. et Ibanescu, L. (2011). An Ontological and Terminological Resource for n-ary Relation Annotation in Web Data Tables. *In 10th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2011)*, pages 662–679, Crete, Grèce.

[Tribout *et al.*, 2012] Tribout, D., Ligozat, A.-L. et Bernhard, D. (2012). Constitution automatique d'une ressource morphologique : Verbagent. *In CMLF 2012*.

[Usbeck et Ngomo, 2015] Usbeck, R. et Ngomo, A.-C. N. (2015). Hawk@ qald5–trying to answer hybrid questions with various simple ranking techniques. *In Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*.

[Uzuner *et al.*, 2010] Uzuner, Ö., Solti, I. et Cadag, E. (2010). Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

[Uzuner *et al.*, 2011] Uzuner, Ö., South, B. R., Shen, S. et DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

[Verga *et al.*, 2016] Verga, P., Belanger, D., Strubell, E., Roth, B. et McCallum, A. (2016). Multilingual relation extraction using compositional universal schema. *In Proceedings of NAACL-HLT 2016*.

[Vickrey et Koller, 2008] Vickrey, D. et Koller, D. (2008). Sentence Simplification for Semantic Role Labeling. *In ACL*, pages 344–352.

[Watanabe *et al.*, 2009] Watanabe, W. M., Junior, A. C., Uzêda, V. R., Fortes, R. P. d. M., Pardo, T. A. S. et Aluísio, S. M. (2009). Facilita : Reading assistance for low-literacy readers. *In Proceedings of the 27th ACM International Conference on Design of Communication*, SIGDOC '09, pages 29–36, New York, NY, USA. ACM.

[Wu et Weld, 2007] Wu, F. et Weld, D. S. (2007). Autonomously semantifying wikipedia. *In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 41–50. ACM.

[Wu et Weld, 2010] Wu, F. et Weld, D. S. (2010). Open information extraction using Wikipedia. *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127. Association for Computational Linguistics.

[Wu et Palmer, 1994] Wu, Z. et Palmer, M. (1994). Verbs semantics and lexical selection. *In Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.

[Wubben *et al.*, 2012] Wubben, S., van den Bosch, A. et Krahmer, E. (2012). Sentence Simplification by Monolingual Machine Translation. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1015–1024. Association for Computational Linguistics.

[Xu *et al.*, 2016] Xu, K., Reddy, S., Feng, Y., Huang, S. et Zhao, D. (2016). Question Answering on Freebase via Relation Extraction and Textual Evidence. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 2326–2336. Association for Computational Linguistics.

[Xu *et al.*, 2015a] Xu, W., Callison-Burch, C. et Napoles, C. (2015a). Problems in current text simplification research : New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

[Xu *et al.*, 2015b] Xu, Y., Ringlstetter, C., Kim, M.-Y., Goebel, R., Kondrak, G., Miyao, Y. et Gini, M. (2015b). A lexicalized tree kernel for open information extraction. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*.

[Yahya *et al.*, 2013] Yahya, M., Berberich, K., Elbassuoni, S. et Weikum, G. (2013). Robust question answering over the web of linked data. *In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1107–1116. ACM.

[Yao *et al.*, 2013a] Yao, X., Van Durme, B., Callison-Burch, C. et Clark, P. (2013a). Semi-Markov Phrase-Based Monolingual Alignment. *In EMNLP*, pages 590–600.

[Yao *et al.*, 2013b] Yao, X., Van Durme, B., Clark, P. et Callison-Burch, C. (2013b). Answer extraction as sequence tagging with tree edit distance. *In Proceedings of NAACL*.

[Yatskar *et al.*, 2010] Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C. et Lee, L. (2010). For the sake of simplicity : Unsupervised extraction of lexical simplifications from wikipedia. *In Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 365–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Yih *et al.*, 2015] Yih, W.-t., Chang, M.-W., He, X. et Gao, J. (2015). Semantic Parsing via Staged Query Graph Generation : Question Answering with Knowledge Base. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 1321–1331. Association for Computational Linguistics.

[Yih *et al.*, 2013] Yih, W.-T., Chang, M.-W., Meek, C. et Pastusiak, A. (2013). Question Answering Using Enhanced Lexical Semantic Models. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1744–1753, Sofia, Bulgaria. Association for Computational Linguistics.

[Zelenko *et al.*, 2003] Zelenko, D., Aone, C. et Richardella, A. (2003). Kernel methods for relation extraction. *Journal of machine learning research*, 3(Feb):1083–1106.

[Zeng *et al.*, 2015] Zeng, D., Liu, K., Chen, Y. et Zhao, J. (2015). Distant supervision for relation extraction via piecewise convolutional neural networks. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), Lisbon, Portugal*, pages 17–21.

[Zeng *et al.*, 2014] Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J. *et al.* (2014). Relation classification via convolutional deep neural network. *In COLING*, pages 2335–2344.

[Zhang *et al.*, 2012] Zhang, C., Niu, F., Ré, C. et Shavlik, J. (2012). Big data versus the crowd : Looking for relationships in all the right places. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers - Volume 1*, ACL '12, pages 825–834, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Zhou *et al.*, 2005] Zhou, G., Su, J., Zhang, J. et Zhang, M. (2005). Exploring various knowledge in relation extraction. *In Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 427–434. Association for Computational Linguistics.

[Zhu *et al.*, 2010] Zhu, Z., Bernhard, D. et Gurevych, I. (2010). A Monolingual Tree-based Translation Model for Sentence Simplification. *In Proceedings of COLING 2010*, pages 1353–1361, Beijing, China.

[Ziegler *et al.*, 2015] Ziegler, J., Gala, N., Brunel, A. et Mathilde, C. (2015). Text Simplification to increase Readability and facilitate Comprehension : a proof-of-concept pilot study. Workshop Brain and Language.