MULTILINGUISME ET TRAITEMENT DES LANGUES NATURELLES

Sous la direction de
Ismaïl BISKRI et Adel JEBALI

Presses de l'Université du Québec

MULTILINGUISME ET TRAITEMENT DES LANGUES NATURELLES

PRESSES DE L'UNIVERSITÉ DU QUÉBEC
Le Delta I, 2875, boulevard Laurier, bureau 450
Québec (Québec) G1V 2M2
Téléphone : 418-657-4399 • Télécopieur : 418-657-2096
Courriel : puq@puq.ca • Internet : www.puq.ca

Diffusion / Distribution :

CANADA et autres pays
PROLOGUE INC.
1650, boulevard Lionel-Bertrand
Boisbriand (Québec) J7H 1N7
Téléphone : 450-434-0306 / 1 800 363-2864

SUISSE
SERVIDIS SA
Chemin des Chalets
1279 Chavannes-de-Bogis
Suisse

FRANCE
AFPUD
SODIS

BELGIQUE
PATRIMOINE SPRL
168, rue du Noyer
1030 Bruxelles
Belgique

AFRIQUE
ACTION PÉDAGOGIQUE
POUR L'ÉDUCATION ET LA FORMATION
Angle des rues Jilali Taj Eddine
et El Ghadfa
Maârif 20100 Casablanca
Maroc

MULTILINGUISME ET TRAITEMENT DES LANGUES NATURELLES

Sous la direction de
Ismaïl BISKRI et Adel JEBALI

2010

Presses de l'Université du Québec
Le Delta I, 2875, boul. Laurier, bur. 450
Québec (Québec) Canada G1V 2M2

Catalogage avant publication de Bibliothèque et Archives nationales du Québec et Bibliothèque et Archives Canada

Vedette principale au titre :

Multilinguisme et traitement des langues naturelles

Comprend des réf. bibliogr.

ISBN 978-2-7605-2569-6

1. Recherche sur Internet. 2. Exploration de données (Informatique). 3. Informatique multilingue. 4. Interdisciplinarité. 5. Web 2.0. I. Biskri, Ismaïl. II. Jebali, Adel.

ZA4230.M84 2010 025.0425 C2010-940648-6

Nous reconnaissons l'aide financière du gouvernement du Canada par l'entremise du Fonds du livre du Canada pour nos activités d'édition.

La publication de cet ouvrage a été rendue possible grâce à l'aide financière de la Société de développement des entreprises culturelles (SODEC).

Intérieur
Mise en pages : Presses de l'Université du Québec

Couverture
Conception : Richard Hodgson

1 2 3 4 5 6 7 8 9 PUQ 2010 9 8 7 6 5 4 3 2 **1**

Dépôt légal – 3e trimestre 2010
Bibliothèque et Archives nationales du Québec / Bibliothèque et Archives Canada
Imprimé au Canada

AVANT-PROPOS

L'invention du Web version 2.0 et l'explosion des médias de communication facilitent plus que jamais l'accès à l'information et la démocratisation de ce dernier. Cependant, cette véritable révolution apporte son lot de problèmes dans des domaines comme celui du repérage de l'information et du forage de données (*data mining*), pour ne citer que ces derniers.

Dans ce contexte, deux stratégies de recherche semblent les plus aptes à relever ces nouveaux défis : le multilinguisme et la pluridisciplinarité.

Le multilinguisme s'impose du fait que toutes les langues participent désormais au Web dans un sens large et non seulement l'anglais comme c'était le cas au début de cette révolution informationnelle. Il s'impose également pour des raisons d'économie d'effort et de robustesse. Ainsi, les applications développées en prenant en compte un plus grand nombre de langues à la fois s'avèrent plus puissantes que celles qui n'ont à l'esprit qu'une seule langue. De plus, toute recherche portant sur les phénomènes langagiers ne peut se passer d'une étude comparative, une méthode léguée par la linguistique comparative du XIX^{e} siècle et fortement privilégiée par les courants linguistiques contemporains dans une perspective de grammaire universelle.

La pluridisciplinarité s'impose du fait que le domaine d'étude nécessite la maîtrise d'outils, de méthodes et de notions provenant de plusieurs disciplines scientifiques. Parmi ces dernières, citons la linguistique (théorique et appliquée), les sciences cognitives, l'intelligence artificielle et l'informatique sans oublier la logique, la psychologie et la philosophie.

TABLE DES MATIÈRES

CHAPITRE 1

LES MARQUEURS DE SUJET DE L'ARABE STANDARD
De l'analyse à l'implémentation

Adel Jebali

RÉSUMÉ

Dans cet article, nous étudions le phénomène des marqueurs de sujet de l'arabe standard. Nous les décrivons et nous les analysons en posant deux questions portant sur leur statut morphologique et leur statut morphosyntaxique. Nous modélisons cette analyse dans le cadre de la théorie HPSG et nous l'implémentons dans le système LKB.

1. INTRODUCTION

Les marqueurs de sujet (désormais MSu) de l'arabe standard (désormais AS) sont des morphèmes qui véhiculent des informations sur le genre, le nombre et la personne du sujet ou du topique. Ils sont agglutinés aux verbes accomplis et inaccomplis en respectant des schèmes morphologiques bien définis. Ils s'attachent ainsi comme suffixes à l'accompli et comme préfixes et/ou suffixes à l'inaccompli. Un exemple de ces unités est fourni en 1, où le MSu *-ta* de 2MS[1] est attaché au radical accompli du verbe *katab* « écrire ». En 2, le MSu *t-i*: de 2FS est attaché au radical inaccompli indicatif du même verbe.

(1)

katab	***-ta***	*maqa:l*	*-a*	*-n*
écrire.ACCO	-2MS	article	-ACC	-INDÉ

« Tu as écrit un article »

(2)

t-	*aktub*	***-i:***	*-na*	*maqa:l*	*-a*	*-n*
2-	écrire.INACC	-FS	-IND	article	-ACC	-INDÉ

« Tu (féminin) écris un article »

L'étude linguistique de ces marqueurs remonte aux premiers grammairiens de la langue arabe, dont notamment Sibawayhi au XIII^e siècle. Ils occupent également une place de choix dans les recherches linguistiques contemporaines sans pour autant constituer un objet de recherche à part entière. Ainsi, nous constatons qu'ils sont étudiés en évoquant des problématiques qui touchent le système d'accord, comme celle des asymétries (Aoun *et al.*, 1994 ; Harbert et Bahloul, 2002 ; Benmamoun et Lorimor, 2006). Jebali (2009) constitue le premier travail entièrement consacré à la modélisation de ces marqueurs, ce qui inclut la description, l'analyse linguistique et l'implémentation informatique.

Les problèmes posés par les MSu sont surtout d'ordre morphosyntaxique. La question qui se pose est la suivante : sont-ils mieux traités comme arguments ou comme marqueurs d'accord ? La terminologie qui sert à les désigner dans la littérature hésite entre les étiquettes disparates de pronoms (voire comme pronoms affixes), de marqueurs d'accord et de clitiques. Ces étiquettes ne sont le plus souvent pas basées sur des critères indépendamment motivés et les auteurs qui en usent ne s'interrogent généralement pas au sujet de l'impact de leur classification sur l'analyse d'autres phénomènes rencontrés en arabe standard et dans les dialectes.

Nous proposons donc une analyse linguistique des MSu basée sur des critères indépendants, nous évaluons l'impact de cette analyse sur d'autres phénomènes de la langue étudiée, nous la modélisons dans le cadre de la grammaire syntagmatique guidée par les têtes (HPSG) (Ginzburg et Sag, 2000 ; Pollard et

1. Nous utilisons les abréviations suivantes : 1 (1re personne), 2 (2e personne), 3 (3e personne), M (masculin), F (féminin), S (singulier), D (duel), P (pluriel), ACCO (accompli), INACC (inaccompli), IND (indicatif), NOM (nominatif), ACC (accusatif), INDÉ (indéfini).

Sag, 1994 ; Pollard et Sag, 1987 ; Sag *et al.*, 2003) et nous l'implémentons dans le système LKB *Linguistic Knowledge Building* (Copestake, 2002) pour en tester la validité.

Le texte qui suit se compose de trois grandes parties. Dans la section 2, nous établissons les faits dont il faudrait tenir compte en ce qui a trait aux MSu. Nous consacrons la section 3 à la modélisation de ces marqueurs en nous basant sur les outils théoriques et formels proposés par les théoriciens de HPSG. Finalement, la section 4 est consacrée à l'implémentation de l'analyse dans un système de compétence, LKB.

2. DESCRIPTION ET QUESTIONS DE RECHERCHE

Nous consacrons cette section à une présentation descriptive des MSu afin d'établir les faits à analyser. Cette présentation est scindée en deux parties suivant les deux aspects des verbes en AS, l'accompli et l'inaccompli. Nous présentons également les questions de recherche soulevées par ces données.

2.1. L'accompli

À l'accompli, les MSu sont suffixés au radical verbal et précèdent ainsi tout autre morphème (par exemple les marqueurs d'objet) qui pourrait être attaché au verbe, comme dans l'exemple 3, où le marqueur d'objet *-ha*: « la » est attaché à droite du MSu *-ta* (2MS), lui-même attaché au radical accompli *katab* « écrire ».

(3)

hal	*katab*	***-ta***	*-ha:?*
est-ce-que	écrire.ACCO	-2MS	-la

« L'as-tu écrite ? »

Une liste complète des MSu de l'accompli est présentée dans le tableau 1.1, où nous avons consigné la conjugaison du verbe *katab* « écrire » à l'accompli. Dans ce tableau, les MSu sont écrits en caractères gras.

Tableau 1.1 – **L'accompli du verbe *katab* « écrire »**

	Singulier	**Duel**	**Pluriel**
1	*katab-**tu***	*katab-**na:***	*katab-**na:***
2M	*katab-**ta***	*katab-**tuma:***	*katab-**tum***
2F	*katab-**ti***	*katab-**tuma:***	*katab-**tunna***
3M	*katab-**a***	*katab-**a:***	*katab-**u:***
3F	*katab-**at***	*katab-**ata:***	*katab-**na***

2.2. L'inaccompli

Nous partons d'un exemple qui illustre la forme la plus simple d'un verbe inaccompli. De la racine /ktb/ « écrire », par exemple, est issu le verbe inaccompli indicatif *jaktubu* « il écrit ». Le radical étant *aktub*, nous constatons la présence d'un préfixe *j-* et d'un suffixe *-u*. Le préfixe est le MSu et il exprime les traits de la 3MS. Le suffixe est le morphème de mode.

Dans certaines formes, la composition de l'inaccompli est plus complexe. À la 2FS, par exemple, l'inaccompli indicatif issu de la racine /ktb/ est *taktubi:na*. Dans cette forme, nous reconnaissons la présence du radical commun *aktub*. Nous pouvons également isoler le morphème -na et postuler qu'il est le morphème de mode pour la simple raison qu'il est tronqué au subjonctif et au jussif dans la forme commune *taktubi:*. Ce morphème est également présent à la 2MP (t*aktubu:na*) et à la 3MP (*jaktubu:na*) de l'indicatif et subit les mêmes changements au subjonctif et au jussif (*taktubu:* et *jaktubu:* respectivement).

Au duel, le morphème que nous pouvons postuler qu'il est le morphème de mode est -ni, visible à l'indicatif de la 2D *taktuba:-ni* « vous [deux] écrivez », par exemple, et tronqué au subjonctif et au jussif dans la forme commune *taktuba:*.

2.3. Le statut morphosyntaxique des MSu

La question morphosyntaxique à propos des MSu se pose dans ces termes : ces marqueurs sont-ils mieux traités en tant qu'arguments[2], comme le soutiennent les grammairiens de la tradition, Akkal (1996), Fassi Fehri (1993) et Lumsden et Haleform (2003), ou en tant que marqueurs d'accord, comme le soutiennent Aoun *et al.* (1994), Fleisch (1979) et Shlonsky (1997) ?

Pour répondre à cette question, il faudrait passer en revue les contextes syntaxiques dans lesquels ces marqueurs sont employés. Nous pouvons classer ces contextes en deux : le redoublement et la dislocation.

2.3.1. Le redoublement des MSu

Le fait d'affirmer qu'un MSu est un argument ou un marqueur d'accord présuppose une analyse du rôle syntaxique joué par le SN (syntagme nominal) ou le pronom indépendant antécédent si ces derniers sont également présents dans la structure. Et puisque c'est souvent le cas en arabe, il faudrait toujours entreprendre une telle analyse.

2. Nous définissons le marqueur d'accord comme étant un élément dépourvu de fonction syntaxique propre. Cet élément réfère à un autre, avec lequel il partage certains traits grammaticaux dans une configuration donnée (Auger, 1994, p. 27).

Le redoublement est entendu ici comme étant la coexistence du MSu et d'un constituant (SN ou pronom indépendant) associé (= co-indicé / coréférentiel) à ce marqueur, à condition que ce constituant soit à droite du MSu et qu'il n'y ait pas de coupure intonatoire qui placerait le constituant en question à la périphérie de la phrase. Cette coupure est souvent notée par une virgule à l'écrit et marquée par une pause à l'oral, comme le notent Philippaki-Warburton *et al.* (2002).

En prenant cette définition en compte, nous constatons que les MSu apparaissent souvent dans des structures que nous pouvons qualifier de redoublement. Ces structures, où le verbe est initial (VSO), sont neutres d'un point de vue pragmatique et le SN n'y est séparé du verbe (et du MSu) par aucune coupure intonatoire. C'est le cas de l'exemple 4, où le SN postverbal *l-walad-u* « le garçon » est postposé au MSu de 3M *-a*.

Nous ne pouvons cependant pas parler de redoublement dans ce cas pour plusieurs raisons. Tout d'abord, le SN et le MSu partagent certains traits, dont le genre (masculin) et la personne (troisième). Le nombre est sous-spécifié sur le marqueur dans les structures VSO et garde toujours une valeur (par défaut) de singulier que le SN soit au singulier, au duel ou au pluriel[3]. Les deux exemples suivants comportent des SN au duel (exemple 5) et au pluriel (exemple 6). On constate que le MSu est le même que dans l'exemple 4, où le SN est au singulier.

(4)	*katab*	***-a***	*l-*	*walad*	*-u*	*maqa:l*	*-a*	*-n*
	écrire.ACCO	-3M	le-	garçon	-NOM	article	-ACC	-INDÉ
	« Le garçon a écrit un article »							
(5)	*katab*	***-a***	*l-*	*walada:ni*		*maqa:l*	*-a*	*-n*
	écrire.ACCO	-3M	les-	garçons.D.NOM		article	-ACC	-INDÉ
	« Les deux garçons ont écrit un article »							
(6)	*katab*	*-**a***	*l-*	*?awla:d*	*-u*	*maqa:l*	*-a*	*-n*
	écrire.ACCO	-3M	les-	garçons	-NOM	article	-ACC	-INDÉ
	« Les garçons ont écrit un article »							

(7)	***man***	*katab*	*-a*	*maqa:l*	*-a*	*-n*	*?*
	qui	écrire.ACCO	-3M	article	-ACC	-INDÉ	
	« Qui a écrit un article ? »						

Ensuite, pour qu'il y ait redoublement, le critère suivant, appelé critère de l'extraction, doit être respecté : un SN qui redouble un MSu ne peut être « extrait », comme le montrent Aoun et Benmamoun (1998). Toutefois, les SN dans les trois derniers exemples peuvent l'être quand ils sont remplacés par des éléments Qu « Wh ». Dans l'exemple suivant, l'élément Qu *man* « qui » entretient une relation à longue distance avec le MSu *-a*, ce qui viole la condition susmentionnée.

3. Cela constitue l'une des asymétries de l'accord.

Nous ne pouvons donc parler de redoublement. Il s'agit en fait d'une relation d'accord entre le verbe et le sujet. L'accord sur le verbe est marqué par un MSu dont le nombre est sous-spécifié. Cette relation équivaut à ce que Bresnan et Mchombo (1987) appellent « accord grammatical[4] ». Un accord anaphorique est observé dans l'ordre dit SVO, auquel nous consacrons la sous-section suivante.

2.3.2. La dislocation

Certains composants co-indicés avec les MSu peuvent être disloqués[5]. Ce sont des SN ou des pronoms indépendants nominatifs et ils peuvent être disloqués à droite ou à gauche. Dans le cas 8, par exemple, le pronom indépendant *hum* « eux » co-indicé avec le MSu de 3MP *-u*: est disloqué à gauche. À l'exemple 9, le SN *?at'-t'alabat-u* « les étudiants » co-indicé avec le MSu de 3MP *-u*: est disloqué à gauche également.

(8) ***hum*** *katab* *-u:* *maqa:l* *-a* *-n*
eux écrire.ACCO -3MP article -ACC -INDÉ
« Eux, ils ont écrit un article »

(9) ***?at'*** ***-t'alabat*** ***-u*** *katab* *-u:* *maqa:l* *-a* *-n*
les- étudiants -NOM écrire.ACCO -3MP article -ACC -INDÉ
« Les étudiants, ils ont écrit un article »

Le tableau 1.2 résume les différences entre l'accord grammatical et l'accord anaphorique en arabe.

Tableau 1.2 – **Accord grammatical et accord anaphorique en arabe standard**

		Accord grammatical	**Accord anaphorique**
MSu	Statut	Marqueur d'accord	Argument (sujet)
	Traits	[-nombre]	[+nombre]
	Pronominal	-	+
SN/ProInd		Argument (sujet), seulement SN	Non-argument (topique, focus)

4. Les tests proposés par les auteurs pour distinguer l'accord grammatical de l'accord anaphorique concourent à justifier cette hypothèse.
5. Ce terme n'implique aucun déplacement syntaxique. La dislocation s'apparente au redoublement, mais s'en distingue par une coupure intonatoire qui détache le SN ou le pronom du reste de la phrase, ce dernier pouvant alors se retrouver à la périphérie gauche ou à la périphérie droite (Philippaki-Warburton *et al.*, 2002, p. 58). Cette structure est associée à des effets « pragmatiques », comme le contraste, l'emphase ou le changement de topique comme le soutient avec raison Auger (1994).

Cette vision de la morphosyntaxe des MSu a le mérite de clarifier certaines questions, dont celle tant débattue des asymétries de l'accord. En fait, ces asymétries suivant l'ordre de surface (VSO/SVO) ne sont en réalité que l'expression de la cohabitation de deux systèmes d'accord : un accord anaphorique et un accord grammatical. En outre, les MSu couvrent deux sortes d'unités : des marqueurs d'accord, dont le seul rôle est de reprendre, de manière redondante, certains traits du sujet ; et des pronoms incorporés (Bresnan et Mchombo, 1987), dont le rôle syntaxique est celui du sujet. Quatre des MSu se trouvent ainsi en situation d'ambiguïté fonctionnelle (Bresnan et Mchombo, 1987) et ce sont les MSu de la troisième personne du singulier : *-a*, *-at*, *j-* et *t-*. Ces marqueurs sont regroupés dans le tableau 1.3.

Tableau 1.3 – **MSu fonctionnellement ambigus**

	Masculin	**Féminin**
Accompli	*-a*	*-at*
Inaccompli	*j-*	*t-*

Nous pouvons ainsi postuler que les MSu ont un statut morphosyntaxique ambigu entre celui d'arguments et celui de simples marqueurs d'accord. Cela n'implique pas que seuls les marqueurs d'accord ont un statut morphologique d'affixes. Nous verrons dans la section subséquente que tous les MSu sont des affixes.

2.4. Le statut morphologique des MSu

Nous avons des raisons de penser que tous les MSu, qu'ils soient des arguments ou des marqueurs d'accord, sont mieux traités comme affixes flexionnels. Les tests proposés par Zwicky (1977, 1985a, 1985b, 1987) et Zwicky et Pullum (1983), ainsi que par Miller (1992) et Anderson (2005) concourent à justifier l'étiquette affixale. Cette approche est appuyée par les tests suivants : les idiosyncrasies, la coordination et la sélection des hôtes.

2.4.1. Les idiosyncrasies

Selon Miller (1992) et Zwicky et Pullum (1983), les idiosyncrasies sont caractéristiques des combinaisons [radical+affixe]. Cette vision de la morphologie (et donc du lexique), comme étant le lieu des irrégularités, est très caractéristique de la grammaire générative, dans le sens le plus large de ce terme. Les affixes, étant des éléments attachés dans le lexique, exhibent plusieurs irrégularités que les éléments introduits par des contraintes syntagmatiques n'exhibent pas. Parmi ces idiosyncrasies, Zwicky et Pullum (1983) ont signalé des idiosyncrasies morphophonologiques et sémantiques, ainsi que la présence des trous arbitraires dans l'ensemble des combinaisons possibles. Miller (1992) a, avec raison, fortement

critiqué la validité des idiosyncrasies sémantiques et a également amoindri l'importance des trous arbitraires. Il a toutefois utilisé ce critère en conjonction avec d'autres tests dans le but de démontrer le statut affixal des clitiques du français. Nous optons pour cette solution et nous exposons ici certains trous arbitraires qui caractérisent les combinaisons [hôte+MSu]. L'existence de ces trous appuie l'idée selon laquelle les MSu présentent des propriétés typiques d'affixes.

Les combinaisons des radicaux avec des MSu connaissent certains trous arbitraires dans les formes possibles. Tout comme en anglais, où aucun affixe du participe passé ne peut s'attacher au radical du verbe *stride*, nous constatons qu'en arabe standard aucun MSu de l'accompli ne peut s'attacher au radical dérivé de la racine verbale /wd3/ « laisser ». La conjugaison accomplie n'étant pas possible pour ce verbe, un autre verbe (avec un sens proche) est employé : *taraka* « laisser ». Ce trou est d'autant plus surprenant que la conjugaison inaccomplie est possible (et par conséquent l'impératif également) : *j-adaQ-u* « il laisse ». Un autre exemple est celui de la racine /wDr/ « laisser », dont on ne peut obtenir la forme verbale conjuguée à l'accomplie.

2.4.2. La coordination

La coordination est très révélatrice de la structure syntaxique ; elle dénote généralement une parité syntaxique entre les éléments coordonnés et peut donc constituer un critère crucial dans la décision concernant le statut des éléments qui entrent dans ce type de relation, ou qui ne peuvent le faire. Selon Zwicky (1977), les règles syntaxiques, l'effacement sous identité par exemple, n'affectent pas les affixes et peuvent affecter les mots. Miller (1992) récupère ce critère, en garde une interprétation en termes de portée (large *vs* restreinte), et le dissout en deux sous-critères :

A. Un item qui ne peut avoir une portée large sur une coordination d'hôtes est un affixe (*Ibid*., p. 155).

C. Quand la répétition sur chaque conjoint est obligatoire, l'item en question est un affixe (*Ibid*., p. 157).

Ces deux critères reviennent à affirmer qu'un affixe ne peut avoir une portée large sur une coordination d'hôtes ou, dans les termes de Zwicky (1977), un affixe ne peut être effacé sous identité. L'application de ce critère démontre que les MSu se comportent comme les affixes à cet égard, comme le montrent les contrastes suivants, où le MSu de 1S *-tu* doit être répété sur son hôte verbal et ne peut avoir une large portée sur une coordination d'hôtes, d'où l'agrammaticalité des exemples 11 et 12.

2.4.3. La sélection des hôtes

Selon Zwicky et Pullum (1983), les affixes entretiennent des relations spéciales avec leurs hôtes. Ils sont ainsi très sélectifs et n'apparaissent généralement qu'avec une classe limitée d'hôtes. En arabe, les MSu sont très sélectifs de leurs radicaux et ne s'attachent qu'à des radicaux verbaux. Nous pouvons par conséquent conclure qu'ils présentent des symptômes compatibles avec un statut affixal.

Le statut des MSu de l'arabe standard est celui d'affixes. Ils s'attachent ainsi à leurs hôtes à un stade présyntaxique et ne doivent par conséquent pas être sensibles aux contraintes syntagmatiques. Leur syntaxe est toutefois celle d'arguments dans certains cas, ce qui pourrait poser un problème à leur formalisation, puisque la théorie doit disposer de moyens qui lui permettent de rendre compte de cette hétérogénéité apparente.

(10) *?akal* ***-tu*** *wa* *farib* ***-tu***
mange.ACCO -1S et boire.ACCO -1S
« J'ai mangé et [j']ai bu »

(11) **?akal* ***-tu*** *wa* *farib*
manger.ACCO -1S et boire.ACCO

(12) **?akal* *wa* *farib* ***-tu***
manger.ACCO et boire.ACCO -1S

3. MODÉLISATION

Les faits suivants constituent le domaine empirique de référence en ce qui a trait à la modélisation des MSu :

1. Les MSu sont des affixes et doivent donc être traités comme tels par la théorie.
2. Ces marqueurs couvrent deux réalités différentes en ce qui a trait à quatre d'entre eux (les marqueurs de 3S) : des arguments et des non-arguments.
3. Les arguments sont des pronoms affixes pleinement spécifiés pour leurs traits d'accord.
4. Les non-arguments, en l'occurrence les marqueurs d'accord, ne sont pas spécifiés pour le nombre et peuvent donc être unifiés avec des SN sujets dont la valeur de l'attribut NBRE est variable.
5. Ces propriétés soulèvent des questions sur l'ordre de surface et son influence sur le MSu sélectionné par la tête verbale et sur l'accord entre le sujet et le verbe.

Nous proposons une modélisation des MSu qui tient compte des propriétés du domaine empirique et qui reste fidèle au modèle adopté, la théorie HPSG, proposant une analyse lexicaliste, dont les ingrédients sont les règles lexicales et une hiérarchie de types à héritage multiple.

3.1. La modélisation de la flexion verbale

Les MSu sont des affixes flexionnels du verbe et ils s'amalgament à d'autres affixes verbaux, comme les marqueurs de mode et d'aspect.

Pour modéliser la flexion verbale, nous proposons trois règles lexicales flexionnelles[6] qui relient un radical (pouvant être considéré comme un lexème) à un mot bien formé (le verbe fléchi). Ces règles ont pour intrant un objet de type radical et pour extrant un objet de type verbe-mot.

La forme générale d'une règle lexicale flexionnelle est la suivante :

(13)
$$\begin{bmatrix} \textit{rflex} - \textit{verbe} & \\ \text{INTRANT} & [\textit{radical}] \\ \text{EXTRANT} & [\textit{verbe} - \textit{mot}] \end{bmatrix}$$

Le type auquel appartient cette règle est *r-flex-verbe*. Trois sous-types sont dominés par ce type : *r-flex-acco*, *r-flex-inacc1* et *r-flex-inacc2*. Les règles du premier type servent à concaténer un suffixe flexionnel au radical de l'accompli, les règles du deuxième type à concaténer un préfixe flexionnel et un suffixe de mode au radical de l'inaccompli et les dernières concatènent un préfixe et un suffixe flexionnel en plus d'un suffixe de mode au radical de l'inaccompli.

(14)
$$\begin{bmatrix} r - flex - acco & \\ \text{INTRANT} & \left[\textit{radical}\ \boxed{1} \right] \\ \text{EXTRANT} & \begin{bmatrix} \textit{verbe} - \textit{mot} & \\ \text{PHON} & \langle \oplus \text{SUFFIXE} \rangle \\ \text{ASPECT} & \textit{accompli} \\ \text{TÊTE} & \boxed{1} \\ \text{STR-ARG} & \langle ... \rangle \end{bmatrix} \end{bmatrix}$$

6. La conception de ces règles, dans notre travail, est dans la lignée de ce qu'ont proposé Sag *et al.* (2003).

(15)
$$\left[\begin{array}{ll} r-flex-inacc1 & \\ \text{INTRANT} & \left[radical\ \boxed{1}\right] \\ \text{EXTRANT} & \left[\begin{array}{ll} verbe-mot & \\ \text{PHON} & \left\langle \text{PRÉFIXE} \oplus \boxed{1} \oplus \text{SUFFIXE-MODE} \right\rangle \\ \text{ASPECT} & inaccompli \\ \text{TÊTE} & \boxed{2} \\ \text{STR-ARG} & \langle ... \rangle \end{array}\right] \end{array}\right]$$

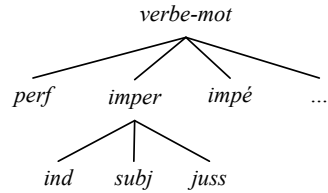
(16)
$$\left[\begin{array}{ll} r-flex-inacc2 & \\ \text{INTRANT} & \left[radical\ \boxed{1}\right] \\ \text{EXTRANT} & \left[\begin{array}{ll} verbe-mot & \\ \text{PHON} & \left\langle \text{PRÉFIXE} \oplus \boxed{1} \oplus \text{SUFFIXE} \oplus \text{SUFFIXE-MODE} \right\rangle \\ \text{ASPECT} & inaccompli \\ \text{TÊTE} & \boxed{2} \\ \text{STR-ARG} & \langle ... \rangle \end{array}\right] \end{array}\right]$$

La valeur de l'attribut STR-ARG (pour structure argumentale) n'est pas spécifiée dans ces règles et fera l'objet d'une discussion plus loin.

3.1.1. Les modes de l'inaccompli

Les marqueurs de mode sont introduits par les deux règles lexicales *r-flex-inacc1* et *r-flex-inacc2*. Cependant, seuls les verbes au mode indicatif, au singulier et à la 1P du mode subjonctif portent des marqueurs de mode. Au jussif et aux autres personnes du subjonctif, les marques de mode sont absentes. Ce comportement suggère une certaine sensibilité des marqueurs de mode aux traits morphosyntaxiques en ce qui a trait au subjonctif. L'indicatif et le jussif sont semblables en ceci que le premier possède des marqueurs de mode alors que le deuxième n'en a pas.

Figure 1.1 – **Une hiérarchie de types pour les verbes de l'arabe**



Pour modéliser ce comportement, nous proposons la hiérarchie de types présentée dans la figure 1.1.

Les contraintes des exemples 17, 18 et 19 sont celles que notre grammaire impose à ces types de verbes inaccomplis :

(17) $\begin{bmatrix} ind \\ \text{MORPH} \ [\text{SUFFIXE-MODE} \ +] \end{bmatrix}$

(18) $\begin{bmatrix} juss \\ \text{MORPH} \ [\text{SUFFIXEMODE} \ -] \end{bmatrix}$

(19) $\begin{bmatrix} subj \\ \text{MORPH} \ [\text{SUFFIXE-MODE} \ \alpha[bool]] \end{bmatrix}$

où α = + si [NBRE *sing*] ou [NBRE *pluriel*] et [PERS *1*], sinon α = -

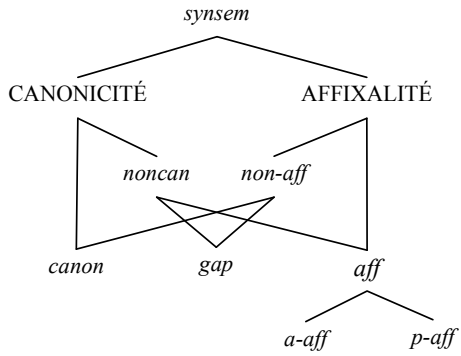
Ces contraintes garantissent que chaque verbe inaccompli sera doté des affixes de mode appropriés à son type. Les verbes jussifs ne contiennent pas de tels affixes du fait de leur appartenance au type *juss*. Les verbes à l'indicatif et au subjonctif comportent les affixes de mode appropriés à leurs types respectifs *ind* et *subj*. Ces affixes interagissent avec les traits morphosyntaxiques exprimés par les MSu, que nous modélisons dans le paragraphe suivant.

3.1.2. Les traits morphosyntaxiques

Nous avons, jusqu'à présent, modélisé l'aspect proprement morphophonologique de l'affixation des marqueurs de mode. Les traits exprimés par les MSu et leur effet sur la structure syntagmatique ne sont pas pris en compte. Une manière d'y arriver est de modifier les règles lexicales flexionnelles pour introduire les traits de genre, de nombre et de personne. Ces traits sont en relation avec la structure argumentale du verbe puisque la réalisation du trait de nombre permet à l'affixe d'apparaître dans la valeur de la liste STR-ARG (en tant qu'argument). Un affixe non spécifié pour le nombre n'est pas un argument et ne constitue donc pas un membre de cette liste.

Pour modéliser cette analyse, nous faisons appel à la typologie de Miller et Sag (1997), à savoir leur distinction entre des types canoniques et des types non canoniques. Dans cette typologie, le type *aff* (pour les affixes) est un type non canonique puisqu'il peut correspondre à des éléments membres de la structure argumentale qui ne sont pas réalisés en tant que dépendants syntaxiques de la tête. Cette typologie est représentée dans la figure 1.2.

Figure 1.2 – **Une hiérarchie partielle des types proposée par Miller et Sag (1997)**



Conjointement à cette hiérarchie partielle, nous proposons d'imposer la contrainte pour l'exemple 20 au type verbe-mot :

(20) $\begin{bmatrix} \textit{verbe-mot} & \\ \text{MORPH} & \begin{bmatrix} \text{FORM} & F_{\textit{praf}}(\boxed{1},\ldots) \\ \text{I-FORM} & \boxed{1} \end{bmatrix} \end{bmatrix}$

Cette contrainte morphologique nous garantit que la valeur de l'attribut FORM est le résultat de l'application d'une fonction phonologique dont l'un des arguments est la valeur de l'attribut I-FORM. Cette contrainte exige ainsi que la valeur de l'attribut FORM soit reliée à la valeur de l'attribut I-FORM par le biais de l'affixation flexionnelle appropriée[7].

Nous proposons que le type verbe-mot possède deux sous-types selon le type de l'affixe flexionnel qui y est attaché : verbe-mot-accord et verbe-mot-MS. Les deux contraintes suivantes sont imposées à ces deux types :

(21) $\begin{bmatrix} \textit{verbe-mot-accord} & \\ \text{TÊTE|ACCRD} & \begin{bmatrix} \text{GRE} & \textit{gre} \\ \text{PERS} & \textit{pers} \end{bmatrix} \\ \text{VAL} & \begin{bmatrix} \text{SUJ} & \langle \boxed{1} \rangle \\ \text{COMPS} & \langle \ldots \rangle \end{bmatrix} \\ \text{STR-ARG} & \langle \boxed{1},\ldots \rangle \end{bmatrix}$

7. Voir Miller et Sag (1997).

(22)
$$\begin{bmatrix} \textit{verbe-mot-MS} & \\ \text{TÊTE|ACCRD} & \begin{bmatrix} \text{GRE} & \textit{gre} \\ \text{PERS} & \textit{pers} \\ \text{NBRE} & \textit{nbre} \end{bmatrix} \\ \text{VAL} & \begin{bmatrix} \text{SUJ} & \langle\ \rangle \\ \text{COMPS} & \langle ... \rangle \end{bmatrix} \\ \text{STR-ARG} & \langle \boxed{\textit{aff}}, ... \rangle \end{bmatrix}$$

Les contraintes sur le type verbe-mot-accord assurent la réalisation d'un sujet lexical (apparaissant à la fois dans la liste de valence et dans la structure argumentale) quand le marqueur de sujet est sous-spécifié pour le nombre.

Les contraintes sur le type verbe-mot-MS assurent la réalisation du sujet (le premier membre de la liste STR-ARG) en tant qu'affixe. Dans ce cas-ci, l'attribut NBRE est présent dans les traits de tête du verbe.

La différence entre ces deux sous-types implique que les verbes arabes munis de MSu de 3MS et de 3FS aient deux entrées lexicales différentes selon leur appartenance à l'un ou à l'autre de ces deux sous-types. Ainsi, le verbe *kataba* « il a écrit », par exemple, est compatible avec deux lectures possibles : dans l'une, le MSu *-a* est un marqueur d'accord et, dans l'autre, ce dernier est un pronom sujet incorporé à la tête verbale et fait partie des arguments de cette tête. Cette double lecture possible se reflète dans les deux entrées lexicales dans les exemples 23 et 24. D'un autre côté, le verbe *katabna* « elles ont écrit », par exemple, n'est compatible qu'avec une seule lecture et n'a, de ce fait, qu'une seule entrée lexicale, celle de l'exemple 25.

(23)
$$\begin{bmatrix} \textit{verbe-mot-MS} & \\ \text{PHON} & \langle \textit{kataba} \rangle \\ \text{TÊTE|ACCRD} & \begin{bmatrix} \text{GRE} & \textit{masc} \\ \text{PERS} & \textit{3} \\ \text{NBRE} & \textit{sing} \end{bmatrix} \\ \text{VAL} & \begin{bmatrix} \text{SUJ} & \langle\ \rangle \\ \text{COMPS} & \langle \boxed{1} \rangle \end{bmatrix} \\ \text{STR-ARG} & \langle [\textit{aff}] \oplus \boxed{1} \rangle \end{bmatrix}$$

(24)
$$\begin{bmatrix} \textit{verbe} - \textit{mot} - \textit{accord} \\ \text{PHON} & \langle \textit{kataba} \rangle \\ \text{TÊTE|ACCRD} & \begin{bmatrix} \text{GRE} & \textit{masc} \\ \text{PERS} & \textit{3} \end{bmatrix} \\ \text{VAL} & \begin{bmatrix} \text{SUJ} & \langle \boxed{1} \rangle \\ \text{COMPS} & \langle \boxed{2} \rangle \end{bmatrix} \\ \text{STR-ARG} & \langle \boxed{1} \oplus \boxed{2} \rangle \end{bmatrix}$$

(25)
$$\begin{bmatrix} \textit{verbe} - \textit{mot} - \textit{MS} \\ \text{PHON} & \langle \textit{katabna} \rangle \\ \text{TÊTE|ACCRD} & \begin{bmatrix} \text{GRE} & \textit{fém} \\ \text{PERS} & \textit{3} \\ \text{NBRE} & \textit{plur} \end{bmatrix} \\ \text{VAL} & \begin{bmatrix} \text{SUJ} & \langle\ \rangle \\ \text{COMPS} & \langle \boxed{2} \rangle \end{bmatrix} \\ \text{STR-ARG} & \langle [\textit{aff}] \oplus \boxed{2} \rangle \end{bmatrix}$$

Par application du principe des traits de tête et du principe de la valence, les effets syntagmatiques de ces entrées lexicales sont les suivants :

- Un verbe dont l'entrée lexicale spécifie le nombre ne peut se combiner avec un sujet lexical (un SN). Si un tel composant est présent, il ne peut faire partie de la liste des arguments du verbe et sera donc disloqué. Le syntagme qui en résulte est de type *syn-t-f* (syntagme-tête-topique).
- Un verbe dont l'entrée lexicale ne spécifie pas le nombre peut se combiner avec un sujet lexical (un SN). Ce dernier fait ainsi partie de la liste des membres de la structure argumentale du verbe et est réalisé comme membre de la liste de valence.

Ainsi, notre grammaire élimine toute possibilité d'unifier la structure de traits qui décrit un SN plein avec celle qui décrit un verbe de type *verbe-mot-MS* dans un syntagme de type *syn-t-s* (syntagme-tête-sujet). Elle permet par contre l'unification d'un SN et d'un verbe-mot-accord dans ce type de syntagme et d'un SN et d'un verbe-mot-MS dans un syntagme de type *syn-t-f*.

Cette analyse repose sur notre conception de l'accord entre le sujet et le verbe en arabe, à savoir la position initiale du verbe et la sous-spécification du nombre sur ce dernier. Nous proposons une modélisation de cette analyse dans la section subséquente.

3.2. L'accord entre le sujet et le verbe

Nous avons avancé une hypothèse selon laquelle l'accord entre le sujet et le verbe ne s'obtient que dans l'ordre VSO en arabe. Le sujet ici est un SN et le verbe porte un marqueur d'accord partiel (non spécifié pour le nombre). Pour rendre compte de cette particularité, nous avons stipulé que les verbes portant de tels MSu possèdent des entrées lexicales différentes des verbes portant des MSu pronominaux (spécifiés pour le nombre). Ainsi, chaque verbe conjugué à la troisième personne du singulier possède deux entrées lexicales et chacune de ces entrées donne lieu à une structure syntagmatique particulière. L'entrée lexicale où le nombre est sous-spécifié donne lieu à une structure où un SN est l'argument sujet. Ce SN s'accorde avec le verbe en genre et en personne seulement. Cette modélisation ne fournit cependant aucun moyen de s'assurer que les phrases agrammaticales, telles que celles des exemples 26c et 26d soient exclues :

(26) a. *daxal* *-a* *l-* *walad* *-u*
entrer.ACCO -3M le- garçon -NOM
« Le garçon est entré »

b. *daxal* *-at* *l-* *bint* *-u*
entrer.ACCO -3F la- fille -NOM
« La fille est entrée »

c. **daxal* *-a* *l-* *bint* *-u*
entrer.ACCO -3M la- fille -NOM

d. **daxal* *-at* *l-* *walad* *-u*
entrer.ACCO -3F le- garçon -NOM

Pour les cas 26c et 26d, le sujet et le verbe portent des informations incompatibles sur le genre et aucune unification n'est possible dans ces contextes. Pour exclure ces phrases, nous avons besoin d'une contrainte qui assure que les traits d'accord, essentiellement le genre, sur le sujet et ceux sur le verbe sont compatibles et peuvent donc être unifiés. Pour y parvenir, nous devons tout d'abord rappeler la théorie de l'accord développée essentiellement par Pollard et Sag (1994), Kathol (1999), Sag et Wasow (1999) et Sag *et al.* (2003).

3.2.1. Le traitement de l'accord en HPSG

HPSG conçoit l'accord dans une perspective non dérivationnelle. En fait, au lieu de le concevoir comme un processus qui copie certains traits du donneur à la cible, la théorie avance l'idée selon laquelle l'accord se réalise quand les deux éléments qui entrent dans ce genre de relation spécifient une information partielle concernant le même objet linguistique : « *Agreement is simply the systematic variation in form that arises from the fact that information coming from two sources about a single object must be compatible.* » (Pollard et Sag, 1994, p. 60).

Dans les premières versions de la théorie, ce partage de structures concerne essentiellement la valeur de l'attribut INDEX (qui est un attribut sémantique réalisé sous l'attribut local CONT)[8]. Dans cette approche, l'accord entre le sujet et le verbe est conçu comme étant le partage des valeurs de l'attribut INDEX de ces deux constituants. Ce partage est illustré dans la structure de traits suivante pour la phrase française *les cigales chantent* :

(27)
$$\begin{bmatrix} \text{PHON} & \langle \textit{les cigales chantent} \rangle \\ \text{SS|LOC|CONT|INDEX} & \boxed{1} \\ \text{FILLES} & \left\langle \begin{bmatrix} \text{FILLE-TÊTE|SS|LOC|CONT|INDEX} & \boxed{1} \begin{bmatrix} \text{NBRE } \textit{plur} \\ \text{PERS } \textit{3} \end{bmatrix} \end{bmatrix}, \begin{bmatrix} \text{FILLE-COMPS|SS|LOC|CONT|INDEX} & \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix}$$

Notons que l'accord entre le sujet et le verbe en français n'implique que deux traits de l'indp. ex., le nombre et la personne.

L'évolution du modèle a amené un changement dans cette théorie de l'accord. Sag et Wasow (1999) et Sag *et al.* (2003) proposent un nouvel attribut AGR (que nous traduisons ici par ACCORD) approprié aux objets de type verbe, nominal et déterminant en anglais. Ces trois catégories sont ainsi regroupées sous le même type lexical *agr-cat* (une catégorie pour laquelle l'accord est approprié). Ce nouveau trait de tête [AGR *agr-cat*] joue le rôle jadis assumé par l'index.

Dans cette théorie, l'accord entre le sujet et le verbe est conçu dans les mêmes lignes que l'accord entre le déterminant et le nom. Le sujet et le déterminant étant regroupés sous le même attribut SPEC (spécifieur), cela permet aux auteurs d'imposer une contrainte lexicale aux lexèmes fléchis[9] de l'anglais, qui appartiennent au type lexical *infl-lxm*, la contrainte de l'accord entre le spécifieur et la tête ou la SHAC (pour *Specifier-Head Agreement Constraint*)[10] :

(28)
$$\begin{bmatrix} \textit{infl - lxm} \\ \text{TÊTE|ACCORD} \ \boxed{1} \\ \text{VAL|SPEC} \quad \left\langle \begin{bmatrix} \text{TÊTE|ACCORD} \ \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix}$$

8. Pollard et Sag (1994) définissent trois sortes d'accord : l'accord des indices, la concordance (l'accord en cas, par exemple) et l'accord pragmatique. Cette distinction est censée rendre compte des cas de l'accord hybride comme dans l'exemple français « vous êtes belle », où l'accord entre le sujet et le verbe est un accord d'indices et celui entre le sujet et le prédicat est pragmatique. Voir Kathol (1999).
9. Cela exclut les lexèmes non fléchis d'être affectés par cette contrainte, tels les noms propres et les pronoms, qui appartiennent à un autre type.
10. Sag *et al.* (2003, p. 238).

Les effets de cette contrainte sont tels que chaque élément du lexique qui sélectionne un spécifieur s'accorde avec ce spécifieur (le sujet ou le déterminant).

3.2.2. La modélisation de l'accord entre le sujet et le verbe

La SHAC est une contrainte lexicale et elle fait donc partie des contraintes propres à une langue donnée (ici l'anglais). Elle devrait de ce fait être modifiée pour rendre compte des faits suivants propres à l'arabe :

1. Il n'y a aucune raison de regrouper le déterminant et le sujet sous la même catégorie, étant donné que les déterminants dans cette langue ne s'accordent pas avec la tête[11]. Nous pouvons ainsi stipuler que l'attribut approprié pour encoder le sujet en arabe est SUJ (et non SPEC).
2. Cet accord implique deux attributs que sont le genre et la personne.
3. La contrainte ne concerne que les verbes en arabe.

Pour rendre compte de ces particularités, nous proposons une contrainte sur l'accord entre le sujet et le verbe, imposée au type lexical verbe-mot-accord et qui prend la forme suivante :

(29)
$$\left[\begin{array}{l} \textit{verbe} - \textit{mot} - \textit{accord} \\ \text{TÊTE|ACCORD} \quad \boxed{1}\left[\begin{array}{ll}\text{GRE} & \textit{gre} \\ \text{PERS} & \textit{pers}\end{array}\right] \\ \text{VAL|SUJ} \qquad\quad \left\langle\left[\text{TÊTE|ACCORD } \boxed{1}\right]\right\rangle \end{array}\right]$$

Cette contrainte impose un partage de structures entre les traits d'accord de la tête verbale et ceux du sujet sélectionné par cette tête. Les traits d'accord en question sont définis comme étant le genre et la personne seulement.

3.2.3. Un exemple d'application

Pour illustrer les effets de cette contrainte sur la sélection opérée par un verbe de son sujet, nous donnons l'exemple suivant :

(30)

daxal	*-at*	*l-*	*bana:t*	*-u*
entrer.ACCO	-3F	les-	filles	-NOM

« Les filles sont entrées »

Deux entrées lexicales sont associées aux verbes conjugués à la 3S, dont l'une seulement est sujette à la contrainte proposée dans l'exemple 29, étant donné que cette contrainte impose des restrictions sur la réalisation du type verbe-mot-accord. L'entrée lexicale du verbe *daxalat*, qui nous concerne ici, est présentée dans la structure de traits partielle suivante :

11. Rappelons que le déterminant défini *Pal-en* arabe est le même quels que soient les traits du nom. Le déterminant indéfini *-n* est également invariable.

(31)
$$\begin{bmatrix} \textit{verbe – mot – accord} \\ \text{PHON} \quad \langle \textit{daxalat} \rangle \\ \text{TÊTE}|\boxed{1}\text{ACCORD} \begin{bmatrix} \text{GRE} \quad \textit{fém} \\ \text{PERS} \quad \textit{3} \end{bmatrix} \\ \text{VAL} \quad \left[\text{SUJ} \left\langle \boxed{2} \left[\text{ACCORD} \quad \boxed{1} \right] \right\rangle \right] \\ \text{STR-ARG} \quad \langle \boxed{2} \rangle \end{bmatrix}$$

Cette entrée lexicale est légitimée par la contrainte de l'exemple 29. Le sujet compatible avec cette entrée lexicale doit contenir une information compatible dans ses traits d'accord, c'est-à-dire [GRE *fém*] et [PERS *3*]. Ce SN peut contenir une information supplémentaire concernant son nombre [NBRE *sing* vs *duel* vs *pluriel*] en préservant la possibilité de l'unification avec l'entrée lexicale du verbe de l'exemple 31. Le SN présent dans notre phrase contient une information compatible puisqu'il est à la 3FP :

(32)
$$\begin{bmatrix} \textit{syntagme} \\ \text{PHON} \quad \langle \textit{lbana : tu} \rangle \\ \text{FILLES} \begin{bmatrix} \text{FILLE-TÊTE}|\text{SS}|\text{LOC}|\text{CAT}|\text{TÊTE} \begin{bmatrix} \text{ACCORD} \begin{bmatrix} \text{GRE} \quad \textit{fém} \\ \text{NBRE} \quad \textit{pluriel} \\ \text{PERS} \quad \textit{3} \end{bmatrix} \\ \text{CAS} \quad \textit{nom} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

La structure de traits suivante représente la phrase qui résulte de cette unification :

(33)
$$\begin{bmatrix} \textit{syntagme} \\ \text{PHON} \quad \langle \textit{daxalatl bana: tu} \rangle \\ \text{SS}|\text{LOC}|\text{CAT} \begin{bmatrix} \text{TÊTE} \ \boxed{1} \\ \text{VAL} \ \langle \ \rangle \end{bmatrix} \\ \text{FILLES} \begin{bmatrix} \text{FILLE-TÊTE} & \begin{bmatrix} \boxed{1}\text{TÊTE}|\text{ACCORD} \ \boxed{2} \\ \text{VAL}|\boxed{3}\text{SUJ} \left\langle \left[\boxed{2}\text{ACCORD} \begin{bmatrix} \text{GRE} \quad \textit{fém} \\ \text{PERS} \quad \textit{3} \end{bmatrix} \right] \right\rangle \end{bmatrix} \\ \boxed{3}\text{FILLE-COMPS} & \left[\text{TÊTE}|\text{ACCORD} \ \boxed{2} \right] \end{bmatrix} \end{bmatrix}$$

Cette structure de traits illustre l'effet de deux principes et d'une contrainte lexicale. Les deux principes sont le principe de traits de tête, qui impose l'unification des valeurs de l'attribut TÊTE de la mère et de la fille, une unification rendue par le chiffre encadré 1, et le principe de la valence, qui annule les éléments de la valence réalisés par les filles (ici le sujet) une fois que ces éléments sont réalisés pour aboutir à un syntagme saturé (la phrase dans notre exemple)

[VAL < >]. L'effet de la contrainte sur le type *verbe-mot-accord* s'observe sur la fille tête, dont les valeurs d'accord sont partagées avec les valeurs d'accord de la seule FILLE-COMPS (le sujet), un partage exprimé par le chiffre encadré 2.

4. IMPLÉMENTATION

La modélisation des MSu que nous avons proposée repose en grande partie sur le lexique. Les entrées lexicales et la hiérarchie des types, en plus des contraintes sur ces types, constituent les éléments essentiels de notre modèle.

Afin de comprendre comment interagissent les différents modules qui composent la grammaire proposée pour légitimer des suites bien formées contenant des MSu, nous avons recours à l'implémentation de cette grammaire dans un système de performance qui permet de la tester. Cette implémentation est faite dans un système informatique conçu pour les grammaires à base de contraintes, à savoir LKB (pour *Linguistic Knowledge Building*). Ce système nous permet non seulement d'analyser des suites bien formées et de rejeter celles qui sont mal formées, mais encore de générer des suites bien formées et rien d'autre que celles-là. La génération est un test fort efficace de la grammaire proposée, puisqu'elle permet de vérifier si les suites produites selon les informations linguistiques fournies au système sont grammaticales ou non et, par conséquent, de vérifier les forces et les faiblesses de la grammaire implémentée.

La grammaire LKB que nous proposons est modulaire et comporte :

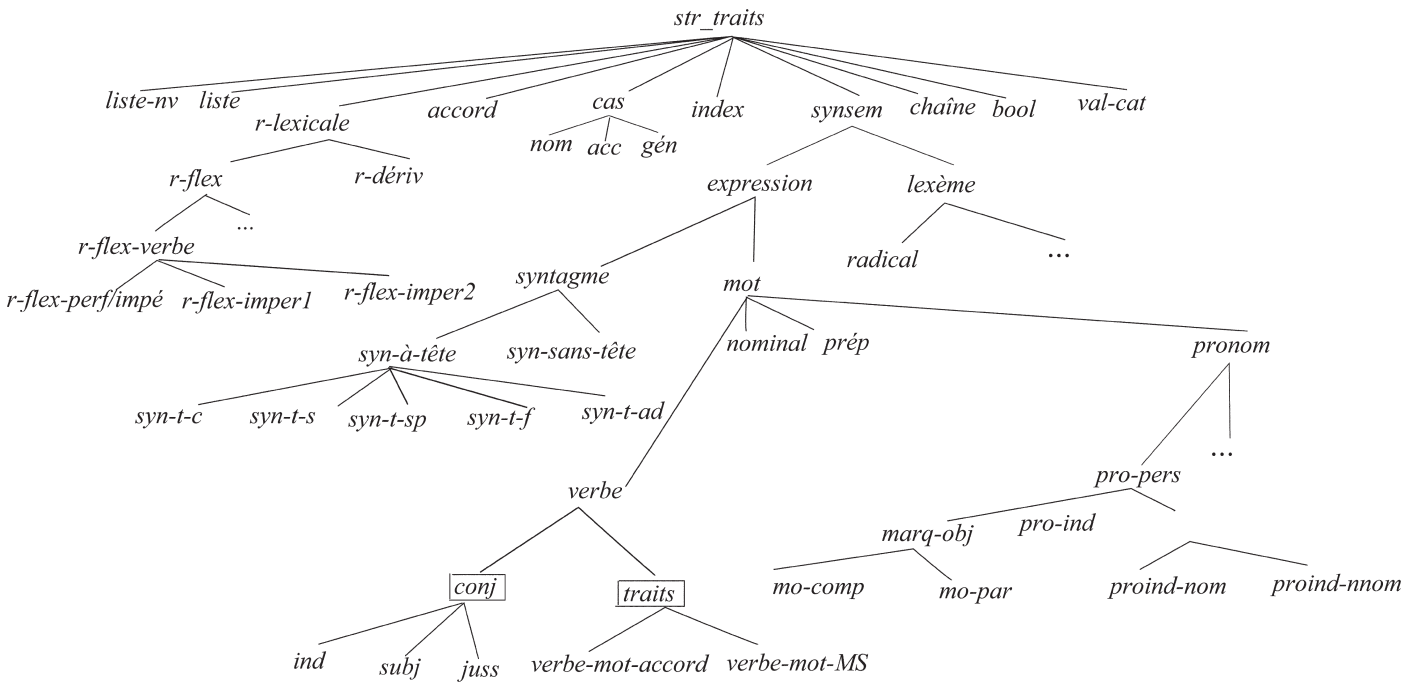
- Une hiérarchie des types ; ce sous-système agit comme un cadre de définition pour le reste de la grammaire puisqu'il détermine les types admis, les conditions d'appropriation des traits et le degré de compatibilité des types.
- Des entrées lexicales ; des radicaux de verbes.
- Des règles flexionnelles ; ces règles créent, à la volée et seulement dans l'analyse (*parsing*) ou à la demande de l'utilisateur, des mots bien formés (des verbes conjugués) à partir des lexèmes fournis dans le lexique.
- Des règles conçues comme des contraintes sur des types syntagmatiques (inclus dans la hiérarchie).

4.1. La hiérarchie des types

La figure 1.3 est une représentation de la hiérarchie des types qui permet d'implémenter les MSu. Cette hiérarchie à héritage multiple comporte également des types syntagmatiques, les types de règles flexionnelles et les types de verbes, entre autres[12].

12. Cette hiérarchie comporte en fait la majorité des types dont nous avons besoin pour implémenter la grammaire de l'arabe standard.

Figure 1.3 – **La hiérarchie des types**



Dans cette hiérarchie, les MSu ne constituent pas de types à part entière et la seule place dans cette structure où nous pouvons constater leur « présence » est sous le type verbe. Ce type possède en fait deux descendants : *verbe-mot-accord* et *verbe-mot-MS*. Les instances de ces descendants contiennent dans leur forme morphologique des MSu après l'application des contraintes appropriées à ces types.

4.2. Les entrées lexicales

Les verbes sont présents dans notre lexique sous forme de radicaux, desquels sont dérivés, par l'emploi des règles flexionnelles appropriées, les verbes-mots employés dans les phrases.

Les exemples suivants représentent les verbes/modèles que nous avons intégrés à notre lexique de base :

```
; écrire

katab:= verbe-transitif-direct &
[ PHON.LIST.FIRST «katab»,
  SEM.RELS.LIST.FIRST.PRD «katab_rel» ].

; dormir

naam:= verbe-intransitif &
[ PHON.LIST.FIRST «naam»,
  SEM.RELS.LIST.FIRST.PRD «naam_rel» ].

; donner

wahab:= verbe-ditran-sn &
[ PHON.LIST.FIRST «wahab»,
  SEM.RELS.LIST.FIRST.PRD «wahab_rel» ].

; craindre/avoir peur

xaaf:= verbe-transitif-indirect &
[ PHON.LIST.FIRST «xaaf»,
  SEM.RELS.LIST.FIRST.PRED «xaafa_rel» ].
```

Pour chaque verbe, dans ces entrées, sont spécifiées les informations concernant son type de valence (intransitif, etc.), sa forme phonologique et son contenu sémantique.

4.3. Les règles flexionnelles

Les verbes contenus dans notre lexique sont introduits sous forme de radicaux. Les règles flexionnelles ajoutent tout affixe nécessaire à la bonne formation des mots verbaux, dont les MSu. L'exemple suivant présente la règle lexicale de la dérivation des verbes accomplis à la 1MS :

```
rlexicale-verbe-1ms-per:=
%suffix (* tu) (naam nimtu) (xaaf xiftu)
verbe-1ms & [TETE.ACCORD accord-masc-1-sing].
```

Cette règle précise le suffixe qui sera concaténé au radical (ici *-tu*), toute forme non régulière (ici pour les verbes *naam* « dormir » et *xaaf* « craindre »), le type lexical auquel appartient la forme obtenue (ici verbe -1MS) et les traits morphosyntaxiques associés à cette forme.

Bien que ce module ait été conçu à l'origine pour encoder la morphophonologie stricte des items lexicaux, nous l'avons étendu pour traiter la morphophonologie en parallèle avec le traitement des traits morphosyntaxiques (de genre, de nombre et de personne), le mode/aspect et les cas morphologiques.

4.4. Les règles de la grammaire

Ces règles sont des structures de traits typées qui décrivent la manière dont les mots et les syntagmes sont combinés pour former d'autres syntagmes. Dans le système LKB, les filles sont encodées dans l'attribut ARGS. La valeur de cet attribut est une liste dont l'ordre des éléments correspond à l'ordre linéaire des filles dans le syntagme.

L'exemple suivant fournit la règle correspondant aux syntagmes dont la tête est un verbe intransitif :

```
regle-tete-complement-0 := tete-initiale-unaire &
[ SUJ #suj,
  THEME #theme,
  ARGS < mot & [ SUJ #suj,
                 THEME #theme,
                 COMPS <> ] > ].
```

Les règles et les entrées lexicales sont traitées de la même façon dans le système : elles sont considérées comme des entrées, c'est-à-dire que la partie qui suit le symbole « := » ne correspond pas à un nouveau type à ajouter à la hiérarchie, mais à un type déjà contenu dans la hiérarchie et duquel l'entrée en question est une instance.

4.5. Les résultats

Les suites contenant des marqueurs de sujet sont analysées et générées correctement que ce soit dans les structures à verbe initial, dans celles avec un SN thématisé, en présence d'un pronom indépendant nominatif ou en l'absence de tout sujet lexical. Le traitement est totalement lexicaliste, encodant les marqueurs de sujet comme faisant partie de l'entrée lexicale du verbe, cette entrée étant elle-même générée par des règles flexionnelles. Cette implémentation est le reflet de notre analyse linguistique de ces unités en tant qu'affixes flexionnels.

Cependant, LKB présente certaines limitations, dont trois affectent notre implémentation des MSu :

1. Le module morphologique est très limité. Les règles flexionnelles ne peuvent être que concaténatives dans un tel système et consistent en l'ajout d'affixes à un radical donné. La morphologie de l'arabe n'étant pas strictement concaténative, ce système ne peut donc l'implémenter convenablement.
2. Ces règles concaténatives ont, de surcroît, l'inconvénient de n'attacher qu'un seul affixe à la fois et ce dernier doit donc être soit un préfixe soit un suffixe. Les MSu de l'arabe, et surtout ceux de l'inaccompli, peuvent se composer d'un préfixe et d'un suffixe, et la seule méthode que nous avons trouvée de les traiter est de choisir entre le préfixe et le suffixe pour qu'il soit attaché à l'entité composée du radical et de l'autre affixe.
3. Les règles flexionnelles ne peuvent s'appliquer à plusieurs lexèmes différents (radicaux) d'un seul mot et ne peuvent donc implémenter les phénomènes de la supplétion. En arabe, le radical de l'accompli est différent de celui de l'inaccompli, mais seul l'un d'entre eux peut figurer comme entrée lexicale, l'autre doit donc être « dérivé » comme forme irrégulière.

5. CONCLUSION

Les marqueurs de sujet de l'arabe standard présentent un sujet d'étude qui fait appel à plusieurs interfaces, dont la syntaxe et la morphologie. En faisant l'effort de délimiter ces deux niveaux d'analyse, nous avons analysé ces marqueurs en définissant leur statut en morphosyntaxe et en morphologie. Nous nous sommes ainsi posé des questions sur leur argumentalité et leur affixalité. Nous avons avancé une hypothèse selon laquelle ils sont toujours des affixes tout en jouant deux rôles morphosyntaxiques différents, celui d'arguments et celui de marqueurs d'accord. Nous avons modélisé cette analyse dans le cadre de la théorie HPSG et nous l'avons, par la suite, implémentée dans le système LKB. Malgré les limitations de ce système, l'implémentation démontre que l'analyse lexicaliste des ces unités est possible et implémentable dans un système de performance.

RÉFÉRENCES

AKKAL, A. (1996). « Word Order and Related Issues in Standard Arabic : A Minimalist Approach », *Recherches linguistiques / Linguistic Research*, vol. 1, n° 1, p. 1-33.

ANDERSON, S.R. (2005). *Aspects of the Theory of Clitics*, Oxford, Oxford University Press.

AOUN, J. et E. BENMAMOUN (1998). « Minimality, Reconstruction, and PF Movement », *Linguistic Inquiry*, vol. 29, n° 4, p. 569-597.

AOUN J., E. BENMAMOUN et D. SPORTICHE (1994). « Agreement, Word Order, and Conjugation in Some Varieties of Arabic », *Linguistic Inquiry*, vol. 25, n° 2, p. 195-220.

AUGER, J. (1994). *Pronominal Clitics in Québec Colloquial French : A Morphological Analysis*, Thèse de doctorat, Philadelphie, University of Pennsylvania.

BENMAMOUN, E. et H. LORIMOR (2006) « Featureless Expressions : When Morphophonological Markers are Absent », *Linguistic Inquiry*, vol. 37, n° 1, p. 1-23.

BRESNAN, J. et S.A. McHOMBO (1987). « Topic, Pronoun and Agreement in Chichewa », *Language*, vol. 63, n° 4, p. 741-782.

COPESTAKE, A. (2002). *Implementing Typed Feature Structure Grammars*, Stanford, CSLI Publications.

FASSI FEHRI, A. (1993). *Issues in the Structure of Arabic Clauses and Words*, Dordrecht, Kluwer Academic Publishers.

FLEISCH, H. (1979). *Traité de philologie arabe*, Beyrouth, Dar El-Machreq Éditeurs.

GINZBURG, J. et I.A. SAG (2000). *Interrogative Investigations : The Form, Meaning and Use of English Interrogative Constructions*, Stanford, CSLI Publications.

HARBERT, W. et M. BAHLOUL (2002). « Postverbal Subjects in Arabic and the Theory of Agreement », dans J. Ouhalla et U. Shlonsky (dir.), *Themes in Arabic and Hebrew Syntax*, Dordrecht, Kluwer Academic Publishers, p. 45-70.

JEBALI, A. (2009). *La modélisation des marqueurs d'arguments de l'arabe standard dans le cadre des grammaires base de contraintes*, Thèse de doctorat, Montréal, Université du Québec Montréal.

KAISER, G.A. (1994). « More About INFL-ection : The Acquisition of Clitic Pronouns in French », dans J.M. Meisel (dir.), *Bilingual First Language Acquisition*, *French and German Grammatical Development*, Amsterdam, Benjamins, p. 131-159.

KATHOL, A. (1999). « Agreement and the Syntax-Morphology Interface in HPSG », dans R. Levine et G. Green (dir.), *Studies in Current Phrase Structure Grammar*, Cambridge, Cambridge University Press, p. 223-274.

LI, C.N. et S.A. THOMPSON (1976). « Subject and Topic : A New Typology of Language », dans C.N. Li (dir.), *Subject and Topic*, Academic Press, New York, p. 457-489.

LUMSDEN, J.S. et G. HALEFORM (2003). « Verb Conjugations and the Strong Pronoun Declension in Standard Arabic », dans J. Lecarme (dir.), *Research in Afro-Asiatic Grammar II*, Amsterdam, John Benjamins Publishing Company, p. 305-337.

MILLER, P.H. (1992). *Clitics and Constituents in Phrase Structure Grammar*, New York, Garland.

MILLER, P.H. et I.A. SAG (1997). « French Clitic Movement without Clitics or Movement », *Natural Language and Linguistic Theory*, vol. 15, nº 3, p. 573-639.

PHILIPPAKI-WARBURTON, I. *et al.* (2002). « On the Status of Clitics and their Doubles in Greek », *Working Papers in Linguistics*, vol. 6, p. 57-84.

POLLARD, C. et I.A. SAG (1987). *Information-Based Syntax and Semantics*, vol. 1, Stanford, Stanford University Press, CSLI.

POLLARD, C. et I.A. SAG (1994). *Head-Driven Phrase Structure Grammar*, Stanford, Stanford University Press, CSLI.

SAG, I.A. et T. WASOW (1999). *Syntactic Theory : A Formal Introduction*, Stanford, CSLI Publications.

SAG, I.A., T. WASOW et E. BENDER (2003). *Syntactic Theory : A Formal Introduction*, Stanford, CSLI Publications.

SHLONSKY, U. (1997). *Clause Structure and Word Order in Hebrew and Arabic : An Essay in Comparative Semitic Syntax*, New York, Oxford University Press.

STUMP, G. (1980). « An "Inflectional" Approach to French clitics », *Ohio State University Working Papers in Linguistics*, vol. 24, p. 1-54.

ZWICKY, A.M. (1977). *On Clitics*, Bloomington, Indiana University Linguistic Club.

ZWICKY, A.M. (1985a). « Cliticization versus Inflection : The Hidatsa Mood Markers », *International Journal of American Linguistics*, vol. 51, nº 4, p. 629-630.

ZWICKY, A.M. (1985b). « Clitics and Particles », *Language*, vol. 61, nº 2, p. 283-305.

ZWICKY, A.M. (1987). « Suppressing the Zs », *Journal of Linguistics*, vol. 23, p. 133-148.

ZWICKY, A.M. et G. PULLUM (1983). « Clicticization vs Inflection : English n't », *Language*, vol. 59, nº 3, p. 502-513.

CHAPITRE 2

INTRODUCTION DE COMBINATEURS DANS DES EXPRESSIONS APPLICATIVES

Adam Joly et Ismaïl Biskri

RÉSUMÉ

Une stratégie d'analyse catégorielle incrémentale de gauche à droite pour éliminer la pseudo-ambiguïté mène bien souvent à la formation de faux constituants. Aussi, des opérations de décomposition de ces faux constituants sont nécessaires. Toutefois peu de travaux ont abordé ce problème. L'une des rares publications dans laquelle une solution, en l'occurrence la décomposition, a été présentée reste celle de Steedman (2000). La décomposition est basée sur le principe de la neutralité paramétrique qu'on peut résumer comme suit : deux types catégoriels associés entre eux par une règle combinatoire (qui ne prend en compte que les types

catégoriels), permettent de déterminer le troisième type. Ce principe rencontre toutefois des limites importantes (voire s'avère inopérant) dans le cas de certaines formes elliptiques de la coordination. Dans notre article, nous proposons une autre méthode de décomposition basée principalement sur la « mémoire » contenue dans les combinateurs de la logique combinatoire.

1. INTRODUCTION

Le modèle des grammaires catégorielles est fondé sur des règles logiques explicites pour un calcul inférentiel logique substitué à une analyse de surface purement linguistique. S'appuyant davantage sur la notion de structure de surface, il aboutit à la notion de forme logique pour exprimer le sens. Ce modèle a l'avantage de représenter l'intrication des unités syntagmatiques en appliquant un opérateur à son opérande, une représentation en soi universelle. Un peu oubliées depuis Husserl (les concepts de catégorèmes et de syncatégorèmes), Lesniewski (les catégories sémantiques), Adjukiewicz, Bar-Hillel et Lambek (calcul de Lambek), on assiste dans les années 1970, 1980 et 1990 à une véritable explosion de travaux et recherches dans le domaine des grammaires catégorielles. On peut citer à cet effet le « collectif » qui pourrait être qualifié de « grammaires catégorielles flexibles », représenté par les modèles de la grammaire universelle de Montague pour une syntaxe catégorielle et une sémantique dénotative, de la grammaire catégorielle combinatoire de Steedman associant analyse syntaxique catégorielle et construction de l'interprétation sémantique fonctionnelle au moyen du lambda-calcul, de la grammaire des opérateurs et opérandes de Harris, de la grammaire catégorielle d'unification, des types intuitionnistes de Martin Löf, de la grammaire catégorielle combinatoire applicative avec l'ajout de métarègles pour piloter les règles de type *raising* et de composition (Biskri et Desclés, 2006), ainsi que d'autres généralisations du calcul de Lambek. Parmi les développements les plus récents, on trouve la version multimodale des grammaires catégorielles combinatoires (Baldridge et Kruijff, 2003) qui introduit des modalités et des restrictions sur l'opérabilité des règles catégorielles pour éliminer des cas d'ambiguïté, ou encore le modèle des grammaires catégorielles abstraites (De Groote et Podogalla, 2004) pour décrire la syntaxe et la sémantique ou encore permettre l'apprentissage de la syntaxe à partir des formes sémantiques (Kanazawa, 2007). Enfin, le modèle *Type-inheritance Combinatory Categorial Grammar* (Beavers, 2004) exploite les performances des grammaires catégorielles tout en ayant une base grammaticale syntagmatique HPSG.

Mis à part les différences dans les approches et les applications, ce qui ressort en particulier de tous les modèles présentés par ce collectif est triple : *i*) leur utilisation de méthodes logiques et mathématiques pour rendre compte de la langue en particulier de la sémantique ; *ii*) leur distinction de plusieurs niveaux logiques de représentations des langues, dont au moins la structure linéaire de l'observable et la structure opérateur-opérande du construit. Cette distinction est

des fois faussement confondue avec la théorie standard exposée par Chomsky qui est loin de reconnaître explicitement des niveaux logiques autres que la structure profonde (Steedman, 2000) ; *iii*) leur flexibilité et adaptabilité à plusieurs langues. À l'instar du français, de l'anglais (Biskri et Desclés, 2006 ; Steedman, 2000), du néerlandais (Steedman, 2000) et de l'allemand avec LEXGRAM, etc., de nouvelles langues sont concernées par le courant des grammaires catégorielles. Les travaux les plus récents concernent des analyses exploratoires pour des langues non indo-européennes avec les constructions relatives en turc (Bozsahin, 2002), les formes de compléments en *–te* en japonais (Kubota, 2007) et les phrases nominales en arabe (Anoun, 2006).

Considérons quelques règles catégorielles pour mieux illustrer ce qui suit :

Application :

$X/Y - Y \rightarrow X$ (>) ; $Y - X\backslash Y \rightarrow X$ (<)

Changement de type :

$X \rightarrow Y/(Y\backslash X)$ (>T) ; $X \rightarrow Y\backslash(Y/X)$ (<T)

Composition fonctionnelle :

$X/Y - Y/Z \rightarrow X/Z$ (>B) ; $Y\backslash Z - X\backslash Y \rightarrow X\backslash Z$ (<B)

La mise en œuvre d'analyseurs catégoriels soulève des défis majeurs. Parmi ces défis on relève le problème de l'explosion combinatoire due principalement à la présence des règles de changements de type ainsi que celui de la pseudo-ambiguïté. Bien que l'utilisation des règles de changement de type soit l'aspect des grammaires catégorielles le plus controversé, celles-ci sont indispensables pour offrir une analyse syntaxique élégante de certains cas de coordination avec ellipses (Steedman, 2000 ; Biskri et Desclés, 2006 ; Moorgat, 1997 ; Dowty, 2000). Nous avons d'ailleurs démontré que des métarègles peuvent contrôler les règles de changement de type et éviter ainsi l'explosion combinatoire (Biskri et Desclés, 2006). Dans le cas de la pseudo-ambiguïté (c'est-à-dire plusieurs arbres syntaxiques qui ne correspondent qu'à une seule interprétation sémantique pour un même énoncé), la littérature scientifique propose l'analyse incrémentale de gauche à droite qui permet de ne conserver qu'un seul arbre de dérivation. Toutefois, cette solution se voit confrontée à la validation des faux constituants (voir l'exemple suivant).

	Charles	mange	son-repas	lentement	
(1)	NP	(S\NP)/NP	NP	(S\NP)\(S\NP)	(>T)
(2)	S/(S\NP)				(>B)
(3)	S/NP				(>)
(4)	S				

L'expression de catégorie S (phrase) *Charles mange son repas* est un faux constituant. Elle est validée comme phrase bien construite alors que *lentement* n'a pas été considéré.

Pour résoudre ce problème, Steedman (2000) propose la décomposition du faux constituant selon le principe de la neutralité paramétrique en prenant comme indices les seuls types catégoriels. L'énoncé du principe de la neutralité paramétrique est : deux types catégoriels associés entre eux par une règle catégorielle (qui ne prend en compte que les types catégoriels) permettent de déterminer le troisième type.

En d'autres termes, pour une règle catégorielle combinatoire telle que A - B → C, si nous connaissons les types catégoriels A et C, nous pouvons déduire B ; ou si nous connaissons les types catégoriels B et C, nous pouvons déduire A ; ou enfin si nous connaissons les types catégoriels A et B, nous pouvons déduire C.

À l'exemple précédent, nous devrions donc effectuer une décomposition à l'étape 5. Nous avons le type catégoriel S de l'expression *Charles mange son repas* et nous avons besoin d'obtenir le type S\N pour qu'il puisse être combiné au type (S\N)\(S\N) de *lentement*. En vertu de la neutralité paramétrique, nous posons x - S\N → S et nous pouvons alors déduire le troisième terme (x) qui est S/(S\N).

	Charles	mange	son-repas	lentement	
(1)	NP	(S\NP)/NP	NP	(S\NP)\(S\NP)	
(2)	S/(S\NP)				(>**T**)
(3)	S/NP				(>**B**)
(4)	S				(>)
(5)	S/(S\NP)	S\NP			(**Décomposition**)
(6)		S\NP			(<)
(7)	S				(>)

La décomposition rencontre toutefois des limites importantes : *i*) l'identification de la catégorie nécessaire n'est associée à aucune stratégie clairement établie ; *ii*) dans le cas de certaines formes elliptiques de la coordination, la méthode s'avère inopérante. Prenons le cas de la phrase *Charles-mange-son-repas-lentement-et-son-dessert-rapidement*, l'analyse va valider l'expression *Charles-mange-son-repas-lentement* de type catégoriel S avant de rencontrer la conjonction *et*. Le second membre de la coordination est du type catégoriel (S\NP)\((S\NP)/NP).

	[…]	son-dessert	rapidement	
(8)		NP	(S\NP)\(S\NP)	
(9)		(S\NP)\((S\NP)/NP)		(>**T**)
(10)		(S\NP)\((S\NP)/NP)		(<**B**)

À cette étape nous disposons de deux types catégoriels. S le type obtenu et (S\NP)\((S\NP)/NP) le type nécessaire. La décomposition de l'expression *Charles-mange-son-repas-lentement*, pour extraire le premier membre de la coordination selon le principe énoncé plus haut, donne le type catégoriel S/((S\NP)\((S\NP)/NP)).

S/((S\NP)\((S\NP)/NP)) - (S\NP)\((S\NP)/NP) → S

Or ce type ne correspond à aucune interprétation sémantique.

2. LA GRAMMAIRE CATÉGORIELLE COMBINATOIRE APPLICATIVE (GCCA)

Dans ce modèle, les unités de la langue sont considérées comme des opérateurs ou des opérandes et seront traduites dans des expressions logiques formelles de la logique combinatoire (Curry et Feys, 1958 ; Shaumyan, 1998). Les modèles linguistiques de la grammaire applicative universelle (Shaumyan, 1998) et de son extension la grammaire applicative et cognitive (Desclés, 1996) insistent de façon très explicite sur la pertinence des niveaux de représentation des langues avec interaction entre ces différents niveaux. Leur principal objectif étant une étude des propriétés formelles des systèmes langagiers et une analyse de leurs structures logiques : *i*) le niveau des structures morphosyntaxiques, où sont exprimées les caractéristiques particulières de la langue (par exemple l'ordre des mots, des cas morphologiques, etc.). C'est la représentation observable de la langue ; *ii*) le niveau des structures prédicatives, où sont exprimés les invariants grammaticaux ainsi que les structures prédicatives sous-jacentes aux énoncés phénotypiques. Ce niveau permet d'exprimer l'interprétation sémantique fonctionnelle des énoncés dans laquelle chaque unité linguistique est un opérateur suivi de ses opérandes ; *iii*) le niveau cognitif, où la signification des prédicats linguistiques est donnée. C'est dans ce cadre général que le modèle de la grammaire catégorielle combinatoire applicative s'inscrit (Biskri et Desclés, 2006).

En plus de la vérification de la bonne connexion syntaxique des énoncés, ce modèle permet de connecter de façon explicite l'expression morphosyntaxique à sa représentation prédicative au moyen de combinateurs de la logique combinatoire auxquels sont associées des règles d'introduction et de suppression (β-réduction). À des fins d'illustration, nous présentons les règles de β-réduction pour Φ, **B** et $\mathbf{C}_*$[1] (u_1, u_2, u_3 et u_4 étant des expressions applicatives typées qui agissent soit comme des opérateurs ou des opérants) :

$((\Phi\ u_1\ u_2\ u_3)\ u_4) \rightarrow u_1\ (u_2\ u_4)\ (u_3\ u_4)$
$((\mathbf{B}\ u_1\ u_2)\ u_3) \rightarrow (u_1\ (u_2\ u_3))$
$((\mathbf{C}_*\ u_1)\ u_2) \rightarrow (u_2\ u_1)$

1. Ici nous nous intéressons seulement aux combinateurs utilisés dans ce chapitre.

Dans le modèle de la grammaire catégorielle combinatoire applicative (voir quelques règles ci-dessous), la prémisse de chaque règle est la concaténation des unités linguistiques avec des types orientés. La conséquence de chaque règle est une expression typée applicative avec l'introduction possible d'un combinateur. Le changement de type d'une unité u introduit le combinateur $\mathbf{C}_*$; la composition de deux unités concaténées introduit le combinateur **B**.

Règles d'application :

$$\frac{[X/Y : u_1] \; - \; [Y : u_2]}{[X : (u_1 \; u_2)]} \; ; \; \frac{[Y : u_1] \; - \; [X\backslash Y : u_2]}{[X : (u_2 \; u_1)]}$$

Règles de changement de type :

$$\frac{[X : u]}{[Y/(Y\backslash X) : (\mathbf{C}_* \; u)]}\mathbf{T} \; ; \; \frac{[X : u]}{[Y\backslash(Y/X) : (\mathbf{C}_* \; u)]}\mathbf{T}$$

Règles de composition fonctionnelle :

$$\frac{[X/Y : u_1] \; - \; [Y/Z : u_2]}{[X/Z : (\mathbf{B} \; u_1 \; u_2)]}\mathbf{B} \; ; \; \frac{[Y\backslash Z : u_1] \; - \; [X\backslash Y : u_2]}{[X\backslash Z : (\mathbf{B} \; u_2 \; u_1)]}\mathbf{B}$$

La logique combinatoire est sans variable. Elle permet d'éliminer les cas de télescopage. En outre, elle permet de systématiser une *réorganisation structurelle intelligente des constituants,* moins gourmande en ressources computationnelles du fait que les combinateurs dans les expressions combinatoires véhiculent un contenu sémantique intrinsèque qui joue le rôle d'une mémoire tout le long de l'analyse. Notons que la *réorganisation structurelle* dont nous parlons est fondamentalement différente de la *décomposition* (Steedman, 2000). La *réorganisation structurelle* utilise comme indices non seulement les types catégoriels mais aussi l'interprétation sémantique fonctionnelle.

Nous présentons maintenant la restructuration. Soit l'analyse suivante (les catégories sont notées [X : u], X étant le type catégoriel et u l'interprétation sémantique) :

(1)	[N : *Charles*] - [(S\N)/N : *mange*] - [N : *son-repas*] - [(S\N)\(S\N) : *lentement*] - [(X\X)/X : *et*] - [N : *son-dessert*] - [(S\N)\(S\N) : *rapidement*]	
(2)	[S/(S\N) : (**C**$_*$ *Charles*)] - [(S\N)/N : *mange*] - [N : *son-repas*] - [(S\N)\(S\N) : *lentement*] - [(X\X)/X : et] - [N : *son-dessert*] - [(S\N)\(S\N) : *rapidement*]	>**T**
(3)	[S/N : (**B** (**C**$_*$ *Charles*) *mange*)] - [N : *son-repas*] - [(S\N)\(S\N) : *lentement*] - [(X\X)/X : *et*] - [N : *son-dessert*] - [(S\N)\(S\N) : *rapidement*]	>**B**
(4)	[S : ((**B** (**C**$_*$ *Charles*) *mange*) *son-repas*)] - [(S\N)\(S\N) : *lentement*] - [(X\X)/X : *et*] - [N : *son-dessert*] - [(S\N)\(S\N) : *rapidement*]	>
(5)	[S : ((**C**$_*$ *Charles*) (*mange son-repas*))] - [(S\N)\(S\N) : *lentement*] - [(X\X)/X : *et*] - [N : *son-dessert*] - [(S\N)\(S\N) : *rapidement*]	
(6)	[S/(S\N) : (**C**$_*$ *Charles*)] - [S\N : (*mange son-repas*)] - [(S\N)\(S\N) : *lentement*] - [(X\X)/X : *et*] - [N : *son-dessert*] - [(S\N)\(S\N) : *rapidement*]	>dec
(7)	[S/(S\N) : (**C**$_*$ *Charles*)] - [S\N : (*lentement* (*mange son-repas*))] - [(X\X)/X : *et*] - [N : *son-dessert*] - [(S\N)\(S\N) : *rapidement*]	<
(8)	[S : ((**C**$_*$ *Charles*) (*lentement* (*mange son-repas*)))] - [(X\X)/X : *et*] - [N : *son-dessert*] - [(S\N)\(S\N) : *rapidement*]	>
(9)	[S : ((**C**$_*$ *Charles*) (*lentement* (*mange son-repas*)))] - [(X\X)/X : *et*] - [(S\N)\((S\N)/N) : (**C**$_*$ *son-dessert*)] - [(S\N)\(S\N) : *rapidement*]	<**T**
(10)	[S : ((**C**$_*$ *Charles*) (*lentement* (*mange son-repas*)))] - [(X\X)/X : *et*] - [(S\N)\((S\N)/N) : (**B** *rapidement* (**C**$_*$ *son-dessert*))]	<**B**
(11)	[S/(S\N) : (**C**$_*$ *Charles*)] - [S\N : (*lentement* (*mange son-repas*))] - [(X\X)/X : *et*] - [(S\N)\((S\N)/N) : (**B** *rapidement* (**C**$_*$ *son-dessert*))]	>dec
(12)	**[S/(S\N) : (C$_*$ *Charles*)] - [S\N : ((B *lentement* (C$_*$ *son-repas*)) *mange*)] - [(X\X)/X : *et*] - [(S\N)\((S\N)/N) : (B *rapidement* (C$_*$ *son-dessert*))]**	
(13)	[S/(S\N) : (**C**$_*$ *Charles*)] - [(S\N)/N : *mange*] - [(S\N)/((S\N)/N) : (**B** *lentement* (**C**$_*$ *son-repas*))] - [(X\X)/X : *et*] - [(S\N)\((S\N)/N) : (**B** *rapidement* (**C**$_*$ *son-dessert*))]	<dec
(14)	[S/(S\N) : (**C**$_*$ *Charles*)] - [(S\N)/N : *mange*] - [(S\N)/((S\N)/N) : (**B** *lentement* (**C**$_*$ *son-repas*))] - [((S\N)\((S\N)/N))\ ((S\N)\((S\N)/N)) : (*et* (**B** *rapidement* (**C**$_*$ *son-dessert*)))]	>
(15)	[S/(S\N) : (**C**$_*$ *Charles*)] - [(S\N)/N : *mange*] - [(S\N)\((S\N)/N) : ((*et* (**B** *lentement* (**C**$_*$ *son-repas*))) (**B** *rapidement* (**C**$_*$ *son-dessert*)))]	<
(16)	[S/(S\N) : (**C**$_*$ *Charles*)] - [S\N : (((*et* (**B** *lentement* (**C**$_*$ *son-repas*))) (**B** *rapidement* (**C**$_*$ *son-dessert*))) *mange*)]	<
(17)	[S : ((**C**$_*$ *Charles*) (((*et* (**B** *lentement* (**C**$_*$ *son-repas*))) (**B** *rapidement* (**C**$_*$ *son-dessert*))) *mange*))]	>
(18)	((**C**$_*$ *Charles*) (((*et* (**B** *lentement* (**C**$_*$ *son-repas*))) (**B** *rapidement* (**C**$_*$ *son-dessert*))) *mange*))	
(19)	(((((*et* (**B** *lentement* (**C**$_*$ *son-repas*))) (**B** *rapidement* (**C**$_*$ *son-dessert*))) *mange*) *Charles*)	(**C**$_*$)
…		
(25)	((∧ (*lentement* (*mange son-repas*)) (*rapidement* (*mange son-dessert*))) *Charles*)	(**C**$_*$)

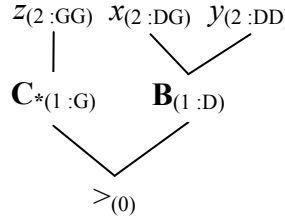
Les étapes 1 à 17 représentent l'application des règles de la grammaire catégorielle combinatoire applicative. Avec ces étapes, nous vérifions la correction syntaxique de la phrase. Les étapes 18 à 25 font partie du niveau prédicatif. Elles permettent de réduire les combinateurs au moyen des règles de β-réduction de façon à construire l'interprétation sémantique fonctionnelle (la forme normale) : ((∧ (*lentement* (*mange son-repas*)) (*rapidement* (*mange son-dessert*))) *Charles*), qui est structurée comme une clause conjonctive. À l'étape 20, le prédicat linguistique *et* est remplacé par sa signification au niveau cognitif Φ ∧ de façon à exprimer sa nature distributive et conjonctive par, respectivement, le combinateur Φ et le connecteur logique ∧.

À l'étape 10 nous disposons du faux constituant ((**C**$_*$ *Charles*) (*lentement* (*mange son-repas*)) de type S et du second membre de la coordination (**B** *rapidement* (**C**$_*$ *son-dessert*)) de type (S\N)\((S\N)/N). Deux étapes sont requises pour extraire le premier membre de la coordination du faux constituant. La première étape repose sur le principe suivant : l'interprétation sémantique fonctionnelle est construite en utilisant les combinateurs introduits dans la structure syntagmatique. Elle est toujours notée dans un format binaire : un opérateur suivi de son opérande. Lorsqu'un faux constituant est obtenu ((**C**$_*$ *Charles*) (*lentement* (*mange son-repas*))), il est nécessaire de tester si le type de l'opérande (*lentement* (*mange son-repas*)) peut être combiné au type de l'unité linguistique résiduelle (**B** *rapidement* (**C**$_*$ *son-dessert*)). Si tel est le cas, on décompose (voir étape 11) sinon on réduit un combinateur. Le procédé est répété jusqu'à aboutissement ou qu'il n'y ait plus de combinateur à réduire. L'expression applicative obtenue à l'étape 11 (*lentement* (*mange son-repas*)) contient le premier membre de la coordination. La seconde étape consiste à déterminer une structure combinatoire équivalente à (*lentement (mange son-repas*)) en l'occurrence ((**B** *lentement* (**C**$_*$ *son-repas*)) *mange*). Il s'agit donc de l'introduction de combinateurs dans une expression applicative sans combinateurs. Une décomposition appliquée à cette dernière expression permet d'obtenir le premier membre de la coordination. La prochaine section décrit cette seconde étape.

3. INTRODUCTION DES COMBINATEURS DANS UNE EXPRESSION APPLICATIVE

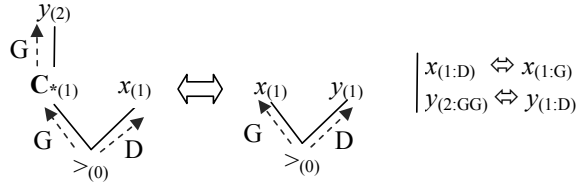
Nous proposerons dans cette section une méthode automatique de restructuration d'un membre d'une coordination de manière à ce que la structure de l'expression combinatoire de ce dernier soit équivalente à celle du second membre. Concrètement, cette solution consistera à étudier le principe que la logique combinatoire est sans variables et que ce sont les combinateurs qui jouent le rôle de « mémoire ». Cette mémoire est celle de l'action du combinateur, intrinsèquement véhiculée lors de l'opération de β-réduction. En effet, il suffit de réduire un combinateur de son expression pour retrouver l'état de la structure précédant son introduction.

Par ailleurs, nous présenterons dorénavant les expressions combinatoires au moyen d'une structure en arbre, afin d'en faciliter la lecture. Nous indiquerons en indice et entre parenthèses le niveau de profondeur des nœuds dans l'arbre, le niveau 0 correspondant à la racine, qui sera par ailleurs toujours une expression applicative (>). Le niveau pourra être également suivi par une suite de directions (« G » pour le premier argument situé à gauche ou « D » pour le second argument situé à droite) à emprunter à partir de la racine pour se rendre au nœud. Par exemple, $\mathbf{C}_{*(2:GD)}$ signifie que le combinateur $\mathbf{C}_*$ se trouve à deux niveaux de profondeur supplémentaires par rapport à la racine, c'est-à-dire à sa gauche, puis à sa droite. Par exemple, considérons l'expression combinatoire suivante : ((**C**$_*$ z) (**B** x y)). L'arbre correspondant serait dans ce cas celui-ci :



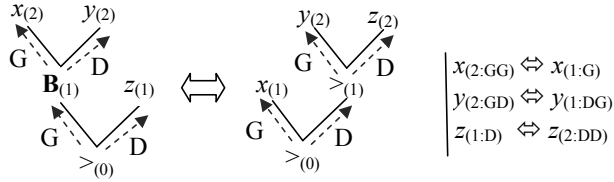
Nous allons maintenant observer ce qui se produit au plan des expressions applicatives lorsque nous effectuons une β-réduction pour les combinateurs **B** et $\mathbf{C}_*$. Nous nous limiterons dans le cadre de nos recherches à l'exploitation de ces deux combinateurs, bien qu'il en existe davantage, mais ceux-ci s'avéreront suffisants pour démontrer la stratégie et rien ne nous empêchera d'utiliser la méthode proposée pour l'appliquer aux autres combinateurs.

Nous débuterons par le combinateur de changement de type $\mathbf{C}_*$. L'expression formelle « littéraire » qui décrit la β-réduction de ce combinateur est : ((**C**$_*$ y) x) → (x y). Sous forme d'arborescence, nous obtiendrons par conséquent ceci :



On peut noter dans le schéma précédent une variation non seulement en termes de niveaux, mais aussi de directions. Plus spécifiquement, après la β-réduction, l'unité x passe de droite à gauche. Puis, afin d'atteindre l'unité y, nous devons aller à gauche à deux reprises, mais une fois que le combinateur $\mathbf{C}_*$ est réduit, le nœud perd un niveau et est directement à la droite de l'application racine.

Dans le cas du combinateur **B** la β-réduction se fait comme suit : ((**B** *x y*) *z*) → (*x* (*y z*)). En utilisant un arbre, cela donne :



La β-réduction du combinateur **B** a un impact sur tous les arguments. Initialement, *x* est à gauche de **B**, qui est lui-même à gauche de la racine. Après la réduction, un niveau est déduit et le chemin qui était « gauche, gauche » devient seulement « gauche ». Dans le cas de *y*, son niveau ne varie pas, mais ses deux directions changent : « gauche, droite » devient « droite, gauche ». Finalement, *z* augmente d'un niveau et nous devons ajouter une direction « droite » supplémentaire.

Ces transformations que subissent le niveau et le chemin des arguments sont liées à la fonction d'un combinateur et jouent un rôle capital : elles constituent la clé pour déterminer la position à laquelle un combinateur doit être introduit. Le second aspect à considérer est que, dans tous les arbres de β-réduction de combinateurs, l'opération d'application avant à la racine de l'arbre ne varie jamais et représente un point de référence que nous appellerons un « délimiteur fonctionnel », puisqu'il délimite la fonction exercée par le combinateur sur ses arguments. Conséquemment, l'introduction d'un combinateur passera par la détermination du nœud dans l'arbre que représente ce délimiteur fonctionnel. Pour ce faire, nous devrons retrouver le chemin original menant au combinateur avant sa β-réduction (ou de façon synonyme, après son introduction). Alors, puisque les combinateurs $\mathbf{C}_*$ et **B** se trouvent à gauche du délimiteur fonctionnel, on retranchera la dernière direction, et on obtiendra l'endroit précis dans l'arbre représentant le second membre de la coordination où on doit effectuer l'introduction. Notons que ce ne sont pas tous les combinateurs qui se retrouvent directement à la gauche de leur délimiteur fonctionnel. À titre d'exemple, le combinateur de distributivité ϕ qui est « deux fois à gauche » de son délimiteur.

Aussi, puisque l'élimination des combinateurs dans une expression combinatoire se fait de droite à gauche (d'un point de vue littéraire), le parcours de l'expression combinatoire pour trouver les chemins d'introduction se fera en sens opposé, soit de la gauche vers la droite. Dans une représentation en arbre binaire comme nous le considérons, cela revient donc à parcourir l'arbre de la racine vers la feuille, de la gauche vers la droite. À chaque fois que nous rencontrerons un combinateur (**B** ou $\mathbf{C}_*$), nous devrons rechercher les nœuds qui représentent chacun de ces arguments et modifier les chemins de ces nœuds conséquemment à ce qui a été présenté plus haut, afin d'obtenir la position de ces arguments avant l'introduction du combinateur. Évidemment, ces opérations ne visent pas à changer le chemin du combinateur « en cours » (le combinateur n'existant pas

avant son introduction, il ne peut avoir un chemin qu'après son introduction), mais seulement ceux de ses arguments, dont peuvent ou non faire partie d'autres combinateurs qui ont été introduits dans l'expression avant ce dernier. Ainsi, au moment où nous atteindrons un combinateur, son chemin sera nécessairement celui après son introduction, ce qui sera suffisant pour déterminer le délimiteur fonctionnel et l'introduire au bon endroit.

Ces observations peuvent être traduites par l'algorithme suivant qui permet de calculer les chemins vers les combinateurs d'une expression combinatoire donnée après leur introduction. L'expression combinatoire est organisée en une structure d'arbre binaire où chaque nœud possède une mémoire contenant le chemin initial jusqu'à ce dernier à partir de la racine (sous la forme d'une liste, par exemple), et la seule entrée de la méthode est un nœud (en commençant par la racine).

```
méthode calculerChemin
     si le noeud courant est un combinateur B alors
          remplacer le chemin « GG » par « G » pour les noeuds du
          premier argument
          remplacer le chemin « GD » par « DG » pour les noeuds
          du deuxième argument
          trouver les noeuds du troisième argument et changer
          le chemin « D » par « DD »
     sinon si le noeud courant est un combinateur C* alors
          remplacer le chemin « GG » par « D » pour les noeuds du
          premier argument
          trouver les noeuds du deuxième argument et changer
          le chemin « D » par « G »
     fin si
     pour chaque noeud enfant, de gauche vers la droite
          appeler calculerChemin pour ce noeud
     fin pour
fin de la méthode
```

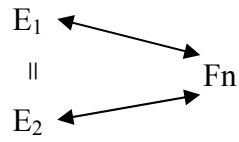
Le procédé s'avère être plutôt simple : il s'agit d'un algorithme récursif qui appelle à tour de rôle chacun des nœuds de l'arbre, de la racine jusqu'aux feuilles, du premier nœud enfant au dernier. Pour chaque nœud, on vérifie s'il s'agit d'un combinateur. Si ce n'est pas le cas, on passe au nœud suivant. Dans un cas positif, on recherche les arguments du combinateur en question et on modifie les chemins selon le combinateur en présence. La recherche des nœuds représentant chaque argument demande à parcourir l'arbre à la recherche de nœuds avec des chemins qui contiennent le chemin jusqu'au délimiteur fonctionnel du combinateur courant, plus le chemin jusqu'à l'argument. Par exemple, si le chemin vers $\mathbf{C}_*$ est (GDG) et le délimiteur fonctionnel se situe à gauche, puis à droite (GD), puisque le deuxième argument doit être à droite du délimiteur fonctionnel, alors les nœuds du second argument sont ceux qui ont un chemin qui débute par (GDD). Ainsi, des chemins valides pourraient être (GDD), (GDDG), (GDDD), et ainsi de suite. Cependant, on n'aura pas à utiliser cette méthode pour chacun des arguments d'un combinateur, étant donné que le premier et le second argument de **B** et le premier argument de $\mathbf{C}_*$ sont leurs nœuds enfants : la propagation des directions à être modifiées est facilement effectuée en agissant directement

sur ces nœuds, sans rechercher un chemin précis. Par contre, le troisième argument de **B** et le second argument de $\mathbf{C}_*$ ne sont pas toujours nécessairement à la même position tels que présentés dans leurs schémas respectifs, à cause des différentes combinaisons possibles entre combinateurs dans l'expression combinatoire (entre autres lorsque plusieurs compositions fonctionnelles se succèdent) et il nous faudra donc utiliser la méthode proposée afin de retracer leurs positions.

Une fois que tous les nœuds de l'arbre ont été passés en revue et, par conséquent, que les chemins vers chacun des combinateurs après leur introduction ont été calculés, nous pouvons procéder à l'opération d'introduction des combinateurs dans l'expression applicative sans combinateurs (en forme normale) à restructurer. Rappelons-nous que ce que nous souhaitons dans le cadre d'une analyse de la grammaire catégorielle combinatoire applicative est de réorganiser la structure du faux constituant contenant le premier membre de la coordination afin qu'elle soit similaire à celle du second membre. Les combinateurs seront introduits dans l'ordre inverse de leur réduction, soit de la droite vers la gauche du point de vue de l'expression combinatoire, en tenant compte des chemins et en ignorant la dernière direction « gauche » pour les raisons expliquées auparavant. L'aboutissement du processus nous mènera à l'une des deux situations suivantes : *i*) soit tous les combinateurs auront pu être introduits aux positions calculées et le premier membre aura alors été correctement restructuré. Dans un tel cas, l'analyse pourra se poursuivre ; *ii*) soit nous aurons une situation où le combinateur ne peut être introduit à l'endroit calculé, car la position n'existe pas ou ne permet pas l'exécution d'une opération de β-réduction pour ce combinateur. Dans ce cas, l'analyse ne pourra pas se compléter et cela signifiera que la phrase n'est probablement pas syntaxiquement bien construite.

Ces conclusions puisent leurs sources dans le théorème de Church-Rosser (figure 2.1). Considérons deux expressions combinatoires différentes, E_1 étant le premier membre de la coordination et E_2 le second. Le théorème stipule que si E_1 possède une forme normale, celle-ci est unique, et par conséquent si la réduction de E_1 mène à une forme normale Fn et que la réduction de E_2 mène également à Fn alors par confluence E_1 et E_2 sont deux expressions combinatoires équivalentes.

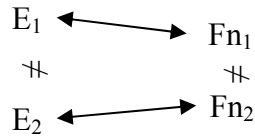
Figure 2.1 – **Théorème de Church-Rosser**



En d'autres mots, notre méthode calcule les étapes d'introduction de combinateurs qui permettent de construire E_2 (le second membre de la coordination) à partir de Fn. Si par la réduction de E_1 nous obtenons Fn alors nécessairement l'introduction des combinateurs dans cette forme normale construira un membre ayant une structure équivalente au second. Nous aurons alors la situation que nous avons décrite en *i*).

À l'opposé, si E_1 et E_2 ne possèdent pas la même forme normale alors, comme le montre la figure 2.2, en aucune façon nous n'arriverons à reconstruire E_1 afin que sa structure soit équivalente à celle de E_2.

Figure 2.2 – **Forme normale différente pour E_1 et E_2**

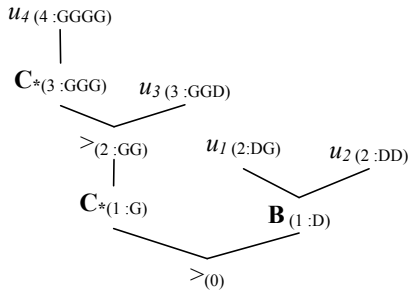


Cela correspond à la situation *ii*). Il se produira dans le cas où la structure syntaxique de la phrase à analyser est invalide. La conséquence en sera que l'application des règles de la grammaire catégorielle combinatoire applicative dans l'expression applicative typée mènera à la construction d'une expression combinatoire erronée pour le premier ou le second membre de la coordination (ou les deux).

4. APPLICATION DE L'ALGORITHME

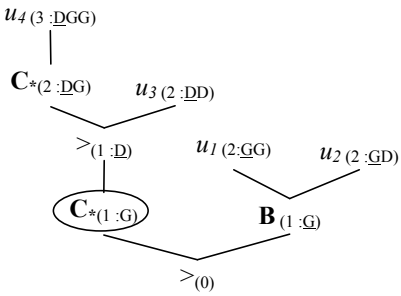
Voyons maintenant l'application de l'algorithme à travers des exemples. Dans le premier cas, nous montrerons étape par étape son exécution pour une expression fictive et nous introduirons les combinateurs dans sa forme normale afin de vérifier si l'algorithme a calculé les bons chemins. Dans le second cas, nous reprendrons l'exemple utilisé précédemment et nous verrons le processus complet de réorganisation structurelle.

Pour le premier exemple, prenons l'expression combinatoire (($\mathbf{C}_*$ (($\mathbf{C}_*$ u_4) u_3) (**B** u_1 u_2)), qui peut être traduite par l'arbre binaire sur la page suivante.

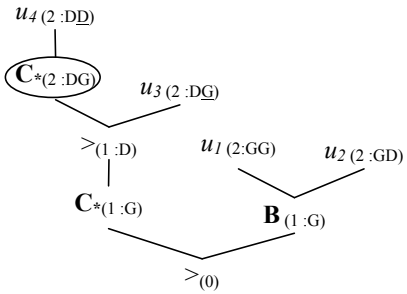


Nous devons exécuter l'algorithme pour chacun des nœuds de l'arbre, de la racine vers les feuilles, de la gauche vers la droite, ce qui donne les étapes intermédiaires suivantes :

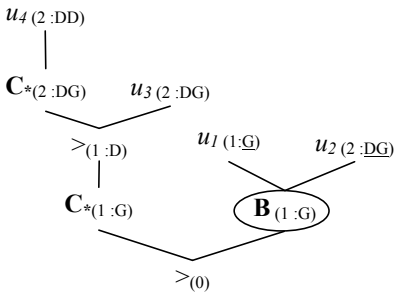
i)



ii)



iii)

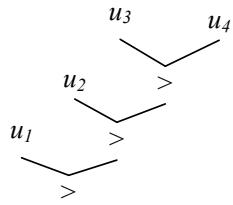


À l'étape *i*), nous atteignons le premier combinateur $\mathbf{C}_*$. Le premier argument comporte tous les nœuds inférieurs au combinateur, c'est-à-dire $((\mathbf{C}_* u_4) u_3)$. Ces nœuds perdent une direction, « GG » devenant « G ». Le délimiteur fonctionnel de $\mathbf{C}_*$ étant la racine, on recherche pour le deuxième argument tous les nœuds dont le chemin débute par « D », ce qui correspond aux nœuds à droite de la racine, dont la première direction des chemins passent de « D » à « G ».

L'étape *ii*) nous conduit au second combinateur $\mathbf{C}_*$, pour lequel le premier argument est u_4, qui verra son chemin perdre une fois de plus une direction. Le chemin actuel calculé de C_* étant « DG », son délimiteur fonctionnel est à droite de la racine, et le deuxième argument s'avère être u_3, « DG » changeant pour « DD ».

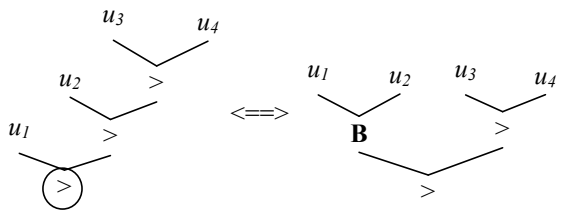
Enfin, à l'étape *iii*), nous croisons le dernier combinateur de l'expression. Les deux premiers arguments de **B** sont les nœuds enfants, respectivement u_1 et u_2. u_1 perd une direction gauche, pour passer de « GG » à « G », tandis que les directions de u_2 sont interchangée, ce qui fait en sorte que « GD » devient « DG ». Il ne reste maintenant plus qu'à trouver le troisième argument de **B**, qui constitue un cas particulier que nous désirons mettre en relief à travers cet exemple. Son délimiteur fonctionnel est la racine, et nous devrions rechercher tout nœud dont le chemin débute par « D » et ajouter une direction droite. $((\mathbf{C}_* u_4) u_3)$ satisferait à cette condition. Cependant, un combinateur ne peut modifier le chemin de nœuds déjà visités. Si nous modifiions tout de même le chemin des nœuds, nous modifierions du même coup le chemin déjà calculé du combinateur $\mathbf{C}_*$, alors que ce dernier n'existerait pas encore au moment d'introduire le combinateur **B**. Pour cette raison, nous ne modifierons pas les directions de la sous-expression $((\mathbf{C}_* u_4) u_3)$.

Nous avons maintenant terminé de calculer l'ordre d'introduction des combinateurs, qui est $\mathbf{B}_{(1:G)}$, $\mathbf{C}_{*(2:DG)}$, $\mathbf{C}_{*(1:G)}$. Ils seront injectés un à un, à partir d'une expression applicative sans combinateurs équivalente, décrite par $(u_1 (u_2 (u_3 u_4)))$ et représentée par l'arbre ci-contre :

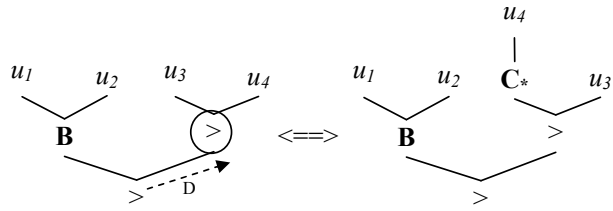


Voici les différentes étapes d'introduction des trois combinateurs :

i)



ii)

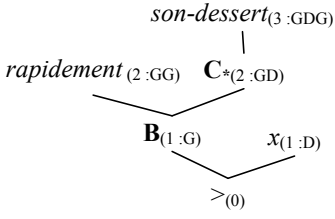


iii)

À l'étape *i*), nous insérons le combinateur **B**. En se rappelant qu'il nous faut ne pas tenir compte de la dernière direction afin de déterminer le délimiteur fonctionnel qui servira de point de référence pour l'introduction du combinateur, nous trouvons que ce délimiteur est la racine. En *ii*), en ignorant la direction gauche, nous devons aller à droite de la racine pour atteindre le délimiteur fonctionnel, d'où l'introduction de $\mathbf{C}_*$ se fera. À l'étape *iii*), nous introduisons le dernier combinateur, $\mathbf{C}_*$, dont le délimiteur fonctionnel s'avère aussi être la racine.

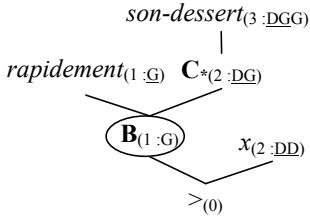
L'expression combinatoire que nous obtenons possède une structure identique à celle de l'exemple, ce qui démontre que l'algorithme a bien fonctionné.

Considérons maintenant le cas de la réorganisation structurelle nécessaire à l'analyse de *Charles-mange-son-repas-lentement-et-son-dessert-rapidement*. L'expression à réorganiser est (*lentement* (*mange son-repas*)). L'« étalon » qui nous permet de déterminer les combinateurs à introduire est le second membre de la coordination (**B** *rapidement* (**C**$_*$ *son-dessert*)). Nous devons, en effet, réorganiser (*lentement* (*mange son-repas*)) de façon à extraire une structure similaire à (**B** *rapidement* (**C**$_*$ *son-dessert*)), afin de pouvoir appliquer la conjonction *et*. Voici donc l'arbre binaire de (**B** *rapidement* (**C**$_*$ *son-dessert*)) :

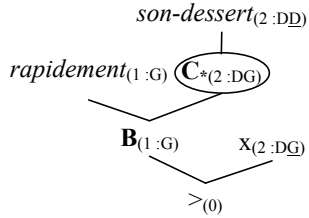


Nous pouvons remarquer que la structure a subi une modification. Une application avant a été ajoutée, avec un nouvel argument *x*. Cela est dû au fait que (**B** *rapidement* (**C**$_*$ *son-dessert*)) n'est pas une expression sémantiquement complète et il manque le dernier argument du combinateur **B** qui, dans le cas d'une coordination de ce type, devrait être *mange*, mais le processus ne connaît pas cette information.

i)



ii)

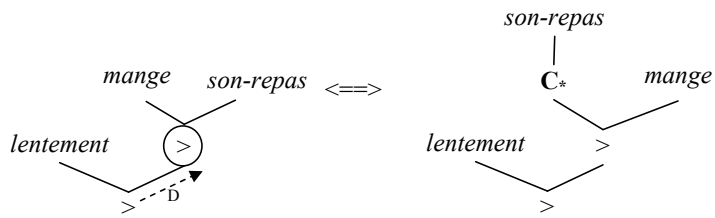


Étape *i*) : le premier nœud que nous atteignons est le combinateur $\mathbf{B}_{(1:G)}$. Ses deux premiers arguments sont *rapidement* et (**C**$_*$ *son-dessert*), qui changeront respectivement de « GG » à « G », puis de « GD » à « DG ». Le troisième argument est le nœud *x* que nous avons ajouté à la structure, dont le chemin « D » devient « DD ».

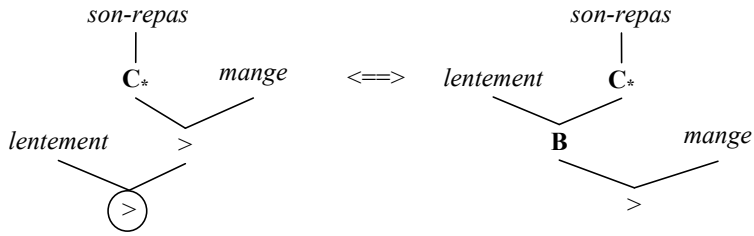
Étape *ii*) : nous arrivons au combinateur $\mathbf{C}_{*(2:DG)}$. Son premier argument est évidemment *son-dessert*. Par conséquent, les directions « GG » dans le chemin se transforment en « D ». Puis, pour le dernier argument, nous recherchons les nœuds dont le chemin débute par « DD », et le dernier « D » deviendra « G ». *x* sera le nœud affecté.

Tous les combinateurs ayant été explorés, nous introduisons alors les combinateurs dans l'expression (*lentement* (*mange son-repas*)) selon l'ordre $\mathbf{C}_{*(2:DG)}$, $\mathbf{B}_{(1:G)}$.

i)



ii)



À l'étape *i*), le délimiteur fonctionnel du combinateur $\mathbf{C}_*$ sera à droite de la racine, puisqu'on doit ignorer la dernière direction. Puis, à l'étape *ii*) celui de $\mathbf{B}_{(1:G)}$ se trouve au niveau 0, car nous devons encore une fois ignorer le dernier « G ». À l'étape *iii*), les deux combinateurs sont introduits et le résultat est un premier membre qui a la même structure que le second. On peut donc poursuivre et terminer l'analyse. L'expression obtenue en *iii*) notée d'une façon linéaire est ((***B*** *lentement* (**C**$_*$ *son-repas*)) *mange*). L'opérateur dans cette expression est le premier membre de la coordination.

5. CONCLUSION

Plusieurs travaux sur les grammaires catégorielles ont été publiés. Toutefois la plupart de ces travaux restent cantonnés à des considérations uniquement théoriques. La mise en œuvre effective des grammaires catégorielles a souvent été négligée. Bien souvent celle-ci rencontre des problèmes majeurs non encore résolus comme le problème que nous avons évoqué dans notre article.

Par ailleurs, la plupart des modèles des grammaires catégorielles utilisent le lambda calcul pour inférer l'interprétation sémantique fonctionnelle. La logique combinatoire malgré sa puissance reste peu utilisée. Nous avons, dans notre article, montré son précieux apport dans un processus de retour arrière qu'on pourrait qualifier d'intelligent.

RÉFÉRENCES

ANOUN, H. (2006). « Towards a Logical Analysis of Nominal Sentences in Standard Arabic », *Actes de ESSLLI 2006*.

BALDRIDGE, J. et G.J. KRUIJFF (2003). « Multi-Modal Combinatory Categorial Grammar », *Actes de EACL 2003*.

BEAVERS, J. (2004). « Type-inheritance Combinatory Categorial Grammar », Genève, *Proceedings of COLING 2004*.

BISKRI, I. et J.P. DESCLÉS (2006). « Coordination de catégories différentes en français », *Faits de Langue*, nº 28.

BOZSAHIN, C. (2002). « The Combinatory Morphemic Lexicon », *Computational Linguistics,* vol. 28, nº 2, p. 145-186.

CURRY, B.H. et R. FEYS (1958). *Combinatory Logic Vol. I,* Amsterdam, North-Holland.

DE GROOTE, P. et S. PODOGALLA (2004). « On the Expressive Power of Abstract Categorial Grammars : Representing Context-free Formalisms », *Journal of Logic and Information*.

DESCLÉS, J.P. (1996). « Cognitive and Applicative Grammar : an Overview », dans C. Martin Vid (dir.), *Lenguajes Naturales y Lenguajes Formales XII*, p. 29-60.

DOWTY, D. (2000). « The Dual Analysis of Adjuncts/Complements in Categorial Grammar », *Linguistics,* n° 17.

KANAZAWA, M., (2007). « Second-order Abstract Categorial Grammars as Hyperedge Replacement Grammars », *Actes de ESSLLI 2007*.

KUBOTA, Y. (2007). « Solving the Morpho-Syntactic Puzzle of the Japanese *-te* Form Complex Predicate : A Multi-Modal Combinatory Categorial Grammar Analysis », *Questions empiriques et formalisation en syntaxe et sémantique,* n° 7, p. 283-306.

MOORGAT, M. (1997). « Categorial Type Logics », *Handbook of Logic and Language*, Amsterdam, North-Holland, p. 93-177.

SHAUMYAN, S.K. (1998). « Two Paradigms of Linguistics : The Semiotic Versus Non-Semiotic Paradigm », *Web Journal of Formal, Computational and Cognitive Linguistics*.

STEEDMAN, M. (2000). *The Syntactic Process*, Boston, MIT Press/Bradford Books.

CHAPITRE 3

UNE MÉTHODE DE *CHUNKING* MULTILINGUE ENDOGÈNE

Jacques Vergne

RÉSUMÉ

Le chunking *consiste à segmenter un texte en* chunks*, segments sous-phrastiques que Abney a défini approximativement comme des groupes accentuels. Traditionnellement, le* chunking *utilise des ressources monolingues, le plus souvent exhaustives, quelquefois partielles : des mots grammaticaux et des ponctuations, qui marquent souvent des débuts et fins de* chunk*. Mais cette méthode, si l'on veut l'étendre à de nombreuses langues, nécessite de multiplier les ressources monolingues. Nous présentons une nouvelle méthode : le* chunking *endogène, qui n'utilise aucune ressource hormis le texte analysé lui-même. Cette méthode prolonge les travaux de Zipf : la minimisation de l'effort de communication conduit les locuteurs à raccourcir les mots fréquents. On peut alors caractériser un* chunk

comme étant la période des fonctions périodiques corrélées longueur et effectif des mots sur l'axe syntagmatique. Cette méthode originale présente l'avantage de s'appliquer à un grand nombre de langues d'écriture alphabétique, avec le même algorithme, sans aucune ressource externe.

1. INTRODUCTION

Le *chunking* est une étape de segmentation courante dans différents types d'application : analyseurs robustes, analyseurs de complexité linéaire (Vergne, 1998, 2000, 2001, 2002), détermination des groupes accentuels et leur mise en relation dans les systèmes de synthèse vocale, pour calculer la macro-prosodie (Vannier *et al.*, 1999), en indexation automatique, le *chunk* comme autre grain indexé en plus du mot et, en alignement sous-phrastique, le *chunk* comme grain aligné.

2. PRINCIPES DU *CHUNKING* ENDOGÈNE

La méthode que nous proposons est fondée sur les propriétés des fonctions longueur et effectif des mots sur l'axe syntagmatique. Ces deux fonctions sont corrélées : périodiques, synchrones, en opposition de phase, et leur période permet de caractériser le *chunk*. Sur une période, la fonction longueur est monotone croissante, et la fonction effectif est monotone décroissante. Ces concepts prolongent les travaux de Zipf : la minimisation de l'effort de communication conduit le locuteur à raccourcir les mots fréquents (Zipf, 1949). La métrique de longueur utilisée par Zipf est le nombre de caractères de la graphie ; la métrique de notre méthode est le nombre de syllabes, ou plus exactement le nombre de noyaux vocaliques, calculable à partir de la graphie ; cette métrique est enracinée dans l'origine orale du concept de *chunk*. L'effectif des mots est mesuré dans le texte à segmenter.

L'algorithme exploite une hypothèse et deux propriétés du *chunk*. Voici l'hypothèse : dans un même texte, à 1 graphie correspond 1 mot ; donc tous les contextes d'une même graphie informent sur le même mot dans un même texte. La première propriété du *chunk* est la suivante : le « virgulot » est défini comme segment délimité par deux ponctuations ; il a été proposé par Hervé Déjean (« entre-ponctuations ») et renommé « virgulot » par Nadine Lucas. Propriété : le *chunk* est un constituant du virgulot ; donc une graphie attestée en début de virgulot est un début de *chunk*, et une graphie attestée en fin de virgulot est une fin de *chunk*. La deuxième propriété exploite la propriété du *chunk* (à l'intérieur d'un virgulot) de pouvoir être caractérisé comme la période des fonctions périodiques longueur et effectif des mots, et leur caractéristique de croissance ou décroissance monotones.

Du fait de l'hypothèse de biunivocité graphie-mot, début et fin de *chunk* sont des attributs des mots du dictionnaire du texte et non pas des attributs des mots occurrences du texte. Pour chaque propriété du *chunk*, l'algorithme consiste en un parcours du texte pour relever des débuts et fins de *chunks*, et renseigner les mots du dictionnaire, puis sortir les résultats en informant les mots occurrences à partir des mots du dictionnaire. L'ordre d'application des deux propriétés est indifférent car elles sont indépendantes.

Cette méthode de segmentation en *chunks* est fondée sur des propriétés numériques, et est valide sur les langues à écritures alphabétiques. Elle est endogène, car elle effectue des calculs sur le texte à segmenter et n'utilise aucune ressource externe au texte analysé.

3. MODÈLE DE STRUCTURE DU *CHUNK* SELON ABNEY ET SELON DÉJEAN

Le concept de *chunk* a été proposé par Steve Abney (1991). Il a été fondé sur des propriétés de la parole : Abney définit le *chunk* comme groupe accentuel. La parole étant contrainte par l'appareil vocal, voyons le *chunk* comme un concept générique sur les langues, un concept du langage. Hervé Déjean (1998) en a proposé un modèle de structure : des débuts et des fins de *chunk* (mots ou morphèmes) entourant un noyau (Déjean, 1998, p. 117); notre méthode utilise ce modèle.

Par exemple, la graphie « Commission » a été relevée dans les *chunks* suivants dans un même texte :

```
                [ Commission européenne ]
          [ la  Commission ]
          [ la  Commission européenne ]
[ dans  la  Commission ]
```

En voici la synthèse :

```
[ dans [ la    [ Commission ]  européenne ]
[ débuts       [    noyau   ]     fins    ]
```

4. HYPOTHÈSE DE LA BIUNIVOCITÉ GRAPHIES ←→ MOTS

La méthode proposée fait l'hypothèse que dans un même texte, on a biunivocité entre une graphie et un mot. Donc tous les contextes d'une même graphie informent le même mot dans un même texte. On sait que cette hypothèse est fausse en toute généralité, mais elle reste valide pour un texte monothématique de quelques milliers de mots. Cette hypothèse constitue une manière de gérer l'articulation entre déductions locales et leur consolidation au niveau global :

ici la consolidation est totale, et toutes les occurrences de la même graphie sont considérées comme étant le même mot, avec des contextes factorisables. Ce choix devra être remis en question pour traiter des textes plus longs.

5. DEUX PROPRIÉTÉS DES DÉBUTS ET FINS DE *CHUNK*

5.1. Propriété 1 : le *chunk* est un constituant du virgulot

Hervé Déjean (1998) a défini l'entre-ponctuations comme étant un constituant délimité par deux ponctuations. Nadine Lucas (2001) a proposé le terme « virgulot », que nous utiliserons désormais. On définit donc la hiérarchie de constituants : le texte est composé de virgulots, eux-mêmes constitués de *chunks*, eux-mêmes constitués de graphies.

Propriété exploitée par l'algorithme :

–une graphie attestée en début de virgulot marque un début de *chunk*,
–une graphie attestée en fin de virgulot marque une fin de *chunk*.

Exemples de virgulots :

*, **in** denen Aale **leben** ,*
*, **bis** die Bewirtschaftungspläne **vorliegen** .*

*. **It** also intends to explore **measures** ,*
*, **before** migrating upstream to spend most of their **lives** .*

*, **en** las aguas centro-occidentales del Océano **Atlántico** .*
*, **donde** transcurre la mayor parte de su **vida** .*

*. **Lasciandosi** trasportare dalla corrente e **nuotando** ,*
*, **dove** si riproducono una sola volta e poi **muoiono** .*

Les premières graphies des virgulots sont des débuts de *chunks* (prépositions, pronoms, …), et leurs dernières graphies sont des fins de *chunks* (noms, verbes, adjectifs, …).

5.2. Propriété 2 : le *chunk* est la période des fonctions corrélées longueur et effectif des mots sur l'axe syntagmatique

Définissons deux fonctions des mots sur l'axe syntagmatique : leur longueur, exprimée en nombre de syllabes, et leur effectif dans le texte à segmenter.

Ces deux fonctions sont entières, périodiques, synchrones, en opposition de phase, donc corrélées, et définissons le *chunk* comme étant leur période. Sur une période, le *chunk*, la fonction longueur est monotone croissante, et la fonction effectif est monotone décroissante.

Prenons un exemple de virgulot :

	, would	migrate	From	the	rivers	on	their	territories
longueur	1	2	1	1	2	1	1	4
effectif	10	3	6	65	2	6	4	1

Sur la fonction longueur, on a les suites monotones croissantes suivantes : [1 2] [1 1 2] [1 1 4]

Sur la fonction effectif, on a les suites monotones décroissantes : [10 3] [6] [65 2] [6 4 1]

Pour les deux fonctions, à une suite correspond une période ; autrement dit, les suites expriment une segmentation ; les deux fonctions sont synchrones, c'est-à-dire que les suites des deux fonctions déterminent (presque) les mêmes périodes ; sur une période (synchrone), les deux fonctions sont en opposition de phase : la fonction longueur est monotone croissante, et la fonction effectif est monotone décroissante ; l'ensemble des propriétés communes de ces deux fonctions nous permet de les dire corrélées ; c'est ce qu'on exprime habituellement en disant que les mots courts sont fréquents, et que les mots longs sont rares.

Remarquons à la suite de Zipf (1949) dans son ouvrage *The Principle of Least-Effort* que l'écriture, comme la parole, constitue une compression optimale ; et notons aussi que l'on retrouve les principes de la compression de fichiers en informatique : ce qui est fréquent est codé court, et ce qui est rare est codé long.

Faisons une observation sur la loi de Zipf, telle qu'elle est connue actuellement : cette loi établit une relation entre les effectifs et les rangs des mots classés par ordre d'effectifs décroissants ; si l'on se limite à cette loi, on oublie les longueurs des mots ; or Zipf a proposé de considérer longueurs et effectifs ensemble, de manière corrélée, en termes d'optimisation (le moindre effort), et c'est à cette origine des concepts de Zipf que nous revenons.

Revenons au mode de calcul de la longueur. Le calcul est effectué à partir de la graphie : la longueur est définie comme le nombre de syllabes, soit comme le nombre de noyaux vocaliques (une suite de voyelles contiguës correspond à un noyau vocalique, et donc à une longueur 1). Ce calcul nécessite en donnée les voyelles de l'alphabet (latin ou grec). Il y a deux cas particuliers : le *y* voyelle ou consonne, et les finales muettes en français et anglais.

Le *y* est voyelle dans « *systèmes* » et consonne dans « *rayon* » ; *y* est consonne par défaut ; *y* est voyelle en début ou fin de mot, ou seul (*usually, by, y*) ; *y* est voyelle entre deux consonnes ; ces règles suffisent à traiter tous les cas pour les 20 langues du corpus.

Puisque la longueur d'un mot est considérée comme le nombre de syllabes prononcées, en français et en anglais, les *e* muets à la fin des mots se terminant par « *e* » ou « *es* », mais pas « *ies* » ou « *ees* », ne sont pas comptés comme syllabes (le cas de « *ent* » muet n'est pas traité, car insoluble au niveau du mot).

Les sigles (suite de majuscules) ont une longueur égale à deux fois leur nombre de lettres (tendance à être en fin de *chunk*). Un nombre est de longueur 1, quel que soit son nombre de chiffres (tendance à être en début de *chunk*).

6. ALGORITHME FONDÉ SUR CES PROPRIÉTÉS

Le dictionnaire du texte est calculé, ainsi que les effectifs de chaque mot. La longueur de chaque mot du dictionnaire est calculée. Du fait de la biunivocité graphie-mot, début et fin de *chunk* sont des attributs des mots du dictionnaire du texte et non pas des attributs des mots occurrences du texte.

Pour la propriété 1, fondée sur le virgulot, le texte est parcouru, et les mots en début ou fin de virgulot sont notés comme début ou fin de *chunk*.

Pour la propriété 2, fondée sur les suites monotones, le texte est parcouru, en notant les frontières entre deux suites monotones, d'où, pour chaque frontière, une fin et un début de *chunk*. Une fonction booléenne « dans la même suite » retourne si deux mots contigus sont dans la même suite monotone. Quatre solutions sont expérimentées : sur la longueur seule, sur l'effectif seul, sur longueur ET effectif, ou sur longueur OU effectif. Les résultats sont très comparables, du fait que les deux fonctions sont fortement corrélées. L'utilisation de la longueur seule permet, en n'utilisant pas l'effectif, de rendre la méthode utilisable sur des textes très courts, tels qu'une requête de moteur de recherche. Par exemple, cette fonction, sur la longueur seule, sur les mots i et i+1, pour traduire le fait que la longueur est monotone croissante, est de la forme :

les deux mots i et i+1 ne sont pas des séparateurs de virgulot ET longueur(i+1) ≥ longueur(i)

Voici la trace du processus sur notre virgulot exemple :

pptés :	virgulot		Suite		résultat				
	d	f	d	f	d	f	long.	effectif	
252	[1,	0]	[2,	0]	[2,	0]	1	10	*would*
253	[0,	0]	[0,	2]	[0,	1]	2	3	*migrate*
254	[0,	0]	[5,	0]	[1,	0]	1	6	*from*
255	[5,	0]	[21,	0]	[2,	0]	1	65	*the*
256	[0,	0]	[0,	1]	[0,	1]	2	2	*rivers*
257	[0,	0]	[6,	0]	[1,	0]	1	6	*on*
258	[0,	0]	[1,	0]	[1,	0]	1	4	*their*
259	[0,	1]	[0,	1]	[0,	2]	4	1	*territories*

Commentaire pour *the :* ce mot a été 5 fois en début de virgulot, 21 fois en début de suite, les deux propriétés en font un début de *chunk* (pour toutes ses occurrences).

Commentaire pour *their* : ce mot a été une fois en début de suite, ce qui en fait un début de *chunk* pour ses 4 occurrences.

D'où la segmentation suivante ([marque un début de *chunk*,] marque une fin de *chunk*) :

, [would migrate] [from [the rivers] [on [their territories] ,

7. QUELQUES VIRGULOTS SEGMENTÉS EN *CHUNKS*

Le corpus de validation de la méthode est composé de communiqués de presse (environ 1000 mots chacun) du site de l'Union européenne (<europa.eu/rapid/>), chaque communiqué étant rédigé en 1 à 22 langues. Les virgulots suivants sont extraits du communiqué IP/05/1018 de 2005 :

[Die Laichgründe] [der Aale] befinden] [sich [im Sargassosee] [im mittleren Westatlantik].

[Eels spawn] [in [the Sargasso Sea [in [the western central Atlantic] Ocean].

[Las *[anguilas] desovan] [en [el Mar [de [los Sargazos] , [en [las aguas] centro-occidentales] [del Océano Atlántico].*

[La zone [de frai] [de l'anguille] [se situe [en mer] [des Sargasses] , [dans [la partie centre-ouest] [de l'océan Atlantique].

[Le anguille] [si riproducono] [nel mar [dei Sargassi] , [nell'Atlantico centro-occidentale].

8. CONCLUSION

En caractérisant le *chunk* de manière purement numérique, à partir des propriétés des fonctions longueur et effectif des mots sur l'axe syntagmatique, cette méthode originale consiste en des calculs sur le texte à segmenter ; elle a l'avantage de s'appliquer à un grand nombre de langues, avec le même algorithme, sans aucune ressource monolingue : des langues d'écriture alphabétique, avec une pratique du mot écrit qui sépare les mots grammaticaux des mots lexicaux,

et compatibles avec un modèle de structure de *chunk* où les mots grammaticaux sont généralement antéposés aux mots lexicaux ; la méthode est prometteuse pour les 22 langues de la Communauté européenne, hormis le finnois[1].

Étant indépendante des spécificités de chaque langue, cette méthode n'est pas « multilangue », ni « multi-monolangue », mais comme elle exploite des propriétés génériques des langues, c'est-à-dire des propriétés du langage, en tant qu'abstraction des langues, on pourrait peut-être simplement l'appeler méthode « linguistique ».

RÉFÉRENCES

ABNEY, S. (1991). « Parsing By Chunks », dans S. Abney, *Principle-Based Parsing*, Dordrecht, Kluwer Academic Publishers, p. 257-278.

DÉJEAN, H. (1998). *Concepts et algorithmes pour la découverte des structures formelles des langues*, Thèse de doctorat, Caen, Université de Caen.

LUCAS, N. (2001). « Étude et modélisation de l'explication dans les textes », *Actes du Colloque L'explication : enjeux cognitifs et communicationnels*, Paris.

VANNIER, G., A. LACHERET-DUJOUR et J. VERGNE (1999). « Pauses Location and Duration Calculated with Syntactic Dependencies and Textual Considerations for T.T.S. System », *ICPhS 1999,* San Francisco, août.

VERGNE, J. (2000). *Coling 2000*, Tutorial : « Trends in Robust Parsing ».

VERGNE, J. (2001). « Analyse syntaxique automatique de langues : du combinatoire au calculatoire » (communication invitée), dans *Actes de la conférence sur le traitement automatique des langues naturelles (TALN 2001)*, Tour, France, p. 15-29.

VERGNE, J. (2002). « Une méthode pour l'analyse descendante et calculatoire de corpus multilingues : application au calcul des relations sujet-verbe », dans *Actes de la conférence sur le traitement automatique des langues naturelles (TALN 2002)*, France, p. 63-74.

VERGNE, J. et E. GIGUET (1998). « Regards théoriques sur le "*tagging*" », dans *Actes de la conférence sur le traitement automatique des langues naturelles (TALN 1998),* Paris, 10-12 juin.

ZIPF, G.K. (1949). *Human Behavior and the Principle of Least-Effort*, Londres, Addison-Wesley.

1. Voir des résultats sur : <www.info.unicaen.fr/~jvergne/chunking_multilingue_endogene/>.

CHAPITRE 4

DESCRIPTION ET ANNOTATION DES ERREURS
Le cas des francophones s'exprimant en anglais

Camille Albert, Marie Garnier, Arnaud Rykner et Patrick Saint-Dizier

RÉSUMÉ

Dans cet article, nous présentons quelques défis soulevés par l'analyse et la correction des erreurs commises par des rédacteurs francophones lorsqu'ils s'expriment en anglais. Nous présentons en particulier notre méthode de constitution d'un corpus d'erreurs et notre approche pour catégoriser ces erreurs. Un ensemble d'annotations est ensuite proposé pour marquer les erreurs, donner leurs caractéristiques et intégrer une ou plusieurs corrections.

1. INTRODUCTION

Les rédacteurs francophones qui doivent écrire en anglais sont souvent confrontés à des problèmes de forme (lexique et syntaxe) et de style, qui pénalisent parfois lourdement l'intelligibilité de leurs textes. Notre projet a pour objectifs 1) d'élaborer des procédures de correction sur des classes d'erreurs qui ne sont pas traitées à ce jour par les éditeurs de textes, même les plus avancés, 2) de modéliser et de réaliser un assistant intelligent qui accompagnerait des apprenants dans leur démarche de perfectionnement de leur compétence en anglais écrit, où interviennent de nombreuses considérations (style, niveau de langue, usages du domaine de spécialité, contexte de la phrase, public, etc.) (Han *et al.*, 2005 ; Lee *et al.*, 2006).

Dans le but d'élaborer un tel assistant, nous analysons des productions spontanées, grand public, peu contrôlées (courriels, forums) ainsi que des productions soignées, davantage professionnelles (publications, rapports). La première tâche est de repérer les erreurs puis de les catégoriser dans une perspective opérationnelle. Cette catégorisation se fait sur la base de traits communs à plusieurs erreurs. Nous considérons des erreurs d'origine grammaticale (par exemple, structure du GV, place de l'adverbe, etc.), lexicale, de style (par exemple, l'usage maladroit du passif) et de calque (à partir des structures françaises). Nous nous attachons ensuite à annoter les erreurs dans les textes en faisant apparaitre plusieurs caractéristiques, dont : la gravité de l'erreur, l'intelligibilité du segment concerné, la ou les corrections (avec des segments de texte corrigés souvent différents selon la correction), la certitude de l'annotateur-correcteur.

L'un des buts majeurs du projet est d'explorer, puis de modéliser les stratégies cognitives utilisées en phase de détection et de correction d'erreurs habituellement déployées par des enseignants de langues ou des traducteurs et de les inclure dans l'assistant intelligent. En effet, la correction de nombre d'erreurs se base sur un processus d'analyse et de décision complexe que nous allons explorer. D'un point de vue opérationnel, nous proposons une analyse de la détection de l'erreur et de sa correction en faisant ressortir les arguments (et les objets abstraits qu'ils manipulent) pour et contre la correction et sa nature, ainsi que le modèle de décision associé. En complément de la correction, l'assistant devra s'adapter dynamiquement au niveau de l'apprenant et à ses besoins, en lui permettant, par exemple, de poser des questions sur la grammaire (règles, exceptions, structures courantes, etc.).

Ce travail comprend en particulier les composantes fondamentales et applicatives suivantes :

1. les aspects linguistiques : le fonctionnement de la grammaire en général, et en particulier en anglais, les liens entre le lexique et la grammaire, les différents paramètres du style selon le type de document (courriels, forums, rapports et publications), la prise en compte de l'utilisateur-apprenant ;

2. les aspects cognitifs : les stratégies déployées par les experts pour déceler et analyser les erreurs et proposer une ou plusieurs corrections, les formes d'argumentation employées ainsi que les processus de décision menant à la correction effective ;
3. les aspects didactiques : les types de dialogues entre l'assistant et l'utilisateur visant à faciliter l'acquisition de compétences en langue, dans le cadre d'une théorisation de l'explication ;
4. les aspects de modélisation : la mise en œuvre de modèles adaptés en particulier en traitement automatique des langues, ainsi qu'en argumentation et théorie de la décision (pour la correction) et en question-réponses (pour l'assistant intelligent et coopératif) ;
5. la réalisation d'un prototype pouvant se greffer sur un éditeur de textes : ce prototype fera l'objet d'une évaluation détaillée par des étudiants et des enseignants en enseignement des langues.

Dans ce document, nous présentons tout d'abord la méthode de constitution du corpus, c'est-à-dire les différents paramètres étudiés et la façon dont ceux-ci sont réalisés dans le corpus. Ensuite, nous abordons, à travers l'analyse de travaux déjà réalisés ou en cours, une méthode de catégorisation des erreurs réalisée à la fois sur une base linguistique et opérationnelle. Enfin, nous proposons un ensemble d'annotations pour marquer ces erreurs et proposer des corrections. Ces annotations, qui feront l'objet d'une analyse quant à leur intelligibilité, adéquation au problème à traiter et utilisation par des annotateurs, sont utilisées dans un cadre inductif pour induire des règles de correction (Albert *et al.*, 2009).

2. CONSTITUTION ET PARAMÈTRES DU CORPUS

La constitution du corpus est une étape majeure pour l'analyse des erreurs. En effet, la qualité du corpus, au plan des paramètres pris en compte, va déterminer en bonne partie la nature des erreurs relevées ainsi que la catégorisation qui sera effectuée. Une des difficultés principales est de déterminer les types de documents qui entrent dans la constitution du corpus : celui-ci doit en effet contenir les erreurs les plus fréquemment rencontrées dans les productions écrites par des francophones s'exprimant en anglais. La diversité des documents, des auteurs et des domaines, la taille des documents et leurs sources (professionnelles *vs* privées) sont autant de critères qui doivent être pris en considération pour notre objectif.

2.1. Méthodologie générale de la constitution du corpus

Nous présentons ci-dessous les principaux paramètres qui nous sont apparus comme fondamentaux pour construire un corpus valide compte tenu du problème que nous étudions. Ces paramètres ont été établis à la suite d'une première

analyse de quelques documents, dont le but était de déterminer les facteurs de variation des types d'erreurs en fonction des types de documents. Selon les documents et les auteurs, on observe une grande disparité dans les types de fautes.

Pour garantir une diversité et une bonne représentativité des erreurs, il convient de prendre en compte les paramètres suivants :

- les registres de langue ;
- les domaines abordés (ceux du voyages et tourisme, de la médecine, de la technologie et des sciences, etc.) ;
- les sources des documents (professionnelles *vs* privées, avec leurs sous-classes) ;
- les auteurs (79 au total) et leur niveau de performance linguistique ;
- le lecteur ou le public visé ;
- le niveau de contrôle présumé des documents.

2.2. Les paramètres de la constitution du corpus

Les différents paramètres pris en compte lors de la constitution de notre corpus sont présentés ci-dessous :

- **Diversité des documents et des niveaux de contrôle** : de façon à être représentatif des erreurs textuelles les plus courantes, notre corpus comporte environ 140 pages de textes, 90 pages pour les documents avec un niveau de contrôle peu élevé, et 50 pages pour les documents soumis à un niveau de contrôle élevé. Nous avons tenté de prendre en considération un large éventail de documents relativement faciles à obtenir : emails, blogs, forums, publications, rapports, etc. Les emails et les notes de forums, de blogs et de pages Web personnelles sont associés à un niveau de contrôle faible. Les publications, les rapports (domaines universitaires et professionnels) et les pages Web professionnelles sont associés à un niveau de contrôle élevé. Pour toutes ces catégories, les situations sont très variables et nous observons l'existence d'un continuum, allant d'un niveau de contrôle très élevé à un niveau de contrôle très faible. Le registre de langue est associé au type de document.
- **Diversité des auteurs** : nous avons considéré environ 35 auteurs pour les documents d'un niveau de contrôle faible, et 25 auteurs pour les documents d'un niveau de contrôlé élevé, provenant de domaines divers et faisant preuve de différents niveaux de compétence en anglais.
- **Diversité des domaines** : La diversité des documents permet de toucher à différents domaines, par exemple le service public hospitalier, l'entreprise (public et privé), le tourisme, l'administration publique, la gastronomie, etc. Cette diversité permet la prise en compte des savoir-faire, des pratiques linguistiques et des modes d'expression propres à certaines communautés.

Les caractéristiques de notre corpus actuel, défini pour une étude de faisabilité, sont résumées dans les tableaux 4.1, 4.2 et 4.3 :

Tableau 4.1 – **Niveau de contrôle**

Type de documents / Niveau de contrôle		**Très contrôlé**	**Contrôle suffisant**	**Peu contrôlé**
Publications		×		
Sites Web	Personnel			×
	Office de tourisme		×	
	Gastronomiques		×	
	Universitaires	×		
	Hôteliers		×	
	Ministère des Affaires étrangères	×		
Emails	Universitaires		×	
	Informatique			×
	Aéronautique			×
	Médical			×

Tableau 4.2 – **Diversité des auteurs**

Type de documents	**Taille (nombre de mots)**	**Diversité des auteurs**
Publications	33564	25
Sites Web	7694	19 (estimation)
Emails	9331	35

Tableau 4.3 – **Diversité des domaines**

Type de documents	**Domaine**	**Taille (en nombre de mots)**
Publications	Universitaire	33564
Sites Web	Personnel	457
	Offices de tourisme	2194
	Gastronomiques	1836
	Universitaires	2380
	Hôteliers	601
	Ministère des Affaires étrangères	226
Emails	Universitaires	2097
	Informatique	1753
	Aéronautique	3018
	Médical	2282
	Familial	181

3. UNE MÉTHODE DE CATÉGORISATION DES ERREURS

À l'interface de la didactique et de la linguistique, l'apprentissage des langues étrangères est un domaine d'étude qui suscite de nombreuses recherches, notamment l'apprentissage de l'anglais. L'analyse des erreurs produites par les apprenants dans la langue cible en est une composante importante. L'hypothèse selon laquelle les erreurs seraient la conséquence des interférences dues aux « habitudes » des apprenants dans leur langue maternelle (ou langue source) a donné naissance au domaine de l'analyse contrastive (*Contrastive Analysis*). Cette hypothèse a été remise en question dans les années 1970, avec le développement du domaine de l'analyse de l'erreur (*Error Analysis*), qui offrait une méthodologie pour l'exploration de la « langue » des apprenants (*learner language*). Cette méthodologie inclut les phases de constitution d'un corpus, de relevé des erreurs, de leur description et de leur explication. La dernière étape explore les raisons expliquant la production d'erreurs (surgénéralisation, mauvaise connaissance d'une règle de grammaire, etc.), et relève notamment des domaines de la psycholinguistique et des sciences de l'éducation. La création de catégories d'erreurs est un élément important de la phase de description des erreurs relevées.

3.1. Explicitation de la méthode et des paramètres pris en considération

Précisons dès à présent qu'il s'agit bien de catégoriser les erreurs selon leur nature propre et non en fonction de la correction qui pourrait être apportée. Selon Ellis (1994, p. 54-56), les erreurs sont le plus souvent classées selon deux méthodes différentes. La première consiste à utiliser ce qu'il appelle des « catégories linguistiques », c'est-à-dire les parties du discours, les structures ou systèmes précis (système des auxiliaires, formation des propositions, etc.), ou encore des niveaux très généraux tels que la morphologie, la syntaxe et le lexique. Izumi *et al.* (2005, p. 76) proposent par exemple une classification des erreurs d'apprenants Japonais en anglais sur la base des parties du discours (nom, verbe, modal, adjectif, adverbe, etc.). Cain *et al.* (1989, p. 25) fondent leurs catégories d'erreurs d'apprenants francophones de l'anglais sur des systèmes tels que la détermination, la deixis, ou encore la transitivité. La seconde méthode consiste à observer les « stratégies de surface » (*surface strategies*) telles que l'omission, l'ajout ou le choix erroné d'un élément. Cerda *et al.* (1999, p. 87) présentent un système de catégories qui utilise ces deux méthodes pour classifier les erreurs d'hispanophones s'exprimant en anglais : les catégories relèvent de niveaux linguistiques généraux mais des distinctions plus fines sont établies à l'intérieur de ces dernières pour refléter les stratégies de surface.

Parfois, les erreurs sont regroupées dans quelques catégories *ad hoc*, c'est-à-dire valables uniquement pour la paire de langues à l'étude, et reflétant des points de friction précis entre langue source et langue cible. Par exemple, Chan (2004, p. 56) choisit d'étudier et de classifier seulement cinq types d'erreurs représentatifs de la production en anglais de locuteurs du chinois de Hong-Kong.

La création de catégories *ad hoc* ne nous semblait pas convenir à une étude visant la description générale des erreurs produites par des francophones s'exprimant en anglais. La méthode de constitution des catégories basée sur l'observation des stratégies de surface ne nous semble pas permettre, quant à elle, de décrire finement les erreurs produites, et offre peu d'informations concernant leur source.

La catégorisation présentée dans cette étude repose donc sur des critères linguistiques. Nous avons choisi de regrouper les erreurs selon le groupe syntaxique (ou syntagme) dont elles relèvent le plus, ce qui garantit d'une part la reconnaissance de ces groupes par le plus grand nombre d'utilisateurs possible, et permet d'autre part de donner une visibilité aux erreurs de syntaxe à l'intérieur de ces groupes. De plus, les catégories ainsi formées sont de même niveau linguistique, ce qui confère une cohérence interne au système de classification. Ce choix de catégorisation est également intéressant car le système ainsi obtenu peut être transféré à d'autres paires de langues. La catégorisation présentée ici inclut aussi des catégories internes plus fines, et qui ont été formées à la suite de l'observation des types précis d'erreurs relevées.

3.2. Présentation de la catégorisation

Les catégories sont illustrées d'exemples, suivis de leur correction. Les lettres (P) et (I) font référence à la source des segments relevés (P pour « publications et rapports », I pour « emails, notes de forums, sites Web (Internet) »). Nous indiquons également la fréquence d'occurrence du type d'erreur à l'aide des lettres (TF) pour « très fréquent », (N) pour « normal », et (R) pour « rare », ces fréquences étant relatives à nos corpus.

GROUPE VERBAL

- ORDRE DES COMPOSANTS PLACÉS À LA SUITE DU VERBE (ARGUMENTS ET AJOUTS)
  - Objet direct séparé du verbe (TF)
    Ontological domains include in our view objects, their properties and relations. (P)
    ✓ In our view, ontological domains include objects…
  - Placement de la particule adverbiale (*phrasal verbs*) (N)
    It does not take into account context. (P)
    ✓ It does not take context into account.

- RÉALISATION DANS LE GROUPE VERBAL DES CONTRAINTES LEXICALES LIÉES AU VERBE
  - Choix de la préposition qui doit suivre le verbe (TF)
    These scores depend from the gold standard. (P)
    ✓ These scores depend on the gold standard.

- Transitivité (TF)
  I'<u>m waiting</u> your answer. (I)
  ✓ I'm waiting for your answer.

■ ADVERBE
- Placement erroné après le verbe principal (TF)
  They exhibit <u>nevertheless</u> the dependency relationships observed in the source parse tree. (P)
  ✓ Nevertheless, they exhibit the dependency relationships…
- Placement maladroit avant le verbe principal (R)
  The value of N is <u>experimentally</u> elaborated. (P)
  ✓ The value of N is elaborated experimentally.
- Utilisation des adverbes de négation (N)
  They are <u>not only</u> constrained to the author's point of view <u>any more</u>. (P)
  ✓ They are not constrained only to the author's point of view anymore.

■ CHOIX DE L'ASPECT (CONTINU, PERFECTIF) (TF)
This summer, our association <u>organizes</u> a trip. (I)
✓ This summer, our association is organizing a trip.

■ CHOIX DES AUXILIAIRES MODAUX (N)
It appears that patients who suffer from FRS <u>would</u> be unable to correctly monitor their actions. (P)
✓ It appears that patients who suffer from FRS are unable to…

■ MORPHOLOGIE :
- Formation des temps composés (*present perfect*) (N)
  You <u>do not have takes</u> action. (I)
  ✓ You have not taken action.

GROUPE NOMINAL

■ ADJECTIF
- Placement de l'adjectif par rapport au nom (R)
  The carrying of weapons is permitted in fifty <u>states different</u>. (I)
  ✓ The carrying of weapons is permitted in fifty different states.
- Apposition abusive de l'adjectif (TF)
  A subset of classes <u>useful</u> to describe our model. (P)
  ✓ A subset of classes which is useful to describe our model.
- Ordre des adjectifs dans une liste (N)
  <u>European academic and industrial</u> partners (P)
  ✓ academic and industrial European partners

- Placement de l'adverbe modifiant l'adjectif (TF)
  A quite detailed analysis. (P)
  ✓ Quite a detailed analysis.

■ Détermination
- Choix de l'article (TF)
  A Merovingian necropolis was built on exact site of the villa. (I)
  ✓ A Merovingian necropolis was built on the exact site of the villa.

■ Construction de la subordonnée complétive du nom (R)
  […] *all the feelings that another is controlling the patient's thoughts* (P)
  ✓ the feeling that another is controlling…

■ Construction NØN
- Construction abusive pouvant générer des problèmes de compréhension (TF)
  Security object granularity (P)
  ✓ The granularity of security objects
- Construction erronée, agrammaticale (TF)
  The objects properties (P)
  ✓ The properties of the objects, the objects' properties

■ Morphologie
- Accord déterminant/nom (N)
  I didn't order this goods (I)
  ✓ I didn't order these goods
- Accord agrammatical adjectif/nom (R)
  News clothes (I)
  ✓ New clothes

GROUPE PRÉPOSITIONNEL (y compris faisant partie d'un GN)

■ Choix de la préposition à utiliser en fonction du cotexte (N)
  They are exchanged and read on their electronic form (P)
  ✓ They are exchanged and read in their electronic form

PHRASE ET PROPOSITION

■ Construction des propositions interrogatives directes (TF)
  It is possible to receive the parcel by the end of August? (I)
  ✓ Is it possible to receive the parcel by the end of August?

- Construction des subordonnées
  - Interrogative indirecte (TF)
    It is necessary to know what <u>is their role</u> in the action expressed by the predicate. (P)
    ✓ It is necessary to know what their role is in the action…
  - Subordonnée à verbe non fini (N)
    They read annotations <u>for evaluating</u> them. (P)
    ✓ They read annotations to evaluate them.

- Placement des ajouts (TF)
  Goals and subgoals are most of the time realized by means of titles. (P)
  ✓ Most of the time, goals and subgoals are realized…

- Construction du comparatif (N)
  […] *<u>as many as possible of incorrect analyses</u>* (P)
  ✓ as many incorrect analyses as possible

- Gestion de l'information/micro-planification (N)
  August 15 in France, <u>it is a holiday</u>. (I)
  ✓ In France, August 15 is a holiday.

- Morphologie
  - Accord sujet/verbe (N)
    The first written mention of Issigeac <u>date</u> from 1008. (I)
    ✓ The first written mention of Issigeac dates from 1008.

LEXIQUE ET AUTRES

- Connaissance des caractéristiques lexicales d'un terme
  - Nom *vs* Verbe (N)
    You will give my <u>apologize</u> to her. (I)
    ✓ You will give my apologies to her.
  - Nom dénombrable *vs* Nom indénombrable (TF)
    A valuable information (P)
    ✓ valuable information
  - Adverbe *vs* Conjonction (N)
    <u>Although,</u> such an excess of mental effort should be reduced at all costs (P)
    ✓ However, such an excess…
  - Signification précise (TF)
    […] *to <u>remind</u> a document*
    ✓ to remember a document

- UTILISATION DE COLLOCATIONS ET D'EXPRESSIONS IDIOMATIQUES (TF)
  Well cordially (I)
  ✓ Yours sincerely

- ORTHOGRAPHE (N)
  A shedulle (I)
  ✓ A schedule

- PONCTUATION (TF)
  The purchase price will be validated by you and me, for the year. (I)
  ✓ The purchase price will be validated by you and me for the year.

3.3. Difficultés liées à la catégorisation des erreurs

La complexité des erreurs peut rendre la tâche de catégorisation relativement difficile. Pour commencer, une même erreur peut parfois être classée dans deux catégories. Par exemple, dans le segment **Our system is able to derive automatically information*, le placement erroné d'un adverbe après un verbe admettant un complément d'objet direct provoque également la séparation du verbe et de son objet : le segment erroné peut donc trouver sa place dans ces deux catégories.

De plus, certains segments erronés présentent plusieurs erreurs juxtaposées ou hiérarchisées, comme dans l'exemple suivant : **Nobody have answer me*. Ici, le segment présente deux erreurs de morphosyntaxe, l'une concernant l'accord du sujet et du verbe (Phrase et proposition – Morphologie), et l'autre la formation d'un temps composé (Groupe verbal – Morphologie). Le choix de l'utilisation de l'aspect perfectif est également une erreur (Groupe verbal – Choix de l'aspect). Ce segment cumule donc trois erreurs relevant de trois catégories différentes.

Enfin, il est parfois malaisé d'évaluer la nature de l'erreur, comme dans le segment suivant : **It is essential to be honest on both sides so that functions*. Ici, la double étiquette de *that* (pronom démonstratif / conjonction de subordination) rend cet erreur ambiguë : il peut s'agir d'une omission du sujet, ou bien de l'omission de la seconde partie de la conjonction *so that*, *that* étant ici traité comme le sujet (ceci peut être le résultat d'un calque du français « pour que cela fonctionne »). Cette ambiguïté n'empêche pas la correction mais peut mener à des difficultés de catégorisation.

Si les erreurs observées peuvent présenter un obstacle à la catégorisation, la nature même de cette tâche se révèle souvent problématique. Pour commencer, tout système de catégorisation est dans une certaine mesure *ad hoc*, puisqu'il repose sur une paire de langues précise. Par exemple, on ne peut douter que les locuteurs natifs du thaï (langue sans morphologie et sans ordre imposé des composants dans la phrase) s'exprimant en anglais produisent des erreurs de nature différente de celles trouvées dans l'expression des francophones.

De plus, il est difficile de proposer un système de catégories pertinent pour tous les types de corpus étudiés ici, car on y trouve des erreurs de natures différentes. Les productions très contrôlées telles les publications présentent des erreurs de syntaxe complexe (subordination, construction NØN, etc.), mais peu d'erreurs concernant la morphologie ou le lexique, notamment car les auteurs peuvent avoir recours à des correcteurs informatiques. Les productions d'un niveau de contrôle faible incluent en revanche de nombreuses erreurs morphologiques et lexicales, tandis que les erreurs syntaxiques touchent plutôt la syntaxe simple, car les auteurs évitent de produire des énoncés très complexes et de multiplier le risque d'erreur. Le modèle de classification que nous avons choisi permet néanmoins de rassembler ces erreurs dans les mêmes catégories principales, puisqu'elles peuvent convenir à tout type de texte. Les distinctions se retrouvent alors au niveau des catégories internes.

Toutefois, il faut garder à l'esprit que le choix de catégories est toujours arbitraire dans une certaine mesure. En effet, notre système n'offre pas la possibilité de classer les erreurs selon qu'elles relèvent de la morphologie ou de la syntaxe, et les erreurs liées au lexique ne peuvent pas toutes être regroupées dans la dernière catégorie (*i.e.* Lexique et autres). Cependant, choisir ce modèle de classification se révèlerait également insatisfaisant, d'une part parce qu'il ne permettrait pas la détermination des erreurs selon leur groupe syntaxique, et d'autre part parce qu'il générerait tout autant de flottements (lexique et grammaire sont parfois difficiles à démêler).

Enfin, la dernière difficulté recensée concerne la granularité, ou le degré de précision à donner au système de catégories. Si des catégories très précises permettent de décrire finement les erreurs rencontrées, elles sont un frein à l'implémentation des résultats dans un système informatique. Il faudrait donc pouvoir adapter la catégorisation aux objectifs de la recherche, tout en leur conservant leur pertinence en tant que système indépendant.

3.4. La catégorisation idéale

La solution à certaines des difficultés présentées précédemment pourrait être de créer un système de catégorisation multidimensionnel, capable de croiser des informations linguistiques générales (morphologie, syntaxe, lexique, etc.), avec les groupes linguistiques dont semblent relever les erreurs. Ce système pourrait également inclure la mention des stratégies de surface (omission, ajout, emploi erroné, etc.). De plus, il serait nécessaire d'évaluer avec le plus de justesse possible les sources de l'erreur (surgénéralisation, transfert syntaxique, ignorance d'une règle), qui pourraient également être une dimension du système, dans le but de faciliter la proposition de corrections pertinentes (Suri et McCoy, 1993).

4. ANNOTATION DES ERREURS

Introduisons à présent le schéma d'annotation que nous avons conçu. Celui-ci est encore au stade de validation et peut faire l'objet de précisions et de modifications. En effet, stabiliser un schéma aussi complexe demande de nombreuses expérimentations en boucle. Ce schéma est une tentative pour refléter, de façon factuelle et déclarative, les différents paramètres qui ont émergé lors de l'analyse des comportements cognitifs déployés par les didacticiens et traducteurs humains lorsqu'ils détectent et corrigent une erreur.

Nous employons un formalisme très standard, XML, enrichi d'attributs dont les valeurs associées ont une granularité élaborée par des didacticiens de notre groupe. La structure des attributs a été conçue de façon à ce qu'ils puissent être utilisés dans un cadre d'argumentation (Albert *et al.*, 2009).

Nous présentons ci-dessous les différentes étapes de l'annotation, en partant de la détection de l'erreur pour aller jusqu'à sa correction. Les étiquettes sont en anglais, ce qui permet d'avoir par la suite un métalangage utilisable dans de nombreux contextes.

4.1. Délimitation et caractérisation de l'erreur

<error-zone> étiquette le groupe de mots liés à l'erreur. Cette zone doit être aussi réduite que possible. Cette étiquette a plusieurs attributs :

Tableau 4.4 – **Annotation des erreurs**

Comprehension	Avec des valeurs de 0 à 4 (0 étant le plus mauvais), indique si le segment erroné est malgré tout compréhensible ou non.	0 : incompréhensible, 1 : compréhensible avec beaucoup d'efforts, 2 : compréhensible avec un peu d'effort, 3 : compréhensible mais peu naturel, 4 : bon
Grammaticality	Avec des valeurs de 0 à 2, indique le niveau de gravité de l'erreur.	0 : agrammatical, 1 : doute quant à la grammaticalité du segment, 2 : grammatical
Categ	Indique la catégorie principale de l'erreur. Si une erreur couvre plusieurs champs, ce qui est assez fréquent, il est possible de spécifier plusieurs valeurs. Le choix d'une catégorie peut être aussi parfois quelque peu ambigu.	Lexicale, syntaxique, stylistique, sémantique, textuelle
Source	Indique, globalement, l'origine de l'erreur.	Calque (lexical ou grammatical), surgénération, manque d'effort, etc.

4.2. Délimitation et caractérisation de la correction

<correction-zone> identifie le fragment de texte sur lequel se fera la correction. Il est égal ou plus grand que <error-zone>. Nous abordons ici le point central. Chaque correction possible de l'erreur est introduite par le tag <correction> avec ses attributs associés. Nous donnons ici chaque attribut avec ses valeurs. En général, un système simple à trois valeurs a été adopté. Celui-ci peut bien entendu évoluer si la faisabilité est possible. La ou les valeurs positives sont soulignées :

Tableau 4.5 – **Annotation des corrections**

Surface	caractérise la taille du texte affecté par la correction	minimal, average, maximal
Grammar	indique, lorsque c'est approprié, le statut de la correction	by default, alternative, unlikely
Meaning	indique si, par la correction proposée, le sens a été altéré ou non	yes, no, somewhat
Change	indique la nature de la correction proposée	syntactic, lexical, stylistic, sémantic, textual
Comp	indique si la correction proposée est un fragment de texte facile à comprendre	yes, average, no
Fix	indique si la correction correspond à une situation locale, idiosyncratique, qui ne peut être étendue à aucune autre situation similaire	yes, no
Qualif	indique le niveau de compétence et de certitude que l'annotateur a de sa correction	low, average, high
Correct	indique le segment de texte corrigé	

À titre d'illustration, voici le cas de la construction NØN incorrecte, comme dans *the meaning utterance* (ou *the goal failure)*, où deux corrections sont possibles :

```
<error-zone comprehension="2" agrammaticality="1"
categ="syntax" source="calque">
     the meaning utterance
     <correction qualif="high" grammar="by-default"
          surface= "minimal" meaning= "not altered" Var-size="+2"
          change="synt" comp="yes"
          correct= "the meaning of the utterance">
     </correction>
     <correction qualif="high" grammar="unlikely"
          surface= "minimal" meaning= "somewhat" Var-size="0"
          change="lexical+synt" comp="average"
          correct= "the meaningful utterance">
     </correction>
</error-zone> </correction-zone> without a context.
```

5. CONCLUSION

Dans cet article, nous avons présenté certaines des étapes préliminaires à la réalisation d'un assistant informatique permettant d'améliorer les compétences et les performances de francophones rédigeant des textes en anglais. Ce projet implique d'établir des procédures et d'étudier les stratégies de détection et de correction des erreurs. Notre corpus exploratoire a été constitué en tenant compte de plusieurs paramètres, dont les différents types des documents et de niveaux de contrôle, ainsi que la diversité des auteurs et des domaines. L'objectif de la prise en considération de ces paramètres est de rassembler et de décrire des erreurs représentatives des productions de francophones dans différents contextes. Les documents qui constituent notre corpus exploratoire proviennent donc de sources très diverses (université, commerce, forums de discussion) et représentent des niveaux de contrôle allant de très fort (p. ex., publications scientifiques) à très faible (p. ex., emails personnels).

L'étape suivante que nous avons abordée est la méthode de catégorisation des erreurs. Nous avons choisi de classer les erreurs principalement selon leur groupe syntaxique le plus probable, en créant également des catégories plus fines à l'intérieur de ces groupes généraux. Ce choix permet à notre système de conserver une cohérence interne et d'être transférable à d'autres langues de structure relativement similaire. De plus, le degré de granularité choisi semble être en accord avec notre objectif de création d'un système informatisé. La catégorisation pourrait être améliorée par la mise en place d'un système multidimensionnel permettant d'inclure et de croiser d'autres types d'information (morphologie, syntaxe, lexique, etc.). Il s'avèrera nécessaire d'analyser les sources psycholinguistiques et cognitives des erreurs afin d'entrer dans la phase d'explication des erreurs et ainsi, à terme, de proposer à l'utilisateur des corrections aussi pertinentes et utiles que possible.

La proposition d'un schéma d'annotation des erreurs relevées permet de refléter les paramètres qui émergent lors de la phase de détection et de correction des erreurs par des correcteurs humains. Le schéma proposé utilise un formalisme XML complété par des balises spécifiques, et inclut les étapes de délimitation et de caractérisation de l'erreur, et de délimitation et caractérisation de la correction. Nous prévoyons d'utiliser ce schéma afin de produire des règles de correction. Les règles de correction seront produites par induction à partir des exemples annotés. Il s'agit d'une forme d'apprentissage rudimentaire, mais bien adaptée à notre problématique. Une des difficultés sera de dégager les bons niveaux d'abstraction tout en gérant de façon fiable des ensembles d'exceptions ou particularismes.

RÉFÉRENCES

ALBERT, C., L. BUSCAIL, M. GARNIER, A. RYKNER et P. SAINT-DIZIER (2009). « Annotating language errors in texts : investigating argumentation and decision schemas » *ACL-LAWIII*, Singapour.

CAIN, A. (dir.) (1989). *L'analyse d'erreurs, accès aux stratégies d'apprentissage : une étude interlangues*, INRP.

CERDA, E., C. VALERO, C. FLYS et G. MANCHO (1999). « Error Analysis and the Teaching of English in Tourism Studies », *III Congres Internacional sobre Llengues per a Finalitats Especifiques*, Universitat de Barcelona, p. 86-89.

CHAN, A.Y.W. (2004). « Syntactic Transfer : Evidence from the Interlanguage of Hong Kong Chinese ESL learners », *The Modern Language Journal*, vol. 88, nº 1, p. 56-71.

ELLIS, R. (1994). *The Study of Second Language Acquisition*, Oxford, Oxford University Press.

HAN, N.R. *et al.* (2005). *Detecting Errors in English Article Usage by Non-native Speakers*, NLE.

IZUMI, E., K. UCHIMOTO et H. ISAHARA (2005). « Error Annotation for Corpus of Japanese Learner English », *Proceedings of the 6th International Workshop on Linguistically Annotated Corpora*, p. 71-80.

LEE, H. et A. SENEFF (2006). « Automatic Grammar Correction for Second-language Learners », *Proceedings of InterSpeech*.

SURI, L. Z., et McCOY, K.F. (1993). « A Methodology for Developing an Error Taxonomy for a Computer Assisted Language Learning Tool for Second Language Learners », *Technical Report* nº 93-16, Department of Computer and Information Sciences, University of Delaware, Newark.

CHAPITRE 5

CONSTRUCTION (SEMI-)AUTOMATIQUE D'UN LEXIQUE SÉMANTIQUE DU FRANÇAIS

Inférences interlinguistiques et morphologie

Nilda Ruimy, Pierrette Bouillon, Bruno Cartoni et Fiammetta Namer

RÉSUMÉ

Nous décrivons deux méthodes complémentaires pour la dérivation semi-automatique d'un lexique du français, suivant le modèle du lexique génératif. La première méthode exploite la similitude entre les langues. Les informations lexicales françaises sont dérivées à partir d'un dictionnaire sémantique électronique de l'italien. À cet effet, nous combinons deux stratégies :

– *pour les mots construits, nous exploitons la cognacité de certains suffixes. Après avoir recherché dans un dictionnaire bilingue la traduction française du mot italien avec le suffixe français correspondant (p. ex.,* costruzione → *construction), nous générons l'entrée lexicale française en y transférant les informations sémantiques de l'entrée italienne ;*

– *pour les mots non construits polysémiques (p. ex.,* frazione *[fraction, hameau]), nous utilisons les indicateurs de sens fournis dans les dictionnaires bilingues (pour* frazione*: (*mat.*); (*centro abitato*)) de manière à déterminer, dans le lexique électronique italien, l'entrée lexicale appropriée qui donnera naissance à l'entrée du sens français équivalent.*

La seconde méthode fait appel aux principes de la morphologie constructionnelle du français et permet potentiellement de coder des mots construits dont les équivalents italiens ne figurent pas dans le lexique source. Ces mots nouveaux, extraits de corpus, sont soumis à l'analyseur dérivationnel DériF dont le résultat fournit une grande partie de l'information lexicale permettant la construction de nouvelles entrées françaises.

1. INTRODUCTION

De nos jours, le multilinguisme est une question essentielle, et de nombreux efforts sont entrepris, notamment en Europe, pour favoriser la communication entre les 27 langues officielles de l'Union européenne. Dans ce contexte, l'Europe élargie a grandement besoin de technologies linguistiques dont le développement repose, entre autres, sur l'existence de larges ressources lexicales. Ces ressources doivent non seulement être disponibles pour toutes les langues de l'Union, mais également être comparables en termes de taille et de qualité. Cependant, la construction de telles ressources lexicales représente aujourd'hui encore un vrai défi, coûteux en matière de prix et de temps. Il serait donc souhaitable, au lieu de multiplier les initiatives séparées, de réutiliser les connaissances et les expériences acquises antérieurement dans le cadre de projets de constitution de ressources lexicales dans une langue donnée, pour induire les informations nécessaires à la constitution d'un lexique dans une autre langue.

C'est dans cet objectif que nous présentons ici une étude de faisabilité qui vise à dériver de manière semi-automatique un lexique sémantique pour le français à partir d'un lexique monolingue de l'italien : PAROLE-SIMPLE-CLIPS (ci-après PSC[1]). Ce lexique comprend un niveau de description sémantique dont l'approche théorique suit les principes du lexique génératif (Pustejovsky, 1995, 2001). Ce cadre de représentation a été un facteur déterminant dans le choix de PSC comme source de dérivation. Les lexiques construits selon cette approche

1. Par économie typographique, nous utilisons ici « PSC » qui n'est cependant pas l'acronyme du lexique.

sont en effet particulièrement utiles dans la perspective qui nous intéresse ici : ils contiennent un grand nombre d'informations et fournissent ainsi une représentation lexicale qui couvre tous les aspects du sens des lexèmes. Dans ce type de lexique, le sens d'un lexème est décrit selon quatre niveaux de représentation sémantique qui en capturent les aspects compositionnels, définissent le type d'événement qu'il dénote, décrivent son contexte sémantique et le positionnent par rapport aux autres mots présents dans le lexique.

Dans l'article qui suit (Ruimy *et al.*, 2005), nous présentons brièvement le lexique PSC, ainsi que différentes méthodologies expérimentées pour dériver un lexique à partir d'autres ressources existantes. Les deux premières approches reposent sur un lexique génératif existant dans une autre langue de même famille ; la troisième permet d'étendre les lexiques *a priori* ainsi développés par des mots construits (en particulier des néologismes), souvent absents des dictionnaires de base, en utilisant des connaissances morphologiques et d'autres ressources langagières générales (dictionnaire et corpus). Nous pensons que ces trois méthodes, ensemble, permettent d'obtenir rapidement des lexiques multilingues utiles pour des applications concrètes. Dans cet article, celles-ci sont d'ailleurs évaluées et comparées de manière à montrer leur apport respectif.

2. LE LEXIQUE PAROLE-SIMPLE-CLIPS

La ressource à partir de laquelle cette expérience a été menée est la plus importante base de connaissance lexicale multi-niveau informatisée de la langue italienne. Le lexique PSC (Ruimy *et al.*, 2002), élaboré au cours de trois projets[2] différents, contient 387 000 entrées phonologiques, 53 000 entrées morphologiques, 64 500 entrées syntaxiques et 57 500 entrées sémantiques, codées suivant des standards internationaux, élaborés lors de la création du modèle PAROLE-SIMPLE (Ruimy *et al*, 1998 ; Lenci *et al.*, 2000a et b). PSC offre ainsi l'avantage d'être compatible avec les onze autres lexiques PAROLE-SIMPLE construits pour toutes les langues européennes, avec qui il partage le même modèle théorique et langage de représentation.

Dans ce lexique, la description d'un lexème inclut toutes les informations phonologiques, morphologiques, syntaxiques et sémantiques qui lui sont propres. Son cadre de sous-catégorisation est décrit en termes d'optionalité, de fonction syntaxique, de réalisation syntagmatique et de restrictions morphosyntaxiques et lexicales de ses compléments. Au niveau sémantique, l'approche théorique

2. Les niveaux morphologique et syntaxique ont été conçus au cours du projet LE-PAROLE. Le modèle linguistique ainsi que le noyau du lexique sémantique ont été élaborés dans le cadre du projet LE-SIMPLE, alors que la description phonologique et l'extension de la couverture lexicale ont été effectuées dans le contexte du projet italien *Corpora e Lessici dell'Italiano Parlato e Scritto* (CLIPS).

adoptée dans le modèle SIMPLE repose essentiellement sur une version revisitée du lexique génératif. Chaque unité sémantique de SIMPLE-CLIPS est pourvue de différents types d'informations très détaillées et utiles pour une exploitation en traitement automatique du langage naturel (TALN), comme la classification ontologique : le lexique est en effet structuré selon un système de types multidimensionnels, basés sur des relations conceptuelles hiérarchiques et non hiérarchiques, selon le principe d'héritage orthogonal (Pustejovsky et Boguraev, 1993). Les propriétés définitoires de chaque type sémantique sont ensuite regroupées dans un *template*, c.-à-d. une structure schématique où sont établies les propriétés ainsi que les contraintes de bonne formation des unités lexicales lui appartenant. Outre la classification ontologique, chaque entrée contient également des informations concernant le domaine d'utilisation de l'unité lexicale décrite, le type d'événement dénoté, les relations de synonymie et de dérivation morphologique, son appartenance à une classe de polysémie régulière, ainsi que d'autres traits sémantiques distinctifs. Les informations les plus caractéristiques de ce modèle sont cependant les relations de la structure des qualia étendue (*Extended Qualia Structure*) et la représentation prédicative.

La structure des qualia étendue permet de modéliser les différentes dimensions du sens d'un mot, ainsi que ses relations avec les autres unités lexicales. Pour exprimer ce type d'informations, un ensemble de 56 relations sémantiques, subsumées par les quatre rôles originaux de la structure des qualia[3] (Pustejovsky, 1995), a été créé dans le cadre du modèle SIMPLE. Ces relations permettent de caractériser une unité lexicale par une relation hyperonymique, de décrire ses propriétés méronymiques et d'indiquer son type d'origine et de fonction en l'associant, par des liens intra- ou inter-catégoriels, à d'autres unités lexicales.

La représentation prédicative expose, quant à elle, le scénario sémantique auquel le terme décrit est associé et permet de caractériser ses participants en termes de rôle thématique et de contraintes sémantiques. De plus, dans la description d'un lexème, tous les niveaux de représentation linguistique sont reliés et les informations sémantiques et syntaxiques, en particulier, sont mises en relation par la projection de la structure argumentale sur sa ou ses réalisations syntaxiques.

La richesse et la granularité de ses informations font du lexique PSC, et en particulier du niveau sémantique SIMPLE-CLIPS, une source précieuse pour la dérivation d'autres ressources lexicales. Cette richesse requiert cependant la mise en œuvre de méthodes robustes et d'outils informatiques permettant l'acquisition de ces données.

3. À savoir, les rôles formel, constitutif, agentif and télique.

3. LA MÉTHODE PROPOSÉE

L'idée maîtresse qui a guidé l'expérience décrite ci-dessous est la suivante : en partant d'une base de données lexicales existante Lex1 (c.-à-d. SIMPLE-CLIPS, construite pour la langue L1 (l'italien), il devrait être possible, en utilisant les informations présentes dans les dictionnaires bilingues, d'inférer semi-automatiquement un lexique sémantique Lex2 pour une langue L2 (dans notre cas, le français). Sur la base des couples d'équivalents de traduction L1-L2 extraits de dictionnaires bilingues[4], les entrées de Lex2 sont générées en assignant au sens de L2 les propriétés sémantiques qui sont associées, dans Lex1, à leur équivalent de traduction L1.

Cette approche n'est pas nouvelle (voir, par exemple, Hong *et al.*, 2004; Fujita *et al.*, 2004). L'originalité réside cependant dans le fait que nous proposons trois volets d'inférences, qui exploitent au maximum les similarités entre les langues française et italienne. Pour les mots morphologiquement construits, nous exploitons d'abord la cognacité d'un ensemble de suffixes italiens et français. Ensuite, pour les mots non construits, ainsi que pour les mots construits qui ne peuvent être gérés par la première méthode, une approche basée sur les indicateurs de sens prend le relai. Finalement, les néologismes construits, extraits de corpus, ainsi que les mots construits n'ayant pas d'équivalents italiens dans le lexique source, sont soumis à un analyseur morphologique du français qui fournit les informations sémantiques nécessaires à la construction d'entrées lexicales. Ces trois approches sont décrites respectivement dans les sections 4, 5 et 6.

4. L'APPROCHE FONDÉE SUR LA COGNACITÉ

La première stratégie exploite le fait que les langues italienne et française partagent un grand nombre de similitudes en termes de structure lexicale et d'informations syntaxiques. Provenant toutes deux de la même langue d'origine (le latin), elles ont un lexique de base qui est en effet relativement similaire. Plusieurs études ont d'ailleurs montré le parallélisme de leur système morphologique, qui a, par exemple, déjà été exploité pour la génération bilingue de néologismes (Namer, 2001) et pour l'acquisition semi-supervisée de nouveaux lexiques (Schulz *et al.*, 2004). La première stratégie proposée ici tire parti de ces similarités et repose sur les deux hypothèses suivantes :

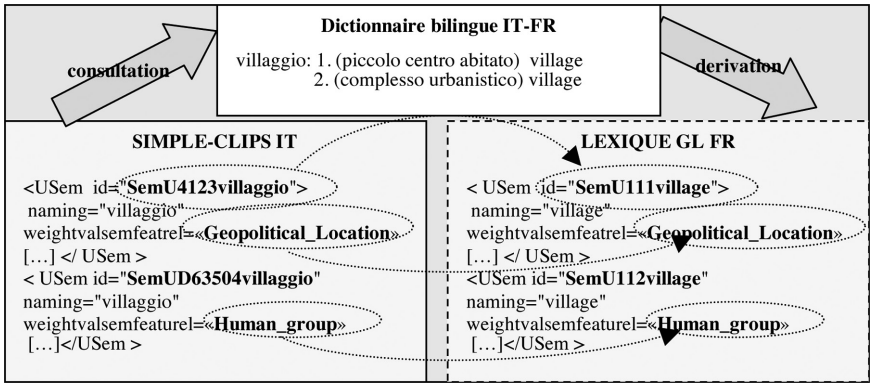
a) les mots morphologiquement construits possèdent un sens largement prévisible à partir de leur structure ;
b) les mots suffixés de l'italien possèdent un ou plusieurs équivalents français, construits avec un affixe équivalent, qui couvrent les mêmes sens que l'élément italien.

4. Dans cette expérience, deux sources bilingues différentes ont été évaluées : *Robert et Signorelli* et *Hachette-Paravia DIF*, tous deux étant des dictionnaires bilingues de grande qualité.

4.1. Méthodologie

La stratégie basée sur la cognacité se résume plus concrètement comme suit. La ressource de base est le dictionnaire bilingue *Robert et Signorelli* à partir duquel sont extraits tous les mots-vedettes italiens se terminant par des suffixes représentatifs des mots construits, avec, pour chaque sous-sens, les différentes traductions possibles. Par exemple, le mot-vedette *torrefazione* (en *-ione*) est extrait avec sa traduction pour les deux sens (torréfaction) : 1. processus de torréfier et 2. maison du café. Nous supposons ensuite que si le mot italien encodé dans SIMPLE-CLIPS possède, dans notre base bilingue, le même équivalent de traduction pour tous ses sous-sens, alors l'équivalent français partagera avec le mot italien toutes les entrées sémantiques de SIMPLE-CLIPS, comme c'est le cas avec *torrefazione*. Comme les langues sont proches, il n'est pas nécessaire qu'il s'agisse réellement de mots construits, synchroniquement analysables ; la méthode s'applique aussi aux mots diachroniquement encore parallèles. Par exemple, le mot en *-aggio villaggio* possède deux sens dans SIMPLE-CLIPS, l'un pour le lieu et l'autre pour l'ensemble des personnes vivant dans le village. Étant donné qu'il est toujours traduit par « village », nous pouvons inférer, sans autre analyse plus approfondie, que « village » partage les mêmes entrées SIMPLE-CLIPS que *villaggio* (figure 5.1). Dans la suite, nous évaluons cette approche pour les trois suffixes nominaux *-tà, -zione* et *-aggio*.

Figure 5.1 – **Méthodologie basée sur les cognats : vue d'ensemble**



4.2. Données de l'expérience

Pour évaluer la pertinence de cette méthode fondée sur la cognacité, nous avons extrait de SIMPLE-CLIPS, de manière aléatoire, 79 lexèmes se terminant par *-aggio*, 80 par *-tà* et 56 par *-zione*. Pour chacun de ces lexèmes, nous avons vérifié

dans le dictionnaire bilingue le nombre de traductions différentes (tableau 5.1). Puis, nous avons regardé, pour tous les mots italiens avec une seule et unique traduction en français couvrant tous leur sous-sens (colonne 2 dans le tableau 5.1), si cette traduction partageait les mêmes sens SIMPLE-CLIPS que le mot italien (tableau 5.2). Le tableau 5.1 montre que, comme prévu, un grand nombre de mots italiens se terminant par un suffixe **cognat** possède, pour tous leurs sens, une seule traduction qui se termine d'ailleurs par le suffixe correspondant en français.

Le tableau 5.2 est également très intéressant et montre que, dans la plupart des cas, cette traduction unique partage avec l'entrée italienne toutes les informations de SIMPLE-CLIPS. Le petit pourcentage d'erreurs est causé par certaines différences de granularité entre SIMPLE-CLIPS et les distinctions de sens du dictionnaire bilingue. Par exemple, SIMPLE-CLIPS propose pour *passaggio* une entrée spécifique pour le domaine du sport (la passe d'une balle). Dans cette acception, le mot devrait avoir une traduction spécifique en français (et distincte formellement du mot italien), qui est absente du dictionnaire bilingue. Étant donné que *passaggio* possède, pour tous les sens de cet exemple, le même équivalent de traduction « passage » dans le dictionnaire bilingue, notre algorithme inférera de manière erronée que « passage » est la traduction correcte également pour ce sens. Mais comme le montrent les résultats présentés dans le tableau 5.2, cette situation est très rare. Nous pouvons donc conclure que l'approche fondée sur les cognats nous permettrait de construire un **devineur de traduction** très efficace. Cependant, pour les mots construits qui possèdent plus d'une traduction, (colonne 3 du tableau 5.1), cette méthode est inadéquate et c'est l'approche basée sur les indicateurs de sens, décrite dans la section suivante, qui prend le relais.

Tableau 5.1 – **Traduction italien-français**

	Lexèmes IT avec même équivalent FR pour tous les sens	**Lexèmes IT avec plus d'une traduction**
-aggio	89,9 %	10,1 %
-tà	77,4 %	22,6 %
-zione	80,4 %	19,6 %

Tableau 5.2 – **Polysémie en français et en italien**

	Lexèmes FR partageant les entrées IT SIMPLE-CLIPS
-aggio	99,97 %
-tà	99,98 %
-zione	99,98 %

5. L'APPROCHE BASÉE SUR LES INDICATEURS DE SENS

Un bon dictionnaire bilingue fournit des équivalents de traduction en langue cible accompagnés, entre parenthèses, de mots (*strumento*), phrases (*colui che vigila*) ou abréviations (*med.*), qui permettent à l'utilisateur de sélectionner l'équivalent le plus correct. Ces indicateurs de sens fournissent des informations syntaxico-sémantiques, utilisées comme indices pour restreindre le sens du mot en langue source, et guider ainsi la sélection de l'équivalent approprié en langue cible (Atkins and Bouillon, 2004), par exemple : *capo* (*promontorio*) → cap, alors que *capo* (*filo*) → fil*; frazione* (*mat.*) → fraction*,* alors que *frazione* (*sport*) → relais.

Pour dériver des entrées sémantiques en L2 à partir du lexique sémantique de L1, nous proposons une approche basée sur les indicateurs de sens. Nous affirmons que, de la même manière qu'ils permettent de guider le choix de l'utilisateur dans un dictionnaire bilingue, ces indicateurs peuvent être utilisés pour la désambiguïsation lexicale sémantique dans la ressource d'origine, c'est-à-dire pour déterminer, dans le lexique sémantique SIMPLE-CLIPS, l'entrée correspondant au sens italien d'un couple de termes italien-français. Nous soutenons également qu'en vertu des similarités sémantiques existant entre équivalents de traduction, il est possible d'assigner, de manière supervisée, les principales propriétés sémantiques d'une entrée italienne à son équivalent français.

5.1. Méthodologie

La méthode basée sur les indicateurs de sens a été testée sur un groupe représentatif de 250 noms et verbes italiens (pour un total de 992 sens), sélectionnés dans SIMPLE-CLIPS en fonction de leur fréquence et en privilégiant les unités lexicales polysémiques. Pour évaluer cette approche, nous avons procédé de la manière suivante (figure 5.2):

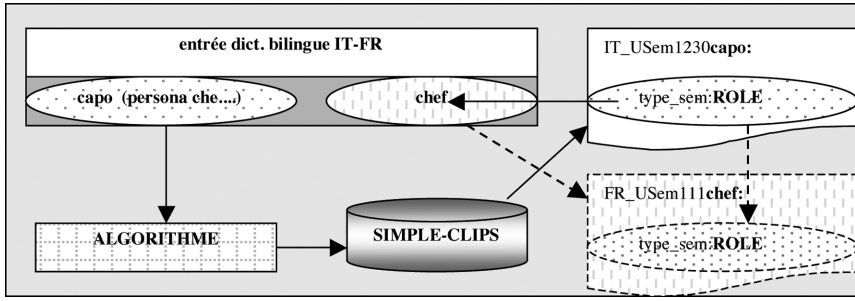
1. Sélection d'un sous ensemble de lemmes dans SIMPLE-CLIPS.
2. Sélection des triplets correspondants (mot **S**ource italien - **I**ndicateur de sens – équivalent **C**ible français, ci-après S-I-C) dans un dictionnaire bilingue[5].
3. Suivant l'étude réalisée par Atkins et Bouillon (2004), analyse et classement des indicateurs de sens en deux catégories, selon leur nature et leur fréquence :
   a) indicateurs de sens morphosyntaxiques, par exemple : sous-classe verbale, sélection de l'auxiliaire, forme plurielle des noms, sujet/objet typique, type de PP ;

5. *DIF*, *Il Dizionario Francese-Italiano Italiano-Francese*, Torino, Paravia-Hachette, 2000.

b) indicateurs de sens inférentiels, par exemple : synonyme, hyperonyme, méronyme, domaine d'usage.

4. Classification ultérieure des indicateurs de sens en fonction de leur pouvoir d'identification de l'entrée sémantique appropriée dans SIMPLE-CLIPS :
   a) indicateurs de sens fournissant des informations directement repérables dans l'entrée lexicale ;
   b) indicateurs de sens nécessitant une conversion dans le métalangage du lexique. Pour ce faire, nous avons mis au point une table de conversion qui met en correspondance les indicateurs de sens et les informations morphologiques, syntaxiques ou sémantiques de PSC.
5. Création et implémentation d'un algorithme à base de règles.
6. Étiquetage supervisé du mot français.

Figure 5.2 – **Méthodologie basée sur les indicateurs de sens : vue d'ensemble**



5.2. L'algorithme

L'étape préliminaire à la dérivation d'une entrée conforme au modèle SIMPLE en L2 (ici, le français) à partir d'une entrée en L1 (l'italien) consiste à déterminer dans le lexique SIMPLE-CLIPS l'unité sémantique (ci-après **USem**) pertinente, dont les propriétés seront attribuées à l'équivalent français, proposé par le dictionnaire bilingue. Dans ce but, une correspondance doit être établie entre les informations contenues dans les entrées SIMPLE-CLIPS et les indices fournis par les indicateurs de sens dans le dictionnaire. Trois types de règles ont donc été implémentées, qui permettent de retrouver, pour chaque triplet S-I-C extrait du dictionnaire bilingue, l'entrée sémantique S adéquate de SIMPLE-CLIPS en fonction de la nature de l'indicateur de sens I fourni dans le dictionnaire.

1. Recherche d'une entrée S SIMPLE-CLIPS dont le contenu inclut I. Dans cette entrée, S et I peuvent être reliés par :
   – une relation sémantique de type synonymique A[6] ou hyperonymique B[7] : par exemple, pour S = *capo* et I = (*testa*), l'entrée appropriée est celle de USem61397*capo*, qui inclut une relation de synonymie avec USem1753*testa*. En revanche, pour S = *capo* et I = (*persona che...*) (personne qui), l'entrée adéquate est USem3615*capo*, où le mot-vedette et la USemD735*persona* sont en relation d'hyperonymie ;
   – une quelconque relation de type qualia C : par exemple, pour S = *scuola* et I = (*movimento*), l'entrée appropriée est USem62940*scuola*, dans laquelle la relation « is_a_follower_of » relie l'entrée avec USem2576*movimento*.
2. Recherche d'une entrée SIMPLE-CLIPS partageant certaines propriétés avec une entrée de I. Les propriétés partagées peuvent être :
   – la cible de la relation hyperonymique D : par exemple, pour S = *comunicare* et I = (*notificare*), une des entrées de chacun de ces deux lemmes a comme cible de la relation hyperonymique USemD5576*dire*. L'entrée S appropriée est donc USem6472*communicare*, où le mot-vedette est relié à USemD5576*dire* par la relation « is_a » ;
   – la classification ontologique (les types de S et I pouvant être soit identiques soit en relation de subsomption) E : par exemple, pour S = *avvertire* et I = (*percepire*), chacun des deux lemmes a un sens appartenant au type sémantique « Experience_event ». L'entrée appropriée est donc celle qui a cette classification ontologique.
3. Recherche d'une entrée SIMPLE-CLIPS contenant des informations converties à partir des indicateurs de sens. Ces informations permettent de restreindre la recherche à :
   – un type sémantique ou une hiérarchie de type spécifique F : par exemple, 1) si I = (*persona che*), la sélection est restreinte à une entrée appartenant à la hiérarchie des êtres humains ; 2) si I = (*rendere…*) (rendre…), pour *illuminare* (*rendere luminoso*), la sélection s'effectue sur une entrée lexicale de ce lemme appartenant à la hiérarchie des événements causatifs. En revanche, si I = (*diventare…*) (devenir…), *scaldarsi* (*diventare caldo*), la recherche sera orientée vers la hiérarchie des événements inchoatifs ;

6. Les lettres majuscules sont reprises dans les différentes règles du tableau 5.3.
7. Les synonymes d'hyperonymes et les hyperonymes d'hyperonymes sont également pris en compte.

- un domaine spécifique G : Si I est une forme abrégée (par exemple *med.*), il est alors interprété comme une information de domaine d'usage et l'on recherche une entrée contenant la valeur correspondante dans la hiérarchie des domaines de SIMPLE ;
- un trait spécifique ou une relation sémantique H : par exemple, 1) si I = (*stare...*) ou (*restare ...*) (être…), l'entrée recherchée est celle où la valeur de l'attribut « Event_type » est « state » ; 2) si I commence par la préposition *per* (pour) ou *di* (de), la recherche est limitée aux entrées de SIMPLE-CLIPS contenant respectivement une relation télique ou constitutive, dont la cible correspond au mot suivant la préposition. Par exemple, *asfalto* (*per rivestire*), l'entrée est celle dont la cible de la relation télique « used_for » est *rivestire* ;
- une structure syntaxique I : C'est ici à travers l'entrée syntaxique correspondante que l'entrée sémantique S est repérée. Par exemple, si I = (intr.aus. *avere, con*), l'entrée appropriée est celle qui est reliée à l'unité syntaxique correspondante dont la sous-catégorisation inclut un complément prépositionnel introduit par « avec ».

Les règles de correspondance A à I ont été pondérées et classées de 1 à 9, selon le degré de « confiance » que les types d'information sur lesquelles elles sont basées confèrent aux résultats. L'ordre d'application des règles est en effet crucial : plus la règle est haut placée, plus le résultat est sûr. La probabilité que l'entrée S retournée soit correcte est plus élevée si la requête est basée sur une information concernant son domaine d'usage ou sa structure syntaxique (règle G ou I) que si elle est basée sur des propriétés (hyperonymie ou type sémantique) partagées avec une entrée de I (règle D ou E). En effet, ces propriétés sont souvent trop peu informatives. De même, la règle C peut produire des résultats insatisfaisants si la relation entre S et I est vague.

En résumé, 96,16 % des couples italien-français étudiés contenaient des indicateurs de sens dans l'entrée bilingue et ont donc pu être traités par l'algorithme. En appliquant les règles à ce sous-ensemble de données, nous avons obtenu un taux de rappel de 64,9 % (c.-à-d. quand l'élément italien d'un couple bilingue a été lié de manière adéquate à une entrée SIMPLE-CLIPS). Comme nous le montrons (en gras) dans le tableau 5.3, c'est à travers l'application des règles au classement le plus élevé que nous obtenons les meilleurs pourcentages.

Tableau 5.3 – **Distribution des taux de succès en fonction des règles**

Identificateur	**A**	**B**	C	D	E	F	**G**	**H**	**I**
Rang	**1**	**2**	9	7	8	6	**3**	**5**	**4**
Taux de succès	**16,6 %**	**26,8 %**	0,92 %	8,9 %	5,8 %	3,9 %	**12,3 %**	**9,2 %**	**15,4 %**

Outre ces résultats, 4,16 % des liens ont été correctement établis par l'application d'une règle par défaut qui relie automatiquement un sens unique du bilingue à une entrée unique dans SIMPLE-CLIPS. Enfin, parmi les 30,94 % d'échecs, des résultats théoriquement possibles ont été retournés dans 7,74 % des cas, qui pourraient être désambiguïsés manuellement.

Le succès de l'application de ces règles présuppose naturellement des niveaux de granularité comparables concernant les distinctions de sens dans les ressources monolingue et bilingue. Lors de l'expérience, les problèmes majeurs rencontrés avec le dictionnaire bilingue *DIF* ont été essentiellement dus au découpage excessif des sens, à une prédominance d'indicateurs de sens sous forme de collocations[8] plutôt que de synonymes pour distinguer les sens des adjectifs[9], ainsi qu'à une certaine incohérence dans le marquage de l'information. En revanche, les échecs dus à l'inexistence de certains sens ou à une description sémantique incorrecte ne sont imputables qu'à des lacunes / erreurs dans l'élaboration du lexique SIMPLE-CLIPS.

Avant de passer à l'évaluation de l'utilisation combinée des méthodes basées sur les cognats et sur les indicateurs de sens, nous présentons la troisième méthode, qui permet d'étendre le lexique existant avec de nouveaux mots construits (néologismes).

6. L'APPROCHE EXPLOITANT LES PROPRIÉTÉS DES RÈGLES MORPHOLOGIQUES DU FRANÇAIS

Cette troisième méthode est complémentaire des deux précédentes dans la mesure où elle est appliquée aux mots construits sans équivalents italiens dans le lexique source. Cette méthode exploite les résultats d'un analyseur morphologique du français, le système DériF.

Il existe dans la littérature des travaux consacrés spécialement à l'étude d'une règle de construction morphologique en français, et dont l'objectif est de faire apparaître les contraintes prototypiques, d'ordre sémantique, syntaxique ou phonologique, pesant sur les lexèmes mis en relation par la règle. L'avantage que retire le traitement automatique des résultats de ces études est immédiat :

8. Inexploitables pour la mise en correspondance avec les informations de SIMPLE-CLIPS. Les adjectifs sont en effet décrits en termes de typage ontologique complexe ainsi que de relation antonymique et synonymique, mais aucune information collocationnelle n'est fournie.
9. L'inadéquation des indicateurs de sens pour les adjectifs ne remet pas en cause notre approche. L'utilisation de ressources bilingues différentes permettrait de remédier à ce problème.

parmi les contraintes imposées par les règles, certaines servent à l'acquisition automatique d'informations lexicales et donc à l'optimisation du contenu du lexique SIMPLE-CLIPS.

6.1. Fonctionnement de DériF

DériF (Namer, à paraître) est un système de règles qui s'appliquent en chaîne sur des lemmes catégorisés. Chaque règle met en relation le lemme qu'il analyse avec le lemme qui constitue sa base morphologique (quand il s'agit de dérivation). La règle engendre également le sens du lemme analysé, en définissant celui-ci par rapport à la base, dont DériF aura calculé la forme et la catégorie. À son tour, la base sert d'entrée potentielle pour une autre règle d'analyse, et ainsi de suite, tant que le résultat produit est un lexème analysable. Outre cet aspect fonctionnel, DériF présente d'autres particularités. Par exemple, si un lexème est constructionnellement ambigu (p. ex., IMPLANTABLE), alors DériF en donne toutes les analyses. D'autre part, une analyse par défaut est toujours proposée, ce qui rend compte de la régularité des néologismes construits. Enfin, outre la pseudo-définition octroyée au lexème construit lors de son analyse, DériF attribue d'autres informations lexicales au lexème analysé et à sa base. Ces informations reflètent les contraintes de la règle morphologique qui met en relation ces deux unités. Nous illustrons ci-dessous cette méthode d'acquisition par l'exemple de la formation des noms déverbaux en *-eur*.

6.2. Annotations sémantiques : les noms déverbaux en *-eur*

Les noms déverbaux en *-eur* sont habituellement décrits comme « agentifs » (Scalise, 1984, p. 228; Bisetto, 1996, p. 447). Plus précisément, l'interprétation agentive (CHANTEUR) ou instrumentale (INTERRUPTEUR) de ces noms est conditionnée par la présence obligatoire d'un **verbe de base agentif** (Busa, 1997, p. 41; Fradin, 2003, p. 49; Kerleroux, 2004, p. 155). Les très rares exceptions à cette règle (NAISSEUR, TRÉBUCHEUR) correspondent à une interprétation perfective ou causative du verbe de base (NAISSEUR = « celui qui fait naître ») (Fradin, 2003, p. 150). Le nom construit désigne, lui, une **entité concrète**, dont la nature inanimée (pour l'instrument) ou animée (pour l'agent) n'est pas prévisible.

L'exploitation de ces informations conduit DériF à distribuer trois séries de traits aux lexèmes reliés par la règle formatrice de noms déverbaux (tableau 5.4) :

1. la pseudo-définition du nom en fonction de la valeur du verbe V de base, suivant la glose « agent ou instrument pour V » ;
2. les contraintes portant sur la nature sémantique du nom en *–eur*, qui renvoie soit à un humain (hum=oui), soit à un artefact (naturel=non, anim=non, concret=oui), et sur la valeur aspectuelle du verbe de base (aspect=dynamique) ;

3. les relations prédicat-argument possibles entre le verbe et le nom (ce dernier, indicé par @2, s'identifie au syntagme nominal sujet agentif du verbe, d'indice @1).

Tableau 5.4 – **Annotation de CARBONISATEUR et CARBONISER**

1. carbonisateur/NOM: « (Agent habituel - Auteur exceptionnel - Instrument) de carboniser »
2. carboniser/VERBE: @1 [aspect = dynamique, sous_cat = <NPagent, ...>]
3. carbonisateur/NOM: @2 [concret = oui, hum = oui, comptable = oui] ||
@2 [concret = oui, anim = non, naturel = non, comptable = oui];
4. rel = NPagent(@2, sous_cat(@1))

Les traits ainsi calculés donnent lieu à des validations croisées, qui visent à en spécifier, compléter ou modérer le contenu. Par exemple, l'analyse de CARBONISER, qui conduit à l'identification du nom CARBONE, est complétée par l'indication suivant laquelle un verbe dénominal en *-iser* est majoritairement **transitif** et dénote un **accomplissement**. Cette information, plus précise que celle que prédit la règle d'analyse des noms en *-eur*, supplante le trait qui caractérise CARBONISER dans le tableau 5.4 par celui que reproduit le tableau 5.5 :

Tableau 5.5 – **Annotation de CARBONISER, après validations croisées**

carboniser/VERBE: @1 [aspect = accomplissement, sous_cat = <NPagent, NPpatient >]

6.3. De DériF à SIMPLE

Les informations fournies par DériF permettent de dériver semi-automatiquement des entrées sémantiques pour les noms construits en *-eur*, conformément au modèle lexical SIMPLE. Dans l'ontologie SIMPLE, ces noms appartiennent principalement à la hiérarchie des humains, et en particulier aux types sémantiques AGENT_OF_TEMPORARY_ACTIVITY (tueur, tuer ; promeneur, se promener), AGENT_OF_PERSITENT_ACTIVITY (skieur, skier ; buveur, boire), PROFESSION (vendeur, vendre) ou à celle des artefacts (type INSTRUMENT).

Prenons, à titre d'exemple, le nom « carbonisateur ». La génération de son entrée lexicale présuppose la sélection du *template* associé au type lexical approprié (*cf.* §2). À cette fin, l'information de DériF concernant la relation sémantique liant le nom et sa base de dérivation (tableau 5.4, ligne 1) est fondamentale. La structure argumentale de la base verbale (Tableau 5.5), l'argument que le nom lexicalise (Tableau 5.4, ligne 4), ainsi que les propriétés sémantiques de ce nom (ligne 3) vont aussi pouvoir être exploités. Dans le cas de l'exemple « carbonisateur », un processus de désambiguïsation (voir Namer *et al.*, 2009, 3.4.2) permet

d'abord d'écarter l'interprétation humaine au profit de l'artefactuelle. Les informations sous-spécifiées (Tableau 5.6) du *template* sélectionné, *i.e.* INSTRUMENT, sont ensuite substituées par les propriétés fournies par DériF (Tableau 5.7).

Tableaux 5.6 et 5.7 – **Template INSTRUMENT et son instanciation lexicale**

Usem	1
Template_Type	[INSTRUMENT]
Unification _path	[CONCRETE_ENTITY | ARTIFACT$_{Agentive}$ | TELIC]
Domain	General
Gloss	//free//
Predicative Representation	<Nil>
Selectional Restr.	<Nil>
Derivation	<Derivational relation>
Formal	isa (1, <instrument>or hyper)
Agentive	created_by (1, <manufacture [CREATION]>)
Constitutive	made_of (1, <Usem>)/opt/ has_as_part (1, <Usem>)opt/
Telic	used_for (1,<Usem:[EVENT]>)
Synonymy	Synonym (1, <Usem: [INSTRUMENT]>) /opt/
Reg. Polysemy	<Nil>

Usem:	**carbonisateur**
Temp._Type:	**[INSTRUMENT]**
Unif._path:	**[CONCRETE_ENTITY | ARTIFACT$_{Agentive}$ | TELIC]**
Domain:	General
Gloss:	==
Pred. Repres.:	**Pred_carboniser:Arg1agent, Arg2patient carbonisateur: agent_nominalisation**
Selectional Restr.:	==
Derivation:	**deverbalNounVerb<carbonisateur, carboniser>**
Formal:	isa **(carbonisateur, instrument / appareil /…)**
Agentive:	created_by **(carbonisateur, <fabriquer [CREATION]>)**
Constitutive:	==
Telic:	used_for **(carbonisateur, < carboniser [EVENT]>)**
Synonymy:	==
Reg. Polysemy	<Nil>

Le processus de migration des informations fournies par DériF permet de générer une entrée sémantique conforme au modèle SIMPLE, qui pourra ensuite être finalisée par l'insertion manuelle des restrictions de sélection sur les arguments du prédicat ainsi que par toute information optionnelle.

7. ÉVALUATION PARTIELLE

En comparant les deux premières méthodes, il apparaît clairement que celle basée sur les cognats est la plus simple à mettre en œuvre et qu'elle procure un plus grand taux de rappel pour le nombre d'entrées SIMPLE-CLIPS qui sont liées à un mot français. Elle reste, en revanche, beaucoup moins précise que la méthode basée sur les indicateurs de sens, étant donné que la seule inférence qui

peut être faite est que le mot FR ayant le suffixe correspondant partage toutes les entrées de SIMPLE-CLIPS du mot italien. Par exemple, « rigidité » partage tous les sens de SIMPLE-CLIPS de *rigidità*. Si l'un des sens dans le dictionnaire possède des synonymes (comme dans *rigidità* (1) « rigidité, dureté » (2) « rigidité, raideur »), ceux-ci ne seront pas liés à leur entrée SIMPLE-CLIPS.

En revanche, l'approche basée sur les indicateurs de sens est beaucoup plus coûteuse et procure un taux de rappel beaucoup plus faible, mais elle permet de relier différents équivalents français à un mot italien ou à une entrée SIMPLE-CLIPS. *A priori*, tous les équivalents de traduction peuvent être reliés. Le tableau 5.8 présente les entrées SIMPLE-CLIPS concernant les noms suffixés en *-aggio*, *-tà* et *-zione* qui ont pu être reliés à au moins un équivalent français, en utilisant l'approche hybride, combinant les deux premières méthodes décrites.

Tableau 5.8 – **Les mots construits traités par les deux premières méthodes**

Approche basée sur les cognats		**Approche basée sur les indicateurs de sens**	
Correct	**Incorrect**	**Correct**	**Incorrect**
82,54 %	0,02 %	11,71 %	5,73 %

Ces résultats appellent quelques commentaires. D'une part, la complémentarité de ces deux méthodes est frappante : parmi les 17,46 % des mots auxquels la méthode basée sur les cognats ne s'applique pas, 11,71 % sont gérés de manière adéquate par la méthode des indicateurs de sens. Le taux de succès pour les mots suffixés étudiés est donc 94,25 %. D'autre part, l'attention doit être également portée sur la fiabilité de l'approche fondée sur les cognats. Si l'on considère, comme Dubois et Dubois (1971, p. 138), que « … le total des mots suffixés forme 68,2 % des termes enregistrés dans la lettre A du *Petit Larousse* », qu'ils pourraient être potentiellement gérés par les deux méthodes, ces résultats sont extrêmement encourageants. Concernant les mots non construits, le taux de succès global (règles d'équivalence et règles par défaut) s'élève à 69 %. À propos de ce taux, il est opportun de souligner deux points. Le premier est qu'il ne se réfère pas à l'analyse d'un échantillon aléatoire de données, mais plutôt à un ensemble d'unités lexicales hautement polysémiques, et donc particulièrement problématiques. Le second est que l'algorithme dépendant fortement des informations fournies par les indicateurs de sens, le taux de succès pourrait être accru si l'on recueillait dans différentes sources bilingues des indicateurs ayant un contenu informationnel plus riche.

Concernant la troisième méthode, sa récente introduction ne permet pas, à ce jour, de fournir des résultats chiffrés. Il apparaît clairement que la sortie de l'analyse morphologique fournit des informations aussi bien pertinentes qu'exhaustives et directement exploitables pour la génération d'entrées de noms construits. Reste à quantifier son apport réel par rapport aux deux autres stratégies, dans le cadre d'une application concrète. Les informations très riches et systématiques fournies par DériF pourraient même être utilisées pour affiner dans SIMPLE les entrées verbales et adjectivales et en particulier la structure des qualia, moins complète que pour les noms

8. CONCLUSION

Dans cet article, nous avons proposé différentes stratégies visant à transférer des connaissances lexicales entre les langues, en développant notamment des nouvelles ressources lexicales à partir de l'existant. À cette fin, une étude pilote a été menée pour étudier les possibilités d'induire un lexique sémantique du français à partir d'un lexique italien multi-niveau. Il est clair que nous ne prétendons pas ici à la complétude du lexique développé par ces stratégies de génération d'entrées. De plus, aucune estimation précise ne peut être faite, pour l'heure, en matière d'effort humain nécessaire pour dériver la nouvelle ressource. Nous sommes cependant convaincus que, comparé au long et complexe processus de construction d'une ressource lexicale, une telle initiative se traduirait par une importante économie en termes de temps et d'effort. La méthodologie combinatoire que nous proposons permettrait d'acquérir rapidement et de façon économique un noyau substantiel d'entrées lexicales, qui seraient ensuite soumises à une phase manuelle de vérification et, éventuellement, de complément. Les résultats d'une telle implémentation représenteraient donc un bénéfice important en matière de coût et de temps, outre l'enjeu majeur concernant la cohérence des informations.

Les approches proposées sont en outre très prometteuses pour d'autres champs d'application. La stratégie basée sur les cognats est adéquate pour tous les couples de langues qui partagent des similarités en termes de formation des mots et d'inventaire lexical. De plus, elle semble particulièrement appropriée pour les langues de spécialité, où les cognats sont très nombreux (Namer, 2007). La stratégie des indicateurs de sens, pour laquelle la similitude entre les langues n'est pas pertinente, est en revanche applicable à tous les couples de langues pour lesquels il existe des ressources dictionnairiques bilingues de qualité. En outre, la méthode récemment intégrée visant à l'acquisition d'information lexicale à partir de l'analyse morphologique des mots construits permet d'envisager un codage rigoureux et économique des néologismes, assurant ainsi une actualisation constante du lexique

Selon nous, l'utilisation de ces stratégies complémentaires pourrait contribuer de manière significative à la production rapide, économique et cependant soignée de lexiques pour les nouvelles langues de l'Union européenne, qui souffrent d'un grand manque de ressources langagières.

RÉFÉRENCES

ATKINS, B.T.S. et P. BOUILLON (2004). « Relevance in Dictionary-Making : Sense Indicators in the Bilingual Entry », dans *Actes del I symposi um Internacional de Lexicografia*, Barcelona, Institut Universitari de Lingüística Aplicada (IULA), p. 39-52.

BISETTO, A. (1996). « Il suffisso *-tore* », *Quaderni patavini di linguistica*, vol. 14, p. 39-71.

BUSA, F. (1997). « The Semantics of Agentive Nominals in the Generative Lexicon », dans P. Saint-Dizier (dir.), *Predicative Forms in Natural Language*, Amsterdam, Kluwer.

DIF IL DIZIONARIO FRANCESE-ITALIANO ITALIANO-FRANCESE (2000). Torino, Paravia-Hachette.

DUBOIS, J. et C. DUBOIS (1971). *Introduction à la lexicographie : le dictionnaire*, Paris, Librairie Larousse.

FRADIN, B. (2003). *Nouvelles approches en morphologie*, Paris, Presses universitaires de France.

FUJITA, S. et F. BOND (2004). « An Automatic Method of Creating Valency Entries Using Plain Bilingual Dictionaries », *TMI-04*, Baltimore, p. 55-64.

HONG, M. *et al.* (2004). « Semi-Automatic Construction of Korean-Chinese Verb Patterns Based on Translation Equivalency », *Proceedings of Workshop on Multilingual Linguistic Resources, MLR2004*, Genève, Coling 2004, p. 80-85.

KERLEROUX, F. (2004). « Sur quels objets portent les opérations morphologiques de construction ? », *Lexique*, p. 85-124.

LENCI, A. *et al.* (2000a). « SIMPLE : A General Framework for the Development of Multilingual Lexicons », *International Journal of Lexicography*, special issue, Dictionaries, Thesauri and Lexical-Semantic Relations, vol. 13, n° 4, p. 249-263.

LENCI, A. *et al.* (2000b). *SIMPLE Linguistic Specifications. LE-SIMPLE (LE4-8346), Deliv. D2.1 & D2.2*, Pise, ILC-CNR et University of Pisa, p. 404.

NAMER, F. (2001). « Génération automatique de néologismes bilingues morphologiquement construits en français et en italien », Tours, TALN 2001, p. 281-296.

NAMER, F. (2007). « Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue », *Traitement automatique des langues,* n° 46, p. 157-181.

NAMER, F. *et al.* (2009). « Morphology-based Enhancement of a French SIMPLE Lexicon », soumis à *GL2009, 5th International Conference on Generative Approaches to the Lexicon*, qui s'est tenue à Pise, 17-19 septembre 2009.

NAMER, F. (à paraître). *Morphologie, lexique et traitement automatique du langage : l'analyseur DériF : TIC et Sciences cognitives*, Londres, Hermes Sciences Publishing.

PUSTEJOVSKY, J. (1995). *The Generative Lexicon*, Cambridge, The MIT Press.

PUSTEJOVSKY, J. (2001). « Type Construction and the Logic of Concepts », dans P. Bouillon et F. Busa (dir.), *The Language of Word Meaning*, Cambridge, Cambridge University Press, p. 91-123.

PUSTEJOVSKY, J. et B. BOGURAEV (1993). « Lexical Knowledge Representation and Natural Language Processing », *Artificial Intelligence*, n° 63, p. 193-223.

ROBERT, et C. SIGNORELLI (2003). *Dizionario francese-italiano italiano-francese*, Paris, LeRobert.

RUIMY, N.P., P. BOUILLON et B. CARTONI. (2005). « Inferring a Semantically Annotated Generative French Lexicon from an Italian Lexical Resource », dans P. Bouillon, K. Kanzaki (dir.), *GL 2005, Third International Workshop on Generative Approaches to the Lexicon*, Genève, p. 218-226.

RUIMY, N. *et al.* (1998). « The European LE-PAROLE Project : The Italian Syntactic Lexicon », *First International Conference on Language Resources and Evaluation Proceedings*, vol. I, Granada, p. 241-248.

RUIMY, N. *et al.* (2002), CLIPS, « A Multi-level Italian Computational Lexicon », *Third International Conference on Language Resources and Evaluation Proceedings*, vol. III, Las Palmas de Gran Canaria, p. 792-799.

SCALISE, S. (1994). *Morfologia*, Bologne, Il Mulino.

SCHULZ, S. *et al.* (2004). « Cognate Mapping – A Heuristic Strategy for the Semi-Supervised Acquisition of a Spanish Lexicon from a Portuguese Seed Lexicon », *Coling 2004*, Genève, p. 813-819.

CHAPITRE 6

MÉTHODES ET CRITÈRES D'ÉVALUATION EN RÉSUMÉ AUTOMATIQUE
Bilan et perspectives

Marie-Josée Goulet

RÉSUMÉ

L'étude présentée dans ce texte a pour principal objectif de décrire les méthodes et les critères d'évaluation utilisés en résumé automatique. Bien que Test Analysis Conference se soit imposé comme modèle d'évaluation, du moins au sein des chercheurs et des développeurs participant à cette campagne, les méthodes et les critères utilisés par ses adeptes ne représentent qu'une partie des possibilités. Par ailleurs, les descriptions des méthodes et des critères d'évaluation dans la littérature sur le résumé automatique ne sont pas aussi détaillées qu'on

le souhaiterait. L'analyse des évaluations de notre corpus a permis de dégager trois méthodes générales et 25 critères d'évaluation des résumés automatiques. En outre, le texte rapporte plusieurs difficultés rencontrées dans l'évaluation des résumés automatiques, comme le désaccord parmi les résumeurs humains dans l'extraction de phrases importantes dans des textes, l'importance de recourir à des systèmes équivalents pour la production de résumés automatiques qui seront comparés et la nécessité de définir de manière précise les critères d'évaluation.

1. INTRODUCTION

L'étude présentée dans ce texte a pour principal objectif de décrire les méthodes et les critères d'évaluation utilisés en résumé automatique. Une évaluation « vise à déterminer un indice de satisfaction, c'est-à-dire à déterminer à quoi quelque chose est bon pour quelqu'un » (Chaudiron, 2004, p. 19). Les résultats des évaluations sont d'un intérêt capital pour les investisseurs, les chercheurs, les développeurs et les utilisateurs d'un système (Hirschman et Mani, 2003). De façon générale, ce sont les chercheurs et les développeurs d'un système qui s'occupent des évaluations. Dans ce texte, le terme *évaluateur* sera employé pour désigner le responsable d'une évaluation.

2. PROBLÉMATIQUE ET MÉTHODOLOGIE

Chaque année depuis 2001, l'institut américain NIST (National Institute of Standards and Technology) organise une campagne d'évaluation internationale pour les résumés automatiques, intitulée *Text Analysis Conference* (TAC, 2009). L'approche de NIST est essentiellement quantitative. L'approche quantitative permet d'automatiser le processus d'évaluation, ce qui représente certainement un avantage non négligeable. L'automatisation de l'évaluation est censée diminuer la subjectivité résultant de jugements humains. De façon générale, l'approche quantitative ne prend pas en considération la satisfaction des utilisateurs du système (Chaudiron, 2004), ce qui semble paradoxal puisque les futurs utilisateurs d'un système sont sans doute les mieux placés pour évaluer ce système, par exemple en fonction de leurs besoins et de contextes d'utilisation variés. Bien que TAC se soit imposé comme modèle d'évaluation, du moins au sein des chercheurs et des développeurs participant à cette campagne, les méthodes et les critères utilisés par ses adeptes ne représentent qu'une partie des possibilités.

Dans son ouvrage sur le résumé automatique, Mani (2001) propose une typologie générale des méthodes d'évaluation. L'auteur oppose par exemple l'évaluation opaque à l'évaluation transparente, l'évaluation intrinsèque à l'évaluation extrinsèque, l'évaluation avec des juges à l'évaluation automatique. Les descriptions des méthodes d'évaluation dans Mani (2001) ne sont pas aussi détaillées qu'on le souhaiterait. Par ailleurs, l'auteur ne recense pas de manière exhaustive les critères d'évaluation utilisés en résumé automatique.

La typologie des méthodes et des critères d'évaluation présentée dans cette étude s'appuie sur l'analyse de 481 données expérimentales, ce qui représente la plus vaste recherche empirique sur l'évaluation de résumés automatiques. Les données analysées proviennent de 27 évaluations de résumés automatiques, menées entre 1961 et 2005 (Rath *et al.*, 1961 ; Edmundson, 1969 ; Morris *et al.*, 1992 ; Brandow *et al.*, 1995 ; Kupiec *et al.*, 1995 ; Minel *et al.*, 1997 ; Salton *et al.*, 1997 ; Klavans *et al.*, 1998 ; Aone *et al.*, 1999 ; Barzilay et Elhadad, 1999 ; Hovy et Lin, 1999 ; Mani et Bloedorn, 1999 ; Marcu, 1999 ; Maybury, 1999 ; Myaeng et Jang, 1999 ; Teufel et Moens, 1999 ; Donaway *et al.*, 2000 ; Nanba et Okumura, 2000 ; Nadeau, 2002 ; Saggion et Lapalme, 2002 ; Châar *et al.*, 2004 ; Farzindar et Lapalme, 2005). Chaque donnée expérimentale a été classifiée dans l'une des 25 catégories, par exemple le type de texte source, le nombre de résumés automatiques produits par texte source, les directives données aux résumeurs humains et le profil des juges.

3. CLASSIFICATION DES MÉTHODES D'ÉVALUATION

Trois méthodes générales d'évaluation ont été recensées dans la littérature :

1. l'évaluation des résumés automatiques sans comparaison ;
2. la comparaison des résumés automatiques avec les textes sources ;
3. la comparaison des résumés automatiques avec d'autres résumés.

Plusieurs variantes de la troisième méthode sont utilisées dans le corpus à l'étude. Dans cette section, nous décrivons ces méthodes.

3.1. Comparaison avec des résumés humains

Les résumés automatiques peuvent être comparés avec des résumés produits par des humains. Ces derniers sont soit des résumés auteurs, soit des résumés professionnels, soit des nouveaux résumés humains. Un résumé auteur est rédigé par l'auteur du texte source, comme on en trouve dans les revues scientifiques notamment. Un résumé professionnel est rédigé par un rédacteur professionnel, dans un service de documentation par exemple. Dans la plupart des cas, les résumés auteurs et professionnels utilisés dans une évaluation existaient avant celle-ci. Quant au nouveau résumé humain, il est produit spécifiquement pour les besoins d'une évaluation. Les nouveaux résumés humains peuvent être produits par extraction ou par reformulation. Pour les résumés par extraction, des humains sélectionnent les éléments importants dans les textes sources. Le plus souvent, les éléments à extraire sont des phrases, mais il peut s'agir de mots clés ou de paragraphes. Pour les résumés par reformulation, comme le nom l'indique, des humains reformulent le contenu des textes sources pour en faire des textes plus concis, ce qui correspond à la méthode traditionnelle de résumé de texte.

L'utilisation de résumés humains dans une évaluation requiert certaines précautions. L'évaluateur doit s'assurer que les résumés humains sont comparables avec les résumés automatiques évalués. Par exemple, si les résumés évalués ont été produits par extraction de phrases, les résumés humains devraient eux aussi être produits par extraction de phrases. Parmi les évaluations de notre corpus, quelques-unes ne respectent pas ce principe. Afin d'illustrer cette affirmation, prenons l'évaluation décrite dans Kupiec *et al.* (1995). Dans cette évaluation, des résumés automatiques produits par extraction de phrases ont été comparés avec des résumés professionnels produits par reformulation. Comme les deux types de résumés ne pouvaient être comparés directement, les résumés professionnels ont été analysés et les phrases de ces résumés pouvant être alignées avec des phrases des textes sources, moyennant des modifications mineures ont été extraites. Cet ensemble de phrases alignées a ensuite servi de base pour la comparaison avec les résumés automatiques. Bref, les résumés professionnels ont été transformés afin de pouvoir être comparés avec des résumés par extraction.

L'utilisation de résumés professionnels dans une évaluation de résumés automatiques peut sembler intéressante. Ces résumés sont sans doute de bonne qualité puisqu'ils ont été rédigés par des professionnels. Par ailleurs, l'évaluateur bénéficie d'une économie de temps et d'argent en recourant à des résumés professionnels déjà rédigés. Toutefois, lors de la « transformation » de ces résumés, n'ajoute-t-on pas une part de subjectivité ? L'alignement des phrases des résumés professionnels avec des phrases des textes sources constitue une tâche subjective, dans une certaine mesure, à moins bien sûr que les phrases alignées soient identiques.

Dans la plupart des évaluations analysées lors de notre étude, les évaluateurs ont opté pour l'utilisation de nouveaux résumés humains, plus précisément des résumés humains produits par extraction de phrases. Cette méthode permet aux évaluateurs de tenir compte de leurs besoins dans la préparation des résumés humains, ce qui représente un avantage certain. Par exemple, les résumés peuvent être produits en fonction d'un thème précis. Du côté des inconvénients, mentionnons l'incertitude face à la fiabilité des résumeurs humains pour l'extraction de phrases. Cette incertitude découle de l'observation suivante : les personnes produisant les résumés humains dans le contexte d'une évaluation n'extraient pas nécessairement toutes les mêmes phrases des textes sources. Dans l'étude de Rath *et al.* (1961), six personnes devaient sélectionner les 20 phrases les plus représentatives du contenu d'un texte et les classer en ordre d'importance. En tout, 10 textes ont été ainsi analysés.

3.1.1. Accord parmi les résumeurs humains

L'accord parmi les résumeurs humains pour la sélection des phrases est peu élevé dans Rath *et al.* (1961), soit 1,6 phrase sur 20 par texte en moyenne. L'accord augmente à 6,4 phrases sur 20 si l'on ne considère que les phrases sélectionnées

par cinq résumeurs sur six. S'appuyant sur ces résultats, les auteurs de l'étude concluent que les humains ne sont pas fiables pour extraire les phrases importantes des textes, car leurs choix divergent beaucoup. Il faut néanmoins admettre que le nombre de phrases communes est considérablement plus élevé en ne considérant que cinq résumeurs sur six, ce qui pourrait suggérer qu'un des résumeurs (ou plusieurs) a fait diminuer l'accord dans cette expérience.

Dans la deuxième partie de leur étude, Rath *et al.* (1961) renchérissent sur la non-fiabilité des humains pour la production des résumés par extraction de phrases. Cinq étudiants ont extrait vingt phrases dans six articles scientifiques. Huit semaines plus tard, les mêmes personnes ont répété l'exercice à partir des mêmes textes. Les résultats de cette expérience sont sans équivoque : en moyenne, les résumeurs n'ont sélectionné les mêmes phrases que dans 55 % des cas. L'expérience a aussi montré que les résumeurs ne se souvenaient des phrases qu'ils avaient sélectionnées la première fois que dans 42,5 % des cas.

Ces résultats ne sont guère étonnants. Huit semaines est une période de temps considérablement longue. Il apparaît donc normal que les résumeurs ne se soient pas rappelé les phrases qu'ils avaient sélectionnées la première fois. En outre, il semblerait que les résumeurs aient été encouragés à ne pas extraire les mêmes phrases : « *they* [en parlant des résumeurs humains] *were instructed not to attempt to select sentences which they had selected previously* […] » (Rath *et al.*, 1961, p. 290). Ainsi, les résumeurs ont tout probablement refait l'exercice en neuf, ce qui expliquerait qu'ils n'aient pas sélectionné exactement les mêmes phrases que la première fois. Qui plus est, cette consigne leur a peut-être fait croire qu'ils n'avaient pas bien effectué l'exercice la première fois, les incitant à sélectionner des phrases différentes. Enfin, il faut admettre que 20 est un nombre élevé de phrases à extraire et à ordonner, peu importe la longueur des textes. Malheureusement, la longueur des textes n'est pas fournie dans Rath *et al.* (1961).

Salton *et al.* (1997) rapportent eux aussi un faible accord parmi les résumeurs humains. Dans leur étude, les résumeurs ont extrait les paragraphes les plus importants des textes. Le recouvrement entre deux résumés humains, c'est-à-dire le nombre de paragraphes communs, est en moyenne de 46 %. Selon Salton *et al.* (1997, p. 354), ces résultats signifient que : « *an extract generated by one person is likely to cover 46 % of the information that is regarded as most important by another person* ». Cette interprétation est discutable. On pourrait en effet arguer que certaines informations se répètent d'un paragraphe à l'autre d'un texte. Dans ce cas, ne serait-il pas souhaitable de calculer l'accord parmi les résumeurs à partir des informations contenues dans les paragraphes? Par ailleurs, bien que neuf résumeurs aient participé à cette étude, chaque texte n'a été analysé que par deux personnes.

D'autres évaluateurs ont obtenu des résultats diamétralement opposés à ceux rapportés dans Rath *et al.* (1961) et Salton *et al.* (1997) pour l'accord parmi les résumeurs humains. Par exemple dans Barzilay et Elhadad (1999), l'accord

est de 96 % pour les résumés représentant 10 % de la longueur des textes sources et de 90 % pour ceux représentant 20 %, ce qui est très élevé. Dans cette évaluation, l'accord parmi les résumeurs a été calculé en fonction du « *ratio of observed agreements with the majority opinion to possible agreements with the majority opinion* » (Barzilay et Elhadad, 1999, p. 118), calcul désigné en anglais par l'expression *percent agreement*. Le pourcentage calculé correspond au nombre de fois qu'une personne est d'accord avec la majorité, autant pour la sélection d'une phrase importante que pour l'exclusion d'une phrase.

Comme les évaluateurs n'utilisent pas tous la même méthode pour calculer l'accord parmi les résumeurs, les comparaisons présentées ci-dessus ne permettent pas de généraliser. Notre analyse fait néanmoins ressortir quelques facteurs qui influenceraient l'accord parmi les résumeurs dans la sélection de phrases importantes, notamment le type de textes sources, les consignes données aux résumeurs et le profil de ceux-ci. Le profil fait référence au niveau de scolarité et aux connaissances personnelles des résumeurs. Advenant le cas où les résumeurs participant à une évaluation ont des profils différents (niveaux de scolarité non comparables ou connaissances jugées non équivalentes), on peut s'attendre à ce que leur opinion diverge davantage quant à l'importance des phrases. De même, le fait de fournir des consignes vagues aux résumeurs favorise la multiplication des interprétations quant à la nature de l'exercice à effectuer. Dans une situation où diverses interprétations d'une consigne cohabitent, des désaccords importants peuvent être observés parmi les résumeurs. Par ailleurs, si les résumeurs ne sont pas familiarisés avec le type de texte source, l'exercice présentera une difficulté supplémentaire, ce qui est susceptible de réduire l'accord parmi les résumeurs. Bref, la prise en compte de ces facteurs peut influencer l'accord parmi les résumeurs et, par conséquent, les résultats de toute l'évaluation. Afin de maximiser la fiabilité des résumés humains, des évaluateurs ont appliqué la méthode de la majorité. Par exemple, Klavans *et al.* (1998) n'ont conservé que les phrases considérées comme importantes par au moins deux résumeurs sur trois. De même, Barzilay et Elhadad (1999) n'ont conservé que les phrases jugées importantes par au moins trois résumeurs sur cinq, et Myaeng et Jang (1999) que les phrases estimées importantes par les deux résumeurs ayant participé à l'évaluation.

3.2. Comparaison avec d'autres résumés automatiques

Au cours d'une évaluation, les résumés automatiques peuvent être comparés avec d'autres résumés automatiques. Ces derniers sont de deux types : des résumés produits par d'autres systèmes ou des résumés de contrôle. Les résumés de contrôle sont produits en extrayant automatiquement des phrases dans les textes sources, soit aléatoirement, soit en tenant compte de l'emplacement des phrases dans ces textes. Les résultats d'une comparaison entre des résumés automatiques et des résumés de contrôle produits par extraction aléatoire peuvent certainement fournir

un indice de la qualité des résumés automatiques. Toutefois, les conclusions que ces résultats permettront de tirer seront limitées, la qualité des résumés de contrôle utilisés dans ce cas étant douteuse. Les résumés de contrôle produits en extrayant des phrases dont l'emplacement est jugé stratégique sont de meilleure qualité que ceux produits en extrayant des phrases aléatoirement. En effet, les chercheurs en résumé automatique reconnaissent que les phrases se trouvant dans l'introduction, dans la conclusion, au début et à la fin des paragraphes, entre autres, sont généralement importantes. Il apparaît donc plus intéressant d'utiliser ce type de résumé de contrôle. Par ailleurs, d'autres critères d'extraction pourraient être envisagés pour la production des résumés de contrôle.

Plusieurs évaluateurs ont quant à eux privilégié l'utilisation de résumés automatiques produits par d'autres systèmes (Brandow *et al.*, 1995 ; Barzilay et Elhadad, 1999 ; Marcu, 1999 ; Saggion et Lapalme, 2002 ; Farzindar et Lapalme, 2005). Cette méthode d'évaluation permet de mesurer la qualité des résumés automatiques produits par différents systèmes, comme dans les campagnes d'évaluation. Les systèmes évalués doivent être équivalents, ce qui ne semble pas toujours être le cas dans les évaluations de notre corpus. Prenons comme exemple deux systèmes, le premier conçu pour produire des résumés de textes techniques d'un domaine précis et le second conçu pour produire des résumés de textes généraux comme *Word 2003*. Si les résumés produits par le premier système étaient jugés de meilleure qualité que ceux produits par le second, nous pourrions conclure que, pour la production de résumés de textes spécialisés, le système conçu explicitement pour produire des résumés de textes spécialisés est plus performant qu'un système conçu pour produire des résumés de textes généraux. Toutefois, les systèmes décrits dans cet exemple n'étant pas équivalents, la comparaison des résumés ne permettait pas de tirer davantage de conclusions.

En terminant, la combinaison de méthodes d'évaluation est souhaitable. L'analyse des évaluations de notre corpus révèle que des évaluateurs ont en effet combiné plusieurs méthodes d'évaluation dans une même expérience (Brandow *et al.*, 1995 ; Hovy et Lin, 1999 ; Saggion et Lapalme, 2002). Dans les cas où résumés humains et résumés de contrôle sont utilisés en parallèle, il serait pertinent de comparer ces résumés entre eux. Cette comparaison permettrait de vérifier si les résumés humains sont de meilleure qualité que les résumés de contrôle.

4. CRITÈRES D'ÉVALUATION DES RÉSUMÉS AUTOMATIQUES

Un critère d'évaluation est une mesure permettant de porter un jugement d'appréciation sur un aspect précis d'un résumé automatique. L'analyse des évaluations de notre corpus a permis de dégager 25 critères pour évaluer la qualité des résumés automatiques. Les critères utilisés se rapportent soit à la qualité du contenu des résumés, soit à la qualité de la langue. Les 18 critères se rapportant au contenu des résumés, présentés ci-dessous, ont été classifiés selon la méthode d'évaluation.

1. Évaluation des résumés automatiques sans comparaison :
   – cohérence des résumés ;
   – concision des résumés ;
   – intelligibilité des résumés ;
   – pertinence des phrases des résumés ;
   – progression logique des informations dans les résumés ;
   – sujet clairement indiqué dans les résumés.
2. Comparaison des résumés automatiques avec les textes sources :
   – acceptabilité des résumés ;
   – présence des idées essentielles des textes sources dans les résumés ;
   – présence des sujets des textes sources dans les résumés ;
   – respect de l'enchaînement des idées dans les résumés ;
   – utilité des résumés dans une tâche donnée ;
   – temps requis pour effectuer une tâche donnée.
3. Comparaison des résumés automatiques avec d'autres résumés :
   – choix du meilleur résumé parmi plusieurs résumés (par exemple le résumé évalué, un résumé humain et un résumé de contrôle) ;
   – précision, la plupart du temps à partir des phrases ;
   – présence de groupes de mots des résumés humains dans les résumés évalués (ROUGE) ;
   – rappel, la plupart du temps à partir des phrases ;
   – recouvrement des informations contenues dans les résumés évalués et dans les autres résumés ;
   – ressemblance des résumés évalués avec les autres résumés.

Les sept critères d'évaluation se rapportant à la langue sont les suivants :

– acronymes définis ;
– clarté des résumés ;
– grammaire ;
– lisibilité des résumés ;
– nombre d'anaphores sans référent dans les résumés ;
– nombre de ruptures argumentatives dans les résumés ;
– orthographe.

Cette présentation sommaire fait ressortir la diversité des critères d'évaluation utilisés en résumé automatique. Ces derniers varient d'une part selon la méthode d'évaluation préconisée et d'autre part selon le type de résumé évalué. Par exemple, certains critères pour la langue ne s'appliquent qu'à des résumés produits par reformulation (génération de texte) tandis que d'autres s'appliquent aux deux types de résumés (extraction et reformulation). Ci-dessous, les critères pour évaluer la qualité de la langue sont présentés selon le type de résumé évalué.

Critères pouvant être utilisés pour les résumés produits par extraction ou par reformulation :

- acronymes définis ;
- nombre d'anaphores sans référent dans les résumés ;
- clarté des résumés ;
- lisibilité des résumés ;
- nombre de ruptures argumentatives dans les résumés.

Critères généralement réservés aux résumés par reformulation :

- grammaire ;
- orthographe.

Parmi les critères utilisés dans la littérature, seules la grammaire et l'orthographe sont réservées aux résumés produits par reformulation. Ces derniers constituent de nouveaux textes générés automatiquement. Ils peuvent donc contenir des erreurs grammaticales ou orthographiques. Les résumés produits par extraction, quant à eux, sont composés d'unités extraites de textes réputés bien écrits sur les plans grammatical et orthographique. Si des erreurs sont présentes dans les textes sources, celles-ci ne devraient pas être attribuées aux résumés automatiques.

L'analyse fait également ressortir le caractère subjectif de certains critères d'évaluation. Parmi les 25 critères d'évaluation recensés, 19 sont plutôt subjectifs et 6 sont plutôt objectifs.

Critères d'évaluation plutôt objectifs :

- nombre d'anaphores sans référent dans les résumés ;
- nombre de ruptures argumentatives dans les résumés ;
- précision ;
- présence de groupes de mots des résumés humains dans les résumés évalués (ROUGE) ;
- rappel ;
- temps requis pour effectuer une tâche donnée.

Critères d'évaluation plutôt subjectifs :

- acceptabilité des résumés ;
- acronymes définis (échelle de valeur) ;
- choix du meilleur résumé parmi plusieurs résumés ;
- clarté des résumés ;
- cohérence des résumés ;
- concision des résumés ;
- grammaire ;
- intelligibilité des résumés ;

- lisibilité des résumés ;
- orthographe ;
- pertinence des phrases des résumés ;
- présence des idées essentielles des textes sources dans les résumés ;
- présence des sujets des textes sources dans les résumés ;
- progression logique des informations dans les résumés ;
- recouvrement des informations contenues dans les résumés évalués et dans les autres résumés ;
- respect de l'enchaînement des idées dans les résumés ;
- ressemblance des résumés évalués avec les autres résumés ;
- sujet clairement indiqué dans les résumés ;
- utilité des résumés dans une tâche donnée.

Les critères faisant précisément référence à un nombre ou à une opération mathématique peuvent être considérés comme étant plutôt objectifs. Pour exemple, prenons le critère du rappel utilisé dans une évaluation où les résumés automatiques sont comparés avec des résumés humains, les deux types de résumés ayant été produits par extraction de phrases. Dans ce contexte, le rappel est une opération mathématique correspondant au nombre de phrases des résumés humains présentes dans les résumés automatiques évalués. Certains pourraient arguer que cette méthode d'évaluation est subjective, car la production des résumés humains utilisés pour la comparaison est elle-même subjective. Toutefois, la classification présentée ici s'appuie sur les critères d'évaluation et non sur les méthodes.

Les critères considérés comme étant plutôt subjectifs sont nombreux. Certains critères réfèrent à des aspects subjectifs en soi, par exemple la cohérence, l'intelligibilité et la pertinence des phrases. D'autres critères ont été utilisés en appliquant une échelle de valeurs, ce qui les rend subjectifs. Par exemple, Saggion et Lapalme (2002) ont demandé à des juges d'évaluer la grammaire et l'orthographe des résumés automatiques, en donnant une note allant de 0 à 5 pour chacun de ces critères. Bien que les aspects grammatical et orthographique puissent sembler plutôt objectifs à première vue, l'utilisation d'une échelle de valeurs, au détriment d'un comptage du nombre d'erreurs, leur confère un caractère hautement subjectif. Ces critères pourraient être plus objectifs si on demandait aux juges de compter le nombre d'erreurs grammaticales et orthographiques dans les résumés automatiques, par exemple en les soulignant ou en les annotant à l'écran.

Les critères utilisés dans les évaluations de notre corpus ne sont pas tous clairement définis. Par exemple, la lisibilité est définie comme : « [l]a facilité de distinction ou de perception du contenu du résumé qui en facilite la compréhension. Ce critère donne une appréciation globale du résumé. On demande si le résumé est clair, assez clair, peu clair ou incompréhensible » (Farzindar et Lapalme, 2005, p. 190). Les deux premières phrases de cette définition semblent

contradictoires. En effet, dans la première phrase on laisse entendre que la lisibilité constitue un aspect du contenu d'un résumé tandis que dans la deuxième phrase on met l'accent sur la qualité globale du résumé. En outre, cette définition ne fait pas mention des données linguistiques traditionnellement associées à la lisibilité, par exemple la longueur moyenne des phrases, la complexité de la construction des phrases et le nombre de mots rares. Bref, nous croyons qu'une attention particulière doit être apportée à la définition des critères d'évaluation en résumé automatique.

5. CONCLUSION

Dans cet article, nous avons présenté une classification des méthodes et des critères d'évaluation utilisés en résumé automatique depuis 1961. La présentation, fondée sur l'analyse de 481 données expérimentales, a notamment tenu compte des différents types de résumés utilisés dans les évaluations et de la nature subjective de certains critères d'évaluation. En outre, le texte a rapporté plusieurs difficultés rencontrées dans l'évaluation des résumés automatiques, comme le désaccord parmi les résumeurs humains dans l'extraction de phrases importantes dans des textes, l'importance de recourir à des systèmes équivalents pour la production de résumés automatiques qui seront comparés et la nécessité de définir de manière précise les critères d'évaluation.

RÉFÉRENCES

AONE, C., M.E. OKOROWSKI, J. GORLINSKY et B. LARSEN (1999). « A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques », dans I. Mani et M.T. Maybury (dir.), *Advances in Automatic Text Summarization*, Cambridge, MIT Press, p. 71-80.

BARZILAY, R. et M. ELHADAD (1999). « Using Lexical Chains for Text Summarization », dans I. Mani et M.T. Maybury (dir.), *Advances in Automatic Text Summarization*, Cambridge, MIT Press, p. 111-121.

BRANDOW, R., K. MITZE et L. RAU (1995). « Automatic Condensation of Electronic Publications by Sentence Selection », *Information Processing Management*, vol. 31, nº 5, p. 675-685.

CHÂAR, S.L., O. FERRET et C. FLUHR (2004). « Filtrage pour la construction de résumés multidocuments guidée par un profil », *Traitement automatique des langues*, vol. 45, nº 1, p. 65-93.

CHAUDIRON, S. (dir.) (2004). *Évaluation des systèmes de traitement de l'information*, Paris, Hermès Science.

DONAWAY, R.L., K.W. DRUMMEY et L.A. MATHER (2000). « A Comparison of Rankings Produced by Summarization Evaluation Measures », dans *Proceedings of the Workshop on Automatic Summarization*, Seattle, p. 69-78.

EDMUNDSON, H.P. (1969). « New Methods in Automatic Abstracting », *Journal of the Association for Computing Machinery*, vol. 16, nº 2, p. 264-285.

FARZINDAR, A. et G. LAPALME (2005). « Production automatique de résumés de textes juridiques : évaluation de qualité et d'acceptabilité », dans *Actes de la conférence sur le traitement automatique des langues naturelles*, Dourdan, p. 183-192.

HIRSCHMAN, L. et I. MANI (2003). « Evaluation », dans R. Mitkov (dir.), *The Oxford Handbook of Computational Linguistics*, Oxford, Oxford University Press, p. 414-429.

HOVY, E. et C.-Y. LIN (1999). « Automated Text Summarization in SUMMARIST », dans I. Mani et M.T. Maybury (dir.), *Advances in Automatic Text Summarization*, Cambridge, MIT Press, p. 81-94.

KLAVANS, J.L. *et al.* (1998). « Resources for the Evaluation of Summarization Techniques », dans *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, Grenade, p. 899-902.

KUPIEC, J., J. PEDERSEN et F. CHEN (1995). « A Trainable Document Summarizer », dans *Proceedings of SIGIR (Special Interest Group on Information Retrieval)*, Seattle, p. 68-73.

MANI, I. (2001). *Automatic Summarization*, Amsterdam, John Benjamins Publishing Company.

MANI, I. et E. BLOEDORN (1999). « Summarizing Similarities and Differences among Related Documents », dans I. Mani et M.T. Maybury (dir.), *Advances in Automatic Text Summarization*, Cambridge, MIT Press, p. 357-379.

MANI, I. et M.T. MAYBURY (dir.) (1999). *Advances in Automatic Text Summarization*, Cambridge, MIT Press.

MARCU, D. (1999). « Discourse Trees Are Good Indicators of Importance in Text », dans I. Mani et M.T. Maybury (dir.), *Advances in Automatic Text Summarization*, Cambridge, MIT Press, p. 123-136.

MAYBURY, M. (1999). « Generating Summaries from Event Data », dans I. Mani et M.T. Maybury (dir.), *Advances in Automatic Text Summarization*, Cambridge, MIT Press, p. 265-281.

MINEL, J.-L., S. NUGIER et G. PIAT (1997). « How to Appreciate the Quality of Automatic Text Summarization ? Examples of FAN and MLUCE Protocols and Their Results on SERAPHIN », dans *Proceedings of Intelligent Scalable Text Summarization Workshop (EACL)*, Madrid, p. 25-31.

MORRIS, A., G. KASPER et D. ADAMS (1992). « The Effects and Limitations of Automatic Text Condensing on Reading Comprehension Performance », *Information Systems Research*, vol. 3, n° 1, p. 17-35.

MYAENG, S.H. et D.-H. JANG (1999). « Development and Evaluation of a Statistically-Based Document Summarization System », dans I. Mani et M.T. Maybury (dir.), *Advances in Automatic Text Summarization*, Cambridge, MIT Press, p. 61-70.

NADEAU, D. (2002). *Amélioration de résumés automatiques produits par extraction de phrases : étude de cas avec Extractor*, Mémoire de maîtrise, Québec, Université Laval.

NANBA, H. et M. OKUMURA (2000). « Producing More Readable Extracts by Revising Them », dans *Proceedings of the 18th International Conference on Computational Linguistics*, Saarbrucker, p. 1071-1075.

RATH, G.J., A. RESNICK et T.R. SAVAGE (1961). « The Formation of Abstracts by the Selection of Sentences », *American Documentation*, vol. 12, n° 2, p. 139-141.

SAGGION, H. et G. LAPALME (2002). « Generating Indicative-Informative Summaries with SumUM », *Computational Linguistics*, vol. 28, n° 4, p. 497-526.

SALTON, G., A. SINGHAL, M. MITRA et C. BUCKLEY (1997). « Automatic Text Structuring and Summarization », *Information Processing and Management*, vol. 33, n° 2, p. 193-207.

TAC (1999). Text Analysis Conference, <www.nist.gov/tac/>.

TEUFEL, S. et M. MOENS (1999). « Argumentative Classification of Extracted Sentences as a First Step Towards Flexible Abstracting », dans I. Mani et M.T. Maybury (dir.), *Advances in Automatic Text Summarization*, Cambridge, MIT Press, p. 155-171.

CHAPITRE 7

LES EXCEPTIONS DANS LES ONTOLOGIES[1]

Un modèle théorique pour déduire des propriétés à partir d'axiomes topologiques

Christophe Jouis, Julien Bourdaillet, Bassel Habib et Jean-Gabriel Ganascia

RÉSUMÉ

Ce chapitre est une contribution à l'étude des ontologies formelles. Il s'agit du problème des entités atypiques dans les ontologies. Nous proposons un nouveau modèle de représentation des connaissances en combinant les ontologies et la topologie générale. Pour représenter les entités atypiques dans les ontologies,

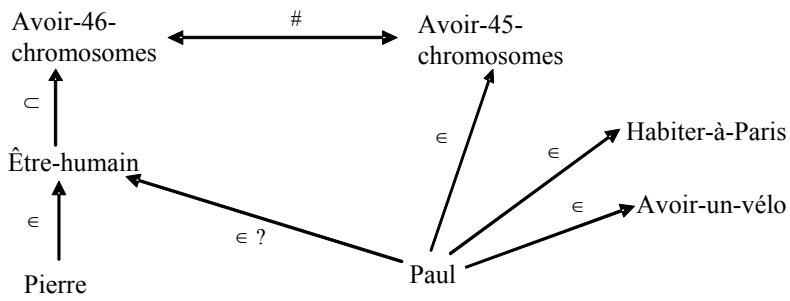
1. Les auteurs remercient Jie Liu pour sa contribution aux travaux présentés dans ce chapitre.

nous introduisons les quatre opérateurs d'intérieur, d'extérieur, de frontière et de fermeture. Ces opérateurs permettent de spécifier si une entité, appartenant à une classe, est typique ou non. Nous définissons un système de relations d'inclusion et d'appartenance topologique dans le formalisme des ontologies, en adaptant les quatre opérateurs topologiques à l'aide de leurs propriétés mathématiques. Ces propriétés sont utilisées comme un ensemble d'axiomes qui nous permettent de définir les inclusions et les relations d'appartenance topologiques. De plus, nous définissons des combinaisons des opérateurs intérieur, extérieur, frontière et fermeture qui nous permettent de construire une algèbre. Notre modèle est implémenté en AnsProlog, un langage récent de programmation logique qui permet d'introduire des prédicats négatifs dans les règles d'inférence.

1. INTRODUCTION

Certaines entités appartiennent plus ou moins à une classe. En particulier, certaines entités individuelles sont rattachées à une classe alors qu'elles ne vérifient pas toutes les propriétés de la classe. Pour illustrer ce phénomène, considérons le réseau ontologique ci-dessous (figure 7.1).

Figure 7.1 – **L'élément [Paul] ne vérifie pas toutes les propriétés de la classe [Être-humain]**



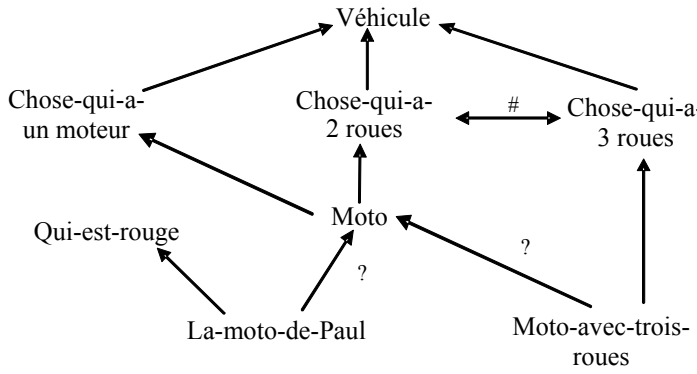
Ce réseau correspond aux énoncés déclaratifs suivants :

1. un être humain a 46 chromosomes ;
2. Pierre est un être humain ;
3. Paul est un être humain ;
4. Paul a 45 chromosomes ;
5. Paul habite à Paris ;
6. Paul a un vélo ;
7. une chose ne peut pas avoir en même temps 46 chromosomes et 45 chromosomes.

Parce que [Paul] est un [Être-humain], il hérite de toutes les propriétés typiques de [Être-humain], en particulier d'avoir 46 chromosomes. Un paradoxe est introduit par l'énoncé 4 parce que « Un être humain a 46 chromosomes » est un fait général mais pas un fait universel. L'énoncé 1 signifie « En général, les êtres humains ont 46 chromosomes mais il y a des exceptions à cette règle ».

Un phénomène similaire peut être observé avec les classes distributives. Certaines sous-classes sont rattachées plus ou moins à une classe plus générale parce que certains de leurs éléments peuvent ne pas vérifier toutes les propriétés de cette classe plus générale. Pour illustrer ce phénomène, considérons le réseau ontologique ci-dessous (figure 7.2).

Figure 7.2 – **L'entité individuelle [La-moto-de-Paul] ne satisfait pas toutes les propriétés de la classe [Moto]. La sous-classe [Moto-avec-trois-roues] ne satisfait pas toutes les propriétés de la classe [Moto]**



Ce réseau correspond aux dix énoncés déclaratifs suivants :

8. une chose qui a un moteur est un véhicule ;
9. une chose qui a 2 roues est un véhicule ;
10. une chose qui a trois roues est un véhicule ;
11. une moto a deux roues ;
12. une moto a un moteur ;
13. il y a des motos avec trois roues ;
14. une chose ne peut pas avoir en même temps deux roues et trois roues ;
15. la moto de Paul est une moto ;
16. la moto de Paul a trois roues ;
17. la moto de Paul est rouge.

Parce que « La moto de Paul est une moto » (énoncé 15), elle hérite de toutes les propriétés de la classe [Moto], en particulier [Chose-qui-a-deux-roues]. Un paradoxe est introduit par l'énoncé 16 parce que « Une moto a deux roues » est un fait général mais pas un fait universel. L'énoncé 11 « Une moto a deux roues » signifie « En général, une moto a deux roues mais il y a des exceptions à cette loi ».

En intelligence artificielle, la solution pour ce genre de problèmes est le raisonnement par défaut : une entité individuelle A appartenant à un concept F hérite des concepts subsumant F sauf indication contraire. Cette technique de raisonnement par défaut conduit par exemple Reiter (1980) à proposer la logique des défauts (Section 2).

Nous dirons qu'une entité individuelle est un élément typique d'une classe si elle vérifie toutes les propriétés de cette classe. De façon similaire, nous pouvons considérer qu'une classe est une sous-classe typique d'une autre classe si tous ses éléments vérifient les propriétés de la classe plus générale. Par exemple, la classe [Moto] est une sous-classe typique de la classe [Chose-qui-a-deux-roues].

Une entité individuelle est un élément atypique d'une classe si elle ne vérifie pas toutes les propriétés de la classe. Par exemple, [La-moto-de-Paul] est un élément atypique de la classe [Moto]. De la même manière, nous considérons une classe comme une sous-classe atypique si tous ses éléments typiques ne vérifient pas toutes les propriétés de la classe plus générale. Par exemple, la classe [Moto-avec-trois-roues] est une sous-classe atypique de la classe [Moto].

Il est important de faire la différence entre les entités individuelles atypiques et les classes atypiques. Quand nous considérons la classe [Moto-avec-trois-roues], nous ne distinguons pas un objet particulier, mais un ensemble d'objets indéterminés qui ont des propriétés typiques.

D'après J.-P. Desclés (Desclés, 1990), en fait, on ne parle pas d'un objet réel, mais d'un « objet typique » qui n'existe pas dans la réalité, mais qui vérifie toutes les propriétés du concept par définition. Nous dirons dans la langue « une moto à trois roues ». C'est ce qu'on appelle la vision en intension d'un concept. La classe associée à un concept, c'est l'ensemble des objets à un instant *t* qui vérifient les propriétés du concept. C'est la vision en extension d'un concept. Cet ensemble peut varier au cours du temps, alors que l'objet typique d'un concept ne change pas : on exprimera cela dans la langue par « les motos à trois roues ».

À l'inverse, une entité individuelle désigne un objet particulier qui a ses propres propriétés en plus des propriétés de la ou des classes auxquelles il est rattaché. On parlera de l'individu « Paul », qui est une entité particulière qui hérite bien sûr des propriétés de la classe des êtres humains mais qui en plus se voit attribuer des propriétés spécifiques : il habite à Paris, il a un vélo, il n'a que 45 chromosomes, etc. On effectue ainsi des déterminations successives à partir de l'objet typique pour obtenir un objet complètement déterminé : une entité individuelle.

Aussi, si nous reprenons l'exemple de « Paul qui a 45 chromosomes », [Avoir-45-chromosomes] est une propriété particulière à l'entité individuelle Paul. Dans le cas de Paul, il ne s'agit pas d'une propriété générale.

Il faut donc distinguer deux types de propriétés atypiques, qui ne sont pas de même nature :

a) celles qui concernent un concept, c'est-à-dire qui sont valables pour tous les éléments typiques de la classe associée au concept ;
b) celles qui concernent un objet particulier, et seulement à lui-même.

Le texte qui suit est organisé comme suit. Dans la section 2, nous présentons les notions clés utilisées dans ce chapitre, en particulier les principes fondamentaux de quelques logiques non monotones. Nous expliquons, dans la section 3, les raisons qui nous ont amenés à utiliser la topologie générale et nous indiquons deux méthodes pour la définir. Dans les sections 4 et 5, nous définissons les six relations d'inclusion et d'appartenance topologiques et nous précisons leurs propriétés. Quelques interprétations, utilisant les relations d'inclusion et d'appartenance topologiques, font l'objet de la section 6. Dans cette section, nous utilisons ces interprétations pour illustrer le phénomène de typicalité décrit dans l'introduction et nous montrons les inférences possibles fondées sur les relations déjà présentées. La section 7 expose un nouveau modèle hybride de représentation des connaissances qui combine les logiques non monotones et la topologie qui permet de manipuler des règles contenant des prédicats négatifs. On y explique aussi l'implémentation du modèle utilisant AnsProlog. Finalement, la conclusion résume notre travail et présente quelques perspectives.

2. RAPPELS SUR LES LOGIQUES NON MONOTONES

2.1. Raisonnement non monotone

La logique classique formalise le raisonnement « correct » c'est-à-dire un type de raisonnement qui respecte la propriété de monotonie. Dans un système formel axiomatique de déduction, cette propriété peut être exprimée de la façon suivante :

– Étant donné $\{F_1, F_2, \ldots, F_n\}$ un ensemble de formules logiques, F et G deux formules logiques, si F peut être déduite de $\{F_1, F_2, \ldots, F_n\}$ et F peut être déduite de $\{F_1, F_2, \ldots, F_n, G\}$, alors le système formel est monotone.

En d'autres termes, si un ensemble de formules $\{F_1, F_2, \ldots, F_n\}$ considéré comme vrai permet de déduire que F est vraie, alors l'ajout d'une nouvelle formule dans cet ensemble ne modifie pas la valeur de vérité de F. La propriété de monotonie est opposée à la notion de raisonnement révisable (c'est-à-dire raisonnement non monotone). Dans un système formel de raisonnement révisable, il est possible (et souhaitable) de déduire de $\{F_1, F_2, \ldots, F_n\}$ que F est vraie et de déduire de $\{F_1, F_2, \ldots, F_n, G\}$ que F est faux.

2.2. Logique des défauts

Un des inconvénients majeurs de la logique monotone réside dans le fait que la base de connaissances ne peut être mise à jour en ajoutant un fait qui n'est pas consistant avec la base. Dans Reiter (1980), l'auteur introduit la logique des défauts qui permet le raisonnement révisable. Elle permet de déduire des faits qui sont vrais dans la plupart des cas, mais qui peuvent aussi être invalidés pour certaines exceptions. Une théorie des défauts est formalisée par un couple (P, V) où P est un ensemble de formules fermées du premier ordre (toutes les variables sont liées) décrivant des faits vrais et V est un ensemble de règles d'inférences appelées défauts. Cela signifie que la logique des défauts utilise des formules de la logique du premier ordre, et en particulier des règles d'inférence qui traitent des exceptions. Les défauts sont des règles qui prennent la forme suivante :

$$\frac{P(X) : J_1(X), J_2(X), \ldots, J_n(X)}{C(X)}$$

Où $P(X)$, $J_1(X)$, $J_2(X)$, …, $J_n(X)$ et $C(X)$ sont des formules bien formées du premier ordre dans lesquelles les variables sont libres. Elles sont interprétées dans le sens suivant : si la fait $P(X)$ est cru vrai et ses justifications $J_1(X)$, $J_2(X)$, …, $J_n(X)$ sont consistants dans la théorie (P, V), alors le conséquent $C(X)$ peut être inféré. Habituellement, seuls les défauts normaux sont utilisés ; ils sont de la forme suivante :

$$\frac{P(X) : C(X)}{C(X)}$$

Où $C(X)$ est à la fois la justification et le conséquent. Dans ce cas, seule une modification de P, qui entraîne une inconsistance de $C(X)$ avec V, peut introduire de la non-monotonie dans la théorie.

Voici un exemple de défaut normal :

$$\frac{\text{oiseau}(x) : \text{vole}(x) \text{ (typiquement, les oiseaux volent)}}{\text{Vole}(x)}$$

Cette catégorie de défauts permet de formaliser un système facilement et maintient la correspondance avec d'autres formalismes. Les défauts non normaux sont caractérisés par certaines différences entre leurs justifications et leurs conséquents. Un cas particulier est celui des défauts semi normaux dans lesquels les conséquents sont (seulement) inclus dans leurs justifications. Par exemple :

$$\frac{\begin{array}{c}\text{oiseau}(x) : \neg\, \text{pingouin}(x), \text{vole}(x) \\ \text{(typiquement, les oiseaux qui ne sont pas des pingouins volent)}\end{array}}{\text{vole}(x)}$$

Le format de ce type de défauts empêche les conflits entre règles de se produire. Les règles de défauts permettent la construction d'extensions de P. Toutes les formules qui peuvent être inférées par l'application successive des défauts, dans lesquels les justifications sont consistantes, sont introduites dans P. Ce processus est répété jusqu'à stabilisation à un point fixe. Chaque extension obtenue par ce processus est un ensemble de formules minimal et consistant pour des déductions classiques et contenant P. Une théorie peut avoir une extension, plusieurs extensions mutuellement exclusives ou aucune extension. Une des difficultés de ce type de processus consiste en la possibilité qu'une extension puisse être stabilisée dans un cycle au lieu d'un point fixe : ce problème est également rencontré dans d'autres logiques non monotones.

2.3. Vue d'ensemble de l'hypothèse du monde fermé

L'hypothèse du monde fermé est une logique non monotone introduite par Ray Reiter en 1978 (Reiter, 1978 ; Morgenstern, 1999 ; Brewka *et al.*, 1997 ; Sombé, 1989). Il suppose qu'un ensemble d'axiomes donné est incomplet. Si une formule n'est pas à l'intérieur de cet ensemble et ne peut pas être inférée à partir de cet ensemble, alors elle est supposé être faux jusqu'à une révision supplémentaire des faits. L'ajout de faits peut nécessiter la révision des conclusions.

L'hypothèse du monde fermé est souvent utilisée dans la consultation de systèmes de bases de données. Par exemple, si une base de données concernant les transports aériens ne contient pas d'information concernant un vol particulier et qu'en même temps ce vol ne peut pas être déduit des faits existants, alors, sous l'hypothèse des mondes fermés, ce vol n'est pas supposé exister.

2.4. Discussion sur la logique par défaut

Les problèmes des logiques par défaut peuvent être énumérés comme suit :

- la difficulté de définir des théories par défaut qui ne contiennent pas des cycles d'inférence ;
- la difficulté de décider si une formule du langage appartient ou non à l'extension de P. Si nous excluons le cas d'une théorie fermée contenant un ensemble fini de défauts normaux, il n'existe pas de processus général de décision ;
- la difficulté de décider s'il existe une extension ou non (même si avec des défauts normaux cela peut être décidé) ;
- la difficulté de choisir parmi plusieurs extensions : *a priori*, toutes les extensions peuvent être justifiées pour le raisonnement.

Les deux premiers points sont directement corrélés à l'approche dont le principe est caractérisé par l'implicite et le non énumératif. En d'autres termes, l'absence de priorités entre plusieurs possibles extensions conduit à de

nombreuses propositions en utilisant des critères heuristiques, des priorités ou des préférences. Par exemple, les extensions minimales d'une théorie peuvent être un critère de sélection. Ces difficultés se rencontrent également dans d'autres approches du raisonnement non monotones.

2.5. Logique autoépistémique (Moore, 1985)

La logique autoépistémique formalise la non-monotonicité en utilisant des énoncés d'une logique modale de croyance avec un opérateur de croyance appelé L. La logique autoépistémique se focalise sur un ensemble stable d'énoncés qui peut être vu comme étant les croyances d'un agent rationnel et comme des extensions stables d'un ensemble de prémisses. Les propriétés des ensembles stables comportent la consistance et une version de la négation introspective : si un énoncé P n'appartient pas à l'ensemble de croyance, alors l'énoncé ¬ L P appartient à l'ensemble de croyance. Ceci correspond au principe selon lequel si un agent ne croit pas un fait particulier, alors il croit qu'il ne croit pas ce fait.

2.6. Circonscription (McCarthy, 1980 ; 1986)

La circonscription formalise le raisonnement non monotone à l'intérieur de la logique classique en limitant l'extension de certains prédicats, c'est-à-dire en les circonscrivant. Par exemple, considérons la théorie contenant les hypothèses suivantes : les oiseaux typiques volent, les oiseaux atypiques (habituellement appelés anormaux) ne volent pas, les pingouins sont atypiques, Opus est un pingouin, et Tweety est un oiseau. Opus doit être dans la classe des oiseaux atypiques qui ne volent pas, mais il n'y a pas de raison pour que Tweety soit dans cette classe ; donc nous concluons que Tweety peut voler. La circonscription d'une théorie est obtenue en ajoutant un axiome de second ordre (ou, dans la théorie du premier ordre, un schéma d'axiomes), limitant l'extension de certains prédicats à un ensemble d'axiomes.

Le système ci-dessous décrit une façon de déterminer les conséquents non monotones d'un ensemble d'hypothèses. Les relations d'inférence (Kraus *et al.*, 1990) généralisent ces approches en considérant l'opérateur d'inférence |~, où P |~ Q signifie que Q est une conséquence non monotone de P, et en formulant des principes généraux de fonctionnement de |~. Ces principes spécifient comment |~ concerne l'inférence |- de la logique classique, et comment des méta énoncés renvoyant à l'opérateur d'inférence peuvent être combinés.

2.7. Révision de croyance (Alchourron *et al.*, 1985)

Le modèle de révision de croyance étudie le raisonnement non monotone à partir d'un point de vue dynamique. Il se focalise sur la façon dont les anciennes croyances sont retirées tandis que de nouvelles croyances sont ajoutées à la base

de connaissance. Il y a quatre opérateurs interconnectés en particulier : la contraction, le retrait, l'expansion et la révision. En général, la révision d'une base de connaissance suit le principe des changements minimaux : le plus d'informations possibles est conservé.

2.8. Implémentations

Les applications informatiques de la non-monotonie sont rares. Les concepteurs rencontrent plusieurs difficultés. Tout d'abord, la plupart des logiques non monotones se réfèrent explicitement à la notion de consistance d'un ensemble d'énoncés. Or, la détermination de la consistance est en général indécidable pour les théories du premier ordre ; de la même manière la logique des prédicats non monotone est indécidable. La détermination des inconsistances est décidable pour la logique des propositions, tandis que pour les raisonnements dans la logique des propositions non monotone, la complexité est exponentionnelle. Ceci empêche le développement d'un système général efficace de raisonnement non monotone (Selman et Levesque, 1993).

Cependant, des systèmes performants ont été mis en œuvre dans des situations limitées. La programmation logique, qui est une technique de programmation utilisant un ensemble d'énoncés logiques sous forme de clauses, utilise une méthode appelée « négation par échec ».

2.8.1. Prolog (Colmerauer et Roussel, 1992)

Prolog est le plus connu des langages de programmation logique. Il est fondé sur un calcul des prédicats du premier ordre, mais est restreint aux closes de Horn dans sa version initiale. Des versions plus récentes de ce langage incluent des prédicats plus complexes, en utilisant la négation par échec.

2.8.2. AnsProlog (Answer Set Prolog) (Baral, 2003 ; Gelfond, 2008)*

AnsProlog* est un type particulier de programmation logique. AnsProlog* est un langage utilisé pour décrire directement les faits (c'est-à-dire le réseau sémantique initial) et les règles d'inférence dans le même formalisme utilisé par Prolog.

2.8.2.1. Syntaxe

Un programme AnsProlog* consiste en une collection de règles de la forme :

$$L_0 \text{ or } \ldots \text{ or } L_k \text{ :- } L_{k+1,}, \ldots, L_m, \text{ not } L_{m+1}, \ldots, \text{ not } L_n.$$

Où les L_{is} sont des littéraux dans le sens de la logique classique.

La règle peut être lue comme ceci : Si L_{k+1}, …, L_m sont vrais, et si not L_{m+1}, …, not L_n peuvent être considérés comme étant faux alors au moins un des L_0 ou … ou L_k doit être vrai.

2.8.2.2. AnsProlog vs Prolog*

AnsProlog* est une alternative déclarative à Prolog. Outre le fait que AnsProlog* admet des disjonctions dans la tête des règles, ce qui suit sont les principales différences entre AnsProlog* et Prolog, d'après (Baral, 2003, p. 4) :

- L'ordre des littéraux dans le corps d'une règle a de l'importance dans Prolog parce qu'il procède de gauche à droite. De la même manière, la position d'une règle dans le programme a de l'importance parce qu'il procède du début du programme jusqu'à la fin. L'ordre des règles et la position des littéraux dans le corps d'une règle n'a pas d'importance dans AnsProlog*. Du point de vue d'AnsProlog*, un programme est un ensemble de règles AnsProlog*, et dans chaque règle AnsProlog*, le corps est un ensemble de littéraux et de littéraux précédés par « not ».
- Une requête en Prolog est un processus « *top-down* » partant de la requête jusqu'aux faits. En AnsProlog*, la méthode de réponse à une requête ne fait pas partie de la sémantique. La plupart des interpréteurs fidèles et complets avec la méthodologie AnsProlog* traitent les requêtes de façon « *bottom-up* » en partant des faits vers les conclusions ou les requêtes.
- À cause des processus « *top-down* » pour la recherche de réponses à une requête, et l'application des règles du début jusqu'à la fin, et aussi l'évaluation des littéraux de gauche à droite dans le corps d'une règle, un programme Prolog peut entrer dans une boucle infinie même pour un programme très simple sans « négation par échec ».
- L'opérateur de coupure (« *cut* ») dans Prolog est externe à la logique, bien qu'il y a eu des tentatives récentes de le caractériser. Cet opérateur ne fait pas partie d'AnsProlog*.
- Il y a certains problèmes, par exemple rester bloqué dans une boucle infinie. En général, nous rencontrons des problèmes avec des programmes qui comportent des récursions au moyen de la négation par échec. AnsProlog* ne rencontre pas ces problèmes, et comme son nom l'indique, le résultat sont des ensembles de résultats sémantiques[2] pour caractériser la « négation par échec ».

2.8.2.3. AnsProlog vs Logique des défauts

AnsProlog est une sous-classe de AnsProlog* avec dans les règles un seul atome dans la tête de la règle et sans négation dans le corps de la règle. La sous-classe AnsProlog peut être considérée comme une sous-classe particulière des logiques

2. ASP = *answer set semantics.*

par défauts qui conduit à une mise en œuvre plus efficace. Nous rappelons que la logique des défauts est un couple (P, V), où P est une théorie du premier ordre et V est une collection de défauts. AnsProlog peut être considéré comme un cas particulier où P = ∅. Du reste, il a été démontré que AnsProlog* et la logique des défauts ont le même pouvoir d'expression. En résumé, AnsProlog* est plus simple que la logique des défauts d'un point de vue syntaxique et cependant a le même pouvoir d'expression, ce qui le rend plus facilement utilisable (Baral, 2003, p. 5).

3. UTILISER LA TOPOLOGIE POUR REPRÉSENTER LES ENTITÉS ATYPIQUES DANS LES ONTOLOGIES

3.1. Pourquoi utiliser la topologie ?

Les réseaux de concepts et de relations sémantiques entre concepts peuvent être représentés sur un plan. Les instances sont des points du plan. Les classes sont alors des lieux du plan, c'est-à-dire des espaces délimités qui sont constitués *a*) d'un intérieur (les éléments typiques qui appartiennent à la classe), *b*) d'un extérieur (les éléments qui ne sont pas dans la classe), et *c*) d'une frontière (les éléments atypiques qui ne vérifient pas toutes les propriétés de la classe, c'est-à-dire qui ne sont ni à l'intérieur, ni à l'extérieur de la classe). Or, littéralement, la topologie signifie l'étude du lieu. Elle s'intéresse donc à définir ce qu'est un lieu (appelé aussi espace) et quelles peuvent en être les propriétés.

3.2. Différentes façons de définir la topologie

Une topologie est fondée sur la théorie des ensembles et par conséquent sur la logique du premier ordre. Du point de vue de la logique du premier ordre, un concept F peut être vu comme un prédicat (ou une fonction) qui s'applique à une variable X et qui retourne une valeur de vérité parmi {vrai, faux}. D'un autre côté, du point de vue de la théorie des ensembles, ceci est équivalent à dire que X appartient à l'ensemble F. Ainsi, un concept F peut être considéré aussi bien comme un ensemble que comme un prédicat (Frege, 1893).

Dans la suite, deux définitions logiquement équivalentes de la topologie sont présentées. La première, qui est la définition « classique », est fondée sur la théorie des ensembles. La deuxième adopte une approche fonctionnelle. Parce que la deuxième approche est plus adaptée aux logiques de déduction, cette deuxième approche sera utilisée dans la suite du chapitre. Ensuite, le raisonnement non monotone sera introduit dans les ontologies pour permettre l'héritage de propriétés aux entités atypiques.

3.2.1. La définition classique de la topologie fondée sur la théorie des ensembles

Classiquement, une topologie sur un ensemble E utilise la notion d'ouvert. Étant donné O une famille de partie de E telle que :

1. $\varnothing \in O$;
2. $E \in O$;
3. Si les $A_i \in O$, alors $\cup A_i \in O$ (les A_i étant en nombre fini ou infini.) ;
4. Si les $A_i \in O$, alors $\cap A_i \in O$ (les A_i étant en nombre fini).

On appellera alors un ouvert de E tout élément de O. Un élément $a \in E$ sera dit *intérieur* à la partie A de E s'il existe un ouvert P tel que : $a \in P \subset A$.

L'intérieur iA de A est l'ensemble des éléments intérieurs à A : c'est un ouvert, l'union des ouverts inclus dans A. Nous pouvons alors définir l'application $i : A \rightarrow iA$ dans l'ensemble 2^E des parties de E par :

$$iA = \bigcup_{\substack{o \subset A \\ o \in O}} o$$

Autrement dit, iA est l'union des ouverts inclus dans A.

Si A est ouvert, iA = A, (puisque A est l'un des ouverts inclus dans A), et réciproquement, si iA = A, comme iA est toujours un ouvert (une union d'ouverts est un ouvert d'après *c*)), c'est que A est ouvert. Nous avons donc la caractérisation :

$A \in O \Leftrightarrow iA = A$.

En outre, nous avons les propriétés :

1. iiA = iA (idempotence) ;
2. $iA \subset A$;
3. a) $i(A \cap B) = iA \cap iB$ et b) $iA \cup iB \subset i(A \cup B)$;
4. $A \subset B \Rightarrow iA \subset iB$ (monotonie).

On définit ensuite les fermés comme étant les complémentaires des ouverts. Par les lois de Morgan, on obtient immédiatement les propriétés de ceux-ci : ils constituent une famille F de parties de E telle que :

(a') $E \in F$
(β∋) $\varnothing \in F$
(c') Si les $A_i \in F$, $\cap A_i \in F$, les A_i étant en nombre fini ou infini.
(d') Si les $A_i \in F$, $\cup A_i \in F$, les A_i étant en nombre fini.

L'application « fermeture », f : A→fA dans les parties de E a la définition duale de celle d'« intérieur » i :

$$fA = \bigcap_{\substack{A \subset X \\ X \in F}} X$$

Autrement dit, fA est l'intersection des fermés contenant A. Ses propriétés, duales de celles de i, sont les suivantes :

5. $A \subset B \Rightarrow fA \subset fB$ (monotonie)
6. $A \subset fA$
7. $ffA = fA$ (idempotence)
8. a) $f(A \cup B) = fA \cup fB$ et b) $f(A \cap B) = fA \cap fB$

Nous venons de voir comment se comporte l'application i par rapport à l'union et l'intersection. Il reste à examiner comment i se comporte par rapport à la troisième opération booléenne, le complémentaire (nous noterons par la suite cA le complémentaire d'un ensemble A de E). En particulier, icA est l'ensemble des points intérieurs au complémentaire de A. Un tel ensemble sera nommé l'extérieur de A (et noté eA) :

$ea = icA$.

Les propriétés (1), (2) et (4) de l'application i ont, compte tenu des lois de Morgan sur le complémentaire, leurs correspondances pour l'application e :

9. $A \subset B \Rightarrow eA \supset eB$;
10. $A \cap eA = \varnothing$;
11. $e(A \cup B) = eA \cap eB$.[3]

Lorsque l'on considère l'intérieur et l'extérieur d'un ensemble A, il reste ce qu'il y a entre les deux, qui n'est ni intérieur, ni extérieur et que l'on appelle la frontière de A, notée bA. Par définition, la frontière de A est le complémentaire de $iA \cup eA$:

12. $bA = c\,(iA \cup eA) = ciA \cap ceA$.

Toutefois, il existe une seconde manière de définir une topologie, moins simple et moins classique que celle qui vient d'être exposée, mais plus « opératoire ». Il s'agit de se donner non plus les ouverts (ou les fermés) mais l'application i (ou l'application f) avec ses propriétés.

3. La propriété 3 d'idempotence n'a pas de correspondance simple.

3.2.2. Une définition opérationnelle de la topologie

Donnons-nous par exemple, E étant un ensemble quelconque, une application f de l'ensemble 2^E des parties de E dans lui-même satisfaisant aux propriétés :

(0) $f \varnothing = \varnothing$;
(1) $A \supset B \Rightarrow fA \supset fB$;
(2) $fA \supset A$;
(3) $ffA = fA$;
(4) $f(A \cup B) = fA \cup fB$.

Nous appelons « fermés » les parties de E invariantes dans f :

$F = \{ X : X \subset E \ \& \ fX = X\}$

La famille F des parties de E ainsi définies satisfait à tous les axiomes des fermés, qui ont été donnés dans la première manière de définir une topologie. Ainsi, se donner f (ou i) et ses propriétés équivaut à se donner F (ou O) et définit une topologie sur E. Une fois que l'on a les fermés, le reste de la construction peut se faire soit comme dans la première manière, soit en définissant directement d'abord l'application i comme transmuée de f par complémentation : $i = cfc$ et les ouverts comme parties invariantes de i. La relation : $ic = cf$ exprime alors qu'une partie est fermée si et seulement si sa complémentaire est ouverte.

Nous avons donc deux façons logiquement équivalentes de définir une topologie. Nous verrons que la manière « opératoire » est plus intéressante en vue de représenter les entités atypiques dans les ontologies.

3.3. Intégration de la logique non monotone

Les systèmes de raisonnement utilisant les logiques non monotones ne sont utiles que si elles peuvent être intégrées avec succès à d'autres formes de raisonnement de sens « commun ». Par exemple, nous pouvons intégrer les logiques non monotones au raisonnement temporel (Hanks et McDermott, 1987) ou aux systèmes multiagents (Morgenstern et Guerreiro, 1993). En général, l'intégration peut nécessiter l'extension à la fois du formalisme non monotone et de la théorie de raisonnement de sens commun.

Dans ce chapitre, nous allons intégrer les logiques non monotones et la topologie générale en vue d'améliorer la modélisation des exceptions dans les ontologies.

4. RELATIONS TOPOLOGIQUES D'INCLUSION ET D'APPARTENANCE

4.1. De la logique non monotone vers la topologie générale

Les logiques non monotones peuvent être vues comme une illustration de la topologie générale exposée précédemment, et elles peuvent être utilisées pour représenter les exceptions dans les ontologies. Étant donné un élément X et un prédicat C (ou une classe, c'est-à-dire une partie de E), nous avons les assertions suivantes :

- Si en logique des défauts la formule C(X) est vraie, sauf indication contraire, alors cela équivaut à dire en topologie que X est à l'intérieur de C. Du point de vue des ontologies, cela revient aussi à dire que X est un élément typique de C.
- Si en logique des défauts la formule C(X) est fausse, sauf indication contraire, alors cela équivaut à dire en topologie que X est à l'extérieur de C. Du point de vue des ontologies, cela revient aussi à dire que X est un élément sans rapport avec C.
- Enfin, si en logique des défauts la formule C(X) est partiellement vraie, c'est-à-dire qu'elle ne vérifie pas toutes les propriétés de C, alors cela équivaut à dire en topologie que X est à la frontière de C. Du point de vue des ontologies, cela revient aussi à dire que X est un élément atypique de C.

Nous avons les trois mêmes équivalences avec les classes / sous-classes typiques ou atypiques.

Dans la suite, X représente un point quelconque et {X} un singleton (c'est-à-dire le plus petit voisinage contenant X). Nous notons S l'ensemble des singletons de E. Enfin, A, B, C représentent des parties quelconques de E qui ne sont pas des singletons.

4.2. Définitions des relations topologiques

a) Appartenance à l'intérieur d'une classe (notée $\in_i$)

Nous définissons ($X \in_i A$) si et seulement si X hérite de toutes les propriétés de A :

$$X \in_i A \Leftrightarrow \{X\} \subset iA.$$

b) Appartenance à l'extérieur d'une classe (notée $\in_e$)

Nous définissons ($X \in_e A$) si et seulement si X ne peut appartenir ni à l'intérieur ni à la frontière de A (et de la même façon récursivement pour les sous-classes de A) :

$$X \in_e A \Leftrightarrow \{X\} \subset eA.$$

c) Appartenance à la frontière d'une classe (notée $\in_b$)

Nous définissons ($X \in_b A$) si et seulement si X est une entité individuelle atypique de A :

$$X \in_b A \Leftrightarrow \{X\} \subset bA.$$

d) Inclusion à l'intérieur d'une classe (notée $\subset_i$)

Nous définissons ($A \subset_i B$) si et seulement si A est une sous-classe typique de B :

$$A \subset_i B \Leftrightarrow A \subset iB.$$

e) Inclusion à l'extérieur de A (notée $\subset_e$)

Nous définissons ($A \subset_e B$) si et seulement si A ne peut être une sous-classe ni à l'intérieur ni à la frontière de B :

$$A \subset_e B \Leftrightarrow A \subset eB.$$

f) Inclusion à la frontière d'une classe (notée $\subset_b$)

Nous définissons ($A \subset_b B$) si et seulement si A est une sous-classe atypique de la classe B :

$$A \subset_b B \Leftrightarrow A \subset bB.$$

Notons que $\in_I, \in_b$ et $\in_e$ sont des sous-ensembles du produit cartésien $S \times D$, tandis que $\subset_i, \subset_b$ et $\subset_e$ sont des sous-ensembles du produit cartésien $D \times D$.

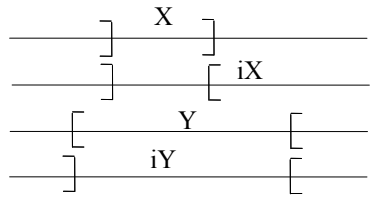
5. RÈGLES D'INFÉRENCE POUR LES SIX RELATIONS D'INFÉRENCE DES RELATIONS TOPOLOGIQUES D'INCLUSION ET D'APPARTENANCE

En utilisant les propriétés des opérateurs i, e, b et f, nous pouvons déduire des règles d'inférence pour les relations d'inclusion et d'appartenance. Pour démontrer (ou invalider) ces propriétés, deux méthodes peuvent être utilisées :

1. La première méthode permet de démontrer qu'une propriété est vraie. Cette méthode utilise les propriétés des opérateurs i, e, b et f comme des axiomes. Par exemple, la relation $\subset_i$ est transitive. En fait, cela revient à démontrer que $(X \subset_i Y) \wedge (Y \subset_i Z) \Rightarrow (X \subset_i Z)$. Or, $iY \subset Y$ (propriété 2 de i, Section 3.2.1, cette propriété est utilisée comme un axiome) ainsi nous avons, $X \subset iY \subset Y \subset iZ$ donc $X \subset iZ$.

2. La seconde méthode permet d'invalider une propriété. Elle consiste à projeter l'espace topologique E sur une droite orientée et de trouver un contre-exemple. Par exemple, la relation $\subset_i$ n'est pas symétrique parce que si nous supposons que $(X \subset_i Y) \wedge (Y \subset_i X)$, alors en projetant cette relation sur la droite orientée, nous avons (figure 7.3):

Figure 7.3 – **$\subset_i$ n'est pas symétrique parce que dans ce contre-exemple $(X \subset_i Y)$ et $(Y \not\subset_i iX)$**



5.1. Réflexivité, symétrie et transitivité des relations d'inclusion

Notons qu'il est inapproprié de considérer ces propriétés pour les relations d'appartenance parce que ces relations sont définies sur SxE et non pas sur ExE.

a) Inclusion à l'intérieur d'une classe $\subset_i$

La relation $\subset_i$ est transitive, non symétrique et non réflexive.

b) Inclusion à la frontière d'une classe

La relation $\subset_b$ n'est pas réflexive, elle est non symétrique et non transitive.

c) Inclusion à l'extérieur d'une classe

La relation $\subset_e$ n'est pas réflexive, elle est non symétrique mais transitive.

De plus, notons que $(F \cap eF) = (G \cap eG) = \varnothing$ (axiome 8, Section 3.2.1). Ainsi, $F \subset_e G \Rightarrow G \cap eF = \varnothing$ parce que $F \subset eG$.

Nous avons par conséquent les règles d'inférence suivantes à partir de $G \subset eF$:

F1: $(G \subset_e F) \wedge (H \subset_i G) \Rightarrow (H \subset_e F)$.

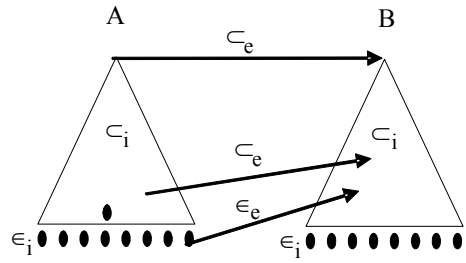
En fait, $H \subset iG \subset G \subset eF$ donc $H \subset eF$.

F2: $(G \subset_e F) \wedge (X \in_i G) \Rightarrow (X \in_e F)$.

En fait, $\{X\} \subset iG \subset G \subset eF$ donc $\{X\} \subset eF$.

Ces deux règles se propagent dans les sous-hiérarchies d'inclusion intérieure de F et de G. Il en est de même pour les feuilles des hiérarchies par l'intermédiaire des relations d'appartenance intérieure (Figure 7.4).

Figure 7.4 – **Propagation des relations $\subset_e$ et $\in_e$ le long des sous-hiérarchies d'inclusion intérieure de F et de G**



5.2. Propriétés de combinaisons des relations topologiques

Dans le but d'identifier exhaustivement les propriétés de combinaisons des six relations topologiques, nous considérons toutes les cases d'un tableau [A … F] x [1 … 6] où un élément Rz du tableau représente la combinaison d'un élément Rx en ligne et un élément Ry en colonne. Par exemple, A1 (si elle existe) contiendra le résultat de la formule logique :

A1: (A $\in$**i** B) $\wedge$ (B $\in$**i** C) $\Rightarrow$ (A Rz? C).

Nous avons le tableau ci-dessous (tableau 7.1).

Tableau 7.1 – **Un tableau [A…F] x [1…6] où une case Rz du tableau représente la combinaison d'un élément Rx en ligne et un élément Ry en colonne**

		1	2	3	4	5	6
	(A Rx B) $\wedge$ (B Ry C) $\Rightarrow$ (A Rz C)	$\in_i$	$\in_b$	$\in_e$	$\subset_i$	$\subset_b$	$\subset_e$
A	$\in_i$	NIL	NIL	NIL	$\in_i$	$\in_b$	$\in_e$
B	$\in_b$	NIL	NIL	NIL	NIL	NIL	NIL
C	$\in_e$	NIL	NIL	NIL	NIL	NIL	NIL
D	$\subset_i$	NIL	NIL	NIL	$\subset_i$	$\subset_b$	$\subset_e$
E	$\subset_b$	NIL	NIL	NIL	NIL	$\subset_b$	NIL
F	$\subset_e$	NIL	NIL	NIL	NIL	NIL	NIL

En d'autres termes, nous avons les règles de combinaison suivantes :

A4: $(X \in_i B) \wedge (B \subset_i C) \Rightarrow (X \in_i C)$

A5: $(X \in_i B) \wedge (B \subset_b C) \Rightarrow (X \in_b C)$

A6: $(X \in_i B) \wedge (B \subset_e C) \Rightarrow (X \in_e C)$

D4: $(A \subset_i B) \wedge (B \subset_i C) \Rightarrow (A \subset_i C)$

D5: $(A \subset_i B) \wedge (B \subset_b C) \Rightarrow (A \subset_b C)$

D6: $(A \subset_i B) \wedge (B \subset_e C) \Rightarrow (A \subset_e C)$

E5: $(A \subset_b B) \wedge (B \subset_b C) \Rightarrow (A \subset_b C)$

Et aussi :

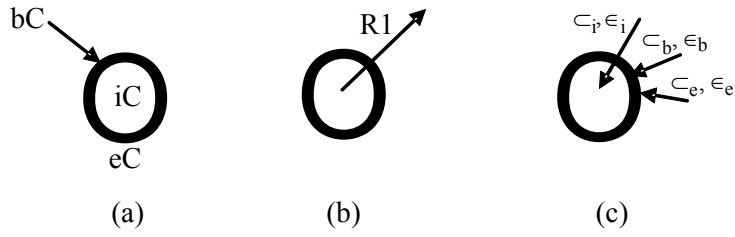
F1: $(A \subset_e B) \wedge (C \subset_i A) \Rightarrow (C \subset_e B)$

F2: $(A \subset_e B) \wedge (X \in_i A) \Rightarrow (X \in_e B)$

6. INTERPRÉTATION TOPOLOGIQUE DES DEUX EXEMPLES

Une entité individuelle est représentée par un point tandis qu'une classe C est représentée sous la forme d'une boule topologique projetée sur un plan, avec son intérieur, sa frontière et son extérieur comme suit (figure 7.5, a) :

Figure 7.5 – ***a*) Représentation d'une classe C; *b*) Le point de départ d'une relation d'inclusion est situé à l'intérieur de C1; *c*) Différents points d'arrivée des arcs représentant des relations**

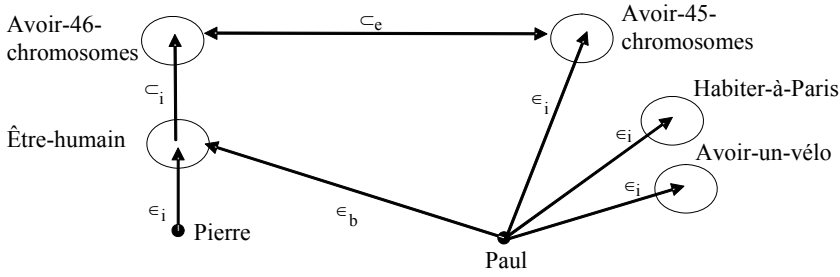


La convention de représentation est la suivante : le point de départ d'une relation R1 d'inclusion entre deux classes (C1 est la classe de départ) est situé à l'intérieur de C1 parce que la relation concerne tous les éléments typiques de C1 et pas nécessairement les éléments atypiques situés à la frontière de C1. Ainsi, nous représentons cette différentiation de cette façon (figure 7.5, b).

Le point d'arrivée d'un arc représentant une relation vers une classe C2 dépendra de la relation. Pour $\subset_i$ et $\in_i$, le point d'arrivée est à l'intérieur de C2, parce que nous avons une classe typique et une entité individuelle typique. Pour $\subset_b$ et $\in_b$, le point d'arrivée est à la frontière de C2, parce que nous avons une classe atypique et une entité individuelle atypique. Pour $\subset_e$ et $\in_e$, le point d'arrivée est à la limite de la frontière de C2, parce que nous avons une classe extérieure à C2 et une entité individuelle extérieure à C2 (figure 7.5, c).

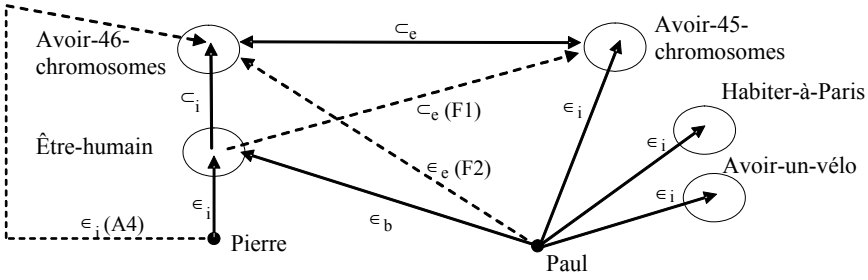
La figure 7.6 représente une interprétation de l'exemple 1 en utilisant nos relations topologiques. En particulier, notons que [Paul] est un élément atypique de la classe [Être-humain].

Figure 7.6 – **[Paul] est un élément atypique de la classe [Être-humain] (exemple 1)**



Dans la figure 7.7 les arcs en pointillé représentent quelques déductions possibles de l'exemple 1 d'après les règles de combinaison définies dans la section 5. Par exemple,

Figure 7.7 – **Quelques déductions possibles dans l'exemple 1**

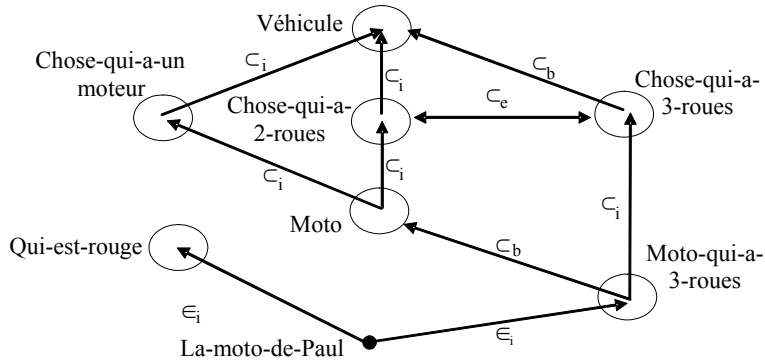


(18) [Pierre] est un élément typique de la classe [Avoir-46-chromosomes] (énoncés 1 et 2 et règle A4) ;

(19) [Paul] ne peut appartenir ni à l'intérieur ni à la frontière de la classe [Avoir-46-chromosomes] (énoncés 4 et 7 et règle F2) ;
(20) [Etre-humain] ne peut être ni une sous-classe typique ni une sous-classe atypique de la classe [Avoir-46-chromosomes] (énoncés 1 et 7 et règle F1).

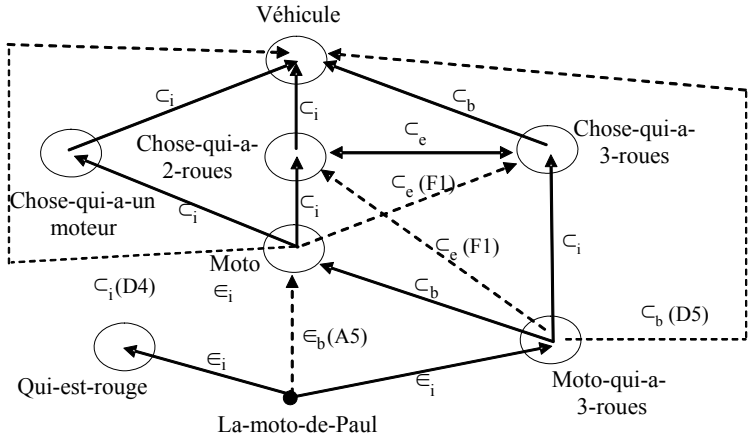
La figure 7.8 représente une interprétation de l'exemple 2 en utilisant nos relations topologiques. En particulier, nous notons que la classe [Moto-qui-a-trois-roues] est une sous-classe atypique de la classe [Moto].

Figure 7.8 – **La classe [Moto-qui-a-trois-roues] est une sous-classe atypique de la classe [Moto] (exemple 2)**



Dans la figure 7.9, les arcs en pointillé représentent quelques déductions possibles dans l'exemple 2 d'après les règles de combinaison définies dans la section 5. Par exemple,

Figure 7.9 – **Quelques exemples de déductions dans l'exemple 2**



Par exemple :

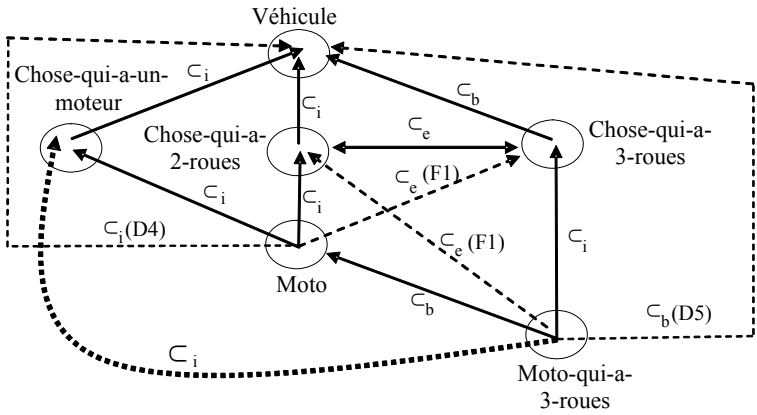
(21) La classe [Moto] est une sous-classe typique de la classe [Véhicule] (énoncés 8 et 12 et règle D4) ;
(22) La classe [Moto-qui-a-trois-roues] est une sous-classe atypique de la classe [Véhicule] (énoncés 10, 13 et règle D5) ;
(23) La classe [Moto-qui-a-trois-roues] ne peut être ni une sous-classe typique ni une classe atypique de la classe [Chose-qui-a-2-roues] (énoncés 13 et 14 et règle F1) ; etc.

7. VERS UN SYSTÈME NON MONOTONE : IMPLÉMENTATION

7.1. Ajout de règles non monotones

Considérons le deuxième exemple. Dans celui-ci, la classe [Moto-qui-a-trois-roues] est une sous-classe atypique de la classe [Moto], parce que la classe [Moto-qui-a-trois-roues] est une sous-classe typique de la classe [Chose-qui-a-trois-roues]. Ceci est incompatible avec la propriété typique des motos qui ont deux roues. Pourtant, puisque nous ne savons pas pourquoi la classe [Moto-qui-a-trois-roues] est une sous-classe atypique de la classe [Moto], toutes les autres inférences sont bloquées. Dans notre exemple, nous voudrions que la classe [Moto-qui-a-trois-roues] hérite de la propriété d'avoir un moteur (figure 7.10).

Figure 7.10 – **La classe [Moto-qui-a-trois-roues] doit hériter de la propriété d'avoir un moteur, par héritage de propriété de la classe [Moto]**



Pour ce faire, il suffit d'ajouter la règle F3 ci-dessous :

F3 : $(F \subset_b G) \wedge (G \subset_i H) \wedge \neg (F \tau_e H) \Rightarrow (F \subset_i H)$.

Nous avons le même phénomène avec les entités individuelles atypiques. Considérons le premier exemple. Paul est un être humain atypique parce qu'il n'a que 45 chromosomes au lieu de 46. Mais, sachant que tous les êtres humains sont des mammifères, nous voudrions que Paul hérite de cette propriété.

Pour ce faire, nous ajoutons simplement la règle F4 ci-dessous :

F4: $(X \in_b F) \wedge (F \subset_i G) \wedge \neg (X \in_e G) \Rightarrow (X \in_i G)$.

Notons que nous avons dans la partie droite des deux règles d'inférence F3 et F4 un prédicat négatif : ($\neg (F \subset_e H)$ pour F3 et $\neg (X \in_e G)$ pour F4). Ces deux règles ne peuvent être déclenchées que si les deux prédicats sont vrais à l'occasion de leurs applications. Ce sont deux règles non monotones. En effet, nous nous trouvons dans le cas de l'hypothèse des mondes fermés (Section 2.3). Par exemple, le prédicat $\neg (F \subset_e H)$ est supposé faux jusqu'à une nouvelle modification de l'ontologie. L'ajout de nouveaux faits peut nécessiter la révision des conclusions de F3 ou F4.

7.2. Implémentation en AnsProlog

Notre modèle a été implémenté en AnsProlog (Baral, 2003) parce que ce langage de programmation logique permet de représenter des énoncés, des exceptions et des énoncés par défaut, et est capable de raisonner avec eux (voir §2.8.2). En particulier, AnsProlog admet des prédicats négatifs dans les règles d'inférence. Les règles d'inférence sont représentées dans AnsProlog directement comme ci-dessous :

A4: $(A \in i\ B) \wedge (B \subset i\ C) \Rightarrow (A \in i\ C)$

```
app_i(O, C) :- app_i(O, B), inclu_i(B, C).
```

A5: $(A \in_i B) \wedge (B \subset_b C) \Rightarrow (A \in_b C)$

```
app_b(O, C) :- app_i(O, B), inclu_b(B, C).
```

A6: $(A \in_i B) \wedge (B \subset_e C) \Rightarrow (A \in_e C)$

```
app_e(O, C) :- app_i(O, B), inclu_e(B, C).
```

D4: $(A \subset_i B) \wedge (B \subset_i C) \Rightarrow (A \subset_i C)$

```
inclu_i(A, C) :- inclu_i(A, B), inclu_i(B, C).
```

D5: $(A \subset_i B) \wedge (B \subset_b C) \Rightarrow (A \subset_b C)$

```
inclu_b(A, C) :- inclu_i(A, B), inclu_b(B, C).
```

D6: $(A \subset_i B) \wedge (B \subset_e C) \Rightarrow (A \subset_e C)$

inclu_e(A, C) :- inclu_i(A, B), inclu_e(B, C).

E5: $(A \subset_b B) \wedge (B \subset_b C) \Rightarrow (A \subset_b C)$

inclu_b(A, C) :- inclu _b(A, B), inclu_b(B, C).

F1: $(A \subset_e B) \wedge (C \subset_i A) \Rightarrow (C \subset_e B)$

inclu_e(A, C) :- inclu _e(A, B), inclu_i(B, C).

F2: $(A \subset_e B) \wedge (C \in_i A) \Rightarrow (C \in_e B)$

app_e(O, B) :- inclu_e(O, B), app_i(O, A).

F3: $(F \subset_b G) \wedge (G \subset_i H) \wedge \neg (F \subset_e H) \Rightarrow (F \subset_i H)$

inclu_i(A, C) :- inclu_b(A, B), inclu_i(B, C), not inclu_e(A, C).

F4: $(X \in_b F) \wedge (F \subset_i G) \wedge \neg (X \in_e G) \Rightarrow (X \in_i G)$

app_i(O, A) :- app_b(O, B), inclu_i(B, A), not app_e(O, A).

Les faits (c'est-à-dire l'ontologie), sont représentés dans AnsProlog par des simples énoncés tels que, pour l'exemple 1 :

app_i(pierre, Etre_humain).

app_i (paul, habiter_a_paris).

app_i (paul, avoir_un_velo).

app_i (paul, avoir_45_chromosomes).

inclu_i(Etre_humain, avoir_46_chromosomes).

inclu_e(avoir_45_chromosomes, avoir_46_chromosomes). Etc.

Nous obtenons les inférences nécessaires en exécutant le programme (figure 7.11).

Concernant l'exemple 2, l'ontologie est représentée par les assertions suivantes :

inclu_i(chose_qui_a_un_moteur, vehicule).

inclu_i(chose_qui_a_deux_roues, vehicule).

inclu_i(moto, chose_qui_a_deux_roues).

inclu_b(chose_qui_a_trois_roues, vehicule).

inclu_e(chose_qui_a_deux_roues, chose_qui_a_trois_roues).

app_i(la_moto_de_paul, moto_qui_a_trois_roues). Etc.

Nous obtenons les inférences nécessaires en exécutant le programme (figure 7.12).

Figure 7.11 – **Les résultats des inférences obtenues avec AnsProlog sur l'exemple 1**

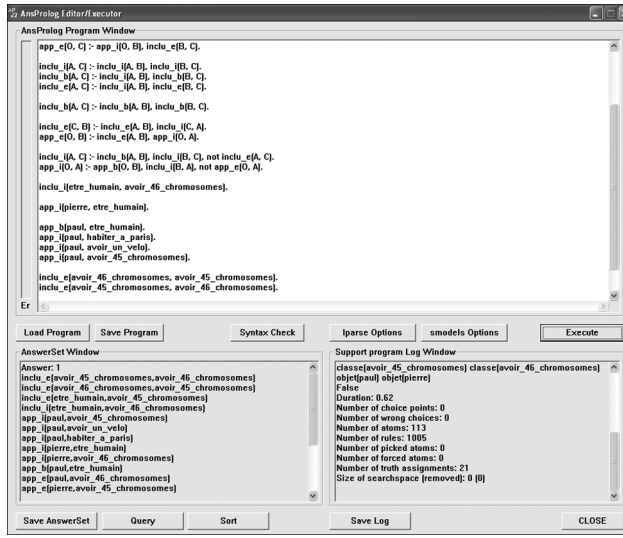
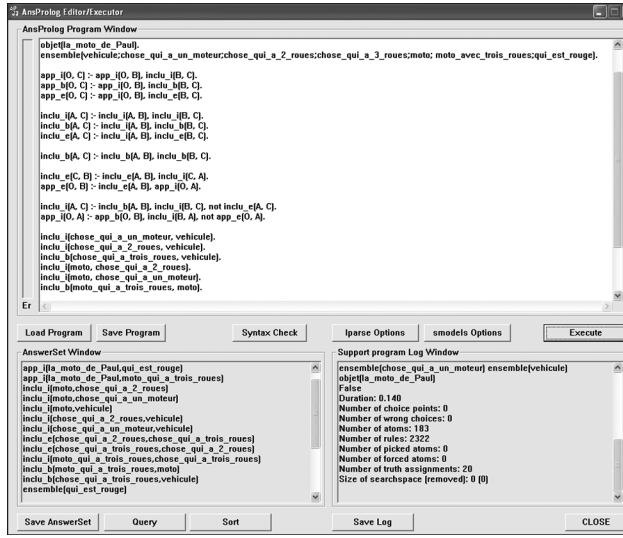


Figure 7.12. – **Les résultats des inférences obtenues avec AnsProlog sur l'exemple 2**



8. CONCLUSION

Dans ce chapitre, nous proposons un nouveau modèle de représentation des connaissances en combinant les ontologies, la logique non monotone et la topologie. Ce modèle peut être utilisé par le concepteur d'une ontologie. Quand une ontologie a atteint une taille importante et qu'une entité atypique est découverte, le coût de la reconstruction de l'ontologie peut être trop élevé. Ce modèle facilite la maintenance de l'ontologie en évitant sa reconstruction.

Pour spécifier si une entité individuelle appartenant à une classe est typique ou non, nous empruntons les concepts topologiques d'intérieur, de frontière, de fermeture et d'extérieur. Pour représenter les entités atypiques dans les ontologies, nous définissons un système de relations d'inclusion et d'appartenance en adaptant les opérateurs topologiques. Nous proposons de formaliser les relations topologiques d'inclusion et d'appartenance en utilisant les propriétés mathématiques des opérateurs topologiques. Or, il y a des propriétés de combinaison des opérateurs intérieur, extérieur, frontière et fermeture permettant la définition d'une algèbre. Nous proposons d'utiliser ces propriétés mathématiques comme un ensemble d'axiomes. Cet ensemble d'axiomes nous permet d'établir les propriétés des relations topologiques d'inclusion et d'appartenance. Notre modèle est implémenté en AnsProlog, un langage de programmation logique récent qui permet des prédicats négatifs dans les règles d'inférence.

Il y a des similarités entre les logiques non monotones et la topologie qui visent à formaliser les entités individuelles dans les ontologies, comme la logique floue (Zadeh *et al.*, 1996) et le raisonnement probabiliste, en particulier les réseaux bayésiens (Wong *et al*., 2003). Ces approches ont pour objectif commun de représenter et de raisonner avec de la connaissance incomplète. Mais notre approche est différente parce qu'elle utilise une méthodologie qualitative plutôt qu'une méthodologie quantitative.

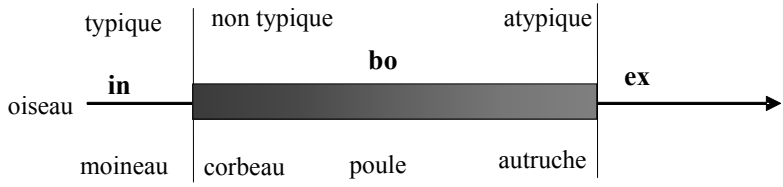
8.1. Perspectives : une échelle de typicalité

Nous pouvons remarquer qu'il existe une troisième façon de définir la topologie, en utilisant la notion de voisinage et une axiomatique appropriée. En topologie générale, un voisinage d'un élément X de E est un ensemble quelconque contenant un ouvert tel que X appartient à cet ouvert. En d'autres termes, un voisinage de X est un ensemble quelconque dans lequel X est à l'intérieur.

Comme perspective nous introduisons une échelle de typicalité. Elle permet de définir des degrés de typicalité où les entités individuelles appartenant à une classe sont plus ou moins typiques. Les éléments les plus typiques sont les éléments pour lesquels il n'existe aucun doute sur leur appartenance dans la classe. Les éléments atypiques ne vérifient pas certaines propriétés de la classe, ce qui implique une réduction de leur degré de typicalité. Par exemple, dans la classe des « oiseaux » un « moineau » est plus typique qu'un « corbeau » parce que pour le sens commun (au moins en France) un oiseau est petit. Une « poule » qui

vole difficilement est encore moins typique qu'un « corbeau », mais plus typique qu'une « autruche » qui ne vole pas. Ces différences nécessitent une échelle de degrés de typicalité. Pour modéliser cette échelle de typicalité, nous introduisons une frontière à épaisseur (figure 7.13).

Figure 7.13 – **Avec une frontière à épaisseur pour la classe « oiseau », nous pouvons introduire des niveaux de typicalité pour « corbeau », « poule » et « autruche »**



Plus il y a d'éléments avec un degré faible de typicalité qui sont rattachés à une classe, et plus la frontière est épaisse. Dans notre exemple, les degrés de typicalité diminuent avec la perte de la propriété de voler ou en fonction du sens commun. Étant donné e, un entier constant arbitrairement posé, qui modélise l'épaisseur de la frontière, il est alors possible de définir l'intérieur in(F, e), l'extérieur ex(F, e) et la frontière bo(F, e) d'une classe F en fonction de e de la façon suivante :

a) Intérieur de F en fonction de e, in(F, e) :

si $e = 0, in(F, e) = in(F)$;

si $e > 0, in(F, e) \subset in(F)$;

$x \in in(F, e) \Leftrightarrow \exists\, n \in N(x) : n \subset in(F, e\text{-}1)$, où N(X) représente l'ensemble des voisinages de X.

b) Extérieur de F en fonction de e, ex(F, e) :

si $e = 0, ex(F, e) = ex(F)$;

si $e > 0, ex(F) \subset ex(F, e)$;

$x \in ex(F, e) \Leftrightarrow \exists\, n \in N(x) : n \cap F = \varnothing \wedge n \subset ex(F, e\text{-}1)$.

c) Frontière de F en fonction de e, bo(F, e):

si $e = 0, bo(F,e) = bo(F)$;

si $e > 0, bo(F) \subset bo(F,e)$;

$x \in bo(F, e) \Leftrightarrow \exists\, n \in N(x), n \cap in(F, e) = \varnothing \wedge n \cap ex(F, e) = \varnothing$.

Ces définitions restent compatibles avec les opérateurs topologiques classiques en posant :

$ex(F) = \cap_{i=0..e} ex(F, i)$;

$in(F) = \cap_{i=0..e} in\ (F, i)$;

$bo(F) = \cap_{i=0..e} bo(F, i)$.

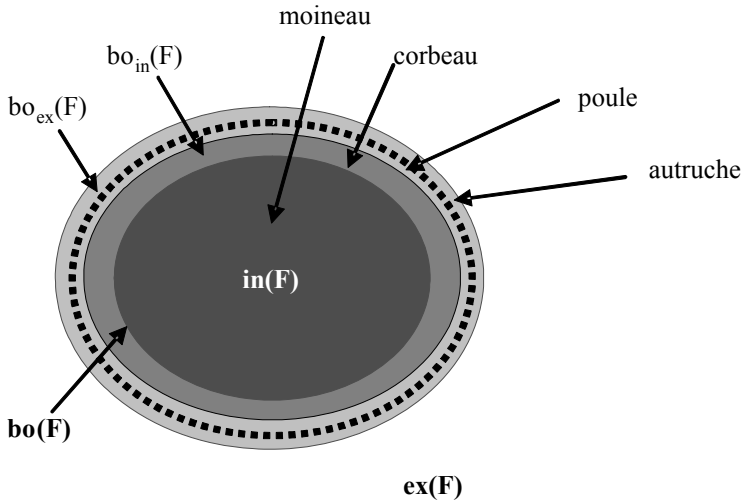
Parce que la frontière a une épaisseur, il est alors possible de définir la frontière intérieure et la frontière extérieure de la façon suivante :

$x \in bo_{in}(F, e) \Leftrightarrow \exists\ n_e \in N(x), \forall\ n \in N(F) : n_e \subset n \wedge n_e \cap in\ (F, e) \neq \varnothing$;

$x \in bo_{ex}(F, e) \Leftrightarrow \exists\ n_e \in N(x), \forall\ n \in N(F) : n_e \subset n \wedge n_e \cap ex\ (F, e) \neq \varnothing$.

La frontière intérieure contient les éléments non typiques, c'est-à-dire les éléments ayant un degré plus faible de typicalité que les éléments typiques de la classe. La frontière extérieure contient les éléments atypiques qui n'héritent pas de toutes les propriétés de la classe (figure 7.14).

Figure 7.14 – **Un concept qui a une frontière à épaisseur**



RÉFÉRENCES

ALCHOURRON, C.E., P. GARDENFORS et D. MAKINSON (1985). « On the Logic of Theory Change : Partial Meet Functions for Contraction and Revision », *Journal of Symbolic Logic*, vol. 50, p. 510-530.

BARAL, C. (2003). *Knowledge Representation, Reasoning and Declarative Problem Solving,* Cambridge, Cambridge University Press.

BREWKA, G, J. DIX et K. KONOLIDGE (1997). « Nonmonotonic Reasoning : An Overview », *Center for the Study of Language and Information, Lecture Notes*, nº 73, Stanford University.

COLMERAUER, A. et P. ROUSSEL (1992). « The Birth of Prolog », dans *The second ACM SIGPLAN Conference on History of programming languages*, p. 37-52.

DESCLÉS, J.P. (1990). *Langages applicatifs, langues naturelles et cognition*, Paris, Hermès.

DESCLÉS, J.-P. et A. PASCU (2005). « Logic of Determination of Objects : The Meaning of Variable in Quantification », dans *Flairs 2005*, édité par AAAI Press, p. 610-616.

FREGE, G. (1893). *Logical Investigations*, dans P. Geach (dir.), Blackwells, 1975.

FREUND, M. *et al.* (2004). « Typicality, Contextual Inferences and Object Determination Logic », *FLAIRS 04*, p. 491-495.

GELFOND, M. (2008). « Answer Sets », dans M. Gelfond, *Handbook of Knowledge Representation*, Londres, Elsevier, p. 285-316.

HANKS, S. et D. McDERMOTT (1987). « Nonmonotonic Logic and Temporal Projection », *Artificial Intelligence,* vol. 33, nº 3, p. 379-412.

JOUIS, C. (2002). « Logic of Relationships », dans C. Jouis, *The Semantics of Relationships*, Dordrecht, Kluwer Academic Publishers, p. 127-140.

KELLEY, J.L. (1975). *General Topology*, Berlin, Springer-Verlag.

KRAUS, S., D. LEHMANN et M. MAGIDOR (1990). « Nonmonotonic Reasoning, Preferential Models, and Cumulative Logics », *Artificial Intelligence,* vol. 44, p. 167-207.

KURATOWSKI, C. (1958). *Topologie*, 2 vol., Varsovie, Panstwowe Wydawnie two Naukowe.

McCARTHY, J. (1980). « Circumscription – A Form of Nonmonotonic Reasoning », *Artificial Intelligence,* vol. 13, p. 27-39.

McCARTHY, J. (1986). « Application of Circumscription to Formalizing Common-sense Knowledge », *Artificial Intelligence,* vol. 28, p. 86-116.

MOORE, R. (1985). « Semantical Considerations on Nonmonotonic Reasoning ». *Artificial Intelligence,* vol. 25, nº 1, p.75-94.

MORGENSTERN, L. (1999). « Nonmonotonic Logics », dans R.A. Wilson et F. Keil (dir.), *MIT Encyclopedia of the Cognitive Sciences*, Cambridge, The MIT Press, <www-formal.stanford.edu/leora/nonmon.pdf>.

MORGENSTERN, L. et R. GUERREIRO. (1993). « Epistemic Logics for Multiple Agent Nonmonotonic Reasoning », dans *Proceedings of the Second Symposium on logical Formalizations of Commonsense Reasoning* (CS-93).

REITER, R. (1978). « On Closed World Data Bases », dans H. Gallaire et J. Minker (dir.), *Logic and Data Bases,* New York, Plenum, p. 119-140.

REITER, R. (1980). « A Logic for Default Reasoning », *Artificial Intelligence,* vol. 13, p. 81-132.

SELMAN, B. et H. LEVESQUE (1993). « The Complexity of Path-based Defeasible Inheritance », *Artificial Intelligence,* vol. 62, nº 2, p. 303-340.

SOMBÉ, L. (1989). *Raisonnements sur des informations incomplètes en intelligence artificielle : comparaison de formalismes à partir d'un exemple,* Éditions Teknea.

WONG, S.K.M., D. WU et C.J. BUTZ (2003). « Probabilistic Reasoning in Bayesian Networks : A Relational Database Approach », *Computer Science*, vol. 2671, p. 583-590.

ZADEH, L.A. *et al.* (1996). *Fuzzy Sets, Fuzzy Logic, Fuzzy Systems*, World Scientific Press.

CHAPITRE 8

COMPARAISON ET COMBINAISON DE DEUX MÉTHODES D'ACQUISITION LEXICALE POUR LA CRÉATION D'ONTOLOGIES MULTILINGUES

Virginie Zampa et Mathieu Lafourcade

RÉSUMÉ

En TAL, il est indispensable de disposer d'informations lexicales, qui sont difficiles à acquérir. Il semble intéressant de combiner des savoirs de spécialités à des connaissances relevant du sens commun, et cela dans un contexte multilingue. Par exemple, le mot « neige » pourra être associé à « froid » et « blanc » comme

caractéristiques relevant du monde, mais également à « travaillée », « poudreuse » ou « fondue » comme phénomène linguistique pour nommer des états possibles. Par ailleurs, relier « neige poudreuse » et « neige fondue » respectivement vers les termes anglais « powder snow » et « slush » permet d'obtenir des ressources multilingues et de conforter les données monolingues. Nous présentons ici deux approches de collecte : l'analyse de la sémantique latente (LSA) et Jeux de mots (JDM). Avec LSA, il suffit de fournir des textes, les relations (proximités sémantiques) non typées entre mots ou textes sont obtenues automatiquement par compilation. JDM est un jeu accessible sur le Web. Les joueurs construisent le réseau lexical, en fournissant des associations (relations typées et pondérées) entre termes. Dans une version multilingue, le terme cible est dans une langue et les propositions doivent être faites dans une autre langue. Nous présenterons ces deux méthodes, leurs utilisations, validations et limites. Puis, nous les comparerons et regarderons PtiClic, né de la combinaison de ces deux approches, dérivé de JDM permettant de renforcer le réseau lexical, jeu accessible sur le Web également.

1. INTRODUCTION

Pour travailler en traitement automatique du langage naturel (TALN), qu'il s'agisse de faire de la correction automatique, de l'indexation de document, etc., il est indispensable d'avoir des informations lexicales. Ainsi, il s'agit de construire une ontologie, de notre point de vue, ensemble de connaissances *a-domaine*, représentant des connaissances à la fois sur le monde et sur la langue. Par exemple, le mot « neige » pourra être associé à « froid » et « blanc » comme caractéristiques relevant du monde, mais également à « lourde » ou « poudreuse » comme phénomène linguistique pour nommer des états possibles. En TALN ces deux aspects sont primordiaux pour la compréhension.

Les informations constituant les ontologies peuvent être collectées et traitées en utilisant différentes méthodes. Nous présentons ici deux approches : l'analyse de la sémantique latente (LSA) et le projet Jeux de mots (JDM).

Avec LSA, il suffit de fournir des textes, les relations entre mots ou textes sont issues de traitements réalisés automatiquement. Le choix du corpus : ce qu'il contient, sa longueur, le type de langage utilisé, etc., est donc primordial. Ainsi, il s'agit d'une analyse automatique permettant d'obtenir des proximités sémantiques entre des mots, entre deux textes, entre un mot et un texte. Ces proximités ont une valeur comprise entre -1 et 1 (valeur expliquée plus bas), mais en aucun cas les relations ne sont typées.

Avec le projet JDM, il s'agit de faire participer un grand nombre de personnes en leur proposant une application ludique accessible sur le Web. Ainsi, à partir d'une base de termes préexistante, ce sont les joueurs qui vont construire le réseau lexical, en fournissant des associations qui ne sont validées que si elles sont proposées par au moins une paire d'utilisateurs. De plus, ces relations typées sont pondérées en fonction du nombre de paires d'utilisateurs qui les ont proposées.

Ces deux méthodes diffèrent donc par de multiples aspects, notamment :

- la constitution du corpus : dans LSA les corpus sont interchangeables (il suffit de modifier la base de textes) alors que dans JDM il est lié aux joueurs ;
- la fabrication des associations entre termes : dans LSA les associations sont issues des contextes dans lesquels les termes apparaissent, alors que dans JDM, il s'agit d'associations libres et spontanées (pouvant potentiellement être fausses) ;
- le niveau de langue utilisé : dans LSA il dépend du corpus retenu mais reste essentiellement soutenu contrairement à JDM dans lequel le niveau de langue varie de soutenu à populaire.

Nous allons donc, dans un premier temps présenter ces deux méthodes, leurs utilisations et validations ainsi que leurs limites. Dans un second temps, nous les comparerons et enfin nous regarderons ce que leur combinaison pourrait nous apporter.

2. DEUX MÉTHODES D'ACQUISITION

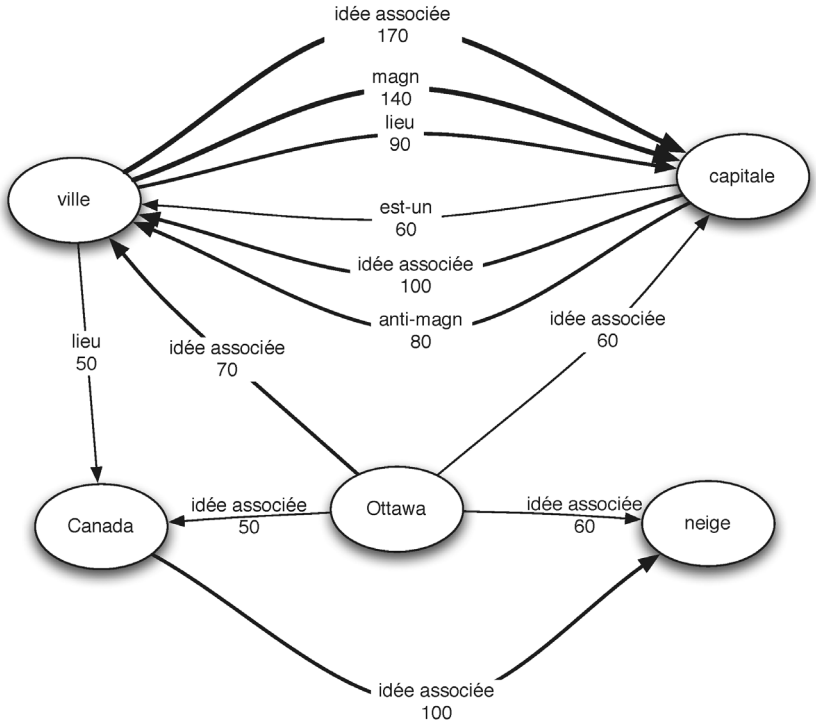
2.1. Jeux de mots

Jeux de mots est un jeu en ligne créé en 2007 (<www.jeuxdemots.org>) qui a pour but la construction d'un réseau lexical. Ce dernier est composé de termes et de relations étiquetées et pondérées. Les relations peuvent être ontologiques (hyper/hyponymie, partie/tout, caractéristiques typiques, etc.), lexicales (synonymes, antonymes, locutions, etc.) ou libres (dans la figure 8.1, voir les relations les plus actives pour le terme « Ottawa » dans le réseau lexical de JDM).

Ainsi, la structure du réseau lexical que nous cherchons à obtenir se fonde sur les notions de nœuds et de relations entre nœuds, comme présentées ci-dessus (Polguère, 2006). D'une façon générale, chaque nœud du réseau est constitué d'une unité lexicale (terme ou expression) regroupant toutes ses lexies et les relations entre nœuds traduisent des fonctions lexicales, telles qu'elles sont présentées par Mel'čuk *et al.* (1995).

Ce type de réseau lexical constitue une ontologie. En effet, il comporte des objets faisant référence à des objets du monde ou à des objets linguistiques. Les relations entre ces objets sont explicitées comme nous venons de le voir ci-dessus. De plus, un point important de notre approche est l'ajout d'une pondération sur chaque relation traduisant implicitement son intensité. Si on le désire, il est ainsi possible à partir du réseau lexical de filtrer les relations en fonction de leur intensité ou de leur type, et ainsi d'obtenir une taxonomie (seules les relations de classes et appartenances sont conservées).

Figure 8.1 – **Mots associés à « Ottawa » dans JDM avec les relations et les poids**



2.1.1. Présentation de la méthode

En pratique, les validations des propositions sont faites par concordance entre paires de joueurs. Ce processus de validation rappelle celui utilisé par von Ahn et Dabbish (2004) pour l'indexation d'images ou plus récemment par Lieberman *et al.* (2007) pour la collecte de « connaissances de bon sens ». Autant que nous sachions, il n'a jamais été mis en œuvre dans le domaine des réseaux lexicaux.

Une partie se déroule entre deux joueurs, en asynchrone, et est fondée sur la concordance de leurs propositions. Lorsqu'un premier joueur que nous appellerons A entame une partie, une consigne concernant un type de compétence est affichée (synonymes, contraires, domaines, etc.), ainsi qu'un terme T tiré aléatoirement dans une base de mots. Par exemple il est demandé au joueur de donner des « idées associées » au terme « télésiège », comme le montre la copie d'écran ci-dessous (figure 8.2). Ce joueur A a alors un temps limité (60 secondes) pour répondre en donnant des propositions correspondant, selon lui, à la consigne appliquée au terme T.

Figure 8.2 – **Partie Jeux de mots en cours**

Remarque : les mots à droite (« neige », « ski », « piste », etc.) correspondent aux mots que l'utilisateur vient de rentrer.

Ce même mot, avec cette même consigne, est proposé par la suite à un autre joueur que nous appellerons B ; le processus est identique. Afin d'accroître l'aspect ludique, pour toute réponse commune dans les propositions de A et B, ces deux joueurs gagnent un certain nombre de points (figure 8.3).

Figure 8.3 – **Résultat de la partie**

Pour le terme cible T, nous mémorisons les réponses communes aux joueurs A et B. Nous ne mémorisons pas les réponses proposées uniquement par l'un des deux joueurs. Cela permet la construction d'un réseau lexical reliant les termes par des relations typées et pondérées, validées par paires de joueurs. Ces relations sont typées par la consigne imposée aux joueurs ; elles sont pondérées en fonction du nombre de paires de joueurs qui les ont proposées. Initialement, les nœuds sont constitués des termes de notre base de départ, mais celle-ci peut s'accroître ; effectivement, si les deux joueurs A et B d'une même partie proposent un terme initialement inconnu, alors ce terme est ajouté à notre base.

2.1.2. Limites

Malgré tout, cette approche comporte certaines limites.

Premièrement, il s'agit d'un vocabulaire actif, forcé par la consigne. Le vocabulaire est dit actif au sens où il fait partie du discours de tous les jours du joueur. Une grande partie du vocabulaire passif (connu mais non régulièrement usité) reste donc non proposée. De plus, il est forcé par la consigne, car le joueur donne ses réponses en fonction de ce qui lui est demandé et non ce qui lui vient en premier à l'esprit à part dans le cas de la consigne « association libre ».

Deuxièmement, certains joueurs ont recours à des ressources externes telles que Wikipédia ou des dictionnaires de synonymes. Cela introduit un biais car il ne s'agit pas du vocabulaire actif, mais cela permet aussi d'introduire dans le réseau du vocabulaire plus soutenu. Ceci compense partiellement la limite précédente.

Troisièmement, les associations sont fortement liées à l'actualité. Le réseau évolue ainsi avec le temps et est représentatif des relations sémantiques à un instant t. Par exemple, actuellement, le terme « Amérique » sera très fortement en association avec « Obama » ce qui n'était pas forcément le cas il y a quelques temps.

Enfin, les connaissances sont issues d'un échantillon non représentatif de la population. En effet, toutes les tranches d'âge ne sont pas représentées, ni toutes les classes sociales.

2.2. L'analyse sémantique latente

LSA (*Latent Semantic Analysis*) a été brevetée en 1988 et publiée et 1990.

2.2.1. Présentation de la méthode

Le but de LSA est de représenter dans un espace multidimensionnel (300 dimensions) les mots de la langue. Grâce à une analyse statistique, le sens de chaque mot est caractérisé par un vecteur. La proximité de sens entre deux mots correspond à la proximité entre les vecteurs.

Pour construire cet espace, LSA prend un ensemble de textes en entrée et construit une matrice d'occurrences qui est réduite par le biais d'une analyse statistique. Ceci permet de faire ressortir les relations sémantiques entre mots ou entre textes.

Grâce à cette méthode, deux mots peuvent être considérés sémantiquement proches même s'ils n'apparaissent jamais conjointement dans un texte. Il suffit qu'ils soient utilisés dans des contextes similaires. Le contexte d'un mot est ici défini comme l'ensemble des mots qui apparaissent conjointement avec

lui. Ainsi, les mots *vélo* et *bicyclette* sont considérés comme sémantiquement proches car ils apparaissent tous deux avec des mots comme *randonnée*, *guidon*, *pédaler*, etc., et ils n'apparaissent qu'occasionnellement avec des mots comme *bouillir*, *imprimante*, *vase*, etc.

Cette notion de cooccurrence est statistique : la méthode fonctionne si un nombre suffisant de textes est utilisé. Il ne s'agit pas simplement d'un comptage d'occurrences, il faut aussi disposer d'une procédure pour établir les liaisons entre mots. Cette procédure est la réduction de la matrice d'occurrences. Le principe est donc le suivant. Dans un premier temps LSA construit la matrice d'occurrences. Il s'agit d'une matrice dont les lignes sont des unités textuelles (l'unité généralement utilisée est le paragraphe) et les colonnes, des mots. L'élément *(i,j)* de la matrice correspond ainsi au nombre d'occurrences du mot *j* dans le paragraphe *i*. L'étape suivante va consister à réduire le nombre de ces dimensions à environ 200. Ce nombre est important car une réduction à un espace de trop grande dimension ne ferait pas suffisamment émerger les liaisons sémantiques entre mots que nous recherchons, et un très petit nombre de dimensions conduirait à une trop grande perte d'informations. Ce nombre adéquat de dimensions est issu de tests empiriques, dans le cas de l'anglais il se situe entre 100 et 300 (Deerwester *et al.*, 1990). Cette réduction des dimensions est réalisée grâce à une décomposition aux valeurs singulières. La réduction à *n* dimensions va consister à ne conserver que les *n* premières de ces valeurs pour reconstituer une matrice approchée, de dimensions n^2. Chaque mot et chaque paragraphe, traité de la même façon dans cette procédure, est ainsi représenté par un vecteur à *n* dimensions.

L'espace sémantique étant construit, il faut choisir la façon de mesurer la proximité entre deux éléments. Les tests empiriques réalisés ont privilégié la méthode du cosinus : la similarité entre deux vecteurs est le cosinus de leur angle qui n'est rien d'autre que leur produit scalaire (pour des vecteurs normés). Cette similarité entre deux vecteurs correspond à la corrélation entre les deux objets (mots, paragraphes) associés aux vecteurs, il s'agit donc d'une valeur entre –1 et 1. L'hypothèse forte du modèle LSA est que quand cette corrélation est positive, elle traduit une similarité sémantique entre les objets liés aux vecteurs. En particulier deux mots dont la corrélation est proche de 1 seront considérés comme sémantiquement proches.

2.2.2. Utilisations, validations, limites

LSA a déjà été utilisé et validé dans différents domaines tels que :

– la recherche d'information. LSA permet de limiter les problèmes de choix de mots clés liés à la synonymie, à la polysémie et à l'inflexion. La recherche se fait sur le sens des mots clés et non uniquement sur leur « forme » (Dumais, 1994 ; Dumais, 1997) ;

- l'acquisition de connaissances. Ces acquisitions peuvent concerner les langues (Landauer et Dumais, 1997 ; Redington et Chater, 1998), les jeux tels que le tic-tac-toe (Lemaire, 1998) ou kalah (Lemaire, 1999) ;
- les EIAH (environnement informatique pour l'apprentissage humain). Dans ces environnements, LSA a été utilisé pour modéliser les connaissances, pour évaluer des copies ou résumés, etc. (Dessus *et al.*, 2000 ; Zampa et Lemaire, 2002 ; Zampa et Raby, 2001 ; van Bruggen *et al.*, 2004 ; Dikli, 2006) ;
- les modélisations cognitives lors de la compréhension des métaphores (Lemaire et Bianco, 2003) ou lors de la rédaction de résumés (Lemaire *et al.*, 2005) ;
- la mesure de la cohérence textuelle (Miller, 2003) ;
- la compréhension de texte (Kintsch, 2000 ; Kintsch, 2001 ; Kintsch, 2002) ;
- la modélisation de la mémoire (Kintsch *et al.*, 1999 ; Denhière et Lemaire, 2004 ; Howard et Kahanna, 2002 ; Howard et Kahanna, 2007).

Mais, même si LSA a été utilisé et validé dans de nombreux domaines et provoque actuellement un certain engouement, il montre certaines limites.

Tout d'abord les connaissances sur les mots dépendent fortement du corpus utilisé pour construire l'espace de connaissance. Par exemple, si le corpus ne contient que des textes issus d'un journal tel que *Le Monde diplomatique*, le vocabulaire, relativement soutenu, ne contiendra pas un certain nombre de mots couramment usités.

De plus, la nature de la relation entre les mots n'est pas spécifiée, ce qui peut poser problème selon l'utilisation voulue. Il peut s'agir d'un synonyme, d'un mot du même domaine, d'une partie, d'une spécification, etc. Par exemple, dans les mots les plus proches de « salade » se trouvent des mots tels que « laitue » (qui est un hyponyme), « vinaigrette » et « persil » (qui sont thématiquement liés), « soupe » (thématiquement lié ou co-hyponyme de « plat »).

Enfin la catégorie morphosyntaxique n'est pas fournie ni traitée ; ainsi *porte* (substantif) et *porte* (verbe conjugué) sont considérés comme le même mot : seule la graphie compte. De ce fait, il n'y a qu'un seul vecteur pour représenter ces deux termes dans l'espace sémantique, les proximités sémantiques de chaque terme pouvant être du bruit pour l'autre.

2.3. Comparaison

Dans cette section, nous comparons pour un mot cible donné trois types de résultats :

- les associations produites par LSA sur un « gros » corpus (ce corpus contient 101 123 graphies différentes et 189 726 paragraphes) ;
- les associations produites par LSA sur une sous-partie du réseau de JDM (10 000 mots cibles les plus fréquents et leurs relations) ;
- les associations issues du réseau JDM.

Nous allons ici (tableau 8.1) nous appuyer sur des exemples précis : tout d'abord le mot « examen ».

Tableau 8.1 – **Comparaison des associations obtenues par LSA et JDM pour le terme *examen***

	LSA				JDM	
	Gros corpus		Sous-partie JDM			
rang	*mot*	*proximité*	*mot*	*proximité*	*mot*	*activation*
1	examen	1,000	examen	1,000	concours	150
2	recel	0,928	bepc	0,988	baccalauréat	140 + 90
3	abus	0,893	examen final	0,984	test	120
4	écroué	0,886	examen blanc	0,984	passer un examen	110
5	escroquerie	0,856	réussir un examen	0,984	note	100
6	factures	0,841	examen médical	0,984	examiner	90
7	blaes	0,841	passer un examen	0,984	épreuve	90
8	courroye	0,829	évaluation	0,876	contrôle	80
9	méry	0,819	études supérieures	0,868	diplôme	80
10	complicité	0,813	fin d'étude	0,868	médical	80
11	supplétif	0,809	études de droit	0,868	médecine	80
12	filippini	0,805	diplôme de fin	0,868	université	80
13	instruction	0,801	longues études	0,868	examen médical	80
14	varces	0,800	études longues	0,868	bac	70
15	incarcéré	0,787	faire des études	0,866	école	70
16	fossorier-longuet	0,784	études	0,865	études	70
17	juge	0,783	brevet	0,865	évaluation	70
18	mis	0,783	diplôme	0,827	santé	70
19	névache	0,774	certificat	0,826	éducation	70
20	escroqueries	0,770	master	0,808	brevet	70

Pour commencer, nous pouvons constater l'influence du corpus. En effet, le « gros corpus » contient entre autres, une année du *Monde*, de ce fait le terme examen renvoie au mot composé « mise en examen » et non à examen au sens de *concours* ou *évaluation*. Les mots les plus proches renvoient à plusieurs noms propres (Blaes, Courroye, Méri, Filippini, Varces, Fossorier-Longuet, Névache) qui ont marqué l'actualité cette année-là. Dans le cadre de la constitution d'une ontologie, ces termes sont moyennement pertinents. Il s'agit ainsi d'un exemple typique des biais liés à un corpus d'actualité. Avec un corpus différent, contenant par exemple, les journaux officiels et bulletins officiels relatifs au ministère de l'enseignement, les mots les plus proches sémantiquement de « examen » auraient sans doute été tout autre. Définir un corpus idéal (représentatif des connaissances d'un humain) reste utopique et même un corpus équilibré (textes de niveaux de vocabulaire différents, traitant de sujets différents, etc.) reste particulièrement

délicat. Il faudrait être capable d'évaluer ce à quoi les gens sont exposés (discours et textes) depuis leur naissance et d'établir un profil d'un individu moyen (sachant que ce dernier n'existe sans doute pas).

Dans le cas de JDM, le corpus est d'une certaine manière constitué par l'ensemble des joueurs. De ce fait, il représente aussi un biais car les joueurs ne sont certainement pas représentatifs de la population. Tout d'abord, puisqu'il s'agit d'un jeu, le public est majoritairement constitué de jeunes adultes. De plus, il s'agit d'un projet de recherche présenté lors de conférences et touchant un public essentiellement universitaire. Ce qui explique les résultats présentés avec le mot « examen ». Prenons maintenant le terme « sida » (tableau 8.2).

Tableau 8.2 – **Mots les plus proches de « sida » et leur proximité dans JDM et LSA**

	LSA			
	Gros corpus		**Sous-partie JDM**	
rang	***mot***	***proximité***	***mot***	***proximité***
1	sida	1,000	sida	1,000
2	virus	0,948	mst	0,952
3	infection	0,900	vih	0,925
4	contamination	0,888	hiv	0,915
5	immunodéficience	0,880	maladie	0,861
6	sexuellement	0,877	contagion	0,854
7	transmissibles	0,872	maladie infantile	0,850
8	dépistage	0,856	maladie grave	0,850
9	infections	0,855	infantile	0,850
10	vih	0,849	varicelle	0,850
11	infectées	0,844	maladie mortelle	0,849
12	infectieuses	0,841	maladie bénigne	0,849
13	vaccin	0,837	hépatite	0,849
14	anti-vih	0,836	choléra	0,848
15	contaminée	0,831	quarantaine	0,847
16	immuno	0,828	infection	0,844
17	contaminé	0,825	rougeole	0,844
18	opportunistes	0,822	incurable	0,842
19	séropositif	0,813	peste	0,842
20	contaminés	0,803	bénigne	0,841

D'une façon générale, le vocabulaire obtenu par LSA semble plus riche que celui acquis via Jeux de mots. Par exemple, un terme comme « immunodéficience » apparaît en 5e position et n'existe pas dans le réseau lexical de JDM.

En effet, par l'intermédiaire du jeu, les connaissances acquises sont des associations plus immédiates ; cela est lié au fait que les joueurs ne sont pas experts du domaine et que le temps est limité. Certaines associations, évidentes, apparaissent très rapidement dans le réseau lexical du jeu. C'est le cas par exemple de l'association sida-maladie qui apparaît en 5e position dans le sous-corpus. Mais cette association est beaucoup plus faible dans le gros corpus (proximité de 0,711) car elle est trop évidente pour être donnée explicitement dans les textes. Ce phénomène semble d'autant plus marqué que le terme cible est spécifique.

Toutefois, dans LSA, la segmentation des textes est faite à partir des caractères de séparation (blanc, virgules, etc.) ce qui interdit l'apparition de termes composés. Ce n'est évidemment pas le cas dans Jeux de mots où les joueurs ont toute liberté dans le choix des termes qu'ils suggèrent. L'ajout d'un prétraitement dans LSA qui consisterait à repérer les termes composés pose plusieurs difficultés, la principale est qu'il faut alors disposer d'une telle liste de termes, or c'est justement ce que nous cherchons à faire ici. Les repérer automatiquement, par des moyens statistiques, avec suffisamment de précision reste difficile. Enfin, un dernier point concernant les biais du réseau issu de JDM réside dans le fait que la connaissance acquise par le système est celle des joueurs. Il est donc possible d'obtenir des connaissances erronées.

3. COMBINAISON DES APPROCHES : PTICLIC

Nous avons montré précédemment que LSA tout comme JDM présentent des biais et ne proposent qu'une couverture partielle du vocabulaire. De plus, ces méthodes ont chacune des parts distinctes de bruits (association qui devrait être plus faible) et de silence (association qui devrait être plus forte). Enfin, dans notre ontologie nous avons besoin que les relations soient typées.

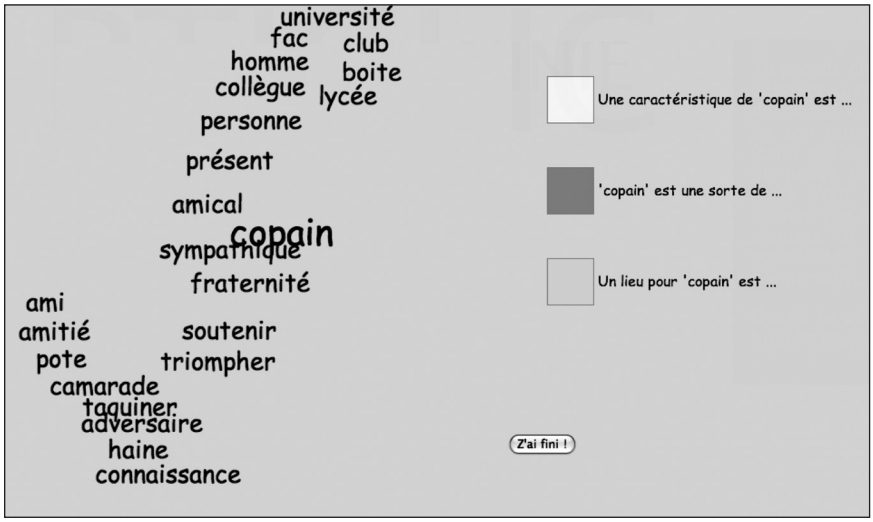
Nous avons donc décidé de combiner les deux méthodes afin que chacune compense les défauts de l'autre, ceci avec PtiClic (<www.lirmm.fr/pticlic/pticlic.php>), dont on peut voir un exemple à la figure 8.4.

3.1. PtiClic vu par le joueur

La consigne donnée au joueur est la suivante :

Tu prends chaque mot et tu le déposes gentiment sur une des zones visibles, en fonction de son rapport au mot cible (celui qui se trouve au milieu du nuage de mots). Certains mots n'ont rien à voir et ne doivent pas être déposés. Pour d'autres il y a plusieurs possibilités, en choisir une bonne suffit. Quand tu as fini, clique sur le bouton en bas.

Figure 8.4 – **Partie PtiClic en cours**



3.2. PtiClic vu par les chercheurs-concepteurs

3.2.1. Sélection du mot source par Jeux de mots et des mots cibles par LSA

Le principe est le suivant : un mot source est sélectionné aléatoirement dans la base de Jeux de mots. Si le mot n'est pas connu par LSA, un autre mot est choisi et ainsi de suite jusqu'à obtenir un mot connu. De ce fait, un mot inconnu de LSA ne sera jamais proposé dans PtiClic. À partir de ce mot source, LSA sélectionne les cinq à vingt mots les plus proches (après un léger traitement : par exemple le mot au pluriel n'est pas retenu) que le joueur doit placer dans une à quatre relations.

3.2.2. Les différentes relations présentes dans JDM

Dans Jeux de mots il existe 27 types de relations qui peuvent être renseignées par les joueurs :

1. idées associées : il s'agit de la relation la plus simple, elle peut contenir toutes les autres relations (p. ex., chat → chien, souris, rat, miauler, lait, griffer, etc.) ;
2. raffinement (p. ex., blaireau → animal /barbier/pauvre type) ;
3. domaine : un domaine auquel peut appartenir le terme (p. ex., service → tennis, restaurant, administration, etc.) ;
4. synonyme (p. ex., vélo → bicyclette, etc.) ;

5. hyperonyme (p. ex., chat → animal, félin, mammifère, animal de compagnie) ;
6. antonyme (p. ex., chaud → froid, glacé, etc.) ;
7. hyponyme (p. ex., animal → lion, chien, vache, etc.) ;
8. partie de (p. ex., cheval → crinière, robe, sabot, etc.) ; remarque : les réponses les plus fournies correspondent régulièrement aux caractéristiques. ;
9. le tout (p. ex., pied → homme, champignon, chaise, etc.) ;
10. relation de locution (p. ex., hallebarde → pleuvoir des hallebardes) ;
11. agent (p. ex., mordre → chien, animal, enfant, etc.) ;
12. patient (p. ex., lire → livre, article, journal, revue, bouquin, etc.) ;
13. lieu (p. ex., où peut-on trouver un ordinateur ? →bureau, administration, etc.) ;
14. instrument (p. ex., que peut-on utiliser pour écrire ? →stylo, plume, crayon, ordinateur, etc.) ;
15. caractéristiques typiques du mot (p. ex., soleil → chaud, rond, rouge, jaune, etc.) ;
16. magne (plus intense que) (p. ex., 1 – qu'est-ce qui est plus intense qu'anguille sous roche ? →cachalot sous gravillon, cachalot sous grain de sable, etc. ; 2 – qu'est-ce qui est plus intense que fièvre ? : grosse fièvre, fièvre de cheval, etc.) ;
17. anti-magne (moins intense que) (p. ex., qu'est-ce qui est moins intense que brûlant ? : tiède, etc.) ;
18. famille (p. ex., chat → chatte, chatterie, chatière, etc.) ;
19. relation inverse de caractéristique : la caractéristique est donnée il faut trouver les termes ayant cette caractéristique (p. ex., rond → ballon, soleil, roue, manège, etc.) ;
20. relation inverse de agent : que peut faire cet agent ? (p. ex., chien → mordre, aboyer, japper, lécher, etc.) ;
21. relation inverse d'instrument : que peut-on faire avec cet instrument ? (p. ex., stylo → écrire, dessiner, griffonner, prendre des notes, etc.) ;
22. relation inverse de lieu : que peut-on trouver dans ce lieu ? (p. ex., restaurant → menus, serveur, plats, apéro, table, chaise, additions, commande, nourriture, etc.) ;
23. lieu_action : dans quel lieu peut-on faire l'action X ? (p. ex., danser → gymnase, opéra, salle de spectacle, dehors, boîte de nuit, discothèque, etc.)
24. action_lieu : quelles actions peuvent être faites dans le lieu X ? (p. ex., salle de spectacle → danser, chanter, jouer, mettre en scène, regarder, écouter, etc.)
25. sentiment : quel sentiment est évoqué par ce terme ? (p. ex., cadeau → joie, bonheur, surprise, etc.)
26. manière : de quelle manière ? (pour un verbe) (p. ex., manger → rapidement, lentement, salement, goulûment, etc.)
27. sens : quel est le sens de … ? (p. ex., tour → pièce d'échec, construction, tour de magie, vacherie, périmètre, etc.)

3.2.3. Les relations de PtiClic

Dans PtiClic, seulement dix de ces relations sont gardées afin que le jeu ne soit pas trop compliqué. Les dix relations maintenues sont : idées associées, synonyme, agent, patient, instrument, lieu, partie de, tout, antonyme, hyponyme.

PtiClic permet ainsi de typer les relations sur des mots qui sont dans un vocabulaire « moins actif », plus précis, que celui de Jeux de mots.

Il est nécessaire de typer les relations avec ce vocabulaire, mais il ne faudrait pas que seule la relation idées associées soit proposée par LSA. Ceci n'est pas le cas. Par exemple, si nous prenons le mot source « ordinateur », nous constatons que onze des vingt-sept relations de Jeux de mots sont couvertes, dont six des dix de PtiClic, avec seulement les 13 mots les plus proches du mot source (tableau 8.3).

Tableau 8.3 – **Mots les plus proches de « ordinateur » et leur relation**

Types de relations	Mots (rang de proximité)
partie de / lieu	logiciel (3), entrée-sortie (5), processeur (7)
idées associées	tout dont ordinateurs (2) et utilisateur (4)
lieu du prédicat / instrument	stocker (6)
lieu, partie de	électronique (8)
caractéristique	numériques (9)
hyponyme	micro-ordinateur (13)
domaine	informatique (14)

Les quatre relations de PtiClic absentes avec *ordinateur* (synonyme, agent, patient, antonyme) sont par contre présentes dans les 8 mots les plus proche de *voler* (tableau 8.4).

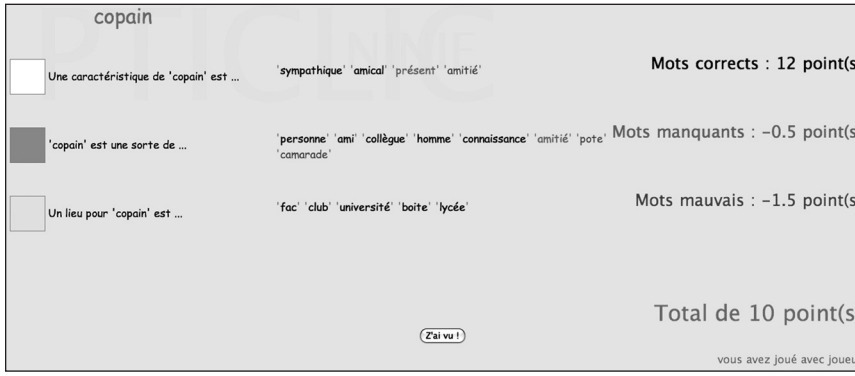
Tableau 8.4 – **Mots les plus proches de « voler » et leur relation**

Types de relations	Mots (rang de proximité)
agent	voleur (2)
synonyme	enlever (4)
antonyme	trouver (7)
patient	sac (8)

3.2.4. Comptage des points et validation dans la BD

Comme nous l'avons déjà montré, le joueur, doit placer les mots cibles qui conviennent dans les catégories proposées. Lorsque deux joueurs ont eu la même partie, le résultat est affiché (figure 8.5) et les points gagnés sont calculés par comparaison entre les accords, différences et oublis. Dans la copie d'écran ci-dessous, les mots en vert correspondent à l'accord entre les joueurs, ils apportent chacun 1 point. Les mots en gris sont ceux qui ont été mis par le joueur 1 mais pas par le joueur 2 ; les mots en rouge sont ceux mis par le joueur 2 mais pas par le joueur 1. Dans les deux cas, chaque mot manquant fait reculer le score de 0,5.

Figure 8.5 – **Résultat de la partie PtiClic**

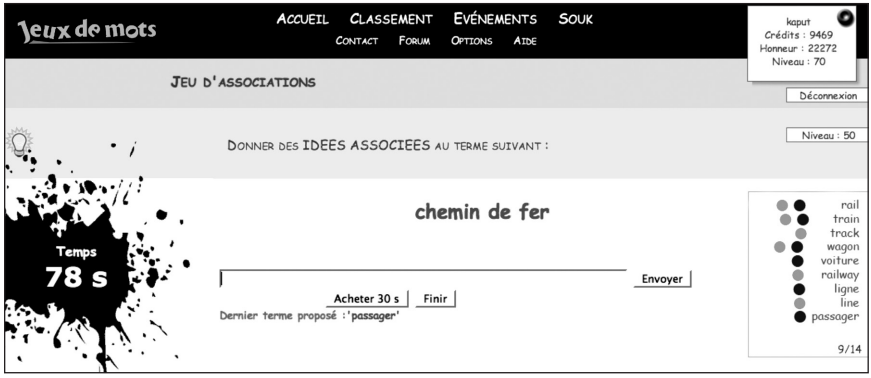


Tout comme dans Jeux de mots, la relation n'est validée dans la base qu'après accord entre paires d'utilisateurs. PtiClic est ainsi composante de JDM agissant sur le même réseau lexical. Contrairement à ce dernier, c'est un jeu en monde clos pour les utilisateurs (le joueur sélectionne parmi des propositions mais ne peut en faire). Ce choix de conception permet d'obtenir des associations sur des termes relevant du vocabulaire passif sélectionnés par LSA, termes qui n'auraient pas spontanément été proposés par les joueurs. L'ajout de PtiClic dans Jeux de mots permet de réduire le bruit (des termes mal orthographiés ou des confusions de sens) ainsi que le silence (de nouveaux termes sont introduits grâce à LSA). PtiClic permet donc de consolider les relations de la base et de densifier le réseau lexical.

4. CONCLUSION… VERS UNE VERSION MULTILINGUE DE JEUX DE MOTS ET PLUS TARD DE PTICLIC !

Actuellement, il est possible de jouer dans plusieurs langues à Jeux de mots (français, anglais, espagnol, japonais, thaï) mais chaque jeu est monolingue (figure 8.6). Il serait intéressant de faire un jeu multilingue pour plusieurs raisons. La première est qu'en comparant les associations obtenues dans chaque langue pour un mot donné, il est possible d'identifier de façon automatique les phénomènes contrastifs. Ensuite, un réseau croisé entre plusieurs langues permet également de renforcer les associations pour chacune des langues. En effet, les relations dans une langue ont tendance à corroborer celles des autres langues. Le scénario utilisateur proposé serait le suivant. Chaque joueur choisirait les langues dans lesquelles il voudrait jouer en fonction du mot proposé. Il pourrait ainsi, quelle que soit la langue du mot source, fournir les réponses dans les langues qu'il a sélectionnées. Le jeu identifierait automatiquement la ou les langues d'appartenance et il serait toujours possible au joueur d'affiner sa réponse en supprimant une langue si une même graphie appartient à plusieurs langues mais ne convient pas dans toutes.

Figure 8.6 – **Partie en cours avec Jeux de mots multilingue**



RÉFÉRENCES

DEERWESTER, S., S.T. DUMAIS, G.W. FURNAS, T.K. LANDAUER et R. HARSHMAN (1990). « Indexing by Latent Semantic Analysis », *Journal of the American Society for Information Science*, vol. 41, nº 6, p. 391-407.

DENHIÈRE, G. et B. LEMAIRE (2004). « A Computational Model of a Child Semantic Memory », *Proc. 26th Annual Meeting of the Cognitive Science Society (CogSci'2004)*, p. 297-302.

DESSUS, P., B. LEMAIRE et A. VERNIER (2000). « Free-text Assessment in a Virtual Campus », dans K. Zreik (dir.), *Proc. Third International Conference on Human System Learning (CAPS'3)*, Paris, Europia, p. 61-76.

DIKLI, S. (2006). « An Overview of Automated Scoring of Essays », *The Journal of Technology, Learning and Assessment*, vol. 5, nº 1.

DUMAIS, S.T. (1994). « Latent Semantic Indexing (LSI) and TREC-2 », dans D. Harman (dir.), *The Second Text RE-trieval Conference (TREC2)*, National Institute of Standards and Technology Special Publication, vol. 500, nº 215, p.105-116.

DUMAIS, S.T. (1997). *Using Latent Semantic Indexing for Information Retrieval, Information Filtering and Other Things*, Cognitive Technology Conference.

HOWARD, M.W. et M.J. KAHANNA (2002). « When does Semantic Similarity Help Episodic Retrieval? », *Journal of Memory and Language,* vol. 46, p. 85-98.

HOWARD, M.W. et M.J. KAHANNA (2007). « Semantic Structure and Episodic Memory », dans T.K. Landauer *et al.* (dir.). *Handbook of Latent Semantic Analysis*, Mahwah, Lawrence Erlbaum Associates.

KINTSCH, W. (2000). « Metaphor Comprehension : A Computational Theory », *Psychonomic Bulletin & Review,* vol. 7, nº 2, p. 257-266.

KINTSCH, W. (2001). « Predication », *Cognitive Science*, vol. 25, nº 2, p. 173-202.

KINTSCH, W. (2002). « On the Notions of Theme and Topic in Psychological Process Models of Text Comprehension », dans M. Louwerse et W. van Peer (dir.), *Thematics : Interdisciplinary Studies*, Amsterdam, Benjamins, p. 151-170.

KINTSCH, W., V.L. PATEL et K.A. ERICSSON (1999). « The Role of Long-term Working Memory in Text Comprehension », *Psychologia,* vol. 42, p. 186-198.

KIPFER, B.A. (2001). *Roget's International Thesaurus*, (6e éd.), Londres, Harper Resource.

LANDAUER, T. K. et S.T. DUMAIS (1997). « A Solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge », *Psychological Review*, nº 104, p. 211-240.

LAPATA, M. et F. KELLER (2005). « Web-based Models for Natural Language Processing », *ACM Transactions on Speech and Language Processing*, vol. 2, nº 1, p. 1-30.

LEMAIRE, B. (1998). « Models of High-dimensional Semantic Spaces », *Proc. 4th Int. Workshop on Multistrategy Learning (MSL 98)*, Desenzano, Italie.

LEMAIRE, B. (1999). « Tutoring Systems Based on Latent Semantic Analysis », dans S. Lajoie et M. Vivet (dir.), *Artificial Intelligence in Education*, Amsterdam, IOS Press, p. 527-534.

LEMAIRE, B. et M. BIANCO (2003). « Contextual Effects on Metaphor Comprehension : Experiment and Simulation », *Proc. of the 5th International Conference on Cognitive Modeling (ICCM'2003)*, Bamberg, Germany.

LEMAIRE, B. *et al.* (2005). « Computational Cognitive Models of Summarization Assessment Skills », dans *Proceedings of the 27th Annual Meeting of the Cognitive Science Society (CogSci' 2005)*, Stresa, Italy, juillet 21-23, p. 1266-1271.

LIEBERMAN H., D.A. SMITH et A. TEETERS (2007). « Common Consensus : A Web-based Game for Collecting Commonsense Goals », *International Conference on Intelligent User Interfaces (IUI'07)*, Hawaii, USA.

MEL'ČUK, I.A., A. CLAS et A. POLGUÈRE (1995). *Introduction à la lexicologie explicative et combinatoire*, Paris, Éditions Duculot AUPELF-UREF.

MILLER, G.A. *et al.* (1990). « Introduction to WordNet : An On-line Lexical Database », *International Journal of Lexicography,* vol. 3, nº 4, p. 235-244.

MILLER, T. (2003). « Latent Semantic Analysis and the Construction of Coherent Extracts », dans G. Angelova *et al.* (dir.), *Proc. of the International Conference RANLP-2003 (Recent Advances in Natural Language Processing)*, septembre, p. 270-277.

POLGUÈRE, A. (2006). « Structural Properties of Lexical Systems : Monolingual and Multilingual Perspectives », *Proceedings of the Workshop on Multilingual Language Resources and Interoperability (COLING/ACL 2006)*, Sydney, p. 50-59.

REDINGTON, M. et N. CHATER (1998). « Connectionist and Statistical Approaches to Language Acquisition : A Distributional Perspective », *Language and Cognitive Processes*, vol. 13, nos 2/3, p. 29-91.

ROBERTSON, S. et K.S. JONES (1976). « Relevance Weighting of Search Terms », *Journal of the American Society for Information Science*, no 27, p. 129-146.

SALTON, G. (1968). *Automatic Information Organization and Retrieval*, New York, McGraw-Hill.

van BRUGGEN, J. *et al.* (2004). « Latent Semantic Analysis as a Tool for Learner Positioning in Learning Networks for Lifelong Learning », *British Journal of Educational Technology*, vol. 35, no 6, p. 729-738.

VÉRONIS, J. (2001). « Sense Tagging : Does It Make Sense? », *Corpus linguistics' 2001 Conference*, Lancaster, Royaume-Uni.

von AHN, L. et L. DABBISH (2004). « Labelling Images with a Computer Game », *ACM Conference on Human Factors in Computing Systems (CHI)*, p. 319-326.

ZAMPA, V. et B. LEMAIRE (2002). « Latent Semantic Analysis for user modeling », *Journal of Intelligent Information Systems*, vol. 18, no 1, p. 15-30.

ZAMPA, V. et F. RABY (2001). « Entre modèle d'acquisition et outil pour l'apprentissage de la langue de spécialité : Le prototype RAFALES (Recueil automatique favorisant l'acquisition d'une langue étrangère de spécialité) », *Asp (Anglais de spécialité)*, nos 31-33, p. 163-179.

CHAPITRE 9

EXXELANT ET MIRTO
Deux exemples d'environnement d'ALAO intégrant des outils TAL

Georges Antoniadis, Claude Ponton et Virginie Zampa

RÉSUMÉ

L'élaboration d'environnements informatiques pour l'apprentissage des langues est l'un des aspects abordés par l'ALAO. La prise en compte par ces environnements des propriétés de la langue, aussi bien aux plans lexical, syntaxique que sémantique, objets de l'apprentissage, a conduit à partir des années 1980 à l'utilisation de procédures et d'outils du traitement automatique de la langue (TAL). L'euphorie du départ a laissé progressivement la place à une intégration mesurée du TAL dans les systèmes d'ALAO, en tenant compte de ses possibilités, de ses limites mais également de la plus-value pédagogique possible.

À travers la description de deux de nos outils, nous présentons dans cet article notre approche de cette problématique de l'intégration du TAL dans les systèmes d'ALAO ainsi que l'apport du TAL pour ces systèmes. Ces deux produits sont le logiciel EXXELANT et la plateforme MIRTO. Nous montrons que le TAL peut et doit être au centre de tout système d'ALAO à condition que son utilisation soit en adéquation avec la pertinence actuelle de ses résultats.

1. TAL ET ALAO

Les premières tentatives d'utilisation du traitement automatique de la langue (TAL) pour l'apprentissage des langues assisté par ordinateur (ALAO) datent des années 1980 avec un pic d'activité durant la période 1985-1995 attesté par bon nombre de symposiums et de travaux européens (Swartz et Yazdani, 1992), nord-américains (Holland *et al.*, 1995) ou français (Chanier *et al.*, 1993). Toutefois, la problématique apprentissage/acquisition des langues secondes (non maternelles) est souvent absente, ce qui empêche tout travail en collaboration entre informaticiens/spécialistes du TAL et didacticiens des langues / pédagogues. Durant la fin des années 1990, cette tendance commence à être corrigée ; des rencontres interdisciplinaires se déroulent et des travaux en collaboration voient le jour. Les systèmes d'ALAO avec lesquels l'apprenant interagit s'enrichissent de quatre composantes : une première gérant les connaissances linguistiques de référence, une autre le dialogue homme/machine, une troisième les connaissances de l'apprenant et/ou ses erreurs et une quatrième relative aux interventions didactiques du système.

Dans ce contexte, les réalisations sont mesurées et les avancées se font à petits pas. Tous les aspects des systèmes d'ALAO sont « revisités », redéfinis, confrontés aux exigences et aux réalités de l'apprentissage des langues. Les travaux actuels (les nôtres compris) essaient d'utiliser toutes les possibilités « réalistes » des résultats du TAL en les adaptant aux objectifs visés.

Selon nous, trois points principaux définissent ainsi la problématique de ce domaine pluridisciplinaire :

– **Définir et évaluer l'apport du TAL pour l'ALAO.** L'apport potentiel du TAL découle de sa problématique et du but qu'il s'est fixé. Lui seul permet de considérer la forme langagière, non pas comme une suite de signes dépourvus d'interprétation, mais comme des éléments d'un système à deux niveaux, forme et sens. Dans le cadre de l'apprentissage des langues, le fonctionnement de chaque niveau doit être considéré, détaillé, manipulé, mis en pratique d'une manière ciblée ; les liens entre les deux niveaux doivent être mis en évidence concrètement, et la polysémie, source de difficultés en apprentissage des langues, doit pouvoir trouver la place et le traitement appropriés. Seule l'utilisation du TAL permet actuellement d'espérer atteindre ce but, de créer des systèmes et des outils permettant aux enseignants des langues (et par extension aux apprenants) de manipuler la

langue telle qu'elle est, et non telle que les techniques de base de l'informatique sont capables de l'appréhender. Concrètement, l'utilisation du TAL n'est possible que lorsque les résultats de ses traitements (aucune technique du TAL n'est fiable à 100%) ne portent pas atteinte à l'intégrité de la situation d'apprentissage. Ainsi, selon nous, le TAL ne doit être utilisé que dans une perspective d'aide, d'assistance à la création et à l'évaluation d'activités, à la recherche de documents pédagogiques, etc.

– **Connaître l'apprenant pour personnaliser son apprentissage et le favoriser.** Cette personnalisation nécessite un diagnostic des activités de l'apprenant pour que soient mises en place des remédiations appropriées. En ALAO, un grand nombre de travaux se rattachent à ces questions comme l'atteste notamment le numéro de la revue CALICO entièrement consacré à ce sujet (Heift et Schulze, 2003). Néanmoins, aller au delà d'une correction du type « vrai/faux » requiert la prise en compte des aspects langagiers des réponses fournies, du diagnostic dans le contexte de l'activité pédagogique et en fonction du profil de l'apprenant, la production de rétroactions tenant compte des paramètres précédents. En ce sens, le TAL peut apporter un ensemble de procédures et d'outils grâce auquel les questions d'analyse des réponses, de diagnostic et de rétroactions peuvent se poser sur d'autres bases, dans une démarche constructive et cumulative.

– **S'assurer que les outils et systèmes proposés ne demandent pas de compétences techniques spécifiques.** En tant que domaines de connaissances, l'informatique comme le TAL, la linguistique comme la didactique, manipulent des concepts qui leur sont propres et demandent l'usage d'outils spécifiques, fondés, le plus souvent, sur ces concepts. Pour le profane, l'utilisation de tels outils peut entraîner de longs apprentissages et suppose l'acquisition d'un minimum de concepts du domaine à l'origine de chaque outil. Dans le cadre de l'apprentissage des langues, si les enseignants sont *a priori* des spécialistes de didactique, leurs compétences ou leurs lacunes en informatique, en TAL voire en linguistique, ne doivent pas être des freins à l'utilisation et l'appropriation didactique de ces outils. En ce sens, tout produit de l'ALAO, destiné *a priori* à des enseignants de langue, doit satisfaire à deux impératifs : son utilisation ne doit demander qu'un minimum de compétences non didactiques tout en permettant le maniement de concepts didactiques en vue de la mise en œuvre de solutions didactiques ou pédagogiques.

Plus de vingt ans après le début des travaux en « TAL et ALAO », même si un certain nombre de prototypes ont été réalisés (Selva, 2002 ; Brun *et al.*, 2002 ; Antoniadis *et al.*, 2005a ; Antoniadis et Chanier, 2005), les systèmes réellement exploités sont pratiquement inexistants. L'avancée insuffisante des recherches du domaine n'est qu'une explication partielle. À notre avis, deux facteurs sont la cause principale de cet état : la pluridisciplinarité du domaine et le coût des ressources et produits issus du TAL. Non standardisés, ces derniers restent encore difficilement utilisables en l'état et demandent souvent d'importantes adaptations

pour être déployés à profit dans le cadre de l'ALAO. En guise d'illustration de notre propos, nous présentons dans la suite de cet article EXXELANT et MIRTO deux exemples d'intégration du TAL dans des systèmes d'ALAO.

2. L'EXEMPLIER EXXELANT

L'utilisation de corpus est largement répandue pour l'apprentissage des langues, qu'il s'agisse de corpus de textes pédagogiques ou de corpus d'apprenants. Le développement des outils informatiques (stockage, traitement) et l'essor d'Internet ont facilité la diffusion et le partage de corpus de grande taille permettant ainsi l'accès à une source pratiquement intarissable d'exemples en langue « authentique ». Le TAL, parce qu'il permet d'accéder aux différents niveaux de la langue, décuple les possibilités d'exploitation de ces corpus.

EXXELANT (*Example Extractor Engine Language Teaching*) est un exemplier (forme de concordancier) développé dans le cadre du projet *Integrated Digital Language Learning*[1] du réseau d'excellence européen Kaléidoscope[2]. Il permet l'interrogation du corpus d'apprenants de FLE (français langue étrangère) FRIDA-bis afin d'en extraire des exemples de phrases comportant certaines caractéristiques notamment en ce qui a trait aux erreurs. Il est destiné à la fois :

- aux enseignants de langue (pour qu'ils repèrent les difficultés spécifiques des apprenants appartenant à un certain groupe linguistique par exemple),
- aux chercheurs en TAL (en vue d'élaborer des outils pour la détection et le diagnostic d'erreur ; les erreurs des apprenants de FLE ne sont pas forcément les mêmes que celles des personnes dont c'est la langue maternelle),
- aux apprenants (pour la mise en évidence des erreurs guidées par l'enseignant afin de favoriser la prise de conscience ainsi que pour l'apprentissage par correction des pairs de façon collaborative – Hegelheimer et Fisher, 2006).

2.1. Le corpus

Le corpus FRIDA-bis est un corpus dérivé du corpus FRIDA (*French Interlanguage Database*) recueilli dans le cadre du projet FreeText (Granger *et al.*, 2001). Il s'agit de textes rédigés par des apprenants de FLE ayant des langues maternelles différentes (néerlandais, allemand, japonais, anglais, etc.) et des niveaux de français variables. Il s'agit de rédactions faites sur un sujet imposé tel que « raconter un fait divers » ou « qu'est-ce que le bonheur ? ». Il est composé de 764 textes, soit 9 466 phrases, 20 474 erreurs et 179 642 mots. Il est au format XML et possède un double système d'annotation.

1. <www.idill.org>.
2. <www.noe-kaleidoscope.org>,

Le premier balisage porte sur les erreurs des apprenants et propose trois niveaux : domaine d'erreur, catégorie d'erreur et catégorie grammaticale ainsi qu'une proposition de correction. Il a été annoté manuellement lors du projet FreeText.

Exemple : Le manager des relations <G><NBR><ADJ>#publiques$ publique </ADJ></NBR></G> explique : …

Dans l'exemple ci-dessus, le mot écrit par l'apprenant comportant une erreur était « publique » (comprise entre les symboles « # » et « $ »). Il a été corrigé par « publiques », le domaine d'erreur est la grammaire (G), la catégorie d'erreur est le nombre (NBR) et la catégorie grammaticale est adjectif (ADJ).

Le second balisage, étiquetage morphosyntaxique, est réalisé automatiquement à l'aide de TreeTagger[3] et porte sur l'ensemble des formes (texte, erreurs et correction).

Par rapport au corpus initial FRIDA, le corpus FRIDA-bis a été nettoyé (le balisage ayant été fait manuellement, il comportait certaines erreurs) puis standardisé au format XML. De plus, quelques données générales ont été ajoutées telles que la provenance de l'apprenant, la longueur des textes (en nombre de mots), la densité d'erreurs (nombre d'erreurs pour 100 mots). Enfin les deux niveaux de balises ont été imbriqués. L'exemple présenté ci-dessus devient :

```
<err ide=»2» dom1=»G» cer1=»NBR» cgr1=»ADJ»>
     <ini>
          <tok lemme=»public» cat=»adj»
          tags=»fem sing»>publique</tok>
     </ini>
     <cor>
          <tok lemme=»public» cat=»adj»
          tags=»fem plu»>publiques</tok>
     </cor>
</err>
```

Chaque erreur possède désormais un identifiant (ide), elle peut contenir plusieurs triplets de balises domaine, catégorie d'erreur et catégorie grammaticale qui seront numérotés (par exemple, sur un même token, il peut y avoir une erreur de genre et une erreur de nombre). Enfin, une erreur a sa valeur initiale écrite entre les balises <ini> et </ini> et sa correction (<cor> </cor>). Pour sa valeur initiale comme pour sa correction, le token, le lemme, la catégorie ainsi que les tags sont renseignés.

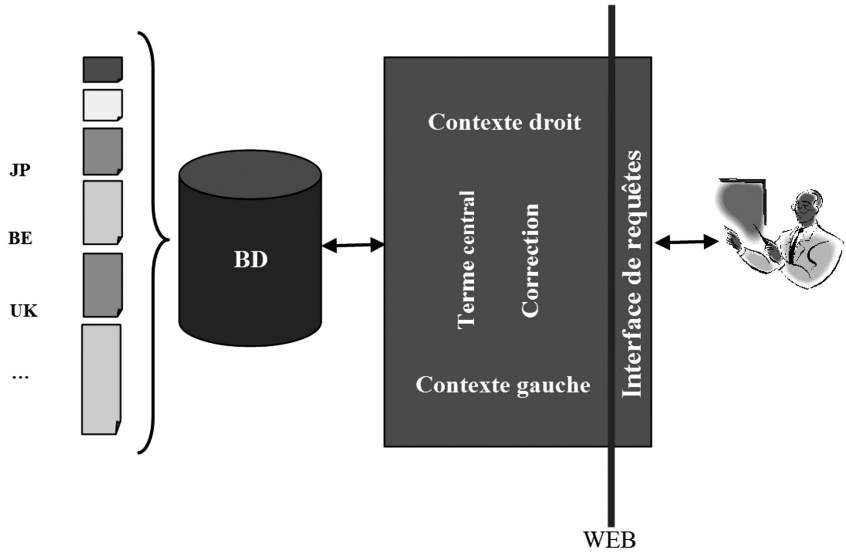
Cette normalisation et cet enrichissement morphosyntaxique de FRIDA nous ont permis de développer une interface d'interrogation riche en vue d'une exploitation didactique fine.

3. <www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>.

2.2. La base de données et l'interface

Nous avons donc décidé d'indexer le corpus, maintenant doublement balisé, dans une base de données mySQL afin de pouvoir l'interroger plus facilement.

Figure 9.1 – **Architecture générale d'EXXELANT**



L'interface, réalisée en PHP (figure 9.1), offre ainsi une interrogation de ce corpus accessible via le Web.

Cette interface est divisée en deux parties : la zone haute de l'écran permet une sélection sur le corpus et la zone basse permet la recherche sur l'expression. Cette dernière zone est subdivisée en trois parties : les contextes gauche et droit ainsi que l'expression elle-même.

- la sélection du corpus permet de préciser :
  - la provenance : il s'agit de la langue maternelle des apprenants. Pour FRIDA-bis, il y a trois possibilités : anglais, néerlandais et divers (japonais, allemand (suisse), yougoslave, etc.) ;
  - la densité d'erreurs (nombre d'erreurs pour 100 mots écrits) comportant 5 choix qui correspondent à une répartition en quartile (faible : [0 : 12[; moyenne : [12 : 17[; moyenne + : [17 : 24[; élevée : [24 : 68[; toutes : [0 : 68[) ;
  - la longueur des textes (nombre de mots contenus dans le texte) à 5 possibilités qui correspondent aussi à une répartition en quartile (tous : 29 à 3 803 ; courts : 29 à 111 ; moyens : 112 à 147 ; moyens + : 148 à 274 ; longs : 275 à 3 803).

- **la recherche d'expression** peut porter sur les critères suivants et être précisée par eux (mais non obligatoirement) :
  - l'erreur. Si le mot est dans une erreur, il est possible de renseigner :
    - le domaine d'erreur (9 domaines, dont grammaire, syntaxe, style) ;
    - la catégorie d'erreur (il y en a 36, elles dépendent du domaine d'erreur ; par exemple le domaine « ponctuation » a trois catégories : confusion, redondance et oubli) ;
    - la catégorie grammaticale (il y en a 55 qui dépendent du domaine et de la catégorie) ;
  - la forme (pour le mot cible, sa correction ou les deux) ;
  - la forme lemmatisée (pour le mot cible, sa correction ou les deux) ;
  - la correction avec sa catégorie et les traits (dont la liste dépend de la catégorie).
  - **les contextes gauche et droit :** qui, en plus des champs identiques à la recherche sur le mot cible ou corrigé, comporte la distance à l'expression (en nombre de mots).

Les deux copies d'écran suivantes présentent la recherche (figure 9.2) et le résultat (figure 9.3) de la forme « je » (contexte gauche) suivie, dans un intervalle de 10 mots, d'une expression dont la forme lemmatisée écrite par l'apprenant est « avoir » et a été corrigée par la forme lemmatisée « être ».

Figure 9.2 – **Recherche dans l'exemplier**

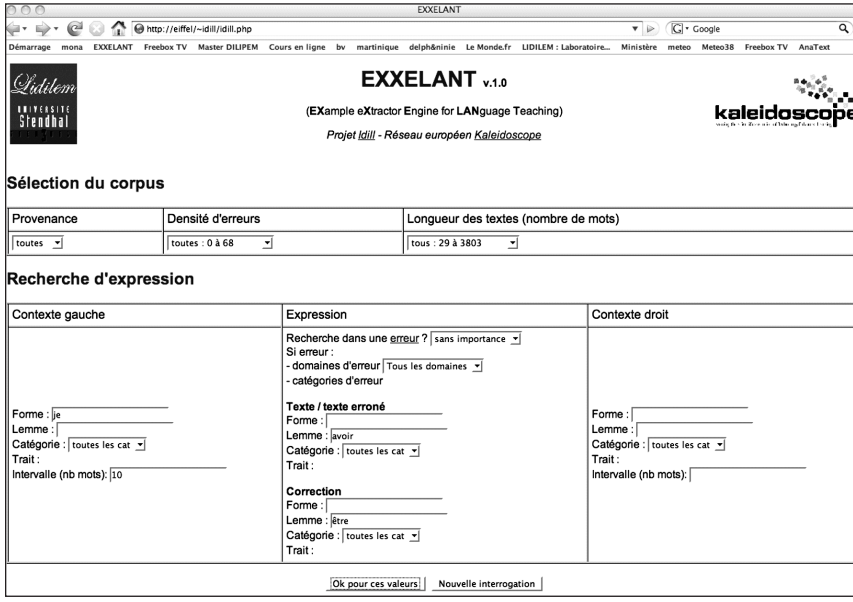
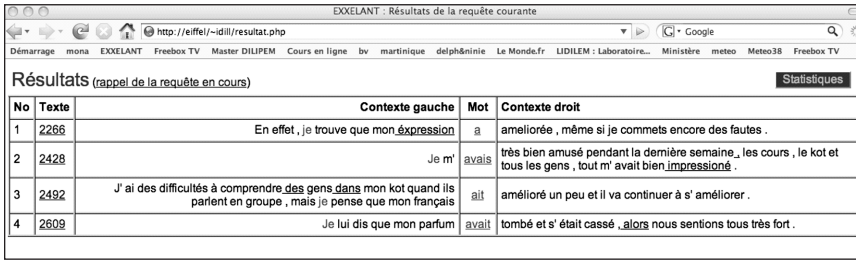
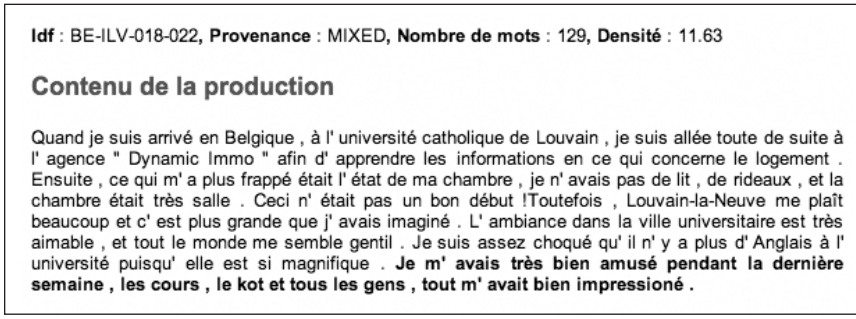


Figure 9.3 – **Résultat de la recherche**



De plus, à partir du résultat de la requête, l'utilisateur peut voir la phrase dans son contexte (c'est-à-dire dans le texte entier) en cliquant à gauche sur le numéro du texte (figure 9.4). La visualisation du texte lui donne aussi accès au sous-corpus auquel appartient le texte (ici MIXED), au nombre de mots compris dans le texte ainsi qu'à la densité d'erreurs.

Figure 9.4 – **Texte complet de la 2e phrase**



Enfin, il peut accéder à quelques résultats statistiques sur cette erreur (figure 9.5) qui lui permettent de la situer par rapport au corpus.

La recherche peut se faire de manière progressive en faisant des requêtes de plus en plus précises en fonction des résultats retournés.

Les utilisations d'un tel système sont variées tant sur un plan de didactique appliquée que sur un plan de recherche en didactique ou en TAL. Pour les chercheurs en TAL, l'exploration de tels corpus devrait permettre une amélioration des modèles de détection et d'évaluation des erreurs pour la production de rétroactions adaptées. Les pédagogues et didacticiens peuvent, entre autres, cerner et quantifier les erreurs typiques des allophones, visualiser les fréquences d'emploi de mots et de structures, comparer des productions entre différents groupes d'apprenants (en fonction de leur langue maternelle, leur niveau), etc. Ceci avant de générer des exercices sur le type d'erreurs précises qu'ils ont choisi. Afin que cet outil évolue, nous comptons ajouter une interface permettant à l'enseignant d'ajouter les textes de ses apprenants à la base de données s'il le désire.

Figure 9.5 – **Statistiques générales**

Requête en cours

(je,,,,10) [(,avoir,,) ; (,Ãªtre,,)] (,,,,)

Provenance : tout, **longueur :** tout, **densité :** tout

Statistiques générales

Nbre de corpus : 1, Nbre de textes : 4, Nbre de phrases : 4, Provenances : 2

Provenance	Nb mots	Occurrences
BE	675	3 (0,44 %)
CH	396	1 (0,25 %)

3. LA PLATEFORME MIRTO

MIRTO est un Environnement Informatisé pour l'Apprentissage Humain (EIAH) au sens où l'entend Pierre Tchounikine :

> Ce type d'environnement intègre des agents humains (élève, enseignant) et artificiels (*i.e.*, informatiques) et leur offre des conditions d'interactions, localement ou à travers les réseaux informatiques, ainsi que des conditions d'accès à des ressources formatives (humaines et/ou médiatisées), ici encore locales ou distribuées (Tchounikine, 2002, p. 4-5).

Le domaine d'apprentissage visé par MIRTO est celui des langues ; nous désignons cette spécialisation par EIAL (Environnement Informatisé pour l'Apprentissage des Langues). L'objectif premier du projet est de mettre au service des enseignants des ressources et des outils issus du TAL afin de leur permettre la conception d'activités et de scénarios[4] pédagogiques exploitant la diversité et la richesse des corpus textuels ; ces activités et scénarios ainsi élaborés étant joués par les apprenants à distance ou en présentiel.

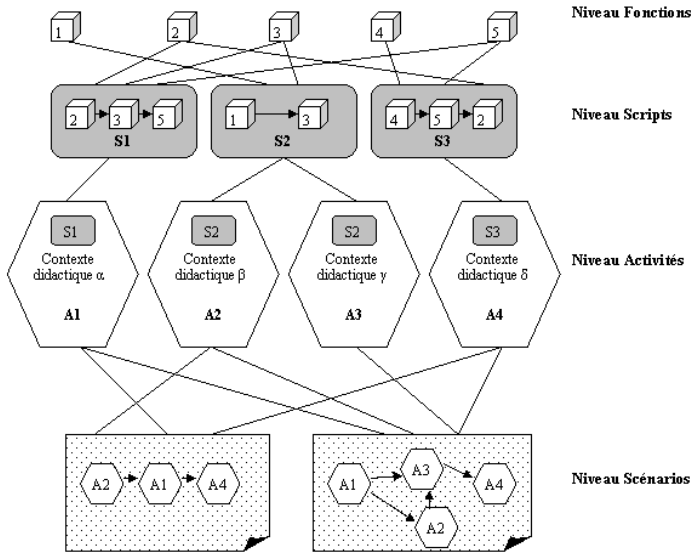
3.1. Architecture globale

L'approche MIRTO est orientée utilisateur dans la mesure où la plateforme est destinée à des enseignants de langue, qui *a priori* n'ont que peu ou pas de connaissances en TAL et en informatique. La nature technique du TAL est transparente pour eux et seuls les aspects didactiques sont visibles et disponibles. Comme le

4. La notion de « scénario » renvoie ici à un enchaînement d'activités dont l'ordonnancement est fonction des résultats et du profil de l'apprenant.

montre le schéma suivant (figure 9.6), quatre niveaux hiérarchiques (fonction, script, activité et scénario) structurent MIRTO. L'encapsulation des niveaux inférieurs par les niveaux supérieurs permet de masquer à l'enseignant de langue, par exemple, le niveau le plus technique (*i.e.* le plus bas) que sont les fonctions.

Figure 9.6 – **Architecture générale de MIRTO**



3.2. Les fonctions

Représentant le niveau le plus bas, les fonctions correspondent à des processus TAL basiques. Les fonctions actuelles implantées dans MIRTO sont essentiellement des logiciels TAL issus des laboratoires de recherche ou du commerce comme, par exemple, des identificateurs de langue, des concordanciers, des programmes capables d'extraire le vocabulaire d'un texte ou de le découper en phrases, des logiciels pouvant assigner une catégorie grammaticale à chaque mot d'un texte, etc. Étant donné la nature technique de ces fonctions, ce niveau n'est pas visible par les enseignants. Les fonctions sont encapsulées au niveau des scripts pour les appliquer à la didactique des langues.

Afin d'améliorer leurs performances, les programmes TAL sont, la plupart du temps, fortement liés à l'application pour lesquels ils ont été créés. Malgré cela, aucun produit TAL ne présente de taux de fiabilité de 100 %. La complexité et la

nature profondément ambiguë de la langue sont quelques-unes des explications à cela. À ce niveau du traitement, le degré de fiabilité des fonctions TAL est bien entendu un problème important car par propagation et amplification des erreurs d'un niveau à l'autre, le risque de produire des activités fortement fautives est élevé. Ce problème est abordé dans MIRTO selon trois axes.

Le premier consiste à appuyer les traitements de MIRTO sur les outils TAL les plus performants (tokenisation, étiquetage morphosyntaxique, etc.). Cela implique bien entendu que certains niveaux de traitement de la langue ne sont pas atteints ou ne le sont que très partiellement (sémantique par exemple) et, par conséquent, que certains types d'activités ne peuvent pas être produites par la plateforme.

Le deuxième axe consiste à faciliter les mises à jour, voire le remplacement, des outils TAL à la base des fonctions. Les avantages de cette approche sont de rendre MIRTO relativement indépendant des outils et d'utiliser uniquement les mieux adaptés.

Ces deux premiers axes permettent de réduire notablement le risque d'erreurs mais ne garantissent pas pour autant une fiabilité totale. Ces fonctions TAL étant utilisées pour la création d'activités, seuls les enseignants sont à même de valider ou pas les activités obtenues automatiquement par MIRTO. Ils ont pour cela la possibilité d'éditer les activités générées et de corriger les éventuelles fautes. L'un des enjeux réside donc dans la production d'activités le moins fautives possible pour que la plateforme conserve tout son intérêt auprès des enseignants.

3.3. La création de scripts

Étant à la frontière entre le TAL et la didactique des langues, l'étape de création des scripts requiert une double expertise. En ce sens, elle doit être transparente pour l'enseignant-utilisateur de MIRTO qui ne perçoit que les résultats, à savoir les scripts eux-mêmes. La création d'un script consiste à définir la séquence de fonctions qui, lorsqu'elle est appliquée à un texte, conduit à un résultat pédagogiquement significatif et exploitable. La création de scripts demande donc des compétences aussi bien en informatique (implantation, encapsulation…) qu'en TAL (potentialité des outils, paramétrage…) et s'effectue en collaboration avec un enseignant de langues. Le rôle de ce dernier est la définition du concept pédagogique que le script doit exprimer et sa dénomination. La partie la plus importante de cette collaboration réside dans la définition des options du script accessibles à l'enseignant-utilisateur (paramètres).

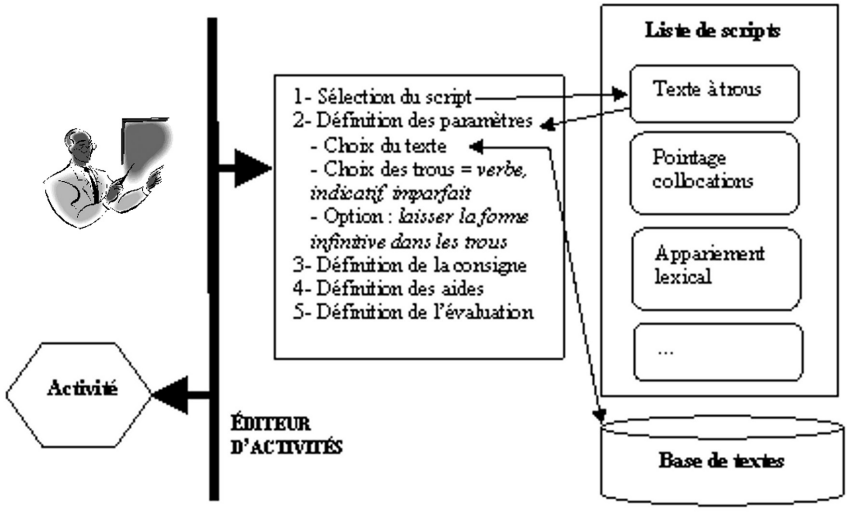
Même si cette procédure reste complexe, on peut remarquer qu'un script créé est partagé par tous les enseignants-utilisateurs de MIRTO et constitue une brique pédagogique pouvant être utilisée, indéfiniment, dans une pléthore de situations pédagogiques.

3.4. La création d'activités

Une activité est la mise en œuvre d'un objectif pédagogique minimal comme : travailler une notion grammaticale particulière, rédiger un paragraphe sur un sujet, réviser des conjugaisons, etc. Elle se réalise par la construction d'un espace de travail pour l'apprenant lui permettant d'atteindre le but visé. Elle correspond à ce qui est traditionnellement désigné comme exercice. D'un niveau purement didactique, la conception d'activités est une tâche opérée par les enseignants de langue à l'aide d'un outil spécifique de MIRTO : l'éditeur d'activités. Ce dernier est un environnement de conception permettant de visualiser et de manipuler des objets et des outils pédagogiques tels que des textes (ou corpus de textes), des scripts et des consignes. La génération d'activités consiste ainsi à appliquer un contexte didactique à un script.

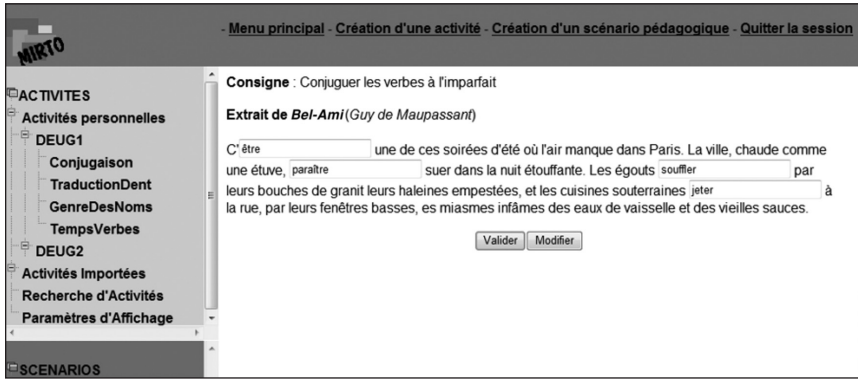
Par exemple, la génération automatique d'exercices lacunaires est considérée comme un script car elle lie les fonctions d'identification de la langue, de tokenisation, d'analyse morphologique et de création des trous en fonction de paramètres didactiques entrés par l'utilisateur. Un script est donc un objet didactique qui constitue l'élément de base de création d'activités puis, éventuellement, de scénarios par les enseignants. Afin d'illustrer cette opération de conception, considérons l'exemple d'un enseignant désirant créer une activité de révision systématique de l'indicatif imparfait au travers d'un exercice lacunaire (figure 9.1).

Figure 9.7 – **Création d'une activité**



La première étape de la création d'une activité concerne le choix de l'outil (le script), adapté aux besoins, suivi du paramétrage de son contexte didactique d'utilisation. Dans notre exemple, la sélection du script « texte à trous » induit un paramétrage, de la part de l'enseignant, sur le texte à utiliser et les lacunes à insérer sur ce texte (choix des formes à enlever : verbes à l'imparfait). Le choix du texte d'une activité répond à certaines visées didactiques. Actuellement, dans MIRTO, l'enseignant le sélectionne dans une base selon des critères classiques comme le titre, l'auteur et la date de création. Toutefois, une base de textes indexée selon des critères didactiques (Loiseau *et al.*, 2008) est actuellement à l'étude et sera prochainement intégrée à la plateforme. Enfin, avant la génération effective de l'activité, il reste à définir des paramètres plus généraux comme la consigne, les aides éventuelles (fiches de cours, dictionnaire, etc.) liées à l'activité et le type d'évaluation. Comme nous l'avons précisé, l'activité générée (figure 9.8) reste modifiable par l'enseignant pour corriger d'éventuelles erreurs ou simplement pour l'adapter à des besoins spécifiques.

Figure 9.8 – **Exemple d'activité**



3.5. Évaluation des activités

Actuellement, seule la partie génération d'activités est développée. Une fois jouées par les apprenants, ces activités donnent lieu à une simple évaluation vrai/faux. Bien évidemment, ce type d'évaluation est nettement insuffisant pour produire des feedback adaptés à l'apprenant, pour lui proposer des activités de remédiation ou simplement pour fournir des indications fines à l'enseignant pour le suivi de ses groupes.

Avec les mêmes hypothèses que pour MIRTO, à savoir l'utilisation des technologies les plus fiables dans une perspective d'assistance, nous travaillons sur des systèmes de détection et de diagnostic des erreurs d'apprenants (Kraif et Ponton, 2007 ; Blanchard, 2007). Ce diagnostic doit à terme permettre la constitution d'un modèle d'apprenants pour :

– la production de feedback adapté à l'apprenant et à la situation ;
– le suivi des groupes et des individus ;
– la création de scénarios (suite d'activités dont l'enchaînement dépend des résultats de l'apprenant).

4. CONCLUSION ET PERSPECTIVES

Si l'on considère le travail d'un enseignant de langues, les tâches ci-après font, entre autres, partie de son « quotidien » pédagogique :

– élaboration de plans pédagogiques en adéquation avec les notions à enseigner et le niveau des apprenants ;
– recherche de textes et/ou d'exemples appropriés ;
– élaboration et/ou recherche d'activités relatives aux notions abordées ;
– évaluation des activités effectuées par les apprenants et remédiation pédagogique en fonction des résultats.

Le plus souvent, ces tâches se conjuguent avec un suivi de la classe et de chaque apprenant, mais aussi d'une adaptation permanente du matériau pédagogique en fonction des progressions individuelles.

De son côté, l'apprenant en autonomie, en semi-autonomie, en présentiel ou à distance est amené à lire ou écouter des documents langagiers, à faire des exercices appropriés, à analyser ses erreurs grâce à des aides et des commentaires pédagogiques, à communiquer avec ses pairs ou avec ses tuteurs, à suivre une progression en fonction d'un parcours préétabli mais aussi de sa progression réelle.

L'ALAO s'intéresse à l'ensemble de tâches ci-dessus ; plusieurs plateformes, logiciels, démarches, etc., tentent d'apporter une aide voire de les automatiser complètement ou partiellement. Néanmoins, le plus souvent, ces solutions butent sur les limites intrinsèques de l'informatique qui, seule, ne peut considérer la sémantique associée aux séquences langagières. En général, ces séquences ne constituent pour elle qu'une concaténation de codes, avec comme seule sémantique celle qui est associée aux propriétés mathématiques de la concaténation, bien loin de la sémantique que considèrent et manipulent les enseignants de langues. Sans qu'il soit la panacée, le TAL permet de dépasser ces limites et peut ainsi apporter une réelle plus-value pédagogique aux systèmes et procédures d'ALAO.

Pratiquement tous les aspects traités par l'ALAO sont concernés par le TAL. Comme EXILLS (Brun *et al.*, 2002) ou ALFALEX (<www.kuleuven.be/**alfalex**/>), EXXELANT et MIRTO constituent des approches tenant compte

des propriétés de la langue pour des tâches telles que la création d'activités ou de scénarios pédagogiques, la détection (Heift et Schulze, 2003 ; Kraif et Ponton, 2007) ou l'évaluation des erreurs des apprenants pour l'élaboration d'un feedback à plus-value pédagogique (Antoniadis *et al.*, 2005b).

La recherche de documents pédagogiques par les enseignants et leur indexation pédagogique sont une autre préoccupation de l'intégration du TAL à l'ALAO. En effet, une telle recherche n'est pertinente que si elle peut se faire sur des critères et des concepts de la didactique des langues. Loiseau (Loiseau *et al.*, 2008) a montré que ce n'est pas le cas avec les normes et les systèmes actuels d'indexation de tels matériaux. Le plus souvent, seuls leurs aspects de « surface » sont considérés. Le TAL peut, ici aussi, permettre d'exploiter, lors de l'indexation et de la recherche, des caractéristiques propres au contenu pédagogique de chacune de ces ressources.

Quelles que soient les « performances » théoriques d'un outil, seule la confrontation avec ses utilisateurs potentiels permet de conclure quant à sa pertinence et apporter si nécessaire les ajustements appropriés. Mis à part la confrontation des fonctionnalités de l'outil au terrain, cette phase de tests permet souvent d'opérer nombre de mesures liées à son utilisation et de développer, en les exploitant, des améliorations ou des nouvelles fonctionnalités significatives. Concernant EXXELANT et MIRTO, les perspectives à court terme vont en ce sens. Ces logiciels devraient être mis à la disposition des enseignants de langues afin d'exploiter par la suite aussi bien le compte-rendu de leur usage que les données collectées par traçage concernant certains aspects de leur utilisation.

En conclusion, le champ couvert par l'application du TAL à l'ALAO est vaste. Il recouvre de nombreux sujets dont certains, comme le traitement de la langue orale, n'ont pas été évoqués ici. En ce qui concerne le traitement de la langue écrite, les nombreuses thématiques abordées sont autant de défis auxquels les chercheurs sont confrontés à l'heure actuelle. Une chose apparaît de manière sûre : si les chercheurs qui travaillent à l'intersection du TAL et de l'ALAO veulent avoir une chance de voir leurs travaux mis en pratique dans l'enseignement des langues, il est indispensable qu'ils accordent une grande importance à la scénarisation pédagogique, aspect encore souvent marginal. La clé réside dans le travail pluridisciplinaire (informatique, TAL, didactique des langues) et la mise en commun de pratiques et de techniques. Néanmoins, tout laisse à penser que l'intégration du TAL à l'ALAO qui permet de prendre en considération les caractéristiques de la langue de référence et celles de l'apprenant, se double actuellement de l'intégration de la didactique des langues, seul moyen pour que l'objectif d'apprentissage soit pleinement atteint par le biais de systèmes et d'outils ALAO conviviaux et performants.

RÉFÉRENCES

ANTONIADIS, G. et T. CHANIER (dir.) (2005). « TAL (Traitement automatique des langues) et apprentissage des langues », *Apprentissage des langues et Systèmes d'information et de communication (Alsic)*, vol. 8, nº 2, numéro thématique.

ANTONIADIS, G., S. ECHINARD et C. PONTON (2005b). « Le TAL au service de l'évaluation automatique des fautes des apprenants. Du vrai/faux à l'évaluation pédagogique », *Actes du 6e colloque des Usages des nouvelles technologies pour l'enseignement des langues étrangères*, Compiègne.

ANTONIADIS, G. *et al.* (2005a). « MIRTO : A New Approach to CALL Systems », dans J.P. Courtiat *et al.* (dir.), *IFIP TC3 Technology Enhanced Learning Workshop (TeL'04)*, Toulouse, World Computer Congress.

BLANCHARD, A. (2007). « Analyse morphologique des réponses d'apprenants », *Actes de RECITAL 2007*, Toulouse.

BRUN, C. *et al.* (2002). « Les outils de TAL au service de la e-formation en langues », dans F. Segond (dir.), *Multilinguisme et traitement de l'information*, Paris, Hermès, p. 223-250.

CHANIER, T., D. RENIÉ et C. FOUQUERÉ (dir.) (1993). *Actes du colloque SCIAL'93 (sciences cognitives, informatique et apprentissage des langues)*, Clermont-Ferrand, Université Blaise Pascal, Préface et sommaire en ligne à <archive-edutice.ccsd.cnrs.fr/edutice-00001148>.

GRANGER, S., A. VANDEVENTER et M.J. HAMEL (2001). « Analyse des corpus d'apprenants pour l'ELAO basé sur le TAL », *Traitement Automatique des Langues*, vol. 42, nº 2, p. 609-621.

HEGELHEIMER, V. et D. FISHER (2006). « Grammar, Writing, and Technology : A Sample Technology-supported Approach to Teaching Grammar and Improving Writing for ESL Learners », CALICO Journal, vol. 23, nº 2, p. 257-279.

HEIFT, T. et M. SCHULZE (2003). *Special Issue of CALICO, Error Analysis and Error Correction in Computer-Assisted Language Learning*, vol. 20, nº 3.

HOLLAND, V.M., J.D. KAPLAN et M.R. SAMS (dir.) (1995). *Intelligent Language Tutors*, Mahwah, Lawrence Erlbaum Associates.

KRAIF, O. et C. PONTON (2007). « Du bruit, du silence et des ambiguïtés : que faire du TAL pour l'apprentissage des langues ? », *Actes TALN 2007*, 5-8 juin, Toulouse.

LOISEAU, M. et G. ANTONIADIS et C. PONTON (2008). « Model for Pedagogical Indexation of Texts for Language Teaching », *3rd International Conference on Software and Data Technologies (ICSOFT 2008)*, Porto.

SELVA, T. (2002). « Génération automatique d'exercices contextuels de vocabulaire », *TALN'2002*, p. 185-194.

SWARTZ, M. et M. YAZDANI (dir.) (1992). *The Bridge to International Communication : Intelligent Tutoring Systems for Foreign Language Learning*, Springer-Verlag.

TCHOUNIKINE, P. (2002). « Pour une ingénierie des environnements informatiques pour l'apprentissage humain », *Revue I3 information – interaction – intelligence*, vol. 2, nº 1, p. 59-95.

NOTICES BIOGRAPHIQUES

CAMILLE ALBERT est titulaire d'un Master 2 Recherche en études anglophones (spécialité : linguistique), délivré par l'Université de Toulouse le Mirail (UTM). Ses recherches concernent l'apprentissage de l'anglais au collège en France. Elle travaille sur le traitement de l'erreur et la correction automatique au sein de l'équipe ILPL (Informatique, linguistique et programmation logique) de l'Institut de recherche en informatique de Toulouse (IRIT) depuis octobre 2008.

GEORGES ANTONIADIS est maître de conférences en informatique, habilité à diriger des recherches, Grenoble 3 (France). Il fait partie du laboratoire LIDILEM (Linguistique et didactique des langues étrangères et maternelles). Il travaille dans le domaine du traitement automatique des langues (TAL) et notamment sur ses applications aux environnements informatiques pour l'apprentissage des langues (EIAL).

ISMAÏL BISKRI est professeur titulaire en informatique à l'Université du Québec à Trois-Rivières, au Laboratoire de mathématiques et informatique appliquées (LAMIA). Il travaille dans le domaine du traitement automatique des langues naturelles (TALN).

PIERRETTE BOUILLON est professeure à l'École de traduction et d'interprétation de l'Université de Genève. Elle y dirige le Département de traitement informatique multilingue et assure pour l'instant un mandat de vice-doyenne. Ses deux principaux champs d'intérêt incluent la sémantique lexicale, et en particulier le lexique génératif, ainsi que la traduction automatique de la parole dans des domaines limités, par des techniques essentiellement linguistiques.

JULIEN BOURDAILLET est docteur de l'Université Pierre et Marie Curie, Laboratoire d'informatique de Paris 6 (LIP6). Il a effectué sa thèse au sein de l'équipe ACASA sous la direction du professeur Jean-Gabriel Ganascia. Sa thèse porte sur le traitement automatique des langues naturelles (TALN), l'algorithmique textuelle et la critique génétique textuelle. Il a travaillé sur la comparaison des brouillons d'écrivains dans le cadre d'une collaboration avec l'Institut des textes et manuscrits modernes (ITEM) de l'École normale supérieure (ENS). Dans ce cadre, il a développé un algorithme d'alignement textuel monolingue implémenté dans l'application MEDITE. Il est actuellement en séjour postdoctoral à l'Université de Montréal, au laboratoire RALI (Recherche appliquée en linguistique informatique).

BRUNO CARTONI est docteur en traitement informatique multilingue de l'Université de Genève. Son travail de thèse et les différents travaux qui en ont découlé portent sur l'inférence entre les langues, et plus particulièrement sur la formalisation de cette inférence, notamment dans le but de la constitution de ressources lexicales informatisées, utilisées dans différents outils du traitement de la langue.

JEAN-GABRIEL GANASCIA est professeur à l'Université Pierre et Marie Curie (Paris 6), depuis le 1er novembre 1988**.** Il dirige l'équipe ACASA au LIP6 (Agents cognitifs et apprentissage symbolique automatique). Il travaille sur trois axes particuliers de l'intelligence artificielle : l'apprentissage, la découverte et la créativité. Initialement centrées sur l'acquisition de connaissances et l'apprentissage symbolique automatique, ses orientations scientifiques ont évolué vers la modélisation cognitive et la conception d'agents intelligents.

MARIE GARNIER est agrégée d'anglais et titulaire d'un Master 2 Recherche en études anglophones (spécialité : linguistique), délivré par l'UTM. Ses recherches concernent notamment l'importance cognitive du corps dans la langue. Elle travaille sur le traitement de l'erreur et la correction automatique au sein de l'équipe ILPL depuis octobre 2008, et est actuellement en préparation de thèse en collaboration avec l'équipe de P. Saint-Dizier et l'UTM (projet CAPRA).

MARIE-JOSÉE GOULET est professeure au Département d'études langagières de l'Université du Québec en Outaouais. Elle est titulaire d'un doctorat en linguistique de l'Université Laval. Ses recherches portent sur l'évaluation des technologies langagières.

BASSEL HABIB est en 3e année de thèse à l'Université Pierre et Marie Curie (Paris 6), Le sujet de sa thèse est « CYBERNARD : la reconstruction rationnelle de la démarche scientifique de Claude Bernard ». Cette thèse porte sur la découverte scientifique et s'intéresse, en particulier, aux travaux scientifiques de Claude Bernard. Elle est inscrite dans le cadre d'un projet mené en collaboration entre trois équipes : l'équipe d'informaticiens ACASA du LIP6, l'équipe d'épistémologues de l'École normale supérieure (ENS) et l'équipe de linguistique de l'Institut des textes et manuscrits modernes (ITEM). Dans ce cadre, il a construit un laboratoire virtuel au sein duquel plusieurs expériences virtuelles peuvent être reproduites.

ADEL JEBALI est professeur adjoint au Département d'études françaises de l'Université Concordia. Il a obtenu un doctorat en linguistique du Département de linguistique et de didactique des langues de l'Université du Québec à Montréal et consacre ses travaux de recherche à l'intégration du traitement automatique des langues (TAL) aux technologies de l'information et de la communication en enseignement (TICE).

ADAM JOLY est étudiant au doctorat en informatique cognitive à l'Université du Québec à Montréal et est affilié au LAMIA à l'Université du Québec à Trois-Rivières. Ses recherches portent sur le traitement automatique des langues naturelles (TALN).

CHRISTOPHE JOUIS est maître de conférences titulaire (professeur associé) à l'Université Paris Sorbonne Nouvelle, depuis 1999, et il est rattaché pour la recherche au LIP6 (Laboratoire d'informatique de Paris 6), depuis 2005. Il travaille depuis plusieurs années sur les représentations sémantiques des textes scientifiques ou techniques. Parmi les questions abordées nous avons les suivantes : 1) la sémantique des relations sémantiques entre concepts (c'est-à-dire, pour chaque relation, le nombre et le type de ses arguments, ses propriétés algébriques, etc.). Une approche possible à ce problème consiste à organiser les relations sémantiques dans une typologie fondée sur des propriétés logiques ; et 2) une contribution à l'étude formelle des ontologies, le problème des entités atypiques.

MATHIEU LAFOURCADE est maître de conférences en informatique à Montpellier 2, au laboratoire LIRMM (Laboratoire d'informatique, de robotique et de microélectronique de Montpellier). Ses thématiques sont l'analyse sémantique de textes à partir d'algorithmes à fourmis, la sémantique lexicale via des réseaux lexicaux et sémantiques de très grande taille et des vecteurs conceptuels, l'acquisition de ressources de façon contributive.

FIAMMETTA NAMER est professeur des universités en sciences du langage, à l'Université de Nancy 2. Ses domaines de recherche sont la morphologie lexicale du français et la construction du lexique, l'interaction morphologie-sémantique lexicale, et le traitement automatique des langues.

CLAUDE PONTON est maître de conférences en informatique à Grenoble 3 au laboratoire LIDILEM (Linguistique et didactique des langues étrangères et maternelles). Il travaille dans le domaine du traitement automatique des langues et notamment sur ses applications aux environnements informatiques pour l'apprentissage des langues (EIAL).

NILDA RUIMY est chargée de recherche de première classe à l'Istituto di linguistica computazionale del consiglio nazionale delle ricerche, à Pise. Elle travaille actuellement dans le domaine de la lexicologie et lexicographie computationnelle monolingue et bilingue, et est responsable scientifique d'une vaste base de données lexicales multi-niveaux de l'italien. Ses champs d'intérêt professionnel sont le traitement automatique du langage, la syntaxe et la sémantique lexicale computationnelle, le multilinguisme.

ARNAUD RYKNER est professeur à l'Université Toulouse-Le Mirail et membre de l'Institut universitaire de France. Il est également directeur du laboratoire Lettres, langages et arts, et est spécialisé dans les études théâtrales et l'esthétique de la représentation. Il est une des personnes à l'origine du projet d'amélioration des corrections/traductions automatiques de texte (projet Transtyler), et travaille en collaboration avec l'équipe Informatique, linguistique et programmation logique (ILPL) depuis 2007.

PATRICK SAINT-DIZIER est directeur de recherches au CNRS en poste à l'IRIT. Il dirige l'équipe Informatique, linguistique et programmation logique (ILPL) depuis plus de vingt ans. Ses principaux axes de recherche sont le traitement automatique des langues naturelles, la sémantique textuelle et les systèmes de question-réponse. Il contribue au projet actuel de création d'un système informatique d'amélioration des productions des francophones en anglais.

JACQUES VERGNE est enseignant-chercheur à l'université de Caen (France). Ses recherches se situent en informatique linguistique et ont toujours privilégié les solutions calcul plutôt que les solutions ressources : analyse syntaxique sans dictionnaire, de complexité linéaire. Son analyseur du français (1998) était constitué d'un *chunker* et d'un relieur de *chunks*, et il a été classé 1er à l'action d'évaluation GRACE (1998). Ses travaux se sont ensuite orientés vers des méthodes d'analyse de plus en plus multilingues et de plus en plus endogènes. Ses travaux actuels concernent l'alignement sous-phrastique multilingue (et non bilingue), et le *chunking* multilingue endogène.

VIRGINIE ZAMPA est maître de conférences en informatique à Grenoble 3, au laboratoire LIDILEM (Linguistique et didactique des langues étrangères et maternelles). Elle travaille dans les domaines des environnements informatiques pour l'apprentissage humain (EIAH), notamment pour l'acquisition des langues étrangères et du traitement automatique du langage (TAL), plus spécialement en analyse sémantique.

Fruit du colloque international *Multilinguisme et traitement automatique des langues*, tenu à Ottawa le 11 mai 2009, les articles recueillis dans cet ouvrage présentent les dernières avancées scientifiques en matière de traitement automatique des langues, et ce, tant sur le plan théorique que de l'application. Une description des méthodes d'évaluation des résumés automatiques, de détection des entités atypiques dans les ontologies ou encore d'autoconstruction assistée d'un lexique sémantique y est notamment proposée. Des exemples d'environnements informatiques intégrant des outils de traitement automatique de la langue sont également présentés, tout comme une application d'autocorrection en cours d'élaboration destinée aux francophones rédigeant en anglais.

La pluridisciplinarité des auteurs, les uns étant spécialistes en informatique, les autres en linguistique ou en sciences cognitives, de même que la multiplicité des langues étudiées par ceux-ci, que ce soit le français, l'anglais ou l'arabe, offrent l'un des rares panoramas portant sur l'analyse et l'implémentation de phénomènes linguistiques.

ISMAÏL BISKRI est professeur titulaire en informatique à l'Université du Québec à Trois-Rivières au Laboratoire de mathématiques et informatique appliquées (LAMIA). Il travaille dans le domaine du traitement automatique des langues naturelles (TALN).

ADEL JEBALI est professeur adjoint au Département d'études françaises de l'Université Concordia. Il a obtenu un doctorat en linguistique du Département de linguistique et de didactique des langues de l'Université du Québec à Montréal et consacre ses travaux de recherche à l'intégration du TAL aux technologies de l'information et de la communication pour l'éducation (TICE).

ONT COLLABORÉ À CET OUVRAGE

Camille Albert, Georges Antoniadis, Ismaïl Biskri, Pierrette Bouillon, Julien Bourdaillet, Bruno Cartoni, Jean-Gabriel Ganascia, Marie Garnier, Marie-Josée Goulet, Bassel Habib, Adel Jebali, Adam Joly, Christophe Jouis, Mathieu Lafourcade, Fiametta Namer, Claude Ponton, Nilda Ruimy, Arnaud Rykner, Patrick Saint-Dizier, Jacques Vergne, Virginie Zampa

www.puq.ca

9 782760 525696

ISBN 978-2-7605-2569-6