ILIMP : Outil pour repérer les occurrences du pronom impersonnel *il*

Laurence Danlos

LATTICE, Université Paris 7, Institut Universitaire de France
Laurence.Danlos@linguist.jussieu.fr

Mots-clés : Pronom impersonnel (explétif), Pronom anaphorique, Lexique-grammaire, Automates, Résolution d'anaphores, Analyse syntaxique modulaire

Keywords: Expletive pronouns, Anaphoric pronouns, Lexicon-Grammar, Automata, Anaphora resolution, Modular syntactic analysis

Résumé Nous présentons un outil, ILIMP, qui prend en entrée un texte brut (sans annotation linguistique) rédigé en français et qui fournit en sortie le texte d'entrée où chaque occurrence du pronom *il* est décorée de la balise [ANAphorique] ou [IMPersonnel]. Cet outil a donc comme fonctionnalité de distinguer les occurrences anaphoriques du pronom *il*, pour lesquelles un système de résolution des anaphores doit chercher un antécédent, des occurrences où *il* est un pronom impersonnel (explétif) pour lequel la recherche d'antécédent ne fait pas sens. ILIMP donne un taux de précision de 97,5%. Nous présentons une analyse détaillée des erreurs et nous décrivons brièvement d'autres applications potentielles de la méthode utilisée dans ILIMP, ainsi que l'utilisation et le positionnement d'ILIMP dans un système d'analyse syntaxique modulaire.

Abstract We present a tool, ILIMP, which takes as input a French raw text and which produces as output the input text in which every occurrence of the word *il* is tagged either with the tag [ANA] for anaphoric or [IMP] for expletive. This tool is therefore designed to distinguish the anaphoric occurrences of *il*, for which an anaphora resolution system has to look for an antecedent, from the expletive occurrences of this pronoun, for which it does not make sense to look for an antecedent. The precision rate for ILIMP is 97,5%. The few errors are analyzed in detail. Other tasks using the method for ILIMP are described briefly, as well as the use of ILIMP in a modular syntactic analysis system.

1 Introduction

En TAL, la résolution d'anaphores est un sujet de recherche fort étudié car c'est une question cruciale pour des applications comme la Recherche d'Informations ou le Résumé Automatique. Parmi les anaphores, les pronoms sont largement traités car fréquents et facilement identifiables. Parmi les pronoms, on peut distinguer un élément (*il* en français, *it* en anglais) avec un emploi impersonnel (explétif) *(il pleut, it rains)* qui se distingue de l'emploi anaphorique *(il est cher, it is expensive)*. Un système de résolution des anaphores doit être capable de repérer les occurrences des pronoms impersonnels avant de s'attaquer aux pronoms anaphoriques et aux autres anaphores. De ce fait, il existe un certain nombre de travaux sur les emplois impersonnels du pronom anglais *it*, citons (Lapin, Leass, 1994), (Kennedy, Bogurev, 1996) et (Evans 2001). Mais à notre connaissance, il n'existe pas de travaux similaires sur le pronom français *il*. Ce travail présente un outil, ILIMP, qui est conçu pour reconnaître toutes les occurrences du pronom impersonnel *il* dans les textes français : ILIMP décore chaque occurrence de *il* de la balise [ANAphorique] ou [IMPersonnel]. Cet outil est à base de règles (comme c'est le cas du système de Lapin et Leass) ; il travaille sur des textes bruts sans annotation linguistique (contrairement au système de Lapin et Leass qui repose sur une analyse syntaxique).

Si ILIMP est un outil s'imposant en amont d'un système de résolution d'anaphores, c'est aussi un outil qui peut s'intégrer dans la chaîne de traitements d'un analyseur syntaxique modulaire. En effet, d'une part, les balises [IMP] et [ANA] sur le pronom *il* peuvent être vues comme un raffinement des étiquettes morpho-syntaxiques traditionnellement utilisées dans les taggeurs : l'étiquette "pronom" serait remplacée par deux étiquettes, "pronom anaphorique" et "pronom impersonnel". Or on sait que plus le jeu d'étiquettes morpho-syntaxiques d'un taggeur est riche, plus un analyseur venant en aval de ce taggeur a des chances d'aboutir à l'analyse syntaxique correcte (Nasr, 2004). D'autre part, nous verrons que des outils dérivés d'ILIMP peuvent être utilisés pour d'autres annotations linguistiques.

La Section 2 présente notre méthode qui est basée, pour l'aspect linguistique, sur le lexique-grammaire du LADL, et pour l'aspect informatique sur UNITEX. La Section 3 décrit la réalisation d'ILIMP, les difficultés que nous avons rencontrées et les choix effectués pour les surmonter. Finalement, la Section 4 donne une évaluation d'ILIMP et discute de son positionnement dans une analyse syntaxique modulaire.

2 Méthode

2.1 Lexique-grammaire

Comme la plupart des phénomènes linguistiques, les constructions impersonnelles reposent sur des conditions tant lexicales que syntaxiques. Par exemple, l'adjectif *violet* ne peut jamais être la tête lexicale d'une phrase impersonnelle, (1a), l'adjectif *probable* ancre une phrase impersonnelle lorsqu'il est suivi d'un complément phrastique, (1b), l'adjectif *difficile* ancre une phrase impersonnelle (resp. personnelle) lorsqu'il est suivi d'une infinitive introduite par la préposition *de* (resp. *à*), (1c) et (1d).

(1)a Il est violet
b Il est probable que Fred viendra
c Il est difficile de résoudre ce problème
d Il est difficile à résoudre (ce problème)

De ce fait, le lexique-grammaire du français développé par Maurice Gross et son équipe (Gross1994, Leclère 2003) est une ressource linguistique appropriée pour ILIMP puisqu'il décrit l'ensemble des têtes lexicales des phrases simples du français avec leurs arguments syntaxiques et les alternances possibles. Nous avons donc extrait (manuellement) du lexique-grammaire tous les items lexicaux qui peuvent ancrer une phrase impersonnelle en enregistrant leur complémentation syntaxique. Présentons un bref aperçu des constructions impersonnelles du français que nous avons recensées. On peut distinguer les constructions intrinsèquement impersonnelles, qui ne peuvent avoir comme sujet que *il,* des constructions avec un "sujet profond extraposé". Parmi les premières, on trouve 45 verbes météorologiques de la table 31i de (BGL, 1976a) *(Il pleut, Il fait beau)*, 21 verbes de la table 17 de (Gross, 1975) *(Il faut du pain /que Fred vienne)* et 38 expressions figées de (Gross, 1993) (Il é*tait une fois, quoi qu'il en soit)*. Pour les constructions impersonnelles à sujet profond extraposé, on peut distinguer celles à sujet phrastique de celles à sujet nominal. Parmi les premières, on trouve 682 adjectifs dispersés dans les tables de (Picabia, 1978) et (Meunier, 1981) *(Il est probable que Fred viendra)*, 88 expressions "être Prép X" des tables Z5P et Z5D de (Danlos 1980) *(Il est de règle de porter un chapeau)*, 21 verbes de la table 5 de (Gross 1975) *(Il plaît à Vanne que Fred vienne)*, et enfin 140 verbes de la table 6 et 92 verbes de la table 9 de (Gross 1975) construits au passif ou au se-moyen *(Il a été dit/se raconte que Fred viendra)*. Les constructions impersonnelles à sujet extraposé nominal ont pour tête lexicale des verbes qui sont dispersés dans les tables élaborées par (BGL, 1976b). On peut distinguer d'un côté des verbes comme *manquer* ou *rester* dont l'emploi en construction impersonnelle est tout à fait courant *(Il manque /reste du pain)*, et de l'autre côté des verbes "inacusatifs" (*Il est venu trois personnes)* ou des verbes construits au passif *(Il a été mangé trois gâteaux)*, dont l'emploi dans une construction impersonnelle relève d'un niveau de langue châtié.

2.2 UNITEX

UNITEX[1] est un outil qui permet d'écrire des patrons linguistiques (expressions régulières ou automates) qui sont localisés dans le texte d'entrée, avec un éventuel ajout d'annotations lorsque les automates sont en fait des transducteurs. Un texte brut donné en entrée à UNITEX est d'abord pré-traité : le texte est segmenté en phrases et les phrases segmentées en tokens. Chaque token est étiqueté avec *toutes* les parties du discours et traits flexionnels enregistrés dans le "full-form" dictionnaire DELAF (Courtois, 2004). Il n'y a pas de désambiguïsation, autrement dit le pré-traitement d'UNITEX n'est pas équivalent à un étiquetage morpho-syntaxique.

Pour réaliser ILIMP, l'idée de base consiste à repérer les constructions impersonnelles grâce à leur tête lexicale et à leur complémentation. Il s'agit donc d'écrire (manuellement) un ensemble de transducteurs comme celui présenté en (2) sous une forme linéaire simplifiée. La balise [IMP] est l'ajout d'information amenée par l'aspect transducteur de (2). Les éléments entre chevrons de (2) se glosent de la façon suivante : <être.V:3s> correspond à toutes les formes du verbe *être* conjugué à la troisième personne du singulier, <Adj1:ms> correspond aux adjectifs masculins singuliers de la classe Adj1 qui regroupe des adjectifs se comportant comme *difficile*, <V:W> correspond aux verbes à l'infinitif.

(2) Il [IMP] <être.V:3s> <Adj1:ms> de <V:W>

---

[1] UNITEX est un logiciel sous licence GPL, dont l'ancêtre est INTEX (Silberstein, 1994). La documentation et le téléchargement de UNITEX se trouvent sur le site http://ladl.univ-mlv.fr.

La balise [IMP] - abréviation de impersonnelle- vient décorer les occurrences de *il* qui apparaissent dans les phrases correspondant au patron de (2). Cette balise vient donc décorer *il* dans (1c). La balise [ANA] - abréviation de anaphorique- est la balise par défaut : elle vient décorer les occurrences de *il* qui n'ont pas été balisées par [IMP]. Cette balise vient décorer *il* dans (1d). Néanmoins, la situation est un peu plus complexe, car il existe une troisième balise [AMB] - abréviation de ambigu- qui sera expliquée dans la section 3.2. Après cet exposé des principes théoriques sous-tendant ILIMP, passons à sa réalisation pratique.

3 Réalisation de ILIMP

3.1 Contexte gauche de la tête lexicale

Dans l'exemple (1c), le contexte gauche de la tête lexicale de la phrase - la séquence de tokens à gauche de *difficile* - se réduit à *Il est*. Mais on trouve fréquemment dans les corpus des phrases comme (3a) ou (3b) dans lesquelles le contexte gauche de la tête lexicale est plus complexe. Dans (3a), ce contexte gauche inclut (de droite à gauche) l'adverbe *très* qui modifie l'adjectif, le verbe *paraître,* à la forme infinitive, qui est "un verbe support" (Danlos, 1992) pour les adjectifs, le pronom *lui* et finalement le verbe modal *peut* précédé par *il*. Dans (3b), le contexte gauche inclut le verbe support *s'avérer* qui est conjugué à un temps composé (*s'est avéré*) et nié (*ne s'est pas avéré*).

(3)a Il peut lui paraître très difficile de résoudre ce problème
b Il ne s'est pas avéré difficile de résoudre ce problème

De ce fait, pour chaque type de tête lexicale (e.g. adjectif, verbe) qui ancre une construction impersonnelle, on doit déterminer tous les éléments qui peuvent figurer dans son contexte gauche et intégrer ces éléments dans les patrons linguistiques. Cette tâche ne se heurte pas à de réelles difficultés, disons que c'est un travail minutieux et coûteux en temps[2]. Par contre, on se heurte à de réelles ambiguïtés avec le contexte droit de la tête lexicale, comme nous allons le montrer. Dans la suite de cet article, les contextes gauches des têtes lexicales sont présentés de façon simplifiée - comme en (2) - pour faciliter la lecture.

3.2 Contexte droit de la tête lexicale

3.2.1 Ambiguïtés syntaxiques

Les ambiguïtés syntaxiques sont légion dans le contexte droit car, comme il est bien connu, une séquence de parties du discours peut recevoir plusieurs analyses syntaxiques. A titre d'illustration, considérons le patron en (4a), dans lequel le symbole Ω correspond à une séquence non-vide de tokens. Ce patron correspond à deux analyses syntaxiques : (4b) dans lequel *il* est impersonnel et l'infinitive sous-catégorisée par *difficile,* et (4c) dans lequel *il*

---

[2] Ce travail peut être réutilisé dans un outil qui repère la tête lexicale d'une phrase simple, outil qui peut s'intégrer dans la chaîne de traitements d'un analyseur syntaxique modulaire. Ce travail revient à mettre sous forme d'automates la notion d'"amas verbal" de (Gerdes, Kahane, 2004).

est anaphorique et l'infinitive fait partie d'un *GN*. Ces deux analyses sont illustrées respectivement dans les phrases (4b) et (4e) - ces phrases diffèrent seulement par l'adverbe *ici/juste*.

(4)a Il est difficile pour Ω <de V:W>
b Il [IMP] est difficile pour $(\Omega)_{GN}$ <de V:W>
c Il [ANA] est difficile pour $(\Omega$ de < V:W>$)_{GN}$
d Il est difficile pour (les étudiants qui viennent ici)$_{GN}$ de résoudre ce problème
e Il est difficile pour (les étudiants qui viennent juste de résoudre ce problème)$_{GN}$

Pour traiter ces ambiguïtés syntaxiques, une solution consiste à déclarer explicitement qu'un patron comme (4a) est ambigu, ce qui revient à décorer l'occurrence de *il* dans (4a) par la balise [AMB] qui doit être interprétée de la façon suivante : "ILIMP ne peut pas déterminer si *il* est anaphorique ou impersonnel". Cependant cette étiquette n'est d'aucune utilité pour les traitements ultérieurs d'un système de résolution d'anaphores ou d'une chaîne de traitements syntaxiques : elle doit donc être utilisée avec modération. Une autre solution consiste à faire appel à des heuristiques basées sur des fréquences. Par exemple, l'heuristique suivante : les phrases qui suivent le patron de (4a) sont plus fréquemment analysées comme (4b) que comme (4c), par conséquent, *il* dans les phrases qui suivent (4a) peut recevoir la balise [IMP], même si cette balise est fausse dans quelques cas. Nous avons adopté cette dernière solution. ILIMP repose donc sur tout un ensemble d'heuristiques que nous avons établies soit à partir de notre connaissance/intuition linguistique soit à partir d'études quantitatives sur les corpus. L'évaluation de ILIMP révèlera si nos heuristiques sont judicieuses (Section 4).

3.2.2 Ambiguïtés lexicales

Dans un très petit nombre de cas (une dizaine), un item lexical peut ancrer une construction impersonnelle ou personnelle avec le même cadre de sous-catégorisation. C'est le cas pour l'adjectif *certain* construit avec un complément phrastique, comme illustré dans la phrase en (5a). Comme les deux lectures de (5a) semblent également fréquentes, *il* dans le patron (5b) reçoit la balise [AMB].

(5) a Il est certain Fred que viendra (*Jean/Cela est certain que Fred viendra)*
b Il [AMB] est certain que P[3]

3.2.3 Autres difficultés

Le dernier type de difficultés s'observe avec des constructions impersonnelles à sujet nominal extraposé qui ne diffèrent en surface que de façon très subtile par rapport à des constructions personnelles. Voir la paire en (6) qui ne diffère que par *du/de* mais où (6a) est impersonnelle et (6b) personnelle. Voir aussi la paire en (6') qui ne diffère que par les noms *valise/ priorité* mais où (6'a) est impersonnelle et (6'b) personnelle. Nous avons tenté d'établir des heuristiques pour distinguer ces cas, sans toutefois nous lancer, par exemple, dans l'utilisation (périlleuse) de traits sémantiques, comme ± concret.

(6)a Il manque du poivre (dans cette maison)

---

[3] *P* symbolise un patron destiné à représenter une phrase. Il est composé d'une séquence non vide de tokens incluant un verbe fini.

b Il manque de poivre, ce rôti
(6')a Il reste la valise du chef (dans la voiture)
b Il reste la priorité du chef (le chômage)

Pour conclure cette section sur la réalisation de ILIMP, disons qu'il est fait fréquemment appel à des heuristiques afin d'éviter une utilisation abusive de l'étiquette [AMB]. Ces heuristiques peuvent mener à des erreurs qui vont être examinées dans la section suivante.

4 Evaluation de ILIMP

Notre corpus de travail a été *Le Monde*. Plus précisément, un corpus de 3.782.613 tokens extraits du corpus *Le Monde 1994*. UNITEX segmente ce corpus en 71.293 phrases. Il contient 13.611 occurrences du token *il* sur 20.549 occurrences de pronoms personnels sujet de la troisième personne *(il, elle, ils, elles)*. Le pronom *il* est donc le pronom sujet de la troisième personne le plus fréquent, avec un taux de 66 %. De ce corpus, 8544 phrases qui incluent au moins une occurrence de *il* ont été extraites et elles totalisent près de 10.000 occurrences de *il* (une phase complexe enchâssant diverses propositions peut inclure plusieurs occurrences de *il*). Ces phrases ont été données en entrée à ILIMP et les résultats - les balises [IMP], [ANA], et [AMB] - ont été évalués manuellement par des amis, collègues et étudiants[4]. Ces évaluateurs devaient se fier uniquement à leur intuition de locuteur : ils ne devaient pas mettre en œuvre leur éventuelle compétence de linguiste pour essayer de détecter des ambiguïtés virtuelles d'occurrences de *il*. Dans ces conditions, l'attribution d'une balise [IMP] ou [ANA] est immédiate dans quasiment tous les cas, la balise [AMB] ne concernant qu'un nombre négligeable de cas (rappelons, section 3.2.2, qu'elle n'est lexicalement justifiée que pour une dizaine de têtes lexicales). Il ressort de cette évaluation un taux de précision de 97,5%. Nous allons examiner en détail les erreurs, en laissant de côté la balise [AMB].

4.1 Erreurs provenant d'ambiguïtés morphologiques

Les erreurs d'ILIMP provenant d'ambiguïtés morphologiques sont (évidemment) comptabilisées comme les autres erreurs provenant de la réalisation d'ILIMP. Ces dernières seront examinées dans les sections suivantes. Pour l'instant, examinons les erreurs dues au fait que le pré-processing d'UNITEX n'effectue pas de désambiguïsation : ce n'est pas un taggeur (Section 2.2). Considérons le patron en (7a) : rappelons (note 3) que P représente une séquence non vide de tokens qui inclut un verbe fini ; <V6:K> couvre les verbes de la table 6 au participe passé, e.g. *choisi*. Le patron en (7a) est destiné à couvrir les phrases impersonnelles comme (7b). Néanmoins, il couvre aussi (7c), où le pronom *il* est donc balisé à tort [IMP]. Cette erreur est due au fait que le dictionnaire DELAF comprend à juste titre deux entrées pour le mot *mètres* - forme finie du verbe *métrer* et pluriel du nom mètre - et qu'UNITEX ne fait aucune distinction entre ces deux entrées. La séquence *l'acier ou le béton pour soutenir une toiture de 170 mètres* correspond donc au patron P (elle contient un verbe fini).

(7)a Il [IMP] <avoir.V:3s> été <V6:K> (ADV) que P

---

[4] A savoir Isabelle Faugeras, Annie Meunier, Christian Leclère, Laurence Delort, Ane Dybro-Johansen, François Lareau, Alexis Nasr, Céline Raynal, Jacques Steinlin, François Toussenel et Mélodie Soufflard, que nous remercions chaleureusement.

b Il a été choisi que les séances se feraient le matin vers 9h
c Il a été choisi plutôt que l'acier ou le béton pour soutenir une toiture de 170 mètres

Tout taggeur devrait attribuer au mot *mètres* de (7c) une étiquette nominale. Si ILIMP travaillait non pas sur du texte brut mais sur la sortie d'un taggeur, l'erreur de balisage de *il* dans (7c) serait donc évitée. Cette stratégie est envisageable[5], mais ILIMP serait alors tributaire des erreurs d'un taggeur. D'une manière plus générale, en admettant qu'un système d'analyse syntaxique repose sur une approche modulaire séquentielle où collabore "en pipe-line" tout un ensemble de modules - taggeur, reconnaisseur d'entités nommées, ILIMP, chunker, etc.- la question se pose de savoir dans quel ordre enchaîner ces modules. Mais laissons cette question ouverte et revenons aux erreurs d'ILIMP travaillant sur un texte brut.

4.2 *il* balisé à tort [IMP] au lieu de [ANA] : 0,3%

Très peu d'erreurs : 33. Ce faible taux d'erreur est surprenant vu le recours fréquent à des heuristiques "brutales". A titre d'illustration, nous avons mis la balise [IMP] dans le patron *il y <avoir.V:3s>* avec ses variantes de contexte gauche. Cette heuristique ne donne lieu qu'à deux erreurs, citées en (8), sur environ 1500 phrases qui suivent ce patron où *il* est correctement balisé [IMP].

(8)a Il s'était réfugié en Angleterre. Il y avait très tôt connu les pianos de ….
b Il revient de Rimini. Il y a donné la réplique à Madeleine.

4.3 *il* balisé à tort [ANA] au lieu de [IMP] : 2%

Plus d'erreurs. Les erreurs de ce type viennent de ce que [ANA] est la balise par défaut : elles viennent donc de lacunes dans l'ensemble des patrons composant ILIMP. Parmi celles-ci, on peut d'abord distinguer celles dues à de la paresse/lassitude/manque de temps. Par exemple, nous avons autorisé des guillemets à certains endroits mais pas partout, de ce fait *il* est balisé à tort [ANA] dans (9a). De même, nous avons écrit certains automates pour traiter les cas avec inversion du sujet, mais nous n'avons pas pris le temps de les écrire tous, d'où l'erreur en (9b). Et nous avons fait l'impasse totale sur toute coordination, d'où (9c).

(9)a Il [ANA] était " même souhaitable " que celui-ci soit issu " de l'opposition ".
b Est-il [ANA] inconcevable que ...
c Il [ANA] est donc indispensable et légitime de les aider …

Un second type d'erreurs provient de lacunes lexicales dans nos patrons. Dans l'état actuel d'ILIMP, il manque principalement des adjectifs ancrant une construction impersonnelle, car les adjectifs n'ont pas été étudiés de façon aussi systématique que les verbes dans le lexique-grammaire. La liste des 682 adjectifs ancrant une construction impersonnelle – à sujet phrastique extraposable - peut donc encore être complétée. Un relecteur anonyme de cet article pose la question de savoir si on peut effectivement recenser les adjectifs ayant un usage dans une construction impersonnelle. Il/Elle note qu'un adjectif comme *myope*, qui *a priori* n'ancre pas de construction impersonnelle, pourrait en fait permettre une phrase impersonnelle comme (10).

---

[5] Elle demande qu'UNITEX soit adapté pour prendre en compte les résultats d'un taggeur, ce qui a été fait par Patrick Watrin pour Tree Tagger.

(10) Il semble tout à fait myope, voire aveugle, de penser que la situation ne peut se détériorer

Si cette phrase est joliment construite et compréhensible, elle nous semble quand même inacceptable (même en tentant de l'améliorer en ajoutant *de sa part* sur le modèle de *Il est idiot de sa part d'avoir refusé cette offre*). Nous sommes plusieurs linguistes à juger (10) aussi inacceptable que **Cette action/idée est myope*[6]. Est-ce à dire qu'il est exclu de trouver une phrase comme (10) en corpus ? Totalement exclu, peut-être pas, car elle présente un joli effet de style. Mais il semble qu'on peut affirmer qu'elle est hautement improbable. De ce fait, *myope* et *aveugle* peuvent être exclus de la liste des adjectifs ancrant une construction impersonnelle. De manière plus générale, nous pensons que la liste des adjectifs (ou verbes) ancrant une construction impersonnelle n'est pas ouverte aux glissements de sens : c'est une liste fermée, donc recensable.

Un troisième type d'erreurs provient de lacunes syntaxiques. En particulier, nous avons considéré obligatoire un sujet phrastique extraposé, alors qu'il existe des cas où il n'est pas réalisé, par exemple dans des expressions en *comme*, comme en (11). Nous avons ajouté certaines de ces expressions, mais nous n'avons pas mené d'étude linguistique pour connaître l'étendue du phénomène, nous avons donc des lacunes syntaxiques.

(11) Comme il / a été annoncé / conviendrait / arrive souvent / est bien connu

Un quatrième et dernier type d'erreurs concerne les constructions impersonnelles à sujet profond nominal. D'une part, comme nous l'avons expliqué en 3.2.3, ces constructions sont délicates à repérer pour des verbes courants comme *manquer* ou *rester*. D'autre part, nous n'avons écrit aucun patron pour repérer les formes impersonnelles avec un verbe au passif ou au se-moyen relevant d'un niveau de langue châtié, voir section 2.1, d'où des erreurs comme en (12).

(12) Il [ANA] s'était formé un cercle d'inimitié autour de cet individu abject …

Les trois premiers types d'erreurs sont évitables à faible coût, mais il n'en est pas de même pour le quatrième type.

4.4 Autres erreurs : 0,2%

Les autres erreurs proviennent de ce que le mot *il* n'est pas employé comme pronom sujet, mais par exemple comme élément d'un nom propre étranger, (13a)[7]. Il y a aussi des fautes de frappe/d'orthographe, (13b) et (13c). Ces erreurs sont imparables en partant d'un texte brut.

(13)a Cela a commencé dans la seconde moitié du 18ème, quand, à Milan, se publie cette revue illuministe appelée Il [ANA] Caffè
b Il [ANA] y vingt-cinq ans
c Puis il [ANA] ont franchi les obstacles dans les bois

---

[6] On peut avancer l'hypothèse d'une corrélation entre la possibilité pour un adjectif d'ancrer une construction impersonnelle et celle d'avoir un sujet abstrait comme *action* ou *idée*, voir la paire *Il est abracadabrantesque de penser que Fred va voter non* et *Cette idée est abracadabrantesque*.

[7] Si ILIMP prenait en entrée un texte où les entités nommées sont reconnues, l'erreur en (13a) serait évitée puisque la séquence *Il Caffè* serait reconnue comme une entité nommée. Dans le présent article, le mot *il* appartient souvent à la méta-langue et n'est alors pas employé comme pronom sujet.

4.5 Evaluation sur des corpus de genre différent

Nous avons aussi réalisé une évaluation d'ILIMP sur des textes littéraires du XIXème siècle concernant 1858 occurrences de *il*. Le taux de bons résultats baisse par rapport au genre journalistique : il passe de 97,5% de bons résultats à 96,8%. Cette baisse est due d'une part à des tournures impersonnelles qui ne sont plus usitées, voir (14), d'autre part à un nombre élevé de phrases impersonnelles avec inversion du sujet comme en (9b), cas que nous n'avons pas pris le temps de traiter systématiquement.

(14) Mais peut-être était-il un peu matin pour organiser un concert, …

Le pourcentage de *il* impersonnel dans ces textes littéraires augmente par rapport au corpus *Le Monde* : il passe de 42% à 49,8%. D'une manière générale, nous nous attendons à des différences importantes de pourcentage de *il* impersonnel selon les corpus[8], mais nous ne nous attendons pas à des différences significatives sur le taux de bons résultats d'ILIMP (surtout si nous prenons le temps de corriger les trois premiers types d'erreurs signalés dans la section 4.2) : nous pensons en effet que les têtes lexicales des constructions impersonnelles forment une liste *fermée* (voir section 4.2) et *stable* quel que soit le corpus.

5 Conclusion et recherche future

La méthode employée dans ILIMP pour repérer les occurrences de *il* impersonnelles ou anaphoriques, qui donnent des résultats satisfaisants, peut être utilisée pour d'autres langues et pour d'autres tâches. Nous avons déjà signalé (note 2) qu'on peut dériver d'ILIMP un outil repérant la tête lexicale d'une phrase simple. On peut aussi envisager d'utiliser ILIMP pour enrichir le calcul des fonctions syntaxiques en identifiant les "sujets profonds extraposés". Par ailleurs, la méthode peut être utilisée pour désambiguïser des mots aussi fréquents et ambigus que *il*. Ainsi, (Jacques, 2005) utilise une méthode similaire à la nôtre dans la première étape du traitement qu'elle propose pour désambiguïser le mot *que* : sans même utiliser la richesse du lexique-grammaire du LADL, elle augmente le taux de précision sur l'étiquetage de ce mot de 14% par rapport à Tree Tagger (passage de 75% de bons résultats à 89%).

On peut s'interroger sur la pertinence de ces "petits" outils qui se cantonnent sur un seul mot ou sur une fonctionnalité bien particulière. Ils sont certes bien modestes par rapport à un système qui produirait, pour n'importe quelle phrase donnée en entrée, une (et une seule) analyse syntaxique, complète, et ce, avec un taux de précision avoisinant les 98%. Mais force est de constater qu'un tel système n'existe pas encore. Nous avancerons donc pour la défense des petits outils le célèbre dicton : *Les petits ruisseaux font les grandes rivières*. Il reste à canaliser ces ruisseaux, i.e. à organiser un vaste effort de recherche pour déterminer comment ordonnancer les "petits" outils dans une chaîne de traitement aboutissant à une analyse syntaxique robuste, complète et correcte.

---

[8] Le quotidien *Le Monde* contient un certain nombre de longs articles racontant la vie et l'œuvre de personnes célèbres. Ces articles, quand ils concernent une personne célèbre de sexe masculin, enchaînent de nombreuses occurrences de *il* anaphorique référant à cet homme. On peut s'attendre à ce que le pourcentage de *il* impersonnel augmente dans les journaux traitant seulement d'actualité ou d'économie.

Remerciements

Je remercie Eric Laporte qui m'a fourni toute la logistique nécessaire pour réaliser ce travail à l'Université Marne-la-Vallée, et Olivier Blanc pour l'assistance technique vitale qu'il m'a apportée. Je remercie aussi Sylvain Kahane et Alexis Nasr pour leurs commentaires fructueux sur cet article.

Références

BGL : BOONS J-P., GUILLET A., LECLERE C. (1976a), *La structure des phrases simples en français: constructions intransitives*. Genève: Droz, 378 p. BGL

BGL : BOONS J-P., GUILLET A., LECLERE C. (1976b), La structure des phrases simples en français: classes de constructions transitives. Rapport de Recherches du LADL n° 6, Paris: Université Paris 7.

COURTOIS B. (2004), Dictionnaires électroniques DELAF anglais et français, *Syntax, Lexis and Lexicon-Grammar. Papers in honour of Maurice Gross* , Lingvisticæ Investigationes Supplementa 24, Amsterdam/Philadelphia : Benjamins, pp. 113–133.

DANLOS L. (1980), *Représentation d'informations linguistiques: les constructions N être Prép X.* Thèse de troisième cycle, Paris: Université Paris 7.

DANLOS L. (1992), Support Verb Constructions: linguistic properties, representation, translation, *Journal of French Linguistic Studies*, Vol. 2, n°1, Cambridge University Press, Cambridge.

EVANS R. (2001), Applying Machine Learning toward an Automatic Classification of *it*, *Literary and Linguistic Computing*, Vol. 16, n°1, pp. 45-57.

GERDES K., KAHANE S. (2004), L'amas verbal au cœur d'une modélisation topologique du français, *Journées de la syntaxe : ordre des mots dans la phrase française, positions et topologie,* Bordeaux, 8.

GROSS M. (1975), *Méthode en syntaxe*, Paris, Hermann.

GROSS M. (1993), Les phrases figées en français, *L'information grammaticale 59*, Paris, p. 36-41.

GROSS M. (1994), Constructing Lexicon-Grammars, *Computational Approaches to the Lexicon*, Oxford, Oxford University Press, p. 213-263.

JACQUES M.P. (2005), *Que* : la valse des étiquettes, *Actes de TALN 05*, Dourdan.

KENNEDY C., BOGURAEV B. (1996), Anaphora for Eveyone; Pronominal Anaphora Resolution without a Parser, in *COLING'96*, Copenhagen.

LAPIN S., LEASS H.J. (1994), An algorithm for pronominal anaphora resolution, *Computational Linguistics*, 20(4), p. 535-561.

LECLERE . C. (2003), The lexicon-grammar of French verbs: a syntactic database, In *Proceedings of the First International Conference on Linguistic Informatics,* Kawaguchi Y. et alii (eds.), UBLI, Tokyo University of Foreign Studies.

MEUNIER A. (1981), *Nominalisations d'adjectifs par verbes supports*. Thèse de troisième cycle, LADL, Université Paris 7.

NASR A. (2004), *Analyse syntaxique probabiliste pour grammaires de dépendances extraites automatiquement,* Habilitation à diriger des recherches, Université Paris 7

PICABIA L. (1978), *Les constructions adjectivales en français*, Genève: Droz.

SILBERZTEIN M. (1994), INTEX: a corpus processing system, in *COLING'94*, Kyoto, Japon, vol. 1, pp. 579-583.