# Vers une identification automatique du chiasme de mots

Marie Dubremetz
Uppsala universitet, Institutionen för lingvistik och filologi - Box 635 - 751 26 Uppsala, Suède
Université Paris Ouest La Défense - 200, Avenue de la République - 92001 Nanterre, France
marie.dubremetz@lingfil.uu.se

RÉSUMÉ

Cette recherche porte sur le chiasme de mots : figure de style jouant sur la réversion (ex. « Bonnet blanc, blanc bonnet »). Elle place le chiasme dans la problématique de sa reconnaissance automatique : qu'est-ce qui le définit et comment un ordinateur peut le trouver ? Nous apportons une description formelle du phénomène. Puis nous procédons à la constitution d'une liste d'exemples contextualisés qui nous sert au test des hypothèses. Nous montrons ainsi que l'ajout de contraintes formelles (contrôle de la ponctuation et omission des mots vides) pénalise très peu le rappel et augmente significativement la précision de la détection. Nous montrons aussi que la lemmatisation occasionne peu d'erreurs pour le travail d'extraction mais qu'il n'en est pas de même pour la racinisation. Enfin nous mettons en évidence que l'utilisation d'un thésaurus apporte quelques résultats pertinents.

ABSTRACT

**Towards an automatic identification of chiasmus of words**

This article summarises the study of the rhetorical figure "chiasmus" (e.g : "Quitters never win and winners never quit."). We address the problem of its computational identification. How can a computer identify this automatically ? For this purpose this article will provide a formal description of the phenomenon. First, we put together an annotated text for testing our hypothesis. At the end we demonstrate that the use of stopword lists and the identification of the punctuation improve the precision of the results with very little impact on the recall. We discover also that using lemmatization improves the results but stemming doesn't. Finally we see that a French thesaurus provided us with good results on the most elaborate form of chiasmus.

MOTS-CLÉS : chiasme, rhétorique, antimétabole, figure de style.

KEYWORDS: chiasmus, rhetoric, antimetabole, stylistic device.

## 1 Introduction

### 1.1 Situation de la recherche

Élevée au rang de science dans l'antiquité, la rhétorique, ou l'étude des techniques de persuasion au moyen du language, a, peu à peu, été délaissée par les recherches linguistiques modernes (Harris et DiMarco, 2009). Parmi les moyens disponibles en rhétorique pour convaincre, on dispose de ce qu'on appelle les figures de style ou figures de rhétorique. Selon Harris et DiMarco (2009) on peut les diviser en deux catégories : les figures de style reposant sur la sémantique

 

(métaphore, métonymie, comparaison...), et les autres qui jouent sur la syntaxe, les phonèmes ou tout autre constituant de la langue (rime, paronomase...). Si les premières sont relativement populaires, les autres restent un sujet de recherche très rare dans l'horizon académique du TAL. Cet article contribue à combler ce vide en proposant une méthode de détection d'une de ces figures de style : le chiasme de mots. Grâce à des observations empiriques et à l'ajout de contraintes simples (« StopWords » , analyse des ponctuations) nous pensons pouvoir améliorer la qualité des algorithmes existants. Enfin grâce aux outils que nous fournit le TAL (lemmatiseur, raciniseur, thésaurus) nous pourrons envisager de détecter une plus grande variété de chiasmes.

Après avoir présenté les applications liées à notre objet d'étude nous présenterons une définition et le corpus utilisé. De là nous ferons l'état de l'art des recherches existantes avant de soumettre, comparer et enfin discuter notre méthode de détection.

### 1.2 Applications : pourquoi la question du chiasme en TAL ?

L'application la plus évidente de ce type de recherche semble, à première vue, l'analyse de discours. Ainsi Gawryjolek (2009) avait déjà mis en pratique son outil de détection des figures portant sur la répétition pour analyser un discours de Barack Obama.

On peut aussi imaginer qu'un outil de détection contribuera, à terme, non pas à analyser le style mais à le générer via l'assistance à la rédaction. En effet les logiciels de traitement de textes nous assistent déjà en suggérant les synonymes. À terme le relevé d'un très grand nombre de chiasmes pourrait permettre la constitution de bases de données des figures de style indexées par mots clefs. Pourquoi alors ne pas imaginer un jour que l'utilisateur pourra écrire tout en se voyant suggérer une figure de style appropriée ? Contrairement à ce qu'on pourrait penser, le chiasme n'est pas qu'une coquetterie de style destinée aux seuls poètes. Il est, au contraire, un procédé rhétorique utilisé dans tout texte argumentatif et ce, même s'il est de nature scientifique. [1]

Enfin, il convient d'évoquer une autre application possible, celle de l'extraction automatique de citations (Bendersky et Smith, 2012). Le chiasme en effet génère souvent d'excellents jeux de mots ce qui valorise le texte d'où il est extrait. Étant donnée la nécessité, aujourd'hui, de traiter un grand nombre d'œuvres écrites, repérer non plus juste les mots clefs mais aussi les parties les plus travaillées d'un texte devient de plus en plus important. L'observation du style apporte une nouvelle dimension qui n'existe pas ou peu dans le traitement automatisé.

À présent que nous avons entr'aperçu l'enjeu comment définir notre objet d'étude ?

## 2 Définition

Le mot chiasme tire son nom de la lettre $\chi$ en référence à la croix qu'elle symbolise. On le définit en effet comme la reprise d'un couple d'éléments en sens inverse :

1. « Très répandu dans les années 70, [le chiasme] a été dénoncé vertement pour la violence qu'il fait à la fonction communicative du langage : « [...]. La recherche du sens c'est le sens de la recherche, etc. Vous pouvez paraître profond avec n'importe quelle banalité». Mais en donnant à penser au lecteur naïf, ce chiasme propositionnel [...], est souvent approprié pour des titres, à la fois par son économie lexicale et par la profondeur apparemment inépuisable des discussions qu'il annonce. Dans *La trouble-fête*, Bernard Andrès épingle le procédé comme typique du jargon universitaire, susceptible d'entraîner considération et subventions. »(Vandendorpe, 1991, p.4)

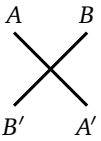


FIGURE 1 – Schéma définitoire du chiasme

On citera ainsi l'exemple le plus classique : « Bonnet blanc, blanc bonnet »

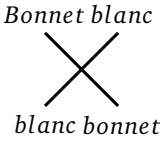


FIGURE 2 – Exemple de chiasme

Il existe toute sorte de chiasmes, ils peuvent porter sur le croisement de phonèmes, de lettres, d'éléments syntaxiques... Nous nous restreindrons à l'étude des chiasmes de mots. Cependant on constate dans cette sous-classe de chiasmes encore beaucoup de variantes qu'il faut définir pour le traitement automatique (Gawryjolek, 2009, p.67). Voici donc, à l'issue des lectures des linguistes [2] et de l'observation des exemples qu'ils donnent, les types de chiasmes de mots que nous avons recensés :

2. Diderot et D'Alembert (1789); Dupriez (2003); García-Page (1991); Greene (2012); Nordahl (1971); Pougeoise (2001); Rabatel (2008); Vandendorpe (1991); Van Gorp *et al.* (2001)

| | Nom | Répétition porte sur : | Exemple |
|---|---|---|---|
| **Les chiasmes de mots** $\begin{matrix} A\ B \\ \times \\ B'\ A' \end{matrix}$ | Antimétabole ou Antimétabole stricte | mots identiques [3] | *Ceux qui ont l'opinion que l'argent peut tout faire pourraient bien être suspectés de tout faire pour de l'argent.* (George Savile) |
| | Chiasme flexionnel ou Antimétabole flexionnelle | mots fléchis | *Lui, veut une alliance des contraires, c'est-à-dire le contraire d'une alliance.* (Le Monde, 13-3-2007) |
| | Chiasme dérivationnel | mots dérivés | *Moderniser l'Islam plutôt qu'islamiser la modernité.* (Jean Daniel) |
| | Antimétalepse | idée ou notion sans forcément de reprise morphologique [4] | *dernière averse [. . . ] neiges d'antan* (Georges Brassens, *Le temps ne fait rien à l'affaire*) |

TABLE 1 – Typologie des chiasmes de mots

Les figures décrites par ce tableau sont très anciennes et connues depuis longtemps. Ainsi, on observe déjà à l'époque de Quintilien (Greene, 2012, Art. « Antimétabole ») l'existence des antimétaboles comportant des mots identiques mais aux flexions différentes. Cependant jamais un nom spécifique n'a été donné à ce phénomène pour le différencier de l'antimétabole « stricte ». Il en va de même pour le chiasme portant sur des mots dérivés. Ce tableau n'est sans doute pas exhaustif puisqu'il est tributaire d'observations empiriques à partir des exemples donnés par les ouvrages de stylistique de référence. Il va cependant permettre à cette recherche d'aller plus loin et de ne pas se heurter au « flottement »(Rabatel, 2008, p.21) qui règne autour de la définition de cette figure : car sans définition formelle, il s'avère difficile d'élaborer un système de détection (Gawryjolek, 2009, p.67).

3. (Dupriez, 2003; Pougeoise, 2001, Art."Chiasme")
4. (Diderot et D'Alembert, 1789, Art. "Antimétabole, Antimétalepse, Antimétathèse")

# 3 Corpus

## 3.1 Mode de constitution

Pour évaluer une méthode de détection des chiasmes, l'idéal serait l'obtention d'un corpus déjà annoté. Cependant le chiasme reste un phénomène linguistique relativement rare ce qui rend la constitution d'un corpus difficile et explique en partie l'absence de ressources pour traiter le phénomène. C'est donc grâce aux relevés précis des exemples donnés par les dictionnaires spécialisés (Pougeoise, 2001; Dupriez, 2003), par la base de donnée JULIBEL [5] (Bradfer *et al.*, 1995) par les quelques linguistes ayant traité spécialement de cette figure (García-Page, 1991; Rabatel, 2008; Nordahl, 1971) ou parfois par la lecture d'extraits de textes d'auteurs célèbres tombés dans le domaine public (Hugo, Musset etc.) que nous avons réussi à constituer ce corpus [6]. Le corpus est en langue française si possible non traduite, il est constitué de 43 chiasmes (22 antimétaboles strictes, 8 chiasmes flexionnels, 6 chiasmes dérivationels, 7 antimétalepses). Lorsque nous en avions la possibilité, nous avons ajouté plusieurs pages du contexte d'origine où est apparu le chiasme et vérifié soigneusement que ce contexte ne contenait pas d'autres chiasmes. Concaténés ces chiasmes avec contextes constituent un texte d'entraînement de 16000 mots. Cette taille est déjà assez grande, nous le verrons, pour donner des comparatifs clairs sur l'efficacité des méthodes de détection. Elle reste toutefois assez réduite pour que nous ayons pu vérifier manuellement le nombre précis de chiasmes dans le texte et leur position. Avec toutes ces données des calculs de rappel et de précision seront praticables.

## 3.2 Contenu

Le corpus sur lequel nous testons est extrêmement varié, à l'image des contextes dans lesquels on peut retrouver cette figure de style. On peut y relever ainsi :

– des textes de littéraires :

> « - Que tu es heureux d'être fou !
> - Que tu es fou de ne pas être heureux ! », (Alfred de Musset, *Les caprices de Marianne*)

– des textes journalistiques :

> « L'économie du discours ne peut se réduire au discours de l'économie. » (*Libération*, paroles rapportées de Michel Debray)

– des textes publicitaires :

> « [A la fin du clip, message au bas de l'écran] "Ne vivez pas pour nettoyer Nettoyez pour vivre " » (RTBF, octobre 2001, clip publicitaire pour *Monsieur Propre*)

Nous mettons à disposition ce corpus sous la forme d'un tableau afin de fournir les métadonnées indispensables à une étude complète. Un chiasme se présentera dans notre tableau sous la forme présentée en table 2.

5. Ressource linguistique destinée à l'enseignement du français `http://home.scarlet.be/lmdp/julibelmoded'emploi.html`, consulté le 08/05/2013

6. Le corpus est mis à libre disposition sur `http://stp.lingfil.uu.se/~marie/chiasme.htm`, consulté le 08/05/2013

| Auteur | Citation | Contexte | Source | Chiasme de catégorie : |
|---|---|---|---|---|
| Cédric Flament | Festival de prix, prix de festival | 58e Festival de Cannes * Festival de prix, prix de festival La Face "cachet" de la Croisette | L'Avenir du Luxembourg 20.05.2005 p.14 | 1 |

TABLE 2 – Présentation d'un chiasme dans notre tableau

La dernière case est particulière à notre étude. À chaque chiasme nous attribuons sa typologie grâce à un numéro qui correspond aussi au degré de difficulté d'identification en TAL. Voici la légende de cette typologie numérotée (Figure 3).

| | |
|---|---|
| Catégorie 1 : | Mots strictement identiques, simple repérage de chaînes de caractères identiques |
| Catégorie 2 : | Mots fléchis, lemmatisation nécessaire |
| Catégorie 3 : | Mots dérivés d'une même racine, racinisation nécessaire |
| Catégorie 4 : | Rapprochement sémantique uniquement |

FIGURE 3 – Typologie des chiasmes : légende

Nous obtenons ainsi un texte d'entraînement dont les chiasmes sont classés selon le type de traitement nécessaire. Cette classification permet de pratiquer des évaluations spécifiques des algorithmes sur un type de chiasme en particulier.

# 4 Les méthodes existantes pour la détection des antimétaboles et leur application sur notre corpus

À ce jour il n'existe que deux recherches en TAL qui traitent du phénomène du chiasme, cependant elles ne traitent que de l'antimétabole et s'inspirent du formalisme de Harris et DiMarco (2009) pour définir la figure recherchée (cf. Table 3).

| Élément | Signification |
|---|---|
| w | Mot |
| ... | Suite de caractères |
| <...> | Frontière de phrase ou de proposition |
| Indice $_{abc}$ | Indique la même identité ou non entre les éléments |

TABLE 3 – Formalisme des figures de style [traduit de Harris et DiMarco (2009, p.3)]

Ce formalisme permet à Harris et DiMarco (2009) de définir les antimétaboles ainsi :

$<W_A \ldots W_B \ldots W_B \ldots W_A>$ (Harris et DiMarco, 2009, p.4)

Cette formule signifie qu'une antimétabole est un ensemble de deux couples de mots identiques disposés en inclusion et séparés éventuellement par n'importe quels autres éléments de la langue (autres mots, ponctuations...).

## 4.1 Algorithme d'identification des doubles paires en inclusion (Gawryjolek, 2009)

En suivant la définition de Harris et DiMarco (2009), Gawryjolek (2009) a mis au point un logiciel qui repère les antimétaboles strictes. Il sélectionne pour cela toutes les doubles paires de mots en inclusion sur une plage spécifiée du texte. Il précise que cette méthode obtient cent pour cent de rappel sur les antimétaboles ce qui, sans surprise, est vérifié sur notre corpus. Il précise aussi qu'il ne filtre aucune paire de mots, même les plus fréquentes, et que la précision est basse sans donner de chiffre ou d'estimation. Nous ne disposons pas du code source de Gawryjolek (2009) ni de la définition de ce qu'il appelle une « plage spécifiée ». Nous avons donc reproduit l'algorithme proposé en limitant la fenêtre de détection à 30 tokens (le plus grand chiasme que nous ayons observé en comportait 23). Concrètement cela signifie que si un chiasme $<W_A \ldots W_B \ldots W'_B \ldots W'_A>$ comporte plus de 30 tokens entre les éléments « $W_A$ » et « $W'_A$ » il est éliminé : c'est une réversion due au hasard sans volonté de provoquer une figure de style, nous appellerons ce phénomène « pseudochiasme ». Dans les conditions que nous avons décrites et sur notre corpus de 16000 mots nous avons obtenu une précision inférieure à 2 % (0.017). Ce manque de précision montre la complexité du problème : une détection efficace des antimétaboles ne peut pas se limiter au repérage de deux couples de mots identiques sans autre forme de sélection.

## 4.2 Méthode par identification de patrons de trois paires de mots (Hromada, 2011)

Hromada (2011) avec le formalisme de Harris et DiMarco (2009) retient une autre définition de l'antimétabole :

$<W_A\ W_B\ W_C \ldots W_C\ W_B\ W_A>$

Cette formulation signifie qu'une antimétabole est une réversion de trois couples de mots successifs (exemple : « Le pouvoir du discours et le discours du pouvoir) séparés au centre par plusieurs caractères symbolisés par « ... ». Grâce à une expression régulière, il va sélectionner les répétitions non pas de deux mais de trois couples de mots successifs en réversion. Les chiasmes que cette expression régulière sélectionne ont aussi pour contrainte de ne jamais comporter de ponctuation forte. Hromada (2011) ne disposait malheureusement pas d'un corpus manuellement annoté pour évaluer ses résultats. Nous avons donc testé aussi son expression régulière sur notre corpus. Les résultats obtenus sont à l'opposé de ceux de l'algorithme de Gawryjolek (2009) puisque, certes, une antimétabole sur deux n'est pas identifiée (11 sur 22) mais la précision en revanche est parfaite. L'expression recherchée présente donc tant de contraintes qu'elle arrive à ne générer, sur notre corpus, aucun faux positif. Nous n'allons pas reprendre l'idée des trois paires de mots à identifier au lieu de deux car cette mise en œuvre est très restrictive. Cependant

cette méthode se fonde sur une bonne observation : celle que le chiasme en plus des deux paires de mots principaux (comme dans notre exemple les termes « pouvoir » et « discours ») entraîne souvent avec lui la réversion d'autres couples de mots secondaires dans la proposition comme les adverbes ou les déterminants (ainsi le mot « du » dans l'exemple).

# 5 Méthode proposée

Les algorithmes que nous venons d'étudier et évaluer sont encore assez limités. Grâce aux évaluations précédentes, nous savons que pour détecter les antimétaboles nous avons à notre disposition le choix entre un algorithme exhaustif mais très peu précis (Gawryjolek, 2009) et un algorithme précis mais au rappel beaucoup moins important (Hromada, 2011). L'équilibre entre rappel et précision est en effet difficile à trouver. De plus, dans ces méthodes, la moindre flexion telle qu'un pluriel suffit à ne plus détecter les chiasmes. Nous allons donc tout d'abord mettre au point un algorithme qui concilie rappel et précision avant d'appliquer des outils de TAL pour couvrir d'autres types de chiasmes.

## 5.1 Les antimétaboles

### 5.1.1 Méthode

Notre algorithme de détection des antimétaboles s'inspire directement de celui de Gawryjolek (2009). Cependant contrairement à ce dernier, nous prenons en compte les données situées entre les mots répétés. Cela nous permet d'ajouter plusieurs contraintes. Le problème en effet de l'algorithme ne repérant que deux paires de mots en inclusion est qu'il repère bien trop de répétitions de mots non pertinentes (par exemple : « Les chats des villes attrapent des souris mais ne les mangent pas »). Voici donc comment nous procédons pour éliminer ces pseudochiasmes :

1. Nous rejetons les répétitions portant sur les mots les plus courants grâce à une liste de « mots vides » (ou « StopWords »). Nous avons introduit une liste de mots à exclure incluant les déterminants et les mots outils les plus courants (articles, auxiliaires, pronoms et quelques adverbes).
2. Nous éliminons les chiasmes en cas de ponctuation forte à l'intérieur des membre gauche ou droit du chiasme. Nous avons prolongé la réflexion de Hromada (2011) sur les ponctuations. Nous pensons que la ponctuation doit être prise en compte mais contrairement à son expression régulière nous allons tolérer la ponctuation forte dans certains cas. Dans nos exemples en effet on observe que la ponctuation si elle est présente à l'intérieur du chiasme, se trouve le plus souvent, dans la partie centrale (exemple : « Le Parti socialiste est un parti sans leader. François Bayrou est un leader sans parti »). Cela signifie que dans un chiasme $<W_A \ldots W_B \ldots W'_B \ldots W'_A>$ il pourra se trouver un signe de ponctuation forte mais seulement entre les mots $W_B$ et $W'_B$. Dans le cas contraire, nous considérerons que le chiasme n'est qu'un pseudochiasme. Cette disposition de la ponctuation se justifie par le fait que le chiasme joue souvent sur un effet de symétrie. Ainsi la position centrale du signe de fin de phrase renforce cet effet.
3. Enfin nous conservons la plage de 30 tokens que nous avions utilisée pour l'algorithme de Gawryjolek (2009) (cf. 4.1).

#### 5.1.2 Résultat

Certes cet ajout de contraintes a baissé le rappel par rapport à l'algorithme de Gawryjolek (2009) : sur 22 antimétaboles à retrouver 21 ont été détectées. Le filtre par « StopWords » en effet est une méthode encore trop rudimentaire pour repérer la phrase de Dumas :

> « Tous pour un, un pour tous. »

Cependant le bénéfice sur la précision est très significatif puisque nous passons de 2 à 72 % de précision. Pour trouver 21 chiasmes notre programme ne relève que 36 extraits au lieu de 1235 extraits pour celui de Gawryjolek (2009).

## 5.2 Les autres types de chiasmes

#### 5.2.1 Les antimétaboles flexionnelles

Pour détecter les chiasmes flexionnels nous avons repris l'algorithme 5.1.1 à la seule différence que nous avons préalablement lemmatisé le texte et les « StopWords » grâce au programme *Treetagger* (Schmid, 1994) associé au lemmatiseur pour le français *Flemm*(Namer, 2000). Nous avons pratiqué le test sur non seulement les antimétaboles flexionnelles mais aussi sur les antimétaboles strictes. Les résultats s'avèrent excellents. 27 antimétaboles sur 30 (22 strictes 8 flexionnelles) ont été trouvées soit un rappel de 90 % et une précision de 46 %. Dans les chiasmes non reconnus on relève :

> « celui qui a le sens de la formule qui ne formule pas beaucoup de sens »

Cette erreur est due non pas à une sous performance du lemmatiseur mais à la nature homonymique du lien entre les deux occurrences du mot « formule ». Ainsi « formule » dans les deux cas a été lemmatisé correctement tantôt sous la forme « formule » (nom) tantôt sous la forme « formuler » (verbe) ce qui empêche l'identification. Dans ce cas précis, la détection avec lemmatisation préalable ne remplace pas la détection sans lemmatisation. À noter, à l'inverse, que les erreurs du lemmatiseur ne gênent pas forcément la détection, ainsi le chiasme :

> « Trop honnête pour être poli que trop poli pour être honnête ! »

lemmatisé par *Flemm* en « Trop honnête pour être **polir** que trop **polir** pour être honnête ! » sera toujours repéré puisque l'erreur provoquée sur « poli » ne change pas l'identité entre les deux occurrences du mot. Ainsi une contre-performance du lemmatiseur n'est pas forcément nuisible à la tâche de détection des chiasmes.

La deuxième erreur de rappel est due à notre filtre par position des ponctuations, l'extrait suivant en effet n'a pas pu être repéré :

> « Faut-il te parler franchement ? ne te riras-tu pas de moi ?
> - Laisse-moi rire de toi, et parle franchement. » Musset, *Les caprices de Marianne*

C'est ici le premier point d'interrogation qui a suscité l'erreur.

Enfin le troisième chiasme non repéré est toujours la devise *des Trois Mousquetaires* (cf. 5.1.2).

### 5.2.2 Les chiasmes dérivationnels

Pour ce type de chiasme la méthode a été la même que pour son homologue flexionnel à la différence que nous avons ajouté un traitement sur les « StopWords » et sur l'ensemble du texte : la racinisation via le programme *Snowball*[7] (Agichtein et Gravano, 2000). Le résultat pour cet outil est défavorable puisque aucun des six chiasmes nécessitant une racinisation n'a été trouvé. Il y a sur ou sous-racinisation de sorte que les termes du chiasme n'obtiennent jamais la même racine. Est-ce qu'un lemmatiseur à base de dictionnaire et non de règles comme snowball serait plus efficace ? A-t-on vraiment en français les ressources suffisantes pour trouver des chiasmes comme « la régification des stars, la starification des rois » ? On est en présence de jeux morphologiques et étymologiques nécessitant parfois une connaissance académique, ce qu'un raciniseur à base de règle n'a pas.

### 5.2.3 Les chiasmes jouant sur les liens sémantiques ou antimétalepses

Théoriquement n'importe quel lien sémantique suffit à relier les quatre mots d'un chiasme. En pratique nous avons observé des liens de co-hyponymie :

> « un bâillon pour la bouche et pour la main le clou », Odon Vallet, *L'évangile des païens*

des liens de synonymie ou synonymie partielle

> « Les désespoirs sont morts, et mortes les douleurs », Albert Samain, *Printemps*

et des relations d'antonymie :

> « Ajoutez quelquefois, et souvent effacez. », Nicolas Boileau, *Art poétique*

Nous n'avons pas trouvé de wordnet français couvrant assez de vocabulaire pour travailler sur nos exemples de chiasmes liant des co-hyponymes ou des antonymes. Nous avons toutefois testé une ressource rarement utilisée en TAL (Rao et Ravichandran, 2009, p.677) : le dictionnaire des synonymes *OpenOffice* (OpenOffice-community, 2011).Ce thésaurus, à l'origine utilisé comme dictionnaire de synonymes dans le célèbre traitement de texte présente l'avantage d'être simple (on le trouve sous la forme d'un fichier texte où est inscrit chaque mot associé à une liste de synonymes) traduit dans de nombreuses langues, gratuit et à libre disposition des chercheurs. Comment ce thésaurus permet-il d'identifier des chiasmes sémantiques ? Voici sur un exemple le procédé algorithmique utilisé pour trouver la synonymie. Considérons l'extrait suivant :

> « Dur avec les faibles, et faible avec les forts. »

1. Tout le texte est préalablement lemmatisé. L'extrait devient ainsi : « dur avec le faible , et faible avec le fort . »
2. « dur » est comparé à tous les mots suivants mais il est procédé à chaque fois à deux comparaisons. La première porte sur la morphologie. En effet, lorsque « dur » est comparé au mot « fort » il y a d'abord vérification des deux chaînes de caractères. « dur » n'est pas la même chaîne de caractère que « fort », il faut donc procéder à la seconde comparaison.
3. La seconde comparaison est la comparaison sémantique. Le thésaurus *OpenOffice* est alors mis à contribution. L'ordinateur vérifie l'entrée « dur » du thésaurus, voici un court extrait de ce

7. `http://snowball.tartarus.org/algorithms/french/stemmer.html`, consulté le 08/05/2013

qui y est écrit (cf. Figure 4).

dur|1
(Adverbe Adjectif Nom) | acerbe | acide | consistant | ferme |[...]
| filandreux | nerveux | **fort** | fortin | forteresse | vigoureux |
robuste [...]

FIGURE 4 – Extrait du thésaurus d' *Open Office* à l'entrée du mot "dur"

Le mot « fort » est contenu dans la liste de synonymes du thésaurus donc « dur » et « fort » sont considérés comme les candidats potentiels d'un chiasme sémantique. Si « fort » n'avait pas été contenu dans l'entrée de « dur » la machine aurait aussi vérifié que « dur » n'est pas compris dans l'entrée de « fort » avant d'éliminer les deux candidats.

Nous ne fournirons pas de calcul de précision sur la détection des antimétalepses car le nombre de chiasmes de ce type à trouver étant très réduit (sept) les résultats n'auraient pas de très grande signification. Cependant nous pouvons établir les problèmes rencontrés. Nous avons pu observer une importante baisse de la précision due à des mots outils de sens proches (tant, comme, alors, tel...) ces mots outils étant considérés comme des synonymes par le thésaurus et étant assez fréquents, il nous a fallu les inclure dans les « StopWords » afin d'obtenir des résultats plus clairs. Au final deux antimétalepses ont été repérées avec succès :

« Les désespoirs sont morts et mortes les douleurs »
« Dur avec les faibles, et faible avec les forts »

Nous observons que ces deux chiasmes sont les seuls repérés sur les sept antimétalepses présentes. Cela dit, ce sont aussi les deux seuls reposant sur un lien de synonymie partielle (« Dur/fort », « désespoir/douleur ») nous n'avons donc pas été victime de lacunes en terme de richesse de vocabulaire du thésaurus. Cette ressource n'étant qu'un dictionnaire de synonymes elle ne pouvait pas, par définition, établir un lien entre des antonymes ou des co-hyponymes. En établissant tous les liens synonymiques nécessaires à notre détection, cette ressource a tenu ses promesses.

# 6 Bilan

Nous avons proposé une méthode permettant d'identifier les chiasmes. Pour ce faire nous avons émis l'hypothèse que par l'ajout de contraintes simples comme le filtrage des mots les plus courants ou la limitation des signes de ponctuation nous pouvions obtenir une meilleure précision. Cette hyothèse est vérifiée. Nous avons ensuite supposé que les outils mis à disposition du TAL français nous aideraient à couvrir une plus grande variété de chiasmes. Cette hypothèse se vérifie largement pour ce qui est du chiasme flexionnel. En revanche, en ce qui concerne le chiasme dérivationnel nous ne sommes pas parvenue avec l'outil de racinisation testé à obtenir de résultats satisfaisants. Enfin l'antimétalepse peut être identifiée automatiquement en français grâce à un thésaurus mais uniquement quand il s'agit de liens de synonymie. Concrétement nous pouvons synthétiser l'apport de notre recherche grâce au tableau ci-dessous (table4).

| | **Recherche doubles inclusions (Gawryjolek, 2009)** | **Recherche triples inclusions (Hromada, 2011)** | **Notre recherche double inclusion avec filtrage + outils TAL** |
|---|---|---|---|
| **Précision** [8] **(extraits justes / extraits relevés)** | <2 % (22 / 1235) | 100 % (11/11) | 58 % (21/36) |
| **Rappel (chiasmes relevés / chiasmes à trouver)** | 100 % (22/22) | 50 % (11/22) | 95 % (21/22) |
| **F-mesure** [9] | 4 % (0.035) | 66 % | 72 %(0.724) |
| **Détecte-t-il les antimétaboles ?** | Oui | Oui (en partie) | Oui |
| **Les chiasmes flexionnels ?** | Non | Non | Oui |
| **Les chiasmes dérivationnels ?** | Non | Non | Non |
| **Les antimétalepses ?** | Non | Non | Oui (sur liens de synonymie) |
| **Multilingue** | Non testé | Oui | Non |
| **Avantages** | – Relevé exhaustif sur les antimétaboles strictes | – Tri manuel des faux positifs quasi-non nécessaire.<br>– Idéal pour fouiller de très grands corpus (>100 000 de mots) | – Relevé des chiasmes jouant sur les mots fléchis<br>– Rappel important<br>– Assez précis pour des analyses semi-manuelles sur corpus de taille moyenne (<100 000 de mots) |
| **Inconvénients** | – Peu précis vérification manuelle nécessaire<br>– Sur les corpus >1000 mots tri manuel très long | – Omission de trop d'antimétaboles pour l'analyse de discours | – Tri manuel encore long si corpus très grand (>100 000 de mots).<br>– Exhaustivité non garantie. |

TABLE 4 – Synthèse des résultats et comparatif des recherches

8. Tous les chiffres donnés dans ce tableau portent sur la détection d'antimétaboles strictes uniquement : les résultats sur antimétaboles flexionnelles et antimétalepses ne pouvant pas être comparés avec d'autres.

9. On émettra cependant une réserve sur les chiffres au regard de notre corpus (cf. commentaire 7.1).

# 7 Discussion

## 7.1 Commentaire sur les résultats

Les résultats du tableau 4 sont à considérer de manière relative. Il est important en effet de rappeler que notre « corpus », utilisé à la fois pour la mise au point et pour l'évaluation de la méthode, même s'il est constitué d'exemples et de contextes réels, ne représente pas la rareté du phénomène du chiasme. Il reste un texte artificiel fondé sur des extraits réels ni plus ni moins. Ainsi, lancé sur un roman d'un 100 000 de mots la précision de notre algorithme comme celle de Gawryjolek (2009) et celle de Hromada (2011) chutera parce qu'il y aura moins de chiasmes à trouver et plus de contexte pour générer des faux positifs.

## 7.2 Perspectives

À notre connaissance cette recherche est la seule, toutes langues confondues, à proposer une ressource avec des chiasmes classés et dont on connaît précisément l'origine ainsi que le contexte. Cette ressource nécessite bien entendu d'être enrichie surtout en ce qui concerne les antimétalepses et les chiasmes dérivationnels trop rarement illustrés dans les ouvrages de référence. Compte tenu des difficultés de collecte évoquées partie 3.1 et des problèmes de définitions auxquels le chercheur en TAL doit préalablement faire face, une liste d'exemples, même modeste, constitue un vrai tremplin vers l'élaboration d'outils automatiques ou semi-automatiques.

À l'amélioration du corpus, il serait intéressant d'ajouter l'amélioration des algorithmes. Cette recherche nous a déjà permis de faire le pas vers la détection des chiasmes flexionnels : compte tenu des résultats encourageants obtenus avec Flemm, il n'y a plus désormais de raison sur un corpus en français de se contenter de la détection d'antimétaboles strictes comme cela était fait auparavant. Notre recherche introduit aussi la détection des chiasmes ne reposant pas sur le rapprochement morphologique. Enfin les erreurs que nous avons relevées ouvrent la voie vers d'autres réflexions : peut-être que sélectionner manuellement les « stopwords » comme nous l'avons fait n'est pas la manière la plus efficace de procéder. En exploitant d'autres informations comme l'étiquetage grammatical nous espérons lors d'une prochaine étude augmenter la précision voire le rappel. Une analyse linguistique plus poussée non seulement des termes principaux (adjectifs, verbes, noms) mais aussi concernant la disposition des termes secondaires (articles, adverbes, autres mots outils) permettrait peut-être une détection plus pertinente (cf. 4.2).

# Remerciements

Merci à mon ancien directeur de recherche le Pr. Marcel Cori de l'université Paris Ouest pour ses conseils avisés lors de la direction de ce travail.

# Références

AGICHTEIN, E. et GRAVANO, L. (2000). Snowball : Extracting Relations from Large Plain-Text

 

Collections. *In Proceedings of the 5th ACM International Conference on Digital Libraries*, pages 85–94, Texas, USA.

BENDERSKY, M. et SMITH, D. (2012). A Dictionary of Wisdom and Wit : Learning to Extract Quotable Phrases. *In Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 69–77, Montréal, Canada.

BRADFER, J., FRANCARD, M. et FAIRON, C. (1995). Base de données Julibel. Centres VALIBEL et CENTAL, `http://julibel.fltr.ucl.ac.be/index.php`, consulté le 08/05/2013.

DIDEROT, D. et D'ALEMBERT, J. l. R. (1789). *Encyclopédie méthodique : ou par ordre de matières, volume 66*. Livre numérique Google `http://books.google.fr/books?id=NchCAAAAYAAJ&lpg=PA198&ots=UEkdoAOE-g&dq=antimetabole%20diderot&pg=PA198#v=onepage&q=antimetabole%20diderot&f=false`, consulté le 08/05/2013.

DUPRIEZ, B. (2003). *Gradus, les procédés littéraires*. Union Générale d'Éditions 10/18.

GARCÍA-PAGE, M. (1991). El "retruécano léxico" y sus límites. *Archivum : Revista de la Facultad de Filología de Oviedo*, 41-42:173–203.

GAWRYJOLEK, J. J. (2009). *Automated Annotation and Visualization of Rhetorical Figures*. Master thesis, Universty of Waterloo.

GREENE, R. (2012). *The Princeton Encyclopedia of Poetry and Poetics : Fourth Edition*. Princeton University Press.

HARRIS, R. et DIMARCO, C. (2009). Constructing a Rhetorical Figuration Ontology. *In Persuasive Technology and Digital Behaviour Intervention Symposium*, pages 47–52, Edinburgh, Scotland.

HROMADA, D. D. (2011). Initial Experiments with Multilingual Extraction of Rhetoric Figures by means of PERL-compatible Regular Expressions. *In Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 85–90, Hissar, Bulgaria.

NAMER, F. (2000). Un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues*, 41:1–23.

NORDAHL, H. (1971). Variantes chiasmiques. Essai de description formelle. *Revue Romane*, 6:219–232.

OPENOFFICE-COMMUNITY (2011). Open Office dictionaries. Thésaurus en français disponible sur : `http://extensions.openoffice.org/en/project/french-dictionary-modern` `http://www.dicollecte.org/download/fr/thesaurus-v2.3.zip`, consultés le 08/05/2013.

POUGEOISE, M. (2001). *Dictionnaire de Rhétorique*. Armand Colin.

RABATEL, A. (2008). Points de vue en confrontation dans les antimétaboles PLUS et MOINS. *Langue française*, 160(4):21–36.

RAO, D. et RAVICHANDRAN, D. (2009). Semi-Supervised Polarity Lexicon Induction. *In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 675–682, Athens, Greece.

SCHMID, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *In Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, Manchester, Great Britain.

VAN GORP, H., DELABASTITA, D. et D'HULST, L. (2001). *Dictionnaire des termes littéraires*. Honoré Champion.

VANDENDORPE, C. (1991). Lecture et quête de sens. *Protée*, 19:95–101.

 