Extraction et analyse automatique des comparaisons et des pseudo-comparaisons pour la détection des comparaisons figuratives

Suzanne Mpouli[1,2] Jean-Gabriel Ganascia[1,2]
(1) Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu 75005 Paris
(2) Labex OBVIL, Université Paris-Sorbonne, 1 rue Victor Cousin 75005 Paris
mpouli@acasa.lip6.fr, jean-gabriel.ganascia@lip6.fr

Résumé. Le présent article s'intéresse à la détection et à la désambiguïsation des comparaisons figuratives. Il décrit un algorithme qui utilise un analyseur syntaxique de surface (*chunker*) et des règles manuelles afin d'extraire et d'analyser les (pseudo-)comparaisons présentes dans un texte. Cet algorithme, évalué sur un corpus de textes littéraires, donne de meilleurs résultats qu'un système reposant sur une analyse syntaxique profonde.

Abstract.

Extraction and automatic analysis of comparative and pseudo-comparative structures for simile detection.
This article is focused on automatic simile detection and disambiguation. It describes an algorithm which uses syntactic chunks and handcrafted rules to extract and analyse similes in a given text. This algorithm, which was evaluated on a corpus of literary texts, performs better than a system based on dependency parsing.

Mots-clés : comparaisons figuratives, comparé, comparant, analyse syntaxique de surface, règles manuelles, analyse syntaxique profonde.

Keywords: simile, tenor, vehicle, chunking, handcrafted rules, dependency parsing.

1 Introduction

Avec le nombre sans cesse croissant de textes numérisés disponibles, se pose la question de leur interrogation automatique afin d'en extraire les éléments relevant du style d'un auteur, d'une œuvre ou d'un genre. C'est dans cette optique que s'inscrit le travail que nous présentons dans cette contribution et qui vise à long terme la reconnaissance automatique des comparaisons figuratives dans les textes littéraires. Bien qu'elles aient été très peu étudiées en TAL, les comparaisons figuratives se rattachent à deux phénomènes qui se rencontrent dans la plupart des langues naturelles : la comparaison et le langage figuré. En effet, du point de vue structurel, les comparaisons figuratives sont des constructions comparatives, c'est-à-dire des formes linguistiques utilisées pour exprimer à quel degré et sur quelle base deux entités au minimum peuvent être considérées semblables ou dissemblables. Cependant, du point de vue sémantique, elles se distinguent des autres types de constructions comparatives (appelées par contraste comparaisons littérales) car elles établissent un parallèle entre des termes n'appartenant pas à la même catégorie sémantique en attribuant à l'un de ces termes des attributs propres à la catégorie sémantique de l'autre (Glucksberg & Keysar, 1990). À titre d'illustration, considérons les deux énoncés suivants :

(1) *Céline est aussi revêche que sa sœur.* (2) *Céline est aussi revêche qu'un cactus*.

Malgré la symétrie parfaite entre ces deux phrases, seule la la phrase (2) peut être classée comme une comparaison figurative car elle crée une image en plaçant au même plan un être humain (Céline) et une plante (cactus) à laquelle est conféré un trait de caractère propre aux humains.

Il apparaît donc que la reconnaissance des comparaisons figuratives dans un texte brut comprend trois tâches principales : l'extraction des structures comparatives et pseudo-comparatives contenues dans le texte, l'identification des constituants de ces structures et la désambiguïsation de ces structures. Afin de mieux décrire le problème de la reconnaissance des comparaisons figuratives, nous présentons, dans la section 2, la structure des comparaisons figuratives ainsi que l'état de l'art en matière d'extraction et d'analyse automatiques des comparaisons et des pseudo-comparaisons. Dans la troisième section, nous décrivons un algorithme qui se focalise sur l'extraction des comparaisons et l'identification de leurs constituants. Puis, nous évaluons cet algorithme qui repose sur des règles manuelles et l'analyse syntaxique de surface (*chunking*) sur un corpus de textes littéraires et le comparons à un analyseur syntaxique profond. Dans la dernière section, nous évoquons brièvement la question de la désambiguïsation des comparaisons figuratives. Pour finir, nous concluons notre travail en présentant quelques perspectives futures.

2 État de l'art

2.1 Structures sémantique et syntaxique des comparaisons figuratives

S'inspirant des travaux de Dumarsais, Soublin (1971) définit un processus en deux étapes pour expliquer la formation des comparaisons figuratives à partir de deux phrases ayant la même structure Syntagme nominal + Verbe + Adjectif qualificatif : insertion d'un outil de comparaison entre les deux phrases, puis suppression du verbe et de l'adjectif après le syntagme nominal suivant cet outil de comparaison. On passera donc ainsi de (3a) « *La fille est calme* » et (3b) « Un *lac est calme* » à (4) « *La fille est calme comme un lac est calme* » et finalement à (5) *« La fille est calme comme un lac* ». Typiquement, l'exemple (5) correspond à la forme canonique de la comparaison figurative (Soublin, 1971). En rhétorique, cette forme canonique se compose de cinq éléments :

- le comparé (« fille ») ou terme source qui est décrit totalement ou partiellement par la comparaison;
- le verbe (« est ») qui introduit le motif ou désigne soit une aptitude, soit un comportement sur lequel porte la comparaison ;
- le tertium comparationis ou motif (« calme ») qui représente l'attribut que les entités comparées ont en commun ;
- l'outil de comparaison ou marqueur (« comme ») qui établit le rapport de similitude ou de différence ;
- le comparant (« lac ») ou terme cible qui sert de point de référence pour la comparaison (Hanks, 2012).

Dans la pratique, tous ces éléments, hormis l'outil de comparaison et le comparant, peuvent être omis. Du point de l'ordre des constituants, le comparant, en qualité de complément, suit nécessairement l'outil de comparaison même si la position du reste des constituants peut varier tout en respectant l'ordre syntaxique prévalant dans la langue, dans le cas du français sujet – verbe – objet. Sur le plan grammatical, ce type de constructions se classe parmi les subordonnées comparatives averbales qui sont des versions elliptiques d'une proposition principale et dans lesquelles le comparant occupe la même fonction que le comparé dans la principale (Fuchs et al, 2008). Il existe ainsi une corrélation entre la fonction grammaticale des éléments de la comparaison et leur rôle sémantique : dans la phrase (5), par exemple, « lac » qui est le comparant, remplace dans la subordonnée le comparé « fille », et les deux substantifs sont des sujets.

La comparaison étant avant tout une question de sens, en plus de « comme » qui est incontestablement le marqueur prototypique de la comparaison figurative, la langue française dispose de plusieurs termes susceptibles d'inférer un rapport de similitude ou de dissemblance. Bouverot (1969) oppose ainsi les comparaisons de type I introduites par les comparatifs ou des outils de la forme « déclencheur + que » (ainsi que, de même que…) aux comparaisons de type II reposant sur des adjectifs, des verbes, des suffixes ou des locutions prépositionnelles. Le Tableau I présente l'ensemble des structures possibles pour les comparaisons de type I et de type II en posant la phrase comme contexte de réalisation.

Comparatifs et locutions prépositionnelles	**Verbes et locutions verbales**	**Adjectifs qualificatifs**
A/ Marqueur + comparant *Il aime briller. Comme les étoiles.*	A/ Marqueur + comparant *Moi, ressembler à une étoile ?*	A/ Marqueur + comparant *Il aime briller. Telle une étoile*
B/ Comparé + marqueur + comparant *Vous êtes revigoré. Souriant. Les yeux comme des étoiles.*	B/ Comparé + marqueur + comparant *Ses yeux ressemblent à deux étoiles scintillantes.*	B/ Comparé + marqueur + comparant *Vous êtes revigoré. Souriant. Les yeux pareils à des étoiles.*
C/ Verbe + marqueur + comparant *Ne jamais filer comme une étoile après le crime.*		C/ Verbe + marqueur + comparant *Ne jamais filer telle une étoile après le crime.*
D/ Comparé + verbe + marqueur + comparant *Sa lame luit comme une étoile.*		D/ Comparé + verbe + marqueur + comparant *Ses yeux sont semblables à des étoiles*.
E/ Adjectif motif + marqueur + comparant *Tout un peuple. Innombrable comme les étoiles !*		
F/ Comparé + adjectif motif + marqueur+ comparant *Une ville exotique, aussi lointaine que les étoiles dans le ciel.*		
G/ Comparé + verbe + adjectif motif + marqueur+ comparant *Son regard brille ainsi qu'une étoile.*		

TABLEAU 1 : Structures des comparaisons figuratives en fonction du marqueur

2.2 Comparaisons figuratives et TAL

En ce qui concerne la recherche sur les comparaisons en traitement automatique des langues, il est possible de délimiter deux phases principales : une première phase linguistique en majorité descriptive et une phase informatique axée sur le développement d'outils pour détecter et analyser des types de comparaisons spécifiques. Bien que relativement récent, le domaine de l'analyse automatique des comparaisons figuratives diffère de celui de la détection des phrases comparatives autant par ses objectifs que par ses méthodes. De manière générale, les approches existantes tentent de tirer parti de la corrélation entre la fonction grammaticale et le rôle sémantique des constituants des comparaisons figuratives. Deux outils ont été testés pour ce faire : GLARF (Meyers *et al*, 2001 ; Niculae et Yaneva, 2013) qui enrichit la sortie des analyseurs syntaxiques en constituants, notamment en identifiant les sujets, les objets et les noyaux des syntagmes, et un analyseur syntaxique profond, TurboParser (Martins *et al*., 2010 ; Niculae, 2013). Dans es deux cas, l'identification repose sur les étapes suivantes :

- Établir une liste de marqueurs ;
- Parcourir les nœuds de la structure en arbre de la phrase jusqu'à trouver un substantif 1 étiqueté comme étant un complément d'un des marqueurs ;
- Identifier dans l'arbre un lien rattachant le marqueur à un verbe ;
- Repérer dans l'arbre un lien connectant le verbe identifié en 3 à un substantif 2 étiqueté comme étant son sujet ;

- Trouver dans l'arbre un lien qui subordonne un adjectif qualificatif au verbe identifié en 3 ;
- Si les étapes 2 à 4 ont des résultats positifs, la phrase est extraite, le substantif 1 est considéré comme étant le comparant, le marqueur comme l'outil de la comparaison, le substantif 2 comme le comparé et le verbe est extrait.
- Si l'étape 5 a un résultat positif, l'adjectif qualificatif est considéré comme étant le motif.

De plus, la comparaison des résultats obtenus avec les deux outils montre que l'analyse syntaxique profonde donne de meilleurs résultats : par exemple, sur un set de 53 phrases, on constate un rappel plutôt haut (71 % contre 43 % avec GLARF) pour une précision assez basse, 24 % (Niculae, 2013). Différentes raisons pourraient être avancées pour expliquer cette performance :

- avec les comparatifs, seules deux structures de comparaisons figuratives sont reconnues sur les sept possibles ;
- la polysémie des marqueurs comme « like » qui peut aussi être une forme verbale ;
- l'exploration ne prévoit pas des structures où le comparé est un complément d'objet direct comme dans la phrase « *l'homme ligota ses frères ainsi que des saucissons* » ;
- l'exploration de l'arbre ne considère pas les verbes juxtaposés ou coordonnés à d'autres verbes et les propositions ayant un pronom relatif pour sujet ;
- la fouille de l'arbre ne tient pas compte des comparaisons figuratives ayant plus d'un comparant ou comparé.
- l'extraction des structures comparatives concerne aussi bien les subordonnées comparatives averbales que des subordonnées comparatives contenant des verbes ou d'autres subordonnées introduites par le marqueur.

Ce dernier point a toute son importance puisqu'il existe une différence sémantique non-négligeable entre les subordonnées comparatives verbales mettant en parallèle deux entités et les subordonnées comparatives verbales qui contrastent deux actions ou processus. Pour finir, les méthodes proposées ne recherchent qu'une seule comparaison par phrase et ignorent donc le reste des comparaisons dans le cas de phrases contenant plusieurs comparaisons figuratives.

Au niveau de la désambiguïsation des comparaisons extraites, la méthode proposée s'appuie sur l'apprentissage automatique et la sémantique distributionnelle pour mesurer la similarité sémantique entre le comparé et le comparant qui est combinée avec d'autres attributs tels que le domaine du comparant et la présence d'un article indéfini avant celui-ci (Niculae & Danescu-Niculescu-Mizil, 2014).

3 Extraction et analyse des comparaisons et des peudo-comparaisons

3.1 Description de l'algorithme

Au regard de notre objectif qui est d'identifier toutes les comparaisons figuratives que renferme un texte littéraire, différents choix méthodologiques ont été faits : identifier les comparaisons de type I mais aussi celles de type II, détecter plus d'une comparaison par phrase le cas échéant, ne pas se limiter aux substantifs comparés, tenir compte de l'ambiguïté des comparaisons figuratives et extraire tous les comparés syntaxiquement possibles. Nous distinguons ainsi quatre groupes de marqueurs de la comparaison :

- les marqueurs traditionnels : *comme, ainsi que, de même que, autant que, plus…que, tel que, moins…que, aussi… que ;*
- les verbes : *ressembler à, sembler, faire l'effet de, faire penser à, faire songer à, donner l'impression de ;*
- les adjectifs qualificatifs : *semblable à, pareil à, tel, similaire à, analogue à, comparable à ;*
- et les locutions prépositionnelles : *à la manière de, à l'image, à l'égal de, à l'instar de, à la façon de.*

L'extraction de comparaisons et de pseudo-comparaisons s'intéresse uniquement aux structures de la forme marqueur + SN ou marqueur,..., SN dans lesquelles le comparant n'est pas un sujet. Afin de mieux circonscrire les phrases concernées, nous avons défini la règle suivante :

Règle 1. Soit un syntagme nominal SN placé immédiatement après un marqueur, le substantif X, noyau de SN, est considéré comme étant un sujet si un verbe conjugué est placé après X et n'est pas séparé de celui-ci par une virgule, un point-virgule, un pronom personnel sujet, un pronom relatif ou une conjonction de subordination.

Une fois le comparant identifié et la phrase extraite, la recherche des constituants de la comparaison se fait vers la gauche uniquement si le marqueur ne se trouve pas en début de phrase, après un signe de ponctuation ou une conjonction de coordination, vers la droite uniquement si le marqueur se trouve en début de phrase, et dans les deux sens s'il est placé directement après un signe de ponctuation ou une conjonction de coordination.

Notre hypothèse de travail principale repose sur le fonctionnement de la syntaxe du français et suppose que si l'on arrive à déterminer la catégorie grammaticale du mot que la structure marqueur + comparant complète syntaxiquement, on peut ainsi inférer la fonction grammaticale du comparé dans la proposition principale et extraire les autres composants de la comparaison. Si ce mot est un verbe, le comparant sera soit le sujet, soit le COD de ce verbe, si c'est un adjectif, le comparant sera forcément le mot que modifie cet adjectif qui peut également être le sujet ou le COD du verbe en fonction de la fonction de l'adjectif et enfin, si ce mot est un substantif, ce substantif est le comparant. De point de vue de la nature des sujets, nous nous sommes limités aux substantifs, aux adjectifs démonstratifs et aux pronoms personnels. Une liste d'indices textuels a été compilée pour vérifier la fonction des constituants recherchés. Par exemple, un adjectif motif ne peut être séparé du marqueur par une conjonction de coordination, un pronom relatif, une préposition ou un syntagme nominal.

3.2 Résultats expérimentaux

L'algorithme présenté dans la section précédente a été testé sur un corpus composé de poèmes en prose écrits par quatre poètes français : Aloysius Bertrand, Stéphane Mallarmé, Charles Baudelaire et Arthur Rimbaud. Ce corpus a été annoté manuellement. Nous nous sommes servis de TreeTagger (Schmid, 1994) pour la tokénisation, l'étiquetage morphosyntaxique et l'analyse syntaxique de surface. Nous avons également écrit des règles qui exploitent la sortie de TreeTagger pour la segmentation en phrases.

Nous avons comparé la performance de notre algorithme à celle d'une version améliorée du système proposé par Niculae (2013) se basant sur des dépendances syntaxiques fournies par le Berkeley Parser (Candito *et al.*, 2010). Les résultats obtenus sont présentés dans le Tableau 2. Pour chaque méthode et chaque classe de constituants, le rappel (vrais positifs/vrais positifs + faux négatifs) et la précision (vrais positifs/vrais positifs + faux positifs) ont été calculés.

	Rp (%)	**Pr (%)**	**VP**	**FP**	**FN**
Comparé	61,9	46,9	163	184	100
Verbe	55,5	52,8	75	67	60
Adjectif motif	58	69,1	83	37	60
Comparant	90,8	96,7	238	8	24

	Rp (%)	**Pr (%)**	**VP**	**FP**	**FN**
Comparé	54,3	50,1	143	142	120
Verbe	64,4	47,8	87	95	48
Adjectif motif	44	69,2	63	28	80
Comparant	87	90	228	23	34

TABLEAU 2 : Évaluation du Berkeley Parser (à droite) et de l'algorithme (à gauche). La précision (Pr), le rappel (Rp), les vrais positifs (VP), les faux positifs (FP) et les faux négatifs (FN) sont indiqués.

Contrairement à notre algorithme, le Berkeley Parser peut directement décider si un comparant détecté est utilisé comme sujet ou non. Il commet cependant au niveau de la reconnaissance de comparants plus d'erreurs d'étiquetage morphosyntaxique (36 % des erreurs) que TreeTagger (14%). Les autres erreurs du Berkeley Parser pour cette tâche sont dues à une mauvaise segmentation de phrase, à un comparant faussement identifié comme étant sujet ou à une dépendance erronée. D'autre part, en ce qui concerne l'identification des verbes, le participe passé pose un problème car dans l'annotation manuelle, en fonction de son emploi, il est tantôt considéré comme adjectif, tantôt comme un verbe.

Soulignons également différentes structures qui sont problématiques pour les deux méthodes :

- le participe utilisé comme nom : « … *coupable à l'égal d'un faux scandalisé* » ;

– les structures « plus de X que de Y » : « *il y a plus de sbires que de citadins* » ;
– l'accumulation de comparaisons dans une même phrase : « *ses cheveux longs comme des saules et peignés comme des broussailles*. »
– l'inversion du sujet : « *cette solide cage de fer derrière laquelle s'agite, hurlant comme un damné, secouant les barreaux comme un orang-outang exaspéré par l'exil, imitant, dans la perfection, tantôt les bonds circulaires du tigre, tantôt les dandinements stupides de l'ours blanc, ce monstre poilu dont la forme imite assez vaguement la vôtre*.»
– les comparés absents : « *Ce soir à Circeto des hautes glaces, grasse comme le poisson, et enluminée comme les dix mois de la nuit rouge, - (son cœur ambre et spunk), - pour ma seule prière muette comme ces régions ...* »
– la présence d'un adjectif non motif avant le marqueur : « *Il est aussi difficile de supposer une mère sans amour maternel qu'une lumière sans chaleur*. »
– un sujet éloigné de son verbe : « *de tous les coins, des fissures des tiroirs et des plis des étoffes s'échappe un parfum singulier, un revenez-y de Sumatra, qui est comme l'âme de l'appartement.* »
– l'accumulation d'adjectifs : « *Les meubles sont vastes, curieux, bizarres, armés de serrures et de secrets comme des âmes raffinées.* »

4 Désambiguïsation des comparaisons figuratives

La phase de la reconnaissance des constituants soulève une question importante pour la désambiguïsation des comparaisons figuratives. En effet, le comparant ou comparé réel dans une phrase n'est pas toujours le noyau du groupe nominal mais souvent son complément comme dans : « *toutes les richesses flambant comme un milliard de tonnerres*. » Cet aspect devrait donc être pris en compte au moment de l'identification des constituants. De plus, sans résolution des pronoms anaphoriques, la désambiguïsation ne peut concerner que les substantifs.

Des structures grammaticales permettent cependant de distinguer les comparaisons littérales des comparaisons figuratives à l'instar de « il y a comme », « c'est comme » et les verbes « élire, nommer, citer, attribuer, considérer, juger, témoigner » employés avec « comme » (Fuchs *et al*., 2008).

En accord avec les pratiques littéraires de description des comparaisons figuratives, nous avons pris le parti de définir une comparaison figurative par un écart entre les traits sémantiques (concret, abstrait, inanimé, animé) et/ou le domaine du comparant et du comparé. Pour tester cette définition, nous avons utilisé le Dictionnaire électronique des mots de Jean Dubois et François Dubois-Charlier[1] qui renseigne entre autres sur l'animéité et le domaine auquel appartient le substantif. Au vu de la polysémie des substantifs, le problème de leur désambiguïsation se pose. Nous avons donc choisi pour chaque substantif de choisir la première catégorie proposée par le dictionnaire qui est en général la catégorie la plus fréquente. Pour déterminer s'il y a un distance sémantique entre deux substantifs, par défaut, nous attribuons une valeur 1 s'ils ont des natures différentes et une valeur 2 s'ils appartiennent à différents domaines. Cette mesure semble marcher assez bien pour certains termes assez proches comme « haine » et « amour » , « charité » et « orgueil » mais pas pour d'autres comme « rue » et « faubourg » ou encore « jour » et « nuit ». On remarque qu'un changement d'animéité est généralement beaucoup plus informatif qu'un changement de domaine mais ne concerne que très peu de couples. Cela laisse supposer qu'une méthode qui s'appuierait sur plus de traits sémantiques ou sur la catégorie naturelle des substantifs donnerait de meilleurs résultats.

5 Conclusion

Notre travail s'inscrivait dans le cadre de la stylistique automatique et avait pour but de présenter ainsi que d'évaluer une méthode pour extraire et analyser automatiquement les (pseudo-)comparaisons. Il nous paraît important dans un premier temps de réduire le bruit, surtout en ce qui concerne la détection des comparés. L'utilisation de ressources linguistiques nous paraît pour ce faire une piste intéressante. Nous songeons aussi à adapter et à tester ce système sur des langues qui possèdent des formes de comparaisons figuratives assez proches de celles du français, comme l'anglais.

[1] Disponible à l'adresse suivante : http://rali.iro.umontreal.ca/rali/?q=fr/dictionnaire-electronique-des-mots-dem

Remerciements

Ce travail a bénéficié d'une aide d'Etat gérée par l'Agence Nationale de la Recherche dans le cadre des Investissements d'Avenir portant la référence ANR-11-IDEX-004-02.

Références

BOUVEROT B. (1969). Comparaison et métaphore. *Le Français moderne* 2, 132-147, 224-238 et 301-316.

CANDITO M., NIVRE J., DENIS P., ANGUIANO E. H. (2010). Benchmarking of statistical dependency parsers for French. *Proceedings of COLING 2010*, 108-116.

FUCHS C., FOURNIER, N., LE GOFFIC P. (2008). Structures à subordonnée comparative en français : Problèmes de représentations syntaxiques et sémantiques. *Lingvisticae Investigationes* 31:1, 11-61.

GLUCKSBERG S, KEYSAR B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review* 97:1, 3-18.

HANKS P. (2012). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review* 97:1, 3-18.

MARTINS A., SMITH N., XING P., AGUIAR P., FIGUEIREDO M. (2001). Parsing and GLARFing. *Proceedings of RANLP*, 110-114.

MEYERS A., KOSAKA M., SEKINE S., GRISHMAN R., ZHAO S. (2010). Turbo parsers: Dependency parsing by approximate variational inference. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 34-44.

NICULAE V. (2013). Comparison pattern matching and creative simile recognition. *Joint Symposium on Semantic Processing, Textual inference and Structure in Corpora*, 110-114.

NICULAE V., DANESCU-NICULESCU-MIZEL C. (2014). Brighter than gold. *JProceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2008-2018.

NICULAE V., YANEVA V. (2013). Computational considerations of comparisons and similes. *Proceedings of the ACL Research Student Workshop*, 89-95.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, 44-49.

SOUBLIN F. (1971). Sur une règle rhétorique d'effacement. *Langue française* 11, 102-109.