

Traitement automatique des métaphores : une approche par marquage textuel

Stéphane Ferrari

LIMSI-CNRS
Groupe Langage & Cognition
GT SEMLEX
BP 133

F-91403 Orsay cédex

ferrari@limsi.fr
téléphone: (1) 69.85.80.18
télécopie: (1) 69.85.80.88

Résumé

Nous exposons ici quelques résultats et perspectives d'une étude effectuée sur un corpus de taille moyenne des marqueurs textuels liés à la métaphore. Confrontée aux travaux précédents dans le domaine, notre approche propose principalement une méthodologie de déclenchement du traitement à apporter à la phrase, fondée sur ce marquage. La validation du modèle nécessite d'une part des évaluations sur gros corpus, pour lesquelles nous développons un outil informatique; d'autre part l'apport de disciplines issues des sciences cognitives, telles la psycholinguistique, pour une étude de l'interprétation avec et sans marquage, et la didactique, dans une perspective de génération de textes explicatifs.

Mots-clés

métaphore, marqueur textuel, traitement automatique de la langue

Introduction

Nous proposons d'étudier les régularités linguistiques qui peuvent accompagner la métaphore dans son énonciation en langue française.

Les précédents travaux sur la métaphore en traitement automatique de la langue, présentés dans la partie 1, ont pour point commun de ne proposer que des méthodes de résolution de la figure, sans en effectuer de repérage *systematique*. D'autres travaux, effectués dans des domaines différents, ont montré l'importance des indices de surface comme *traces textuelles* d'un processus langagier, intentionnel ou cognitif sous-jacent. Nous nous sommes tournés vers ce type d'approche afin d'améliorer la détection automatique des métaphores.

Une analyse de corpus a permis de mettre en évidence un certain nombre de régularités dans l'énonciation des métaphores, présentées dans la partie 2. Cette étude se poursuit par la mise en œuvre d'un outil de détection automatique¹ permettant de travailler sur de gros corpus électroniques, présenté dans la partie 3. Nous développerons les perspectives d'intégration d'un repérage utilisant les marqueurs textuels dans un système complexe de traitement de la langue. Nous insisterons à cet effet sur la nécessité d'une collaboration entre disciplines issues des sciences cognitives pour la validation du modèle.

1 Approches classiques et détection

En traitement automatique de la langue, deux orientations majeures coexistent quant à la notion de *dé-*

tection dans les travaux relatifs aux métaphores.

L'une consiste à lancer un traitement spécifique de la figure après avoir détecté une malformation de la représentation du sens issue d'un traitement classique. Elle se fonde généralement sur une résolution du lien² entre la source et la cible de la métaphore. Ce type d'approche présente l'inconvénient majeur d'ignorer les expressions ambiguës, telles :

ex. 1 *La bourse s'est littéralement effondrée.*

pour lesquelles le contexte permettrait de favoriser une interprétation métaphorique. De plus, des reproches peuvent être émis à l'égard des temps de traitement relatifs d'une part au sens littéral, d'autre part au sens métaphorique, la différence entre ceux-ci allant à l'encontre des résultats qu'obtiennent les psycholinguistes à ce sujet [5].

Une seconde orientation consiste à exploiter des connaissances sur des sens métaphoriques *conventionnels*. Dans ce cadre, la stratégie proposée par Martin [7] se fonde sur un modèle où le traitement des métaphores est indissocié du traitement « littéral », exploitant une représentation de métaphores conceptuelles [6]. Cette méthode répond alors de manière plus satisfaisante aux exigences exposées dans [5]. Cependant, le traitement de métaphores nouvelles ne repose plus alors que sur les connaissances relatives aux métaphores conventionnelles déjà présentes dans la base de connaissances, ce qui laisse supposer que ce modèle peut encore être confronté à des métaphores pour lesquelles il ne dispose ni de moyen de détection, ni de moyen d'interprétation. En outre, l'absence de connaissance relative à la détection des métaphores

1. Ce travail s'inscrit dans un projet de recherche soutenu par l'AUFELF-UREF

2. Voir [3] et [4] pour une classification plus fine des types de traitements en fonction de la nature du lien métaphorique

A1.2 positions «attribut du COD»

A1.2.1 $GN_1 V_1 GN_2 (moins|plus|aussi) Adj (que|qu') GN_3,$

A1.2.2 $GN_1 (le|la|les|l') V_1 (moins|plus|aussi) Adj (que|qu') GN_3,$

où V_1 verbe d'action comme *croire, juger, faire, estimer, rendre, trouver*.

cible : $GN_2|le|la|les|l'$

source : GN_3

tension : $\mathcal{V}^*(Adj) \cup \{p(GN_3)\} \setminus (\{p(GN_2)\} \cap \{p(GN_3)\})$

FIG. 1 - A1 : « Comparaisons syntaxiques », de groupes nominaux

conduit *a posteriori* à une désambiguïsation systématique de représentations multiples qui pourrait, dans certains cas, être évitée.

Selon nous, le déclenchement d'un traitement particulier de la métaphore nécessite des hypothèses nouvelles. En effet, si l'on considère l'exemple 1, les traitements précédents échouent en interprétant dans le sens littéral d'un bâtiment s'effondrant, ignorant la métonymie³. Nous proposons de ce fait un outil de repérage exploitant dans un premier temps un marquage textuel. Ainsi, dans l'exemple 1, la présence de l'adverbe « littéralement » peut laisser penser, intuitivement, que l'auteur de la phrase attire l'attention sur la littéralité du verbe.

D'une manière générale, un locuteur peut fournir des indices de surface concernant non seulement le contenu de son discours mais encore la forme de celui-ci, ainsi que le montrent de précédents travaux [8], [2], [9] relatifs au domaine de l'explication. Ces marqueurs constituent alors des indices évidents pour l'interprétation du discours. Pour la métaphore, le corpus du journal « Le Monde » fournit l'exemple « frappant » illustrant cette notion, et où il n'est question en réalité que de situation économique :

ex. 2 *il apparaît peu probable que l'on retrouve, du moins à court terme, la combinaison des "détonateurs" - pour reprendre la métaphore d'un expert français - à l'origine de l'explosion du 19 octobre.[...] En fait, et pour poursuivre la métaphore précédente, toute la question reste de savoir si l'explosif à l'origine de l'effondrement du 19 octobre est toujours présent.* Les travaux actuels sur la métaphore n'exploitant aucunement ce genre de marqueurs, nous avons tout d'abord étudié l'existence de telles régularités. Les résultats de cette étude sont décrits dans la partie suivante.

2 Marquage textuel

Pour l'étude des régularités liées à la métaphores, une première phase s'est effectuée sur des exemples pris dans la littérature et divers journaux. Dans le cadre d'une collaboration avec des collègues psychologues

du groupe « Cognition Humaine » du LIMSI, un corpus explicatif de 26 textes de 200 mots en moyenne a été recueilli, afin d'étudier plus particulièrement *les analogies et métaphores explicatives dans l'enseignement* [10]. Les régularités rencontrées dans le corpus sont de deux types : la mise en relation de concepts, par le biais d'une comparaison explicite, d'une identification ou d'une opposition, et l'utilisation de marqueurs spécifiques d'un type d'emphase.

Lorsqu'une mise en relation explicite de concepts a lieu, la comparaison n'est pas un attribut intrinsèque des métaphores résultantes, celles-ci pouvant aussi opérer par substitution. On constate donc que la forme de surface reste indépendante du mode d'opération. Lorsqu'une comparaison explicite a lieu, nous parlons encore de métaphore au sens des métaphores conceptuelles introduites par Lakoff [6]. La raison en est que le traitement informatique est indifférencié pour les comparaisons et pour les véritables métaphores linguistiques, introduisant chacune une métaphore conceptuelle. L'emphase s'effectue, quant à elle, par l'emploi d'adverbes et d'adjectifs spécifiques, et ne mène pas nécessairement à une mise en relation explicite de concepts. Elle est principalement utilisée pour introduire les métaphores *in absentia*, par opposition aux précédents marqueurs introduisant des métaphores *in presentia* [11] où la cible est explicite. Elle peut aussi apparaître conjointement aux autres types.

La classification que nous proposons se fonde, pour les trois groupes concernant la mise en relation de concepts, sur des constructions syntaxiques récurrentes, ceci dans le but d'une automatisation du repérage. Nous avons choisi de représenter ces structures avec d'éventuelles contraintes sur certains éléments, accompagnées des attributs **source**, **cible** et **tension** indiquant, lorsque cela est possible, quels éléments de la structure sont porteurs respectivement de la source, de la cible et de la nature de la relation entre celles-ci. Notation : $\{p(x)\}$ est mis pour la notion intuitive d'« ensemble des propriétés de x », et $\mathcal{V}^*(x)$, inspiré de la notation mathématique du voisinage ouvert d'un point, représentera la notion de l'ensemble des concepts proches de x , x exclu.

Les structures syntaxiques permettant d'exprimer une comparaison explicite (*plus ... que, moins ... que, aussi ... que, autant...*) correspondent au groupe **A**.

3. Le modèle « met* » de Fass [3], se contente ici d'une interprétation littérale. Fass propose un traitement en parallèle pouvant alors produire la solution recherchée, mais ne résoud pas le problème du choix de la solution la plus pertinente.

groupe de marqueurs	A	B	C	D
nombre de métaphores	2	17	9	3

FIG. 2 - répartition des métaphores selon le groupe de marqueurs

Dans la figure 1, l'attribut **tension** reflète la notion intuitive qu'une éventuelle métaphore portera vraisemblablement sur des attributs ou qualités proches de ce qu'exprime l'adjectif comparatif, qui soient relatifs à la source et ne fassent pas partie des attributs connus de la cible. Ceci traduit le fait qu'une métaphore conceptuelle existe lorsque la comparaison ne porte pas uniquement sur ce qu'exprime le comparateur. Ainsi, dans l'exemple suivant de comparaison de groupes verbaux :

ex. 3 *Pierre s'est jeté sur son repas plus vite qu'un lion sur sa proie*

la rapidité de l'action n'est pas le seul objet de la comparaison. La cible et la source de la métaphore ne sont pas seulement les **actions** comparées mais les **situations complètes**, où chaque élément est susceptible d'apporter un élément de sens supplémentaire.

La comparaison peut en outre être introduite par le sens d'un des éléments de la phrase, dit alors **marqueur lexical**, comme par exemple certains verbes, tels *ressembler, sembler, paraître*. Les constructions utilisant ces marqueurs lexicaux constituent le groupe **B** pour lequel le comparateur est sémantiquement vide quant aux qualités comparées. L'identification et l'opposition, groupe **C**, permettent aussi un repérage de la source et de la cible de la métaphore présumée. Ce sont des processus langagiers très proches de la comparaison en ce sens qu'un terme est mis en relation avec un autre de manière explicite. Une forme particulière a été reconnue dans l'exemple typique :

ex. 4 *Un truc qu'on a fait, c'est... non pas de ne parler que de généralités, mais aller mettre les mains dans le cambouis en réalisant des études de cas*⁴

Elle permet à la fois une opposition au premier terme et une identification au second. Elle permet alors d'accentuer les traits à transporter de la source à la cible, lorsqu'il y a effectivement métaphore.

3 Perspectives et ouvertures

Dans le corpus recueilli dans le cadre de l'explication, 31 métaphores ou comparaisons implicites présentaient l'une des régularités exposées dans cet article. Leur répartition selon les groupes est présentée

dans la figure 2, où le groupe **D** correspond aux marqueurs d'emphase. La méthode de recherche actuelle sur corpus électronique ne permet pas encore d'analyse comparative de ce résultat, car elle ne couvre pas tous les cas proposés.

La prédominance d'utilisation de marqueurs lexicaux (groupes B, C et D), ainsi que l'emploi de métaphores *in praesentia* (cas des structures A, B et C de mises en relation explicite), reflètent d'autant l'intention du locuteur. Ce constat va dans le sens de l'hypothèse sur laquelle se fondait notre travail, mais nécessiterait une validation par le biais d'expérimentations psycholinguistiques, dont le but serait notamment d'évaluer dans quelle mesure la compréhension d'une métaphore devient plus aisée lorsque celle-ci est exprimée avec l'une des régularités précédemment énoncées.

Dans le cas d'un marquage lexical, il apparaît en outre que certains marqueurs sont prédominants. Il convient, de ce fait, d'affiner l'étude en étudiant plus particulièrement la distribution en fonction du marqueur lexical. Nous remarquons que le cadre explicatif des métaphores relevées peut être mis en corrélation avec le type de marqueurs employés. Cet aspect de l'étude peut être mis à profit dans le cadre de la génération automatique, pour laquelle l'emploi des bons marqueurs permettrait la formulation de métaphores remarquables pour le locuteur, dans une énonciation naturelle [10]. La didactique serait ici d'un grand secours pour étudier de quelle manière introduire les analogies et métaphores présentes dans l'enseignement.

Dans le cadre d'un outil d'aide au traitement automatique des métaphores, les marqueurs précédents peuvent être utilisés pour constituer un outil de présomption de présence de la figure. Un prototype est actuellement développé, « stk », qui utilise en entrée des extraits préformatés de manière automatique du corpus « Le Monde sur CD-ROM ». Il permet actuellement d'extraire les articles contenant une chaîne spécifique (p.e. l'adverbe « littéralement », afin d'étudier le contexte d'occurrence des marqueurs, de segmenter le texte en vue d'un étiquetage, puis fait appel à l'étiqueteur de Brill [1], avec une option d'optimisation que nous avons mise au point. L'étiqueteur de Brill a été entraîné pour cette occasion sur des extraits du corpus préalablement étiquetés à la main et utilise comme lexique une adaptation de BDLEX⁵. Il est ainsi pos-

4. extrait du compte-rendu oral sur les activités du groupe de travail n° 8 (GT SCOOP) du GDR-PRC « CHM » par M. Beaudoin-Lafon à la Journée du pôle Interfaces Multimodales, Université des Sciences et Technologies de Lille, mercredi 7 décembre 1994.

5. Lexique d'environ 275 000 formes fléchies du français développé dans le cadre de l'« Action BDLEX du GRECO CNRS de la Communication Parlée »

sible d'étudier la fréquence d'emploi des marqueurs dans le cadre de métaphores (p.e. l'adverbe « littéralement » fournit ici 35 emplois métaphoriques sur 36), ceci afin de **valider** les marqueurs et de leur accorder une *valeur de qualité*. Conjointement, nous mettons au point un étiquetage spécifique rendant compte des fréquences d'occurrences des termes et de la présence d'un marquage — en cours de réalisation.

« stk » est développé dans le but d'une intégration dans un système complexe de traitement de la langue écrite. Dans cette optique, notre approche consiste à exploiter le marquage textuel conjointement à d'autres outils [4] selon la méthodologie globale :

- les métaphores conventionnelles peuvent être traitées comme indiqué par Martin dans [7],
- les métaphores nouvelles seront traitées par un agent spécialisé

1. détermination par la syntaxe et/ou le résultat du marquage de la cible éventuelle,
2. traitement d'une éventuelle analogie,
3. extraction de propriétés prototypiques,
4. proposition d'une ou plusieurs représentations du sens, établissement d'une valeur de pertinence pour chacune d'entre elles, en fonction du marquage et des résultats de l'évaluation des marqueurs (*valeur de qualité*).

Conclusion

L'étude menée a permis de mettre en évidence un certain nombre de régularités liées à l'énonciation des métaphores en français. Le repérage de métaphores n'est pas rendu possible par la seule détection des régularités proposées dans cet article. Néanmoins, la plupart des régularités repérées permettant une détection simple de la source et de la cible potentielles, ces structures, utilisées en combinaison avec d'autres outils, peuvent déboucher sur un repérage robuste.

Les premiers résultats de l'analyse sur gros corpus laissent penser que certains marqueurs lexicaux spécifiques rendent envisageable la mise au point d'un outil automatique allant dans le sens de la thèse défendue dans [5], selon laquelle le temps de traitement de la figure doit rester comparable à celui du sens dit littéral. La méthodologie à suivre dans ce sens consiste à exploiter les résultats de l'évaluation des différents types de marquages et des fréquences d'occurrences sur de grands corpus pour guider le choix du traitement à apporter aux éléments de la phrase et la pertinence des représentations de sens issues de chaque type de traitement. Une validation de ces hypothèses est actuellement amorcée par la mise au point de « stk », un outil informatique d'aide à l'évaluation des régularités et

préparatoire d'un traitement ultérieur du texte. Notons enfin que le marquage peut être utilisé pour la génération, aspect non étudié ici.

Références

- [1] E. Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, 1992. ACL.
- [2] M.P. Daniel, L. Nicaud, V. Prince, and M.P. Pery-Woodley. Apport du style Linguistique à la Modélisation Cognitive de l'Élève. *Lecture Notes in Computer Sciences*, 608:252–260, 1992. Proceedings of the International Conference on Intelligent Tutoring Systems (ITS-92), Montréal.
- [3] D. Fass. met : A Method for Discriminating Metonymy and Metaphor by Computer. *Computational Linguistics*, 17(1):49–90, 1991.
- [4] S. Ferrari. A corpus-based approach for metaphor processing. In *Proceedings of the AISB96 2nd Tutorial and Workshop Series - LEDAR Workshop*. University of Sussex, Brighton, UK. 31 March - 2 April 1996.
- [5] R.J. Gerrig. Empirical Constraints on Computational Theories of Metaphor : Comments on Indurkha. *Cognitive Science*, 13:235–241, 1989.
- [6] G. Lakoff and M. Johnson. *Metaphors we live by*. University of Chicago Press, U.S.A., 1980.
- [7] J.H. Martin. Computer Understanding of Conventional Metaphoric Language. *Cognitive Science*, 16:233–270, 1992.
- [8] M.P. Pery-Woodley. Textual clues for user modeling in an intelligent tutoring system. Master's thesis, University of Manchester, U.K., 1990.
- [9] V. Prince. Indices linguistiques pour la construction d'un modèle automatique d'analyse et de production des explications. In Sup. Telecom, editor, *Actes de l'atelier de recherche GENE du PRC IA*, pages 141–153. 15-16 décembre, 1994.
- [10] V. Prince and S. Ferrari. A Textual Clues Approach for Generating Metaphors as Explanations by an Intelligent Tutoring System. In *Proceedings of TALC96, Conference on Teaching And Language Corpora*. Lancaster University, U.K., 9-12th August, 1996. to be published.
- [11] P. Ricœur. *La métaphore vive*. Seuil, Paris, France, 1975.