Désambiguïsation sémantique des expressions polysémiques en traitement automatique du langage

LIMAME-BEN SALAH Dalila
UR Sémantique, Syntaxe & Pragmatique (Sousse/ Tunisie)
Centre Lucien TESNIERE (Besançon/France)

Résumé :

Les expressions figées constituent un problème de poids pour le traitement automatique, du fait de leur caractère polylexical, figé, non compositionnel et polysémique. Nous nous proposons de présenter « une » méthode permettant de contrer ces différents obstacles.

Le système proposé est un système dôanalyse conduisant à la reconnaissance des expressions figées polysémiques par le biais de filtres. Certes, certains indices (morphologique et syntaxique) permettent de définir si une expression est libre ou figée, mais parfois ces indices ne suffisent pas à déterminer et à identifier lôexpression potentielle. La catégorisation de ces unités polysémiques et leur actualisation en contexte ont lôavantage de réduire la polysémie laissant place uniquement à une seule des différentes acceptions (réelle ou virtuelle).

1. Introduction

Les unités minimales porteuses de sens, autrement dit les mots se répartissent en deux groupes : les mots polysémiques[1] et les mots monosémiques[2]. Les expressions figées nommées également expression idiomatique, séquence, synthème, locution, mots composés, lexie complexe, phraséologie, praxème, proverbe (é) ont une double caractéristique : lôune est de nature formelle, combinatoire, associative, lôautre est de nature sémantique. La première suppose quôune lexie 1 se combine avec une lexie 2 pour produire un sens inhérent et un sens afférent. Les langues naturelles renferment un nombre considérable dôexpressions plus ou moins figées : *tête de Turc, perdre la boussole, prendre le taureau par les cornes, casser sa pipe***é** Le traitement automatique dôexpressions figées est problématique tant au niveau de la reconnaissance quôau niveau de la génération (en traduction automatique)[3] . En définissant séparément les composants dôune expression *prendre, taureau, cornes*, le calcul du sens ne pourra être que littéral, compositionnel. Or, force est de remarquer que le sens dôune expression nôest *aucunement* le résultat du sens de ses composants. La représentation du sens lexical est une donnée importante en TAL car elle réduit voire annule les ambiguïtés et ceci par un calcul

[1] Ils constituent la majorité et permettent de limiter le lexique dôune langue naturelle.

[2] Ils représentent lôexception, ainsi les dénominations propres aux domaines de spécialité (vocabulaires spécialisés) tels que le vocabulaire professionnel, scientifique ou technique possède un sens univoque et une valeur référentielle unique.

[3] Les systèmes de traduction automatique connaissent des carences notamment en ce qui concerne le cadre sémantique.

sémantique en contexte. Il est indéniable que le problème de la polysémie reste au cî ur de lôétude des expressions.

2. Le figement

2.1. R.Martin (1997)

La locution est un syntagme figé, situé au-delà du mot et en deçà de la phrase figée

La phrase figée, pour lui, englobe les proverbes ou la phrase idiomatique en dôautres termes les phrases toutes faites spécifiques à une certaine situation du type *le jeu nôen vaut pas la chandelle*. Les propriétés caractéristiques du figement se résume à des restrictions sélectionnelles, la non-compositionnalité et à sa valeur intensionnelle[4]. La restriction sélectionnelle revient à parler de restriction combinatoire, certaines unités lexicales acceptent des associations variables. Ainsi le verbe *prendre* accepte un nombre considérable de combinaison (*prendre au collet, prendre de la terre avec une pelle, prendre sur son compte, prendre un verre…*), en revanche le verbe *hocher* admet un champ combinatoire fort limité (*hocher la tête*). Plus le champ combinatoire est restreint plus on entre dans la sphère des expressions figées. Il est à signaler que certaines unités nôapparaissent quôau sein même dôune expression (*convoler en justes noces, se mettre martel en tête*).

Une unité lexicale est figée sémantiquement lorsque le sens est non compositionnel, R. Martin définit la non compositionnalité à travers deux processus : *lôenrichissement sémantique* et *la démotivation étymologique*. Le premier renvoie à un plus sémantique, à lôau-delà des mots. Ils sont empreints dôune culture

[4] renvoie aux tests transformationnels

(*avoir le feu vert, avoir un cî ur dôartichaut*), dôun passé (*le talon dôAchille*). La métaphorisation est le procédé le plus usité qui permet la création dôexpressions. Ainsi, le sens métaphorique (non compositionnel ou afférent) peut parfois concurrencer le sens compositionnel.
Dôautre part, une expression telle que *porter le chapeau*[5] reproduit un fait historique qui nôest plus dôactualité.

2.2. Critères du figement

Le degré de figement des expressions sôétablit par le biais de différents critères. Dôune part, par le polylexicalité, une expression est constituée de plusieurs unités lexicales, par le blocage des propriétés transformationnelles et distributionnelles (paradigatique).. Ainsi, lôexpression *mettre les pieds dans le plat*

Paul a mis les pieds$_0$ dans le plat$_1$

Pronominalisation	Paul les$_0$ a mis dans le plat*
	Paul a mis les pieds dedans*
Passivation	Les pieds ont été mis dans le plat par Paul*
Extraction	les pieds, Paul a mis dans le plat ??
	Dans le plat, Paul a mis les pieds ?
Relativisation	les pieds que Paul a mis dans le plat*
Clivage	Ce sont les pieds que Paul a mis dans le plat
Insertion	Paul a mis les [petits] pieds dans le [grand] plat ?
Nominalisation	la mise des pieds dans le plat*

Le blocage des propriétés transformationnelles renforce le caractère figé de lôexpression.

[5] à lôépoque médiévale, le condamné portait un chapeau grotesque.

3. La polysémie

La notion même de polysémie est à définir étant donné que certains linguistes (tel que D. LePesant, 2003) considèrent comme étant polysémiques des constructions possédant au moins deux sens, le sens figuré et le sens propre ; pour dôautres (tel que S. Méjri, 2003) la polysémie des expressions figées ne doit pas être perçue de la même façon que pour les expressions monolexicales. Ainsi, S. Méjri parle de dédoublement sémantique : une signification littérale et une signification globale et non de polysémie. De plus, il attribue le terme de polylexicalité aux *séquences figées* et le terme de polysémie aux *unités unilexicales*. Dôautre part, parler de la polysémie dôunité monolexicale nôest pas parler dôunité polylexicale, en dôautres termes nous pouvons nous poser la question de savoir si le sens de *mettre les pieds dans le plat* renvoie à la polysémie du verbe *mettre* ou des noms *pieds* ou *plat*. La polysémie des unités monolexicales[6] rapproche des paradigmes sémantiques dôune seule et même unité ce qui est différent de la polysémie des expressions figées polylexicales. Cette dernière suppose un déséquilibre entre signifié et signifiant, le poids du signifiant est plus conséquent que celui du signifié.

Deux interprétations sont à concevoir : une première interprétation, celle de lôexpression littérale, compositionnelle et une seconde interprétation, celle de lôexpression prise dans sa totalité lôexpression polylexicale. Seulement, il est également possible de rencontrer des expressions possédant deux voire trois

[6] Unicité du signifiant et multiplicité du signifié, ceci implique une variété considérable des significations.

significations globales. Cest le cas de lexpression *avoir du coffre* où lunité lexicale *coffre* est polysémique et qui signifie avoir du souffle et, avoir du courage ou encore de lexpression *mettre les pieds dans le plat* possède certes une signification compositionnelle mais à ses côtés nous rencontrons deux autres significations[7] : aborder une question délicate avec une franchise brutale et, commettre une gaffe[8].

La polysémie est loin dêtre un fait marginal, bien au contraire elle est inhérente à toute langue naturelle. La désambiguisation sémantique dune expression polysémique[9] seffectue par le contexte (phrastique ou textuel) et lisotopie sémantique. Le contexte immédiat dune unité lexicale permet de déceler le sens adéquat de ses différentes occurrences. La récurrence de sèmes communs permet de mettre en relation la signification potentielle de lexpression avec la signification des éléments qui la précédent ou qui la suivent.

4. Système et sous-systèmes

Dans ce qui suit nous présentons un système de reconnaissance (par contraintes) permettant de définir létat dune expression (libre ou figée) mais également de définir son sens en contexte.

[7] Le sens dit figuré ou non compositionnel nest pas forcément une métaphore du sens propre mais peut être également une métonymie. Autrement dit, le sens non compositionnel constitue une extension sémantique.

[8] Définitions du *Petit Robert*, 2002

[9] mais également dune unité monolexicale.

4.1. Filtres ou contraintes syntaxiques

En utilisant des critères formels (transformationnels et distributionnels), nous avons pu créer plusieurs matrices.
Une expression libre se prête aux diverses contraintes transformationnelles[10] et distributionnelles[11]. Lôauditeur dôune expression établit une suite dôhypothèses par laquelle le sens sera défini en relation avec une série de contextes de vérité. En énonçant une expression figée, le locuteur sôexprime sans établir de correspondance entre le contenu sémantique de lôexpression énoncée et son énoncé. Le problème qui se pose lors de lôinterprétation dôune expression figée est celui du rendu sémantique, dôune interprétation fidèle[12]. Les structures syntaxiques représentent un domaine construit et analysable par conséquent nous avons élaboré une catégorisation des expressions selon diverses structures syntaxiques

[10] Les contraintes transformationnelles sont lôextraction, la pronominalisation, la relativisation, la substitution, lôinsertion dôéléments, la négation, lôinachèvement (ce critère reste une donnée intéressante pour la traitement automatique, notamment concernant les proverbes ou tournures (*au petit bonheur*é Ÿ *au petit bonheur la chance*)

[11] les contraintes distributionnelles sont le paradigme, lôaspect verbal, la négation, la catégorie du sujet selon quôil sôagisse dôun animé ou dôun inanimé (*Paul vole de ses propres ailes*Ÿ *cet oiseau vole de ses propres ailes*), lôactualisation (en dôautres termes la détermination : lôabsence dôarticle renvoie à un état archaïque, à une non actualisation ; la présence de lôarticle est un indicateur soulignant lôétat de figement *mettre les points sur les i* Ÿ *mettre les points sur le i**)

[12] le caractère de ressemblance, *il pleut des cordes*, la lexie *cordes* évoque toutes les caractéristiques aspectuelles de lôobjet corde ; cette expression exprime bien plus que lôobjet en soi (épaisseur, horizontal). Il sôagit dôune économie langagière.

possibles. Les expressions sont réparties en trois catégories :
Catégorie I regroupe les expressions sans contrainte (littérale ou idiomatique) : *tourner la page* admet toute transformation.
Catégorie II regroupe les expressions possédant des contraintes pour les expressions idiomatiques allant de 1 à 5 contraintes. Ainsi, *perdre la main* admet toutes les transformations hormis la relativisation qui n'est envisageable que dans le cas d'une expression littérale. Il est des cas tel que *tuer le ver*, cette expression n'admet que deux transformations (l'affirmation et l'insertion), les autres transformations sont valides pour l'expression littérale uniquement.
Catégorie III se scinde en deux groupes :
Groupe 1 présente des expressions ayant des contraintes au niveau du sens littéral.
Groupe 2 présente des expressions ayant des contraintes au niveau du sens littéral et du sens idiomatique.

4.2. Filtres ou contraintes distributionnelles

Cette nouvelle série de contraintes (sur le sujet et sur le déterminant) va permettre ou bien de confirmer les premières hypothèses ou de les infirmer.

4.2.1. Sujet

Le sujet constitue un élément de base dans la reconnaissance d'une expression idiomatique. Une expression telle que *avaler des couleuvres* ne peut prétendre à une interprétation idiomatique avec un sujet

ANH[13] (*ce chien a avalé des couleuvres*) mais avec un sujet AH ou NANH.
Chercher la petite bête possède un sens littéral avec un sujet ANH.
Pour *mordre à l'hameçon,* l'interprétation littérale se déduit de la catégorie du sujet ainsi un sujet ANH (tel que *poisson)* renvoie à la littéralité en revanche un sujet AH renvoie à l'idiomaticité.

4.2.2. Déterminant

Le déterminant est l'élément d'actualisation qui se produit sous la forme de l'article (défini ou indéfini) ou de la catégorie (article ou adjectif (numéral, possessif ou démonstratif)).
Article défini vs article indéfini
Ainsi, *dérouler le tapis rouge* peut prétendre à une interprétation idiomatique ou littérale alors que *dérouler un tapis rouge* ne peut admettre qu'une interprétation littérale.
Article singulier vs article pluriel
Tirer les ficelles admet les deux interprétations alors que *tirer la ficelle* admet une seule interprétation possible (littérale)
Absence d'article vs présence d'article
Acheter chat en poche est une structure archaïque, idiomatique, en revanche *acheter un chaté* est une structure libre.

[13] ANH : animé non humain ; AH : animé humain ; NANH : non animé non humain

4.2.3.Sémantisme

Certaines expressions outrepassent ces premiers filtres : *retourner sa veste* et *vider son sac* ne possèdent pas de contraintes que ce soit sur le plan syntaxique ou sur le plan distributionnel. Par conséquent, une analyse plus approfondie doit être menée pour ces expressions restées dans le flou.

Le cadre phrastique en général se révèle insuffisant pour la reconnaissance automatique de lôexpression idiomatique puisque bien souvent les traducteurs automatiques peinent face à ces expressions et donnent une traduction mot à mot ce qui fausse lôinterprétation. Un contexte plus large est plus opportun ; certains éléments textuels procurent des indices permettant de lever toute ambiguïté sémantique.

5. Désambiguïsateur

Notre système permet dôidentifier une expression telle que *prendre une veste* possédant un sens littéral et un sens idiomatique et de lever lôambiguïté , qui plus est il permet également la reconnaissance dôexpressions polysémiques et lôinterprétation sémantique en contexte. Prenons lôexemple *marcher sur les pieds*. Lôanalyse morphologique repère les mots clefs inclus dans une expression par le biais dôune base de données lexicales. Cette dernière nous informe sur les lexies ótemps, nombre, partie du discours, etc), leur reconnaissance se fait au moyen de leurs racines et de leurs flexions : *pieds* représente un mot clef dans lôexpression idiomatique dans sa forme pluriel uniquement ; lôunité prise au singulier ne pose aucun problème dôanalyse dans le cadre syntagmatique énoncé et constitue ainsi une expression

libre. Notre base de données lexicales prend la forme : <pied> <pieds, nom, masculin, pluriel, +mot clef>
Le mot clef est mis en relation avec les mots collocatifs pouvant entrer dans le cadre dôune expression polysémique.

Casser sa pipe mot clef : <pipe>

Eléments collocatifs : verbe *casser* possessif : *sa, leur*
Si le système rencontre le mot *pipe*, il poursuivra sa recherche au niveau des mots collocatifs *casser* et les possessifs. Dans le cas où le système ne trouve aucun élément collocatif, il déduira une lecture littérale. Certaines expressions potentielles élimine dôemblée lôinterprétation idiomatique dans la mesure où le mot clef ne possède pas les critères morphologiques requis. *Plomb* ne peut prendre la forme pluriel dans aucunes de nos expressions.
Lôétape suivante sôintéresse à lôanalyse syntaxique en dôautres termes à lôétiquetage du texte en faisant ainsi usage dôun analyseur morphologique, le système Labelgram (Cardey & Greenfield, 2003). A chaque mot correspond une fonction, la suite de fonctions constitue un patron syntaxique qui permettra la mise en relation avec la base de données syntaxiques. Chaque expression rassemble ses propres structures syntaxiques selon les possibilités transformationnelles.

<lever le pied>
< dverô, dô, dpiedô> <X, liste des structures syntaxiques>

dès lors quôune instanciation du patron syntaxique est possible, dans ce cas la séquence est syntaxiquement

valide et constitue alors une expression polysémique potentielle. Les expressions polysémiques refermant des structures syntaxiques spécifiques sont reconnues quêelles soient littérales ou idiomatiques. Les matrices transformationnelles donnent accès à une reconnaissance (dûne certaine catégorie) dêexpression.

Danser devant le buffet >
P*aul ne danse pas devant le buffet*

la négation indique une expression libre, lêexpression idiomatique rejette la structure négative. Toute expression potentielle ayant été identifiée subit une analyse des structures transformationnelles et une analyse distributionnelle qui vient appuyer celle-ci. En effet, une expression potentielle peut valider tous les tests transformationnels, autrement dit ne posséder aucune contrainte. Lêexpression sera reconnue au niveau structural comme étant une expression potentielle non catégorisée, par conséquent les matrices distributionnelles seront un moyen de catégoriser lêexpression. Ainsi, *mordre à lõhameçon* fait partie de la Catégorie I, la structure affirmative de lêexpression *Paul a mordu à lõhameçon* [N, Vbconj, Prép, N] est acceptée que ce soit pour le sens idiomatique ou le sens littéral. Lêéchec dêidentification va conduire le système à poursuivre sa recherche au moyen dûne analyse distributionnelle puisque certaines expressions ont des contraintes au niveau du sujet et/ou du déterminant. Ici, le sujet AH impliqué par Paul renvoie après vérification de la matrice distributionnelle sujet à une expression validée idiomatique.

Paul tient la chandelle passe à travers tous les filtres cités précédemment, le système va analyser le contexte immédiat de l'expression en relevant les mots avoisinants figurant à droite et à gauche de l'expression les unités permettant de désambiguïser le sens d'une expression. Ayant établi une liste d'éléments avoisinants potentiels, ces unités sont vérifiées dans cette liste comprenant des noms mais également des verbes.
Avaler des couleuvres peut se trouver à proximité d'unités telles que *nourrir, alimenter, bouche, ventre, festin*é
Verbe {nourrir, alimenter, mastiquer, affamer, engloutir, mâcher, dévorer}
Nom {bouche, ventre, faim, nourriture, repas, collation, festin, dîner, déjeuner}
Si, après vérification, le système valide l'unité *festin* donc il en déduira qu'il s'agit d'une expression littérale.

Conclusion
Le système présenté a pour but d'extraire des expressions «potentielles», de les analyser et également de permettre la génération en vue d'une traduction automatique. Il est encore aujourd'hui difficile de délimiter les expressions figées puisque certaines sont plus ou moins figées, et de les chiffrer puisque d'un linguiste à l'autre les chiffres varient.

Références

CARDEY S. & GEENFIELD P. (2003), *Disambiguating and tagging using systemic grammar*, Proceedings of the 8th international symposium on social communication, Santiago de Cuba, pp.559-564

ENJALBERT, P. (2005), *Sémantique et traitement automatique du langage naturel*, Paris : Hermès.

GIGUET, E., VERGNE, J. (1997), From part of speech tagging to memory-based deep syntactic analysis, *Proceedings of International Workshop of Parsing Technologies,* Massachussetts : MIT

KLEIBER, G. (1994), *Problèmes de sémantique: la polysémie en question*, Paris : PUF

LIMAME D., ALSHARAF H., CARDEY S., GREENFIELD P., SKOURATOV I. (2003), Fixedness, the complexity and fragility of the phenomenon: some solutions for natural language processing, in Acts of XVII International Congress of Linguists, Prague

MARTIN, R. (1997), Sur les facteurs du figement lexical, in *Locution entre langue et usage*, Fontenay-aux-roses : ENS Editions.

MARTINS-BALTAR, M. (éd.), (1997), *La locution entre lexique, syntaxe et pragmatique*, Paris : Klincksieck.

MEJRI, S. (éd.), (2003), *Syntaxe & Sémantique 5 : Polysémie et polylexicalité,* Caen : Presses Universitaires de Caen.

VICTORRI, B. & FUCHS, C. (1996), *La polysémie : construction dynamique du sens*, Paris : Hermès.