Que peut-on attendre d'un corpus du Web pour caractériser les facettes de l'instrumentalité ?

Patrick Saint-Dizier, IRIT-CNRS stdizier@irit.fr
Sina Zarriess, Université de Potsdam, Allemagne, zarriess@rz.uni-potsdam.de

1. Objectifs

La notion d'instrumentalité couvre un domaine conceptuel très large. Selon WordNet, l'instrumentalité réfère à un artéfact ou à un ensemble d'artéfacts qui sont instrumentaux (= qui se comportent comme des instruments) dans le but d'accomplir un certain objectif. Dans (Mari et Saint-Dizier 2002), nous identifions une triple relation : agent-instrument-action où les relations entre agent et instrument, instrument et action et agent et action sont complexes. On peut en percevoir l'étendue dans le contraste, en usage direct, entre :

Jean coupe du pain avec un couteau
Jean mange la soupe avec une cuillère ,

Où le couteau a un rôle plus 'agentif' que la cuillère : c'est lui qui coupe le pain, la cuillère n'étant qu'une sorte de récipient, elle concourt simplement au transport de la soupe. Dans ce dernier exemple, l'action de l'instrument se trouve dans une configuration métonymique, car son rôle n'est pas explicite.

En fait, presque tout objet concret ou abstrait, voire personne, peut-être utilisé comme instrument dans une action donnée. Si certains instruments sont plus prototypiques que d'autres pour accomplir une certaine action, l'étude de corpus que nous décrivons ici, loin d'être exhaustive, nous montre malgré tout l'étendue et la grande diversité des usages, qu'ils soient 'normés' ou liés à un processus de créativité langagier particulièrement notoire sur le Web. On peut ainsi *couper (certains matériaux) avec un marteau*, ou *écrire avec son coeur*. Dès lors, on réalise que l'analyse de la notion d'instrumentalité, si l'on en veut une image utilisable dans des cadres opérationnels, va nécessiter une analyse de ce que l'on peut espérer des usages rencontrés sur le Web. Notre objectif applicatif étant le développement d'un système question-réponse coopératif sur le Web pour les instruments, il est clair qu'il nous faut une analyse orientée par la tâche du corpus. Nous travaillons donc essentiellement selon une perspective qualitative.

Dans le présent travail, nous nous centrons sur l'analyse sémantique et conceptuelle des prépositions qui dénotent l'instrumentalité. Celles-ci varient assez largement d'une langue à l'autre (cf.(Kawtrakul et ali. 06) où nous étudions les marques instrumentales et les restrictions sur les instruments dans 12 langues : Français, Italien, Allemand, Espagnol, Urdu, Kashmiri, Hindi, Bengali, Thai, Malay, Arabe et Berbère). Cette analyse plurilingue a permis de dégager un certain nombre de composantes particulièrement riches de cette notion. Cependant, cette analyse, basée sur un petit corpus et sur les pratiques des auteurs, est relativement centrée sur les usages les plus prototypiques. En complément à cette analyse, nous avons conduit une analyse approfondie de corpus pour le français sur le Web, lieu où tous les usages et les excès sont possibles, de façon à mieux dégager une certaine dynamique générative de ces usages et la façon de les exploiter en question-réponses. Ce travail sera poursuivi sur l'allemand, en contraste.

Notre objectif, à travers une analyse de corpus Web est (1) de repérer un nombre assez large d'usages hors des prototypes et (2) de caractériser manuellement leur fonctionnement par rapport aux usages normés. Il est clair que le Web ne nous donnera pas des distributions exactes, il foisonne d'expressions inusitées, il laisse de côté, ou est peu représentatif, au contraire, d'usages très standards, mais il permet au moins de prendre la mesure de la diversité, et, partant, de nous guider dans une analyse plus fine du phénomène. En cela, le corpus Web est très différent du langage normé que l'on trouve dans des textes, en particulier liés à des domaines (juridique, financier, procédural). Nous aurons ainsi quelques paramètres de la diversité de la notion d'instrumentalité dans une langue peu 'controlée', en particulier en ce qui concerne :

(1) les prépositions mises en jeu (où certaines, par exemple, forcent un type instrumental pour un objet qui n'en n'est pas un naturellement pour l'action considérée),
(2) les instruments mis en jeu pour réaliser une action, caractérisée par le verbe d'action et l'objet, thème de l'action (ex. couper – verre),
(3) les différents emplois prototypiques ou non d'instruments pour réaliser une action donnée,
(4) les instruments métaphoriques (humains, chemins, etc.)

Dans chacun de ces cas, notre analyse se fera par l'identification de groupes d'usages que nous caractériserons essentiellement par des restrictions de sélection définies elles aussi inductivement. Notre approche s'est faite en collectant deux groupes de données : celles qui correspondent à des usages relativement prévisibles, et celles qui sont beaucoup plus productives et qui correspondent à des situations peu courantes. Le premier groupe permet de cerner les contours fondamentaux de la notion, tandis que le second permet d'en percevoir la productivité au quotidien.

2. La méthode

Nous employons une méthode relativement simple et bien connue, dite de 'bootstrapping' (ou inductive) basée sur des mots inducteurs soumis à un moteur de recherche de type Google. Nous soumettons un groupe incomplet, l'objectif étant d'analyser ce qui est remonté en lieu des éléments non précisés. En réponse, nous analysons manuellement autant de réponses que possible. Les réalisations se trouvent directement dans les 'snippets', petits textes remontés en même temps que les liens. Nous considérons le groupe :

Verbe – objet – préposition – instrument.

Nous testons les groupes suivants :

(1) ***Verbe + préposition donnés*** : vise à retrouver l'instrument, éventuellement aussi l'objet. Nous choisissons des verbes caractéristiques d'actions simples (*ouvrir*, *couper*, *battre*) ainsi que des verbes un peu plus abstraits (*écrire*, *expliquer*), ce point permet de mesurer la prototypicalité des instruments ainsi que le continuum ou le recouvrement qui existe entre instrument et manière. Nous avons étudié ces verbes sur un plan conceptuel de par le passé et pourrons donc alors comparer les méthodes et résultats.
(2) ***Verbe + instrument donnés*** : si l'instrument est prototypique par rapport au verbe et à l'action sous-entendue, nous aurons une caractérisation des prépositions utilisées (éventuellement par domaine) et des objets. Si l'instrument n'est pas prototypique, nous obtenons alors des usages spécialisés sur les objets ainsi que les prépositions utilisées. La créativité sur les objets est très importante. Nous remontons aussi des usages techniques.

(3) ***Préposition + instrument donnés*** : si l'on prend des instruments relativement génériques (marteau, couteau), nous obtenons alors la caractérisation des procédures dans lesquelles ils peuvent intervenir. Ceci nous permet d'identifier des classes d'usages d'un côté et des objets concernés de l'autre. Là aussi, l'approche avec deux groupes d'usages évoquée ci-dessus est utile car il est important de structurer les usages relativement attendus (1er groupe) et aussi d'analyser les comportements 'génératifs' ou inventifs du second groupe de façon indépendante.

3. Les résultats dégagés

Lorsque l'on dégage des résultats ou émet des conclusions à partir d'une analyse de corpus, en particulier du Web, il convient d'être prudent : il faut bien mesurer l'importance des données, les biais introduits (par exemple par le style Web lui-même ou par le profil des auteurs), ainsi que la façon dont on formule les conclusions, qui passe souvent par le filtre d'une certaine subjectivité (les idées que l'on teste ou que l'on a naturellement sur le phénomène analysé).

Nous présentons ci-dessous, à gros traits, les principales conclusions que nous formulons à partir de cette analyse, conclusions qui complètent, éclairent, voire amplifient les analyses conceptuelles menées préalablement.

3.1 Les prépositions

AVEC : il est clair que *avec* est la préposition prototypique, en français, de l'instrument (elle apparaît dans environ 75% des cas). Ceci nous place dans une situation particulière, d'autres langues ayant plusieurs prépositions prototypiques. *Avec* est le support de nombres d'usages directs, simples, il introduit un très grand nombre de formes plus ou moins directes :

- instruments indirects (*écrire avec un ordinateur*), parties du corps, matériaux, etc.
- très grand nombre d'usages métaphoriques, et 'excentriques' (*battre avec une fleur*), ainsi que des usages spécialisés (*battre les oeufs*),
- très grand nombre de formes métonymiques.

AU MOYEN DE, A L'AIDE DE : sont des prépositions plus formelles. Celles-ci, outre qu'elles sont moins fréquentes, se limitent à des usages plus classiques, proches de ceux rencontrés dans des documents textuels formels. Ces prépositions sont aussi plus fréquentes dans des domaines particulier (juridique, financier, médical). Dans d'autres contextes, elle jouent le rôle de 'forceur' de type quand l'instrument n'est pas très prototypique de l'action (*couper du verre au moyen d'un marteau*), l'instrument acquiert ainsi son statut.

VIA, A TRAVERS, et autres prépositions spatiales sont utilisées métaphoriquement pour figurer un emploi parfois prolongé de l'outil. Ces usages répondent aux schémas de métaphores établis depuis Lakoff et Johnson, du type : 'l'action vue comme un déplacement, les buts vus comme des cibles, les acteurs vus comme des voyageurs'.

GRACE A, A CAUSE DE : ont une coloration respectivement positive et négative : elles donnent une orientation, éventuellement argumentative, à l'usage de l'instrument.

3.2 Les champs conceptuels

La notion abstraite d'instrument est relativement complexe et a des connexions fortes avec d'autres notions telles que la causalité, la manière ou la spatialité. L'analyse de corpus Web ne permet pas de retrouver toutes les distinctions conceptuelles déjà établies, elle les nuance, les enrichit fortement, en repousse les limites. Elle permet de construire des ressources descriptives sur cette notion, via des généralisations sur les types (verbes, objet, instrument)

observés. Cette analyse est donc précieuse pour éviter de rester sur quelques cas élaborés à partir de notre intuition. Il convient néanmoins, dès lors que l'on veut ensuite faire un traitement TAL en Question-réponse, d'écarter certaines expansions trop particulières ou à connotation trop poétique. Ces décisions sont, bien entendu, arbitraires.

Examinons à présent brièvement les éléments stabilisés que nous avons pu identifier :

(1) Continuum entre instruments concrets et abstraits (souvent métaphoriques) pour une action donnée : les distributions observées sont très riches, et, par exemple, pour écrire une lettre, les instruments vont de '*stylo'*, à '*des larmes*' via ordinateur, de l'encre rouge, des éléments psychologiques et épistémiques. Ce continuum, et les difficultés évidentes d'interprétation, vont malgré tout permettre la spécification de restrictions de sélection autour de noyaux prototypiques, qui est la philosophie de notre projet PrepNet.
(2) Continuum entre instrument et manière : *écrire en abrégé, expliquer avec des arguments faux* : ces deux exemples sont à la fois des instruments et des manières. On peut souhaiter garder l'ambivalence, que nous estimons réelle, on peut aussi souhaiter forcer une interprétation par exemple à partir de la préposition et du contraste, ici en/avec (manière/instrument).
(3) Continuum entre instrument et cause, dans la mesure où il est évident qu'un instrument est, à divers degrés, la cause non volitive d'une action.
(4) Niveaux de langue : comme évoqué ci-dessus, le choix des prépositions dépend du domaine et du niveau de langage : une préposition 'lourde' (plus qu'un mot) est marquée d'une forme de focalisation.
(5) Orientation positive ou négative : sont particulièrement marquées, en particulier, l'emploi d'une préposition à connotation positive avec un instrument à coloration négative (*éduquer, expliquer, grâce à une punition*) : il est alors intéressant d'analyser les rapports entre ces termes et l'intention de l'auteur.

Ces constatations se retrouvent en corpus sur d'autres langues, avec des différences liées en particulier aux prépositions ou autres marques existant dans chacune de ces langues. Notre tâche est à présent de regrouper les usages rencontrés autour de noyaux prototypiques, afin de les contraindre et d'en expliquer le fonctionnement et l'interprétation qui s'ensuit. Ceci est très crucial pour pouvoir répondre correctement à des questions instrumentales, même simples. Nous nous limitons à la construction de noyaux prototypiques, il n'est guère possible d'élaborer des structures plus complexes, a priori, telles que des ontologies d'instruments, tant les usages sont variés, tant les actions sont elles aussi variées et nécessitent des instruments différents.