CELINE VAGUER
UMR 7114 – MoDyCo
Université Paris X – Nanterre
France
cvaguer@free.fr

Une base de données comme moyen de communication scientifique?

Notre intervention se focalisera sur les procédés de communication du point de vue de la linguistique. En effet, lorsque l'on fait de la recherche, quelle qu'elle soit, mais plus particulièrement ici en linguistique, se pose à tout linguiste la question de la *circulation de l'information scientifique* et ce, dans un souci de "rentabiliser" le travail de constitution de corpus et de manipulation sur les données ainsi rassemblées. Par le passé, les données rassemblées n'étaient guère justifiées que par l'intuition – surtout avec la grammaire générative – et parfois même, on ne savait pas clairement sur quels énoncés reposaient les résultats affichés par le linguiste (*cf.* Gross 1975[1]). Cette absence empêchait donc d'évaluer la pertinence des analyses ou des résultats. C'est à partir de ce constat et pour pallier ce manque, que nous avons entrepris de rassembler les attestations relevées dans diverses sources d'accès aux données (presse, littérature, article linguistique…) dans une base de données. La constitution d'une base de données présente bien des avantages pour la recherche linguistique (stockage illimitée d'informations, organisation des informations de façon significative, récupération des informations selon des critères de sélection, diffusion de l'information contenue dans la base…).

L'objet de notre communication se centrera sur *les modalités de la communication des corpus*, analyses, résultats qui forment l'objet de la linguistique. Pour ce faire, nous présenterons dans un premier temps la base de données que nous avons constituée en vue d'avancer dans la caractérisation syntaxique, distributionnelle et sémantique des constructions verbales de type "V *dans* GN" tels que *Le garçon se coula dans la cale, Zef s'engouffra dans le métro, Paul fouille dans sa mémoire*… Puis, nous montrerons en quoi ce type d'étude peut permettre d'améliorer les analyseurs syntaxiques (et ce faisant les logiciels de traduction automatique), et de créer des lexiques-grammaires dans le cadre de l'apprentissage d'une langue étrangère (*s'engouffrer dans* <*lieu*> = : *(precipitarse + meterse) en* vs *engouffrer* <*aliment*> = : *zamparse, engullir*).

1. Présentation des caractéristiques de la base de données Zephyr-V

La constitution d'une base de données pour rassembler un corpus et les données qui lui sont associées (syntaxique, lexicale et sémantique) n'est pas une pratique naturelle en linguistique (lorsqu'elle n'est pas spécialisée en Traitement Automatique du langage naturel – T.A.L.N.) ; c'est pourquoi nous précisons les apports de ce type d'organisation pour une recherche centrée sur les constructions verbales en *dans* mais aussi pour la communication du corpus lui-même ainsi que des analyses et résultats qui en sont proposés.

La base de données permet de regrouper logiquement un ensemble d'informations que l'établissement du corpus simple (entendu comme un ensemble d'énoncés rassemblés en vue de leur analyse) ne permet pas. Ces informations qui constituent les différentes tables de la base de données sont de divers ordres :

1.1. La table « source » contient les informations relatives à la source dont est extrait l'énoncé. Nous avons souhaité garder en mémoire ce type d'information car étant donné la diversification des sources (environ 170 qui relèvent aussi bien de la littérature que de la presse, d'ouvrages linguistiques, de corpus oraux…) cela peut permettre d'envisager des études plus ciblées : la diversification permet d'une part de couvrir un large éventail des emplois de *dans* (car les sources retenues ne sont pas également riches en attestation de la forme qui nous intéresse), d'autre part d'envisager une réutilisation de notre corpus. En effet, il se peut que d'autres chercheurs souhaitent – dans le cadre d'une étude portant, par exemple, sur les constructions verbales – sélectionner tous les énoncés issus d'une source de même nature (presse, littérature contemporaine…), ou plus spécifiquement issus d'un journal particulier (*Le Monde*, *Libération*, *L'Humanité*…) ou encore issus d'œuvres d'un auteur précis (E. Orsenna, P. Besson…). Mais on peut également envisager de sélectionner les énoncés sur une période temporelle (par exemple 1959–1990) et ce, quelle que soit la nature de la source ou sur des sources bien ciblées ; ce qui laisse possible des comparaisons des emplois (quantitatives et/ou qualitatives) selon les périodes et les sources étudiées… Toutes ces possibilités sont permises par une base de données et faciles à mettre en place (une requête suffie à extraire, rassembler l'ensemble de ces informations).

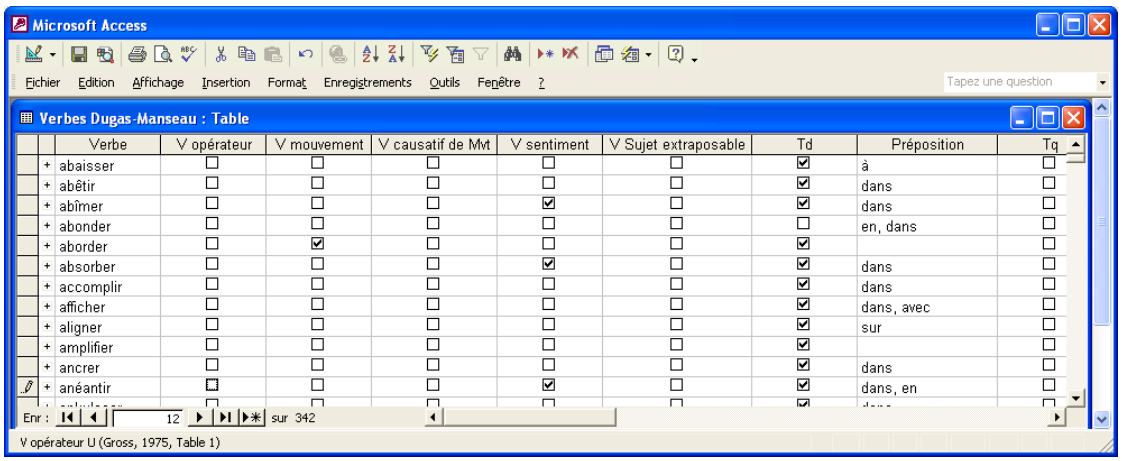
1.2. Notre étude se focalisant sur la complémentation verbale, nous avons regroupé dans la table « Analyse syntaxique du SP » l'ensemble des informations syntaxiques permettant d'identifier les syntagmes prépositionnels (SP) introduits par la préposition *dans*. Ainsi a-t-on appliqué les tests syntaxiques retenus comme

[1] M. Gross se donne l'objectif d'une description systématique des mots axée sur les propriétés syntaxiques (distributionnelles et transformationnelles) rendant compte de leur emploi dans les phrases (phrases simples réduites aux seules informations pertinentes, formées par un opérateur et ses arguments), fondée sur le principe heuristique que l'unité syntaxique comme sémantique de base est la phrase élémentaire et non le mot. Or, M. Gross ne propose aucune attestation pour illustrer chacune des entrées de ses tables. On doit donc les restituer et il n'est pas rare de trouver des exemples qui ne correspondent pas aux propriétés définitoires retenus par lui. Ainsi *amerrir*, répertorié dans la table 2, ne semble pas pouvoir se construire avec un infinitif complément (?? *L'avion amerrit récupérer les naufragés*), propriété retenue pour identifier pour identifier des verbes de <mouvement>. On peut également voir ces tables "validées" dans différentes recherches, en particulier *in* B. Lamiroy (1984) pour une étude centrée sur les verbes de mouvement ou encore dans un article de 2004 pour une analyse de la structure de la phrase en français de Belgique. On peut également consulter les travaux de J.-P. Boons, A. Guillet & C. Leclère pour s'en rendre compte (puisque ces derniers, prenant en compte les critiques adressées à M. Gross, proposent une illustration pour chaque entrée).

pertinents (suppression, détachement, déplacement…) sur l'ensemble des énoncés et tiré les conclusions de ces applications (« 0 », « 1 », « 2 », respectivement équivalents à « impossible », « possible », « douteux »). De cette analyse, nous concluons : soit le SP est ajout, soit le SP est argument, soit il peut avoir une double analyse (ajout et argument) que nous motivons. Par exemple, le SP du verbe *s'abêtir* peut être identifié soit comme *ajout* (il peut être supprimé et détaché en tête de phrase) et *s'abêtir* de signifier « devenir bête, stupide » dans un énoncé tel que *J'ai peur que le grand homme ne finisse par s'abêtir dans sa haine*, soit comme *argument* (il ne peut ni être supprimé, ni être détaché) et *s'abêtir* relève plutôt de l'interprétation « le sujet ne devient pas bête mais perd le goût de la vie » comme l'illustre *Elle n'avait cessé de s'abêtir dans une profonde absence*. Ne pas prendre en compte ce type d'analyse peut avoir des conséquences sur les traductions (comme nous le montrerons *infra*). Mentionner l'ensemble de ces analyses permet là encore d'envisager de nouvelles recherches pour qui s'intéresse à la syntaxe : par l'application des tests, nous avons forgé des énoncés qui peuvent se prêter à des nouvelles analyses comme, par exemple, l'analyse des constructions détachées : pourquoi certaines SP en *dans* sont possibles dans cette position et non d'autres ? On peut également envisager de focaliser les recherches sur l'étude descriptive (*i.e.* syntaxique, distributionnel, sémantique) d'un type d'emploi (*ajout* ou *argument*…), par exemple ou encore, d'analyser la complémentation verbale dans *Le Monde, Libération*, dans les œuvres de *P. Besson*, de *E. Orsenna*… en voyant de ce fait comment les auteurs utilisent tel type de complément introduit par la préposition *dans*.

1.3. Il s'ensuit de cette analyse syntaxique, une analyse distributionnelle et sémantique regroupée dans les tables « Identification des constituants » et « Propriétés des noms ». L'analyse distributionnelle vise à caractériser les constructions verbales *i.e.* à identifier quels verbes, noms, déterminants entrent dans la construction et quel effet de sens est ainsi obtenu. Le SP argument peut être identifié soit comme un complément circonstanciel de lieu (*Il s'engouffra dans le métro*), de cause (*Les peuples s'abêtissent [dans + à cause de] la paresse*), de cause-durée[2] (*Elle n'avait cessé de s'abêtir [dans + à cause de + au cours de] une profonde absence*), de moyen (*Il se drape [dans + à l'aide de + au moyen de] un linceul*), soit comme une « apposition » (*La principale nouveauté réside dans la création d'un nouveau taux d'imposition* = : *La création d'un nouveau taux d'imposition est la principale nouveauté*), soit comme un emploi métaphorique (*Celui qui va vous tuer croupira dans une prison*). Là encore du point de vue de la communication du corpus, de son analyse et des résultats, l'ensemble de ces informations permettent d'envisager d'autres exploitations comme, par exemple, des études plus systématiques sur les noms qui entrent dans l'interprétation du SP comme un complément circonstanciel de cause, ou sur l'étude distributionnelle des constructions suivant qu'elles sont caractérisées par le même sens : quels types de verbes, noms et déterminants se trouvent associés quand le SP a trait au lieu, au temps… On peut aussi envisager des études quantitatives sur l'emploi de la préposition *dans* ou de telle construction en *dans* associée à tel sens (spatial, causal, temporel…) : quel est l'emploi le plus souvent usité de la construction dans le journal *Le Monde*, à l'oral, à l'écrit, dans tel genre (presse, littérature…). Il s'agirait en quelque sorte de voir quel emploi est le plus représentatif ou le plus fréquent dans les performances.

Conclusion : La constitution d'une base de données présente des avantages pour son concepteur mais aussi pour d'autres linguistes puisque l'on peut diffuser l'information qu'elle contient. Notre objectif est de pouvoir diffuser cette base de données en la rendant accessible *via* Internet. En effet, si elle est centrée, certes, sur les constructions verbales en *dans*, elle peut être utilisée à d'autres fins que nous avons déjà évoquées. En effet, un chercheur en syntaxe se trouvera satisfait de trouver ici un ensemble d'énoncés pourvus de leur analyse (chose rarissime) et de pouvoir rassembler, par une simple requête, les données sur lesquelles il souhaite travailler. De même, un chercheur en linguistique comparée se trouvera à son tour satisfait de trouver ici rassemblé un ensemble d'énoncés en français, mais encore un chercheur travaillant sur les prépositions se trouvera satisfait de voir mentionnées pour chaque verbe les prépositions que celui-ci sélectionne (l'ensemble de ces informations se trouvent rassemblées dans la table « Verbes Dugas-Manseau » dont une illustration est proposée ci-dessous) et de trouver déjà rassemblés des énoncés illustrant divers emplois de *dans* (il ne lui restera qu'à rassembler les énoncés contenants une autre préposition (*à, sur, en*…) !).



Verbe	V opérateur	V mouvement	V causatif de Mvt	V sentiment	V Sujet extraposable	Td	Préposition	Tq
abaisser	☐	☐	☐	☐	☐	☑	à	☐
abêtir	☐	☐	☐	☐	☐	☑	dans	☐
abîmer	☐	☐	☐	☑	☐	☑	dans	☐
abonder	☐	☐	☐	☐	☐	☐	en, dans	☐
aborder	☐	☑	☐	☐	☐	☑		☐
absorber	☐	☐	☐	☑	☐	☑	dans	☐
accomplir	☐	☐	☐	☐	☐	☑	dans	☐
afficher	☐	☐	☐	☐	☐	☑	dans, avec	☐
aligner	☐	☐	☐	☐	☐	☑	sur	☐
amplifier	☐	☐	☐	☐	☐	☑		☐
ancrer	☐	☐	☐	☐	☐	☑	dans	☐
anéantir	☐	☐	☐	☑	☐	☑	dans, en	☐

Illustration de la table « Verbes Dugas-Manseau »

[2] Cette dénomination a été proposée par D. Leeman (1985 & 1987).

2. Débouchés

Etant donné le nombre d'attestations et l'ensemble des analyses proposées, ce type de corpus peut fournir énormément de données en vue d'une réutilisation dans différentes recherches. On peut en effet y voir un corpus permettant des tris selon les auteurs… mais aussi un corpus permettant l'élaboration de lexiques et leur exploitation dans le domaine de l'acquisition de la langue, tout comme dans celui du traitement automatique des données.

Du point de vue du traitement automatique des données, le fait d'avoir identifié chacune des constructions de la sorte *Croupir dans [<objet> ; <lieu>]* peut être utile pour la levée des ambiguïtés (notamment celles qui sont dues à la polysémie). Ainsi les prédicats verbaux peuvent être caractérisés par le type d'arguments propres à chacun de leurs emplois : « la séparation des sens au moyen des propriétés syntaxiques qui s'associent à chacun des sens, constitue le seul moyen de désambiguïser dont dispose les 'machines' » (Lamiroy 1998 : 12). Ce type d'analyse est donc indispensable. Autrement dit, chaque emploi de *se draper, s'embarbouiller* se trouve décrit de manière appropriée selon le type d'argument-objet qu'il sélectionne (<état> ou <vêtement> pour (1) *se draper*, <parole> ou <vêtement> pour (2) *s'embarbouiller*) :

(1) *se draper* :	N_0 : hum ; N_1 : <état>	*Gladys se drape dans sa dignité*
	N_0 : hum ; N_1 <vêtement>	*Il se drape dans son manteau*
(2) *s'embarbouiller* :	N_0 : hum ; N_1 <parole>	*Jujube s'embarbouille dans ses phrases*
	N_0 : hum ; N_1 <vêtement>	*Elle s'embarbouille dans ses loques*

Ce type d'étude a son utilité dans le domaine de la traduction car on peut envisager d'améliorer, d'affiner les analyseurs syntaxiques :

(1) *se draper* :	N_0 : hum ; N_1 : <état>	(F) : *Gladys se drape dans sa dignité*
		(E) : *Gladys se encastilla en su dignidad*
	N_0 : hum ; N_1 <vêtement>	(F) : *Il se drape dans son manteau*
		(E) : *Se envuelve en su abrigo*
(2) *s'embarbouiller* :	N_0 : hum ; N_1 <parole>	(F) : *Jujube s'embarbouille dans ses phrases*
		(E) : *Jujube se embrolla en sus palabras*
	N_0 : hum ; N_1 <vêtement>	(F) : *Elle s'embarbouille dans ses loques*
		(E) : *Se enreda en sus andrajos*

Pour ce faire, il faut associer à chaque nom le codage de la classe à laquelle il appartient (<état>, <vêtement>… ce classement des noms est opéré par certains linguistes – cf. en particulier G. Gross et ses « classes d'objet ») et introduire une règle que l'on peut qualifier de « cotextuelle » : lorsque le verbe *se draper* a pour complément un nom d'<état>, il se traduit en espagnol par *encastillarse* et non *envolverse* (emploi réservé aux noms de <vêtement>). Là encore, pour les traductions, savoir ce type de choses peut permettre de les affiner puisque, si on a la même forme verbale (*se draper dans, s'embarbouiller dans*) en français pour deux types d'emplois, ce n'est pas le cas en espagnol, ce qui peut provoquer des erreurs de traduction si les énoncés ne sont pas suffisamment spécifiés – comme on a pu s'en rendre compte par l'observation de quelques traductions obtenues au moyen de logiciels de traduction automatique, grand public (*Systran*, *Reverso*), accessibles par Internet. En effet, les deux logiciels proposent la même traduction pour *se draper* = : *envolverse* et ce, quelle que soit la nature (<état> ou <vêtement>) de l'argument-objet. De même, aucun des deux logiciels ne semble connaître la forme verbale *s'embarbouiller* (elle doit donc être absente de leur dictionnaire). Ainsi proposent-ils les formes *embarbou* et *embarbouille* pour éviter de laisser vide la place occupée par le prédicat verbal.

- Traductions obtenues à l'aide du logiciel de traduction automatique Systran :

(1) se draper :	Gladys se envuelve en su dignidad
	Se envuelve en su abrigo
(2) s'embarbouiller :	Jujube se embarbou en sus frases
	Se embarbou en sus andrajos

- Traduction obtenue à l'aide du logiciel de traduction automatique Reverso :

(1) se draper *:*	Gladys se envuelve su dignidad
	Se envuelve su abrigo
(2) s'embarbouiller *:*	Azufaifa se embarbouille en sus frases
	Se le embarbouille en sus andrajos

Conclusion : Par l'étude des constructions verbales « V *dans* GN » nous avons pu d'une part avancer dans la caractérisation syntaxique, distributionnel et sémantique de la structure mais cela nous a aussi permis d'autre part de rendre compte des subtilités de la langue dont les analyseurs syntaxiques présents dans les logiciels de traduction automatique ne tiennent et ne rendent pas compte. Or, comme nous avons pu le mettre en évidence, cela est d'une grande utilité pour une 'meilleure' traduction mais aussi pour l'établissement d'index bilingue qui

pourraient servir dans le domaine de l'acquisition de la langue puisque, si l'on a une forme en français, il en existe deux en espagnol selon le type d'argument-objet : « le processus d'analyse automatique du langage naturel, qui est d'une complexité extraordinaire, sera d'autant plus performant que les informations fournies par les dictionnaires et grammaires électroniques seront complètes » (Lamiroy 1998 : 18). Ainsi, notre étude peut-elle servir à compléter les dictionnaires existants et à préciser les grammaires électroniques déjà disponibles. L'index que nous proposons (*infra*) pourrait servir aussi bien pour les apprenants de l'espagnol que pour les apprenants du français afin de pallier les manques existants et d'affiner l'acquisition du lexique pour une meilleure utilisation, en vue d'emplois plus appropriés. En effet, les variantes sont souvent opaques pour les locuteurs et les mentionner ne peut permettre d'accroître leur performance et d'affiner leur maîtrise de la langue.

		N_0	N_1	Enoncés (F)	Enoncés (E)
s'abîmer	*naufragar, zozobrar*	-hum	<lieu>	*Le navire s'abîme dans les flots*	*La embarcacion naufraga / zozobra en las olas*
	sumergerse, ahundirse, sumirse	hum	<état>	*Elle s'abîme dans ses pensées*	*Se sumerge / se ahonda / se sume en sus pensamientos*
se draper	*encastillarse*	hum	<état>	*Gladys se drape dans sa dignité*	*Gladys se encastilla en su dignidad*
	envolverse	hum	<vêtement>	*Il se drape dans son manteau*	*Se envuelve en su abrigo*
s'embarbouiller	*embrollarse*	hum	<parole>	*Jujube s'embarbouille dans ses phrases*	*Jujube se embrolla en sus frases*
	enredarse	hum	<vêtement>	*Elle s'embarbouille dans ses loques*	*Se enreda en sus andrajos*
s'emberlificoter	*trabarse, enredarse*	hum	<lieu>	*On s'emberlificote dans un enchevêtrement de ronces*	*Se traba / se enreda en un encabestramiento de espinos*
	embrollarse	hum	<état>	*Gustave Flaubert est emberlificoté dans sa suffisance*	*Gustave Flaubert esta embrollado en su suficiencia*
s'embourber	*atascarse*	-hum	<lieu>	*Il s'embourbe dans les sables mouvants*	*Se atasca en las arenas movedizas*
	liarse, caer	hum	<état>	*Ils ne veulent pas s'embourber dans le bas du classement*	*No quieren liarse por la parte baja de la clasificación.*
germer	*brotar*	-hum	<aliment>	*Un pépin germe dans la pomme*	*Una pipa brota en la manzana*
	germinar nacer	-hum	<état-pensée>	*Une idée germe dans son esprit*	*Una idea germina / nace en su espíritu*
mouiller	*mojar*	hum	<liquide>	*Il s'est mouillé dans une flaque d'eau*	*Se mojo en un charco de droga*
	comprometerse	hum	<état>	*Il s'est mouillé dans une histoire de drogue*	*Se compromete en una historia de droga*

Références bibliographiques

BOONS J.-P., GUILLET A. & LECLERE C. (1976a) *La structure des phrases simples en français. Constructions intransitives*, Genève/Paris, Droz.

GROSS, G. (1994) « Classes d'objets et traitement de la synonymie », *Supports, opérateurs, durées,* Annales littéraires de l'Université de Besançon 516, Série Linguistique et Sémiotique, vol.23, Paris, Les Belles Lettres : 93-102.

GROSS G. & VIVES R. (1986) « Les constructions nominales et l'élaboration d'un lexique-grammaire », *Langue française*, 69, Paris, Larousse : 5-27.

GROSS, M. (1975) *Méthodes en syntaxe*, Paris, Hermann.

GUILLET A. & LECLERE C. (1992) *La structure des phrases simples en français : Constructions transitives locatives*, Genève/Paris, Droz.

HABERT B., NAZARENKO A. & SALEM A. (1997) *Les linguistiques de corpus*, Paris, Armand Colin / Masson.

LAMIROY, B. (1983) *Lingvisticae Investigationes. Supplementa* 11 : *Les verbes de mouvement en français et en espagnol. Étude comparée de leurs infinitives*, Amsterdam/Philadelphia, John Benjamins Publishing Co.

LAMIROY, B. (1998) « Le lexique-grammaire. Essai de synthèse », *Travaux de linguistique*, 37, Bruxelles, Duculot : 7-24.

LAMIROY B. & KLEIN J. (2004) « La structure de la phrase en français de Belgique », *in* C. Leclère, E. Laporte, M. Piot & M. Silberztein (eds.) *Lingvisticae Investigationes. Supplementa 24* : *Lexique, Syntaxe et Lexique-Grammaire*, Amsterdam/Philadelphia, John Benjamins Publishing Co. : 343-371.

LEEMAN, D. (1985) « Tentative de caractérisation d'un complément circonstanciel : *Dans mon affolement, je lâchai mon panier de cerises* », *Linx*, 12, Centre de Recherches Linguistiques de l'Université Paris X : 97-146.

LEEMAN, D. (1987) « A ma grande surprise », *Revue québécoise de linguistique*, 16 : 2, Montréal, Université de Québec : 225-265.

VAGUER, C. (2003) « *Corpus*, vous avez dit *corpus* ! De la notion de *corpus* à la création d'un "corpus informatisé" », Communication aux *3èmes Journées de la linguistique de corpus*, Lorient (11-13 septembre 2003) (à paraître aux Presses Universitaires de Rennes).

VAGUER, C. (2004a) « Constitution d'une base de données : les emplois de *dans* marquant la "coïncidence" », *Revue Française de Linguistique Appliquée*, IX-1, juin 2004, Amsterdam, Éditions "De Werelt" : 83-97.

VAGUER, C. (2004b) *Les constructions verbales « V* dans *GN ». Approches syntaxique, lexicale et sémantique*, Thèse de l'Université Paris 10 – Nanterre, sous la direction de Danielle Leeman.

Logiciels de Traduction Automatique :

- Systran : *http://www.systranbox.com/systran/box*

- Reverso : *http://reverso.net/text_translation.asp*