CONSTRUCTION COGNITIVE D'UN MOTIF : COOCCURRENCES TEXTUELLES ET ASSOCIATIONS MEMORIELLES

Dominique Longrée (Uliège), Sylvie Mellet, Frédéric Lavigne (Université Côte d'Azur)

Depuis [Sinclair 1991], il est reconnu que tout locuteur use d'un « prêt-à-parler », ensemble d'expressions toutes faites dont la source et le stockage ne se situent pas seulement dans la mémoire individuelle de chaque locuteur particulier, mais relèvent des ressources linguistiques partagées par une communauté de locuteurs. Depuis un peu plus d'une dizaine d'années, diverses recherches ont montré que ces ressources incluaient des structures textuelles qui n'entraient pas dans le champ de la phraséologie au sens strict, d'où la proposition d'étendre le domaine de celle-ci [Legallois & Tutin 2013]. [Longrée & Mellet 2013] ont montré que ces structures textuelles et les formes phraséologiques pouvaient l'une et l'autre être décrites comme les occurrences d'une nouvelle unité linguistique, le motif. On s'intéresse ici aux modes de stabilisation, de mémorisation et de reconnaissance des motifs textuels, à travers une double approche : d'abord par le biais d'une expérience de psychologie cognitive basée sur l'effet d'amorçage sémantique, ensuite par une étude statistique cooccurrentielle en corpus. Les objets choisis pour l'étude sont des motifs textuels latins composés de trois éléments lexicaux. La convergence remarquable des résultats de l'étude expérimentale et du traitement quantitatif des données textuelles apporte un enrichissement original et très prometteur aux modèles qui tentent de rendre compte du processus d'association sémantique.

motifs textuels, association sémantique, psycholinguistique, acquisition, mémorisation

Since [Sinclair 1991], it is generally recognized that speakers make use of a collection of "ready built" expressions whose origin and storage are not only located within the individual memory of each speaker, but are part of a set of common resources shared by a distinct group of speakers. For more than ten years now, researches have shown that this set includes textual structures which cannot be considered as "phraseological" in a narrow sense. Hence it was suggested spreading the field of the phraseology [Legallois & Tutin 2013]. [Longrée & Mellet 2013] showed that both these textual structures and the phraseological units can be described as occurrences of a new kind of linguistic unit, the motif. This paper deals with the mechanism grounding the stabilization, the memorization and the recognition of the textual motifs. It will combine two different methods applied to a set of Latin motifs made of three lexical items: on the one hand, from the point of view of a psycho-cognitive approach, an experiment allowing to evaluate the multiple semantic priming effects, on the other hand, from a statistical point of view, a study of collocations in a reference corpus. Both methods, psychological and statistical, give noticeably convergent and promising results which enhance considerably the theoretical models trying to describe the processing of semantic association.

Textual ‘motifs’, semantic association, linguistic psychology, acquisition, lexical storage

1. Introduction

Depuis [Sinclair 1991], il est reconnu que tout locuteur use d’un « prêt-à-parler », ensemble d’expressions toutes faites dont la source et le stockage ne se situent pas seulement dans la mémoire individuelle de chaque locuteur particulier, mais relèvent des ressources linguistiques partagées par une communauté de locuteurs et alimentent, notamment, le lexique d’une langue. C’est le fameux « principe idiomatique » dont l’étude a trouvé deux champs d’application principaux : la phraséologie (champ lexical) et les grammaires de construction (champ lexico-syntaxique). [Longrée & Mellet 2013] ont proposé d’étendre l’analyse du principe idiomatique au champ discursif en étudiant des expressions complexes significativement récurrentes, qui sollicitent simultanément le niveau lexical et le niveau grammatical, et auxquelles une fonction discursive stable confère le statut de véritables marqueurs de structuration des discours. Ces expressions, appelées « motifs » textuels, élargissent le champ d’application de la phraséologie. En dépit de quelques variations de surface qui rendent difficile leur extraction automatique, les motifs sont immédiatement reconnaissables comme tels par un locuteur natif et, assez rapidement, par un apprenant langue seconde, ce qui oblige à s’interroger sur leurs modes de stabilisation, de mémorisation et de reconnaissance.

D’un point de vue linguistique, il s’agit de comprendre le fonctionnement du langage dans sa dimension idiomatique la plus complexe, celle qui associe la dimension syntagmatique textuelle à la dimension paradigmatique. L’impact de cette analyse sur les études et pratiques en acquisition des langues secondes pourrait être important. Or plusieurs questions se posent sur le rôle, dans la mémorisation, de la fréquence du motif et de la force de l’association entre les termes qui le composent. L’apparition du premier élément du motif déclenche-t-elle l’attente des termes suivants sur la simple base de phénomènes statistiques de cooccurrences des mots deux à deux (probabilité conditionnelle d’apparition du troisième terme sachant chacun des deux précédents indépendamment l’un de l’autre) ? Ou, au contraire, le schème abstrait sous-jacent au motif textuel, sa structure grammaticale profonde et la fonction discursive qui le caractérise confèrent-ils aux premiers termes du motif le statut d’une amorce complexe spécifique induisant une attente forte des termes suivants (probabilité conditionnelle d’apparition du troisième terme sachant la combinaison préalable des deux précédents) ? Et,

dans le cas où le schème abstrait sous-jacent joue un rôle important, tous les types de segments phraséologiques (idiomes, motifs ; lexicaux, lexico-grammaticaux ; strictement figés, avec variations) présentent-ils les mêmes propriétés au regard de la mémorisation et de la reconnaissance ?

D'un point de vue cognitif, les psychologues sont intéressés à l'analyse et la modélisation de ces motifs à deux titres au moins : il s'agit pour eux de compléter le modèle selon lequel l'amplitude de l'effet d'amorçage sémantique peut être prédit par la force d'association entre l'amorce et la cible [voir Brunel & Lavigne 2009 pour une revue des questions sur le sujet], d'une part en y introduisant la linéarité d'un motif textuel (ou d'une formule idiomatique) et d'autre part en allant au-delà de la simple association entre deux mots. En dépassant le niveau de l'association binaire, qui est à la base de la plupart des modèles actuels, la prise en compte des motifs permet de mieux cerner les processus dynamiques d'amorçage sémantique et, par là-même, les mécanismes de la compréhension discursive ; en effet, le processus de décodage sémantique dépend du contexte lexical antérieur [Hagoort, Hald, Bastiaansen & Peterson 2004 ; Hald, Steenbeek-Planting & Hagoort 2007 ; Menenti, Petersson, Scheeringa & Hagoort 2009] en raison du processus d'activation prélexicale ou prédictive déclenché par les mots présentés antérieurement (*priming*). Contextualiser ce processus et, par là-même, intégrer de nouveaux paramètres à son analyse, devrait permettre un enrichissement des modèles d'amorçage sémantique en considérant que l'amorçage d'un troisième mot (la cible) dépend de la combinaison des deux mots amorces précédents et pas seulement de la somme des deux activations que ces amorces génèrent indépendamment l'une de l'autre.

Dans le cadre de ce double intérêt du linguiste et du psychologue cognitiviste, cette étude consiste à confronter les données statistiques extraites d'un grand corpus textuel et les résultats d'une expérience psycholinguistique d'amorçage sémantique d'un mot cible par deux mots amorces, avec pour objectif de formuler une hypothèse sur la forme prise en mémoire par l'association récurrente de trois éléments formant un motif discursif ; ce faisant, on pourra approfondir le lien entre l'exposition du sujet parlant à la récurrence observée de certains termes cooccurrents dans les textes – phénomène que la statistique descriptive tend à présenter comme statique –, et du fonctionnement de réseaux associatifs en mémoire, qui relève du processus (donc d'une dynamique) et dont l'activation permet de faire fonctionner le « prêt-à-parler ».

L'étude portera sur des données latines. Le choix d'une langue morte peut paraître paradoxal. Il a été motivé par le fait que l'existence de formules récurrentes et structurantes dans les textes latins est depuis longtemps attestée par la philologie classique ; certains de ces motifs ont donné lieu à des études approfondies[1] ; les apprentis latinistes les rencontrent régulièrement dans leur cursus de lettres classiques et leur apprentissage du latin repose pour une bonne part sur cette exposition récurrente et sur la maîtrise des conditions d'emploi (y compris génériques) de telles formules. Une expérience de psychologie cognitive avec des participants ayant suivi ce cursus jusqu'à un niveau avancé peut donc présenter un certain intérêt. En outre, la confrontation entre les résultats de celle-ci et les données statistiques textuelles peut s'appuyer sur un corpus représentatif de textes latins classiques lemmatisés et étiquetés[2], corpus disponible et exploitable par le logiciel d'exploration et de traitement statistique Hyperbase[3].

2. Motifs *vs* segments répétés

Le motif est une construction lexico-grammaticale [Gledhill & Frath 2007] associée à un nombre restreint de formes, et dont la fonction sémantique et discursive reste comparable d'une réalisation à l'autre.

Sur le plan formel, un motif se définit par l'association récurrente de n éléments du texte muni de sa structure linéaire [Legallois 2006], laquelle donne une pertinence aux relations de successivité et de contiguïté [Longrée, Luong & Mellet 2008 ; Mellet & Longrée 2009]. Ainsi, si le texte est formé d'un certain nombre d'occurrences des éléments A, B, C, D, E, un motif pourra être la micro-structure récurrente ACD ou bien encore AAA, etc., sans qu'on préjuge ici de la nature des éléments A, B, C, D, E en question. En effet, la notion de motif est conçue comme un moyen de conceptualiser le caractère multi-niveau de certaines formes récurrentes qui

---

[1] La plus connue est sans doute celle de [Chausserie-Laprée 1969] sur les « clichés de liaison ».

[2] Corpus du Laboratoire d'Analyse Statistique des Langues Anciennes de l'Université de Liège. Voir sa présentation à l'adresse htpp:// www.cipl.ulg.ac.be/Lasla/tlatins.html

[3] Logiciel développé par Étienne Brunet au sein du laboratoire BCL (http://www.unice.fr/bcl/) ; version web à l'adresse www.hyperbase.unice.fr

sollicitent à la fois le lexique, les catégories grammaticales et la syntaxe, éventuellement la prosodie, la métrique.

Outre cette caractéristique formelle, certains motifs ont un rôle fonctionnel : ces motifs offrent alors un cadre collocationnel accueillant un ensemble de paramètres susceptibles d'accompagner la structuration textuelle et/ou de caractériser des textes de genres divers. La stabilité de cette fonction textuelle permet la reconnaissance du motif en dépit des éventuelles variations de forme qui peuvent l'affecter jusqu'à un certain point. Ces variations, en nombre limité, sont la permutation dans l'ordre de deux termes, la commutation d'un terme avec un autre terme au sein d'une liste paradigmatique restreinte, la suppression ou l'ajout d'un terme facultatif, ou encore la variation morphologique telle que l'alternance singulier / pluriel ou présent / passé [Longrée & Mellet 2013]. Voici un exemple de motif latin transitionnel narratif ; ce motif textuel, qui ne se rencontre que chez les historiens, permet au narrateur de changer de décor, de passer d'un lieu à un autre et d'ouvrir ainsi un nouvel espace narratif. Il associe la conjonction *dum* (« pendant que »), un démonstratif neutre pluriel sujet en fonction anaphorique « ces choses, ces événements », un complément circonstanciel locatif « à tel endroit » et la forme verbale *geruntur* (« se déroulent ») ; le prototype en est :

(1)	*Dum haec in Gallia geruntur*
	« Tandis que ces événements se déroulent en Gaule »

et voici quelques exemples de variantes :

- permutation dans l'ordre de deux termes

(1)	a.	*Haec dum in Gallia geruntur*

- commutation au sein d'un paradigme lexico-sémantique :

(1)	b.	*Dum ea in Gallia geruntur* (anaphorique *ea* au lieu du démonstratif *haec*) *Dum haec in Gallia aguntur* (verbe *ago* au lieu de *gero*)

- suppression du circonstanciel

(1)	c.	*Haec dum [Ø] geruntur*

- variante morphologique sur le circonstanciel (au cas locatif ou à l'ablatif prépositionnel ou à l'accusatif prépositionnel)

(1)	d.	*Dum haec Romae geruntur*
		« Tandis que ces événements se déroulent à Rome »
(1)	e	*Dum haec in Gallia geruntur*
		« Tandis que ces événements se déroulent en Gaule »
(1)	f	*Dum haec ad Ilerdam geruntur*
		« Tandis que ces événements se déroulent aux portes d'Ilerda »

La micro-structure d'un motif combine donc à la fois des éléments de stabilité assurant sa mémorisation et sa reconnaissance et des éléments de transformation assurant le jeu inhérent aux divers usages en discours.

Par son aptitude à développer des variantes, par son sémantisme et sa fonction stables, le motif textuel est une structure plus complexe, plus ouverte et plus souple que les « segments répétés » [Salem 1987] qui se définissent de manière purement formelle par la stricte récurrence à l'identique d'une série de formes contiguës, série qui n'est pas toujours porteuse de sens.

Néanmoins, pour l'expérience réalisée en psycholinguistique, nous avons sélectionné des motifs relativement simples, à savoir des motifs de trois termes contigus, à fonction de structuration textuelle ou pas. Il s'agit donc d'associations ternaires qui, concrètement, coïncident avec des segments répétés à trois éléments (ou tri-grammes) – porteurs de sens, néanmoins. Ce choix a été guidé, d'une part, par la nécessité de trouver en corpus un nombre suffisant d'occurrences attestées des formes soumises au sujet lors de l'expérience psycho-cognitive et, d'autre part, par l'état de l'art sur les associations sémantiques et le *priming*, par rapport auquel le passage à des associations ternaires et l'intégration de la linéarité textuelle constituent déjà une complexification suffisante des paramètres étudiés. Les tri-grammes sont donc constitués de deux termes initiaux, A1 et A2, qui servent d'amorce au motif et qui entraînent l'apparition du troisième terme appelé cible (C) par référence à la théorie psycholinguistique du priming (voir Lavigne *et al*. 2011, 2012, 2013 pour des revues de questions).

3. Présentation du matériel d'étude et de ses propriétés cooccurrentielles dans le corpus

Sur la base de nos compétences de latinistes et grâce à une recherche automatique avec Hyperbase dans le corpus du LASLA[4], appliquée aussi bien aux formes lexicales qu'aux étiquettes grammaticales associées à chaque forme des textes, nous avons sélectionné une liste de tri-grammes qui répondent donc aux critères suivants : (i) être composés de seulement trois termes (A1-A2-C) ; (iii) être porteur d'un sémantisme complet et reconnaissable ; (iii) offrir un nombre suffisant d'occurrences dans le corpus pour permettre le traitement statistique et notamment le calcul de la force d'attraction cooccurrentielle entre chacun des termes deux à deux ainsi qu'entre le dernier terme et le couple formé par les deux premiers.

La liste ainsi obtenue comprend 72 tri-grammes (voir annexe 1).

Les relevés d'occurrences ont donné lieu à un tableau récapitulatif dont un extrait, concernant les trigrammes suivants, est présenté ci-dessous (tableau 1) :

(2)	a.	*Quibus*	*rebus*	*cognitis*
		Relatif de liaison ABL.PL.	'Choses' ABL.PL.	'Connaître » Participe passé passif ABL.PL.
		« Ayant appris ces choses »		
(2)	b.	*Quod*	*cum*	*auditum [est]*
		Relatif de liaison NOM.SG.	CONJ. 'quand'	'Entendre' PARF.PASSIF 3SG.
		« Quand on eut appris cela »		
(2)	c.	*Eo*	*cum*	*uenisset*
		Adverbe lieu anaphorique.	CONJ. 'quand'	'Venir' SUBJ. PQP 3SG.

[4] La base « Latin » du DVD Hyperbase rassemble des textes allant de Plaute à Tacite, soit un ensemble de 18 auteurs différents et un total de 1 709 956 occurrences correspondant à 24 288 vocables différents.

		« Alors qu'il en était arrivé là »		
(2)	d.	*Vt*	*ita*	*dicam*
		CONJ COMP.	ADV. 'ainsi'	'Dire' SUBJ. PST. 1SG.
		« Pour ainsi dire »		
(2)	e.	*Si*	*eis*	*uideatur*
		CONJ 'si'	Pr. anaphorique DAT.PL.	'Sembler' SUBJ. PST. 3SG.
		« S'il leur semble bon, s'ils en jugent ainsi »		
(2)	f.	*Si*	*Ita*	*feceris*
		CONJ 'si'.	ADV. 'ainsi'	'Faire' FUT.ANT.ACT 2SG.
		« Si tu fais ainsi »		

Pour chacun de ces motifs, le tableau 1 affiche le nombre d'occurrences de chaque terme pris séparément (A1 ; A2 ; C), le nombre d'occurrences de la séquence contiguë des deux premiers termes (A1-A2), puis de la séquence du premier et troisième termes séparés par autre chose que l'élément attendu dans le motif (A1-X-C), et enfin le nombre d'occurrences du motif complet (A1-A2-C) ;. Ces données quantitatives donnent des premières indications intéressantes sur la force d'association entre chacun des termes pris séparément ainsi qu'entre le couple A1-A2 et le dernier terme C.

Tableau 1 : extrait du tableau des 72 tri-grammes soumis à l'expérimentation

A1	A2	C	A1-A2	A1-X-C	A2-C	A1-A2-C
Quibus	Rebus	Cognitis	Quibus rebus	Quibus X cognitis	Rebus cognitis	Quibus rebus cognitis
2566	1426	82	91	17	27	16

Quod	Vbi	Auditum	Quod ubi	Quod X auditum	Vbi auditum	Quod ubi auditum
11662	2162	36	54	6	4	4
Eo	Cum	Venisset	Eo cum	Eo X uenisset	Cum uenisset	Eo cum uenisset
2296	13894	141	20	9	21	9
His	Paratis	rebus	His paratis	His X rebus	Paratis rebus	His paratis rebus
1718	34	1426	2	43	4	2
Vt	Ita	Dicam	Vt ita	Vt X dicam	Ita dicam	Vt ita dicam
14876	2862	556	59	61	48	45
Si	Eis	Videatur	Si eis	Si X uideatur	Eis uideatur	Si eis uideatur
11327	387	303	8	3	2	2
Si	Ita	Feceris	Si ita	Si X feceris	Ita feceris	Si ita feceris
11327	2862	77	74	8	5	5

La lecture de ce tableau partiel et du tableau général présenté en annexe 1 suggère deux remarques préliminaires.

Première remarque : les motifs commencent le plus souvent par une forme très fréquente ; le terme C est plus fréquent que le terme A1 seulement dans 13 motifs sur les 72 motifs retenus. Par voie de conséquence, au sein de la majorité des motifs la cooccurrence est asymétrique entre les éléments A1 et C, en ce sens que l'élément A1 ne consacre qu'une faible part de ses effectifs globaux à la cooccurrence avec les termes A2 et C pour former le motif. Au contraire la proportion des effectifs de l'élément C consacrée à cette cooccurrence formant motif est généralement bien plus importante. Ce type d'asymétrie caractérise les formules en voie de figement[5].

[5] Sur l'asymétrie de la cooccurrence et le lien avec le figement, voir [Luong *et al*. 2010].

Deuxième remarque : parmi les associations ternaires que nous avons retenues pour cette étude, certaines semblent relativement figées. Les occurrences du motif complet épuisent alors une grande partie, voire la totalité du potentiel d'enchaînement de la séquence A1-A2 (que nous appellerons désormais « double amorce ») : ainsi, pour une double amorce en *his paratis* (2 occurrences), on a 2 occurrences du motif complet (*his paratis rebus*) ; ou encore sur 59 doubles amorces en *ut ita*, 45 produisent le motif *ut ita dicam*. En revanche, la marge de liberté après une double amorce en *quod ubi* est beaucoup plus grande (4 occurrences du motif seulement pour 54 de la double amorce). C'est cette plus ou moins grande liberté d'enchaînement après les amorces que nous avons tenté d'évaluer par deux méthodes complémentaires ; nous commençons par présenter l'expérimentation psycholinguistique d'amorçage sémantique multiple (avec double amorce).

4. Etude expérimentale

4.1 Protocole expérimental

Il s'agit de vérifier l'impact cognitif des associations attestées en corpus textuel et de voir si une convergence se dessine – ou pas – entre les résultats fournis respectivement par l'approche textométrique et par une approche expérimentale. Le focus de cette étude sera de déterminer si l'effet d'amorçage sur le troisième terme du motif, la cible C, est explicable à partir du schéma classique d'activation par chacune des amorces (A1, A2) considérées séparément (via des associations A1-C et A2-C) ou s'il est nécessaire d'introduire dans le modèle la notion de double amorce (A1-A2), dont les effets ne se réduiraient pas à la composition des effets de A1 et de A2. Il y aurait là un effet spécifique lié non seulement à la fréquence des associations A1-C et A2-C mais aussi à la fréquence du motif ternaire A1-A2-C en tant que motif textuel ou segment phraséologique.

La méthode utilisée est celle d'une tâche de décision lexicale permettant de mesurer l'effet d'amorçage sémantique (*semantic priming*). Le participant, après avoir été exposé à une double amorce, doit décider si une suite de lettres qui lui est présentée ensuite (la cible) est ou non un mot latin. Le temps nécessaire à cette prise de décision est classiquement plus rapide si la force d'association entre les deux mots amorces et la cible est grande. Dans notre cas d'étude, par exemple, après un amorçage par le groupe *dum haec,* le temps nécessaire pour reconnaître que *geruntur* est un mot latin devrait être plus court que pour reconnaître *nuntiata*. L'expérience se situe donc dans le paradigme de la chronométrie mentale selon lequel plus un traitement

cognitif est facile, plus la tâche est effectuée rapidement. L'hypothèse est ici que si la cible fait partie d'un motif avec les deux amorces, alors l'amorçage (son activation) devrait être plus fort que si l'association de la cible à chaque amorce ne forme pas motif. Un tel résultat enrichirait les modèles d'amorçage sémantique en ajoutant un mécanisme d'amorçage combinatoire (entre les trois mots) supplémentaire à l'amorçage entre paires de mots.

Les participants à l'expérience, au nombre de 30, sont des étudiants de master de l'Université de Liège et de l'Université de Bruxelles. Ils ont tous la même formation en latin et une pratique comparable des textes de référence.

Chaque participant a été exposé au même nombre de triplets constitués d'une double amorce et d'une cible, à savoir 72 triplets répartis en deux types : ceux avec association forte et ceux association faible entre les paires constituées de la double amorce et de la cible (la force d'association a été estimée en fonction du taux de cooccurrence de ces paires en corpus : les motifs dont la fréquence en corpus est supérieur ou égal à 25% des effectifs de la double amorce relèvent d'une association forte). La comparaison des temps de décision lexicale sur des cibles associées fréquemment *vs* rarement avec les amorces permet de mesurer l'amorçage sémantique multiple (Lavigne *et al*., 2011).

Pour chaque force d'association entre paires (*i.e.* entre la cible et la double amorce), forte ou faible, la même proportion de cibles formant motif avec la double amorce (comme *geruntur* après *dum haec*) et de cibles ne faisant pas motif (car empruntées à un autre motif, comme *rebus* de *His paratis rebus* après *dum haec*) a été présentée aux participants.

Enfin, pour que les sujets aient une vraie tâche de décision lexicale à effectuer, des leurres (pseudo-mots latins) leur ont également été présentés en guise de cibles.

4.2 Résultat principal

Le principal résultat de cette expérience fait apparaître une interaction entre la fréquence du motif (A1-A2-C) et la fréquence des associations par paires (*i.e.* A1-C, A2-C), avec un impact sur le temps de reconnaissance des mots cibles. Autrement dit, la cible est activée lorsque les deux amorces lui sont associées, formant une paire avec la cible *et* que les trois mots constituent un motif ; il s'agit là d'un effet d'amorçage spécifique au motif de trois mots, non réductible aux seules activations entre paires de mots deux à deux. La fréquence du motif global est donc une

variable statistiquement pertinente qui amplifie l'activation de la cible (C) par les deux amorces. L'expérience prouve ainsi que la fréquence des paires A1-C et A2-C n'a d'impact sur l'activation de la cible C que lorsque ces trois termes sont inclus dans un motif stabilisé ; aucun impact significatif n'est observé lorsque les trois termes ne forment pas motif. Il s'agit donc d'un nouvel effet d'amorçage, propre aux structures récurrentes de type motif, qui se superpose à l'effet d'amorçage classique par association par paires de mots et qui fait intervenir aussi la fréquence du trigramme.

Ces résultats expérimentaux enrichissent les modèles de l'amorçage sémantique et nous incitent à reprendre de manière plus précise et approfondie une recherche cooccurrentielle et quantitative en corpus pour valider par des outils de linguistique textuelle quantitative les résultats de notre expérience psycholinguistique. Ce retour au texte nous permettra de réintégrer dans l'étude du fonctionnement des motifs leur aptitude à développer des variantes, propriété fondamentale que les contraintes de l'expérience psycholinguistique nous ont obligés à laisser momentanément de côté.

5. Les cooccurrents spécifiques des amorces

5.1 Prise en compte des variantes de chaque motif

Comme on l'a dit précédemment, un motif est une structure stable dont divers éléments assurent la mémorisation et la reconnaissance tandis que d'autres éléments, variables, assurent le jeu inhérent aux divers usages en discours.

Voici les éléments de variation relevés en corpus pour les motifs présentés au § 2 et qui nous servent d'échantillon.

Quibus rebus cognitis :

anaphorique au lieu du relatif de liaison : *his rebus cognitis* ;

singulier au lieu du pluriel : *qua re cognita, ha re cognita* ;

absence de l'anaphorique : *Ø re cognita* ;

variation lexicale sur le participe : *nuntiata* « annoncée » ou *animaduersa* « remarquée » au lieu de *cognita* « apprise, connue » : *ha re nuntiata, ha re animaduersa* ;

Quod ubi cognitum [est] :

variation lexicale sur le verbe : *audio* « entendre dire » au lieu de *cognosco* « apprendre » : *quod ubi auditum est*

Eo cum uenisset :

relatif de liaison quo au lieu de l'anaphorique *eo* : *quo cum uenisset* ;

variation lexicale sur le verbe : préfixé *peruenisset* au lieu du simple *uenisset*

Vt supra memoraui :

variation lexicale sur le verbe : *demonstro* « montrer » et *dico* « dire » au lieu de *memoro* « rappeler » : *ut supra demonstrauimus* / *diximus* ;

variation sur la personne verbale : première du singulier au lieu de la première du pluriel : *ut supra memoraui*, *demonstraui*, *dixi* ; ou passif impersonnel « comme il a été montré plus haut » : *ut supra demonstratum [est], dictum [est]*

variation sur le subordonnant : relatif à la place de *ut* : *quem supra demonstrauimus*, *de quibus supra diximus*.

Si eis uideatur :

Variation sur le temps verbal : imparfait *uideret* au lieu du présent *uideatur*, en contexte passé.

Pour chacun des motifs, nous reproduisons les valeurs chiffrées du tableau 1 en y ajoutant les occurrences de leurs variantes attestées.

Ces données quantitatives seront exploitées pour évaluer le lien cooccurrentiel entre chacun des termes pris séparément ainsi qu'entre la cible et les amorces du motif. La prise en compte des variantes, outre qu'elle fait sens linguistiquement, permet d'atteindre des effectifs suffisants pour autoriser l'usage de tests statistiques.

5.2 Tableau récapitulatif : nombre d'occurrences en valeurs absolues[6]

Tableau 2 : extrait du tableau des 72 tri-grammes soumis à l'expérimentation, augmenté des principales variantes des motifs

Dum	Haec	Geruntur	Dum haec	Dum X geruntur	Dum haec geruntur
1415	3606	37	31	2	18
		Aguntur		Dum X aguntur	Dum haec aguntur
		27		0	5
		Total des variantes		Total des variantes	Total des variantes
		64		2	23
Quibus	Rebus	Cognitis	Quibus rebus	Quibus X cognitis	Quibus rebus cognitis
2038	1336	65	76	2	12
His			His rebus	His X cognitis	His rebus cognitis
1582			129	0	12
Qua	Re	Cognita	Qua re	Qua X cognita	Qua re cognita
2273	1464	137	39	1	6
Hac			Hac re	Hac X cognita	Hac re cognita
			64	0	4
					Ø re cognita
					12
		Nuntiata		Qua X nuntiata	Qua re nuntiata
		32		0	9
				Ea/Hac X nuntiata	Ea/Hac re nuntiata

[6] On rappelle que le corpus de référence comporte 1 709 956 mots.

				3	1
					Ø re nuntiata
					2
		Animaduersa		Qua X animaduersa	Qua re animaduersa
		10		0	7
Total des variantes	Total des variantes	Total des variantes	Total des variantes	Total des variantes	Total des variantes
5893	2800	244	308	6	65
Quod	Vbi	Cognitum [est]	Quod ubi	Quod X cognitum est	Quod ubi cognitum est
10608	2073	20	41	1	3
		Auditum est		Quod X auditum est	Quod ubi auditum est
		8		0	4
		Total des variantes		Total des variantes	Total des variantes
		28		1	7
Eo / Quo	Cum	(Per) venisset	Eo cum / Quo cum	Eo X (per)venisset	Eo cum (per)venisset / Quo cum (per)venisset
542 / 389	8299	135	20 / 8	0	10 / 4
Vt	Ita	Dicam	Vt ita	Vt X dicam	Vt ita dicam
13536	2541	539	54	5	44
Vt	Supra	Memoraui / Memorauimus	Vt supra	Vt X memoraui / memorauimus	Vt supra memoraui / memorauimus
13536	310	41	45	1	6
		demonstrauimus / demonstratum est		Vt X demonstrauimus / demonstratum est	Vt supra demonstrauimus / demonstratum est
		48		2	13

		Dixi / Diximus		Vt X dixi / diximus	Vt supra dixi / diximus
		393		15	13
		Dictum est		Vt X dictum est	Vt supra dictum est
		66		1	11
		Total des variantes		Total des variantes	Total des variantes
		548		19	43
Si	Eis	Videatur/ Videretur	Si eis	Si X uideatur/uideretur	Si eis uideatur/uideretur
10734	366	566	8	(3 avec X = Ø)	7
Si	Ita	Feceris	Si ita	Si X feceris	Si ita feceris
10734	2541	71	70	2	6

5.3 Exploitation par le test des cooccurrents spécifiques

Il s'agit ici de déterminer quels sont les mots qui, dans le corpus, se trouvent significativement attirés dans la sphère de l'amorce du motif (d'abord de A1 seul, puis de la double amorce A1-A2). La méthode utilisée pour ce faire est d'extraire du corpus tous les paragraphes contenant l'amorce, de dénombrer les effectifs de tous les mots cooccurrents de cette amorce qui apparaissent dans les paragraphes en question et de comparer les effectifs ainsi observés en corpus aux effectifs théoriquement attendus dans le cadre d'une distribution aléatoire. La significativité de l'écart entre les effectifs calculés et les effectifs observés est mesurée par un écart réduit [Muller 1973 : 69]. On estime que lorsque cet écart est supérieur à 2.5 (+ 2.5 pour les excédents et – 2.5 pour les déficits), il est significatif et mérite d'être pris en compte par l'analyste.

Les excédents significatifs mettent en évidence les mots spécifiquement attirés par l'amorce et donc remarquablement présents dans son contexte immédiat. Il est à noter que, dans cette méthode, les termes dont on étudie la cooccurrence ne sont pas nécessairement contigus – ce

qui n'est pas gênant pour des motifs qui, par définition, acceptent des expansions en leur sein (*dum haec geruntur* / *dum haec in Gallia geruntur*).

On va voir que cette liste de termes spécifiques inclut toujours la cible du motif.

Mais on va constater aussi que la différence entre la liste des cooccurrents privilégiés du seul mot A1 et ceux de la double amorce A1-A2 est parfois remarquable.

***Dum haec geruntur* (« tandis que ces événements se déroulent »)**

La forme *geruntur* est le premier cooccurrent spécifique de la double amorce *dum haec* (A1-A2) avec un écart de 14.37. C'est aussi le premier cooccurrent de la variante *dum ea* avec un écart de 4.22 et même du syntagme *dum* + démonstratif avec un écart de 13.88.

Après l'amorce simple par la conjonction *dum* (A1), *geruntur* est rétrogradé en deuxième position dans la liste des cooccurrents spécifiques (ce qui atteste cependant d'une force d'association toujours très importante avec un écart de 11.52 très significatif). Réciproquement – et c'est là un point tout à fait remarquable – les cooccurrents privilégiés de la cible *geruntur* intègrent en très bonne position les termes *dum*, *haec* et *ea* ; il en va de même pour la cible *aguntur* : nous observons donc une réversibilité de la force d'association qui atteste du caractère fortement phraséologique de la formule.

Revenons aux cooccurrents spécifiques des amorces : si la variante *aguntur* est bien le deuxième cooccurrent spécifique de la double amorce *dum haec*, avec un écart de 7.09, elle est rétrogradée au 23ème rang de la liste des cooccurrents de l'amorce simple *dum*.

La conclusion est que la double amorce *dum haec* initie un motif quasi figé et enclenche par là-même non seulement le choix du lexème verbal qui sera prédicat de la subordonnée, mais aussi celui de la voix verbale, du mode, du temps et de la personne : *geruntur* principalement, *aguntur* secondairement.

***Quibus rebus cognitis* (ayant appris ces choses »)**

Le terme C (*cognitis*) est le premier cooccurrent spécifique de la double amorce A1-A2 avec un fort écart réduit de 9.72 (le second cooccurrent de la liste offre un écart sensiblement plus faible de 7.81). En revanche, C n'est que le 15ème cooccurrent spécifique de A1 (*quibus*) considéré seul, avec un écart réduit de 6.52.

La cible *cognitis* a pour premiers cooccurrents *rebus* (13.52), *his* (7.19), *Caesar* et *quibus* (6.20). L'attraction est là aussi réciproque.

La variante *his rebus* a également pour premier cooccurrent *cognitis* (8.43).

Au singulier, la double amorce *qua re* a pour premiers cooccurrents spécifiques *animaduersa* (« ayant remarqué », avec un écart de 9.99), *nuntiata* (« ayant reçu la nouvelle de », 9.52), *Caesar* et *cognita* (« ayant appris », 5.77). Après la variante *hac re*, on trouve *cognita* en 3ème position.

En revanche, la liste des cooccurrents spécifiques du mot *qua* (A1) considéré isolément contient la forme *animaduersa* (6.0) au 15ème rang seulement ; la forme *nuntiata* se trouve au 52ème rang (4.58) et la liste ne contient pas la forme *cognita*.

***Quod ubi cognitum [est]* (« quand il eut appris cela »)**

*L*a liste des cooccurrents spécifiques de *quod ubi* (A1-A2) est très parlante ; dans l'ordre d'importance des écarts réduits, on relève : *auditum* (« quand il eut entendu », 5.79), *cognitum* (5.5), *animaduertit* (« quand il eut remarqué », 3.94), *conspexit* (« quand il eut perçu », 3.51). En revanche, aucune forme de ces verbes *AVDIO*, *COGNOSCO*, *ANIMADVERTO* et *CONSPICIO* ne se trouve dans la liste des cooccurrents spécifiques associés au seul *quod* (A1).

***Eo cum uenisset* (« alors qu'il en était arrivé là »)**

La forme *uenisset* est le premier cooccurrent spécifique de *eo cum* avec un écart réduit de 10.08 et le treizième cooccurrent après la variante *quo cum* avec un écart de 4.17. Il est rétrogradé à la 32ème place dans la liste des cooccurrents spécifiques du seul *eo* (6.23) et il ne figure pas dans celle de *quo* (3.78).

***Vt ita dicam* (« pour ainsi dire »)**

La forme *dicam* est le premier cooccurrent du syntagme *ut ita*, avec un écart réduit record de 17.52. Il tombe à la 60ème place après le seul *ut* (écart de 8.02). Le figement très marqué de cette formule, que nous avions déjà signalé, se confirme ici. Cela explique sans doute l'absence de variantes.

***Vt supra memoraui* (« comme je l'ai rappelé plus haut »)**

Ni *memoraui*, ni les variantes *memorauimus* (« nous l'avons rappelé »), *demonstraui* (« je l'ai montré »), *demonstrauimus* (« nous l'avons montré »), *demonstratum [est]* (« il a été montré »), *diximus* (« nous l'avons dit »), *dictum [est]* (« il a été dit ») ne sont des formes présentes dans la liste des cooccurrents spécifiques de la seule conjonction *ut*. En revanche, les cinq premiers cooccurrents spécifiques de la double amorce *ut supra* sont les suivants : *dictum* (8.28), *demonstratum* (7.76), *demonstrauimus* (7.40), *dixi* (7.29), *memorauimus* (5.44). *Diximus* est en 8ème position avec un écart réduit de 4.17. La double amorce joue donc ici pleinement son rôle d'amorçage. *Memoraui*, *memorauimus*, *demonstrauimus*, *demonstratum* ont comme premier cooccurrent *supra* (A2) alors que *ut* (A1) n'apparaît pas ou n'apparaît que très loin dans les listes. Et réciproquement *supra* (« plus haut ») a comme cooccurrents spécifiques l'ensemble de ces formes verbales (avec cependant des scores d'écarts réduits moindres que pour la double amorce) ; il est à noter cependant que ce motif implique nécessairement son intégration dans une subordonnée, soit relative, soit introduite par *ut*. L'adverbe *supra* et la construction syntaxique jouent donc un rôle pivot dans la construction de ce motif, sans pour autant que la force d'association qui lit les trois termes du motif soit la somme des forces d'association des paires.

***Si ita feceris* (« si tu fais ainsi »)**

La forme *feceris*, qui est seulement au 73ème rang dans la liste des cooccurrents spécifiques de *si* considéré isolément, se trouve au 1er rang dans celle de *si ita*. Réciproquement, *si* est le premier cooccurrent de *feceris* (8.63).

***Si eis uideatur* (« s'il leur semble bon »)**

Ce motif présente une particularité intéressante : sans doute en raison d'un moindre figement du motif, la forme verbale *uideatur* n'est qu'au 9ème rang (3.98) de la liste des cooccurrents spécifiques de la double amorce *si eis*. La variante *uideretur* (imparfait au lieu du présent) est au 6ème rang dans la liste de *si eis* (5.2). Ni l'une ni l'autre ne figurent dans celle de *si* considéré isolément (A1). La différence d'amorçage reste donc importante, mais il est indéniable que la force d'association cooccurrentielle entre la double amorce et la cible est ici moins forte que dans d'autres motifs.

Cet ensemble de données chiffrées semblent indiquer que les forces d'association sont aussi importantes dans les principales variantes textuelles d'un motif que dans la forme standard qui a été retenue pour l'expérimentation. Considérer l'ensemble des variantes d'un motif conforte les conclusions sur l'existence d'une micro-structure sous-jacente dont la stabilisation repose sur des forces d'association intégrant moins les effets d'amorçage classique entre paires de mots, que les effets spécifiques liés à l'existence du motif ternaire. Ces constats tendent à faire penser que l'existence de variantes, loin de nuire à la mémorisation et la reconnaissance du motif, contribue au contraire à les consolider.

Conclusion

Les résultats expérimentaux convergent avec les données en corpus : ce résultat est tout à fait remarquable. Il tend à prouver les éléments suivants.

- Les textes et les discours donnent à voir, sous forme de réseaux cooccurrentiels, les associations sémantiques enregistrées en mémoire et contribuent simultanément à construire ces associations, à les enrichir, à les stabiliser ; conjoindre les deux sources d'information – corpus et expérimentation cognitive – pour mieux appréhender le fonctionnement de ces réseaux se révèle particulièrement pertinent ; cela est particulièrement vrai dans le cas d'une langue morte et seconde dont l'apprentissage se fait par exposition aux textes transmis. De plus, l'amorçage sémantique en langue maternelle est lui aussi proportionnel à la fréquence d'occurrence des paires amorce-cible (voir par exemple Van Petten 2014). On peut donc penser que l'exposition d'un enfant aux discours qui le font baigner dans sa langue maternelle produit ce même effet.

- Ces réseaux cooccurrentiels reposent sur des associations lexicales qui s'organisent selon des micro-systèmes phraséologiques fortement structurés et structurants. Ainsi, nous avons prouvé que certains motifs à trois termes (A1-A2-C) affichent des propriétés telles que, tant sur le plan textuel que cognitif, ils doivent être analysés comme des paires associant, selon des modalités propres, une double amorce [A1-A2] avec une cible C, plutôt que comme des tri-grammes A-B-C constitués uniquement de relations entre chacun des termes pris deux à deux. Sur le plan de la linguistique textuelle, ce résultat confirme a posteriori le bien-fondé de la notion de « motif », définie au sens strict, et dont nous pensons utile de faire une nouvelle unité textométrique.

- Enfin, l'effet d'amorçage caractéristique des motifs ternaires enrichit les modèles cognitifs des psychologues en montrant que les motifs sont traités en temps réel lors de la lecture des mots [Lavigne *et al*., 2014, 2016]. Il serait donc utile que les linguistes l'intègrent davantage dans leurs propres modèles, notamment pour analyser les effets d'attente ou d'attente déçue dans la dynamique phrastique et textuelle, en lien avec les processus de réception et de décodage et le sentiment de complétude phrastique.

Bibliographie :

Brunel, N. & F. Lavigne. 2009. Semantic priming in a cortical network model. *J Cogn Neurosci* 21(12). 2300–2319.

Chausserie-Laprée, J.-P. 1969. *L'expression narrative chez les historiens latins, Histoire d'un style.* Paris: E. de Boccard.

Depecker, L. 1999. Monème, synthème et phrasème. Essai d'introduction du concept de phrasème dans la théorie fonctionnaliste. *La Linguistique* 35(2). 43-62

Firth, J.R. 1957. *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Gledhill, C. & P. Frath. 2007. Collocation, phrasème, dénomination : vers une théorie de la créativité phraséologique. *La Linguistique* 43(1). 63-88.

Hagoort, P., L. Hald, M. Bastiaansen & K.M. Petersson. 2004. Integration of word meaning and world knowledge in language comprehension. *Science* 304(5669). 438–441.

Hald, L.A., E.G. Steenbeek-Planting & P. Hagoort. 2007. The interaction of discourse context and world knowledge in online sentence comprehension. Evidence from the N400. *Brain Res* 1146. 210.

Lavigne, F., L. Dumercy & N. Darmon. 2011. Determinants of multiple semantic priming: a meta-analysis and spike frequency adaptive model of a cortical network. *J Cogn Neurosci* 23(6). 1447-1474.

Lavigne, F., L. Dumercy, L. Chanquoy, B. Mercier & F. Vitu-Thibault. 2012. Dynamics of the semantic priming shift: behavioral experiments and cortical network model. *Cogn Neurodyn* 6(6). 467-483.

Lavigne, F., L. Chanquoy, L. Dumercy, & F. Vitu-Thibault, F. 2013. .Early dynamics of the semantic priming shift ». *Adv Cogn Psychol* 9(1). 1-14.

Lavigne, F., F. Avnaïm & L. Dumercy. 2014. Inter-synaptic learning of combination rules in a cortical network model. *Front Psychol* 5. 842. doi:10.3389/fpsyg.2014.00842

Lavigne, F., D. Longrée, D. Mayaffre & S. Mellet. 2016 : Semantic integration by pattern priming: experiment and cortical network model. *Cognitive Neurodynamics* 10(6). Springer Verlag, doi:10.1007/s11571-016-9410-4.

Legallois, D. 2006. Des phrases entre elles à l'unité réticulaire de textes. *Langages* 163. 56-70.

Legallois, D. & A. Tutin (eds). 2013. Vers une extension du domaine de la phraséologie. *Langages* 189.

Longrée, D., X. Luong & S. Mellet. 2008. Les motifs : un outil pour la caractérisation topologique des textes. In S. Heiden & B. Pincemin (eds), *JADT 2008*, Actes des 9èmes Journées internationales d'Analyse statistique des Données Textuelles, vol. 2, 733-744. Lyon: Presses universitaires de Lyon. http://lexicometrica.univ-paris3.fr/jadt/jadt2008/pdf/longree-luong-mellet.pdf

Longrée, D. & S. Mellet. 2013. Le motif : une unité phraséologique englobante ? Etendre le champ de la phraséologie de la langue au discours. *Langages* 189. 65-79.

Longrée, D. & S. Mellet. 2018 [sous presse]. Towards a topological Grammar of Genres and Styles: a way to combine paradigmatic quantitative analysis with a syntagmatic approach**.** In D. Legallois *et al*. (eds), *Grammar of Genres and Styles*, 141-163. Bruxelles: de Gruyter. https://doi.org/10.1515/9783110595864-007

Luong, X., E. Brunet, D. Longrée, D. Mayaffre, S. Mellet & C. Poudat. 2010. La cooccurrence, une relation asymétrique ? ». In S. Bolasco *et al*. (eds), *JADT 10*, Actes des 10èmes Journées internationales en Analyse statistique des Données Textuelles, vol 1, 321-332. Milan: LED. http://lexicometrica.univ-paris3.fr/jadt/jadt2010/allegati/JADT-2010-0321-0332_028-Luong.pdf

Mellet, S. & J.P. Barthélemy. 2007. La topologie textuelle : légitimation d'une notion émergente. *Lexicometrica* (numéro spécial « Topographie et topologie textuelles »), article consultable sur http://www.cavi.univ-paris3.fr/lexicometrica/numspeciaux/special9/mellet.pdf

Mellet, S. & D. Longrée. 2009. Syntactical Motifs and Textual Structures. *Belgian Journal of Linguistics* 23 ('New Approaches in Textual Linguistics'). 161-173.

Menenti, L., K.M. Petersson, R. Scheeringa, & P. Hagoort. 2009. When elephants fly: differential sensitivity of right and left inferior frontal gyri to discourse and world knowledge. *J Cogn Neurosci* 21(12). 2358-68. doi: 10.1162/jocn.2008.21163

Muller, C. 1973. *Initiation aux méthodes de la statistique linguistique*. Paris: Hachette Université.

Salem, A. 1987. *Pratique des segments répétés*. Paris: Klincksieck.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Van Petten, C. 2014. Examining the N400 semantic context effect item-by-item: relationship to corpus-based measures of word co-occurrence. *Int J Psychophysiol* 94. 407–419.