Les séries linéaires dans le discours : marques, opérations et structures sous-jacentes.

Agata Jackiewicz

Les séries linéaires dans le discours : marques, opérations et structures sous-jacentes.

Agata Jackiewicz

Laboratoire LaLICC (Paris IV, CNRS)
Université de Paris-Sorbonne (ISHA)
96, bd Raspail, 75006 Paris –France

Agata.Jackiewicz@paris4.sorbonne.fr

Le développement de nouveaux modes d'accès aux documents numériques, que nous connaissons aujourd'hui, notamment dans le domaine de la fouille automatique de textes, conduit à mettre l'accent sur l'exploitation de la structuration interne des documents et de sa signalisation. De nouveaux besoins viennent en particulier de la médiation des savoirs (Souchier 1998), où l'on cherche entre autres à (i) « recycler » des textes (littéraires, philosophiques…) pouvant être « consommés » sur une longue période, (ii) vulgariser des textes scientifiques très spécialisés pour en assurer une plus large diffusion. S'il ne s'agit pas de modifier le contenu de ces textes, on peut néanmoins chercher à en faciliter la réception, notamment par une segmentation (plus) adéquate et divers enrichissements attachés à celle-ci.[1] Cette phase de transformation serait sans doute grandement facilitée par un outil informatique d'assistance capable de signaler au sein d'un texte donné ses différentes articulations marquées « de l'intérieur », par le discours lui-même, et surtout celles dont le balisage ténu, hétérogène ou incomplet peut induire un lecteur pressé en erreur.

En nous appuyant sur plusieurs études linguistiques portant sur l'auto-représentation de l'énonciation (Authier-Revuz, 1995), le métadiscours (Borillo, 1985 ; Luc et Virbel 2001), l'activité épilinguistique (Culioli 1987), les marqueurs d'intégration linéaire et les énumérations (Turco et Coltier 1988 ; Péry-Woodley 2000), la prise en charge énonciative (Desclés et Guentchéva 2000), la reformulation (de Gaulmyn 1987 ; Rossari 1997), ainsi que la citation (Mourad 2001), nous cherchons à proposer un ensemble de critères (idéalement, une carte de valeurs sémantiques) permettant de rendre compte d'une manière unifiée des principales opérations méta-énonciatives rencontrées dans des textes écrits. Ces critères sont destinés à une procédure automatique permettant d'annoter certains objets textuels ou certains lieux et/ou modes d'articulation de ces objets, l'annotation pouvant être exploitée notamment pour la navigation intra-documentaire ou la synthèse sélective de textes (Crispino et al. 2004). L'ensemble du travail s'inscrit dans la lignée des recherches sur l'encadrement du discours (Charolles 1997) et emprunte le cadre formel de la méthode d'exploration contextuelle (Desclés 1997).

Dans la présente contribution, nous nous proposons de montrer à partir d'une expérience complète (allant d'une analyse linguistique à l'implantation), menée sur l'objet textuel « série linéaire » (Jackiewicz 2002 ; Jackiewicz et Minel 2003 ; Couto et al. 2004), comment la tâche informatique de repérage d'objets et de structures de discours (i) oriente le travail linguistique de collecte et d'analyse de marques d'organisation textuelle, (ii) questionne le cadre méthodologique, faisant notamment apparaître la nécessité de représenter et de traiter des « faisceaux » d'indices contextuels, relatifs à plusieurs opérations discursives en présence,

[1] C'est ce que font par exemple certains journaux ou revues lorsqu'ils publient des textes rédigés par des non collaborateurs (experts, témoins…). La segmentation en sections, le choix et l'insertion de titres et de divers outils (chapôs, accroches, clés…) sont assurés par la rédaction.

(iii) ouvre de nouvelles voies de recherche, notamment sur des rapports existant entre certaines structures discursives (entre la sériation et la reformulation, par exemple).

Saisir l'organisation discursive d'un texte, c'est identifier les segments, leur hiérarchie, leurs relations, puis en déduire la nature des structures ou configurations sous-jacentes. Or, les études linguistiques classiques focalisent leur attention, soit sur les marques (le plus souvent, les connecteurs), soit sur les relations de discours sémantico-pragmatiques entre segments adjacents ou proches. Bien qu'elles apportent des descriptions fines de certains procédés de liage ou d'encadrement, ces études ne « remontent » pas jusqu'aux structures textuelles partiellement engendrées par les procédés en question.

En effet, reconnaître dans un texte une structure ou configuration discursive particulière, par exemple une série, c'est savoir répondre à plusieurs questions qui touchent à la nature de l'objet textuel lui-même, pour lesquelles les outils de la linguistique de la phrase ne seront que de peu d'aide. Qu'est-ce qu'une série ? Quels sont ses éléments constitutifs ? Ses composants obligatoires ? Quelles contraintes une série impose-t-elle à ses composants, en particulier aux items ? Qu'est-ce qu'une série bien formée ? Quand est-ce qu'une série est considérée comme déviante ? Existe-t-il des normes de formation d'une « bonne » série ? Ces normes ou pratiques dépendent-elles des genres textuels ? Comment la nature discursive des items influence-t-elle les caractéristiques structurelles de la série ? Le choix du balisage discursif (*en premier lieu, premièrement, tout d'abord…*) de la série a-t-il un impact sur sa morphologie ?

Dans le discours, la sériation apparaît comme une opération complexe faisant appel à la fois au balisage des objets textuels, à leur groupement et à leur mise dans un certain ordre. Le balisage trace dans le matériau textuel se déroulant linéairement les frontières des unités, permettant ainsi leur identification et leur différenciation. Les relations syntaxiques et la ponctuation, qui mettent en évidence des unités infra-phrastiques, au-delà de la phrase se voient relayées par la présence de marques discursives (sortes de « balises ») dotées de capacités d'indexation et d'articulation. Le groupement assure l'équivalence des objets réunis au sein de la série. Satisfaisant un ou plusieurs critères communs, ces objets sont considérés comme étant de même nature ou de même importance. L'action d'ordonner, quant à elle, introduit une relation d'ordre entre les objets de la série, selon une ou plusieurs variables (enchaînement temporel, importance, fonction…). On constate que les marques d'intégration linéaire, indicatrices privilégiées de la sériation, peuvent assurer, selon leur nature et leur configuration précises, chacune de ces trois opérations. Mais, quelle que soit leur orientation dominante, elles s'accompagnent toujours d'indices complémentaires, contribuant à des degrés divers à l'expression des articulations à l'œuvre dans des séries. On peut donc dire que des faisceaux de marques linguistiques de nature différente concourent ensemble au signalement d'une structure en série, dont les variantes (cf. le paragraphe suivant) résultent de la manière particulière dont se composent les opérations de balisage, de groupement et de mise en ordre des items.

Dans (Jackiewicz 2002)[2], nous avons montré que les séries textuelles balisées par des marques discursives étaient: (i) typiquement constituées de 2 ou de 3 items (au maximum 10 items attestés dans les corpus), qui matériellement peuvent aller du syntagme à une suite de paragraphes (articulant ainsi des documents entiers) ; (ii) structurées à un niveau, plus rarement à deux niveaux (emboîtement); (iii) précédées par une amorce explicite ou reconstruite *a posteriori* ; (iv) parfois clôturées par une rétro-évaluation. Sur le plan de la signalisation discursive, (i) les items sont généralement introduits par des marqueurs

[2] Ce premier travail a été mené sur deux ensembles textuels : (i) Le Monde diplomatique sur cédé-rom (1984-1998) ; (ii) la base Frantext.

d'intégration linéaire (MIL) hétérogènes[3], dont certaines configurations peuvent être « incomplètes » ou partiellement relayés par d'autres marques (connecteurs argumentatifs…) ; (ii) l'amorce explicite contient généralement un classifieur (souvent présent également dans certains introducteurs des items) et un quantifieur, qui assurent d'une manière complémentaire aux MIL le signalement des items (leur paradigme commun, leur nombre) ; la rétro-évaluation, quand elle est présente, est introduite par des marques de reformulation (*en bref, somme toute, au fond…*).

Ce résumé des caractéristiques saillantes des séries linéaires dans le discours, qui de par sa forme fortement modale donne l'impression d'une variabilité difficile à dominer, réfère dans les faits à une somme de données relatives, d'une part à la morphologie de la structure en série, et d'autre part aux différentes possibilités de la signaler. Le signalement discursif, qui en premier lieu tient compte de la morphologie (plus une série est longue et/ou plus ses items sont de taille importante, plus le signalement est régulier), semble être également soumis à l'influence d'autres facteurs (genre, contexte social…), dont il est toutefois difficile d'évaluer les dépendances réciproques possibles. L'impact du genre textuel, qui peut s'exercer nettement dans des cas bien précis (certains textes littéraires, certains textes techniques) tant sur la morphologie que sur le signalement, nous semble globalement moins prégnant que les qualités dialogiques de l'auteur ayant à cœur de rendre l'architecture de son texte lisible et facilement interprétable. On peut montrer par exemple que des journalistes censés respecter les mêmes règles éditoriales et écrivant des articles de même « type » (horizons-enquête, analyse…), font en matière des articulations textuelles des choix fort différents, privilégiant pour certains une structuration discursive faible, avec recherche d'effets stylistiques grâce à l'emploi de marques « hétérogènes », et pour d'autres, une structuration forte, marquée d'une manière homogène et systématique[4].

Sur le plan méthodologique, le processus d'identification des structures discursives interroge les principes de la méthode d'exploration contextuelle (Jackiewicz et Minel 2003), destinée à formaliser les connaissances linguistiques relatives à ces structures. En effet, cette méthode postule qu'il existe, pour une tâche donnée, un indicateur privilégié (marqueur le plus saillant), qui peut être un élément lexical, grammatical ou typographique. Dans l'implantation informatique issue de (Jackiewicz 2002), c'est l'introducteur du premier item de la série qui a été considéré comme un indicateur. Ce choix, comme nous l'avons constaté lors de l'évaluation, engendre nécessairement des silences. En réalité, aucun élément, pris isolément, ne peut remplir ce rôle d'indicateur. La nécessité de recourir à des complexes de marqueurs linguistiques, vient de la co-existence de plusieurs opérations discursives différentes, qui ensemble engendrent une structure d'un certain type. Une modélisation rendant clairement compte de cette imbrication d'opérations devrait permettre de trouver une solution informatique théoriquement plus adéquate et, sans doute, plus performante. Quant à l'existence de nombreux paradigmes de marques stylistiquement différents, mais logiquement équivalents, nous constatons que les données collectées systématiquement sur notre premier corpus d'analyse révèlent l'ensemble (sinon, une très forte majorité) de procédés formels de constitution de balises[5], permettant en particulier de prédire des séries d'occurrences non rencontrées dans ces corpus, mais réalisables en pratique.

[3] Par exemple : *premièrement* / *en deuxième lieu* / *enfin* ou *tout d'abord / deuxième facteur / en troisième lieu / enfin, dernier élément.*

[4] Ces observations proviennent de l'analyse des textes du journal le *Monde* (édition complète des années 2002 et 2003 sur cédé-rom), l'analyse réalisée par confrontation avec les règles éditoriales de ce journal, consignées dans (Le Monde 2002).

[5] Ce résultat vient d'une double vérification, obtenue (i) par introspection, (ii) par confrontation avec le corpus du Monde, cité ci-dessus.

Principales références

J. Authier-Revuz, *Ces mots qui ne vont pas de soi. Boucles réflexives et non-coïncidences du dire*, t.1 et t.2, Larousse, 1995.

A. Borillo, « Discours ou méta-discours », in DRLAV, Métalangue, métadiscours, métacommunication, n°32, pp. 47-61, 1985.

M. Charolles, *L'encadrement du discours : Univers, champs, domaines et espaces*, Cahier de Recherche Linguistique 6, Université de Nancy2, 1997.

J. Couto, O. Ferret, B. Grau, N. Hernandez, A. Jackiewicz, J.-L. Minel, S. Porhiel, « RÉGAL, un système pour la visualisation sélective de documents », *Revue d'Intelligence Artificielle*, vol. X, n° X/2004, à paraître.

G. Crispino, A. Jackiewicz, J.-L. Minel, « Spécification et implantation informatique d'un langage de description de structures discursives », *TALN 2004*, Fès, Maroc, 2004.

A. Culioli, « La linguistique : de l'empirique au formel », in *Pour une linguistique de l'énonciation, Opérations et représentations* (1990), t.1, 1987.

J.-P. Desclés, « Systèmes d'exploration contextuelle », *Co-texte et calcul du sens*, Claude Guimier (Ed.), Presses Universitaires de Caen, Caen, p. 215-232, 1997.

J.-P. Desclés et Z. Guentchéva, « Enonciateur, locuteur, médiateur », A. Becquelin et Ph. Erikson (éds), *Les rituels du dialogue*, Editions de l'Harmattan, 2000.

M.-.M. de Gaulmyn, « Reformulation et Planification métadiscursives », in Cosnier et Kerbrat-Orecchioni (éds) *Décrire la conversation*, pp. 167-1999, Presses Universitaires de Lyon, 1987.

J. Goody, *La raison graphique. La domestication de la pensée sauvage*. Les Editions de Minuit, 1979.

A. Jackiewicz, « Repérage et délimitation des cadres organisationnels pour la segmentation automatique des textes », *CIFT'02*, Hammamet, Tunisie, p. 95-107, 2002.

A. Jackiewicz, J.-L. Minel, « L'identification des structures discursives engendrées par les cadres organisationnels », *TALN 2003*, Batz sur Mer, juin 2003.

Le Monde (Journal quotidien), « *Le style du Monde »*, Guide rédactionnel, 2002.

C. Luc, et J. Virbel, « Le modèle d'architecture textuelle : fondements et expérimentation ». *Verbum,* 23(1), 103-123, 2001.

G. Mourad, *Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique des citations*, Thèse de doctorat, Paris, Sorbonne, 2001.

M.-P. Péry-Woodley, *Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle*, Mémoire d'habilitation, Université de Toulouse-le Mirail, 2000.

C. Rossari, *Les opérations de reformulation*, Peter Lang, 1997.

E. Souchier, « L'image du texte. Pour une théorie de l'énonciation éditoriale », *Cahiers de Médiologie,* n°6, 1998.

G. Turco et D. Coltier, « Des agents doubles de l'organisation textuelle, les marqueurs d'intégration linéaire », *Pratiques*, n°57, 1988.