CORPUS

Corpus

8 | 2009
Corpus de textes, textes en corpus

Etude des textes en corpus et problèmes d'échelle

Nadine Lucas

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/1690
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 15 novembre 2009
Pagination : 197-220
ISSN : 1638-9808

Référence électronique
Nadine Lucas, « Etude des textes en corpus et problèmes d'échelle », *Corpus* [En ligne], 8 | 2009, mis en ligne le 08 juillet 2010, consulté le 19 avril 2019. URL : http://journals.openedition.org/corpus/1690

Etude des textes en corpus et problèmes d'échelle

Nadine LUCAS
CNRS, GREYC, Université de Caen Basse-Normandie[1]

1. Introduction

Le recours à l'informatique soulève de nouveaux problèmes dans la définition des objets d'étude de la linguistique. Parler du texte ou du discours pose le problème d'un tout élastique dont on ne précise ordinairement ni la taille ni la mesure. Les termes de micro-structure et macro-structure ne font que réitérer la question des unités cette fois produites par l'analyse (Rastier, 2006). On la retrouve aussi sous les appellations grain d'analyse ou granularité (Prince, 1999).

Grize relève que l'hypothèse d'un tout réduit à un élément n'est pas viable si l'on s'intéresse aux relations et à leur schématisation, et propose de définir des configurations sous-tendues par des relations multiples (Grize, 1990 : 73).

> Sauf à de rares exceptions comme « un train peut en cacher un autre », une schématisation n'est pas faite que d'un seul énoncé. Elle ne l'est pas non plus d'une simple succession d'énoncés. C'est un système, une structure diront certains, dont les éléments soutiennent entre eux des relations multiples. Ainsi les énoncés sont organisés en configurations de dimensions variables, lesquelles configurations à leur tour se composent pour constituer un tout. Le tout peut être de la taille de *A la recherche du temps perdu* ou « Bains interdits. Eau polluée ».

Le problème qui se posait alors pour relier la linguistique et l'informatique *via* la logique se pose toujours près de vingt ans plus tard. L'influence de l'informatique semble forte dans les modèles en hiérarchie inclusive. L'hypothèse d'une homothétie

1 Laboratoire mixte CNRS / UCBN / Ensicaen, UMR 6072.

Corpus n°8 « Corpus de textes, textes en corpus »
(2009), 197-220

de structure entre le texte et la proposition affleure dans la formalisation de cette époque (Adam, 1992 : 34).

> [# Texte # [séquence(s) [macro-propositions [proposition(s)]]]]

En linguistique, les relations et les unités abstraites sont mises en avant sans que les fenêtres d'observation soient toujours explicitées, laissant à l'observateur le soin de trouver une bonne résolution. Ainsi Fløttum définit un objet théorique, le genre des discours, une unité textuelle construite, le passage polyphonique (PP), des points de vue (pdv) également construits et une démarche ascendante à partir de relations (Fløttum, 2002).

> Le PP est une unité non typée, qui peut correspondre à une *phrase simple ou complexe* dans certains cas et à un *ensemble d'un nombre limité de phrases* dans d'autres cas.

Coutinho, pour sa part, tient compte de la disposition observable et place la compositionnalité du côté de la mise en texte dans un genre (Coutinho, 2004 : 36-37).

> Parmi les composantes du genre, […] la [compositionnalité] impliquera l'identification d'unités et de sub-unités textuelles (unités macrostructurelles de différents niveaux, unités aux formats séquentiels plus ou moins prototypiques, segments subordonnés ou autonomes, aux contours morpho-syntaxiques) et la *disposition*, plus ou moins (im)prévisible, de ces unités – l'auto-suffisance d'une unité, quand c'est le cas, le type d'enchaînement qui unit plusieurs unités […], des places fixes ou prévues attribuées à des unités, dans le texte ou le paratexte, des rôles métatextuel et/ou rhétorique que ces unités jouent, dans le développement textuel.

En informatique, les structures ne sont pas élastiques, et, qui plus est, l'habitude a été prise de traiter des termes de la relation plutôt que les relations et de les réifier. L'étiquetage a souvent pris le pas sur le calcul. Nous nous plaçons dans la branche des calculs dits endogènes, qui imposent de définir des espaces de

recherche pour mener des opérations d'algorithmique du texte (Lucas, 2009).

Notre pratique s'appuie sur un corpus multilingue et multigenre. Les textes étudiés manuellement au cours des années sont très nombreux, autour de 2 000, nous ne pourrons en présenter que quelques uns, en français, dans le sous-corpus des écrits de vulgarisation scientifique et sans aller jusqu'aux ouvrages faute de place.

Nous arguons avec Coutinho que, pour être épistémologiquement fondées et utiles, les méthodes doivent proposer une définition distincte des observables – comme données d'entrée – et des unités relatives au cadre théorique d'interprétation mobilisé – comme construits en sortie. Par analogie avec la physique, nous appelons « référentiel linguistique » le cadre théorique muni d'opérations de mise en relation. Pour poursuivre l'analogie, nous posons qu'un cadre théorique est intéressant pour le texte quand il définit ces opérations en quelque sorte « en adimensionnel », mais en précisant des ordres de grandeur et pour chacun, les mesures utiles.

La comparaison avec la physique est suscitée par les exigences de l'informatique linguistique. Dans la dernière décennie, les progrès accomplis ont permis d'objectiver les relations traitées et de calculer sur les relations. Il faut donc définir des tokens ou unités de compte, des unités de traitement, des unités qui peuvent servir d'indices, et mesurer, comparer, ordonner. Le problème qui reste entier, cependant est le passage d'un ordre de grandeur à un autre.

2. Relations multiples

Pour poser quelques jalons et préciser le vocabulaire, nous nous appuyons sur la méthode distributionnelle (Harris, 1952 ; Hockett, 1958). Une sélection est un espace de recherche et en particulier un sous-espace, une position définie comme début, milieu ou fin, dans un espace de recherche donné. Autrement dit, il faut préciser l'unité typographique examinée (début de quoi ?) et la portée de la marque (quel est l'empan du début ? de la fin ?).

Du point de vue informatique, ces méthodes sont bien gérées (*inter alia* Déjean, 1998). Rien n'empêche de travailler

de la même façon pour tous les segments typo-dispositionnels du texte. On manie en fait un objet qui est caractérisé à la fois par sa position et sa forme. Admettons la proposition de Pike : un tagmème est une unité théorique liant forme et position, servant de marque dans un texte (Pike, 1958). Pour certains, le tagmème est une sélection de taille fixe. Pour d'autres, dont nous sommes, le tagmème est proportionnel à la longueur et à la densité du texte traité.

Concernant le texte, le travail fondateur de Fahnestock consiste à montrer comment les figures de style répertoriées par la rhétorique sont nécessaires à la construction du sens et donc naturellement utilisées de nos jours encore. Leur étude à l'échelle des articles ou des traités, dans le genre académique, montre ainsi l'indissociabilité du fond et de la forme en rhétorique (Fahnestock, 1999).

Pour illustrer le problème sans plus attendre, voici un exemple, où nous avons respectivement encadré le début et la fin du texte, surligné le début du corps de texte (le chapeau de l'article) et souligné le début des paragraphes. Notons du reste que cet exemple, qui est assez court, fait partie d'un dossier de presse accessible en ligne. Nous avons numéroté les sections et les paragraphes. L'exemple est destiné à montrer que tout tient à ce qu'on reconnaît comme début ou fin, en somme la question de la mesure adéquate qui capte le sens, ou un phénomène interprétable.

Exemple 1. FOV 173 Article annoté en fonction des positions liminaires utiles

Poséidon, dieu ébranleur et fondateur en terre d'Athènes

§1 Poséidon dont l'image est presque toujours celle du « dieu de la mer » est en fait un dieu très actif sur la terre athénienne. C'est le thème central de la thèse que Sonia Darthou, membre du Centre Louis Gernet de recherches comparées sur les sociétés anciennes, a consacrée aux mythes de fondation d'Athènes à travers la figure de Poséidon, tout en abordant la facette très destructrice de ce dieu à travers son pouvoir sur les séismes et les raz-de-marée.

S1 **Une image trop maritime**

§2 Poséidon a été souvent relégué du côté d'une nature imprévisible, cantonné dans la violence des tempêtes et des tremblements de terre ou la fougue des chevaux. Cela a quelque peu enfermé ce dieu dans une image qui interdisait de lui donner un quelconque rôle « politique ». À Athènes, il a été particulièrement marginalisé dans les recherches face à la déesse Athéna, souffrant d'une image trop exclusivement maritime. Or, Poséidon s'avère un dieu éminemment actif sur la terre athénienne, car les citoyens se revendiquent autochtones, c'est-à-dire nés de la terre, et font de cette naissance leur axe de fondation.

S2 **L'Ébranleur de la terre**

§3 La Grèce est une région de forte sismicité où les tremblements de terre s'avèrent fréquents, violents et destructeurs. La mythologie relaie cette sismicité du territoire en attribuant à un dieu, Poséidon, la responsabilité de ces fléaux. On l'appelle d'ailleurs l'Ébranleur de la terre, et ses épithètes poétiques ou cultuelles montrent bien la violence de son pouvoir. Elles sont en effet formées sur des verbes qui signifient pousser, remuer, vibrer ou secouer : dans les mythes, Poséidon frappe le sol, fouette la terre, bouge les fondations, fait s'écrouler les murailles, les palais ou les maisons, crée des failles et des raz-de-marée qui peuvent engloutir des villes entières parfois. La terre qui doit pourtant être stable pour porter les cités des hommes devient alors un jouet entre les mains de Poséidon : on dit d'ailleurs qu'elle danse sous les coups de son trident vengeur. Les hommes ont même peur que la terre n'éclate dans les airs.

S3 **Le Teneur de fondations**

§4 Poséidon fait donc l'objet d'un culte pour apaiser sa colère et l'amener à révéler son versant plus positif : et c'est alors également à lui que les citoyens s'adressent pour garantir les bonnes fondations de leurs cités. Dans ce cas, ils préfèrent l'appeler le Stable ou le Teneur de fondations.

Ce résultat met en valeur une répétition d'un nom qui correspond à un procédé stylistique commun dans les langues latines (l'anaphore poétique). D'un point de vue informatique, il faut préciser dans quel espace on recherche une répétition, donc plus systématiquement délimiter les débuts et les fins en leur donnant une mesure. Nous avons respectivement encadré le début et la fin du texte, surligné le début et la fin du corps de

texte et souligné le début et la fin des paragraphes, puis évidé les milieux remplacés par […].

Exemple 2. FOV 173 Article annoté en fonction des positions liminaires

Poséidon, dieu ébranleur et fondateur en terre d'Athènes

> **Poséidon dont l'image est presque toujours celle du « dieu de la mer » est en fait un dieu très actif sur la terre athénienne. C'est le thème central de la thèse que Sonia Darthou, membre du Centre Louis Gernet de recherches comparées sur les sociétés anciennes, a consacrée aux mythes de fondation d'Athènes à travers la figure de Poséidon, tout en abordant la facette très destructrice de ce dieu à travers son pouvoir sur les séismes et les raz-de-marée.**

Une image trop maritime

> Poséidon a été souvent relégué du côté d'une nature imprévisible, cantonné dans la violence des tempêtes et des tremblements de terre ou la fougue des chevaux. […] Or, Poséidon s'avère un dieu éminemment actif sur la terre athénienne, car les citoyens se revendiquent autochtones, c'est-à-dire nés de la terre, et font de cette naissance leur axe de fondation.

L'Ébranleur de la terre

> La Grèce est une région de forte sismicité où les tremblements de terre s'avèrent fréquents, violents et destructeurs. . […]. Les hommes ont même peur que la terre n'éclate dans les airs.

Le Teneur de fondations

> Poséidon fait donc l'objet d'un culte pour apaiser sa colère et l'amener à révéler son versant plus positif : et c'est alors également à lui que les citoyens s'adressent pour garantir les bonnes fondations de leurs cités. Dans ce cas, ils préfèrent l'appeler le Stable ou le Teneur de fondations.

Le choix présenté ici est assez fruste : puisque l'on part de l'observable typographié, le début du corps est un paragraphe et le début d'un paragraphe est une phrase. De plus, et nous y reviendrons, nous n'avons pas considéré que les intertitres étaient des paragraphes.

On a bien sûr déjà trouvé des positions intéressantes en rhétorique, des points d'ancrage qui permettent de relier la disposition observable à la modélisation du discours. Le choix de la mesure est souvent implicite en linguistique. Comme les enseignements de la rhétorique sont en grande partie perdus, les phénomènes observables sont souvent reliés directement à des interprétations. Ils sont également nommés par des termes comme *anaphore*, *répétition*, *co-référence*, mais rarement par l'arsenal des termes techniques décrivant des configurations différentes de répétition (Silva rhetoricae ; Fahnestock, 1999). Ainsi, on souligne spontanément les mots ou groupes de mots qui se répètent, ou des membres de phrase de même structure (interrogative ou citative par exemple).

Du point de vue informatique, il faut cependant toujours définir des espaces de recherche par des critères simples et qui s'appliquent en toutes circonstances, sauf à sombrer dans des difficultés énormes (Déjean, 1998). Ainsi, dans notre exemple, on définit les liminaires, en début ou en fin d'une unité typodispositionnelle en tant que « sélections de la marque » ou espaces de recherche en position remarquable, espace dans lequel on trouve ou non des formes à comportement remarquable, notamment des formes répétées. On peut alors extraire les tagmèmes répétés et les nommer selon qu'ils sont observables en début d'unité (anaphore dite aussi anaphore poétique) ou en fin (épiphore). On peut chercher les mots répétés dans un espace ponctué, nommé virgulot pour faire court.

Exemple 3. FOV173 Article analysé par les virgulots liminaires Anaphore (soulignement) épiphore (double soulignement)

Poséidon, dieu ébranleur et fondateur en terre d'Athènes

Poséidon dont l'image est presque toujours celle du « dieu de la mer » est en fait un dieu très actif sur la terre athénienne.

Poséidon a été souvent relégué du côté d'une nature imprévisible, […] c'est-à-dire nés de la terre, et font de cette naissance leur axe de fondation.

Poséidon fait donc l'objet d'un culte pour apaiser sa colère et l'amener à révéler son versant plus positif : […] ils préfèrent l'appeler le Stable ou le Teneur de fondations.

Quoique l'usage s'en soit perdu, il est également utile de différencier des niveaux de granularité ; l'anastrophe et l'épistrophe correspondent à la mesure paragraphe, tandis que l'anaphore et l'épiphore concernent des mots en début et fin de phrase. Autrement dit : l'anaphore et l'épiphore sont détectées, en admettant que nous explorons techniquement les premiers espaces ponctués des phrases dans chaque paragraphe ; mais il est vrai aussi qu'il existe une répétition partielle entre le premier et le dernier paragraphe du corps d'article (en gris) et le titre de cet article. Ce type de répétition partielle est une forme d'accord isomorphe.

Si nous adoptons la notion d'épistrophe en l'associant à l'espace de recherche de la phrase, pour la recherche de répétitions dans notre exemple, nous trouverons non pas une répétition étendue mais un mot répété, un cinquième *Poséidon* en position liminaire.

Exemple 4. FOV 173 marques d'accords de phrases liminaires

Poséidon, dieu ébranleur et fondateur en terre d'Athènes

> Poséidon dont l'image est presque toujours celle du « dieu de la mer » est en fait un dieu très actif sur la terre athénienne.
>
> Poséidon a été souvent relégué du côté d'une nature imprévisible, cantonné dans la violence des tempêtes et des tremblements de terre ou la fougue des chevaux. […] Or, Poséidon s'avère un dieu éminemment actif sur la terre athénienne, car les citoyens se revendiquent autochtones, c'est-à-dire nés de la terre, et font de cette naissance leur axe de fondation.
>
> Poséidon fait donc l'objet d'un culte pour apaiser sa colère et l'amener à révéler son versant plus positif : et c'est alors également à lui que les citoyens s'adressent pour garantir les bonnes fondations de leurs cités. Dans ce cas, ils préfèrent l'appeler le Stable ou le Teneur de fondations.

Enfin, si l'on observe les relations répétées, le bornage *Poséidon fondation* est visible deux fois, au niveau du paragraphe (§2) et au niveau du corps d'article. Notre but n'est pas de citer tous les termes de rhétorique possibles, correspondant à ces configurations, mais simplement de

montrer que ce qui est vu à un certain grain ne l'est pas au grain du dessous ou du dessus.

Une distinction peut s'opérer aussi entre points de vue, grammaire du discours ou stylistique. On comprend que si l'on travaille dans une langue donnée et sur des collections importantes de textes, les formes remarquables que l'on repère par leur fréquence à une position remarquable dans une langue donnée sont de préférence des connecteurs de discours ou autres mots-outils repérables à l'échelle des mots. Mais que les connecteurs de haut niveau sont moins nombreux, puisque les unités de haut niveau sont peu nombreuses (Lucas et Crémilleux, 2004). Pour capter des phénomènes, il est plus utile de noter des traits, des types de phrase, à l'échelle des phrases et des constructions à l'échelle de l'article.

Pour aller un peu plus loin, nous présentons une réflexion sur l'analyse de texte ou de discours à travers l'exemple de l'exposition didactique, qui peut s'envisager à différents grains d'analyse et à travers des théories différentes. Nous posons la question de la taille du texte, de son style collectif à travers sa disposition et ses marques. Nous interrogeons la pertinence du modèle en relation avec la notion d'échelle ou de résolution.

3. Relations situées en contexte

L'analyse syntaxique pose par commodité ou convention que la phrase est une unité de sens, dans laquelle signifiant et signifié coïncident. Pour illustrer cette coïncidence, on a recours à des exemples de grammaire dont la phrase est la fenêtre d'observation. La proposition est le « référentiel », une unité construite, qui permet d'interpréter les relations tant syntaxiques que sémantiques. Puisque la phrase peut se constituer d'une seule proposition, les ensembles se confondent ainsi aisément dans les exemples canoniques.

L'unité observable « stylistique » ou prosodique de la phrase peut se confondre avec la proposition, mais les frontières se confondent aussi récursivement, puisque nous pouvons trouver un seul syntagme dans une phrase et un seul mot dans un syntagme, au final donc un seul mot dans une phrase. Les conditions posées

par les théoriciens sur le syntagme (l'existence d'une *relation* dite syntagmatique) et sur la proposition (l'existence d'une *relation* constitutive, dite prédicative, sujet-verbe dans notre tradition) sont à prendre en compte comme limites à l'interprétation d'une phrase (Gary-Prieur, 1985).

La même difficulté sur les observables concerne les autres unités typographiques du texte. L'unité d'observation immédiate, paragraphe, section, partie, chapitre, peut être formalisée, analysée, mais l'attribution de la qualité prédicative, ou de toute autre qualité fonctionnelle en système tient à des relations qui doivent être définies (Saumjan, 1971).

Nous tenterons de définir les unités du texte, observables et construites, en précisant les conditions auxquelles les relations restent interprétables. Les articles utilisés comme exemples sont pour la plupart accessibles en ligne.

Les propriétés du modèle référentiel

Quoique le problème se pose pour tous les modèles, nous en choisissons un seul, pour montrer en quoi le référentiel influe sur la perception et la pertinence des liens reconnus à différents grains ou ordre de grandeur. Soit un modèle de référence thème rhème, pour lequel on précise que le thème est en facteur pour les constituants du rhème (Sgall *et al.*, 1995 ; Garnier, 2001 ; Lucas & Giguet, 2005). L'hypothèse est que la relation thème rhème est constitutive de la construction. Cette relation est sémantiquement glosée par le postulat « le rhème développe le thème » et une relation d'anaphore ou de co-référence est établie du rhème vers le thème. La construction est asymétrique. Ces notions ne stipulent rien quant aux tokens d'entrée (les opérandes) ni quant aux marques.

Plus exactement, Sgall *et al.* synthétisent les observations du nouveau cercle de Prague relatives au paragraphe avec un thème positionnellement marqué en tchèque. Garnier synthétise du point de vue thème et rhème les théories japonaises sur la phrase, morphologiquement marquée. Nous aurions pu multiplier les références aux constructions de type thème-rhème dans les écrits didactiques (Jones, 1977 ; van

Dijk, 1992). Nous avons tenté de généraliser les opérations pour pouvoir faire des analyses quelle que soit la longueur du texte.

Prenons maintenant quelques articles de vulgarisation scientifique, dont nous voulons faire l'analyse. Quel est le bon niveau d'observation ? Il faut d'abord préciser si le titre rentre ou non dans la fenêtre d'observation, quelles sont les unités manipulées (Ho-Dac, 2007). Nous admettons que le titre général fait partie de la fenêtre d'observation. Nous ne tenons pas compte des illustrations. Bien sûr, ces postulats biaisent fortement le résultat, et c'est ce que nous voulons montrer. En étant plus explicite, ce choix nous écarte du style individuel et de l'énonciation pour nous permettre de généraliser en langue, de construire un paradigme par idiome.

Quels sont les critères qui peuvent être retenus ? L'interprétation d'un article de presse fait fréquemment appel à des notions telles que titre, chapeau, corps de texte. Ces constituants, commodes dans le genre journalistique, renvoient à la disposition observable ; mais si l'interprétation est immédiate, informatiquement les segments sont mesurables en plusieurs « mesures » assimilables aux paragraphes ou aux phrases ou aux mots.

Dans une perspective différentielle ou structurale, la construction ordonnée thème-rhème est caractérisée par des traits singuliers (qui opposent le thème au rhème) et des traits communs (qui assurent la cohérence). Pour construire un tagmème différentiel pour le thème, on cherche une forme remarquable liée à la position initiale du texte. Cette marque est discriminante car elle est absente dans le rhème. Pour l'exercice ici mené, nous cherchons s'il existe une forme remarquable en position remarquable, un tagmème hapax, qui discrimine effectivement le segment thème du segment rhème ; et d'autre part, des tagmèmes répétés permettant d'établir un accord entre thème et rhème dans une construction.

4. Paramètres de variation liés à la taille du texte

Les études de discours se contentent en général de textes courts, souvent d'articles, sans pousser jusqu'aux ouvrages, faute de place. Nous prendrons ici avantage de l'accès en ligne du

matériau pour envisager des articles assez longs et un dossier de presse de 13 pages. Nous présenterons aussi un opuscule sur les marées de 32 pages et un second dossier de presse d'une centaine de pages sur les séismes et tsunamis. Pour limiter les paramètres de variation, nous avons choisi une seule langue, le français et un même genre (ou style collectif), la vulgarisation scientifique. Les documents dont deux sont issus du Cnrs sont comparables en registre, mais les auteurs individuels sont différents.

Ordre dans une construction simple

Commençons par l'ordre de grandeur des articles. Soient trois textes expositifs de taille différente. Lorsque l'ordre de grandeur est précisé, le même modèle relationnel ou référentiel peut être utilisé.

Exemple 6. FOV 192e1 Article court *Journal du cnrs*

LE LITTORAL SOUS BONNE GARDE

> Fleuron du modèle français de développement durable, le Conservatoire du littoral fut créé en 1975 avec pour mission d'acquérir les espaces naturels fragiles ou menacés. « *Puis sa politique a évolué vers la conservation des sites patrimoniaux, pour finalement se concentrer sur la gestion de ses 100 000 hectares, soit 400 ensembles naturels représentant 880 kilomètres de rivages maritimes* », explique Bernard Kalaora, conseiller scientifique du conservatoire. Aujourd'hui, le conservatoire conseille les collectivités locales, gérantes de 90% des acquisitions. Et trois grands dossiers sont prioritaires : contrôler la fréquentation des lieux, définir des objectifs de sauvegarde de la biodiversité et enfin, se doter d'une stratégie de gestion des effets du changement climatique, comme le recul du trait de côte et la submersion.

Dans la fenêtre d'observation convenable pour cet article, en encadré, la mesure immédiate car donnée par la typographie est celle de la phrase, que l'on peut subdiviser en virgulots (unité ponctuée). En observant les phrases liminaires, on remarque que la dernière phrase est coordonnée. C'est une caractéristique des textes expositifs. Pour un petit texte, *Et* peut donc être un indice important. Au milieu du texte on peut relever une citation directe. La première phrase débute par un

segment nominal apposé (premier virgulot), caractéristique d'un segment thématique.

Article moyen

Dans la fenêtre d'observation de l'article FOV 190 *Le littoral : un espace essentiel à l'humanité*, la mesure immédiate, donnée par la typographie, est celle du paragraphe. Il y a 6 paragraphes, subdivisés en phrases. Les liminaires ne présentent *a priori* pas de formes remarquables. En revanche, on notera que le dernier paragraphe est constitué d'une phrase infinitive, sans verbe conjugué. C'est donc l'article qui porte un ajout, sous forme récapitulative (# indique une coupure de plusieurs paragraphes).

Exemple 7. FOV 190 paragraphes liminaires

> La vision des écosystèmes marins côtiers qui s'offre à nous aujourd'hui est très contrastée : aux images idylliques de ces lieux de rencontre entre terre et mer se heurtent celles d'une réalité qui nous renvoie aux conséquences d'un développement des activités humaines qu'il faut aujourd'hui qualifier de non durable.
> [#]
> Observer, expérimenter, modéliser : un trio fondamental pour prévoir et anticiper, mais aussi pour expliquer et conserver les richesses d'un espace essentiel à notre humanité et en construire une gestion raisonnée et durable.

Article long

L'exemple FOV 192 *Homme Littoral, des relations houleuses* est trop long pour être reproduit, on le trouve à l'url http://www2.cnrs.fr/presse/journal/3491.htm. Il compte un chapeau (1§) + 5 sections titrées (14 §) + 2 encarts + 2 photos.

Dans cet exemple, la mesure immédiate donnée par la typographie est celle de la section, les sections titrées regroupant un ou plusieurs paragraphes. On retrouve l'interrogation rhétorique, dans le chapeau. Mais ce n'est pas une marque hapax, elle apparaît plusieurs fois en cours de texte. Les formes interrogatives ne sont pas ici des formes contrastées, mais bien des formes répétées.

En revanche, il existe bien une forme remarquable hapax et distinctive dans la section thématique, c'est une phrase exclamative et suspensive placée au début du texte.

Exemple 8. FOV 192 1er § chapeau

> Ah, le charme des bains de mer !… Les plaisirs du littoral attirent toujours autant – si ce n'est plus.

Dossier de presse

Il est légitime de considérer qu'un dossier de presse est une sorte de texte polyphonique, le produit d'un acte d'énonciation. L'interdépendance des textes est sensible tant sémantiquement que discursivement. Dans les exemples précédents, quelques anomalies apparaissent, au niveau des articles individuels. Elles sont peu sensibles dans les articles du milieu de dossier mais très caractéristiques dans l'article d'ouverture du dossier (en fonction thématique).

Le document ci-dessous occupe la position initiale dans un dossier qui se mesure en sections. C'est une très courte section de deux §, qui a un rôle d'ouverture thématique et de cataphore. C'est un catalogue au sens technique d'ouverture de discours et au sens de liste.

Exemple 9. FOV 191 Article thématique de « l'enquête » http://www2.cnrs.fr/presse/journal/3464.htm

> **Surpopulation, érosion, pollution, montée des eaux...**
>
> **Alerte sur le littoral**
>
> Cet été, vous ferez peut-être partie des centaines de millions de touristes dans le monde qui se délasseront sur leur plage préférée. Mais remarquerez-vous que le littoral que vous fréquentez est en train de se modifier ? Érosion des côtes, changement climatique, montée des eaux, forte pression des activités de l'homme – tourisme, pêche, industrie –, autant de facteurs qui pèsent sur la morphologie et les écosystèmes côtiers…
>
> Pour *Le journal du CNRS*, économistes, sociologues, géologues, climatologues et biologistes dressent un état des lieux parfois inquiétant de ces territoires fragiles, mais qui n'abritent pas moins de 40% de la population mondiale et 80 % de la biodiversité marine. Mettez-vous dans le bain.

La forme distinctive observable à grain fin est la phrase impérative, liminaire à l'échelle de la section. La répétition d'un élément ou d'une forme est difficilement détectable à l'échelle du dossier si l'on garde pour principe de rejeter les illustrations. Mais si l'on tient compte du document avec ses informations

graphiques, on notera que l'illustration de la section thématique est issue de la dernière section. Il y a donc répétition d'image.

Exemple 10. FOV 192 Fin marquée du segment thématique de l'article *Homme et littoral*

> […] Du coup, le bord de mer montre peu à peu ses faiblesses. Il faut le protéger certes, mais comment jongler avec ces conflits d'usage pour respecter l'environnement côtier ? Les politiques publiques sont-elles adaptées à un développement durable ?
> Les chercheurs du CNRS dressent le constat d'une situation dangereuse pour les côtes. Et font le point sur les solutions mises en place pour les préserver.

La dernière phrase de l'introduction est coordonnée, ce qui marque une fin locale pour ce petit texte, mais par rapport au dossier, on note une double interrogation, suivie d'une double réponse dilatoire. Cette interrogation rhétorique IR va se répéter mais sous des formes différentes, en particulier disjointes.

Exemple 11. FOV 192 Reprise distante de l'interrogation rhétorique et de phrases coordonnées défectives.

> **Les remous du trafic maritime**
>
> Et côté mer ? Car le littoral ne s'arrête pas à sa facette continentale : le transport maritime est un axe incontournable de l'économie de la côte.
> [#]
> […] Un premier exemple dans la baie de Saint-Brieuc, où le gisement de coquilles Saint-Jacques était quasiment épuisé voilà quinze ans : grâce à un plan de gestion contraignant – horaires de pêche, taille, quotas, etc. – associant le comité local des pêches, des scientifiques et l'administration maritime, le gisement a pu retrouver sa productivité. Et l'économie locale repartir.

On observe aussi une répétition de structure, la double question suivie à distance d'une double réponse se retrouve dans la dernière section du dernier article.

Exemple 12. FOV 193 Section finale

> **LES EFFETS DU RÉCHAUFFEMENT**
>
> Et qu'en est-il du réchauffement global ? N'est-il pas susceptible de bouleverser radicalement les si fragiles équilibres littoraux ?

> [#]
> Or si le niveau de l'eau monte trop vite, le corail n'arrivera pas à suivre. Et plus de corail, plus d'atolls »…

Changement d'ordre de grandeur

En se reportant à un dossier de presse plus épais (101 pages), la mesure immédiate car donnée par la typographie et la disposition est celle du chapitre, que l'on peut subdiviser en sections (celles-ci étant à leur tour subdivisées en sous-sections). Le lecteur pourra vérifier deux ou trois phénomènes qui ont probablement dissuadé plus d'un linguiste de faire de la macro-syntaxe. En effet, si l'on veut suivre les formes déjà repérées, il faut abandonner les positions remarquables, limitrophes. Si au contraire on reste attaché méthodologiquement aux positions, on est amené à perdre la manifestation morphologique qui caractérisait les segments de texte assurant un rôle d'ouverture thématique.

Ordre de grandeur ou échelle

Parler d'échelle, disent les géographes, « c'est justement admettre qu'autre chose que la taille change, quand change la taille » (Lévy et Lussault, 2003). Une intéressante discussion sur l'échelle, le niveau, la hiérarchie est proposée dans un article traitant non seulement de la géographie mais des sciences humaines (Verdier, 2004). Elle met en valeur la problématique de la rupture ou de la continuité de la structure perçue à différents niveaux, avec des déformations de représentations parfois. Ce qui nous amène à réfléchir aux ordres de grandeur.

On suppose qu'un ordre de grandeur reste représentable au moyen d'une échelle, avec des échelons, mais que dans cet ordre, la structure reste homothétique. Mais parfois on doit admettre aussi que la valeur des mesures ou des indices n'est plus la même et qu'il faille entrer dans un autre ordre de grandeur. Les relations retenues dans un référentiel à un nouvel ordre de grandeur conservent leur géométrie, les propriétés du modèle relationnel peuvent être retrouvées, mais il faut former de nouveaux tagmèmes, avec de nouvelles mesures, pour interpréter les phénomènes de mise en relation d'unités textuelles en conservant la relation au sens.

L'exemple du dossier *Alerte sur le littoral* montre ainsi l'impossibilité de tenir un modèle récursif pour une analyse en thème et rhème, si l'on admet que l'éditorial du numéro fait partie du dossier, ou annonce le contenu du dossier. C'est pourtant à ce constat que l'on aboutit à partir de critères lexicaux calculés par les méthodes dominantes en informatique linguistique (Labadié & Prince, 2008). On peut d'ailleurs l'admettre en vertu d'un jugement humain.

Mais alors les propriétés du modèle thème et rhème ne sont plus respectées, en particulier parce que l'éditorial est long et qu'il n'est pas marqué de manière cohérente en même temps que contrastée par rapport au reste du dossier. Il n'y a pas d'interrogation rhétorique, ni de citation. La construction thème rhème est en revanche clairement marquée si l'on considère seulement le chapeau et les deux articles longs rassemblés sous la rubrique « Enquête ». L'article introductif court est en facteur pour le reste du texte, qui développe en effet deux explications conçues par deux journalistes; et cela est satisfaisant sémantiquement.

Dans le cas du gros dossier de presse *les bleus de la Terre*, qui compte 70 articles, on voit apparaître des constructions locales à l'intérieur des chapitres. Le travail du rédacteur en chef est en effet de créer une structure lisible à l'intérieur d'un dossier mis en forme par une dizaine de rédacteurs à partir d'une cinquantaine de contributions d'auteurs. Les délinéaments sont exagérés, comme les routes sur les cartes sont représentées de manière exagérée par rapport à la surface réelle des routes.

Ainsi, l'organisation du dossier à partir du sommaire simplifié est plus facile à lire qu'à partir du sommaire détaillé.

Exemple 13. Sommaire simplifié du dossier

```
# Sommaire
# Avant-propos
# Une leçon humaine et scientifique
# Phénomènes et stigmates
# La Terre en observation
# Atlas des risques sismiques
# L'Indonésie, un an après...
# Des récits, des pensées et des hommes
```

Pédagogie et médias
À découvrir

La structure expositive commence par le chapitre 1 *Avant-propos* en facteur pour l'ensemble du rhème et se termine au chapitre 7 avec un titre énumératif coordonné, *Des récits, des pensées et des hommes*.

A l'échelle de ce dossier, il est quasiment impossible de ne pas tenir compte des illustrations. L'avant-propos est discriminé par la présence d'une carte animée. A un grain plus fin, les liminaires de cet article sont saturés par des procédés qui sont repris dans les différents chapitres, l'interrogation et la citation notamment. Le trait du catalogue demeure : des énumérations, des listes.

Exemple 14. FOV 119 Liminaires de l'avant-propos

> Le Déluge, les Dix plaies d'Égypte, les destructions de Jéricho, Sodome et Gomorrhe, l'éruption du Vésuve sur le village de Pompéi… les récits de grandes catastrophes naturelles se font depuis la nuit des temps. Au-delà du mythe présent dans presque toutes les cultures, les colères de la Terre ont toujours jalonné l'histoire de l'humanité.
> [#]
> Si la science peut le plus souvent fournir une explication rationnelle et de plus en plus pointue à ces événements extrêmes en analysant leurs causes physiques, environnementales, géopolitiques, humaines, peut-elle aussi aider la planète à se protéger des catastrophes naturelles, en particulier des tremblements de terre, des conséquences du mouvement des plaques lithosphériques et des tsunamis ? Peut-elle lui permettre de les anticiper, de les éviter ou d'y faire face et d'en atténuer les effets ?
> Au chevet de la planète Terre, les scientifiques de toutes disciplines confondues tirent les leçons des catastrophes d'aujourd'hui pour faire face aux menaces de demain. Et mettent en œuvre des stratégies de prévention et d'information. De l'observation à la gestion du risque, notre planète est sous haute surveillance. Avec toutefois une certitude pour le commun des mortels : la nature reprend souvent le dessus…

Ainsi, il faut changer de fenêtre d'observation pour percevoir dans l'avant-propos assurant la fonction thématique du dossier les traits du catalogue et ceux de la catastrophe en

position initiale. Les deux paragraphes de fin constituent une interrogation rhétorique sur deux paragraphes.

Si l'on se penche sur le sommaire détaillé, on remarque que certains types d'articles comme les interviews sont des formes exploitées pour créér des bornes, des transitions entre aspects du problème. Par exemple, l'accord isomorphe entre le chapitre 6 et le chapitre 7 est manifesté par une séquence semblable formée d'un article sous forme d'interview et d'un article sous forme de commentaire, aux positions respectivement antépénultième et pénultième. La transition est ici faite sous une forme de paire d'articles de fin.

Exemple 15. Organisation des chapitres et accord de fin d'exposé : interview + commentaire

6 L'Indonésie, un an après...

Une frontière de plaques complexe
Frontière à haut risque
• Le séisme de Sumatra : un nouvel éclairage de la Terre
Sumatra ou le défi lancé aux scientifiques
SAGER, voyage vers l'épicentre du séisme
Le séisme de décembre 2004 passé au crible grâce au GPS
• Détection des tsunamis : une course contre la montre
Les blessures de la grande bleue en 3D
2,5 à 5 km_3 d'eau par kilomètre de plage ! Jusqu'à 110 mètres d'altitude Des bouleversements écologiques hors normes
« TSUNARISQUE » : de la prévention avant toute chose

Priorité à Cilacap, principal port menacé Expert ès-« coulées de débris »
Animaux et tsunamis : l'échappée belle
Aceh ou la colère de Dieu ?

• Le tsunami : un tournant dans la vie des Sri-Lankais
Reconstruire Vellaveediya
La question de l'eau au Sri Lanka
L'eau source de mort

7 Des récits, des pensées et des hommes

• Poséidon, dieu ébranleur et fondateur en terre d'Athènes
Catastrophe et littérature de colportage
Mémoires des catastrophes naturelles Politique et catastrophe
La « mort collective » sous la loupe de la sociologie

Catastrophe « asiatique » : à qui la faute ?
Internet et SMS au secours des catastrophes
De Lisbonne à Sumatra, qu'avons-nous appris ?
La France et les risques naturels : peut mieux faire
• L'humanitaire expert : des victimes plus abstraites

Alors qu'à l'échelle d'un article, « Poséidon », nous avions ignoré les intertitres, nous nous fions maintenant au type d'article, à l'échelle du chapitre, et à la forme des titres, à l'échelle du dossier. C'est ainsi que l'on peut retrouver une bonne résolution pour un texte didactique d'une page, d'une trentaine de pages ou d'une centaine de pages, à la condition de se départir des critères exploités pour des tokens plus petits. De nouveaux traits sont nécessaires pour effectuer de nouveaux regroupements.

Il faut noter que l'espace de recherche est plus étendu pour les gros documents, ce qui permet de rechercher des marques à partir de patrons proportionnels au segment marqué, mais aussi des changements d'ordre local à l'intérieur des structures. Les marques sont morphologiques ou syntaxiques (positionnelles).

5. La résolution optimale

La résolution optimale par modèle

Tout modèle est présenté à l'aide d'exemples, qui donnent une bonne résolution. Le modèle de proposition est illustré par des exemples convaincants, mais il ne convient pas à toutes les phrases. De la même manière, Adam propose une typologie des discours en fonction des types de séquences, en prenant soin de montrer des cas difficiles à côté des cas favorables illustrant les critères stipulés.

Les textes présentés ici sont des exemples positifs. Mais on voit bien que la délimitation du corps de texte est toujours faite en fonction du modèle que le linguiste a en tête, en l'occurrence celle de la construction expositive. Par exemple, on ignore la préface. Avec un modèle de l'énonciation, la préface aurait un autre statut et serait traitée comme indice important de la relation au lecteur.

La résolution optimale pour un texte

A contrario, partant d'un texte, ou d'une collection de textes, on peut chercher le référentiel le mieux adapté pour en donner une interprétation convaincante. On ne tient compte alors que les constructions typiques, clairement marquées. On peut être amené à ignorer les niveaux d'analyse intermédiaires, par exemple sauter l'organisation des sections, ignorer les intertitres, pour se focaliser sur celle des paragraphes, comme on l'a fait avec l'exemple 1.

Cette solution heurte les théoriciens, mais elle est raisonnable en pédagogie. Dans le cadre de la modélisation informatique, elle est également nécessaire pour définir les conditions favorables et tenter de déduire les conditions défavorables à l'interprétation, de sorte que les extrapolations dangereuses soient spécifiées. Cela revient à tenir compte du genre par exemple du genre journalistique pour faire un profilage des textes (Habert *et al*., 2000)

Conclusion

Nous avons présenté une réflexion sur l'analyse de texte ou de discours qui peut s'envisager à différents grains d'analyse et s'appuyer sur des théories différentes, plus ou moins exigeantes en termes d'opérations et de cohérence de critères. A travers l'exemple de l'exposé didactique, nous avons tenté de montrer l'incidence de la longueur du texte en choisissant une approche descendante à « tagmème de grain variable ». Les marques sont de portée variable, en relation avec leur distribution.

L'unité d'observation est une unité segmentée d'avance par le scripteur. Dans la mesure où l'on parle de relations constitutives d'une structure, la fenêtre d'observation doit contenir plus d'une unité d'entrée ou « token ». L'unité de sortie est une unité construite, définie par le métalangage du linguiste. Mais si l'on admet la prééminence des relations sur les unités, et la détermination du global sur le local, c'est la dynamique du modèle qui doit être opératoire : elle doit au bout du compte expliquer la structure de l'occurrence étudiée, sans violenter le sens intuitivement perçu. Plutôt que d'unité

linguistique, il semble préférable de parler d'opérations métalinguistiques.

Les modèles pédagogiques ou informatisés qui pourraient être construits à partir de ces référentiels, seraient idéalement des modèles linguistiques munis d'opérations et de grandeurs caractéristiques. Ils seraient plus satisfaisants s'ils permettaient de gérer l'échelle, et donc s'ils ne font pas l'économie de la mesure et de l'ordre de grandeur.

Références bibliographiques

Adam J-M. (1992). *Les textes: types et prototypes – récit, description, argumentation et dialogue*. Paris, Nathan.

Coutinho A. (2004). « Schématisation (discursive) et disposition (textuelle) », in *Textes et discours : catégories pour l'analyse.* J.-M. Adam, J.-B. Grize & M. Ali Bouacha (éds.). Dijon : Editions universitaires de Dijon, pp. 29-42.

Déjean H. (1998). « Inférence automatique de contextes distributionnels », *5e conférence annuelle sur le Traitement Automatique des Langues Naturelles*, TALN'98.

Fahnestock J. (1999). *Rhetorical Figures in Science*. Oxford / New York : Oxford University Press.

Fløttum K. (2002). « Polyphonie au niveau textuel », *Romansk Forum* 16 (2) : 339-350.

Gary-Prieur M.-N. (1985). De la grammaire à la linguistique: l'étude de la phrase. Paris : Armand Colin.

Grize J.-B. (1990). *Logique et langage.* Paris : Ophrys.

Habert B. *et al.* (2000). « Profilage de textes : cadre de travail et expérience », *in* M. Rajman & J.C. Chappelier (eds), *JADT 2000, 5èmes Journées internationales d'Analyse des Données Textuelles*, vol. 1, pp. 163-170.

Harris Z. (1952). « Discourse analysis », *Language* 28 : 1-30.

Hockett C. F. (1958). *A Course in Modern Linguistics*. New York : MacMillan.

Ho-Dac L. M. (2007). La position initiale dans l'organisation du discours: une exploration en corpus. Thèse Université Toulouse le Mirail, Toulouse.

Jakobson R. (1985). *Verbal Art, Verbal Sign, Verbal Time*, édité par K. Pomorska & S. Rudy. Minneapolis : University of Minnesota Press.

Jones L.K. (1977). *Theme in English expository discourse*. Lake Bluff, Ill. : Jupiter Press.

Labadié A. et Prince V. (2008). « Finding Text Boundaries and Onding Topic Boundaries : Two Different Tasks ? », *6th International Conference on Natural Language Processing* (GoTAL'08). Gothenburg : Springer-Verlag : 260-271.

Lucas N. et Crémilleux B. (2004). « Fouille de textes hiérarchisée, appliquée à la détection de fautes », *Document numérique* 8 (3) : 107-133. En ligne http://www.cairn.be/.

Lucas N. et Giguet E. (2005). « UniTHEM, un exemple de traitement linguistique à couverture multilingue », *Cide 8 Conférence internationale sur le document électronique*. Paris : Europia, pp. 115-132.

Lucas N. (2006). « Le discours rapporté en sciences humaines et son ellipse en sciences exactes », in Lopez-Muñoz J. M., Marnette S. et Rosier L. (éds.), *Dans la jungle du discours rapporté: genres de discours et discours rapporté.* Cadix : Presses de l'Université de Cadix, pp. 205-215.

Lucas N. (2009). « Discourse Processing for Text Mining », in Prince V. et Roche M. (éds.), *Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration*. Hershey, PA, USA : IGI Global (Medical Information Science Reference), pp. 229 -262.

Pike K. (1958). « On tagmemes, née grammemes », *International Journal of American Linguistics* 24 : 273-278.

Prince V. (1999). « Analyse de structures de surface pour la construction d'un modèle automatique d'analyse et de production d'explications », *Interactions et Cognition* 9-10 : 107-146.

Rastier F. (2006). « La structure en question », *Texto!* En ligne http://www.revue-texto.net/

Saumjan S. K. (1971). *Principles of Structural linguistics.* The Hague : Mouton.

Sgall P. (1994). « La linguistique fonctionnelle et structurale de Prague et sa continuation à l'époque de la description formelle », *Cahiers de l'I.L.S.L*. 6 (Fondements de la recherche linguistique: perspectives épistémologiques).

van Dijk T.A. (1986). « News schemata », in Greenbaum S. et Cooper C. (éds.), *Studying writing*. Beverly Hills : Sage.

Verdier N. (2004). « L'échelle dans quelques sciences sociales: Petite histoire d'une absence d'interdisciplinarité », in Orain O. *et al.* (éds.), *Géographie, échelles et temporalités en géographie*. Paris : Centre national d'éducation à distance, pp. 25-56.
http://halshs.archives-ouvertes.fr/halshs00104485/en/.

Corpus

FOV 190, Trousellier M., Le littoral : un espace essentiel à l'humanité (Editorial)
http://www2.cnrs.fr/presse/journal/3490.htm#global
enquête http://www2.cnrs.fr/presse/journal/3464.htm

FOV 191 Grousson M. et Olivier A. (2007). « Alerte sur le littoral », *Journal du Cnrs* 210-211 Juillet-août 2007 http://www2.cnrs.fr/presse/journal/3452.htm

FOV 192 Olivier A., Homme et littoral : les liaisons houleuses *ibidem* http://www2.cnrs.fr/presse/journal/3491.htm

FOV 192e 1, encadré Olivier A., « Le littoral sous bonne garde ». *Ibidem.*

FOV 193.Grousson M., Les mille et une métamorphoses du littoral *ibidem* http://www2.cnrs.fr/presse/journal/3490.htm

FOV 119-189 Cnrs Dossier de presse *Thema* N°8, 4e trimestre 2005 *Les bleus de la Terre : frisson, tremblement, tsunami*
http://www2.cnrs.fr/presse/thema/661.htm

FOV 119 Les bleus de la Terre. Frisson, tremblement, tsunami (Avant-propos) *Ibidem :* 3.

FOV 173 Poséidon, dieu ébranleur et fondateur en terre d'Athènes. *Ibidem :* 82.