



Appréhender dynamiquement les textes à plusieurs niveaux de détail

Pourquoi les titres ?

Nom du porteur :

Christian Jacquemin

Laboratoire de rattachement :

LIMSI-CNRS

Thématique de recherche :

Visualisation dynamique de textes

Noms des partenaires et Laboratoires de rattachements des partenaires :

Lydia-Mai Ho-Dac, Marie-Paule Jacques, Marie-Paule Pery-Woodley, Thomas Porquet et Josette Rebeyrolle, (ERSS), Mustapha Mojahid et Jacques Virbel (IRIT), Brigitte Grau, Michèle Jardino et Jean-Philippe Merienne (LIMSI), Massih-Reza Amini et Patrick Gallinari (LIP6), Thierry Baccino et Bérénice Closson (LPEQ)

La visualisation dynamique de textes ne prend vraiment son sens que si l'on a affaire à des documents de grande taille et que le rapport entre la partie visible d'un document et sa longueur totale est important. L'approche la plus simple dans ce cas est de ne montrer qu'une fenêtre sur le document (dimensionnée par la taille de la partie visible) et de laisser l'utilisateur parcourir d'avant en arrière le document. Cette approche naïve n'est pas satisfaisante car elle n'offre aucune vision synthétique du document et n'exploite pas les capacités représentatives et discursives de certaines parties privilégiées du discours. Si l'on peut isoler des segments textuels qui représentent le contenu de segments inférieurs temporairement masqués ou mal visibles, on peut alors construire des outils d'accès au document qui reposent sur des paradigmes de présentation graphique bien établis en visualisation de l'information tels que la distorsion (la perspective, la loupe, la profondeur de champ, l'affichage à plusieurs échelles, la diminution de la taille des zones périphériques), le zoom et le déplacement.

La fonction première des titres étant d'annoncer un contenu, il était naturel que les études linguistiques se focalisent sur ces éléments. Les études linguistiques des titres forment le socle de ce projet. Elles s'appuient sur des corpus de la taille d'un ouvrage pour que des régularités apparaissent tant sur leur forme que sur leur fonction. Sur ces bases, trois directions de recherches complémentaires ont été développées. La réalisation d'analyseurs pour l'extraction automatique de relations entre titres et corps du texte ou entre titres eux-mêmes pour assister l'analyse linguistique et pour préparer des applications industrielles dans ce domaine pour l'accès aux bibliothèques numériques. La conception d'outils de visualisation de documents offrant des niveaux de détail dans les rendus variant en fonction de l'intérêt du lecteur. Et l'analyse des stratégies visuelles de lecture d'un document ainsi mis en espace pour voir comment un lecteur parcourt un document lorsqu'il peut naviguer entre plusieurs échelles de représentation. Ce projet a donc permis de rassembler des études linguistiques sur les titres et

leur fonction dans l'organisation du discours textuel, des études en informatique sur l'extraction de ces informations par des méthodes d'apprentissage automatique et sur leur visualisation à plusieurs niveaux, et des études en psychologie cognitive sur les stratégies de lecture dans des documents offrant plusieurs niveaux de représentation et des passages entre ces niveaux.

Les nouvelles lectures

Sous la poussée des jeux, les technologies des cartes graphiques ont beaucoup évolué et offrent désormais des techniques d'affichage pour la visualisation dynamique de documents intégrant la transparence, la spatialisation, le flou, le rendu à plusieurs niveaux de détail, l'orientation tridimensionnelle. Ces technologies offrent donc de nouvelles possibilités d'accès et de parcours (par ex. Small, 1990). Celles-ci autorisent des appréhensions évolutives en profondeur et en largeur du texte en fonction du texte lui-même et/ou de son parcours par l'utilisateur-lecteur. Il peut s'agir d'une forme de prise de connaissance efficace dans le cas de documents longs pour lesquels la forme "volumen" (vs "codex") ou la structuration hypertextuelle peuvent ne s'avérer que partiellement satisfaisantes.

Des recherches visant à compléter ou remplacer la présentation textuelle habituelle des documents, ou à prendre en compte des contextes particuliers de production (par ex. les travaux collaboratifs médiés par l'ordinateur CSCW) ont aussi par ailleurs contribué à ouvrir un champ intéressant de réflexions, et de "design" textuel entièrement nouveau (par ex. Tonfoni 1994, Yazdani & Barker 2000, Kirschner P.A et all. 2002, Tufte 1990). D'autres propositions intéressantes peuvent aussi venir du monde de l'art, numérique (par ex. Couchot 2005) ou plus ancien (les expériences visuelles de S. Mallarmé, G. Appolinaire, M. Leiris, M. Butor, D. Roche, ou les recherches typographiques futuristes, constructivistes, psychédélics, popartistes, etc.).

Néanmoins, l'immense majorité des documents, de facture "classique" ou en référence implicite à l'écrit sur papier, présentent des modes de structurations et de manifestations visuelles directement déterminées par ce contexte. Ces aspects ont fait l'objet de travaux diversifiés (Virbel 2002), mais il apparaît que l'approche de la dynamisation de la visualisation de texte se heurte alors à un problème important qui peut être résumé de la manière qui suit.

Le constat de départ est que la face visuelle des textes, i.e. : le mode de présentation de l'information textuelle, comporte pour l'essentiel cinq composantes principales : le contexte technique ou économique de production, l'impact de valeurs culturelles, esthétiques ou expressives, le genre rédactionnel, l'efficacité communicationnelle et la contribution au sens du texte.

Ces cinq composantes entretiennent de multiples dépendances, et sont par le fait intriquées les unes aux autres. Ces dispositifs dans leur ensemble sont associés à une acception statique et inerte du texte en tant qu'objet intangible, mais ils sont si profondément intériorisés dans les pratiques, des auteurs comme des lecteurs (et positivement sanctionnées par une expérience multiséculaire), qu'il paraîtrait aventureux de les ignorer dans la dynamisation de l'affichage, ou même de sous-estimer leur impact dans l'accès aux textes. Il est donc nécessaire de les réanalyser en fonction de leur rôle et de leur économie propres, dans la perspective de leur manipulation et de leur éventuel retraitement à des fins de dynamisation. Si la conception classique du texte ne permet pas d'organiser une dynamique de sa présentation, elle permet en

revanche d'anticiper sur des dynamiques d'utilisation telles que la lecture suivie exhaustive de première prise de connaissance, mais aussi la relecture, le parcours de type "feuilleter", la "superlecture" (dite aussi lecture en diagonale), la consultation rétrospective, la recherche inédite ciblée, etc.

Effets d'annonce

Les considérations sur les nouvelles formes de lecture nous ont conduit à accorder au titre une place privilégiée dans ce projet sur la dynamisation et la variation des niveaux de détails dans la présentation et l'accès aux documents numériques. Dans les documents longs et structurés, les titres ont un rôle triple : ils délimitent des zones de texte, ils fournissent des informations sur la nature ou le contenu des segments ainsi délimités, et ils permettent d'établir des liens entre ces segments.

Les titres de presse, également les titres de tableaux ont fait l'objet de diverses études. En revanche, il existe peu de travaux sur le fonctionnement des titres dans des documents longs possédant une structure hiérarchisée en sections et sous-sections titrées. Nos travaux ont été construits sur des analyses de corpus. Le premier but poursuivi est de définir une grammaire des titres: quelle est la structure d'un titre et quelles sont les régularités observées ? Il s'agit ensuite d'analyser la fonction des titres dans les documents longs, en particulier dans leur annonce du contenu et leur organisation de la rhétorique du texte. On peut également considérer les titres comme un contenu autonome et analyser les liens entre titres en tenant compte de leur niveau dans la hiérarchie du document. Enfin, on peut également étudier les liens entre titre et texte : liens vers l'avant, comment sont repris les éléments du titre (dans le segment titré), mais aussi liens inverses, les annonces dans le texte précédant le titre. Ces dernières études déterminent, entre autres, la pertinence de certains passages du segment titré par rapport à un titre donné.

Reconnaissance et apprentissage automatique des titres

Dans ce projet deux approches complémentaires ont été mises en oeuvre pour que des programmes réalisent des collectes automatiques de titres ou de relations entre titres et contenus textuels. Dans la première approche que nous qualifierons de symbolique descriptive, l'expertise linguistique humaine acquise à l'analyse de corpus documentaires est entrée dans le programme, sous formes de règles ou de patrons de reconnaissance. Les algorithmes s'appuient sur ces données pour retrouver dans d'autres documents des occurrences linguistiques de forme proche de celles fournies. Ils ont permis de réaliser les objectifs suivants:

- Caractérisation formelle des titres (« grammaire des titres ») ;
- Caractérisation fonctionnelle des titres dans les documents longs ;
- Etude des liens entre titres (« la titraille comme texte »), en tenant compte de leur niveau dans la hiérarchie du document ;
- Etude du lien entre titre et texte : reprises d'éléments du titre (dans le segment titré) mais aussi annonces (dans le segment précédant le titre), pour déterminer entre autres la pertinence de certains passages du segment titré par rapport à un titre donné.

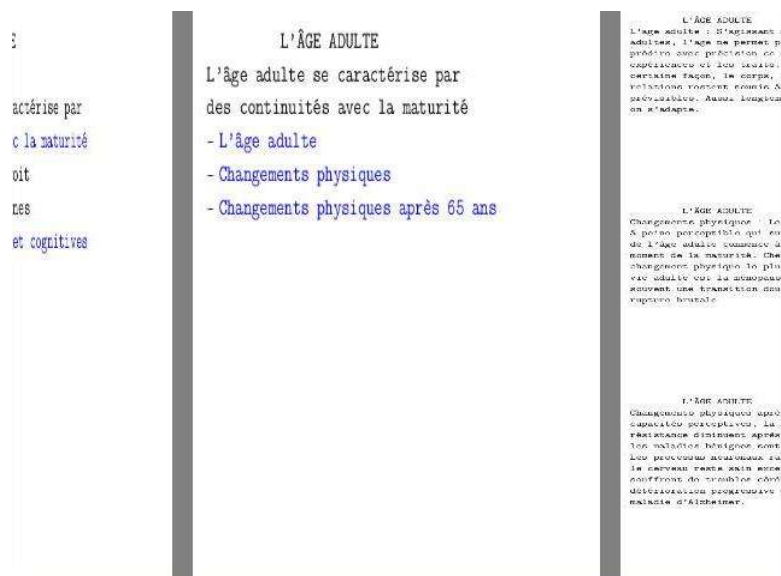
La deuxième approche est basée sur des techniques d'apprentissage pour la segmentation automatique de texte et permet d'attribuer des titres à des passages textuels. Il s'agit donc

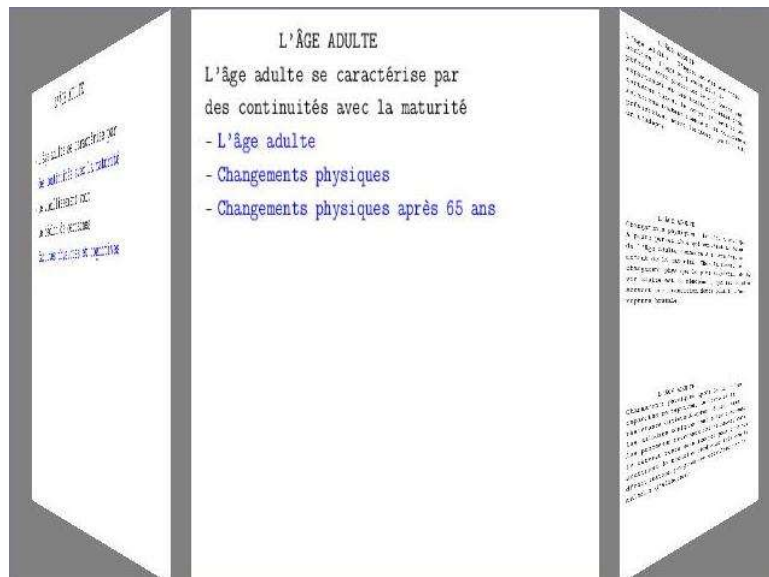
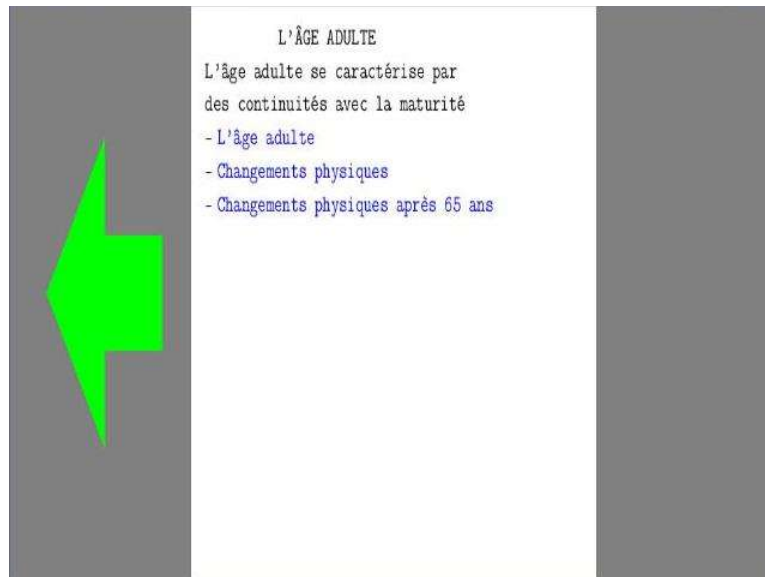
d'apprendre à associer automatiquement des titres à des passages textuels que l'on pourrait proposer automatiquement en recherche d'information. La méthode se fait en deux passes. Le système découvre d'abord différents concepts présents dans un texte, chaque concept étant défini par un ensemble représentatif de mots. Le texte est ensuite segmenté suivant des paragraphes en utilisant une technique de partitionnement basée sur la vraisemblance classifiante. Ce système de segmentation basé sur l'apprentissage dit non-supervisé, donc où il n'est plus nécessaire d'étiqueter manuellement un corpus pour apprendre.

La méthode de segmentation par apprentissage considère le paragraphe comme unité de base, elle comporte trois étapes successives. On apprend tout d'abord les concepts, chacun étant défini comme un ensemble représentatif de mots. Dans une deuxième étape, on caractérise les paragraphes dans l'espace de ces concepts. Cette étape permet de représenter les paragraphes d'une manière concise en réduisant considérablement la dimensionalité du problème par rapport à la représentation sac de mots. On trouve finalement les différentes thématiques présentes dans la collection en regroupant les paragraphes "sémantiquement" proches au sens de ces concepts. Ce modèle par apprentissage automatique fournit donc une collection de segments étiquetés suivant leurs concepts.

Outils de lecture ou jeux vidéo ?

La compréhension d'un document électronique est étroitement liée à la qualité de l'interface visuelle et aux modes de présentation de l'information qu'elle implique. Il est souvent nécessaire de savoir si l'aspect visuel/spatial ou le contenu de l'interface module la prise d'information du lecteur et plus largement la navigation à l'intérieur du document. Afin de tester cela, trois interfaces d'un même texte ont été construites ayant chacune un mode de présentation spatiale différent : une interface simple, une interface plane et une interface 3D.





Ces interfaces ont servi de bases à des expérimentations de psychologie cognitives reposant sur des mesures oculométriques qui repèrent les positions du regard sur un document, ainsi que les saccades oculaires, les déplacements rapides du regard entre deux fixations.

Les 3 interfaces ont une page centrale plane de même dimension qui représente la surface de lecture. L'interface 3D et l'interface plane possède des pages contextuelles (droite et gauche) affichées respectivement en perspective ou à plat. Les pages contextuelles servent à repérer la page lue à l'intérieur du texte. Alors que lire un texte page par page (interface simple) ne nécessite que de suivre linéairement les lignes rédigées par l'auteur, la lecture d'une interface visuelle 3D/plane nécessite la mise en correspondance de la page lue par rapport à l'ensemble du texte. Il s'agit donc pour le lecteur de hiérarchiser l'information traitée et de l'intégrer au contexte global.

Pour un niveau de compréhension équivalent (aucune différence dans le taux d'erreurs aux questions ni dans aucune autre mesure), le nombre moyen de pages lues est moins important pour l'interface 3D que pour les interfaces plane et simple. Cette différence est la même quel que soit le thème du texte. Cela suggère que la mise en contexte des informations lues par une visualisation textuelle en 3D peut aider à l'intégration des informations au cours de la progression dans le document. Cette intégration qui consiste à lier constamment les informations arrivant sous le regard avec le texte déjà lu et même prévu (celui qui n'a pas encore été lu mais annoncé dans la hiérarchie) est facilitée par les pages contextuelles qui situent le paragraphe courant dans l'ensemble du document. En d'autres termes, le lecteur repère plus facilement l'information qu'il lit à l'intérieur du document entier grâce à la présence des pages contextuelles (gauche, droite) ce qui entraîne notamment moins de relectures. L'interface 3D permet de représenter la structure textuelle dans un espace visuel tridimensionnel qui s'applique efficacement à des documents à structure hiérarchique.

Perspectives

Les recherches développées dans ce projet ouvrent sur de nombreuses problématiques nouvelles pluridisciplinaires pour lesquelles la poursuite des coopérations impliquant des acteurs en linguistique, informatique, psychologie et sociologie seront profitables. Ces projets conduiront à une meilleure connaissance des mécanismes cognitifs sur la lecture augmentée sur support électronique, à la réalisation de nouveaux terminaux pour l'accès aux documents qui utiliseront des métaphores graphiques intuitives, et à l'automatisation et au raffinement des outils d'indexation et d'analyse de documents afin de produire des représentations à plusieurs niveaux, etc. Les travaux réalisés ici ont fait déjà quelques pas dans cette direction.

Références

Baccino, T. (2004). La lecture électronique, Presses Universitaires de Grenoble, Coll. Sciences et Technologies de la Connaissance. (254 pages).

Caillet M., Pessiot, J.-F., Amini, M.-R. & Gallinari, P. (2004). Unsupervised Learning with Term Clustering for Thematic Text Segmentation, *Actes de la 7ème Conférence Internationale en Recherche d'Information Assisté par Ordinateur, RIAO*. pp. 1-11.

Jacques, M.-P., Ho-Dac, L.-M. & Rebeyrolle, J. (2004). Quelques aspects méthodologiques d'une étude de la fonction discursive des titres en corpus. *Actes Journée ATALA Modéliser et décrire l'organisation discursive à l'heure du document numérique, Semaine du Document Numérique*, 22 juin 2004, La Rochelle.

Merienne, J.-Ph. et Jacquemin, C., (2003). Large XML Document Manager and Visualizer. In *Proceedings, EUROGRAPHICS 2003*, Granada, Spain.

Virbel J (ed.) (2002) Inscription Spatiale du Langage. *Actes des Journées « Cognitive »*, Toulouse, IRIT, janvier 2002.