ISBN 2-909285-31-6

15, avenue de Ségur,
75007 Paris, France.
Téléphone (Fr) 01 45 51 26 07 - (Int.) +31 1 45 51 26 07
Télecopie (Fr) 01 45 51 26 32- (Int.) +31 1 45 51 26 32
e-mail: production@europia.org
http://europia.org

Document Électronique Dynamique

Le Multilinguisme

Actes du huitième Colloque International sur le Document Électronique : CIDE.8

25 - 28 mai 2005, Beyrouth, Liban

Coordinateur :

Khaldoun ZREIK
Université de Caen

Préface

Depuis 1998, CIDE propose un cycle de manifestations scientifiques pluridisciplinaire sur le thème du document électronique, avec pour objectif de confronter les points de vue des différentes disciplines concernées, et de diffuser les résultats des laboratoires académiques ou industriels qui contribuent à en améliorer les usages.

CIDE'98 a eu lieu à Rabat (Maroc) autour des problèmes posés par l'utilisation du document électronique comme outil spécifique dans le processus de communication écrite (que le document électronique soit enjeu ou simple support de cette communication). La réflexion s'est organisée autour de : la nature hétérogène des objets contenus dans le document, la richesse des éléments structurels et leurs contributions à la description du document et enfin la nature du support physique d'inscription.

CIDE'99, qui a eu lieu à Damas (Syrie) a souhaité que la réflexion soit portée sur la dimension dynamique du nouvel document électronique et plus particulièrement sur le document Web.

CIDE2000, a eu lieu à Lyon (France) pour insister sur la conception et l'utilisation de systèmes d'information documentaire ainsi que sur les aspects dynamiques du document. Les sujets développés ont concerné notamment le cycle de vie du document et ses aspects ergonomiques et perceptifs.

CIDE2001, s'est tenu à Toulouse (France) avec comme objectif d'aborder les dimensions cognitives du document électronique. Dans ce cadre, le document électronique – lieu et support de communication – s'offre comme un terrain privilégié d'investigation, par les sciences cognitives, des différents processus dont le document est l'objet.

En 2002, CIDE5 a eu lieu à Hammamet (Tunisie) avec comme thématique principale la perception et l'exploitation du document électronique dans le cadre de l'éducation et la formation. Un intérêt particulier a été octroyé aux travaux de recherche visant l'exploitation des documents mobiles dans la mise en oeuvre de systèmes éducatifs hybrides.

CIDE.6, a choisi Caen (France, 2003) pour aborder les méthodes traitant de la plurimodalité et du multimédia dans le document électronique. Ces méthodes trouvent leurs sources dans l'ingénierie du document et certains domaines des sciences humaines et de la psychologie. Ces disciplines contribuent par des modèles, des traitements automatiques et des expériences, à l'analyse des différentes modalités d'expression et de leurs relations.

CIDE.7, a été organisé à La Rochelle (France, 2004) avec comme thématique centrale les « approches sémantiques pour le document électronique ». Sur ce thème, des progrès significatifs ont été réalisés au cours des dernières années aussi bien sur le document textuel que sur le document de type hypermédias (extraction d'information et indexation de documents sonores et vidéo par le contenu, résumé d'oeuvres...).

Cet ouvrage rassemble l'ensemble des contributions présentées sur Le Multilinguisme à CIDE.8 du 25 au 28 mai 2005 à Beyrouth (Liban).

CIDE.8 a pour objectif de resserrer les liens entre l'ingénierie documentaire et l'ingénierie linguistique tout en considérant les différentes dimensions des documents électroniques à savoir : cognitive, structurelle et technologique.

Quatre laboratoires universitaires ont collaboré étroitement à l'organisation de cette manifestation : le GREYC (UMR 6072 CNRS – Université de Caen et ENSI Caen), le PPF Modescos (Université de Caen), le LaLICC (Université de Paris IV Sorbonne) et le PSI (Université de Rouen, INSA de Rouen) représentés respectivement par Jacques Madelaine et Khaldoun Zreik (GREYC), Christine Durieux et Jean Vivier (PPF Modescos), Ghassan Mourad (co-président du CIDE.8, LaLICC) et Jacques Labiche (PSI).

CIDE.8 a donné un intérêt particulier aux thèmes relatifs à la question du multilinguisme dans la conception et la perception du document multilingue tels que :

- L'indexation des documents électroniques multilingues,
- Le formatage et multicodage dans les documents électroniques multilingues,
- Les traductions de textes en ligne : prise en compte des exigences du multilinguisme,
- La recherche et extraction d'information dans des documents électroniques multilingues,
- La catégorisation des documents électroniques multilingues,
- L'écriture et la lecture du document multilingues.

Un des objectifs de CIDE.8 est d'aller à la rencontre de nos collègues libanais et proches orientaux, nous avons donc organisé des tutoriels (Patrick Andries et Jacques Ducloy), des ateliers ainsi qu'une session spéciale sur la diffusion d'information scientifique et technique en collaboration avec l'INIST (groupe ARTIST animé par Jacques Ducloy).

CIDE.8 a été organisé dans des circonstances exceptionnelles, aussi le comité d'organisation tient-il à remercier très chaleureusement pour leur support et coopération :

- Les animateurs des tutoriels, des ateliers et surtout les auteurs,
- Les membres des comités de programme et de lecture,
- Le CERTIC - Caen,
- Le laboratoire Paragraphe de l'Université de Paris 8,
- Le laboratoire LaLICC de l'université de Paris IV Sorbonne,
- Le laboratoire GREYC (UMR 6072 CNRS),
- Le RTP Doc (CNRS),
- L'EPFL de Lausanne – Suisse.

Le comité de programme de CIDE.8, souhaite exprimer sa profonde reconnaissance à :

- Lydie Sauvé (GREYC, Caen) qui a assuré avec perfection et ténacité les tâches de secrétariat, de préparation des actes et d'organisation du colloque.
- Arnaud Daret et Christophe Turbout (CERTIC - Caen) qui, après avoir réalisé le système de gestion en ligne du colloque, ont été continuellement à notre écoute pour garantir le bon déroulement du colloque.

Le coordinateur de `CIDE.8`
Khaldoun Zreik

Table des matières

Session 1. Documents multimédia et multisupports

Session 2. Recherche et extraction d'information

Session 3. Multilinguisme et sciences cognitives

Session 4. Catégorisation et indexation

Session 5. Formatage et multicodage

Conférence invitée

♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦

Comité de programme

P. Andries, Consortium Unicode, Canada
C. Balliu, Haute Ecole de Bruxelles, Belgique
G. Bastin, Université de Montréal, Canada
H. Blanchon, CLIPS-CNRS, France
J. Caelen, CLIPS-CNRS, France
J. Ducloy, INIST-CNRS, France
C. Durieux, Université de Caen, France
B. El Eter, Université de Beyrouth, Liban
C. Fluhr, CEA, France
M. Gaio, Université de Pau, France
J. Labiche, Université de Rouen, France
O. Larouk, Université de Rouen, France
J. Madelaine, Université de Caen, France
G. Mourad, Université de Paris IV, France
M. Politis, Université Ionienne de Corfou, Grèce
I. Saleh, Université de Paris VIII, France
G. Serasset, CLIPS-CNRS, France
Y. Toussaint, LORIA, France
C. Turbout, CERTIC, France
C. Vanoirbeek, EPFL, Suisse
J. Vivier, Université de Caen, France
M. Zoater, Université de Beyrouth, Liban
K. Zreik, Université de Caen, France

Comité d'organisation

A. Daret, CERTIC, France
I. Kalaoun, Europia, France
G. Mourad, Université de Paris IV, France
L. Sauvé, Université de Caen, France
C. Turbout, Université de Caen, France
K. Zreik, Université de Caen, France

Session 1

Documents multimédia et multisupports

Base de connaissances multimédia pour le cinéma d'animation

Daniel Beauchêne[1], Françoise Deloule[1], Bogdan Ionescu[2], Patrick Lambert[3]

[1] *Equipe Condillac – LISTIC – Université de Savoie*
Campus scientifique, 73376 Le Bourget du Lac Cedex - France
{daniel.beauchene,francoise.deloule}@univ-savoie.fr

[2] *Université Politehnica - LAPI*
Bucarest -Roumanie
bodgan.ionescu@univ-savoie.fr

[3] *LISTIC – Université de Savoie*
ESIA, BP 806, 74016 Annecy Cedex - France
patrick.lambert@univ-savoie.fr

Résumé :

Dans le domaine particulier du cinéma d'animation, nous proposons de combiner deux approches pour construire une base de connaissance multimédia : l'une, terminologique, s'appuie sur les péri-textes et une ontologie du cinéma d'animation, l'autre est basée sur l'extraction de caractéristiques issues des images. L'objectif de nos travaux est de proposer des outils qui serviront à la construction d'une plate-forme d'analyse des films et de navigation dans la base des films.

Mots-clés : bases de données multimédia, fusion d'information, bases de connaissances terminologiques, indexation vidéo.

1. Introduction

Le travail présenté est le résultat d'une collaboration entre le Laboratoire Informatique, Systèmes de Traitement de l'Information et de la Connaissance (LISTIC) et le Centre International du Cinéma d'Animation (Cica), organisateur du Festival International du Film d'Animation qui a lieu annuellement à Annecy.

Dans cette collaboration, les objectifs du Cica sont :

- Un objectif « patrimonial » de conservation et de mise à disposition de l'ensemble des films disponibles. La constitution d'une base numérisée permet d'obtenir un support de sauvegarde fiable et un accès facile. Des outils informatiques peuvent alors être mis en place pour permettre la recherche ou la navigation dans la base constituée.
- Un objectif d'exploitation : permettre une utilisation et une exploitation nouvelle de cette base par la construction d'outils de caractérisation et d'analyse des films. Cela présente un intérêt fondamental à la fois pour les professionnels, les cinémathèques, les enseignants ou même le grand public.

Après une présentation des données disponibles (paragraphe 2), nous décrivons comment nous modélisons les données terminologiques (paragraphe 3) et les données image (paragraphe 4). Enfin, nous présentons les travaux en cours sur la fusion des informations issues des deux sources.

2. Les sources de connaissances

Les sources de connaissances utilisées dans ce projet sont fournies par le Cica. Elles sont constituées d'une part d'une base de données qui décrit chaque film à travers des champs essentiellement textuels, d'autre part des films eux-mêmes.

2.1 La base de données

Cette base contient des informations sur près de 21 000 films d'animation et 55 000 professionnels du secteur, complétées par des milliers de photographies, des dossiers documentaires, livres et magazines. Elle référence 14 000 films sur divers supports et s'enrichit chaque année à l'occasion du festival de près de 1 500 nouveaux titres. Ces films sont décrits par :

- Le titre de l'œuvre (dans la langue originale, en français et en anglais), la nationalité et l'année de production,
- Les noms des « auteurs » : scénario, graphismes, sons, etc.,
- Des indications sur les techniques utilisées, l'âge du public visé, le genre du film, sa durée, son support,

- Un synopsis en français et en anglais.

Ces données sont assez complètes pour les films qui ont été présentés au festival, beaucoup moins pour d'autres films. D'autre part, les indications sur les techniques utilisées et les synopsis ont été le plus souvent remplies par ceux qui ont inscrits le film au festival, mais aussi parfois par les personnes qui ont saisi les données dans la base. Ainsi, le texte du champ synopsis peut selon les cas, contenir une accroche ou un résumé du film.

De même, la qualité des traductions est variable selon les périodes de saisie. On a donc un corpus dont la qualité n'est pas homogène.

2.2 Fiche descriptive d'un film

Dans l'état actuel de la base, certains éléments des fiches descriptives des films (en se limitant aux films présentés au festival) sont accessibles sur le site du Cica (www.annecy.org) à la rubrique Animaquid.

Une de ces fiches est donnée en exemple ci-dessous.


FICHE FILM

23 rue des Martyrs

Synopsis

Évocation de la mort de Théodore Géricault et de sa passion fatale pour les chevaux.

Identité

Réalisation : Luc Perez
Pays : France
Année : 1997
Durée : 00 h 04 mn 00 s
Technique(s) utilisée(s) : Rotoscopie
Procédé : Couleur
Catégorie : 1997 Courts métrages
Version : Sans dialogue ni commentaire
Public(s) visé(s) : Adultes

Générique

Producteur(s) : Farid Rezkallah

Production : 24 Images - Videogram Productions

Figure 1 : une fiche film sur le site du Cica

2.3 Films numérisés

La base de données contenant les films au format numérique est en cours de constitution. Elle est actuellement composée d'une cinquantaine de films au format *mpeg2*. Les images ont un format de 704x326 pixels, chaque pixel étant codé sur 24 bits. Ces films couvrent une large palette de techniques et de genres.

3. Connaissances terminologiques

Nos objectifs ici sont d'une part de faciliter la mémorisation et l'exploitation de ces bases importantes d'informations en perpétuelle évolution et d'autre part d'utiliser les connaissances terminologiques propres au domaine étudié. Pour cela nous combinerons les approches terminologiques et ontologiques.

Nous voulons construire des terminologies métier consensuelles, cohérentes au sens de la consistance logique de l'ensemble des significations, partageables et réutilisables. C'est en effet au travers des documents produits et utilisés (notamment les fiches films), que le Cica peut expliciter son savoir-faire et son métier. Cependant, différents experts interviennent à différentes périodes, créant une certaine hétérogénéité dans les informations et connaissances exprimées.

De ce fait, nous souhaitons faire évoluer ces bases d'informations vers des bases de connaissances terminologiques (BCT) [MEYE92] constituées :

- Des textes qui sont les sources d'informations qui caractérisent le sens des termes (signification en contexte),
- Des termes, mots ou expressions du langage du métier,
- Des concepts dénotés par ces termes, et qui ont leur propre signification (signification décontextualisée, donc indépendante de l'expert).

Ces concepts, une fois déterminés, seront organisés selon une approche ontologique dans le modèle OK que nous présentons par la suite.

Pour ce qui est de l'approche retenue nous pouvons nous appuyer sur les travaux réalisés dans ce domaine [SLOD95]. Nous retenons trois approches principales de construction : les terminologies textuelles [BOUR99] pour lesquelles les expressions du métier, connaissances et relations conceptuelles se trouvent dans les textes, les terminologies conceptuelles [NF ISO 704 2001] pour lesquelles la définition des concepts et de leurs relations est primordiale, et les terminologies ontologiques pour lesquelles le concept est au centre de l'ontologie [ROCH99], celle-ci reposant à la fois sur les termes du métier et sur les principes épistémologiques liés à la définition de l'ontologie. Les principes sous-jacents du modèle ontologique OK répondant à nos objectifs, nous permettent d'extraire des concepts des bases d'information à notre disposition et de construire une terminologie métier consensuelle, cohérente, réutilisable et partageable.

3.1 Le modèle OK

Dans le cadre d'une terminologie ontologique, la signification d'un mot est définie comme étant la connaissance décontextualisée qu'il dénote (alors que le sens est une signification actualisée en fonction du contexte et de la situation).

Cette connaissance dépend directement de considérations épistémologiques du modèle ontologique, qui n'appartient pas au domaine linguistique. Nous avons retenu le modèle ontologique OK pour Ontological Knowledge [ROCH01] qui repose sur des principes pluridisciplinaires comme l'épistémologie (distinction de l'essence des choses de ce qui les décrit), la linguistique (le signe est arbitraire, la langue est un système qui structure les mots), l'intelligence artificielle (impact du langage de représentation sur la modélisation) et de la logique (sémantique logique et ensembliste des connaissances, formalisation *a posteriori)*.

Le modèle OK repose sur une construction en plusieurs étapes :

La première étape consiste à identifier les **catégories OK**. La notion de catégorie permet de regrouper les termes dont les définitions sont liées. Un domaine est donc décrit par plusieurs catégories qui peuvent ensuite être combinées pour les différentes utilisations du système. Par exemple ici, les techniques, les supports (papier, verre), les outils (pinceaux, crayon, pastel...), les genres, les publics…, forment autant de catégories distinctes.

Dans chaque catégorie, les termes sont alors répartis selon leur nature et permettent de définir des **concepts**, des **différences**, des attributs et de prendre en compte certaines relations conceptuelles. Nous présentons ici ces principales notions.

Le **concept** désigne une connaissance portant sur une pluralité de choses, à l'opposé d'une connaissance singulière. Le concept de départ porte en général le nom de la catégorie. Puis chaque concept est défini par rapport au précédent en terme de *sorte-de*, relation de subsomption, puis d'un « complément » précisant ce qui le différencie du précédent. Par exemple les <ombres chinoises> sont une *sorte de* <découpage>.

Les **différences** constituent les atomes de signification élémentaires à partir desquels se construisent et se divisent les concepts. Les différences que nous prenons en compte portent sur l'**essence** des choses. De ce fait un concept possède ou ne possède pas une caractéristique essentielle, et nous permet de construire une ontologie reposant sur une arborescence, sur laquelle nous pourrons appliquer des éléments de logique (formalisation a posteriori, sémantique logique et ensembliste des connaissances). Les <ombres chinoises> se différencient des <éléments découpés> par le mode d'éclairage.

L'**attribut** permet de décrire l'état du concept et non son essence (par exemple : le matériau permettant de construire les ombres chinoise est un attribut).

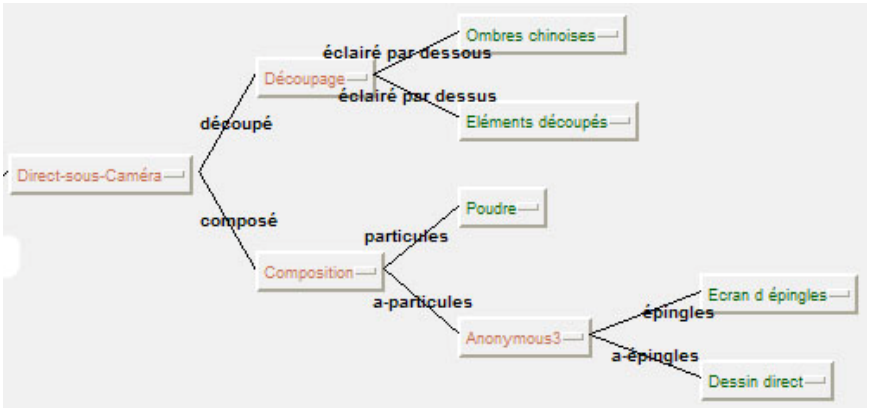


Figure 2 : Extrait de la catégorie des Techniques

3.2 La méthode de construction de la terminologie

Notre base de connaissances a pour but de rassembler et de mettre à disposition les connaissances contenues dans les bases de données textuelles associées aux films. Nous nous appuyons, pour la construire, sur l'ensemble des textes, sur la compréhension et l'utilisation du langage métier et sur la mise en lumière des concepts nécessaires à la mise en œuvre de cette base de connaissances.

La première étape de la construction d'une terminologie consiste à comprendre le métier qui doit être modélisé et à définir les besoins en terme d'application. Puis il faut construire un lexique des termes employés par les différents protagonistes. Pour cela, nous cherchons à identifier un corpus de textes de référence représentatif du métier, avec l'aide d'un expert du domaine.

Sur ce corpus, nous essayons, à l'aide d'outils de sélection et de statistiques, de faire émerger des candidats termes, c'est-à-dire des termes utilisés de façon spécifique dans ce domaine ou métier. Nous disposons d'un outil, Linguistic Craft Workbench (LCW) de la société Ontologos-Corp qui nous permet d'extraire différents éléments du texte, d'un simple substantif à une expression composée et en éliminant les « termes standard » du langage. A ce niveau, nous devons veiller à différencier le vocabulaire en fonction des utilisations et des profils des différents acteurs (intervenant du Cica, étudiants, préparation de festival, festivalier, etc.).

A partir d'une liste qui peut être assez importante, nous devons regrouper, avec l'aide d'un expert, ces termes dans une catégorie existante ou créer une nouvelle catégorie. Pour chaque terme, il faut déterminer s'il s'agit d'un concept, d'une différence, d'un attribut ou d'une valeur d'attribut, etc. et le placer dans l'ontologie en construction. Cette construction se fait à l'aide de deux approches combinées.

D'une part, on cherche à mettre en avant des relations de subsomption entre les concepts, qui peuvent dans un premier temps donner lieu à des regroupements de concepts puis à création d'un arbre n-aire. D'autre part, on effectue un travail sur des grilles de différences qui reprennent les caractéristiques essentielles des concepts. On peut alors regarder pour chaque concept si une caractéristique a un sens et dans ce cas, s'il la possède ou non.

Au travers de ces deux approches, on peut alors construire l'arbre de Porphyre, sous jacent au modèle OK. Tout au long de ce travail, la validation des propositions par un expert est bien sûr incontournable.

Dans le cas de l'étude menée au Cica, des retours vers les bases d'informations et parfois vers les films ont été nécessaires pour définir les termes dénotant les concepts et leurs synonymes.

Enfin, l'ontologie construite à partir d'un extrait de la base de film doit être validée par des experts et par la totalité du corpus. Les éléments qui ne peuvent être placés doivent être étudiés, pour insertion d'un synonyme ou création d'un nouveau concept et des différences essentielles associées.

Comme nous l'avons dit au paragraphe 1, le corpus sur lequel nous travaillons présente des défauts (informations manquantes, différences d'interprétation de ce qu'est un synopsis, qualité des traductions, etc.), la présence des experts et la possibilités de retour au film lui-même a permis, malgré ces défauts, de construire une terminologie consensuelle, voire d'améliorer la qualité de la base de données. En effet, si les candidats-termes extraits des données textuelles apportent une aide à la construction de la terminologie-métier, celle-ci provient en réalité de l'interaction entre experts et candidats-termes.

4. Connaissances image

Dans le domaine de l'analyse des images, la thématique appelée « indexation » mobilise fortement les chercheurs du domaine depuis bientôt dix ans [DELB99] [SMEU01] [SANT01]. L'objectif est d'arriver à extraire de chaque image un ensemble de descripteurs (ou index) permettant la caractérisation de l'image. Ces descripteurs sont ensuite utilisés pour mettre en place des mécanismes de recherche ou de navigation dans de grandes bases d'images. Le passage de l'image fixe à la vidéo est plus récent [LEFE03] [HUET03], et rendu plus complexe par l'augmentation considérable de la masse de données à traiter et par la prise en compte de l'information de mouvement.

Dans notre application, les images constituant un film contiennent une grande quantité d'informations pouvant également contribuer à la construction de la base de connaissance. Cependant, l'exploitation de ces informations présente deux difficultés majeures :

- Les techniques d'analyse fournissent des caractéristiques de « bas niveau », souvent numériques, qu'il est difficile de traduire en descripteurs de nature

sémantique. On parle alors de « fossé sémantique ». Ainsi, on sent bien que la connaissance de la distribution des couleurs utilisées dans un film présente un lien avec l'atmosphère se dégageant de ce film, mais ce lien est difficile à formaliser.

- Les descripteurs issus des images sont en général définis sur une autre base terminologique que celle issue de l'analyse des textes. L'analyse des images pourra fournir des résultats du type : film à dominante rouge, ou présentant une grande variété de couleurs, etc., caractérisations qu'il n'est pas simple de relier au genre « comédie » par exemple.

Les descripteurs image que nous nous proposons de déterminer sont souvent proches de ceux qui sont proposés dans le format MPEG7 [MPEG7]. Cependant la norme MPEG7 se limite à la définition d'une syntaxe de description, mais ne standardise pas la production de ces descriptions. Par contre, la démarche envisagée pourrait facilement s'adaptée à l'exploitation de vidéos au format MPEG7.

4.1 Méthodologie

La méthodologie utilisée se décompose en trois étapes principales :

⇨ La *réduction des couleurs* : dans une image le codage de la couleur est effectué sur plus de 16 millions de niveaux. Une première étape incontournable consiste à réduire la palette à quelques dizaines de couleurs.

⇨ La *segmentation temporelle* : il s'agit de trouver les différents plans composant le film.

⇨ La *caractérisation du film*, bâtie à partir des deux premières étapes. La caractérisation que nous proposons se fait dans trois directions :

- La caractérisation du *rythme* décrivant la répartition des changements de plan dans le film,
- La caractérisation *couleur* donnant la distribution des principales couleurs utilisées dans le film,
- La caractérisation de type « image statique ».

4.2 La réduction des couleurs

La réduction couleur joue un rôle très important pour le reste de l'analyse. L'objectif est de définir une palette réduite suffisamment pertinente pour ne pas perturber le reste de l'analyse. Deux types se solutions sont envisageables :

- Une réduction adaptée à chaque film. Ces solutions sont en général de meilleure qualité, mais restent attachées à chaque film.
- Une réduction fixe, commune à tous les films.

Pour permettre des comparaisons simples entre films, nous avons choisi la deuxième solution. Après différentes études, nous avons sélectionné une palette de

215 couleurs obtenue par diffusion d'erreur [WURD04]. La figure ci-dessous présente la palette utilisée, puis une image sans réduction couleur et après la réduction couleur proposée. Les différences sont à peine perceptibles.


a – Palette réduite


b – Image originale


c – Image en couleurs réduites

Figure 3 : Réduction des couleurs

4.3 La segmentation temporelle

L'analyse d'un film demande la connaissance de son découpage en plans. La technique utilisée repose sur la recherche des changements de plan. Il existe différents types de changements de plan :

- Les changements de plan simples (cuts). Le passage d'un plan à l'autre se fait sur deux images. On utilise une mesure de différence entre images pour détecter ces changements. La mesure de différence employée est une différence entre histogrammes couleur calculés sur la palette réduite. La différence est ensuite comparée à un seuil, calculé automatiquement [IONE04]. Pour ne pas être trop sensible aux déplacements à l'intérieur des images, quatre histogrammes, correspondant aux quatre cadrans de l'image, sont utilisés. Une décision majoritaire est alors prise.

- Les changements de plan par effets spéciaux. Dans les films d'animation, on distingue principalement 3 types d'effets spéciaux :

  ⇨ Les transitions lentes de type *fade*. Le passage d'un plan à l'autre se fait par transformation progressive de l'image vers une image uniforme (noire ou constante) – *fade out* – puis réapparition progressive de l'image suivante par le mécanisme inverse – *fade in* -. Ce type de transition est détecté en recherchant des décroissances (ou des croissances) lentes de la moyenne de l'intensité des images successives.

Figure 4 : Exemple de « Fade in »

⇨ Les changements brefs de couleurs. Spécifiques aux films d'animation, c'est un effet visuel qui se caractérise par un brusque changement de couleur à l'intérieur d'un même plan, sur une durée très courte (quelques images). La figure ci-dessous illustre un tel effet. La recherche de changements de plan simples détecte deux transitions face à ce type d'effet. La similarité des images situées avant et après ces deux transitions proches permet de détecter ces effets.

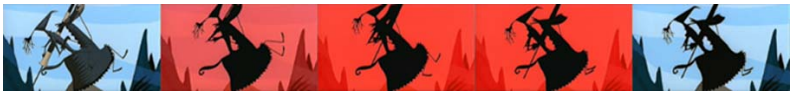


Figure 5 : Exemple de changement bref de couleurs

⇨ Les fondus enchaînés. Ces transitions sont les plus difficiles à détecter. La technique utilisée repose sur l'analyse de l'amplitude des contours [LIEN01]. Un fondu enchaîné se traduit par une décroissance lente de l'amplitude des contours de la dernière image du plan suivi d'une croissance lente de l'amplitude des contours de la première image du plan suivant.

Le tableau suivant donne quelques résultats de détection. Les séquences testées ont été segmentées manuellement pour permettre une mesure des taux de détection. La durée totale des séquences représente plus de 4h de films.

	Nombre de transitions	% non détection	% fausse détection
Cuts	3 166	7.4 %	4.0 %
Fade	103	4.0 %	15.0 %
Chgt bref	125	11.7 %	6.7 %

Tableau 1 : résultats de la détection des changements de plans

4.4 La caractérisation des films

A partir de la connaissance de la segmentation temporelle, il est possible de fournir des caractéristiques descriptives de chaque film. Pour le moment trois pistes sont en cours d'exploration.

- La caractérisation du rythme du film.

  La densité et les positions relatives des transitions fournissent des indicateurs simples donnant une bonne idée du rythme de la séquence. La figure ci-dessous propose une représentation graphique des transitions. En rouge sont figurés les changements de plans et en bleu la segmentation temporelle après élimination des effets spéciaux.

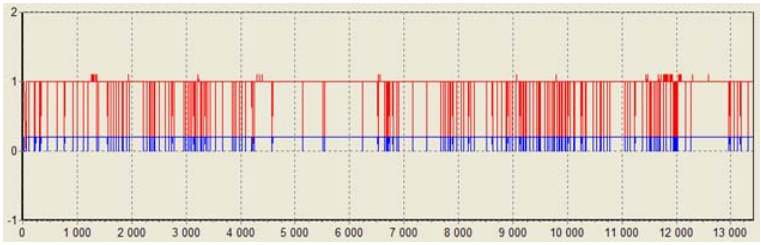


Figure 6 : Représentation graphique des changements de plans

  Il est possible alors de définir des indicateurs numériques du rythme de la séquence :

  ⇨ n_t = nombre de transitions à l'instant t, calculé sur une fenêtre temporelle d'une minute centrée sur l'instant t.

  ⇨ N = moyenne des n_t

  ⇨ δ_t = écart moyen entre transitions à l'instant t, calculé sur une fenêtre temporelle d'une minute centrée sur l'instant t.

  ⇨ Δ = moyenne des δ_t.

- La description des couleurs du film

  En utilisant la palette réduite, on calcule l'histogramme de chaque plan. Ensuite, ces histogrammes sont moyennés en pondérant l'importance d'une couleur dans un plan par la durée du plan. Sur l'histogramme final apparaissent les principales couleurs du film et leur proportion respectives. Ci-après, l'histogramme du film « A Viagem ».

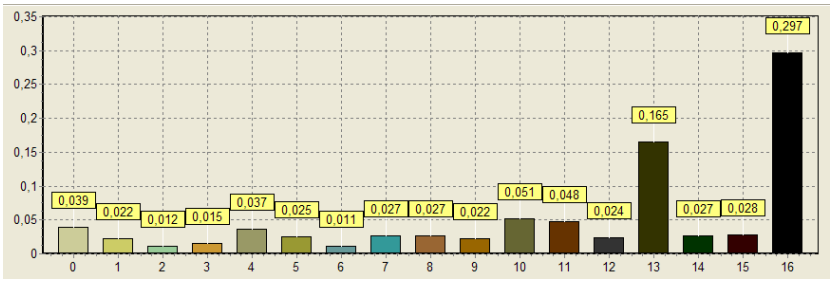


Figure 7 : Distribution couleur globale d'un film

- Les descripteurs de nature statique.

  Il est toujours possible d'isoler quelques images représentatives (par exemple le milieu des plans les plus importants) et d'utiliser les techniques d'analyse des images statiques. En particulier, on peut définir des attributs de texture qui peuvent permettre d'identifier les techniques de réalisation

5. Vers la fusion des connaissances

Les deux approches utilisées apportent des éléments de connaissance de natures très différentes à propos de chaque film. Nous attendons de la fusion de ces connaissances l'émergence de nouveaux concepts permettant de caractériser, par exemple, l'atmosphère (ou ambiance) qui se dégage d'un film. Ces travaux sont encore en cours de développement, mais on peut déjà annoncer les directions vers lesquelles il semble possible de chercher des liens entre les caractérisations de type texte et de type image.

5.1 Le texte aidant les images

Comme nous l'avons déjà indiqué, la difficulté avec les caractéristiques de type image est de les hisser à un niveau de description sémantique. Pour atteindre cet objectif, une solution peut consister à s'appuyer sur les résultats fournis par les caractéristiques textuelles. La procédure, qui s'apparente alors à une classification supervisée des images [BURG98], la supervision étant apportée ici par le texte est alors la suivante :

A – Phase d'apprentissage :

- Catégorisation d'un corpus de films à l'aide de la terminologie mise en place,
- Extraction des caractéristiques image sur ce même corpus,

- Construction de l'espace des caractéristiques image et positionnement des caractéristiques extraites du corpus,
- Étiquetage des points de cet espace par les termes issus de l'approche textuelle,
- Recherche de classes homogènes. A ce niveau, il est probable que la plupart des caractéristiques ne se grouperont pas naturellement en classes. La solution sera dans la recherche de nouvelles caractéristiques ou dans l'utilisation de techniques de classification adaptées.

B – Exploitation :

La phase d'apprentissage étant effectuée, la classification obtenue pourra ensuite être utilisée pour catégoriser des films en exploitant uniquement les données images (situation où le film n'a que peu ou pas de descriptif textuel). On peut également envisager que cette classification vienne en soutien de l'analyse textuelle pour confirmer certains choix.

Selon ce schéma, la terminologie utilisée est celle apportée par le texte, l'image venant renforcer ou suppléer le texte.

Bien sûr, l'application de cette démarche demande un choix des descripteurs image qui soit en lien avec la terminologie issue du texte. Par exemple, il semble que les caractéristiques attachées au rythme et à la couleur doivent pouvoir se fusionner à une terminologie sur les genres.

5.2 L'analyse conjointe des descripteurs

Dans cette deuxième approche, le texte et l'image sont d'abord exploités indépendamment de manière à fournir leurs propres caractéristiques. Ceci suppose que l'analyse des images parvienne à fournir des caractéristiques dont la nature sémantique soit suffisante. On peut pour cela s'aider des transcriptions numérique / symbolique proposées par la logique floue [BOUC90].

Ensuite, en se plaçant dans l'espace de dimension N des caractéristiques (N est le nombre total de caractéristiques de type texte et image), on procède à une analyse factorielle dans le but de faire émerger des axes principaux caractéristiques. L'intervention des experts est alors indispensable pour donner du « sens » à ces axes factoriels.

6. Conclusion

Dans ce papier, nous avons proposé une méthodologie de construction d'une base de connaissances sur les films d'animation à partir de deux sources différentes : les péri-textes et les images. Un certain nombre de descripteurs ont été présentés. Cet ensemble de descripteurs doit encore être enrichi. Dans le domaine du texte,

l'utilisation des critiques de films, la gestion des synonymes et la prise en compte de l'évolution des termes dans le temps permettront d'affiner la base de connaissances terminologiques. Dans le domaine de l'image, l'extraction des caractéristiques de mouvement, la mise en place d'opérateurs spécifiques (recherche de visages, suivi de personnages, etc.) donneront une meilleure pertinence des informations sémantiques issues de cette approche.

Enfin, les techniques de fusion envisagées doivent être validées par des expérimentations sur une large base de films, en accord avec les experts du domaine.

7. Bibliographie

[BOUC90] B. Bouchon-Meunier, S. Després, D. Dubois, O. Gascuel, A. Genoche, H. Prade, "Interface entre symbolique et numérique", Actes des 3ème journées nationales du PRC-IA, Hermès, p. 89-138, Paris 1990.

[BOUR99] Bourigault D., Slodzian M., "Pour une terminologie textuelle", *Revue Terminologies Nouvelles*, n° 19, déc. 1998 - juin 1999.

[BURG98] C.J.C Burges, "A tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery, Kluwer Acadelic Publishers, vol. 2, n° 2, p. 12-16, 1998.

[DELB99] A. Del Bimbo, "Visual Information Retrieval", *Morgan Kaufmann Publishers Inc*., San Francisco California, USA, 1999.

[HUET03] B. Huet, I. Yahiaoui and B. Merialdo, "Comparison of multi-episode video summarization algorithms", *EURASIP Journal on applied signal processing Special issue on multimedia signal processing*, 2003(1), January 2003.

[IONE04] B. Ionescu "Contribution à l'étude d'un système d'indexation vidéo", *Rapport interne LISTIC*, n° 04-05, déc. 2004.

[LEFE03] S. Lefèvre, N. Vincent, C. Proust,"Architectures et Outils pour la recherche d'Evènements dans les Séquences Vidéo", *Congrès INFORSID en Bases de Données et Recherche d'Information*, Nancy, France, juin 2003.

[LIEN01] R. Lienhart, "Reliable Transition Detection In Videos: A Survey and Practitioner's Guide", *International Journal of Image and Graphics* (IJIG), vol. 1, n° 3, p. 469-486, 2001.

[MEYE92] Meyer I., Skuce D., Bowker L., Eck K., "Towards a new generation of terminological resources: an experiment in building a Terminological Knowledge Base", Proceedings of the 14th International Conference on Computational Linguistics (COLING 92), p. 956-960, 23-28 August 1992.

[MPEG7] http://www.chiariglione.org/mpeg/standards/mepg-7/mpeg-7.htm.

[NF ISO 704 2001] *AFNOR*, avril 2001, ISSN 0335-3931.

[ROCH99] Roche C., Marty J.C., Lacroix S., "Ontologie et terminologie : le modèle OK", *Revue Terminologies Nouvelles*, n° 19, déc. 1998 - juin 1999.

[ROCH01] Roche C., "From Information Society to Knowledge Society: the Ontological Issue", C*ASYS'2001*, Liège, Belgium, 13-18 August 2001.

[SANT01] S. Santini, "Exploratory Image Databases: Content-Based Retrieval", *Academic Press*, 2001.

[SLOD95] Slodzian M., "Comment revisiter la doctrine terminologique aujourd'hui ?", *Banque des mots*, n° spécial 7/1995.

[SMEU01] A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, "Image databases at the end of the early years", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22(12), p. 1349-1380, 2001.

[WURD04] "Worldnet User's Reference Desk", (www.wurd.com/pwp_color.php).

Le concept de genre comme point de départ pour une modélisation sémantique du document électronique

Ioannis Kanellos, Thomas Le Bras, Frédéric Miras, Ioana Suciu

ENST de Bretagne – département LUSSI - France
Ioannis.kanellos@enst-bretagne.fr

Résumé :

L'article propose une analyse de la question du genre à l'aune du document électronique (DE). Il défend l'idée que le concept de genre, souvent négligé, se trouve à la source de toute tentative de modélisation sémantique du DE. Il articule ainsi secondairement une critique du projet d'un WEB authentiquement sémantique, en cherchant à dévier le thème du sens vers des considérations interprétatives, où la notion de pratique de lecture exige d'être clarifiée. Il ouvre, enfin, sur une méthodologie visant les typologies de genres de DE dans des domaines particuliers. Il propose ainsi un ensemble de directions de recherche sur la formalité du concept de genre de DE.

Mots-clés : Genre, typologie des genres, point de vue, pratique, document électronique, stratégie de lecture, sémantique de l'interprétation.

1. Le genre : entre traditions et recherches actuelles

Présente en moult domaines – et pas seulement en littérature –, la question du genre délimite au fond un projet de salubrité épistémologique :

- Tout d'abord, il s'agit de décrire la variété des formes complexes qui peuplent les pratiques de notre quotidien en les rapportant à des formes simples, générales, peut-être aussi premières. C'est une attitude récurrente dans la pratique scientifique, depuis toujours. Souvenons-nous que Platon, dans un fameux dialogue (le *Sophiste*) pose déjà la question des genres de

l'être, tout en expliquant l'importance et la nécessité d'une telle question. Puis, c'est le tour de son élève, Aristote, qui, cherchant précisément les conditions d'une langue qui dit l'être, en propose les fameuses *Catégories*, source de tant de malentendus en Sciences Cognitives. Elles constituent, au fond, les genres généralissimes de la prédication.

- Ensuite, il s'agit de disposer des moyens de reconnaissance des formes qui composent nos espaces empiriques. D'induire, aussi, rationnellement, une normalisation dans les dynamiques qui président nos représentations et un monde phénoménal qui se dresse autour de nous comme contexte.
- Enfin, il s'agit de se munir des normes de régulation de l'existant, par capitalisation et retour d'expérience, ainsi que des techniques de bon bricoleur cognitif capables d'adapter nos schémas de compréhension et d'action à des situations nouvelles.

Le DE ne devait pas se tenir longtemps à l'écart de telles problématiques, somme toute communes à l'exercice de toute science. Il ne le pouvait pas non plus. Même s'il devait parfois le faire de manière irrespectueuse par rapport aux traditions critiques répertoriées. Après tout, le DE, avant d'être « électronique », est document et son héritage au moins textuel lui donne droit à pas mal d'éléments d'une dot riche sur la question des genres.

Comme toujours, les définitions des dictionnaires donnent des vues trop générales, non spécialisées sur la notion de genre. *Le Petit Robert* voit le genre comme une idée générale d'un groupe d'êtres ou d'objets présentant des caractères communs, une subdivision d'une famille, ou une catégorie d'œuvres définie par la tradition (d'après le sujet, le ton, le style). Espèce, sorte, type, mais aussi façon, mode, attitude, forme et parfois race en font le voisinage du terme sur le plan du sens. *Larousse* en propose une variante et précise que la catégorie-genre sert à rassembler des œuvres qui répondent à des critères pragmatiques, formels ou thématiques semblables. Mais aussi, que le genre est, dans d'autres contextes, manière de s'exprimer, de vivre ou de se comporter en société. Ailleurs, on trouvera également une idée de descendance dans la charge sémique du genre. Pour les sciences de la vie, il s'agira d'espèce, genre et famille.

Il y a probablement quelque chose de commun dans ces tentatives de définition de la notion de genre. Disons, le concept d'origine. Et, associée, l'idée que c'est en se rapportant aux origines qu'on connaît mieux les choses. Probablement.

Plusieurs années après l'étude des [Crowston & Williams 1997], sur 1 000 pages WEB, qui visait déjà un classement des genres auxquels elles appartenaient, nous arrivons aujourd'hui aux mêmes conclusions. En effet, les auteurs mettaient à l'époque en évidence 48 genres « différents », dont 60 % étaient une pure reproduction de genres non électroniques, 30 % des adaptations de genres papier à l'électronique, 5% de nouveaux genres et 5 % inclassables.

Les critères (nombreux) utilisés pour arriver à ce type de résultats dépendent toujours des objectifs de la classification. Certains hantent encore nos imaginaires classificatoires : ce sont, en général, ceux liés au style du DE, une notion incertaine,

qui semble exprimer un jugement global, et qui renoue avec des problématiques anciennes ([Riegl 1893], [Wölfflin 1915]) ou déjà établies voire faisant désormais référence ([Adam 1992], [Combe 1992], [Fowler 1992], [Genette 1986], [Swales 1990], [Schaeffer 1989] entre autres). La notion de style, appréhendée comme un langage utilisé par l'auteur de façon à être compris du lecteur, se rapproche du concept de genre. [Schmid-Isler 2000] reprend cette idée de manière explicite trois ans plus tard. Ce style-langage semble déjà mettre en valeur une vision « interprétative » des informations partagées par une même communauté.

Plutôt empiriques, ces études ont récemment connu un accroissement important, au point de voir s'octroyer des sessions consacrées tous les ans à des colloques internationaux[1]. Peut-être non sans raison. Le terme désormais consacré dans la langue de la mondialisation est celui de « *digital genre* ».

Considérant les genres comme des émergences de conventions discursives, [Schaeffer 1989] en distingue essentiellement trois : *constituantes*, *régulatrices* et conventions *traditionnelles*.

- Les conventions constituantes auraient pour fonction d'instituer la communication et de lui donner une forme. Elles concerneraient les actes communicationnels et permettent, par exemple, d'opposer les actes expressifs, persuasifs et assertifs, voire de montrer certains effets de parole.
- Les conventions régulatrices rajouteraient des règles à la communication première. Ce sont des particularités dans la forme du discours qui viennent s'insérer dans cet acte de communication (contraintes métriques, phonologiques, stylistiques et même de contenu).
- Les conventions traditionnelles, enfin, moins strictes que les précédentes, auraient pour objet le(s) sens du discours, issu de son rapport à des productions discursives stabilisées. Ces dernières conventions opposent habituellement un texte actuel à un texte passé, censé lui servir de modèle. Plus incertaines que les autres, elles supportent l'inscription d'un document à un modèle sur la base d'un air de ressemblance détecté.

Schaeffer s'occupe essentiellement de textes. Les catégories qu'il propose s'appliquent vraisemblablement à un document statique, mais se révèlent inadéquates pour le DE. Le multimédia dans la formation des DE semble effectivement nécessiter une rénovation de l'idée de genre. Transposant vaguement les catégories de Schaeffer dans des préoccupations de « *cybergenre* », [Shepherd & Polanyi 2000] prennent leurs distances avec une écriture qui contiendrait seule toute l'information, et proposent de caractériser le genre propre au DE par trois éléments, également « constitutifs » :

- Le *contenu* (information), qui s'organise suivant une structure *matérielle* (mise en page, etc.), souvent suffisante lors d'une première et rapide lecture

[1] Comme, par exemple, le Hawai International Conference on System Sciences (HICSS).

rapide pour deviner le genre et *logique* (titre, auteur, date, etc.), qui apporte de l'information sur l'organisation intellectuelle du document.

- Le *contenant* (support, médium), qui détermine le mode d'accès, d'appropriation ou de « lisibilité » de l'information (par l'homme ou la machine).
- Le *contexte de production*, qui retrace l'intention de publication, dans un cadre, une fonction ou une activité donnée. Jouant un rôle essentiel dans le processus de lecture du document, le contexte peut être en partie reflété tant dans le contenu que dans le contenant.

En plus de ces trois dimensions, qui font respectivement appel aux aspects matériels, sémantiques et pragmatiques d'un DE, [Mas 2002] propose d'inclure également d'autres caractéristiques dans la définition du genre numérique, comme la *pertinence* et la *fonction* du document. [Schmid-Isler 2000] et plus tard, [Shepherd & Watters 2004] soutiennent que ce qui différencie le DE d'un document « traditionnel », c'est exactement et essentiellement sa *fonctionnalité*. La tripartition initiale de [Shepherd & Polanyi 2000] (contenu, contenant et contexte) devient système, et même système de signes, mieux reconnaissable dans un nouveau triplet : le *contenu*, la *forme*, et la *fonctionnalité*. Si les deux premières catégories de critères (suffisantes pour un document traditionnel) correspondent respectivement au contenu et au contenant, il y a un déplacement des espaces occupés par chacune. Ainsi, le contenu concerne les attributs, les thèmes et les topiques, et devient le porteur fondamental de la signification du DE, alors que la forme concerne essentiellement l'investigation du style (concept qui ne s'est guère clarifié depuis; cf. relativement [Rastier 1994]). Toutefois, c'est la fonctionnalité qui permet à l'utilisateur non seulement de parfaire sa compréhension, mais aussi de correctement utiliser un DE.

La fonctionnalité, sorte d'agrégat de fonctions dans la *doxa*, fait cependant sortir la problématique du genre d'un DE hors d'un désir de rationalité, dans la mesure où, suivant toujours les mêmes auteurs, elle peut convoquer des facteurs économiques, politiques, techniques, esthétiques, éthiques, etc. Elle peut même dériver de considérations ayant trait au but, à l'utilité, de facultés intellectuelles, prioritairement l'attention, de se trouver en correspondance avec notre conception d'efficience (pratique, communicative, sémiotique…), etc.

Hybrides et peu envahissants, sinon déjà ambigus et inopérants, les concepts de fonction et/ou de fonctionnalité, semblent aussi douteux que la notion de contexte, thème butoir de nombreuses recherches interdisciplinaires. Convoqués dans la caractérisation du concept de genre par les derniers auteurs, ils annoncent probablement la n-ième réécriture de l'histoire concernant le progrès scientifique du point de vue fonctionnaliste. Innocemment, [Breure 2001] répétera que la présence de la fonction est une caractéristique de la théorie moderne des genres (*op. cit.* p. 35). Cela ne devrait toutefois pas étonner : le concept de genre, nécessaire et incontournable à terme dans toute forme de recherche, est parmi les plus obscurs, en mobilisant d'emblée des fonctions cognitives complexes et étendues, qui dépassent

vite l'individualisation de la connaissance pour s'ouvrir d'emblée à des régimes de régulation sémiotique de facture intersubjective. On regrettera bien sûr le recours à des concepts pour lesquels on peut douter de l'existence d'une science assurée, même à terme (comme le contexte) ou trop opaques (comme la fonction), et qui mériteraient des affinements et des formulations plus détaillées.

Dans un travail antérieur, [Shepherd & Watters 1998] voyaient déjà dans les genres des formes de communication attendues et relativement (i.e. culturellement) stables, n'admettant pas toujours une définition précise. Prêts à s'affranchir du joug de la formalité, les auteurs soutiennent que tout mode de vie implique la formation de routines confectionnées par plusieurs genres, et qui attestent d'une adaptation. Les genres possèdent généralement des rôles complémentaires dans une même activité. Souvent même, ils s'impliquent mutuellement ou reçoivent des déterminations des structures et des processus sociaux, pour régler l'adéquation toujours exigible entre les moments de la production et les séances de consommation sémiotique.

[Erickson 2000], également en rupture discrète avec une intention positiviste, voit le genre comme un « pattern of communication, created by a combination of the individual (cognitive), social and technical forces, implicit in a recurring communicative situation ». Le genre vise à structurer la communication, en donnant précisément des moyens pour scénariser des attentes partagées. Il facilite, ainsi, « the burden of production and interpretation ».

La question du point de départ de la formation des genres rappelle probablement le célèbre dilemme de la poule et de l'œuf. Le modelage des genres se fait, en réalité, toujours sur d'autres genres qui préexistent. Même si les « règles » d'un genre ne sont jamais parfaitement codifiées, l'identification d'instances de chaque genre crée des précédents, dont l'efficience cumulée génère, à la longue, des schèmes d'écriture, pouvant produire de nouvelles instances. Corrélativement, en réception, il crée des attentes pour l'interprétation des instances subséquentes. Ainsi, le genre produit, en quelque sorte, à la fois un cadre et une « tension » pour la communication, et impose une sorte de « pression », en conformant les échanges à des patterns déjà disponibles, reconnaissables et généralement partagés.

En étudiant l'existence ou la non-existence de formes analogues disponibles sur un support « traditionnel » (notamment sur papier), les travaux de [Shepherd & Polanyi 2000], [Marcoccia 2004], [Mas 2002] et [Shepherd & Watters 2004] mettent l'accent sur deux processus en œuvre qui président la genèse du DE (pour la terminologie, cf. aussi [Breure 2001]) :

- La *réplication* (reproduction) des genres par rapport aux genres préexistants, qui se retrouvent dans d'autres médias.
- L'*émergence* de nouveaux genres, qui n'ont pas d'équivalent sur support analogique ou dans un autre média.

Le premier cas comprend entre autres les documents bureautiques, comme la lettre, la forme du contrat, les notes de services, les plans, les cartes, mais aussi les

documents audiovisuels etc. Dans la deuxième catégorie, on pense à la page WEB, le courrier électronique, les tableurs, les fichiers de base de données, les documents hypertextuels, les documents composites etc. Revenant au rôle définitoire de la fonctionnalité, l'émergence de nouveaux genres s'explique souvent comme la conséquence d'une augmentation des fonctionnalités (par exemple [Shepherd & Watters 2004]). Sans doute, rien ne peut exclure des scénarios de génération spontanée, qui échapperaient aux deux processus ci-dessus.

Entités sensibles, les genres subissent, bien sûr, les inexorables mutations de la culture qui les utilise : régulations et compétitions, influences et interactions font le lot perpétuel des genres qui ne restent pas étrangers aux problématiques de la co-formation et de la co-évolution dynamiques. Qu'ils concernent un DE ou un autre produit de l'activité sémiotique humaine, on ne saurait un seul instant soutenir l'immuabilité des genres. Avant même d'arriver à une considération des différences des pratiques qui les affectent, les genres sont toujours situés dans un temps, dans un espace et dans une société. D'un autre côté, les mutations des nouvelles technologies importent de grandes opportunités de métamorphose, bien plus que les mutations de la langue ou d'autres codes sémiotiques, plutôt traditionnels et relativement stables (voir aussi [Shepherd & Watters 2004], notamment pp. 13 et 15, et [Mas 2002] *passim*). Les nouvelles technologies transforment en effet les pratiques, donc les discours et les activités associés aux pratiques, et, par conséquent, aussi, les genres et les usages qui leur sont attachés.

Plus mûre et certainement assurée, la pensée de [Rastier & Pincemin 1999] sur les genres synthétise et affine beaucoup de ces idées en les plaçant proprement dans une réflexion foisonnante répondant d'une tradition de critique herméneutique ([Rastier 1987], [Rastier 1989] entre autres). Même si elle ne concerne pas spécifiquement les genres numériques, nous pouvons l'étendre sans difficulté :

1. Si la typologie des DE ou des textes peut assumer un principe de plaisir, la théorie des genres doit obéir à un principe de réalité. Il ne faut pas assimiler typologie des DE et genres, au risque d'oublier que la définition d'un type dépend de l'analyste : pour les besoins d'une cause ou d'une application.
2. Les genres sont définis non par un critère, mais par un faisceau de critères. Ils doivent leur caractère d'objectivité à cette multiplicité des critères. Plus précisément, un genre se définit, tant au plan du signifié qu'à celui du signifiant par la cohésion d'un faisceau de critères et son incidence sur la textualité.
3. Un genre est défini comme un mode d'interaction normé entre composantes. Chaque système de genres semble rester autonome et évolue selon ses propres lois, comme en témoignent les phénomènes de co-évolution diachronique.
4. Ce n'est pas la typologie des DE, mais celle des genres de DE qui nous importe. Or cette typologie est subordonnée à celle des productions sémiotiques utilisées sur lesquelles s'appuient les échanges dans le réseau. Le genre d'un DE serait ainsi le facteur fondamental de la *sémiosis*

documentaire réalisée par les technologies d'information et de communication.

5. La caractérisation raisonnée des genres reste un préalable à la constitution de corpus pleinement utilisables pour des tâches de description. Quels que soient les critères choisis, on ne peut tirer grand-chose d'un corpus hétérogène, car les spécificités des genres s'annulent réciproquement, et les disparates qui demeurent ne peuvent être interprétées pour des fins de caractérisation des DE.

Nous chercherons à donner forme applicative à ces considérations.

2. Méthodologie pour la modélisation du genre de DE

Résumons, pour fixer les idées, quelques caractères récurrents dans les diverses approches de la notion de genre :

- L'idée d'*origine*, conduite par celle de la réécriture/relecture de l'original, qui rend la nouveauté permise en écriture et reconnaissable et compréhensible en lecture.
- Celle de *parcours*, à travers une matière formelle multiple, pouvant s'ouvrir à des réalisations nouvelles de DE, en accord avec une pratique (qui décline contextes, normes, valeurs, objectifs et profils partagés). C'est toujours cette notion de parcours qui se trouvera à la source d'une relation de réécriture (entre un DE modèle et un DE produit ; cf. [Rastier 1995]).

Ce qui opposerait un genre origine et parcours à un genre (proto)type ou classe, est, probablement, l'accent et la priorité que nous accordons au différentiel fondamental entre dynamisme et statisme. Tout comme une fonction en mathématiques, qui est vue tantôt comme mécanisme de correspondance ou de transformation, tantôt comme résultat ou image, et parfois même ensemble, le genre peut aussi apparaître tantôt comme type ou classe, tantôt comme programme de (ré)écriture ou de lecture. Relativement au DE, la distinction peut être portée par la notion de parcours. Ainsi :

- Comme objet à établir, le genre peut être un parcours en acte, un mécanisme de construction qui assemble des éléments issus de différentes qualités informationnelles (contenu, contenant, contexte…).
- Comme objet déjà établi, le genre peut être envisagé comme produit achevé par le parcours.

Réécriture et parcours, d'un côté, intégration de traits relevant de plusieurs topiques de l'autre, le genre numérique appelle à une modélisation à la fois modulaire et modulable, susceptible d'instancier des variétés comme résultats de

parcours alternatifs[2].Peu importe la voie suivie, l'opérationnalisation du concept de genre d'un DE exige des caractérisations reposant sur des items quantifiables relevant de plusieurs qualités informationnelles (cf. aussi, [Biber 1992], [Agre 1998]). En effet, toute information pouvant être reconnue et exploitée dans un DE n'est pas de même qualité, niveau ou importance. Par exemple, les caractéristiques typographiques ne peuvent pas être considérées comme les caractéristiques logiques ou conceptuelles, ni même comme des critères sociaux : le plan de l'expression ne peut se substituer au plan du contenu, même si ces deux plans entretiennent des rapports indissociables, le plan de l'ergonomie ne peut assurer les causes (matérielles, finales, efficientes, formelles) qui font l'identité d'un DE.

La question est d'importance. Car, en réalité, elle cache une autre, probablement de fond : celle d'un modèle global du DE, modèle qui manque toujours aussi cruellement. Les différents langages de balisage ne constituent pas un modèle, et, de toute façon, ils ne peuvent capter qu'une partie du contenu d'un DE, pas toujours la plus significative d'ailleurs. Les diverses tentatives de caractérisation du genre d'un DE, que nous venons de commenter rapidement, malgré leur aspect parfois élémentaire, qui nous fait revisiter l'histoire du développement des concepts de la Linguistique contemporaine (contenu, contenant et contexte reprennent au fond la tripartition convenable entre Sémantique, Syntaxe et Pragmatique dont les insuffisances ne sont plus à démontrer), attestent, au moins, d'une conscience acquise : que toute caractérisation du concept de genre passe par une composition de caractéristiques, d'une sorte de faisceaux de traits cohérents, comme le proposent déjà [Rastier & Pincemin 1999][3]. Tel un vecteur de caractéristiques, la modélisation du genre doit faire la part à la fois de cette polymorphie et de cette irréductibilité.

Ce dernier constat est angulaire. Car il est à la source d'une méthodologie pour la caractérisation du genre du DE. En effet, il devient directement programme :

- À un premier niveau, nous devons poser les caractéristiques que nous considérons pertinentes pour être prises en compte.
- À un deuxième niveau, nous devons expliciter la formation des faisceaux : les dimensions informationnelles mutuellement exclusives. Convenons de les appeler *points de vue* (PdV). En réalité, la matière véritable du genre n'est pas composée par les informations du premier niveau, mais par celles,

---

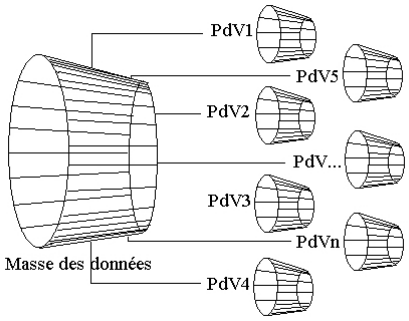
[2] De tout ce que nous avons pu trouver, nous ne connaissons pas de tentative de formalisation du genre reposant sur la notion de parcours. La plupart des travaux sur le parcours de documents électroniques ne font même pas allusion au genre.

[3] Une variante de cette idée se trouve aussi dans [Pedauque2003]. En effet, cette tripartition s'identifie sous la terminologie de « forme, signe et medium » et se confond avec la notion de point de vue, notion reprise ici (cf. plus bas), mais dans une perspective différente et certainement plus générale. On retrouve dans cette référence une lointaine analogie entre la « spécialité », vue comme une sorte de domaine spécifique qui s'attarde sur un des points de vue et le genre du document électronique qui sera abordé ensuite. Cependant, [Pedauque2003] semble peu intéressé par la problématique du genre, (le terme « genre » n'est jamais cité). La rhétorique du genre doit se construire a posteriori. Et c'est dommage.

surdéterminées par leur qualité informationnelle, de ce second. Déjà structurées, elles promulguent leurs différences en des potentialités définitoires des genres.

- Le genre serait ainsi un parcours (vision procédurale) et/ou le résultat d'un parcours (vision résultante) opéré sur ce deuxième niveau.

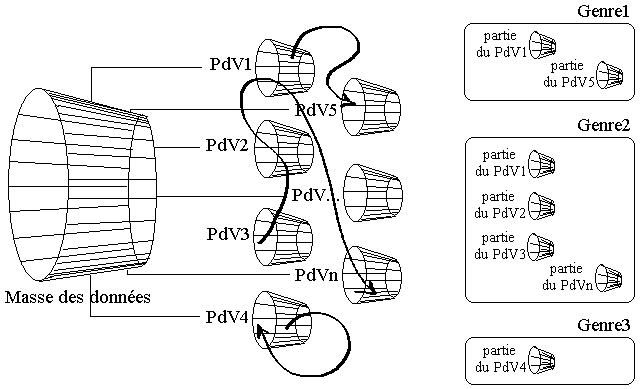
La figure (3.1) détaille les deux premiers niveaux mentionnés.



(3.1)

Le genre est un processus qui « visite » certains points de vue où il glane un ensemble de traits. Sorte de pré-lecture, il vise à *situer* le DE dans une pratique stabilisée, déterminée par une classe de lectures pertinentes au sein d'une communauté. L'effet remarquable d'une telle opération de mise en situation est la réduction de la complexité du DE, en limitant les possibilités de son traitement. Et donc, l'augmentation de la performance et de l'efficacité d'une lecture détaillée, souvent orientée, toujours en projet. Il sélectionne, d'une certaine manière, les stratégies pertinentes pour l'appropriation du DE suivant un ensemble de PdV. Il constitue, dans cette forme d'approche (3.2), des éléments d'un troisième niveau :

Simple, une telle représentation (qui s'inspire librement de [Kanellos *& al.* 2000]), procède par intégration successive au travers de ces trois niveaux et peut devenir le siège de nombreuses possibilités de formalisation de notions qu'on vient de visiter. Elle peut, de plus, étendre nos conceptions sur la question de genre tout en constituant autant de programme de recherche. Nous en esquissons certains.



(3.2)

3. Le genre numérique : formalisation et axes d'étude

Quelques considérations préliminaires, tout d'abord :

- Dans une constante volonté de tenir compte au moins de certains aspects de la constitution procédurale d'un genre, on conviendra que l'identité d'un genre est sensible à l'ordre du parcours qui le constitue (au moins en attribuant des valeurs d'importance à l'ordre établi par le parcours). Ainsi, deux genres contenant les mêmes parties peuvent être considérés comme différents au sens de leur « histoire de constitution » (cf. Définition 1).
- L'utilité d'un genre maximal, établi par l'intégralité de tous les PdV, semble douteuse. Même si l'on souhaite une telle monstruosité, on admettra que de tels genres maximaux sont nombreux (dans la mesure où différents parcours peuvent aboutir à sélectionner l'intégralité des tous les PdV).
- Un genre peut contenir des parties issues des PdV arbitrairement grandes (mais strictement incluses dans les PdV concernés : cf. le point précédent). Procédure de choix, un genre se constitue comme image d'une pratique de lecture efficiente *(i.e.* réduite en complexité). Il peut aussi bien contenir un grand nombre de PdV ou seulement un. Même moins : un seul attribut glané d'un PdV. Dans ce cas, on retrouve la problématique typologique (départage du corpus sur la base d'un seul critère, d'un seul PdV).
- Il n'y a pas de contraintes de sélection des parties de PdV entre genres : une même partie peut se trouver intacte dans deux genres différents.
- Il n'y a pas non plus de contrainte sur le nombre ou la complexité des genres qui dépendent de la richesse structurale du niveau des PdV.

- Un genre peut décliner des sous-genres. Ils sont aisément définissables par des considérations d'inclusion de type ensembliste, et en constituer des structures d'ordre partiel (cf. aussi, ci-dessous). Cependant, la variété peut être plus grande, dans la mesure où on peut ou non tenir compte de l'orientation du parcours à l'origine de la formation d'un genre.
- On peut imaginer que la constitution d'une expertise étale des genres plutôt nombreux et petits (contrairement à une conception naïve du domaine, qui procéderait sur peu de genres mais massifs).

Définition 1 :

Deux genres sont égaux *intentionnellement* s'ils contiennent les mêmes parties de PdV et dans le même ordre. Ils sont égaux *extensionnellement* s'ils possèdent les mêmes parties de PdV.

Clairement, l'égalité intensionnelle entre genres est plus fine que l'égalité extensionnelle. Une telle définition ouvre déjà sur un projet d'étude des formes d'égalité entre genres dans le sens de [Zaldivar-Carrillo 1995] entre autres. De l'autre côté, une telle définition met en avant la possibilité d'un projet d'investigation sur la sériation des genres sur un axe allant de l'égalité *intensionnelle* (niveau d'indiscernabilité maximal) à la *rupture* (niveau d'étrangeté maximal), passant par diverses formes d'*opposition* entre genres.

Définition 2 :

Deux genres sont dits en *rupture*, lorsqu'ils n'ont pas de parties issues de mêmes PdV.

Autrement dit, deux genres en rupture établissent deux conceptions fort différentes de DE. Par exemple, les genres 4 et 3 ou 4 et 1 dans la figure (3.2). La relation de rupture est seulement symétrique. Deux genres en rupture ne sont pas directement complémentaires. La complémentarité s'obtient de deux manières : relative et absolue. Elle est *relative* lorsque les deux genres contiennent des parties issues des mêmes PdV qui sont mutuellement complémentaires par rapport aux PdV concernés. Elle est *absolue* lorsque les parties en question reconstituent la totalité des PdV. La deuxième possibilité semble être plutôt l'exception.

Définition 3 :

Deux genres sont en opposition lorsqu'ils ne sont ni égaux (extensionnellement ou intentionnellement), ni en rupture.

En d'autres termes, deux genres en opposition partagent quelque chose en commun. La notion d'opposition est plurielle et décline de grandes quantités de variétés. Par exemple, deux genres sont dits en *opposition étroite* lorsqu'ils sont formés par les mêmes parties de PdV dans le même ordre de parcours, mais ayant des valeurs d'attributs définitionnels différents. C'est, par exemple, le cas de deux lectures qui s'établissent exactement de la même manière, mais

avec des valeurs attribuées différemment. On peut en faire la généalogie quantitative d'une telle notion (dépendant du nombre des différences détectées). Deux genres sont en *opposition étendue* lorsqu'ils sont établis par la même structure de parcours (ordre de sélection), sur les mêmes PdV, mais contenant des parties différentes de ces PdV. On peut aussi jouer à loisir sur les possibilités d'une telle notion, somme toute combinatoire et ensembliste, dont la limite n'est pas la notion de rupture entre les genres mais celle d'une différentiation par PdV : deux genres peuvent aussi différer sur un ou plusieurs PdV, et ceci de moult manières.

Ces définitions peuvent aussi se trouver à la source d'une classe plus ou moins intéressante de relations de similitude. Voire, même, de relations enrobées par une métaphore de proximité géométrique. Bien entendu, et comme toujours, il s'agira de quantifier la partie symbolique commune, et en distinguer des variétés de similitude. Nous en faisons également l'économie, en nous contentant des choses de base : dans une métrique de similarité, l'égalité intentionnelle en constitue le degré maximal, alors que les diverses formes de rupture entre genres le degré zéro. Les degrés de similitude doivent cependant être relatifs, concernant des lignées de genres (cf. ci-dessous), et non absolus, dans la mesure où les structures hiérarchiques ne forment pas toujours un bon ordre. L'attribution de valeurs arbitraires résout sans doute le problème dans le cadre d'une obligation applicative, mais reste toujours discutable et, à nos yeux, de peu de signification.

Il est possible de donner aussi substance formelle aux phénomènes de génération et, plus généralement, de transformation des genres. Un genre génère « naturellement » des *lignées*, *i.e.* des sous-genres spécifiques, définis par restriction des parties des PdV qui le constituent. Un genre peut aussi être à l'origine d'un nouveau genre soit par *complétion* soit par *élagage*. De même, ces opérations peuvent être soit relatives au PdV (extension à des informations d'un PdV déjà sélectionné), soit trans-PdV (extension à des informations d'un PdV non encore sélectionné). Il est ainsi aisé de formaliser les cas que nous avons vus à la section précédente. La transformation d'un genre peut être vue comme une succession d'ajouts et de suppressions. Cependant, ces opérations empruntent également des parcours privilégiés qui sont instanciés dans la structure de préséance des parties de PdV constitutives des genres. Ainsi, il peut y avoir des ajouts et/ou des suppressions qui aboutissent à des structures génériques extensionnellement égales, mais pas aux mêmes structures intensionnelles, puisque reflétant des parcours différents.

Enfin, il est possible d'investir cette même formalité dans la constitution des circonstances de rédaction et de lecture faisant partie des pôles extrinsèques non seulement d'un DE isolé, mais d'une société de DE (disons, d'un *inter-DE*, sorte de généralisation de l'intertexte), constituant un corpus homogène notamment ([Kanellos 1999a]). En effet, une famille de genres dessine l'extériorité d'un corpus de DE en induisant divers degrés de cohérence même s'il est entaché d'hétéroclite. Une telle « inter-généricité », situe la diversité des DE appartenant au corpus en stabilisant « leur vocabulaire de construction, leurs formes d'organisation, leurs

contenus attendus, leurs modes rédactionnels » ([Rastier & Pincemin 1999]). En d'autres termes, une famille de genres peut donner cadre, légitimité et intelligibilité à un corpus de DE.

4. Pour un WEB authentiquement sémantique

Le WEB *Sémantique*, qui annonce un nouveau « rapprochement » de la forme et du sens, en cherchant des moyens pour réinvestir des capitaux sémantiques dans la bourse même des DE, peut aussi être compris comme la revanche des exclus : plusieurs chercheurs d'une IA déjà vieillie y trouvent sans doute aujourd'hui un nouveau terrain d'exercice et de création d'écoles, pour servir l'évolution de leurs concepts et techniques appuyant encore l'espoir. Non sans raison, peut-être. Cependant, la sémantique n'est pas une, et les projets qu'elle est capable de déployer dans les problématiques foisonnantes autour du DE ne peuvent se suffire aux dérivés syntaxiques qu'on nomme souvent abusivement « sémantiques ». Le WEB Sémantique réactualise le désir de sens d'antan, démontre sa nécessité, apporte de l'évidence à l'argument du contenu dans les opérations fondamentales du traitement de l'information… Mais il est loin, très loin d'une réalisation complète ou simplement satisfaisante des exigences scientifiques qu'il préconise comme horizon.

Tout au contraire même : il complexifie un tel désir en le subordonnant désormais à des notions difficilement maîtrisables portées par la multimodalité ([Kanellos 1999b]) et le multilinguisme qui affectent les masses de données et qui ont une préférence irrépressible pour l'entropie. Cependant, la partageabilité, qui institue les communautés de pratiques peut encore opérationnaliser ce WEB d'un imaginaire d'information et de communication enfin humanisées. En extrayant, notamment, des localités opératoires à la manière que le groupe social donne localement substance à la fonction globale de la société. La société est un terme abstrait, hypothétique : seul le groupe social participe aux processus de notre expérience qui règlent l'action individuelle sur les normes et les valeurs en vigueur aussi par d'autres. La massivité des données devient moins inquiétante ou désespérante si nous arrivons à construire un projet de cohérence locale, reconnaissable par des pratiques que nous revendiquons. C'est à cette exigence que répond le genre. Et l'ensemble de notre argument dans cette contribution.

Sorte de structure acquise, validée et partagée, le genre permet de simplifier les traitements. Il est le moyen qui assure deux médiations dont nous avons constamment besoin pour améliorer nos performances sémiotiques trouvant siège et mesure d'évaluation aux DE ([Rastier & Pincemin 1999]) :

- La médiation symbolique, qui articule l'individuel et le social. Dans une telle socialisation se superposent une socialisation de facture typique, et une autre, faite de représentations dérivées de la première et élargie par la communication à travers le réseau. Précisément grâce à des DE.
- La médiation sémiotique, qui sépare le physique du représentationnel.

Certes, il n'y a pas que des lois du genre. Cependant, il apparaît toujours comme le facteur historique de normalisation de tous ces langages émergents et hybrides, sur et autour des pratiques de communication impliquant des DE ([Orlinowski & Yates 1994], [Orlinowski & Yates 1998]). Totalisant et commandant désormais les performances sémiotiques complexes, ces pratiques coordonnent genres et usages comme les schèmes du comprendre et du faire, nécessaires pour une réponse individuelle située, à la fois adaptée à un environnement et partageable par un groupe. Ils constituent l'élément de fond de la formation de l'épistémè.

La conscience d'une telle possibilité (nécessité) passe aujourd'hui par des préoccupations d'objectivation, dont la forme la plus remarquable est celle d'ontologie. Le concept d'ontologie, sans doute incontournable, ne saurait toutefois épuiser l'exigence de la partageabilité, dans la mesure où cette dernière annonce plus qu'une normativité consensuelle faite d'objets et de relations : la capacité de lire un fragment de monde de la même manière que d'autres dans lesquels nous installons notre action. Ceux précisément qui nous importent. La capacité, aussi, d'adresser le même type de stratégies de lecture, qui sont gages non pas d'une univocité, mais certainement d'une conformation des sens possibles. D'une cohérence aussi, dont la norme n'est plus à rechercher à l'objectif et à l'immuable, au vrai ou même à l'incontestable, mais à des éléments fondamentaux qui composent l'intersubjectivité.

Partager des documents électroniques n'est rien – ou peu – si l'on ne peut les accompagner de « réflexes interprétatifs » reconnaissables par tous ceux avec qui nous faisons (ou nous souhaitons faire) société. La sémantique qui accompagne l'identification du genre est le socle de toute lecture adaptée et partageable, qui fait muter l'horizon de l'objectivité, toujours à rechercher, en un dessein collectif institué précisément à travers l'intersubjectif. Résumée, c'est toute l'idée de notre travail. Qui n'en constitue pas moins un projet : celui d'une caractérisation des genres du DE. Une caractérisation urgente, tant que la fonction culturelle du DE nous importe encore. Un travail à plusieurs, un projet qui ne peut se limiter aux seuls formalismes syntaxiques. La modélisation des genres du DE est un devoir « civique » dans toute société s'appuyant sur le WEB et les moyens de communication que celui-ci permet.

La méthodologie qui accompagne une telle vision du concept de genre est librement généralisable. Elle se base, au fond, sur le préalable d'une reconnaissance d'autres genres : ceux qui distinguent entre elles, dès la mise en place d'un modèle, des qualités informationnelles. Un départage qui projette en réalité, sur le tissu formel, un élément omniprésent dans toute performance sémiotique humaine. La richesse formelle et les possibilités d'identification non extensionnelle qu'elle permet sont l'effet émergent des déterminations additives, engagées par la hiérarchisation des niveaux de constitution. Sa généralité, sa généricité aussi, sont impliquées par le fait qu'elle donne une schématisation claire et exploitable de l'eidétique possible composée de multimodalité et de profondeur variable des informations. De la modularité qui recoupe la modalité également. Par sa capacité de traduire des programmes d'initialisation des projets sémantiques, aussi.

Par le genre, il y a encore lieu d'espérer que la sémantique ne restera pas au niveau de la lettre. Celle qui tue.

5. Références

[Adam 1992] Adam, J.-M. : *Les textes : types et prototypes*, Paris, Nathan, 1992.

[Agre 1998] Agre, Ph.: "Designing Genres for New Media: Social, Economical and Political Contexts", *CyberSociety 2.0: Revisiting CMC and Community*, Sage, 1998.

[Bauer 1986] Bauer, H.: "Form, Struktur, Stil", *in* H. Belting (Ed), *Kunstgeschichte*, Reimer Verlag, 1986.

[Bergquist & Ljungberg 1999] Bergquist, M. & Ljungberg, J.: "Genres in Action: Negotiating Genres in Practice", *Proceedings of HICSS'99*, Hawaii, (1999).

[Biber 1992] Biber, D.: "The multi-dimensional approach to linguistic analysis of genre variation: an overview of methodology and findings", *Computers and the Humanities*, 26 (5-6), 1992, p. 331-345.

[Breure 2001] Breure, L.: "Development of the Genre Concept", 2001.

[Combe 1992] Combe, D. : *Les genres littéraires*, Paris, Hachette 1992.

[Crowston & Williams 1997] Crowston, K., & Williams, M.: "Reproduced and emergent genres of communication on the World Wide Web", *Proceedings HICSS'97*. Maui, Hawaii, vol. VI, 1997, p. 30-39.

[Erickson 2000] Erickson, T.: "Making Sense of Computer-Mediated Communication (CMC): Conversations as Genres, CMC Systems as Genre Ecologies", *Proceedings of the 33th Hawaii International Conference on Systems Science*, 2000.

[Fowler 1982] Fowler, A.: *Kinds of Literature. An Introduction to the Theory of Genres and Modes*, Oxford, Oxford University Press, 1982.

[Genette 1986] Genette, G. (éd.) : *Théorie des genres*, Paris, Seuil, 1986.

[Kanellos 1999a] Kanellos, I. : « De la vie sociale du texte. L'intertexte comme facteur de la coopération interprétative », *Cahiers de Praxématique* 33, 1999, p. 41-82.

[Kanellos 1999b] Kanellos, I. : « À propos de l'héritage critique du document électronique : multimodalité sémiotique, stratégies de lecture et intertextualité », *Actes CIDE'99*, Damas, Syrie, juillet 1999, p. 97-109.

[Kanellos & *al.* 2000] Kanellos, I., Thlivitis, Th. & Léger, A.: « Indexation anthropocentrée d'images au moyen de textes», *InCognito*, n° 17, p. 33-43.

[Dehors & Le Bras 2005] Le Bras, Th. & Dehors, S.: « Path analysis among multimedia resources : a typology », ICMTL 2005.

[Marcoccia 2003] Marcoccia, M. : « La communication médiatisée par ordinateur : problèmes de genres et de typologie », *Journée d'étude : « Les genres à l'oral »*, Université Lumière-Lyon 2, 18 avril 2003.

[Margolin & Buchanan 1998] Margolin, V. & Buchanan, R.: *The idea of Design. "A design Issues" Reader*, MIT Press, Cambridge, 1998.

[Mas 2002] Mas, S. : « Propos généraux sur la notion de document », in *Définition, caractéristiques et organisation des documents administratifs électronique*, séminaire de doctorat en Science de l'information, Université de Montréal.

[Orlinowski & Yates 1994] Orlikowski, W., & Yates, J.: "Genre repertoire: the structuring of communicative practices in organizations", *Administrative Science Quarterly* 39, 4, 1994, p. 542-574.

[Orlinowski & Yates 1998] Orlikowski, W., & Yates, J.: *Genre systems: structuring interaction through communicative norms*, Cambridge, MA, MIT, 1998.

[Rastier & Pincemin 1999] Rastier, F. & Pincemin, B. : Des genres à l'intertexte, *Cahiers de Praxématique* 33, 1999, p. 83-97.

[Rastier 1987] Rastier, F. : *Sémantique interprétative*, Paris, P.U.F., 1987.

[Rastier 1989] Rastier, F. : *Sens et textualité*, Paris, Hachette, 1989.

[Rastier 1994] Rastier, F. : « Le problème du style pour la sémantique du texte » *in* Molinié, G. & Cahné, P. (éd.), *Qu'est-ce que le style ?*, Paris, P.U.F., 1994.

[Rastier 1995] Rastier, F. : Communication ou transmission, *Césure*, 8, 1995, p. 151-195.

[Riegl 1893] Riegl, A.: Stilfragen, Grundlegung zu einer Geschichte der Ornamentik, 1923.

[Pédauque 2003] Pédauque, R.T. : Document : forme, signe et médium, les re-formulations du numérique (*cf.* http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/05/11/).

[Schaeffer 1989] Schaeffer, J.-M. : *Qu'est-ce qu'un genre littéraire ?*, Paris, Seuil, 1989.

[Schmid-Isler 2000] Schmid-Isler, S.: "The Language of Digital Genres – A Semiotic Investigation of Style and Iconology on the WWW", in *Proceedings of the 33rd Hawaii International Conference on System Sciences*, 2000, IEEE, p. 57-66.

[Shepherd & Polanyi 2000] Shepherd, M. & Polanyi, L.: "Genre in Digital Documents", in *Proceedings of the 33rd Hawaii International Conference on System Sciences*, 2000.

[Shepherd & Watters 1998] Shepherd, M., & Watters, C.R.: "The evolution of cybergenres", *Proceedings of HICSS'98*. Hawaii, vol. II, 1998, p. 97-109.

[Shepherd & Watters 2004] Shepherd, M. & Watters, C.: "IdentifyingWeb Genre: Hitting a Moving Target", http://www.cs.dal.ca/~watters/www2004WorkShop/pdfs/4.pdf.

[Swales 1990] Swales, J.-M.: *Genre Analysis. English in Academic and Research Settings*, Cambridge, Cambridge University Press, 1990.

[Wölfflin 1915] Wölfflin, H.: *Kunstgeshichtliche Grundbegriffe*, München 1915.

[Zaldivar-Carrillo 1995] Zaldivar-Carrillo, V.-H. : *Contributions à la formalisation de la notion de contexte : La notion de « théorie » dans la représentation des connaissances*, Thèse de doctorat, Université de Montpellier II, 1995.

Développer la communication orale multilingue sur le web, créer et partager des corpus de parole multilingues, avec la plate-forme évolutive ERIM

Georges Fafiotte

GETA, CLIPS-IMAG (UJF – Université de Grenoble 1),
385 rue de la Bibliothèque, BP 53, 38041 Grenoble Cedex 9 - France

georges.fafiotte@imag.fr

Résumé :

La première étape de notre travail a donc été de proposer une typologie des différentes stratégies d'oralisation envisageables et des relations que celles-ci entretiennent. Ce cadre est sous-tendu en particulier par le Modèle d'Architecture Textuelle (MAT) [VIRB89] et la théorie de la fonction représentative de l'intonation pragmatique du français de [ROSS99].

L'action de développement de ressources de communication orale multilingue sur réseau, dont les réalisations actuelles et les retombées possibles sont décrites dans l'article, s'est déployée parallèlement à l'intégration du français en Traduction Automatique de Parole (TAP) multilingue, effectuée au CLIPS-IMAG pour les projets internationaux C-STAR et NESPOLE!. Dans le cadre de deux projets, ERIM (Environnement Réseau pour l'Interprétariat Multimodal) et ChinFaDial (architecture ouverte de systèmes de TAP, avec collecte de dialogues parlés spontanés français-chinois), nous avons prototypé différentes facettes d'une plate-forme évolutive.

Ces environnements traitent plusieurs aspects de la communication orale spontanée bilingue non finalisée sur le web : interprétariat humain à distance, collecte de corpus de dialogues spontanés bilingues traduits, interaction multimodale entre interlocuteurs, et intégration d'aides automatiques aux intervenants. Ces dernières expérimentent un serveur de TAP utilisant des composants du marché, et des aides en ligne aux interprètes ou aux locuteurs.

Nous présentons les premiers résultats : les prototypes ERIM-Interprète et ERIM-Collecte sont opérationnels. Des corpus sont collectés en français-chinois, français-vietnamien et français-hindi d'abord, qui seront accessibles sur un site serveur (DistribDial), de même qu'une version robuste d'ERIM-

Collecte. Il est prévu d'intégrer les différentes plates-formes en un système multifonctionnel ERIMM, « plate-forme laboratoire » d'aide à la communication multilingue multimodale sur réseau, dont une variante pourra s'étendre à la formation à distance (e-training) à l'interprétariat.

Mots-clés : Interprétariat à distance sur réseau, collecte de corpus oraux bilingues, dialogues spontanés, communication multilingue, collecte linguistique patrimoniale, mutualisation de ressources.

Abstract:

Our effort in developing resources for network-based multilingual oral communication, (the current achievement and expected fallout of which are described in the paper), took place in parallel with the commitment of our laboratory (CLIPS-IMAG) for integrating the French language into multilingual Speech Machine Translation (SMT), within the C-STAR and NESPOLE! international projects. In the framework of two projects, ERIM (Network-based Environment for Multimodal Interpreting) and ChinFaDial (studying open architectures for SMT, and collecting French-Chinese spontaneously spoken dialogues), we have prototyped in recent years different facets of an evolving platform. They handle several aspects of spontaneous general-purpose bilingual spoken dialogues on the web: distant human interpreting, collection of spontaneous bilingual spoken dialogues corpora with translation, multimodal user interaction, integration of machine aids. Such aids experiment with server-based SMT using commercial products, and online aids to interpreters and to speakers.

The paper presents first results: ERIM-Interp and ERIM-Collect prototypes are now operational. Corpora are collected, first in French-Chinese, French-Vietnamese, French-Hindi, and should be available soon on the web (DistribDial server), with a hardened version of ERIM-Collect. Besides, all platforms should be integrated into one multifunctional ERIMM laboratory-system for enhancing multilingual multimodal distant communication, a version of which could extend to distant e-training in human interpreting.

Keywords: Web-based interpreting, bilingual spoken corpora collection, spontaneous dialogues, multilingual communication, resource mutualisation.

1. Introduction

Le développement des outils de téléphonie et de visioconférence ouverts aux transactions sur Internet suscite l'essor d'activités nouvelles de commerce électronique et de téléservice, avec dialogue libre ou de négociation, demande de renseignements, recherche d'informations, etc. Maintenant à l'oral comme

précédemment à l'écrit, le multilinguisme devient un enjeu central de ces interactions à distance avec, comme questions sensibles, un équilibre à construire entre les grandes langues véhiculaires parmi lesquelles le français, et également un réel maintien de la diversité des langues d'origine des conversants, par exemple grâce à des services d'interprétariat intermittent à distance.

Dans ce contexte, nous constatons :

- D'une part le manque notoire de grands corpus de dialogues parlés bilingues en accès libre – ressources utiles à diverses recherches sur le patrimoine linguistique, et indispensables également au développement de systèmes de Traduction Automatique de Parole (TAP),
- D'autre part l'importance de l'étude de l'impact de la multimodalité sur la communication multilingue, à l'heure où les équipements de télécommunication mobile monolingue intègrent ce type de dispositif.

Ces motivations nous ont conduits à étudier, modéliser et prototyper un ensemble de ressources génériques et de plates-formes orientées vers les aides à la communication orale multilingue et multimodale sur le web, en contribution aux attentes de l'ingénierie linguicielle.

À l'issue d'une phase de développement intensif et d'expérimentation, l'article explicite nos motivations, présente les plates-formes ERIM (Environnement Réseau pour l'Interprétariat Multimodal) maintenant disponibles, puis celles en cours de développement et en projet. Il rend compte ensuite des premières utilisations pour la collecte de dialogues parlés spontanés français-chinois, français-vietnamien, français-hindi, français-tamoul, en domaine finalisé (projets ChinFaDial, VTH-Fra.Dial), et esquisse les développements attendus en vue notamment d'une intégration sur plate-forme unique.

2. Motivations, enjeux

Le projet ERIM (Environnement Réseau pour l'Interprétariat Multimodal) répond à plusieurs lignes de motivation.

2.1 Aider la communication multilingue sur réseau

Nous souhaitions au départ proposer aux interprètes humains des scénarios innovants et de nouvelles modalités d'intervention sur le web, permettant le travail à distance (par exemple pour l'insertion professionnelle de handicapés) et facilitant les interventions ponctuelles à la demande. Ces situations nouvelles peuvent incidemment favoriser, après accord des intervenants et avec garantie d'anonymat, une collecte de grands corpus de dialogues bilingues spontanés, traduits, sur des domaines en général ciblés (au gré des situations des traductions).

Par ailleurs, l'arrivée de la multimodalité dans les équipements de télécommunication mobile (c'est à dire en communication parlée, mais enrichie d'échanges d'informations écrites brèves, d'images, de schémas, de tracés libres sur documents graphiques partagés, avec marquage dynamique des informations visuelles, désignation tactile, et prochainement analyse d'expressions faciales pour l'enrichissement de la génération et de la prosodie en traduction automatisée…), cette évolution appelle la création de plates-formes d'expérimentation sur de telles ressources, pour une étude systématique de facteurs ergonomiques et cognitifs : en d'autres termes pour évaluer l'impact du « multimodal » sur la traduction assistée de conversations spontanées bilingues, dans différents contextes d'application.

2.2 Une plate-forme pour l'ingénierie de la TA de Parole

Notre motivation est née également de l'observation des conditions de l'intégration du français, dans les projets internationaux C-STAR II et III (Consortium for Speech Translation Advanced Research, maintenant en allemand-anglais-coréen-chinois-français-italien-japonais) [sitCST], et au projet européen quadrilingue NESPOLE ! (NEgociating through SPOken Language in E-commerce) [sitNES], projets dans lesquels notre laboratoire était responsable de la traduction depuis et vers le français parlé.

Avec pour finalité une aide à l'ingénierie linguicielle (et particulièrement à l'ingénierie de la TA de Parole), nous souhaitions faciliter une expérimentation fine permettant l'évaluation comparative de différentes souches ou versions de composants de TAP, leur réglage ("tuning"), en reconnaissance de parole, traduction et synthèse, avec également l'expérimentation différentielle sur différentes ressources multimodales à intégrer ou non au cadre d'une TAP.

Il s'agissait donc dans ce contexte de créer une plate-forme générique ouverte à l'insertion (plug-in) de composants, et facilitant la capture de données.

2.3 Développer les ressources linguistiques multilingues

Nous constatons l'absence de grands corpus de parole spontanée en dialogues bilingues traduits, en français et différentes langues – corpus essentiels pour conduire des recherches linguistiques comme pour produire des modèles linguistiques des langues parlées, importants en TA de l'oral (modèles très différents de ceux des dialogues monolingues et des textes écrits).

Il est, pour chaque langue, de multiples variantes de la langue parlée spontanée, tant sont prégnants et fréquents les phénomènes élocutoires liés à l'utilisateur (faux départs, interruptions, reprises, raccourcis…), les variations de style, de traits syntaxiques particuliers (anaphores, style indirect…), les tournures orales usuelles ou idiomatiques. Il est intéressant par exemple d'étudier l'incidence de ces traits de parole spontanée sur les dialogues, leur reconnaissance et leur traduction, en situation spécifiquement multilingue.

Nous voulons donc pouvoir collecter, à coût réduit, de grands corpus de parole traduite « français - langue L », et les proposer en accès libre aux chercheurs intéressés, sur le web.

Enfin, nous souhaitons à terme mettre à disposition certains des logiciels développés (à commencer par la plate-forme de collecte), en libre accès sur le web, pour favoriser un volontariat de production de ressources linguistiques (corpus bilingues de parole spontanée, en toutes langues) partageables au sein de la communauté des chercheurs en TA de Parole ou en linguistique des langues parlées.

3. Les plates-formes ERIM

La famille des plates-formes ERIM [FafBoi03] comprend actuellement deux environnements opérationnels : ERIM-Interprète pour l'interprétariat multimodal à distance (et base des autres composants), et ERIM-Collecte pour la collecte de corpus de dialogues parlés spontanés bilingues traduits. ERIM-TA, prototype orienté vers un service de Traduction partiellement Automatique de Parole aidée par le locuteur (et plate-forme « banc d'essai de composants de TAP ») est en cours de développement avancé.

ERIM-Aides (vers des aides en ligne à la communication multilingue sur réseau) est en début de prototypage, et ERIM-Formation (pour l'e-training à distance, en interprétariat bilingue) en test de faisabilité sur la plate-forme générique de base.

3.1 ERIM-Interprète : interprétariat multimodal à distance

3.1.1 Objectifs

Il existe sur le marché des environnements « propriétaires » fonctionnant sur réseau, de type cabine d'interprète, analogues aux environnements de traduction fixes destinés aux grandes conférences multilingues (ONU, Communauté Européenne). Mais leur code source n'est pas accessible, pour des développements orientés vers la recherche.

De plus, nous envisageons des scénarios différents de ceux de l'interprétariat classique :

- « Interprétariat intermittent à la demande » : les locuteurs essaient de converser en utilisant la connaissance qu'ils ont de la langue de leur interlocuteur, ou dans une langue véhiculaire, ou connue des deux ; lorsque cette communication s'avère impraticable, ou pour des séquences « sensibles » de leur échange, ils font appel momentanément aux services d'un interprète disponible sur le web, qui peut les aider ;

- Téléconférence ("conference call") : les interlocuteurs prennent un rendez-vous avec un interprète pour un créneau de durée donnée.

Parallèlement à la modélisation et au prototypage de ressources orientées vers des services d'interprétariat à distance, un autre objectif pour cette première plate-forme est l'étude expérimentale de l'incidence, sur les dialogues bilingues ou multilingues, de diverses combinaisons de ressources multimodales. Cette approche expérimentale requiert des fonctionnalités de capture et d'enregistrement de données, à l'origine également de la variante ERIM-Collecte.

3.1.2 Conception

La plate-forme ERIM-Interprète est le socle de l'environnement ERIM, et gère la communication entre interlocuteurs. L'architecture générique, commune avec celle d'ERIM-Collecte, est présentée en figure 1 : elle comprend un serveur de communication, deux stations locuteurs, une station interprète. Les locuteurs s'adressent le plus souvent à l'interprète, mais peuvent se parler directement s'ils le souhaitent.

En multimodalité, sont actuellement possibles l'échange de textes courts, le partage sur un « tableau blanc » de documents textuels et graphiques avec pointages, marquages et soulignements libres, et la communication vidéo.

3.1.3 Avancement

L'implémentation est multi plate-forme (c'est-à-dire indépendante des systèmes d'accueil, Windows, MacOS, ultérieurement Linux), et générique.

Elle permet des configurations évolutives : le serveur de communication sur station séparée (ou non), 2 stations locuteurs (ou plus) distantes, 1 station interprète distante (ou plusieurs) ; 2 processus interprètes sont possibles sur une même station, par exemple en situation de bilinguisme avec traductions « symétriques » ; 2 processus locuteurs également, en situation de « visite » d'un locuteur à l'autre.

Des validations à longue distance sont en cours, avec étude des performances en fonction des liaisons utilisées.

3.2 ERIM-Collecte : collecte de dialogues spontanés traduits

3.2.1 Objectifs

L'importance de grands corpus réalistes est essentielle pour la construction de systèmes de TA de Parole. Ces systèmes requièrent des corpus acoustiques maintenant largement produits à partir de communications sur le web. Il est besoin également, pour établir des modèles de langues parlées en situations réelles, de corpus parallèles d'énoncés transcrits alignés. Très peu de ressources de ce type sont

produites (NEC, ATR, ELRA et quelques autres), et aucune n'est en accès libre. Pourquoi ces corpus sont-ils « propriétaires » ? Parce qu'ils sont coûteux à collecter, et encore plus à transcrire et annoter : les rendre disponibles gratuitement semble déraisonnable si on les a beaucoup « travaillés ».

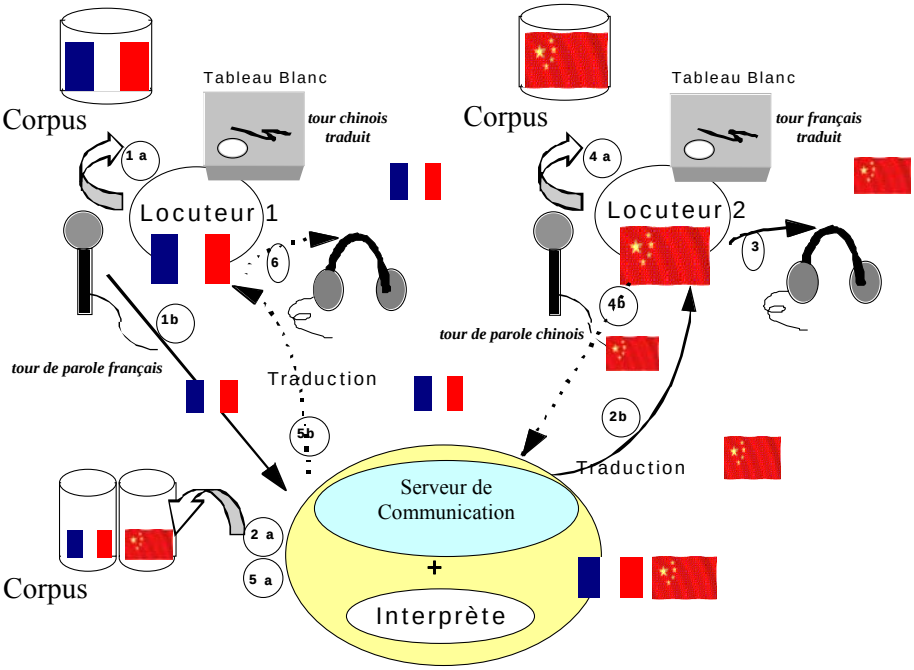


Figure 1 : ERIM-Collecte, en configuration l locuteurs et i interprètes (ici l=2, i=1)

Face à cette situation, nous avons développé ERIM-Collecte pour :

- Collecter à coût réduit, à partir de dialogues réels, des données « brutes » les plus multimodales possible, que d'autres équipes traiteront ensuite,
- Proposer à des volontaires de produire ces données gratuitement … en l'échange d'un accès libre aux plates-formes ERIM nécessaires et aux services qu'elles proposent.

3.2.2 Conception

ERIM-Collecte est une extension d'ERIM-Interprète (cf. fig. 1), avec enregistrement systématique des actes et données de l'interaction pour tous les participants (deux locuteurs ou plus, un interprète ou plus).

Dans la situation du schéma de la figure 1, l'interlocuteur français parle (1) (tour de parole en un ou plusieurs énoncés), avec enregistrement local (1a), et transmission au serveur de communication, qui les diffuse vers un salon virtuel établi pour le dialogue (1b). L'interprète écoute ce tour de parole, le traduit en chinois (2). La traduction est enregistrée localement (2a), en même temps que diffusée (2b). L'interlocuteur chinois écoute la traduction (3), puis répond (4). De nouveau, son tour de parole est enregistré localement (4a) et diffusé (4b). En (5), l'interprète le traduit en français ; la traduction est enregistrée localement (5a), et transmise (5b) au destinataire (6).

L'enregistrement est fait localement lors de la conversation. En fin de dialogue, les descripteurs et fichiers produits localement sont transmis à un serveur de collecte, où ils sont regroupés et structurés.

3.2.3 Avancement

Sur la version actuelle ERIM/3-Collecte, sont enregistrés tours de parole et textes courts (bimodalité). La capture d'évènements du tableau blanc et des objets impliqués (fichiers visuels partagés, tracés libres, url…) est en cours d'intégration, comme celle de la vidéo, si celle-ci est souhaitée (par exemple pour des études ultérieures d'expressions faciales). La figure 2 présente l'interface actuelle de la plate-forme, pour un locuteur.

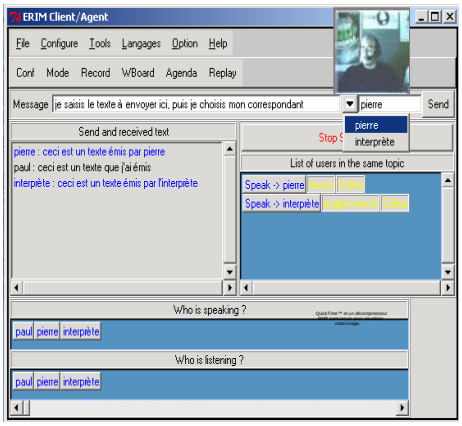


Figure 2 : Écran Locuteur

Une ressource de réécoute sélective, le module « Replay », permet de reconstituer tout ou partie du dialogue, chronologiquement ou avec extraction de versions monolingues. Il facilite un suivi visuel de l'échange (cf. fig. 3). Il fonctionne également en version bimodale (tours de parole, échange de textes courts), et une extension est en cours au tableau blanc, pour des corpus comprenant la collecte des événements de multimodalité. Il sera accessible sur un site web (DistribDial) d'accès aux corpus produits, pour des utilisateurs de la communauté scientifique.

En revanche ERIM-Collecte, destinée par choix initial à la constitution de corpus de données brutes (fichiers descripteurs de session et tours de parole) ne propose pas de ressource particulière d'aide intégrée à la transcription, ni à l'annotation. Ces dernières sont réalisables hors ERIM.

L'architecture ouverte de la plate-forme générique facilitera des développements futurs. Nous prévoyons d'utiliser par exemple la ou les reconnaissances vocales intégrées d'ERIM-TA (cf. 3.3) pour produire des pré-transcriptions en version instantanée brute, à la mesure certes de la qualité des composants de reconnaissance disponibles.

Les versions successives d'ERIM-Collecte ont été utilisées, d'abord à Grenoble et à Pékin, puis récemment à Grenoble, DaNang et Bombay, pour la collecte de premiers corpus de dialogues spontanés traduits (en réservation hôtelière).

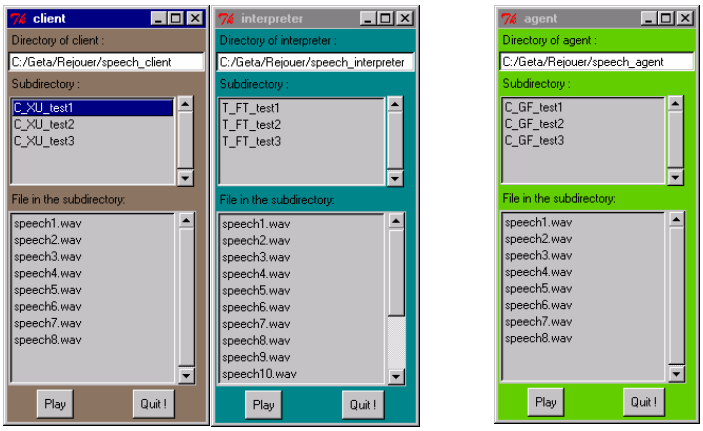


Figure 3 : Réécoute des énoncés Locuteurs, et Interprète (au centre)

3.3 ERIM-TA : Traduction partiellement Automatique de Parole, en interaction avec le locuteur

3.3.1 Objectifs

Pour cette variante, l'objectif était l'intégration générique, par plug-in, de séries de composants de TAP issus de la recherche académique (Reconnaissance, Traduction, Synthèse), pour leur mise au point en évaluation comparative, contrastive, avec la production « humaine » d'un interprète.

L'enjeu global est pour nous une « Traduction partiellement Automatique de qualité », de parole (« tchat » parlé) et de texte (de type SMS), sans recours à un interprète humain, mais en introduisant un niveau adéquat de contrôle par l'utilisateur (système synergique de TAP).

3.3.2 Conception

Pour que les couvertures lexicale et grammaticale soient assez larges, nous avons choisi dans un premier temps d'intégrer des composants commerciaux (de reconnaissance de parole, de traduction, et de synthèse de parole) tous accessibles sur serveur. Ce sont, dans le premier prototype, des produits disponibles respectivement chez Philips, Linguatec, et ScanSoft.

3.3.3 Avancement

Un premier prototypage est effectué, qui reste à intégrer au socle ERIM. Le locuteur émet un énoncé de parole en langue source, et l'énoncé textuel « reconnu » lui est soumis pour validation ; il pourra si nécessaire re-soumettre son énoncé vocalement (ou s'il préfère, modifier au clavier l'énoncé textuel proposé). Après validation, le composant de traduction produit un texte en langue cible, ensuite synthétisé par la synthèse vocale.

Dans certains cas (par exemple d'un locuteur émetteur partiellement bilingue), cette traduction écrite peut être renvoyée ; une rétro-traduction peut aussi être lui soumise pour validation, et jouer un rôle d'indicateur, certes imparfait, d'une non-pertinence de la traduction. En cas de réelle difficulté, le locuteur peut reformuler oralement (voire au clavier) son énoncé en langue source, pour contourner une possible inaptitude de la TA. Une validation provoquera la synthèse vocale.

Si une reconnaissance vocale s'avère fiable, elle peut fournir une aide instantanée, suggérée d'ailleurs par certains interprètes professionnels (cf. 5.1), pour un rappel ciblé ou un éclairage ponctuel ("flash check") de termes ou d'énoncés particuliers utilisés par un locuteur dans un des derniers tours de parole.

ERIM-TA enregistrera toutes les productions (de reconnaissance, de traduction et rétro-traduction, et de synthèse), et toute reformulation.

Le paradigme d'une TA de Parole « aidée par l'utilisateur » semble a priori se prêter à des utilisations de « tchat » oral multilingue de qualité, et plus globalement de communication multilingue avec traduction de qualité, notre objectif final.

4. Collecte et partage de corpus de dialogues spontanés traduits

4.1 ChinFaDial : corpus de dialogues français-chinois

ERIM-Collecte a été utilisé dans le cadre du projet LIAMA ChinFaDial [FaBoSeZo04], en partenariat avec le NLPR (National Laboratory for Pattern Recognition — Institut d'Automatique de l'Académie des Sciences de Chine, à Pékin), pour la collecte de dialogues parlés bilingues français-chinois, spontanés et traduits, entre deux locuteurs monolingues (cf. fig. 4 - la transcription, ultérieure, est ici manuelle).

顾客/Client (7)
我在火车站; 火车站离你们旅馆不知道远不远、怎么走?
(Je suis à la gare, je ne sais pas comment me rendre à l'hotel à partir de la gare.)

Agent/代理（7）
Alors c'est extrèmement simple,en sortant de la gare vous tournez à droite et c'est à 80 mètres en face de l'autre côté de la place.
（很简单、如果你出了火车站以后向右转、只要走到80米左右。就是我们的旅馆。）

顾客/Client (8)
好谢谢那就一会儿见
(Merci bien, alors à tout à l'heure)

Agent/代理（8）
Merci, bonsoir Monsieur,à tout à l'heure.
（谢谢…谢谢、那么、这是…一会见）

Figure 4 : Dialogue français-chinois (extrait), entre hôtelier français et client chinois

La plupart des collectes utilisent pour le moment les réseaux locaux des laboratoires partenaires, avec trois participants dans un même bâtiment ou des bâtiments proches. La collecte à longue distance est en cours de validation.

Les situations sont réelles (réservation hôtelière), en dialogue spontané. Une dizaine d'heures de dialogues oraux ont été enregistrées lors de sessions au CLIPS et au NLPR, et constituent un premier corpus de données brutes, non transcrites.

Les collectes se poursuivent. D'autres corpus sont collectés, dans des régions francophones des pays concernés, en français-vienamien, français-hindi (Center for Indian Languages Technology, à l'IIT-Bombay) et français-tamoul (Pondichéry), dans le cadre du projet VTH-Fra.Dial soutenu par l'AUF (réseau LTT). Des actions complémentaires pourraient concerner la collecte contrastive de parlers régionaux, avec éventuelle médiation d'un interprète.

4.2 Vers la création collaborative de ressources linguistiques

Nous souhaitons contribuer à promouvoir des actions de mutualisation de ressources pour l'ingénierie de la TA de Parole [Fafi04] : corpus parallèles de dialogues parlés spontanés, à collecter, puis à transcrire et annoter, en participation collaborative.

Les corpus déjà collectés, de dialogues parlés spontanés bilingues traduits, devraient être disponibles en 2005 sur un site du GETA-CLIPS, pour des recherches en linguistique, ou ultérieurement en didactique des langues, et pour alimenter l'ingénierie de la TA de Parole.

Ce site et l'environnement DistribDial devront faciliter l'enrichissement libre des corpus par d'autres chercheurs, par ajout de transcriptions et/ou d'annotations en fichiers parallèles (selon un format homogène) —à rendre accessibles également sur le web.

Les deux plates-formes ERIM-Interprète (d'aide à l'interprétariat humain à distance sur le web), et ERIM-Collecte (pour la collecte de corpus de dialogues parlés bilingues traduits), seront mises à disposition en accès libre sur un site web, après des tests complémentaires de robustesse et d'utilisabilité, pour une utilisation collaborative ou pour toute activité de recherche ou de préservation de patrimoine linguistique.

5. Développements ultérieurs

5.1 ERIM-Aides : aides en ligne à la communication multilingue sur réseau

5.1.1 Objectifs

Dans notre scénario d'interprétariat intermittent à la demande, il est proposé aux interprètes de passer d'une conversation à une autre et d'un sujet à un autre (à la manière d'interprètes dits de cocktail). C'est une activité difficile, pouvant créer un réel stress lexical. Des aides automatiques en ligne seraient bienvenues, par exemple avant intervention, pour une mini-immersion lexicale prenant en compte la

spécificité du domaine concerné ou un éventuel « contexte lexical » lié à des utilisateurs réguliers ; ou encore après intervention pour la prise de notes personnelles ou contextualisées.

L'étude d'une éventuelle aide instantanée « pendant » l'activité de traduction, et une expérimentation en situation avec la plate-forme ERIM-Aides, devront prendre en compte les surcharges instantanées, perceptives, cognitives, mnésiques, qui caractérisent l'activité d'un interprète. Des interprètes professionnels ont suggéré la consultation possible d'un terme particulier, à la volée, à pointer dans la transcription instantanée du dernier ou de l'avant-dernier tour de parole.

On souhaite également pouvoir disposer d'aides automatiques pour les locuteurs (par exemple faiblement ou partiellement bilingues, ce qui concrètement est parfois le cas), pour les aider à se débrouiller sans interprète, en cas de nécessité.

5.1.2 Conception

Un premier type d'aide « à la communication » comprend des ressources :

- Pour se voir en se parlant (locuteurs, interprète),
- Pour partager des données, éventuellement modifiables, pointables ou marquables (tableau blanc, visuels partagés…), consultables a posteriori,
- Pour planifier et gérer des rendez-vous avec un agenda sur site serveur.

D'autres aides, de type « linguistique », peuvent inclure :

- L'accès à des fiches terminologiques thématiques bilingues, à des dictionnaires électroniques à saisie au clavier ou à entrée vocale, par exemple par détection de mots automatique (word-spotting) suivi d'un filtrage, d'une recherche en dictionnaire, et présentation de synthèse sur fenêtre unique,
- Une reconnaissance de parole atténuant les difficultés de compréhension orale en l'absence d'interprète, ou produisant une trace ou un historique de la conversation, que peut également consulter l'interprète avant intervention,
- De la TA de Parole partiellement automatique, avec à terme une désambiguïsation interactive à la LIDIA [BoiBlan95, Blan04].

5.1.3 Avancement

Les aides de communication sont prototypées. Dans une première version l'agenda est global sur un site d'accès ERIM, chaque utilisateur y accédant via une vue personnalisée.

Les premières aides linguistiques vont être prototypées en interfaçant des ressources dictionnairiques existantes, en accès libre sur le site Papillon [sitPAP].

Une ressource de reconnaissance vocale (Philips), déjà interfacée avec ERIM-TA, pourra être intégrée, par exemple pour consultation de la transcription instantanée de termes utilisés dans les derniers tours de parole.

5.2 ERIM-Formation : e-training en interprétariat, sur le web

Une variante ERIM-Formation prévoit de proposer, à des étudiants en interprétariat bilingue ou à des interprètes en perfectionnement, différents modes de formation à distance (FAD) sur le web, de type e-training, pour une activité en ligne « en direct » ("live"), ou « de doublage ». L'architecture générique des plates-formes actuelles permettra rapidement de simuler son fonctionnement, et d'effectuer des tests de faisabilité et d'utilité, en configuration multi-interprètes. Ce dispositif peut favoriser également l'investigation de nouvelles activités d'Apprentissage des Langues à Distance.

Nous prévoyons un dispositif de type « laboratoire de langues sur le web, pour interprètes », avec par exemple un interprète professionnel (et/ou professeur) assurant la continuité et la fluidité du dialogue bilingue entre deux locuteurs monolingues sur le réseau ; bénéficiant de cette situation, un cercle d'interprètes « juniors » peuvent (sans entendre les traductions du professionnel) s'exercer à distance, ou éventuellement intervenir chacun à son tour sur proposition d'un médiateur de traduction. Toute intervention est enregistrable (avec l'accord des intervenants) et consultable, et un travail pédagogique intéressant peut s'ensuivre.

Cette ressource rejoint de fait l'approche collaborative que privilégie le projet ERIM, car il est attendu de la part d'étudiants avancés en interprétariat ou d'interprètes en perfectionnement, qu'ils acceptent de se porter volontaires, pour coopérer à l'activité collectrice sur ce type d'outil « d'apprentissage à distance par la pratique », contribuant à la création peu coûteuse de corpus de dialogues spontanés, rendus anonymes (cf. 4.2).

Il est possible également d'envisager des situations d'utilisation plus institutionnelles, où des étudiants « seniors » en interprétariat acceptent d'assurer des traductions bilingues dans le cadre d'un service multilingue d'interprétariat « grand public », bénévolement ou en échange d'une validation académique de cette activité (par exemple aux Jeux Olympiques de Beijing 08).

6. Évolution de la plate-forme, prospective

L'environnement ERIM s'est construit par prototypage exploratoire puis développement incrémentiel de plusieurs classes de ressources complémentaires, en cohérence fonctionnelle. Il devrait nous permettre d'explorer plus directement des situations réalistes d'usage de systèmes de TAP futurs, tels que nous les concevons : médiatisés et synergiques, intégrant des ressources de traduction « par la machine » (cf. 3.3) et « aidée par l'utilisateur », ce dernier disposant d'aides en ligne (cf. 5.1).

Parallèlement à la diffusion d'une plate-forme à usage ciblé (pour la collecte de données brutes, puis leurs enrichissements contributifs), nous effectuerons des évaluations d'ERIM-Interprète et ERIM-Collecte chacune en situation d'utilisation réelle soutenue. Elles permettront d'ajuster la spécification et le développement de ERIM-Aides, ERIM-TA, et également ERIM-Formation.

Nous souhaitons également une unification et une intégration des différents composants d'ERIM ici présentés, en une « plate-forme laboratoire » unique ERIMM (Environnement Réseau pour l'Interprétariat Multilingue Multimodal), regroupant un ensemble d'aides à la communication multilingue sur réseau.

7. Conclusion

Dans notre recherche en conception et production d'outils et de ressources pour la communication orale multilingue sur le web, et parallèlement à une action d'intégration du français en Traduction Automatique (TA) de Parole multilingue (projet international C-STAR en sept langues, projet européen quadrilingue NESPOLE!), nous avons développé plusieurs plates-formes génériques en choisissant dans un premier temps une approche duale, complémentaire de celle de la TA : centrée sur l'aide de la machine aux interprètes humains et aux locuteurs. Ce développement s'est fait dans le cadre des projets ERIM (Environnement Réseau pour l'Interprétariat Multimodal) et ChinFaDial (conception et prototypage de systèmes de TA de Parole, avec collecte de corpus parlés spontanés français-chinois).

Nous avons présenté trois plates-formes dont deux sont opérationnelles et la dernière en intégration finale, qui permettent d'aider la communication bilingue sur le web, pour des dialogues spontanés a priori non finalisés : ce sont ERIM-Interprète pour l'interprétariat humain sur réseau, et ERIM-Collecte, qui réduit le manque de données utiles pour développer de meilleurs systèmes de TA de Parole ; ERIM-TA, ressource générique destinée à la Traduction partiellement Automatique de Parole (TpAP) de qualité, utilise dans un premier temps des produits logiciels, de reconnaissance, de traduction et de synthèse, à couverture large et disponibles sur serveurs, avec contrôle par l'utilisateur (feedback, validation directe). ERIM-TA assiste aussi l'ingénierie des systèmes de TA de Parole, et constitue un banc d'essai générique en situation réelle, ou de réglage fin, pour leurs composants.

Une plate-forme ERIM-Aides, proposant aux interprètes et locuteurs des aides en ligne (aides de communication, aides lexicales), est en début de réalisation.

De premiers corpus « bruts » de dialogues bilingues parlés, spontanés et traduits, ont été collectés en français-chinois et français-vietnamien, en domaine finalisé (réservation hôtelière), et d'autres collectes suivent dans ces langues.

Ces corpus seront accessibles sur un site web du CLIPS, pour des transcriptions, annotations, ou recherches linguistiques, et nous souhaitons inciter tous les contributeurs intéressés à participer à des créations collaboratives de

ressources linguistiques, dans différents couples de langues, dont le français en langue pivot.

Notre recherche se poursuit par la collecte et la distribution de données concernant d'autres langues (français-hindi, français-tamoul), par l'enrichissement fonctionnel des quatre premières plates-formes (notamment en multimodalité et production de corpus multimodaux), et le renforcement de leur « robustesse » logicielle dans des situations ou avec des logistiques d'utilisation difficiles.

Nous préparons leur unification en un environnement ERIMM, plate-forme « laboratoire » de collecte, d'expérimentation et d'aide à l'ingénierie de systèmes de Traduction partiellement Automatique de Parole multilingue et multimodale.

Nous visons enfin le développement d'un « labo de langue sur le web pour l'interprétariat », ERIM-Formation, qui pourrait également contribuer aux collectes.

Il nous apparaît que de tels outils (plates-formes d'interprétariat, de collecte, de formation) et leur mise à disposition sur des sites en téléaccès libre, au delà du développement synergique qu'ils permettent – celui de services de communication multilingue multimodale sur le web, notamment entre langues peu dotées et grandes langues véhiculaires –, pourraient efficacement sous-tendre d'une part des efforts de production de ressources linguistiques (documents multimédia multilingues et multimodaux) en vue de recherches fondamentales et d'actions de préservation de patrimoines linguistiques, et d'autre part des actions de développement de ressources humaines – par exemple en favorisant l'apprentissage de l'interprétariat bilingue à distance, ou en facilitant la découverte ou la pratique des langues sur le web.

Remerciements

Ces travaux ont été soutenus par le CLIPS-IMAG (Université Joseph Fourier - Grenoble 1, INPG, CNRS), par la Région Rhône-Alpes (projet ERIM), le laboratoire franco-chinois LIAMA (projet ChinFaDial), et le réseau LTT de l'AUF (projet VTH-Fra.Dial).

Zhai JianShe (Université de Nankin, Chine) en résidence au CLIPS, puis Julien Lamboley (INSA de Lyon) ont contribué de manière essentielle au développement des plates-formes. Qu'ils soient ici remerciés, ainsi que les collègues du CLIPS et de MultiCom (à Grenoble), du National Laboratory for Pattern Recognition (CAS-IA à Pékin, Chine), de l'Université de Danang (Vietnam) et de l'IIT-Bombay (Inde), qui ont participé aux expérimentations et collectes.

8. Références

[Blan04] Blanchon H. (2004), Comment définir, mesurer et améliorer la qualité, l'utilisabilité et l'utilité des systèmes de TAO de l'écrit et de l'oral – une bataille contre le bruit, l'ambiguïté, et le manque de contexte. *HDR UJF, 20/12/04*, 320 p.

[BoiBlan95]Boitet C., Blanchon H. (1994), Multilingual Dialogue-Based MT for Monolingual Authors: the LIDIA Project and a First Mockup. *Machine Translation, vol. 9(2)*, p. 99-132.

[Fafi04] Fafiotte G. (2004), Building and sharing multilingual speech resources, using ERIM generic platforms. *COLING-MLR 2004, Geneva, 28/08/04*, 8 p.

[FafBoi94] Fafiotte G., Boitet C. (1994), Report on first EMMI Experiments for the MIDDIM project in the context of Interpreting Telecommunications. *MIDDIM report TR-IT-0074 GETA-IMAG & ATR-ITL*, Aug. 1994, 11 p.

[FafBoi03] Fafiotte G., Boitet C. (2003), ERIMM, a platform for supporting and collecting multimodal spontaneous bilingual dialogues. *IEEE NLP-KE2003, Beijing, 26-29/10/ 2003*, 6 p.

[FaBoSeZo04] Fafiotte G., Boitet C., Seligman M., Zong C.-Q. (2004), Collecting and Sharing Spontaneous Speech Corpora: the ChinFaDial Experiment. *LREC 2004, Lisboa, 26-285/04*, 8 p.

[FafZhai99] Fafiotte G., Zhai J.-S. (1999), A Network-based Simulator for Speech Translation. *NPLRS'99, Beijing, 5-7/11/99*, p. 511-514.

[LokYatMor94] Loken-Kim K.-H., Yato F., Morimoto T. (1994) A Simulation Environment for Multimodal Interpreting Telecommunications. *IPSJ-AV workshop, March 1994*, 5 p.

[sitCST] Site web C-STAR: http://www.c-star.org

[sitNES] Site web NESPOLE! : http://nespole.itc.it

[sitPAP] Site web PAPILLON: http://www.papillon-dictionary.org

Session 2

Recherche et extraction d'information

Acquisition et comparaison en ligne de l'écriture d'enfants bilingues

Iyad Zaarour[1], Zekhnine Saliha[2], Laurent Heutte[3], Daniel Mellier[2]

[1] *Laboratoire LPM – Université Libanaise - Liban*
ayadzcij@hotmail.com

[2] *Laboratoire Psy-co – Université de Rouen - France*
daniel.mellier@univ-rouen.fr

[3] *Laboratoire PSI, CNRS FRE 2645 – Université de Rouen*
laurent.heutte@univ-rouen.fr

Résumé :

La présente communication envisage la situation créée par le bilinguisme arabe français pour étudier dans quelle mesure la planification motrice et son contrôle s'appliquent indifféremment aux deux langues ou bien est instanciée différentiellement selon le contexte linguistique, allographique et spatial.

Mots-clés : écriture, bilinguisme, modélisation des connaissances, sciences cognitives.

1. Introduction

Bien que les troubles graphomoteurs de l'écriture n'entrent pas dans les nomenclatures nosographiques internationales des troubles du développement, ils sont considérés d'une part comme indice des troubles de l'expression écrite rédactionnelle (le writing anglais) et sont, d'autre part, associés aux troubles de l'attention, à l'épilepsie et aux troubles neurologiques consécutifs à des lésions cérébrales.

Considérant le potentiel invalidant de la dysgraphie pour les apprentissages scolaires quand la formation graphique est trop lente ou mobilise exagérément les ressources attentionnelles, il est utile d'améliorer les modalités d'évaluation clinique qui s'avèrent inconsistantes ou insuffisamment détaillées. En effet, l'écriture est examinée soit comme un produit fixe en évaluant la lisibilité, la régularité, la spatialité d'un document figé (Ajuriaguerra, 1989, 1990 ; Charles, Soppelsa, Albaret, 2003) soit comme une action procédurale mobilisant une suite de connaissances et de processus représentées dans les variables dynamiques du tracé. Les deux approches, qui sont rarement associées, sont nécessaires pour connaître les processus qui induisent le défaut de lisibilité, admis en priorité comme signe d'appel principal des troubles de l'écrit.

Le modèle de la formation graphomotrice des lettres et mots proposé par Zesiger (2003) (cf. figure 1) permet de faire l'hypothèse que la formation graphique peut être principalement perturbée au titre d'un déficit de la planification - programmation motrice (Wann, 1986) ou d'un déficit de l'exécution motrice (Smits Engelsman et Van Galen, 1997).

Le défaut de planification du geste entretient la dysfluence (discontinuité du mouvement), la lenteur, la longueur des pauses et la sur-utilisation de la rétroaction visuelle. Le déficit d'exécution se manifeste quant à lui par l'irrégularité spatiale, temporelle et cinématique avec sur-représentation de mouvements de haute fréquence. Ces deux défauts ne sont pas exclusifs.

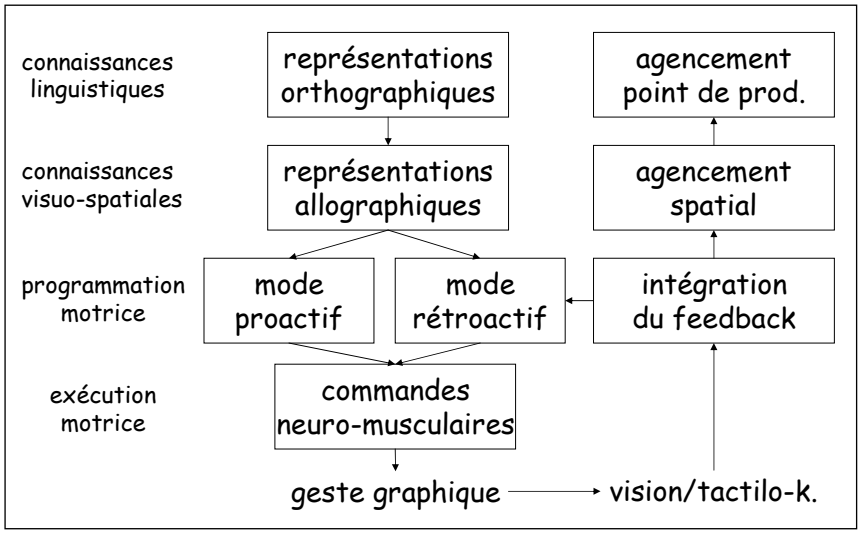


Figure 1 : Modèle de l'écriture selon (Zesiger, 2003)

Considérant l'intérêt d'utiliser l'apport des recherches en reconnaissance de formes pour automatiser les examens cliniques, le laboratoire PSI et le laboratoire Psy-Co (Université de Rouen) se sont associés pour mettre au point des outils d'acquisition et d'analyse du graphisme des enfants d'âge scolaire. Le programme de recherche a permis d'abord d'isoler les primitives à extraire (Remi, 2002), puis d'optimiser la catégorisation des profils d'écriture (Zaarour *et al.* 2003, Zaarour *et al.* 2004)). Il est poursuivi par une recherche Interreg III avec l'Université de Kent pour mettre au point un outil diagnostic directement utilisable en consultation et applicable au suivi rééducatif (http://www.meddraw.org).

La présente communication envisage la situation créée par le bilinguisme arabe français pour étudier dans quelle mesure la planification motrice et son contrôle s'appliquent indifféremment aux deux langues ou bien est instanciée différentiellement selon le contexte linguistique, allographique et spatial. L'étude consiste à recueillir le graphisme de figures géométriques puis d'écriture d'enfants normalement scolarisés et exposés aux deux langues orales et écrites. Un autre protocole, encore en cours de réalisation, compare la fluence quand le scripteur écrit du texte dans chacune des deux langues.

La première partie de l'étude a consisté à décrire les stratégies motrices d'écriture d'enfants bilingues arabe-français en adoptant un modèle graphique probabiliste (i.e. les réseaux bayésiens) et en repliquant les analyses développées par Zaarour *et al.* 2004. Nous décrivons le protocole expérimental et les premiers résultats obtenus dans les sections suivantes.

2. Procédure de recueil des observations

Les scripteurs sont invités à tracer des figures géométriques extraites du « test de structuration visuelle » de Bender, puis à copier un texte en français puis en arabe pendant 5 minutes. Les tracés sont enregistrés sur Tablet PC ACER 2003 et traités automatiquement par un logiciel permettant l'acquisition en ligne de tracés manuscrits (dessins et écriture) et l'extraction de primitives. (Amara, 1997 ; Remi 2002).

Plus précisément, le protocole expérimental comprend six tests : deux dessins issus de l'épreuve de Bender consistant au tracé d'une figure contenant un cercle et un carré et une composée de pennures de flèches alignées en frise ; l'écriture du nom et du prénom ; l'écriture d'un mot isolé ; l'écriture d'une phrase en copie et en mémoire. Les élèves fréquentent les cinq premières classes primaires. Ils sont bilingues, n'ont redoublé aucune classe, n'ont pas de trouble avéré du langage oral ou écrit et ont un âge conforme à leur niveau scolaire. Ils ont été observés trois fois à 6 mois d'intervalle afin d'évaluer la stabilité de leur stratégie d'écriture.

Le calcul des variations de vitesse, d'accélération ; des types de segmentation des tracés ; de leur organisation dans l'espace de la feuille (orientation, linéarité), autorise à décrire les règles d'action qui président à la planification-programmation

graphique. La planification est évaluée par les variations de vitesse, le nombre de pauses. L'observation est complétée par un relevé en parallèle du nombre de consultations visuelles manifestes du modèle.

La qualité de l'exécution est donnée par les indices de régularité spatiale et temporelle des tracés. Elle est parallèlement estimée en termes de lisibilité par des juges indépendants et non informés de l'étude. Ces derniers cotent la régularité, l'aisance à lire et la ressemblance des tracés dans les deux langues.

Le taux de concordance calculé entre les indices donnés par l'analyse automatique et l'estimation des juges permet d'assurer la fiabilité de l'analyse en temps réel.

3. Modélisation

Le recours à des modèles graphiques probabilistes, plus spécifiquement les réseaux bayésiens, est un moyen de représentation des connaissances permettant de préciser graphiquement les dépendances probabilistes entre les propositions et les événements (Pearl, 1988). Dans le contexte de notre étude, il est utilisé comme un outil d'apprentissage qui cherche à établir un modèle de classification en se basant sur la théorie des probabilités. La phase suivante consiste à étiqueter les stratégies globales les plus fréquentes, puis à évaluer l'effet de la variable stratégie sur chacun des tests.

Une stratégie d'écriture est donc représentée par une catégorie de variables cachées et non mesurables. Nous construisons un modèle global hiérarchique pour lier les stratégies locales aux stratégies globales et modéliser la dépendance probabiliste entre variables et stratégies. Ce modèle hiérarchique, appris à partir des données réelles recueillies, permet de décrire les stratégies des élèves bilingues tout en autorisant une modélisation temporelle et dynamique des stratégies d'écriture motrices indépendantes de l'âge scolaire.

Considérant que nous représentons une stratégie d'écriture par une variable cachée, l'ordre de grandeur de la valeur maximale d'une telle variable est détecté grâce aux k-means. La structure du réseau (i.e. les relations de dépendance directe entre les variables) est partiellement définie par les experts (cf. figure 2). L'apprentissage des paramètres (i.e. les probabilités conditionnelles) du réseau est assuré par l'algorithme Expectation-Maximisation (Dempster *et al.* 1977). A partir de l'inférence bayésienne, nous effectuons une classification automatique, nous calculons ensuite la distribution de probabilités de chaque variable cachée afin de pouvoir étiqueter les clusters obtenus, puis nous étudions l'évolution temporelle de ces clusters en fonction du niveau scolaire.

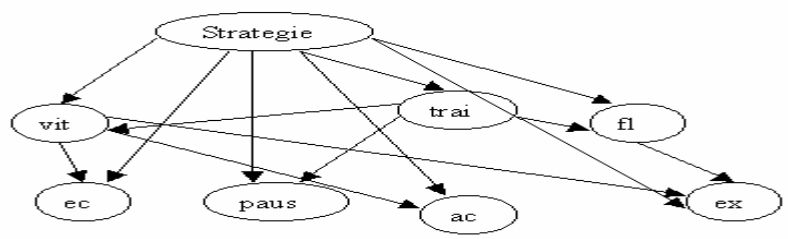


Figure 2 : Modèle générique

4. Classification

Nous avons construit un modèle Global Hiérarchie, où chacune des épreuves de dessin ou d'écriture est liée à une variable cachée qui représente la stratégie locale d'écriture de ce test (cf. figure 3).

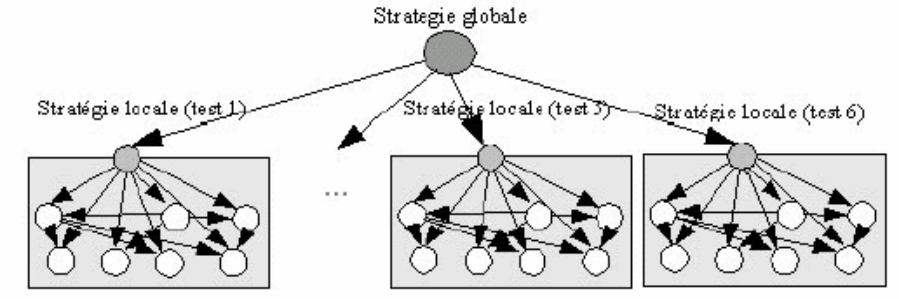


Figure 3 : Modèle hiérarchique global

Après avoir estimé que le nombre d'états de la stratégie globale est égal à 4, nous avons procédé à une classification (clustering) de tout notre modèle. Nous opérons ensuite test par test en comparant pour chaque test les distributions de probabilité des caractéristiques avec les différents états de la variable cachée. Nous calculons ensuite la loi marginale de chacun des états des variables cachées (i.e. stratégies).

La figure 4 présente ainsi la distribution de probabilité de la caractéristique "Accélération" en fonction de la stratégie locale dédiée au test d'écriture du nom.

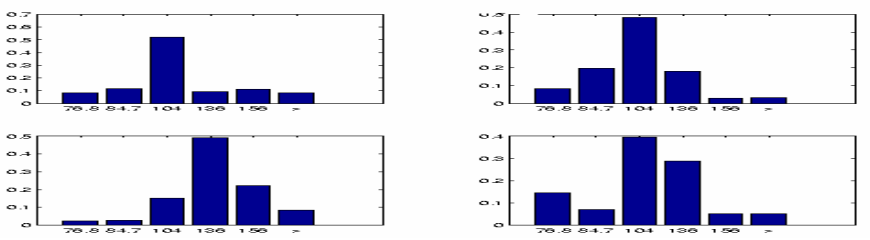


Figure 4 : Test du Nom : Prb marj(Accélération|stratégie locale=1,..,4)

Le calcul de la probabilité de la stratégie globale ayant donné une stratégie locale extrait, pour chaque test, deux stratégies locales qui regroupent 86 % de la population. Ces deux stratégies locales liées aux deux stratégies globales "Glob1" et "Glob4", (cf. figure 5), correspondent pour les psychologues, aux praxies d'écriture des élèves normo-scripteurs (NS, "Glob4"), et à celles des élèves normo-scripteurs plus avancés (NSA, "Glob1").

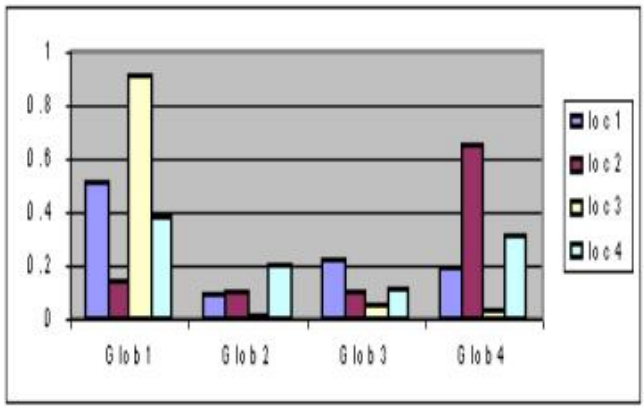


Figure 5 : Prb{Globale|Locale) - Test du Nom

5. Conclusion

Cette première analyse des données a exploité la puissance du réseau bayésien, et surtout de l'inférence bayésienne pour prendre en compte l'incertain et la variabilité en ne modélisant pas la valeur d'une caractéristique mais sa distribution de probabilité. Les groupes découverts constituent un tout cohérent, étiquetable et informatif au niveau des caractéristiques d'écriture. De point de vue cognitif, chaque stratégie locale évoque des compétences particulières au niveau des tests d'écritures proposés. L'intégration d'une variable cachée globale agissant sur les variables locales a pour but d'étudier l'effet d'un aspect global qui agit sur les stratégies locales et qui pourrait être le fruit d'une représentation centrale spécifique. Il apparaît que les états "Glob2" et "Glob3" de la stratégie globale sont des états intermédiaires qui se différencient par la vitesse d'écriture.

6. Références bibliographiques

Ajuriaguerra, J. de, Auzias, M., Denner, A. (1990), La rééducation de l'écriture, Neuchâtel et Paris, Delachaux et Niestlé.

Ajuriaguerra, J. de, Auzias, M., Denner, A. (1989), L'écriture de l'enfant, tome I : L'évolution de l'écriture et ses difficultés, Neuchâtel et Paris, Delachaux et Niestlé,

Amara, M., Courtellement, P., Brucq, D. de, Devinoy, R. (1997), A Software Tool for the Analysis of Handwriting: Writing and drawing Application", IGS'97, Italy, p. 107-108.

Charles, M., Soppelsa, R., Albaret, J.M. (2003), BHK - Echelle d'évaluation rapide de l'écriture chez l'enfant, Paris, Editions EAP.

Dempster, A., Laird, N., Rubin, D. (1977), Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, B 39: 1-38.

Pearl, J. (1988), In: Probabilistic reasoning in intelligent systems: Networks of plausible inference, Brachman (Ed), Morgan Kaufmann.

Rémi, C., Frélicot, C., Courtellemont, P. (2002), Automatic analysis of the structuring of children's drawing and writing, Pattern Recognition 35, p. 1059-1069.

Smits-Engelsman, B.C.M., & Van Galen, G.P. (1997), Dysgraphia in children: Lasting psychomotor deficiency or transient developmental delay?, Journal of Experimental Child Psychology, 67 (2), p. 164-184.

Wann, J.P. (1986), Handwriting disturbance: developmental trends. In H.T.A, Whiting & M.G. Wade (Eds), Themes in motor development, Dordretcht, Martinus Nijhoff Publishers, p.207-223.

Zaarour, I., Heutte, L., Eter, B., Labiche, J., Mellier, D., Zoaeter, M. (2003), Une modélisation probabiliste de l'évolution des stratégies d'écriture des élèves en scolarité primaire, Premier Congrès International sur les Méthodes Numériques Appliquées, CIMNA'2003, Beyrouth, Liban, novembre 2003.

Zaarour, I., Heutte, L., Leray, P., Labiche, J., Eter, B., Mellier, D. (2004), Clustering and bayesian network approaches for discovering handwriting strategies of primary school

children, International Journal of Pattern Recognition and Artificial Intelligence, 18(7), p. 1233-1252.

Zesiger, P. (2003) ; Acquisition et troubles de l'écriture ; Enfance, 1, p. 56-64.

Réflexion sur un outil d'aide à l'interprétation des besoins dans le domaine du conseil en systèmes d'information

Sabrina Boulesnane, Laïd Bouzidi,

[1] *SICOMOR – Centre de Recherche de l'IAE – Université Jean Moulin – Lyon 3*
6 cours Albert Thomas – 69008 Lyon - France

boulesnane_sabrina@yahoo.fr
bouzidi@univ-lyon3.fr

Résumé :

L'hétérogénéité des profils des différents acteurs d'une entreprise rend de plus en plus difficile l'identification exacte d'une demande de service. Ceci est d'autant plus vrai lorsqu'il s'agit d'une entreprise spécialisée dans le domaine de l'audit et du conseil en systèmes d'information. L'activité de conseil couvre un spectre très large touchant plusieurs domaines et spécialités et faisant intervenir plusieurs acteurs ayant des profils différents. Autrement dit, le problème qui se pose ne se situe pas au niveau des services proposés et offerts par ces cabinets de conseil, mais plutôt dans l'expression des besoins formulés par les PME et l'interprétation faite par les conseillers. La formulation des requêtes est en règle générale biaisée par plusieurs facteurs, nous citons les profils des personnes chargées d'identifier et de décrire le service demandé. Dans cet article, nous allons nous attacher à définir une approche nous permettant de nous orienter vers des outils susceptibles de proposer une aide aux cabinets de conseil pour faciliter l'appréhension des dites requêtes. Nous adoptons pour cela une démarche méthodologique orientée usager, nommée « approche tridimensionnelle ». Les trois dimensions qui constituent le socle de cette approche sont : la dimension cadre ou activité, dans laquelle est décrite toute l'activité en mettant en relief les problèmes réels qui en découlent. La dimension humaine permettant l'identification et l'analyse des profils des différents acteurs qui interviennent dans le processus tant au niveau l'expression des besoins qu'au niveau de leurs interprétation. La dimension technique permettant d'identifier les outils de représentation de l'information. Ces trois dimensions sont précédées d'une présentation du contexte dans lequel notre étude du terrain a été réalisée.

Mots-clés : profil des utilisateurs, requête, audit et conseil en système d'information, expression des besoins, l'approche tridimensionnelle, dimension cadre, dimension humaine, dimension technique.

1. Introduction

Dans tous les domaines, les systèmes d'information occupent une place privilégiée au sein des entreprises. En effet, le management moderne se caractérise par l'optimisation des systèmes d'information et l'intégration de plus en plus forte et nécessaire des nouvelles technologies [MARC 97], [CHAR 02]. Le management et la gestion d'entreprise se font à travers les systèmes d'information et les outils technologiques qui représentent aujourd'hui « la matière première » sur laquelle l'efficacité et la performance des entreprises sont évaluées. Mais l'intégration des nouvelles technologies pour l'optimisation des systèmes d'information est une « affaire » de spécialistes. Souvent, les PME ne disposent pas d'experts dans le domaine et font appel à des cabinets de conseil censés les aider et les orienter en matière de systèmes d'information et de nouvelles technologies. L'identification des besoins réels des PME et leur formulation constituent une réelle problématique car elles restent en grande partie à la charge des PME [THOR 00].

L'environnement d'appartenance socioprofessionnel des clients étant assez hétérogène par rapport au métier d'audit, la difficulté de compréhension entre les deux pôles de la communication (client/ prestataire) reste une réalité incontestable. La question qui se pose à ce niveau est du type : est-ce que l'individu désigné au sein de l'entreprise cliente pour exprimer les besoins est forcément compétent dans le domaine des systèmes d'information ? La réponse est évidemment non. Le « sujet » parlant n'a pas forcément une expertise et une connaissance suffisante et surtout approfondie dans le domaine et par conséquent ne cible pas systématiquement par l'emploi d'un terme donné la réalité référentielle exacte. Ceci engendre des conséquences néfastes et lourdes à gérer tant au niveau de l'état d'avancement de la mission qu'au niveau de la gestion de la relation client.

Ces différents constats ont été recensés lors d'études faites sur le terrain. C'est dans un contexte pragmatique que la problématique que nous traitons est née. Une enquête réalisée auprès d'entreprises spécialisées dans le domaine de l'audit et du conseil en systèmes d'information, révèle l'existence d'une très importante difficulté de discernement et d'interprétation des requêtes des clients. Cette problématique et les conséquences qu'elle engendre ne sont pas exclusivement associées à l'audit des systèmes d'information mais se retrouvent dans différents contextes : de l'indexation d'une source documentaire et son identification à la représentation d'une connaissance et son interprétation. Le problème reste le même, il s'agit de faire correspondre deux ou plusieurs « mondes différents » à travers un ensemble de concepts pouvant être considérés comme étant substituables et/ou complémentaires dans un contexte donné. Notre problématique se singularise par le fait qu'elle couvre

un spectre de champs disciplinaires large : l'informatique, la linguistique et le domaine du métier traité, ici, l'audit et le conseil en systèmes d'information. Notre article est organisé en trois grandes parties. La première partie sera consacrée au cadre méthodologique que nous adoptons et déclinons sur le domaine d'analyse de la problématique et de nos propositions de solutions. Nous présentons, en effet, l'approche théorique dite « approche tridimensionnelle » [BOUZ 01]. On accorde dans cette méthodologie une importance particulière à l'acteur humain pris dans une activité donnée et aux outils technologiques qui lui permettent d'intégrer et de s'adapter efficacement à son environnement professionnel. La seconde partie, « analyse linguistique », correspond au traitement de notre problématique d'un point de vue linguistique. Le corpus à traiter est constitué essentiellement des termes qui ont été relevés lors de notre enquête. Enfin, la troisième partie, « analyse technologique » propose d'inventorier les divers outils techniques potentiels de représentation de l'information. Ces outils permettent d'assister l'utilisateur tant au niveau de sa recherche d'informations que dans l'identification des besoins de sa clientèle. En somme, sans prétendre trouver une solution, il s'agit plus de positionner la problématique à travers trois volets : le volet contextuel, le volet linguistique et le volet technique et de proposer une réflexion pouvant amener à la conception de démarches et d'outils d'aide à l'identification des besoins.

2. Analyse du contexte

Dans le domaine de l'audit et du conseil en systèmes d'information, deux groupes d'acteurs interviennent dans le processus : les acteurs clients et les acteurs conseillers. La difficulté réside dans le fait que chaque groupe possède un référentiel. Celui des acteurs clients est essentiellement orienté métier tandis que celui des acteurs conseillers est plus spécialisé dans le domaine des systèmes d'information et des nouvelles technologies. Il s'agit en fait de construire un sous référentiel commun permettant une adéquation entre le besoin exprimé et son interprétation. Cette adéquation n'étant pas évidente à reconstruire. Pour tenter d'atteindre cet objectif, nous allons essayer de répondre à un ensemble de questions. Nous citons : Quels sont les facteurs responsables de cette divergence ? Quel est le cadre méthodologique qui permet de mettre en relief les outils d'analyse de la problématique ? Comment décrire le référentiel de chaque population ?

2.1 Le domaine d'application : l'audit et le conseil en système d'information

La profession de consultant[1] permet aux entreprises de faire face aux changements dus aux évolutions sociologiques et technologiques. Les professionnels de ce domaine se donnent comme objectif d'assister le client dans l'évaluation de la

[1] Cette profession se dit pour les cabinets de conseil.

performance et de la croissance globale de son entreprise en intégrant les NTIC[2] [CARL 92].

Plusieurs critères conditionnent le succès d'une mission d'audit. Sans prétendre à l'exhaustivité, nous relevons le profil des différents acteurs tant au niveau des aspects techniques qu'au niveau du métier. Les conseillers ont tendance, de par leur spécialisation à avoir un profil plus technique. A l'inverse, les acteurs clients ont plus un profil fonctionnel lié à leur métier. Le problème que nous tentons d'aborder ne se situe pas au niveau de la forme de la requête mais au niveau du fond. En effet, le métier de conseil, tout en demeurant « technique », nécessite de la part des professionnels de l'audit une véritable capacité de compréhension et d'assimilation des préoccupations principales des PME et notamment sur le plan de l'expression des besoins. [DERR 92].

Nous proposons une démarche privilégiant l'acteur humain sans pour autant ignorer les aspects techniques et les spécificités fonctionnelles liées au métier. Cette approche méthodologique est nommée « approche tridimensionnelle » [BOUZ 01].

2.2 Cadre méthodologique : l'approche tridimensionnelle

L'approche tridimensionnelle se fonde sur le paradigme orienté usager [AMOS 99], [BENA, HUBE, MOTH 02]. Elle a été développée par le Professeur L. Bouzidi[3]. Le principe fondateur de cette approche est de porter une attention particulière sur l'acteur humain et d'analyser les outils technologiques potentiels capables d'optimiser son intégration dans une activité professionnelle donnée. Facteur humain, environnement professionnel et aspects techniques sont les composantes « clés » et le socle de notre approche. Ils correspondent successivement à la dimension humaine, la dimension activité et la dimension technique. Nous allons tenter d'appréhender les idées directrices des deux premières dimensions, la troisième sera abordée à la fin de l'article.

2.2.1 La dimension cadre ou activité

Dans le cadre de nos travaux, analyser les différentes fonctions assumées dans l'organisation, identifier les objectifs et les finalités visés et décrire les flux informationnels qui circulent et qui sont produits dans l'entreprise, sont autant d'éléments significatifs qui correspondent aux différentes facettes à travers lesquelles il est possible d'appréhender les principaux processus engagés dans un contexte entreprenarial.

---

[2] NTIC : Nouvelles Technologies de l'Information et de la Communication.

[3] Equipe de recherche SICOMOR, Centre de Recherche de l'IAE, Université Jean Moulin Lyon 3.

Cette dimension cadre permet de représenter une activité à travers l'analyse de plusieurs niveaux. En effet, le niveau organisationnel permet de préciser l'environnement global dans lequel une activité est assumée. Le second axe, dit fonctionnel, s'attache à étudier l'activité en identifiant les différentes fonctions qui la composent. Le niveau structurel est lié aux ressources humaines et matérielles nécessaires pour assurer l'activité. Enfin, étant donné que toute activité n'est pas statique dans le temps, une évolution liée au contexte socio-économique doit être prise en considération afin de représenter et d'analyser toute l'évolution de l'activité.

Au niveau de l'organisation de l'activité d'audit en systèmes d'information, la dimension cadre est constituée des grands pôles structurants de l'activité. Une cellule conseil serait décomposée en trois axes principaux : le niveau prospection, le niveau analyse et étude et le niveau développement et mise en œuvre. La fonction prospection touche à l'aspect marketing et commercial et a trait à l'étude de l'offre et de la demande. La phase suivante relève de l'analyse des informations recueillies par les prospecteurs en tenant compte des ressources humaines en particulier celles qui seront déployées lors de la mise en œuvre des solutions.

La figure 1 représente la dimension cadre et l'interaction entre ses différentes composantes.

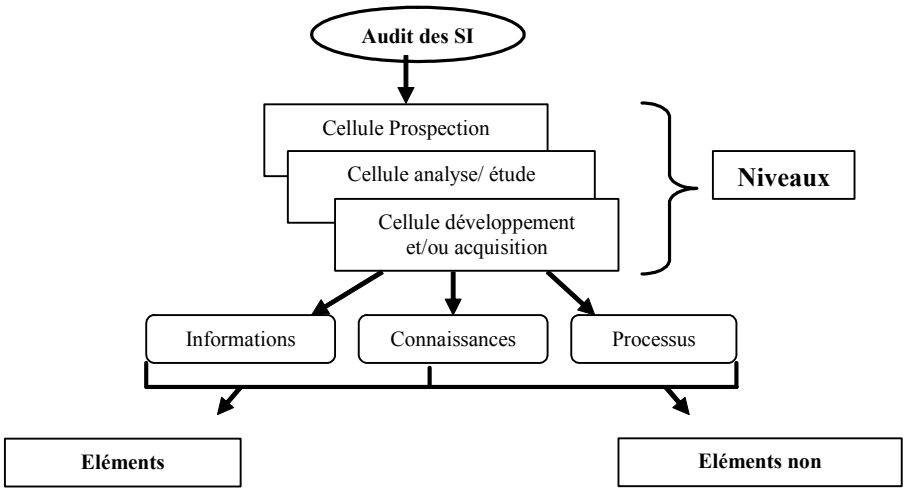


Figure 1 : La dimension cadre

Tenant compte des différentes entités représentées dans la figure ci-dessus, nous allons tenter de formaliser les phases générales de déroulement d'une mission d'audit. Nous nous focaliserons dans cette démarche sur l'enjeu et l'impact des problèmes d'interprétation sur la communication inter-individuelle, sur l'acheminement des informations entre le demandeur et le donneur d'ordres et enfin sur l'aboutissement et la réalisation de la mission. A ce niveau, chronologiquement, l'expression des besoins par le client se fait dans un premier temps. Ces besoins sont ensuite analysés par les employés chargés de la prospection. Des problèmes d'interprétation sont à prévoir ; la requête est reformulée par le cabinet et n'est que provisoire. Une réunion entre le prestataire et son client permet d'éclaircir le contenu de la requête initiale, avant la mise en œuvre et l'application des solutions proposées.

La figure 2 représente les phases de déroulement d'une mission d'audit et les problèmes en terme d'interprétation qui peuvent interrompre son bon fonctionnement.

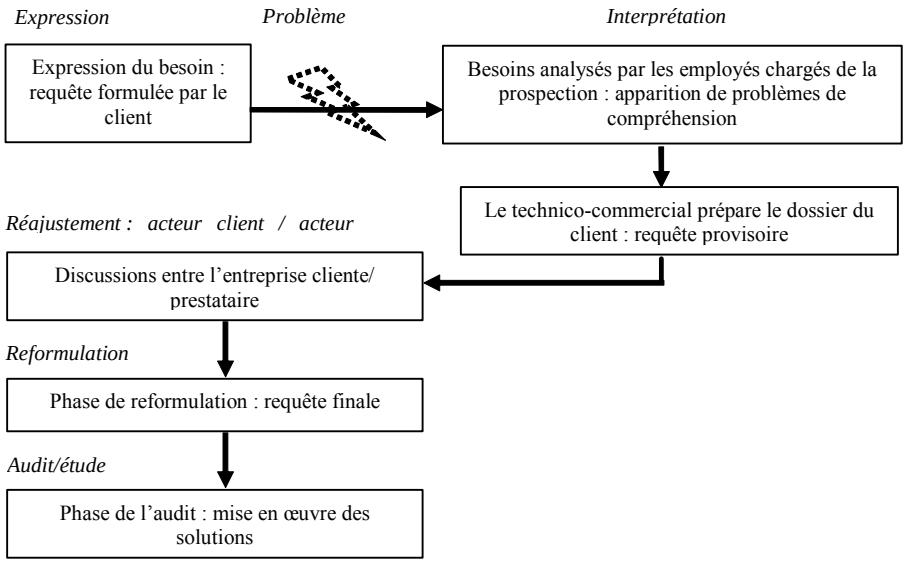


Figure 2 : Problèmes d'interprétation d'une mission d'audit

2.2.2 La dimension humaine

A - Plan d'identification et de classification des usagers

Dans un contexte d'expression des besoins, l'entreprise désigne, parmi les acteurs qui la composent, un individu qui sera chargé de s'exprimer auprès du cabinet de conseil et d'exposer la situation en précisant avec le prestataire le budget, le planning à suivre et les résultats attendus. Dans la plupart des cas, cet utilisateur rencontre des difficultés à se faire comprendre et, de là, à communiquer et échanger les connaissances sur son environnement. L'acteur humain représente l'élément essentiel qui conditionne l'échec ou la réussite du processus d'identification de la requête.

Il existe une variation d'utilisation selon chaque individu. En partant de notre enquête, nous sommes attirés par des particularités et des situations différentes dans lesquelles un locuteur s'exprime selon son domaine d'activité, la fonction qu'il occupe dans l'entreprise, son cadre environnemental et socioculturel, le degré de maîtrise du vocabulaire spécifique aux systèmes d'information [BENA 97]. En somme, tout se situe au niveau du savoir et du savoir-faire de l'individu, c'est-à-dire de sa compétence et de sa performance dans un domaine donné.

Dans la même perspective, l'approche orientée usager permet de traiter le comportement des différents types d'acteurs impliqués dans le processus d'expression des besoins [INGW 92], [LECO 01]. En se basant sur cette approche, nous avons établi une subdivision des acteurs en trois grandes classes majeures. Nous avons retenu la catégorie « utilisateur occasionnel », la catégorie « utilisateur averti » et la catégorie « utilisateur spécialisé ». Un utilisateur occasionnel se définit comme étant un individu qui utilise occasionnellement un outil technologique. La périodicité d'utilisation des outils technologiques est limitée et ciblée plutôt vers l'outil bureautique. On entend cependant par un utilisateur averti une personne initiée à l'essentiel du domaine de l'informatique. Elle maîtrise des notions qui sont traditionnellement connues, sans pour autant avoir une qualification qui lui permet d'atteindre un niveau de connaissances approfondies. Enfin, un utilisateur spécialisé, comme son nom l'indique, est un individu capable de maîtriser parfaitement son domaine : il est expert dans le domaine de l'informatique et des systèmes d'information.

B - Déclinaison de la dimension humaine :

On se propose à partir de cette classification de décliner la « dimension humaine » sur le domaine étudié. L'objectif est de cerner les besoins informationnels de l'utilisateur afin d'améliorer l'exactitude, l'identification et la précision de ses besoins réels. La figure 3 nous donne une idée synthétique de ces constats.

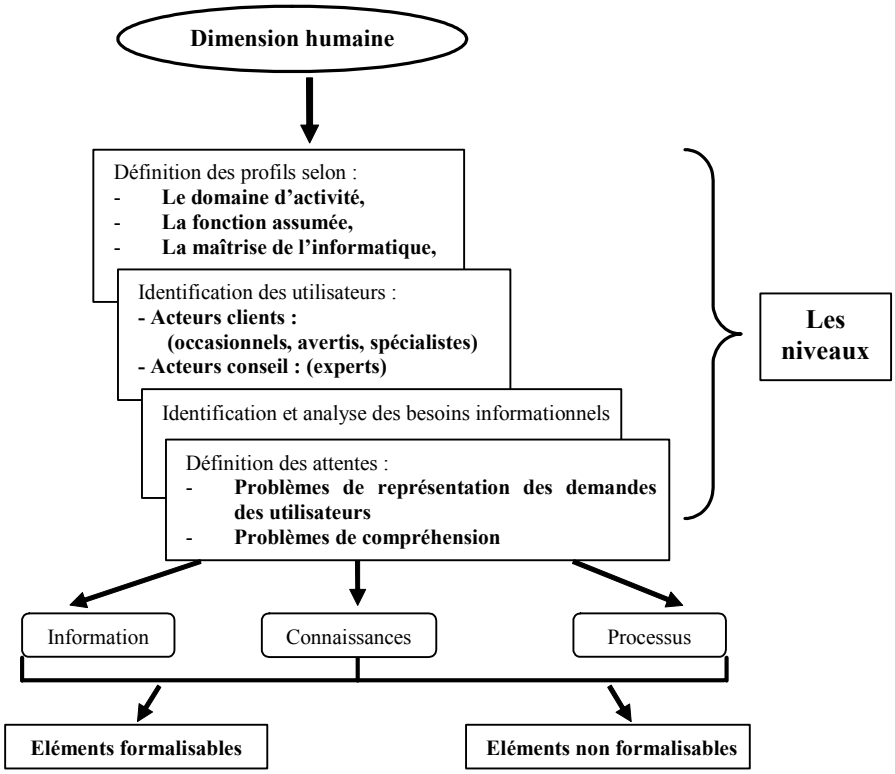


Figure 3 : La dimension humaine

L'analyse globale de la dimension humaine permet d'accorder au profil de l'usager de l'information une place capitale dans la problématique posée. Cependant, il existe d'autres pistes de recherches intéressantes que nous avons citées en introduction. En effet, la thématique que nous abordons peut également se poser au niveau de la pertinence du langage. Nous avons recensé des situations, où pour un même concept, l'utilisateur non expert emploi des désignations multiples, c es dénominations n'étant pas appropriées au contexte référentiel exact. Comme conséquence, les partenaires de la communication ne donnent pas un sens identique à un mot donné ; une discordance entre le sens produit par l'énonciateur et celui du récepteur peut naître de ce fait [BOUL 04]. En somme, la problématique tourne autour des mots, de cette unité de base du langage.

Quels moyens méthodologiques adopter pour qu'un système (le langage) aussi complexe puisse être étudié ? Est-il possible à travers une sensibilisation aux problèmes du langage d'amener les entreprises à réfléchir sur des outils techniques qui réduiront d'une part l'amalgame dans l'emploi des termes par le client et d'autre part la mauvaise interprétation faite par le prestataire ? La première question fera l'objet de la partie « analyse linguistique », la seconde question, quant à elle, sera abordée dans la dernière partie.

3. Analyse linguistique

Lorsque les informations sont communiquées par le client, la modélisation de ces besoins fait appel à des méthodes et des outils que nous organisons en deux niveaux. La description linguistique qui s'attache à dégager les règles sémantiques reliant les termes constitue le premier niveau. La représentation par les graphes qui permet de poursuivre l'analyse en formalisant les résultats obtenus constitue le deuxième niveau.

3.1 Recueil de données : démarches empiriques de constitution du corpus

C'est à partir d'une enquête réalisée auprès d'entreprises spécialisées dans le domaine de l'audit et du conseil en systèmes d'information que le corpus a été réalisé. Pratiquement, nous sommes partis de documents existants qui ont été fournis par les entreprises d'accueil. Ces documents que nous qualifions de « documents sources » contiennent des requêtes formulées par certaines entreprises clientes ayant des difficultés pour se faire comprendre. Le filtrage de ces documents nous a permis d'extraire une liste de termes qui constituent notre corpus d'analyse. Ces termes font l'objet de confusion dans l'emploi, ils ont une valeur sémantique très forte dans notre contexte d'étude.

Sans prétendre être exhaustif, il s'agit d'une analyse qui couvre des besoins exprimés au cours d'une période et dans un contexte donné, caractérisé par l'explosion des tentatives d'intégration des NTIC dans la gestion des systèmes d'information en Rhône-Alpes[4]. Voici un extrait des termes constitutifs du corpus d'analyse : *Applicatif, Architecture de systèmes d'information, Architecture client-serveur, Audit des systèmes d'information, Bases de connaissances, Bases de données, Concepteur de solutions informatiques, Conduite de projets, Conseil en systèmes d'information, Consultant, Datamining, Datawarehouse, EDI (Echange de Données Informatisées), ERP (Entreprise Ressources Planning), Extranet, GED (Gestion Electronique de Documents), Groupware, Ingénierie des systèmes*

[4] Enquête réalisée sur des documents sources en 2003, affinée en 2004.

d'information, Ingénierie conseil, Internet, Intranet, Logiciel, Management des systèmes d'information.

3.2 Outils d'analyse

Afin d'affiner les relations sémantiques qui existent dans ce corpus, nous allons nous appuyer sur un glossaire permettant d'identifier, de par la définition des concepts, les relations que ces concepts entretiennent a priori. La prise en considération du contexte et du profil permet de mettre en relief, outre ces relations, des liens que l'on qualifie de « relations liées à l'usage ».

3.2.1 Elaboration d'un dictionnaire des concepts utilisés

Le dictionnaire est structuré selon des définitions simplifiées. Pour chaque notion, nous avons retenu l'usage le plus commun. A titre d'exemple, le glossaire s'organise comme suit :

ENTREE : Architecture de système d'information.

Définition : le mot architecture est synonyme de structure, organisation, agencement. Il s'agit de la structure du système d'information, son organisation en matière de données et de traitements.

ENTREE : base de données.

Définition : pour que les ordinateurs puissent opérer, il est nécessaire de leur fournir en entrée des données structurées de façon précise et cohérente. On peut également définir cette expression comme un ensemble de données évolutives, organisées ou structurées pour en faciliter l'utilisation via un programme spécialisé d'accès aux données de la base [LAMI, SILE 01].

3.2.2 Elaboration d'une grille à double entrée

Le filtrage et le traitement des documents sources nous ont permis de rassembler des groupes de mots qui se caractérisent par une synonymie "liée à l'usage". Ils sont employés indifféremment par des locuteurs non experts et sont censés désigner pour eux une réalité singulière. Les résultats de cette analyse nous ont amenés à élaborer une grille à double entrée. Pour l'achèvement et la réalisation de cette grille, nous nous sommes appuyés sur deux ressources principales. D'une part, les documents primaires constituent évidemment la source de nos constats. D'autre part, une série d'entretiens avec les acteurs du cabinet de conseil chargés du traitement de ces missions. Ces entretiens nous ont permis d'affirmer nos hypothèses et de valider la répartition des différents termes dans la grille. Nous proposons à ce stade de l'étude d'analyser un extrait de la grille à double entrée.

CONCEPTS	UTILISATEURS NON EXPERTS (liste des synonymes liés à l'usage)
Applicatif	-Logiciel - progiciel
Audit des systèmes d'information	-Conduite de projets - Etude d'opportunité -Conseil en système d'information -Management des systèmes d'information -Ingénierie des systèmes d'information -Sécurité des données -Cahier des charges
Bases de données	-Datamining -Datawerahouse -SGBD (systèmes de gestion de Bases de données) -Systèmes à base de connaissances

Tableau 1 : Extrait de la grille d'analyse

L'objectif de cette grille est de nous permettre de « définir » un dictionnaire de « synonymes liés à l'usage » que les deux groupes d'acteurs peuvent utiliser. Sa pertinence pour les cabinets de conseil est liée à son utilisation, soit pour identifier avec certitude les besoins ainsi exprimés, soit pour lever le doute en affinant les besoins par des questions additives.

Au vu de ce tableau, la relation de synonymie liée à l'usage est fortement utilisée par les locuteurs non experts. Nous verrons plus en détail l'analyse des autres relations potentielles s'ajoutant à celle-ci.

3.3 Détermination de la nature des liens inter concepts

A priori, on peut relever, en technologie comme dans d'autres domaines, des relations de niveau morphologique, syntaxique et sémantique et des relations de natures différentes de type hyponymie, antonymie, synonymie, polysémie. Tenant compte de ces niveaux et/ou natures, nous allons procéder à une classification terminologique qui se fonde essentiellement sur la répartition des unités linguistiques (les termes) dans des catégories qui ont les mêmes propriétés linguistiques. Le principe est d'aboutir à un regroupement par « familles de mots » sur la base du sens « lié à l'usage », produit par un locuteur non expert dans le domaine. Ce dernier emploie plusieurs expressions pour référer au même contexte référentiel. Est-ce que cet emploi reste valide d'un point de vue proprement linguistique ? Quelle est la pertinence de l'analyse des relations liées à l'usage dans notre contexte d'étude ? Le langage représente-il l'outil « clé » pour répondre à la problématique posée par les cabinets de conseil ? Quel est l'impact de la « non

maîtrise » du langage de spécialité sur l'identification d'une demande ? Autant de questions qui nous interpellent et nous orientent vers la prise en compte du langage et des situations de « dérivation » dans l'emploi.

Analyse de la typologie des liens sémantiques

On distingue dans notre corpus des relations qui englobent les niveaux : hyperonymie, synonymie. On recense également des relations morphologiques dont on retient les niveaux suivants : la composition, la dérivation, les emprunts, les mots valises. Voyons les idées directrices de chaque type de relation.

Micro structuration sémantique : paire synonymique, hyponymique

La classification obéit à une logique qui consiste à regrouper un ensemble de termes sur la base de l'appartenance à la même relation sémantique. La première relation sémantique a été identifiée à travers l'échantillon des termes suivants : *SGBD*[5], *base de données*, *fichier*. En effet, il apparaît, à la lecture des documents sources[6], que l'utilisateur emploi l'expression *SGBD* censée, pour lui, inclure le sens de *base de données* et de *fichier*. C'est ce qui devrait correspondre normalement à la relation d'hyponymie[7] [LANG 98], [TOUR 00].

Afin de vérifier la validité de cet emploi, nous nous proposons de tester les implications admises et celles rejetées en se basant sur les définitions établis dans le glossaire (cf. section 3.2.1). On relève qu'une *base de données* est par définition l'ensemble des données relatives à l'entreprise, mémorisées dans un ordinateur et exploitées par un utilisateur. Un *SGBD* quant à lui est un logiciel qui permet à un utilisateur d'interagir avec une base de données. La première notion représente un ensemble de données, la seconde représente le mode de traitement de ces données. Nous avons en face de nous une « classe » de signifiants qui renvoie à des signifiés totalement différents et en même temps complémentaires. Ce traitement linguistique nous mène à conclure qu'il ne s'agit aucunement d'une relation de type hyponymique.

---

[5] SGBD est l'abréviation de l'expression Système de Gestion de Base de Données.

[6] Les documents sources représentent la matière « première » qui nous a permis d'élaborer le corpus d'analyse. Ce sont des textes écrits qui résument en détail la requête de certains clients (cf. section 2.1.)

[7] On définit l'hyponymie comme étant un rapport sémantique qui unit un lexème appelé « hyponyme », à un lexème appelé « hyperonyme » par la relation d'inclusion sémantique. On entend dire par inclusion quelque chose comme être compris dans ou être contenu dans. Ainsi, un élément X peut être appréhendé comme étant englobé dans ou faisant partie d'un élément Y. En terme logique, on peut rendre compte de ce type de relation par une représentation du type : $Y \subset x$.

La relation sémantique se démarque également par la présence de la synonymie [NYCK 99]. En partant du tableau à double entrée (cf. section 3.2.2), une subdivision est à prévoir sur l'ensemble des termes concernés par cette relation. En premier lieu, nous remarquons un groupe de synonymes en « langue » qui représente les synonymes que l'on retrouve traditionnellement dans les dictionnaires de spécialité. L'exemple qui nous semble le plus parlant est la synonymie entre *ERP*[8] et *système intégré*. En effet, dans la mesure où un *ERP* est l'appellation en anglais de *système intégré* et du moment que la variation entre ces deux mots n'affecte que le niveau de la désignation et de la dénomination et ne touche pas à un niveau plus profond, c'est-à-dire au sens ; on peut s'accorder à qualifier la relation entre ces deux notions de relation de synonymie en « langue ». En second lieu, on repère une liste de synonymes en « contexte » qui représente les substitutions de concepts que les usagers font en exprimant leurs besoins. L'exemple qui nous permet de rendre compte au mieux de cette relation est la synonymie liée à « l'usage » entre *système d'information* et *système informatique*. L'équivalence de sens entre ces deux termes est exclusivement liée à l'usage du moment qu'un *système d'information* est par définition l'ensemble des informations et des procédures qui définissent l'activité d'une organisation et que le *système informatique* représente un moyen qui permet l'optimisation de ce système d'information.

Morphologie lexicale

La dérivation et la composition sont les deux grandes voies de formation morphologique des mots [AINO 97]. La première forme un mot à partir d'un autre en y ajoutant un ou plusieurs affixes. La seconde forme un mot en assemblant plusieurs mots [LEHM, MART 98]. En ce qui concerne la première relation, des termes comme *Intranet*, *Extranet* et *Intranet* entretiennent une relation de nature morphologique et de type dérivationnel. Le radical *net*, renvoie à l'idée de réseau, les dérivés, quant à eux, ont été obtenus à travers une opération de préfixation. De par la procédure de formation de ces termes, nous pouvons avancer un constat très intéressant. On dira qu'étant donné que l'affiliation entre le sens du mot global et le sens de chacun des morphèmes qui le composent, c'est-à-dire du radical et du préfixe, permet d'obtenir le sens global du terme, la substitution dans l'emploi de l'un de ces termes par l'autre peut être due à l'origine morphologique. Autrement dit, le rapport d'immédiateté entre la forme et le sens de ces mots peut représenter une piste plausible quant à l'interprétation de notre problématique. Nous pensons qu'en règle générale, le sujet parlant ne maîtrisant que la signification du terme *Internet*, du fait qu'il soit le réseau mondial d'interconnexion, se retrouve

[8] ERP est l'abréviation de Entreprise Ressources Planning.

confronté à une situation où lui échappe le sens des dérivés, d'où le recours à la substitution.

3.4 Représentation des concepts et des liens par une structuration adaptée : arborescence mettant en relief les liens sémantiques et terminologiques (thésaurus)

La recherche d'information concerne les mécanismes qui facilitent l'accès à une base d'informations. Il existe plusieurs modèles de représentation, le choix du modèle conditionne la pertinence des réponses du système, donc la satisfaction de l'utilisateur [BERR, BOUG 02], [LAINE 01].

Notre travail portera sur la représentation par le modèle des graphes conceptuels. Ce modèle propose une structuration arborescente mettant en relief les termes et les relations inter concepts, ce qui permet à l'utilisateur de visualiser et d'appréhender les informations d'une façon plus objective [BERG 83].

3.4.1 L'approche par les graphes : vers un méta modèle

Le principe est d'inscrire les unités linguistiques dans un ensemble de relations d'ordre sémantique ou morphologique. Ceci nous permet de voir pour chaque terme son voisinage immédiat et/ou profond, de telle sorte que le sens de l'unité qui fait l'objet de confusion ne soit pas confondu avec le sens des unités avec lesquelles elles entretiennent d'éventuelles relations. On obtient des réseaux, les graphes, et s'appuyant sur une théorie particulière qui est la théorie des graphes [LABE 81].

Nous proposons d'appliquer la théorie des graphes pour définir un méta-modèle, puis de décliner ce dernier sur les données qui constituent réellement notre thésaurus. Pour pouvoir concourir à cet objectif, nous considérons le thésaurus comme étant un graphe. Sur le plan logique, les réseaux sémantiques sont des graphes. Les nœuds de ce graphe représentent l'ensemble des termes relatifs au domaine des systèmes d'information et qui sont reliés par des arcs représentant les relations sémantiques ou morphologiques qu'entretiennent ces termes [SACH 74].

La première relation, notée « r1 », traduit l'existence d'une relation que l'on qualifie d'hyperonymie. Ce lien se caractérise par une schématisation verticale. La seconde relation, notée « r2 », désigne la relation de synonymie où l'on observe bien que le sens de la manipulation se fait horizontalement. Nous représentons le méta modèle dans la figure ci-dessous.

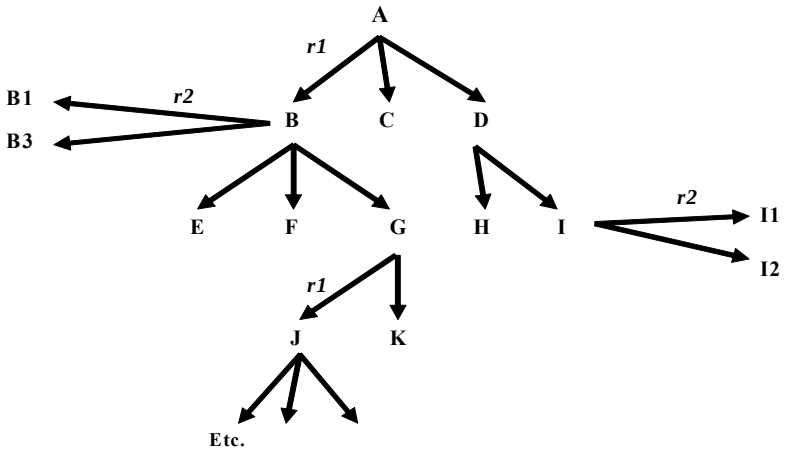


Figure 4 : Application de la théorie des graphes : le méta-modèle

3.4.2 Instanciation du méta-modèle : le thésaurus final

Le méta-modèle contient les éléments de base qui seront exploités afin d'établir le thésaurus. Son instanciation résulte du fait de remplacer les symboles désignant, d'une part, l'ensemble des termes (A, B, C, D) et, d'autre part, les relations (r1, r2) par les vraies valeurs correspondantes.

Valeurs du « méta modèle »	**Valeurs du « modèle »**
X (l'ensemble A, B, C, etc.)	Système d'information, SGBD, SIAD, base de données, logiciel, Internet, télécommunication, Datamining…
r1	De hyperonyme à hyponyme
r2	De synonyme à synonyme

Tableau 2 : instanciation du méta modèle

La relation sémantique la plus évidente à considérer est la relation hiérarchique entre l'hyperonyme et son hyponyme. A titre d'exemple, un *SIAD*[9] est, par définition, un système d'information. On peut considérer l'expression *SIAD* comme l'hyponyme et l'expression *système d'information* comme l'hyperonyme, tout en

[9] SIAD est l'abréviation de Système d'Information d'Aide à la Décision.

respectant le sens de la relation qui est bien évidemment vertical. Ce constat reste valable pour *système d'information et de décision*, *système expert* et *logiciel*. De même, dans un contexte d'expression des besoins, le sujet parlant a tendance à employer les termes *base de données* et *datamining* et à les considérer comme des synonymes en contexte. Le sens de cette relation est unilatéral. Par contre, il existe des paires synonymiques dont l'implication est double. Nous citons les expressions *base de données* et *système de gestion de bases de données* qui sont des « quasi-synonymes » substituables dans tous les discours. En terme logique, cette relation est symétrique : {il existe au moins deux termes A et B/ A est synonyme de B et B est synonyme de A}. En terme sémantique, le signifié de A = signifié de B et le signifié de B = signifié de A. Enfin, dans le graphe, nous pouvons considérer les relations de niveau morphologique. Par exemple, *Internet*, *Extranet* et *Intranet*, de par le procédé de formation, sont à retenir. Nous considérons l'expression *système d'information* comme nœud principal. Sa définition et son contenu sémantique riche la classe au sommet de la représentation hiérarchique. On peut continuer de faire ressortir les éléments de la hiérarchie pour obtenir la figure 5.

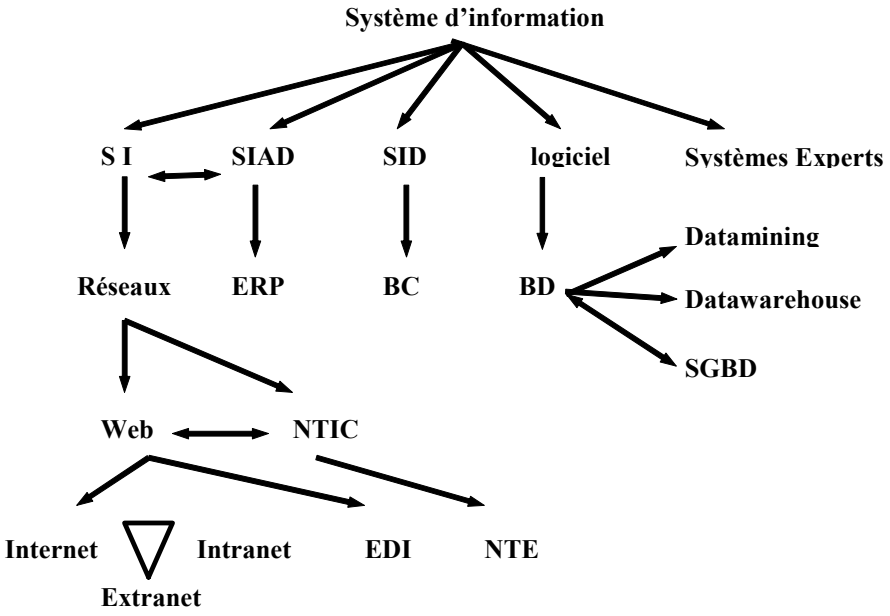


Figure 5 : Extrait du modèle représentant le thésaurus

Les abréviations correspondent à : SI (Système Informatique) ; SID (Système d'Information et de Décision) ; BC (Base de Connaissances) ; BD (Base de Données) ; SGBD (Système de Gestion de Base de Données) ; EDI (Echange de Données Informatisées) ; NTE (Nouvelles Technologies Educatives).

Il faut signaler que ce type de représentation est certes répandu. L'innovation réside dans le fait que nous essayons de l'utiliser pour structurer des relations liées « à l'usage ».

En somme, une représentation sous forme de graphe est intéressante, néanmoins, d'autres formes de représentation des informations peuvent être considérées, telles que les réseaux de neurones, les structures de treillis. De plus, l'exploitation et la représentation de cet ensemble de données et de relations peuvent être optimisées par l'utilisation d'un outil d'aide à l'interprétation que nous essayons de développer dans la partie suivante.

4. Vers une réflexion sur un outil informatisé

4.1 Limites d'une exploitation « manuelle » : intérêt d'une informatisation

Le principe est de définir le contour d'un outil technologique pouvant nous permettre de valider nos hypothèses et notre démarche. Pour ce faire, deux scénarii « conceptuels » résumeront l'essentiel de la démarche technique.

Scénario n° 1

On propose d'envisager une situation concrète où le client exprime son besoin à travers une requête d'où sont extraits des mots clés que l'acteur conseil juge pertinents et suffisants pour couvrir l'ensemble sémantique du sujet dont il est question. A partir de ces mots clés, la seconde étape se résume par la consultation du thésaurus (représenté dans la figure 6 par la base d'informations). En réponse et lors de la phase de sélection, notre base de données nous fournit l'ensemble des termes constitutifs du voisinage des concepts de la requête. Enfin, l'acteur conseil pourra se permettre d'affiner ce que vise son client. Dans ce scénario, l'utilisateur final (le cabinet de conseil) est censé procéder par une indexation des termes jugés pertinents. Ceci mène à la constitution d'une base de connaissance où des groupes de termes sont étiquetés par des relations sémantiques ou morphologiques qui les identifient.

La figure 6 résume la prise en considération des éléments linguistiques.

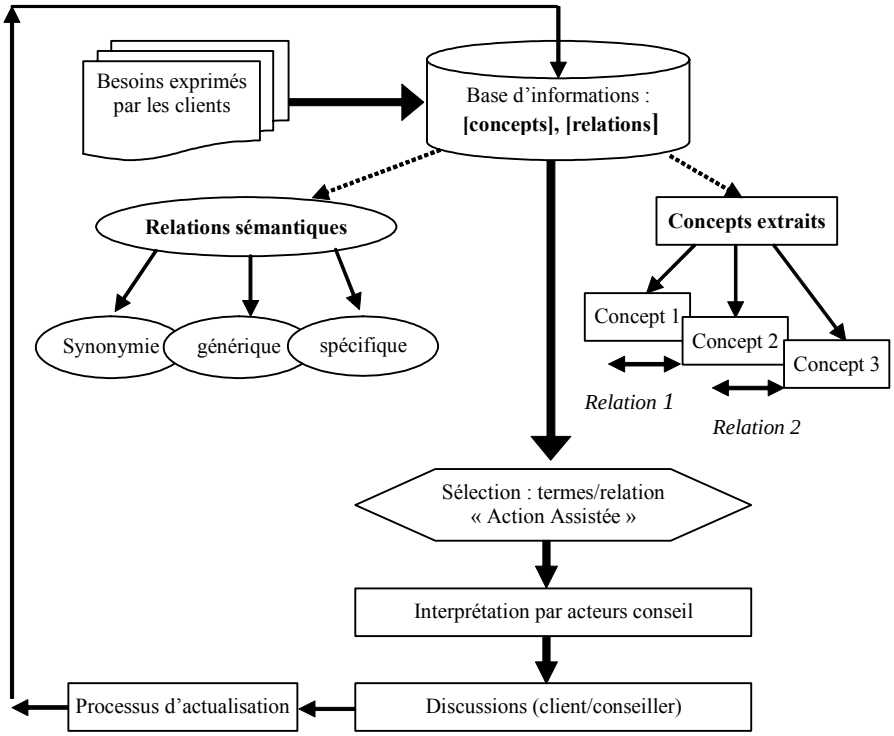


Figure 6 : Prise en compte des éléments linguistiques

Scénario n° 2

Le second scénario complète le premier par la prise en considération de la dimension humaine (c'est ce qui correspond à « modélisation des profils potentiels » dans la figure 7). Elle se matérialise par l'identification des acteurs exprimant leurs besoins. Le profil est identifié par des critères dont certains sont difficilement appréhendables. Nous pouvons citer le degré de maîtrise du domaine, la fonction assumée. En somme, l'injection du profil du locuteur débouche sur un système « personnalisé » d'aide à l'interprétation des besoins, qui tient compte à la fois des termes, des relations sémantiques mais aussi des caractéristiques du locuteur.

Le second scénario est représenté dans la figure 7.

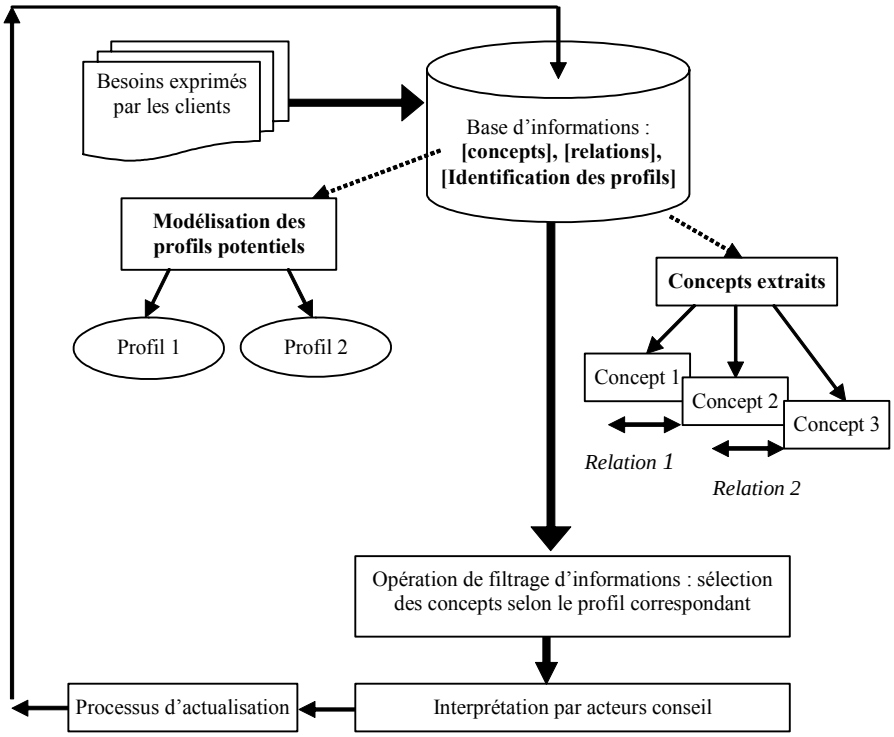


Figure 7 : Vers un système d'aide à l'interprétation des réponses

4.2 Déclinaison de la dimension technique

Dans un contexte de représentation des informations, déterminer l'outil technique permet d'améliorer considérablement la pertinence des résultats. L'approche tridimensionnelle, par sa composante technique, nous permet de définir le contour technique d'un outil pouvant s'adapter à notre contexte d'étude. L'analyse de cet aspect technique est à la fois organisationnelle, fonctionnelle et opérationnelle. L'analyse fonctionnelle concerne l'identification des différents types d'outils d'accès et de traitement de l'information. Elle est dite fonctionnelle car elle permet de définir les fonctions réalisées par le système en décrivant ses performances mais aussi les limites de son adaptabilité et de son utilisation.

Le choix et la mise en oeuvre du système dépendent des ressources matérielles et humaines de l'entreprise. Son intégration appelle une réorganisation des postes de travail, un réaménagement des tâches et une estimation du budget nécessaire à l'implantation et à la mise en place. Enfin, l'analyse opérationnelle concerne la mise en œuvre et l'exploitation de cet outil technique par des acteurs humains. Il s'agit

entre autre de répondre aux besoins des utilisateurs tant au niveau de la facilité d'accès aux informations qu'à la convivialité d'exploitation du système.

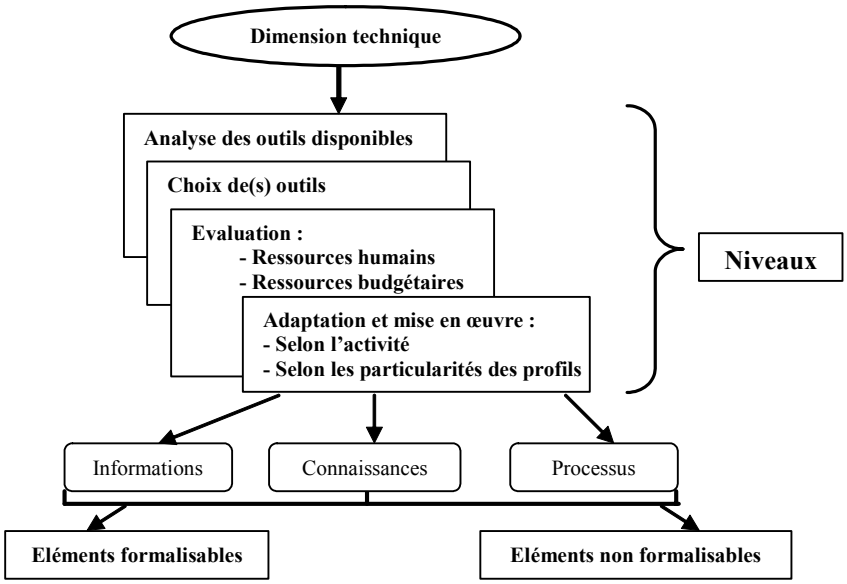


Figure 8 : La dimension technique

4.3 Architecture du système proposé : interaction des trois dimensions

La dimension activité nous permet de représenter l'activité conseil à travers l'identification des différents pôles structurants le domaine. Nous avons retenu les axes suivants : le pôle prospection, le pôle analyse et enfin, le pôle développement. La dimension humaine permet d'aborder notre problématique à travers l'analyse des particularités des profils de deux populations : les acteurs clients et les acteurs conseil. La dimension technique résume les composantes essentielles du système informatique que l'on projette. Les termes, les relations et la typologie des acteurs sont les éléments fondateurs.

L'approche tridimensionnelle nous permet au final d'avoir une architecture globale du fonctionnement du système d'aide à l'interprétation des besoins que nous proposons et ce, à travers l'interaction de ces trois dimensions [BOUZ 01]. La figure 9 décrit l'architecture du système final.

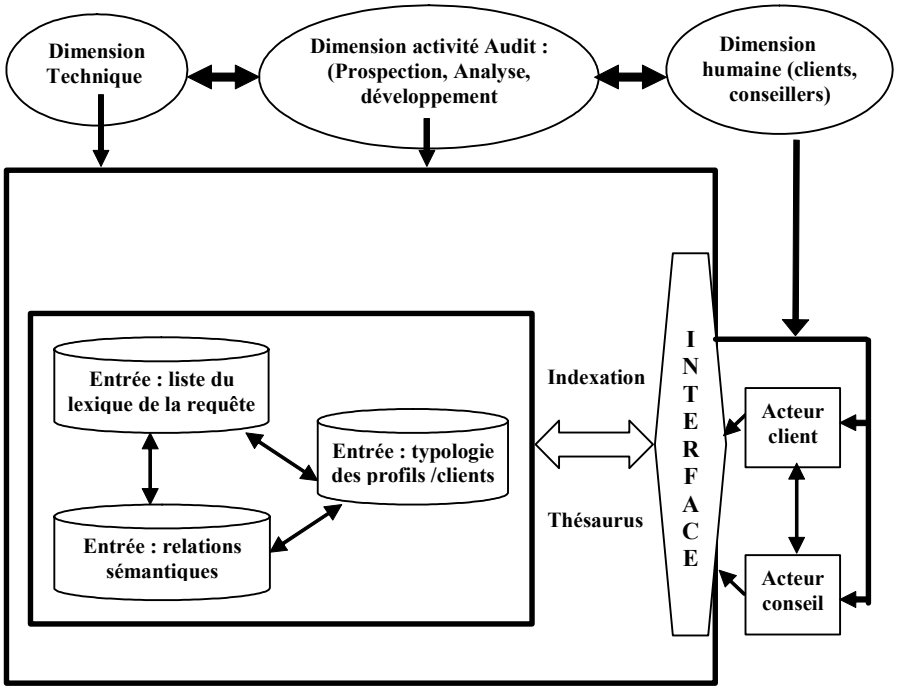


Figure 9 : Architecture du système proposé : interaction des trois dimensions

4.4 Architecture du système proposé : interaction des trois dimensions

L'idéal dans le système que nous proposons serait d'envisager un schéma global de fonctionnement où l'on intègre des informations sur : Les termes, les relations et le profil du client dans la même base de donnée. Le système devrait proposer à l'utilisateur en fonction du profil du client à la fois le terme sur lequel porte la confusion et le voisinage en terme de relations liées à l'usage.

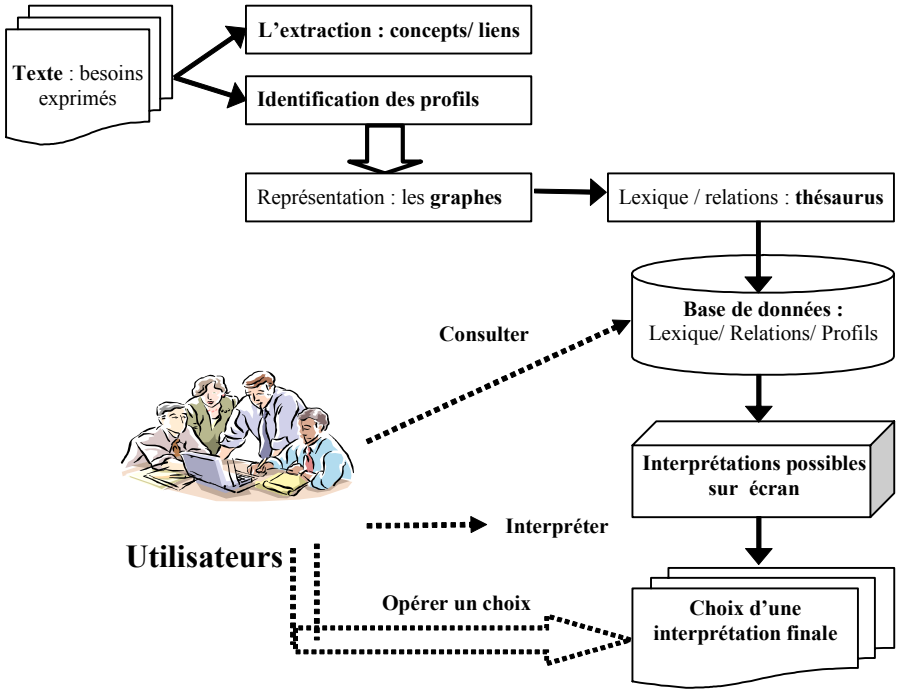


Figure 10 : Schéma global de fonctionnement du système d'aide à l'interprétation des besoins

5. Conclusion

En matière de recherche, de représentation et d'interprétation des informations, plusieurs critères sont à prendre en considération. Ils se déclinent à travers les trois dimensions qui constituent le fondement de la conception d'un système d'aide à l'interprétation. Au niveau de la dimension humaine, l'identification des profils des acteurs et leur rôles, au niveau de la dimension activité le degré de pertinence des informations et l'analyse des processus et au niveau de la dimension technique, les outils d'optimisation de ces mêmes processus.

Nous avons décliné notre problématique et notre démarche par la proposition d'une solution dont nous tentons de valider les hypothèses sur un terrain plus large dans le même domaine d'activité.

6. Bibliographie

[AINO 97] Aino Niclas-Salminen, « La lexicologie », Edition Armand Colin, 1997.

[AMOS 99] Amos D., « Modélisation de l'utilisateur et recherche coopérative dans les systèmes de recherche d'informations, Organisation des connaissances en vue de leur intégration dans les systèmes de représentation et de recherche d'information », Congrès d'ISKO- France, Lille, 1999.

[BENA 97] Ben Abdellah N., « Analyse et structuration des documents scientifiques pour un accès personnalisé à l'information : vers un système d'information évolué », Thèse de Doctorat, Université C. Bernard - Lyon 1, 1997.

[BENA, HUBE, MOTH 02] Benammar A., Hubert G., Mothe J., « Proposition à l'intégration des profils dans le processus de recherche d'information », Institut de recherche en informatique de Toulouse, 2002.

[BERG 83] Berge C., « Graphe », Gauthier, Villars, 1983.

[BERR, BOUG 92] Berrut C., Boughanem M., « Recherche et filtrage d'information », Hermès science publication, Paris, 2002.

[BOUL 04] Boulesnane S., « Analyse terminologique des concepts dans le domaine des systèmes d'information », Mémoire de DEA, 2004.

[BOUZ 01] Bouzidi L., « Un système d'aide à l'accès aux connaissances : apprentissage, décision et recherche d'information ». Habilitation à Diriger des Recherches, Université Jean Moulin Lyon 3, 2001.

[CARL 92] Carlier A., « Stratégie appliquée à l'audit des systèmes d'information : les approches méthodologiques et l'audit qualité », Hermès, 1992.

[CHAR 02] Charpentier P., « Organisation et gestion de l'entreprise », Edition Nathan, 2002.

[DERR 92] Derrien Y., « Les techniques de l'audit informatique », Dunod, 1992.

[GOND, MINO 79] Gondron M., Minoux M., « Graphes et algorithmes », Eyrolles, 1979.

[INGW 92] Ingwersen P., "Information retrieval interaction", Taylor and Graham, London, 1992.

[LABE 81] Labelle J., « Théorie des graphes », Modulo éditeur, 1981.

[LAINE 01] Laine-Cruzel S., « Conception de systèmes de recherche d'information : accès aux documents numériques scientifiques », Habilitation à Diriger des Recherches, Université Claude Bernard Lyon 1, 2001.

[LAMI, SILE] Lamizet B., Silem A., « Dictionnaire encyclopédique des sciences de l'information et de la communication », 2000.

[LANG 90] Revue Language, « L'hyperonyme et l'hyponyme », n° 98, juin 1990.

[LECO 01] Le Coadic Y.F., « Usages et usagers de l'information », ADBS, Nathan Université, Information et documentation, 2001.

[LEHM, MART 98] Lehmann A., Martin-Berthet, « Introduction à la lexicologie, sémantique et morphologie », édition Nathan, 1998.

[MARC 97] Marciniak R., « Système d'information, dynamique et organisation », édition Economica, 1997.

[NYCK 99] Nyckees V., « La sémantique », Edition Belin, Solange Ghernaouti-Hélie et Arnaud Dufou : De l'ordinateur à la société de l'information, Collection Que sais-je, 1999.

[SACH 74] Sach A., « La théorie des graphes », Paris, 1974.

[THOR 00] Thorin M., « Audit informatique », Hermès science publication, 2000.

[TOUR 00] Touratier C. « La sémantique », Paris, A. Colin, 2000.

Catégorisation des hyperdocuments multilingues : le système hyperling

Khaldoun Zreik, Tuan-Dang Nguyen

GREYC – UMR 6072 CNRS – Université de Caen
14032 Caen Cedex - France

{zreik,tnguyen}@info.unicaen.fr}

Résumé :

La prise en compte du multilinguisme est un élément déterminant dans le domaine du Web sémantique. La présence ou non de multiples langues sur un site Web engendre trois types de problèmes dont l'ignorance pourraient nuire à la qualité des résultats obtenus d'une démarche du Web Mining : i) la redondance, si le site propose simultanément des traductions en plusieurs langues, ii) les parcours bruités lors d'un passage d'une langue à une autre via les vignettes (génération de graphes, conceptuellement, non signifiant), iii) la perte de l'information par la négligence de la spécificité structurelle (même implicite) de chaque langue. Dans cet article, nous abordons l'aspect multilinguisme dans un contexte de catégorisation des sites Web multilingues. Nous tâcherons d'apporter quelques éléments de réponses à des questions de base telles que : la représentation d'hyperdocuments multilingues ; la modélisation des données en une structure homogène ; la qualité de la recherche d'information dans un contexte multilingues et enfin la définition de la notion de centre de gravité pour départager des langues dominantes ?

Mots-dés : Apprentissage, Catégorisation, Fouille des Sites Multilingues, Hyperdocument, Multilinguisme.

1. Introduction

La question de la spécificité culturelle et tout particulièrement celle de la spécificité linguistique influencent encore les travaux de déploiements massifs des procédures de standardisation et de mondialisation. Ceci peut être expliqué par le fait que le nombre des sites multilingues ne cesse de croître malgré leur coût de développement qui est nettement plus élevé que celui du développement des sites monolingues. Cependant, nous constatons que le domaine de recherche dans le domaine de fouilles de sites Web multilingues est très peu exploré. Dans cette contribution nous proposons une approche de catégorisation structurelle pour la reconnaissance et la recherche de documents Web multilingues.

Pour la reconnaissance de la caractéristique « multilingues » des sites nous introduisons, dans un premier temps, les notions de : relations, interrelations, et frontières entre des « régions » qui seraient formées par différentes langues sur le site (s'il y en avaient plusieurs). Aussi, nous formulons une hypothèse, dont la plausibilité ne serait qu'expérimentale, qui consiste à considérer que la spécificité de chaque langue peut émerger (entre autre) dans la structuration et la localisation des informations complémentaires. Autrement, dans un système multilingue, les distances entre les chemins de parcours d'information de même langue devraient être inférieures aux distances entre les chemins de parcours d'information de langues différentes.

Dans notre approche nous observons que la **structure** d'un document Web incorpore des informations qui sont indispensables pour toute démarche d'optimisation de la recherche d'information ou des fouilles de sites web. La crédibilité de cette hypothèse peut être renforcée par le développement accéléré de documents structurés. Pour illustrer notre propos nous avons développé une méthode distributionnelle (le système Hyperling) capable de déterminer, sans aucune connaissance linguistique préalable et explicite, le nombre des langues dominantes sur un site Web multilingues. Hyperling est développé dans le cadre d'un projet de recherche global sur la recherche et l'extraction d'information à partir de documents électroniques dynamiques monolingues ou multilingues. Extraire une information dans ce contexte est défini comme un processus de repérage, formalisation et de traitements (indexation et classification) des structures de données pouvant comporter d'information pertinente (Kosala R., Blockeel H., 2000).

2. La catégorisation des hyperdocuments multilingues

Pour le traitement des hyperdocuments (sites web) multilingues nous considérons, entre autres, deux types d'approches : celles qui sont basées sur le Traitement Automatique des Langues Naturelles (TALN) et celles qui émanent des domaines de l'apprentissage automatique et de la fouille des ressources électroniques structurées. Ces deux approches seraient complémentaires (dans une

problématique de découverte de connaissance à partir d'un hyperdocument) ou bien totalement séparées.

En phase de prétraitements des ressources d'information (des sites Web), Hyperling s'appuie sur des méthodes de d'analyses structurelles et statistiques (approuvées par diverses communautés d'apprentissage et de fouille des données). Ces méthodes possèdent un potentiel non négligeable pour débrouiller plusieurs questions soulevées dans un contexte de multilinguisme (Fürnkranz, 1999) avant d'invoquer, si nécessaire, des méthodes spécialisées en TALN. Les hyperdocuments, qui sont méthodiquement conçus et réalisés, renseignent amplement sur la structure de l'information qu'ils portent :

- Le balisage HTML permet de reconnaître l'importance sémantique de chaque élément de l'hyperdocument
- La représentation d'un lien entre les hyperdocuments peut identifier une catégorie d'appartenance de l'hyperdocument cible.

Il est indiscutable que la représentation des hyperdocuments a un impact déterminant sur la qualité des résultats obtenus par des procédures de classification ou de catégorisation. Ainsi Hyperling établit, à partir des deux points cités ci avant, une représentation symbolique et statistique de l'hyperdocument :

- Le vocabulaire (regroupements de mots),
- La structure interne (pointeur structurels : balises HTML ou XML)
- Les caractéristiques quantitatives des composants spéciaux (tableau, légendes d'images...)
- Les liens (pointeurs de dépendances) entre les hyperdocuments.

Cette représentation, dans sa définition, conserve, sans qu'ils soient explicités, des spécificités linguistiques et sémantiques de l'hyperdocument. Le choix de cette présentation qui est fortement liée aux traitements que subiront les ressources en question, a pour objectif d'optimiser la qualité de la numérisation (vectorisation) de ces ressources. Cette étape est préalable à celle de la classification.

2.1 Approches simplifiées

La majorité d'approche de classification ou de catégorisation opère sur des données qui sont prétraitées et représentées par des structures robustes et simplifiées, par exemple la représentation tabulaire (dans les arbres de décisions) ou la représentation vectorielle (pour la catégorisation conceptuelle, apprentissage paramétriques). Dans Hyperling nous optons pour une représentation numérique vectorielle. Donc la numérisation d'un hyperdocument sera rendue sous forme de vecteurs.

La performance des méthodes de numérisation qui sont basées sur la fréquence des mots, semble être assez problématique lorsqu'il s'agit de problèmes de classification des documents spéciaux de type brevets ou des pages Web

(Chakrabarti S., 2001). En outre, d'autres études (Wai-Chiu Wong, Ada Wai-Chee Fu, 2000) ont montré que 94,6 % des pages Web contiennent moins de 500 mots distinctifs et la plupart des mots présente une fréquence inférieure à 2,0. Néanmoins, la grande majorité des approches d'indexation, classification, catégorisation et de recherche/restitution d'information (IR : Information Retrieval) adopte ce critère dans le prétraitement des ressources en question, c'est-à-dire leur représentation interne (Honkela T., Kaski, S., Lagus K., Kohonen T. 1997 ; Ahonen H., Heinonen O., Klementtinen M., Verkamo A., 1998 ; Billsus D., M. Pazzani, 1999 ; Hofmann T., 1999 ; Scott S., Matwin S., 1999, etc.). Plusieurs autres chercheurs ont démontré les limites de la performance de cette représentation lors de la classification des documents textuels et hypertextuels (Dengel A., Dubiel F., 1995 ; Asirvatham P. A., Ravi K., 2001 ; Ma L., Shepherd J. Nguyen A., 2003 ; Theobald M., Schenkel R., Weikum G. ; Bratko, A., Filipi, B., 2004, etc.). Certains chercheurs ont proposé d'enrichir cette représentation par l'introduction d'un critère tenant compte des relations entre les hyperdocuments pour améliorer la performance des méthodes de classification (Chakrabarti S. *et al.* 1998). La prise en compte des relations entre les hyperdocuments (les hyperliens), est inspirée des travaux de recherches sur les réseaux sociaux (Wasserman S., Faust K., 1994 ; Kautz H., Selman B., Shah M., 1997) qui proposent d'établir un modèle topologique pour classifier des pages Web (Chakrabarti S., Dom B, Indyk P., 1999)[1]. Les premières expérimentations de cet approche ont été effectuées dans le domaine de la recherche d'information sur l'Internet (Bharat K., Henzinger M. R., 1998 ; l'algorithme PageRank - Brin S., Page L., 1998 ; Chakrabarti S. *et al.* 1998 ; l'algorithme HITS - Kleinberg J. M., 1998).

Pour catégoriser/classifier des hyperdocuments, nous constatons que l'ensemble des travaux, cités ci-dessus, est basé essentiellement sur des critères statistiques et pseudo structurels, à savoir : la fréquence des mots, la représentation interne (analyse de certaines parties de la structure telles que les méta-data, titres, etc.) et les hyperliens.

Pour modéliser les hyperliens (Džeroski, Lavrač, 2001) proposent des prédicats représentant des relations entres deux pages et l'appartenance d'un mot à une page, par exemple : « link_to »(page1, page2) et « has_word »(page, word). Pour classifier les pages hypertextuelles (Fürnkranz J., 1999, 2001) propose l'introduction de la notion de «hyperlink sets». L'idée étant d'exploiter les textes associés aux hyperliens, c'est-à-dire : les titres, les mots présent dans le paragraphe contenant un hyperlien et les textes relatifs aux ancres.

Partant de ces travaux (Džeroski, Lavrač, 2001, Fürnkranz J., 1999, 2001) nous proposons une représentation interne (vectorisation) des hyperdocuments donnant une importance accrue à la notion d'hyperlien (directe ou indirecte) dans l'analyse

[1] Il est à noté que la structuration des hyperliens a fait l'objet de nombreux travaux de recherches (Bharat K., Henzinger M. R., 1998 ; Brin S., Page L., 1998 ; Chakrabarti S. et al. 1999 ; Fürnkranz, 1999). Ces travaux ont donné lieu à un axe de recherche et développement en plein essor depuis 2000, à savoir la fouile des structures Web (Web Structure Mining) (Chakrabarti S., 2000).

de la structure des pages Web multilingues. Cette méthode a été brièvement introduite dans (Nguyen T. D., Zreik K., 2004, 2005).

2.2 Représentation sous forme de graphe orienté

Hyperling, assimile un site Web (constitué d'un ensemble de documents html ou équivalent pouvant avoir des hyperliens entre eux) à une base de donnée semi structurée (Abiteboul S. *et al.* 1997 ; Konopnicki D., Shmueli O., 1998). Cette base peut être localisée sur une seule machine ou sur plusieurs machines réparties sur un ou plusieurs lieux. Cependant, un langage assertionnel de type SQL (Structured Query Language), qui représente l'interface traditionnelle d'interrogation des bases de données relationnelles, n'est pas le mieux adapté (Mendelon A., Wood P., 1995) pour explorer l'information sur une telle structure de données. Cette inadéquation sera plus évidente s'il s'agit d'un site Web multilingues. D'ailleurs, la crédibilité de ce constat est renforcée par le développement massif des travaux en Web sémantique.

Pour Hyperling nous avons retenu le concept du « Web Graph » introduit par (Broder A. *et al.* 2000 ; Mumar R. *et al.* 2000a, 2000b). Les diverses ancres y représentent les noeuds et les arcs sont les hyperliens. La complexité de tel graphe réside essentiellement dans la définition des ancres d'une part et dans la portée des hyperliens d'autre part. Formellement, et pour une ancre donnée, Hyperling ne s'intéresse qu'à son identification et il ne traite pas sa composition. Chaque ancre est représentée par un objet formel.

2.3 Hypothèse de convergence structurelle dans un environnement multilingue

Le fonctionnement d'Hyperling s'appuie sur une hypothèse qui peut paraître assez forte. Elle observe que, dans un site multilingue, il existe beaucoup de relations « internes » ou « locales » entre les parties d'hyperdocuments d'une même langue. De même, notre hypothèse considère qu'il y a peu de relations « externes » entres les parties d'hyperdocuments qui sont rédigées en différentes langues.

Cette hypothèse a été consolidée par la réalisation du système d'analyses statistico-distributionnelles « Hyperling » regroupant le contenu d'un site Web en plusieurs catégories. Le nombre des catégories dominantes représentera le nombre de langues principales sur le site. Les premières expérimentations de ce système ont confirmé la plausibilité de cette hypothèse (Nguyen, T. D, Zreik, K. 2004).

2.4 Représentation des objets formels

Chaque ancre ou objet O_i est représenté sous la forme $O_i=(W_i,ID_i)$, où :

- ID_i est son code d'identification,
- $W_i=(w_{i1},w_{i2},...,w_{in})$ est le vecteur ou graphe associé à O_i. Chaque vecteur est composé de n éléments w_{ij}, où $j\in\{1,2,...n\}$ et n est le nombre total des objets formels dans le graphe. La valeur w_{ij} exprime la fréquence de la coexistence de l'objet O_j avec l'objet O_i dans un contexte distributionnel défini par le graphe.

Ainsi, cette méthode de représentation exige la prise en compte du contexte de chaque objet formel (cf. section 3.1).

3. Le système HYPERLING

L'architecture du système Hyperling est très classique. Elle est composée de trois modules principaux (cf. figure 1).

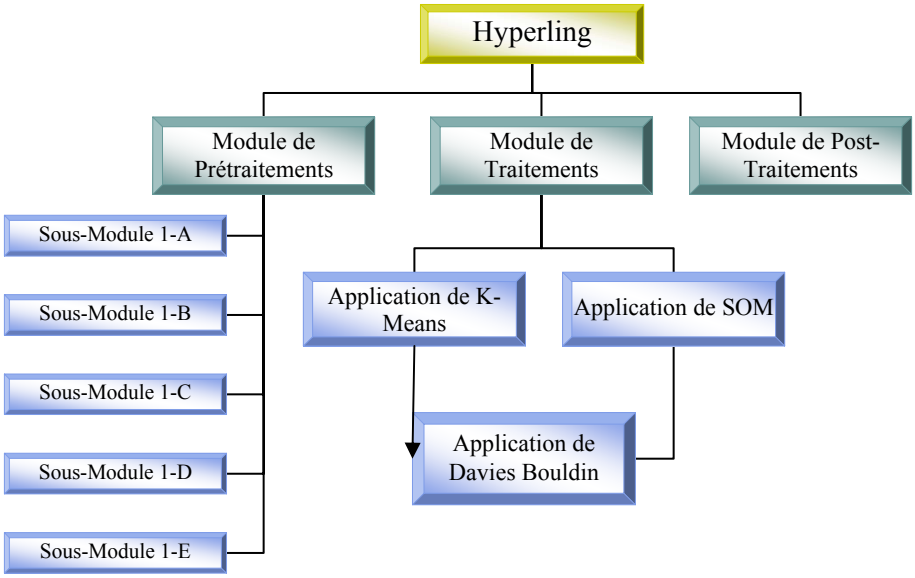


Figure 1 : Schéma général d'Hyperling

⇨ Module 1 : Prétraitement

Ce module est le point clé de succès des processus de catégorisation et d'extraction des caractéristiques multilingues d'un site web. Il comprend 5 sous modules complémentaires.

- Sous-module 1-A : pour parcourir les hyperliens du site pour chercher l'information.
- Sous-module 1-B : pour représenter les hyperliens parcourus en graphe orienté.
- Sous-module 1-C : pour déterminer les contextes distributionnels pour chaque objets formels (nœud du graphe construit)
- Sous-module 1-D : pour représenter sous forme de vecteurs les contextes distributionnels qui ont été extraits.
- Sous-module 1-E : pour réduire la complexité du système en optimisant les dimensions des vecteurs obtenus.

⇨ Module 2 : Traitement

Une fois que l'ensemble des vecteurs ont été extraits et optimisés formellement (une représentation ordonnées et des dimensions réduites) Hyperling propose deux programmes de catégorisation qui sont élaborés à partir de K-means (MacQueen, J., 1967 ; Bottou L., Bengio, Y., 1995) et de Self-Organizing Map (Kangas J. A, Kohonen T., Laaksonen T., 1990 ; Kohonen T., 1995 ; Honkela T., Kaski S., Lagus K. et Kohonen T., 1997). Hyperling propose d'en appliquer l'un ou l'autre ou les deux si on souhaite comparer la qualité des résultats obtenus, par défaut Hyperling applique la méthode issue de K-Means.

Pour interpréter les résultats obtenus par les modules de catégorisation conceptuelle et pour conclure sur les catégories dominantes, nous avons renforcé Hyperling par la méthode Davies-Bouldin (Vesanto J., Alhoniemi E., 2000).

⇨ Module 3 : Post-traitement

Les résultats obtenus par le module de traitement sont soumis à un certain nombre de règles « d'explication » dont le but est de ne pas retenir que les catégories informationnelles, dites déterminantes.

Le nombre de catégories qui sont considérées comme dominantes représentera le nombre de langues présentes sur le site. En figure 2 nous illustrons l'architecture de fonctionnement du système Hyperling.

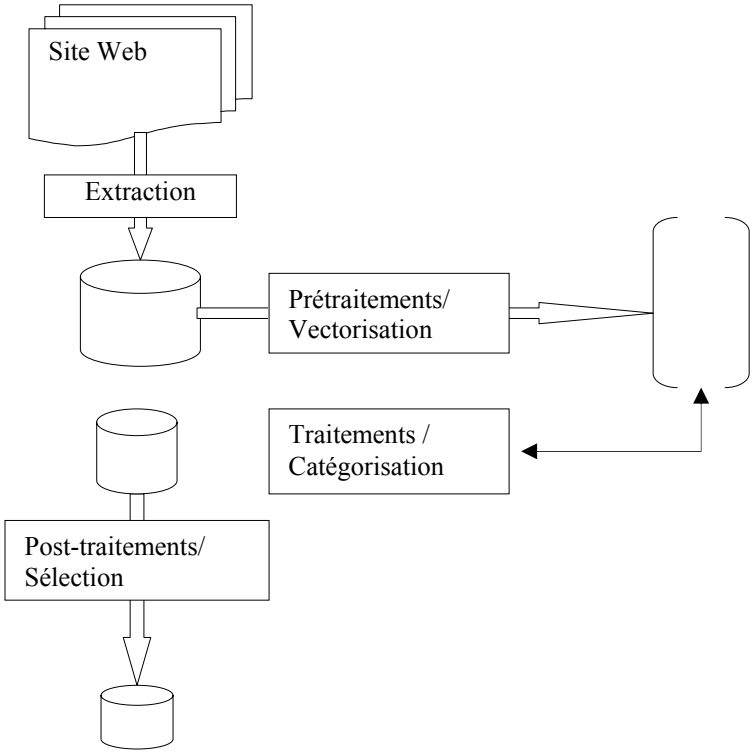


Figure 2 : Architecture de fonctionnement d'Hyperling

3.1 Définition et vectorisation des contextes distributionnels

Définition 1 (**Ensemble** H)

Etant donné l'objet formel $O_i, 1 \le i \le n$, qui est un nœud dans le graphe, nous définissons l'ensemble H comme constitué par tous les chemins partant de l'objet initial O_0 du graphe orienté et passant par O_i.

Définition 2 (**Contextes distributionnels d'un objet formel**)

Etant donné l'objet formel $O_i, 1 \le i \le n$, qui est un nœud dans le graphe orienté, nous définissons l'ensemble C qui détermine les contextes distributionnels de O_i comme suit :

$\forall O_j \in C, i \neq j, 1 \le i, j \le n, \exists h \in H : O_j$

Pour générer tous les chemins qui relient des objets formels, Hyperling respect les contraintes suivantes :

- Tous les chemins commencent par le noeud initial.
- Chaque objet formel est identifié par au moins un chemin.

Pour optimiser les processus et éviter les cycles Hyperling considère que :

- Un objet formel peut paraître une seule fois dans un chemin.
- Un objet formel peut appartenir à plusieurs chemins.

3.2 Optimisation de la représentation des vecteurs

Les projections au hasard, dont la méthode Random Mapping proposée par Kaski S., 1998), ont récemment émergé comme méthode puissante pour la réduction de dimensions des vecteurs obtenus. Les résultats théoriques indiquent que la méthode préserve bien les distances entre les catégories de vecteurs (à définir ultérieurement). Les résultats de la projection des données sur une espace de dimensions réduites peuvent être comparables aux méthodes conventionnelles de réduction de dimensions pratiquées, entre autres, par les méthodes d'analyses de composant principal.

Comme nous l'avons mentionné, la méthode de la projection au hasard préserve la similarité (du point de vu de distances) entre les vecteurs des données. De même, les expériences ont montré que l'application des projections au hasard est sensiblement moins coûteuse que d'autres méthodes, par exemple, celle de l'analyse de composant principal (Bingham E., Mannila H., 2001). Pour ces deux raisons, Hyperling adopte la méthode Random Mapping (Kaski S., 1998).

3.3 Catégorisation et mesures de distance

Hyperling dispose des méthodes de catégorisation dérivées de K-Means (Botton L., Bengio, Y., 1995) et de Self-Organizing Map (Honkela T., Kaski S., Lagus K. et Kohonen T., 1997). Ces deux méthodes sont présentées assez brièvement dans les sections qui suivent (cf. 3.3.1 et 3.3.2). Quelle que soit la méthode choisie, la mesure de la distance entre les vecteurs se réfère aux mesures suggérées par le modèle classique de l'espace vectorielle proposé par (Salton G., McGill M. J., 1983). Pour Hyperling, nous avons retenu la distance Euclidienne pour mesurer la distance et par conséquent la similarité entre les vecteurs.

Nous avons obtenus dans le cadre d'une série d'expérimentation ces deux méthodes (K-Maens et S.O.M) sur le même jeu de sites. Les résultats obtenus sont très similaires et confirment l'hypothèse de convergence.

3.3.1 K-Means

K-Means (MacQueen, J., 1967) propose un algorithme de catégorisation « dur ». Grâce à son efficacité et la facilité de son application (simplicité), K-Means est indiscutablement l'algorithme le plus connu et le plus utilisé dans sa catégorie.

Très brièvement, K-Means peut être définit comme suivant :

- Un ensemble de vecteurs en entrée : $D_m = \{z_1,...,z_m\}, z_i \in R^n$.
- Des prototypes qui représentent par les centres de gravités des catégories à définir : $\mu_k \in R^n, k = 1...K$.
- Une mesure de distance $s(z) = \arg\min_k \|\mu_k - z\|^2$
- Un objectif : pour un vecteur d'entrée z, trouver le prototype «gagnant» (le plus proche) selon $s(z)$.

3.3.2 Self-Organizing Map (SOM)

L'algorithme SOM **(**Kangas J. A, Kohonen T., Laaksonen T., 1990) définit une projection non linéaire à partir d'une espace de R^n sur un tableau de 2-dimensions qui contient M neurones. Les vecteurs en entrée de n-dimensions sont notés r_i. Chaque neurone est connecté à un vecteur de référence de n-dimensions w_i (appelé vecteur référentiel). SOM cherche pour chaque vecteur en entrée le neurone gagnant. Le cas échéant, c'est le neurone le plus proche du vecteur (i.e. le neurone ayant la distance la plus petite parmi celles produites entre le vecteur en question et tous les vecteurs référentiels). L'adaptation des vecteurs référentiels s'effectuera pour le neurone gagnant et ses voisinages. Ces voisinages apprennent aussi à partir de chaque vecteur d'entrée. Cet apprentissage local, se répète autant de fois pour aboutir à un ordre dit global. Ce dernier garanti que les vecteurs proches dans l'espace d'origine de n-dimensions apparaissent aux neurones voisins sur une carte de 2-dimensions. Chaque étape de l'apprentissage consiste en les étapes suivantes :

- Choisir au hasard un vecteur d'entrée, calculer la distance entre ce vecteur et tous les vecteurs référentiels des neurones ;
- Choisir le neurone gagnant, qui a la distance la plus petite par rapport au vecteur d'entrée ;
- Adapter les vecteurs référentiels au neurone gagnant j et à ses neurones voisins, le voisinage du neurone gagnant est défini par la fonction Gaussienne ou par les positions géométriques.

4. Expérimentations

Pour tester et valider nos hypothèse ainsi que les modules implantés dans Hyperling, Nous avons examiné trois sites Web qui sont : http://www.wto.org, http://www.undp.org et http://www.unicef.org. Ces sites sont à vocation internationale par conséquent sont multilingues. Le tableau ci-dessous donne les caractéristiques relatives aux traitements de ces sites.

Sites web consultés	**Taille des sites (en date de)**	Nombre objets formels trouvés	Nombre objets formels vectorisés	**Nombre langues ou catégories dominantes**
WTO http://www.wto.org	≈ 344 MB (10/2004)	27 569	5 557	3 : Anglais, Français, Espagnol
UNDP http://www.undp.org	≈ (10/2004)	11 037	2 162	3 : Anglais, Français, Espagnol
UNICEF http://www.unicef.org	≈ 160MB (5/2003)	12 174	4 582	3 : Anglais, Français, Espagnol

Caractéristiques et résultats obtenus par l'expérimentation

Les trois phases relatives à l'extraction de nombre de langues dominantes sur un site, à savoir le prétraitement, le traitement et le post-traitement sont presque toutes automatisées. Les outils proposés pour les différents modules sont de grande efficacité mais ils restent un peu limités sur le plan ergonomie d'interface. En effet ils ont été développés dans le cadre d'un prototype de recherche pour tester et valider de l'hypothèse de regroupement structurel des hyperdocuments multilingues. Même le nombre d'itérations à effectuer par les méthodes de catégorisation peut être définit par défaut.

Le temps d'exécution d'Hyperling est étroitement lié au volume du site, au nombre d'objets structurels et surtout au nombre d'itérations effectuées par les méthodes de catégorisation.

Le développement d'une interface permettant une interaction plus ergonomique de l'utilisateur d'Hyperling améliorera indiscutablement sa performance en temps et en calcul.

5. Conclusion et perspectives

Hyperling propose une méthode d'exploitation de l'information structurelle de type hyperdocuments. Se référant au Web Graph, Hyperling propose un modèle de représentation basé sur des entités structurelles comme les objets formels présents sur un site et les hyperliens (relations formelles) les reliant.

Nous avons élaboré quelques hypothèses de convergence, en relation avec la définition des objets structurels, que nous avons validées dans un premier temps par l'application de deux algorithmes de catégorisation issus des méthodes K-Means et SOM.

Les premiers résultats obtenus, sur 3 sites multilingues sont très favorables et ne montrent pas d'anomalie. Aussi, ces expérimentations nous ont permit de dresser quelques limitations et un certain nombre de points à étudier afin d'améliorer la performance d'Hyperling.

L'approche retenue est basée essentiellement sur le constat de la distribution statistique ne permettant que d'observer les objets qui sont très fréquents. Or, il serait possible de reprocher cette limitation à Hyperling. En effet l'élargissement de traitement, c'est-à-dire la non prise en compte de critères d'optimisation, pour couvrir l'ensemble des objets structurels, pourrait surmonter cette limitation. La prise de telle mesure serait au détriment de temps de calcul qui risquerait, selon la nature de site en question, d'être plus coûteux. La question d'optimisation de la complexité des algorithmes utilisés devrait pouvoir clarifier l'ampleur de tel élargissement.

La précision des frontières entre les catégories peut être mieux affinée en renforçant le module de post-traitement. Une approche de classification supervisée s'avère être assez appropriée.

Nous tenons à rappeler le caractère expérimental de ce projet. Les algorithmes de traitement, issus essentiellement des travaux en apprentissage automatique, demeurent approximatifs : Ils peuvent montrer la pertinence de la convergence des données statistiquement fréquents mais ils ne sont pas encore capables de démontrer ou de garantir la complétude et la certitude totales des résultats.

La performance d'Hyperling n'a pas été conçue pour traiter des sites monolingues. Comme dans ce type de site les données sont souvent éparses la recherche de catégories dominantes, par Hyperling, devient fastidieuse et les résultats peu signifiant. Nous avons accordé la priorité à ce point et nous espérons avoir des résultats dans un avenir très proche.

Actuellement nous sommes entrain d'étudier les relation d'équivalences et de complémentarités entre les catégories dominantes extraites après le traitement pour savoir s'il s'agit des traductions ou compliments d'information, cas des sites distribués qui sont développés et maintenus dans un cadre collaboratif. Pour cela nous adoptons également une approche statistico-structurelle.

6. Références bibliographiques

Abiteboul S., Quass D., McHugh J., Widom J., Weiner J., "The Lorel Query language for semistructured data". Intl. J. on Digital Libraries, 1(1):68-88, 1997.

Ahonen H., Heinonen O., Klementtinen M., Verkamo A. "Applying data mining techniques for descriptive phrase extraction in digital document collections". Avanced in Digital Libraries (ADL'98), Santa Barbara, California, USA, April 1998.

Asirvatham P.A., Ravi K.K., "Web Page Classification based on Document Structure". International Institute of Information Technology Hyderabad, India, 2001.

Bharat K., Henzinger M.R., "Improved algorithms for topic distillation in a hyperlinked environment". In Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98), 1998.

Billsus D., Pazzani M., "A hybrid user model for news story classification". International Conference on User Modeling (UM'99), Banff, Canada, 1999.

Bingham E., Mannila H., "Random projection in dimensionality reduction: applications to image and text data". In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 2001.

Bottou L., Bengio, Y., "Convergence properties of the K-means algorithm". In Tesauro, G., Touretzky, D., and Leen, T., editors, Advances in Neural Information Processing Systems 7, p. 585–592. MIT Press, Cambridge, MA, 1995.

Bratko, A., Filipi, B., "Exploiting Structural Information in Semi-structured Document Classification", International Electrotechnical and Computer Science Conference (ERK'2004), 2004.

Brin S., Page L., "The autonomy of a large-scale hypertextual Web search engine". Proc. 7th WWW Conf., 1998.

Broder A., Kumar R., Maghoul F., Raghavan S., Rajagopalan P., Stata R., Tomkins A., Wiener J., "Graph structure in the web : experiments and models". International World Wide Web Conference, Amsterdam, The Netherlands, May 2000.

Chakrabarti S., "Data mining for hypertext: A tutorial survey". ACM SIGKDD Explorations, 1(2):1-11, 2000.

Chakrabarti S., Dom B., Indyk P., "Enhanced hypertext categorization using hyperlinks". In Proceedings of the ACM SIGMOD International Conference on Management of Data, Seattle, WA, June 1998. ACM Press.

Chakrabarti S., Dom B., Gibson D., Kleinberg J., Kumar S., Raghavan P., Rajagopalan S., Tomkins A. "Mining the link structure of the world wide web". IEEE Computer, 32(8):60-67, 1999.

Dengel A., Dubiel F., "Clustering and classification of document structure - a machine learning approach". International Conference on Document Analysis and Recognition, 1995.

Džeroski, Lavrač N. 2001, "Relational Data Mining: Inductive Logic Programming for Knowledge Discovery in Databases". Springer-Verlag, 2001.

Fürnkranz J., "Web Structure Mining - Exploiting the Graph Structure of the World-Wide Web". ÖGAI Journal 21(2):17-26, 2002.

Fürnkranz J., "Exploiting structural information for text classification on the WWW". In D. Hand, J.N. Kok, and M. Berthold, editors, Advances in Intelligent Data Analysis:

International Symposium (IDA-99), Amsterdam, The Netherlands, 1999. Springer-Verlag.

Hofmann T., "The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data". International Joint Conference on Artificial Intelligence (IJCAI-99), 1999.

Honkela T., Kaski S., Lagus K., Kohonen T., "WEBSOM self-organizing maps of document collections". In Proceedings of Workshop on Self-Organizing Maps (WSOM'97), Espoo, Finland, June 4-6, 1997.

Kangas J.A., Kohonen T., Laaksonen T., "Variants of self-organizing maps". IEEE Transactions on Neural Networks, 1(1):-99, 1990.

Kaski S., "Dimensionality reduction by random mapping: Fast similarity computation for clustering". In Proceedings of International Joint Conference on Neural Networks (IJCNN'98), 1998.

Kautz H., Selman B., Shah M., "The hidden web". AI magazine, 18(2):27-36, 1997.

Kleinberg J.M., "Authoritative sources in a hyperlinked environment". ACM-SIAM Symposium on Discrete Algorithms, 1998.

Kleinberg J.M., Kumar R., Raghavan P., Rajagopalan S., Tomkins A.S., "The Web as a graph: Measurements, models, and methods". International Conference on Computing and Combinatories (COCOON). Springer-Verlag, 1999.

Kohonen, T., "Self-Organizing Maps". Springer, Berlin, Heidelberg, 1995.

Konopnicki D., Shmueli O., "Information gathering on the World Wide Web: the W3QL query language and the W3QS system". Trans. On Database Systems, 1998.

Kosala R., Blockeel H., "Web Mining Research: A Survey". ACM SIGKDD Explorations, 2000.

Kumar R., Raghavan P., Rajagopalan S., Sivakumar D., Tomkins A., Upfal E., "The web as graph". ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Dallas, TX, May 2000a.

Kumar R., Raghavan P., Rajagopalan S., Sivakumar D., Tomkins A., Upfal E., "Stochastic graph models for the web graph". Annual Symposium on Foundation of Computer Science, Redondo Beach, CA, Nov 2000b.

Ma L., Shepherd J. Nguyen A., "Document Classification via Structure Synopses". Australasian Database Conference (ADC 2003), Adelaide, South Australia, February 2003.

MacQueen J., "Some methods for classification and analysis of multivariate observations". In Le Cam, L. M. and Neyman, J., editors, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, p. 281-297, Berkeley, California. University of California Press, 1967.

Mendelon A., Wood P., "Finding regular simple paths in graph databases". SIAM J. Comp., 24(6):1235-1258, 1995.

Nguyen T.D., Zreik K., "Multilingual Hyperdocument Recognition: a document mining approach". International Conference on Information & Communication Technologies: from Theory to Applications (ICTTA04), Syria, April 2004.

Nguyen T.D., Zreik K., « HYPERLING : Système de reconnaissance et de classification des hyperdocuments multilingues ». Article accepté, International Conference in Computer Science – "Research, Innovation and Vision of the Future" (RIVF 2005), Can-tho, Vietnam, February 2005.

Salton G., McGill M.J., “Introduction to Modern Information Retrieval”. McGraw-Hill, New York, 1983.

Scott S., Matwin S., “Feature engineering for text classification”. International Conference on Machine Learning (ICML-99), 1999.

Theobald M., Schenkel R., Weikum G., “Exploiting structure, annotation, and ontological knowledge for automatic classification of XML data”. WebDB Workshop, 2003.

Vesanto J., Alhoniemi E., “Clustering of the Self-Organizing Map”, Student Member, IEEE, 2000.

Wai-chiu W., Ada Wai-chee F., 2000, “Incremental Document Clustering for Web Page Classification”. Chinese University of Hong Kong, July 2000.

Washerman S., Faust K., “Social Network Analysis”. Cambridge University Press, 1994.

Session 3

Multilinguisme et sciences cognitives

UniTHEM, un exemple de traitement linguistique à couverture multilingue

Nadine Lucas, Emmanuel Giguet

GREYC – UMR 6072 CNRS – Université de Caen
14032 Caen Cedex - France

{nadine.lucas,giguet}@info.unicaen.fr}

Résumé :

Un logiciel d'analyse thématique à couverture multilingue est présenté. Le programme prend en entrée un texte HTML et renvoie en sortie le texte colorié en fonction des thèmes traités, en proposant une vue de la hiérarchie des sous-thèmes. Ce logiciel appelé UniTHEM accepte des langues à écriture alphabétique (langues latines, anglais, … russe) mais aussi les écritures à graphie liée (chinois, japonais). Les limites actuelles de couverture tiennent à des particularités de format d'une part, à la longueur du texte d'autre part. En effet, les textes structurés par des intertitres ne sont pas analysés comme tels. Ces limites montrent que la démarche n'est pas statistique ni basée sur des mots-clés. Elle s'appuie sur un modèle théorique de l'exposition, mis en relation avec des traits stylistiques, ce qui permet l'exploitation de la mise en forme matérielle du document, qui est relativement invariante. Les indices exploités sont communs à des familles d'écriture. Les ressources sont limitées aux séparateurs graphiques. Ces données permettent de constituer une hiérarchie des unités thématiques traitées par recoupements successifs des contextes. La qualité des analyses obtenues est satisfaisante. Les problèmes relatifs à l'évaluation de tels outils sont évoqués.

Mots-clés : recherche d'information, documents multilingues, analyse de texte, mise en forme matérielle, TAL robuste, thématique, Unicode.

Abstract:

This paper introduces a language-free topic parser. The task is to highlight the theme-topic structure and the hierarchy of subtopics in a text. It is performed on newspapers and magazines in French, English and various European

languages, then extended to different writing systems, such as Cyrillic and Chinese. Resources are common separators, punctuation and repeated segments. The algorithm relies on a linguistic model that allows to link stylistic features to the topic structure. The layout of text provides information on the stylistic features.

Keywords: robust text parsing, cross language information retrieval, layout retrieval, topic subtopic hierarchy, Unicode.

1. Introduction

La recherche des "thèmes" traités dans un document est une tâche ordinairement basée sur un traitement statistique des mots-clé censés représenter un concept (ou de termes représentant un concept) [Segond 2002 ; Toussaint 2004]. Ceci présuppose que le mot est une unité universelle [Salton 1989 ; Grefenstette 1998]. Cette vision occidentale domine largement le domaine de la recherche d'information. Cependant, nombre de "langues" ou d'écritures n'utilisent pas le mot graphique, entre autres, le chinois et la majorité des écritures d'Asie, ainsi que l'arabe. On peut s'étonner que la vision dominante ne soit pas davantage contestée par des auteurs issus de traditions scripturales différentes, mais elle l'est parfois (poliment) [Ogawa 1995 ; Chen 1997 ; He 2002].

L'approche ici présentée s'écarte de la vision dominante en informatique en ce que le mot n'est pas posé comme pivot ou atome de sens. L'unité thématique dans un texte n'est pas caractérisée lexicalement, mais plutôt spatialement comme un passage de texte traitant (en principe) d'une même idée. Cette conception est assez proche de la vision de Hearst connue sous le nom de *text tiling* [Hearst 1994, 1997]. Mais, contrairement à celle de Hearst, notre approche est fondée sur un modèle de linguiste. Quoique le résultat soit visuellement similaire, les présupposés sont différents. Le modèle de référence est celui de Yamada (1873-1958), un linguiste japonais qui définit des opérations de mise en discours sans s'appuyer sur le concept de mot (lequel n'est pas matérialisé en japonais). Il présente des opérations qui peuvent être exprimées comme une série de contraintes qui limitent *de facto* un exposé qui serait sinon un développement infini. On va donc délimiter des passages de texte, comme ayant une cohésion interne, mais il faudra aussi hiérarchiser ces passages, puisqu'il est possible qu'un exposé contienne des développements subordonnés ou incis, qui interrompent momentanément la progression du discours. L'intérêt principal d'une telle approche basée sur des considérations stylistiques est qu'elle fait appel à la perception (de la mise en forme du texte). Elle a une portée générale et s'applique sans modification profonde à des textes en chinois ou dans une autre écriture qui ne connaît pas le mot graphique, mais aussi aux langues occidentales.

Du point de vue informatique, notre travail se situe dans le paradigme des analyseurs robustes, à ressources limitées et à couverture multilingue. Considérant le défi du multilinguisme, nous avons recensé les techniques pouvant s'appliquer à des textes non caractérisés en langue. La présente étude s'appuie sur les traitements sans dictionnaire [Vergne 2001], certains concepts comme l'apprentissage robuste générant de nouvelles connaissances (techniques dites d'induction) [Kushmerick 1999] ou la déduction contextuelle [Muslea *et al.* 2002a, b], techniques ordinairement appliquées à l'exploration de la toile, que nous exploitons aussi dans l'analyse des textes. Quoique l'objectif soit celui de la recherche d'information multilingue (*cross language information retrieval*) [Oard 1997 ; Grefenstette 1998; Collins & Singer 1999], nous ne nous focalisons pas sur la recherche de mots ni d'entités nommées. Nous extrayons plutôt les connecteurs [Déjean 2000].

Partant d'un logiciel d'étude destiné à détecter et à colorier les « unités thématiques » dans des articles en français, nous nous sommes attachés ici à tester les possibilités d'adaptation de ce logiciel à un corpus multilingue et même « multi-script », c'est-à-dire traitant des documents non alphabétiques et sans mot graphique. Nous présentons dans la section 2 le logiciel THEMA dans un contexte français, puis ses avatars dans la section 3. Le premier, EuroTHEM, est destiné à des textes européens (Iso-latin), ce qui ne nécessite pas de modifications profondes du logiciel. Le second avatar, UniTHEM, est adapté à des textes que l'on peut traiter sous Unicode, incluant donc les textes en graphie liée, informatiquement gérables sous utf-8. Nous discutons dans la dernière section des caractéristiques permettant la re-définition du multilinguisme dans la perspective du TAL appliqué aux documents.

2. Le fil du discours selon THEMA

2.1 Principes généraux

Le logiciel THEMA est à l'origine un logiciel d'étude, développé pour détecter des unités thématiques. L'objectif de ce travail était de mettre en œuvre une stratégie d'analyse textuelle à l'échelle du document dans des articles courts en français [Pinatel, 2003]. Cette analyse est basée sur des critères rhétoriques et stylistiques, en cela elle ressemble à celles de Kando et Karlgren [Kando 1997 ; Karlgren 2000]. Mais contrairement à ces travaux, nous avons cherché à limiter la dépendance au lexique, qui est un obstacle à la généricité. Le modèle de référence que nous avons choisi est celui de Yamada [Yamada 1936], linguiste et sémioticien souvent évoqué par des chercheurs japonais en RI [Hirakawa 1989 ; Sakamoto & *al.* 2002]. Ce modèle met l'accent sur les procédés d'exposition dans le discours. Il met en valeur le fil du discours, et pour cela traite de divers échelons de la hiérarchie des thèmes (l'enchâssement de thèmes subordonnés) et des unités repérées par rapport au titre, les thèmes dérivés notamment. Pour la première étape, le logiciel devait traiter d'un corpus de textes de vulgarisation en français, donc de textes d'abord facile.

Ce logiciel présente un certain nombre de défauts et de limitations. Cependant, il est suffisamment robuste pour produire des analyses correctes dans la majorité des cas et supporter la comparaison avec des logiciels plus classiques de détection des thèmes [Hernandez & Grau 2002]. Les ressources linguistiques utilisées sont légères. Pour le cas où une marque attendue n'est pas détectée dans le texte entrant, des procédures d'induction, exploitant les ressources endogènes (présentes dans le texte à analyser) sont mises en œuvre, suivant la méthode distributionnelle [Déjean 2002]. Ces propriétés permettent d'envisager une couverture multilingue. Nous avons cherché à tirer parti de la norme ISO-CEI 10646, couramment appelée Unicode, une révolution technologique qui permet d'aborder le traitement informatique de l'écrit avec un degré d'abstraction suffisant [Andries 2002].

2.2 Description du logiciel de base

Les documents fournis au format HTML sont segmentés. La mise en forme matérielle ou MFM [Virbel & Pascual 1996] a une grande importance, puisque le titre et le corps de texte doivent être correctement délimités. Il est nécessaire d'initialiser le traitement en extrayant le titre. Nous retenons comme informations pertinentes la position et la mise en forme différentielle d'une sous-chaîne de caractères, qui caractérisent le titre de l'article. Par différentielle, on entend tout simplement une MFM différente du reste de l'article. De même, la position est remarquable, le titre étant soit le premier élément de l'article, soit un élément isolé placé dans le premier tableau. Ce segment (la chaîne de caractères qui est censée contenir le titre) est posé comme thème de niveau 1 ou G pour global.

Notons toutefois que la segmentation typographique n'est qu'une étape. Le logiciel peut renvoyer comme résultat que le titre et le chapeau d'un article forment le segment thématique au niveau global. Dans la version ici présentée de THEMA, la segmentation interne au corps de texte est restée primitive, autrement dit les textes sont segmentés en unités typographiques fixes : paragraphes, phrases et virgulots (unités délimitées par une virgule). Une version ultérieure du logiciel traite des unités typographiques telles que les sections et chapitres, mais pour l'expérience présente nous nous centrons sur l'aspect multilingue et traitons de textes journalistiques courts.

La segmentation du texte est faite par le module *TextTokenizer*, en unités fixes, paragraphe, phrase et virgulot. Ce module permet également de reconnaître et de garder en mémoire la mise en forme matérielle ainsi que la hiérarchie des composants. Leurs attributs éventuels de MFM (en-ligne) tels que graisse, italique, couleur, sont également mémorisés ainsi que les caractéristiques d'alignement.

La segmentation en paragraphe est une segmentation peu dépendante des familles d'écriture, c'est une délimitation du document HTML. Les codes HTML mal formés forment un obstacle à l'analyse, mais le programme Tidy[1] de Dave

[1] http ://www.w3.org/People/Raggett/tidy.

Raggett permet d'en corriger un grand nombre. La segmentation en paragraphe peut être modifiée dans une fenêtre interactive, les paragraphes que l'on ne considère pas comme faisant partie du corps de texte (mentions éditoriales en particulier) peuvent ainsi être décochés et exclus du traitement ultérieur. Cette fonctionnalité n'est pas utilisée dans les exemples proposés, sauf mention contraire. On verra donc le résultat automatique.

Les opérations sous-jacentes au modèle (mise en facteur d'un thème dans une unité thématique) sont implémentées dans un algorithme de mise en relation correspondant au modèle de Yamada, baptisé *Thematisation*. Nous ne nous attarderons pas ici sur les principes de base de l'algorithme. La structure des exposés est représentée par une série de contraintes sur un développement infini.

Les ressources en mémoire sont les ponctuations et une cinquantaine d'items qui forment une base d'indices dits morphologiques stockés dans une base de données MySQL. Nous avons en effet établi pour le français les marques spécialisées pour la détection d'exposés (marques d'ouverture et clôture de thème, marques de subordination et disjonction). Ces marques sont hiérarchisées. En effet, si l'on trouve deux phrases coordonnées par *De plus*, cela n'a pas la même valeur que deux phrases coordonnées par *Et*. Deux paragraphes coordonnés par *De plus* n'ont pas le même statut que deux phrases coordonnées par ce connecteur, même si en français, la marque de coordination est la même.

Le logiciel ne comporte pas de diagnostic de langue puisqu'il est censé travailler sur le français. Il est important de noter que les ressources en mémoire sont exploitées lorsqu'elles sont présentes dans le texte entrant, mais qu'elles ne sont pas la seule source de connaissances. La déduction contextuelle informée par le modèle, ou si l'on veut, l'induction des marques de bornage (*wrapper induction*) est implémentée.

2.3 La détection des thèmes

Les unités thématiques sont calculées et présentées sur trois niveaux d'inclusion, autrement dit, pour un article normal le premier niveau G subdivise l'article en titre et corps de texte. On suppose ici que le titre est en fonction thématique détachée, et que le corps de texte est le développement (ou propos) qui apporte de l'information (au niveau global). Le corps de texte est subdivisé ensuite en sous-thèmes au niveau G1 et si besoin est, également à un troisième niveau G11 (si le texte est long ou dense). A ces trois niveaux successifs, les informations pertinentes sont analysées automatiquement, à partir d'indices graphiques (MFM), morphologiques (les marques), d'indices positionnels (début et fin) et d'indices de niveau (trois échelons correspondant aux trois grains ou mesures de texte).

Thème G	**Coraux** en mal de **squelette**	
	Thème G1	"Nous avons montré, pour la première fois que, outre son **rôle** dans le réchauffement global, **l'augmentation** de la **concentration** de gaz carbonique (CO2) dans l'atmosphère a un effet négatif sur un écosystème **marin** : les **récifs** coralliens" explique **Jean-Pierre Gattuso**, chargé de recherche au laboratoire Océanographie biologique et écologie du plancton marin de Villefranche-sur-Mer.
	Thème G11	Comment? En réduisant, tout simplement, la possibilité pour les coraux de se constituer un squelette calcaire **à** partir de calcium et de carbonate (CO3). En effet, l'augmentation de CO2 provoque une diminution de la concentration en CO3 néfaste au **processus** de **calcification** : "Nous estimons que celle-ci a diminué de 8% depuis 1880 et qu'elle pourrait encore baisser de 20% d'ici 2065" poursuit le chercheur.

Figure 1 : Le fil du discours déroulé sur trois niveaux d'inclusion

La répétition de certaines sous-chaînes (en gras dans l'exemple ci-dessus) constitue un critère, mais il est important de pouvoir situer des répétitions par rapport à une mesure de texte. En ce sens, les répétitions ne sont pas à entendre comme dans les techniques tf-idf, mais dans une acception stylistique (anaphore et épiphore). La démarche consistant à exploiter des indices morphologiques et des positions est commune aux systèmes de type syntaxique de complexité réduite [Vergne 2001]. Lorsque les indices morphologiques sont absents, l'algorithme emprunte aux systèmes d'apprentissage de règles grammaticales [Déjean 2002]. Cet apprentissage n'est pas itératif sur un corpus, il ressemble donc davantage aux principes dits d'induction, utilisé pour le dépouillement de sites.

Les résultats sont proposés sous plusieurs formes. La « structure compacte » permet de donner une vue globale du texte avec les débuts des unités thématiques, dont le développement est évidé si le texte est long. Les unités thématiques de rang 1, 2 et 3 sont ensuite présentées séparément, pour permettre la vérification de pertinence des relations hiérarchiques. Les unités de rang 1 sont plus longues et moins nombreuses, et on procède ainsi par raffinement successif jusqu'aux plus petites structures détectées qui sont incluses dans les développements. La structure développée représente les unités thématiques et leurs inclusions pour l'ensemble du texte. Nous présentons ici les structures abrégées dites structures compactes.

STRUCTURE THEMATIQUE COMPACTE

Unité Thématique G , niveau 1		
THEME	Coraux en mal de squelette	
RHEME	*Unité Thématique G1 , niveau 2*	
	THEME	"Nous avons montré, [...] chargé de recherche au laboratoire Océanographie biologique et écologie du plancton marin de Villefranche-sur-Mer.
	RHEME	*Unité Thématique G11 , niveau 3*
		THEME Comment? [...] "Nous estimons que celle-ci a diminué de 8% depuis 1880 et qu'elle pourrait encore baisser de 20% d'ici 2065" poursuit le chercheur.
		RHEME Or, [...] PEDRO LIMA

Figure 2 : Résultat de la structuration thématique en français

On y voit que le thème le plus général G (le titre, en bleu clair) est informé par le corps de texte (en vert clair), qui comprend un seul thème de niveau 2 correspondant à une reformulation plus détaillée (bleu plus soutenu), le rhème étant constitué également d'une unité thématique. On voit ici que le dégradé des thèmes n'est pas accompagné par un dégradé des rhèmes, dans un texte monothématique. Autrement dit, les segments thématiques et les segments de clôture restent tous au niveau 3 (vert le plus foncé). Le nom d'auteur n'est pas ramené au niveau du titre. Il s'agit d'une maladresse dans l'algorithme et dans la présentation des résultats. Mais pas seulement.

En effet, le modèle de Yamada n'est pas un emboîtement en poupées russes. Il permet en revanche une distinction entre un texte monothématique (exemple 2) et un texte plurithématique comme dans l'exemple 3 où le décrochement G12 signale l'articulation entre le thème 1 et le thème 2. Notre parti pris dans THEMA a été de mettre en valeur le fil du discours. Une comparaison est proposée avec un autre modèle d'analyse dans la section 4.

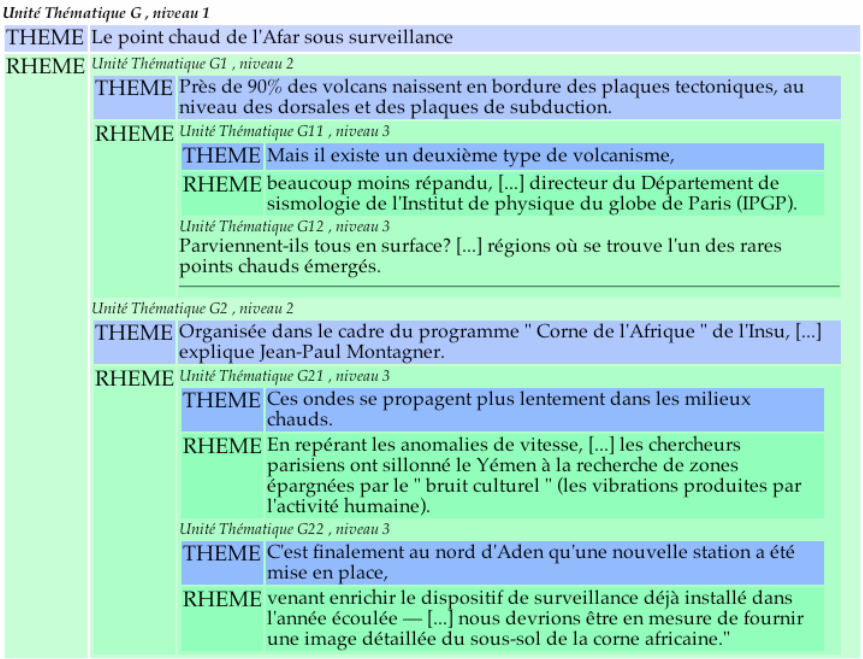


Figure 3 : Résultat de la structuration thématique en français

3. Version « multilingue »

3.1 Langues latines et EuroTHEM

La première extension de THEMA a été la couverture aux langues latines (espagnol, portugais, italien, roumain), ce qui ne nécessitait pas de modification du point de vue informatique. En effet, les tests de détection de connecteurs (établis pour le français) ne sont pas conçus comme décisifs. La variation stylistique est assez grande dans le corpus traité (en français) pour que les formes attendues ne soient pas « obligatoires ». Comme on l'a vu, les traitements par défaut exploitent la répétition d'une forme x inconnue à une position y donnée, c'est donc le procédé stylistique qui est capté et exploité. Cette recherche est contrainte par l'examen des positions remarquables, en position de préfixe (au début d'une unité typographique) ou alors en position de suffixe (en fin d'une unité typographique). On spécifie alors un petit espace de recherche et on recense les chaînes répétées, par exemple dans la dépêche en italien (fig. 4), *partiti*.

STRUCTURE THEMATIQUE COMPACTE

Unité Thématique G , niveau 1	
THEME	Blitz dei partiti: pioggia di soldi in arrivo
RHEME	*Unité Thématique G1 , niveau 2*
	THEME ROMA - Pioggia di miliardi pubblici sui partiti.
	RHEME *Unité Thématique G11 , niveau 3*
	THEME Ieri infatti,
	RHEME con un voto perfettamente bipartisan, [...] compiuto ieri in commissione affari costituzionali.
	Unité Thématique G12 , niveau 3
	THEME "E' stato votato in trenta secondi - racconta il parlamentare - all'unanimità.
	RHEME Mi sono astenuto solo io". [...] il provvedimento diventa immediatamente operativo senza dover passare per l'aula parlamentare.
	Unité Thématique G2 , niveau 2
	THEME Secondo il racconto di Boato, [...] mentre si discutevano questioni di importanza secondaria.
	RHEME *Unité Thématique G21 , niveau 3*
	THEME In poco tempo,
	RHEME il provvedimento è stato presentato e votato, [...] e che il rimborso per ogni voto passa da due a cinque euro.
	Unité Thématique G22 , niveau 3
	THEME Aumenta di cinque volta anche il rimborso per le elezioni regionali,
	RHEME che sale da cinque centesimi a dieci.

Figure 4 : Résultat de la structuration thématique en italien

3.2 Ecritures alphabétiques

Pour étendre la couverture de THEMA aux langues à écriture latine, nous avons dû séparer la détection des paragraphes et celle des unités ponctuées (phrase, virgulot). En effet, la structuration en paragraphe est invariante, tandis que la segmentation en phrases est dépendante d'une graphie liée à une famille de langues. En contexte multilingue, les difficultés de segmentation en phrases tiennent essentiellement aux variations dans les abréviations et dans l'utilisation de la capitalisation [Mikheev 2002 ; Kiss et Strunk 2002, 2004]. L'algorithme proposé par ces auteurs n'est pas implémenté dans notre logiciel, mais cette amélioration est envisagée. La segmentation effectuée par le module "texttokenizer" permet cependant de traiter des textes en anglais, allemand, néerlandais, langues slaves, ainsi que dans les langues latines.

Dans l'étape d'analyse, le logiciel exploite comme précédemment les répétitions pour affecter une valeur de préfixe ou suffixe à des chaînes de caractères. Cette procédure à fondement stylistique permet de délimiter des unités thématiques à partir de ressources strictement endogènes (contenues dans le texte).

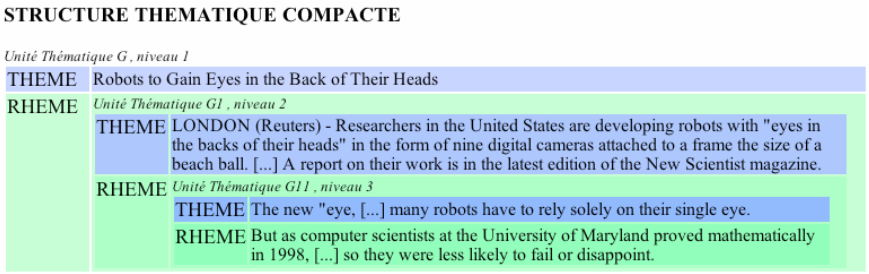


Figure 5 : Résultat de la structuration thématique en anglais

Le russe, plus exactement l'écriture cyrillique, qui détache les mots, ne pose pas de problèmes particuliers du point de vue informatique. Les textes sont segmentés en paragraphes, puis en phrases et en virgulots. Les codes des glyphes correspondant au point de fin de phrase et à la virgule sont les mêmes qu'en Iso-latin.

Figure 6 : Résultat de la structuration thématique en russe

3.3 Version multi-script UniTHEM

Le traitement des caractères en utf-8 est à ce stade indispensable. La détection automatique du jeu de caractères entrant est suivie d'une conversion automatique vers utf-8. Les entités HTML sont également converties vers utf-8. L'ajout de cette fonction dans UniTHEM a pour conséquence un ralentissement de la procédure de segmentation, par rapport à la version en Iso latin, les fichiers sont aussi plus lourds. Le traitement des caractères en utf-8 et les classes d'équivalence de ponctuation nous permet d'aborder des écritures dites à graphie liée, pour des articles en chinois et japonais, qui utilisent des idéogrammes et ne séparent pas les mots. Le coréen, qui ne sépare pas les mots mais utilise un syllabaire n'est pas présenté ici faute de place. Les principes sont les mêmes, mais il est nécessaire de définir des classes de ponctuations [Giguet *& al.* 2000].

Les documents en japonais et chinois sont segmentés en paragraphes, puis en phrases et en virgulots, ce qui nécessite de prendre en compte les codes des glyphes correspondant au point de fin de phrase et à la virgule. On prend également en compte les ponctuations spécifiques de ces écritures (point de mot équivalent au trait d'union). On notera que la segmentation en mots est tout simplement omise, car elle n'est pas indispensable à l'analyse. Contrairement aux traitements lexicographiques, basés sur la consultation de dictionnaires, notre approche exploite des chaînes de caractères répétés à certaines positions et la disposition d'ensemble du texte.

De la même manière que précédemment, nous retenons comme informations pertinentes la position et la mise en forme différentielle de segments de textes ou sous-chaînes, qui délimitent le titre de l'article. Ce segment est posé comme thème de niveau 1. Comme indiqué ci-dessus, dans l'étape d'analyse, le logiciel exploite les répétitions d'une suite de caractères (entre le titre et le corps de texte, et à l'intérieur du corps de texte) pour affecter une valeur à des segments bornés par des préfixes ou suffixes. Les résultats sont tout à fait corrects, avec les mêmes imperfections ou biais que dans les langues latines, mais aussi les mêmes atouts.

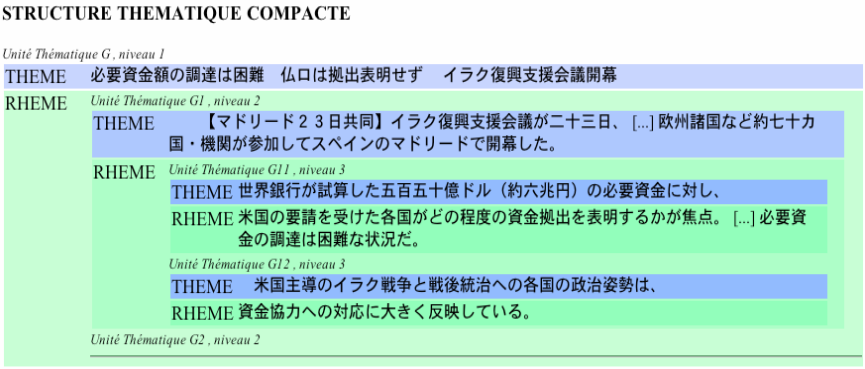


Figure 7 : Résultat de la structuration thématique en japonais

La figure 7 montre que le thème le plus général G est informé par le corps de texte, lui-même subdivisé en un thème de niveau 2 (le 1er §) et un rhème explicatif en deux parties (G11 objectifs et G12 commentaire). Les unités les plus petites au niveau 3 sont subdivisées en virgulots.

Dans la figure 8 on voit que le thème le plus général G (le titre, en bleu clair) est informé par le corps de texte (en vert clair), lui-même subdivisé en deux sous-thèmes (récit des faits et commentaire). Le dernier § est traité comme un ajout ou digression. La signature apparaît comme dernier rhème (les clôtures n'étant pas différenciées des rhèmes).

Figure 8 : Résultat de la structuration thématique en chinois

Figure 9 : Résultat de la structuration thématique en arabe

Les textes en arabe pris sur les sites de presse ne présentent pas une graphie liée. Ils ne nécessitent pas de traitement particulier, à part la gestion des indications de changement de sens d'écriture (*right to left mark*). La visualisation gère le calage à droite des paragraphes.

4. Résultats et évaluation

Le logiciel UniTHEM donne satisfaction pour le traitement des unités thématiques indépendamment des langues et des écritures. Les exemples cités ci-dessus proviennent d'internent et n'ont pas été retouchés manuellement par décochage des paragraphes. Les possibilités d'exploitation des documents sous Unicode nous semblent parfaitement mises en valeur. L'avantage d'UniTHEM est de présenter une vision synthétique du contenu des articles à travers la structure compacte. Cela permet de saisir l'information très rapidement, fournissant ainsi une aide appréciable à la lecture. Une nouvelle voie est ouverte pour aller au-delà des traitements purement statistiques et des traitements linguistiques lexicaux. Le logiciel a été intégré à la plate-forme « wims » [Giguet 2005] et il fonctionne en ligne.

La couverture des formats de documents est imparfaite. Nous sommes limités par le format d'entrée HTML et nous avons encore des difficultés avec certains documents que l'utilitaire Tidy ne permet pas de corriger.

L'évaluation qualitative de tels outils d'analyse thématique pose des problèmes, car le consensus entre différents courants ou experts n'est pas obtenu aussi facilement que sur l'analyse grammaticale par exemple. Notre algorithme étant déterministe, il y a toujours une solution. Cette solution est discutable. Le choix du modèle de référence peut être critiqué, d'autant qu'il n'est pas beaucoup vulgarisé.

Nous avons soumis des résultats d'analyse à des lecteurs externes (langue maternelle à tester) pour juger de leurs réactions sur des ensembles de 10 dépêches ou articles tirés aléatoirement (structure développée et structure compacte). La question posée était « l'analyse est-elle correcte, permet-elle de souligner le fil du discours ? ».

Langues /jugement	**Nbre de textes**	**dont courts**	**non**	**oui**	**ne sait pas**	**nb oui/10**
Chinois	10	5	3	5	2	5/10
Arabe	10	2	8	1	1	1/10
Allemand	10	7	2	8		8/10
Français	20	12	2	17	1	9/10
total	50	26	15	31	3	6/10

Tableau 1 : Jugement des lecteurs sur la pertinence de l'analyse

Le ratio de « oui » indique un bon taux de satisfaction pour les textes courts. Une limitation patente est que l'analyse est faite sur trois niveaux seulement, ce qui est pénalisant pour les textes longs (plus de 10 §). Cela est net pour l'arabe, où la moyenne des paragraphes par article de notre échantillon avoisinait 20. Les hésitations s'observent sur les textes particuliers (interviews, éditoriaux). Enfin, concernant la visualisation, nous envisageons une représentation des segments de clôture différente de celle des rhèmes profonds, car ce défaut suscite aussi des critiques.

Les commentaires d'auteurs d'articles en français rejoignent ceux des lecteurs, les textes longs étant invariablement jugés moins bien analysés que les textes courts. En complément, nous avons soumis les résultats d'analyse à un analyste de presse (3 articles analysés finement pour le roumain). Les commentaires sont positifs, ils discutent surtout du modèle pour l'organisation des unités incluses.

A titre de comparaison pour le traitement informatique, nous avons analysé un texte en français « Le vin jaune » déjà présenté par [Hernandez & Grau 2002]. Par rapport au résultat proposé par ces auteurs (voir annexe, structure incluse en gris), on remarque que la relation disjointe entre le premier et les deux derniers paragraphes (commentés comme introduction et conclusion) n'est pas représentée dans notre analyse. Contrairement au modèle en poupées russes, notre analyse favorise le suivi du fil du discours en partant du titre. La relation « introduction-conclusion » est constitutive de l'unité thématique, aussi le segment conclusif apparaît dans le rhème. Par ailleurs, dans notre analyse, le thème de niveau 2 englobe deux paragraphes, qui introduisent une séquence explicative. Le modèle invoqué par Hernandez & Grau, au contraire, favorise la relation question-réponse, et traite la réponse comme unité distincte.

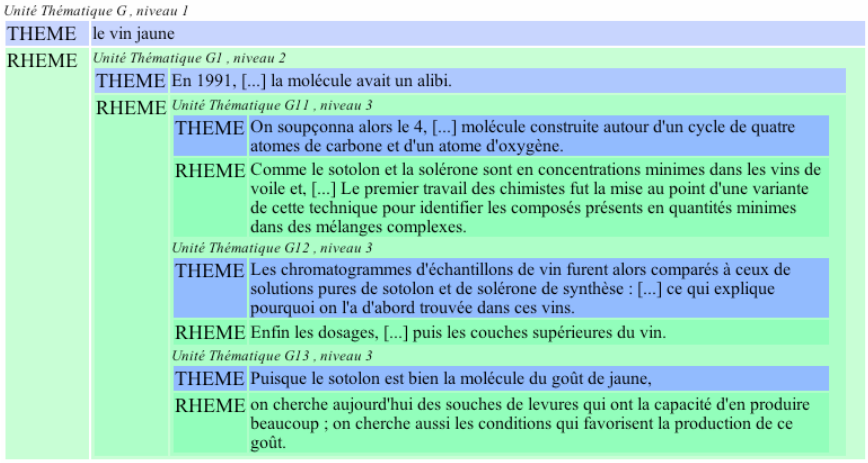


Figure 10 : Résultat de la structuration thématique sur le vin jaune

Les textes que nous avons soumis à l'analyse sont des dépêches de presse ou des articles de journaux et de magazines, des textes fortement structurés du point de vue thématique. La MFM est riche et diversifiée. L'algorithme délimite des unités thématiques à partir de ressources endogènes uniquement. Il serait donc possible de désactiver la consultation des indices morphologiques pour le français. Cependant, disposer d'indices a pour avantage une plus grande rapidité du traitement, car le calcul des répétitions est plus coûteux que la détection d'une forme spécifiée. Il serait préférable de rajouter un diagnostic de langue et des ressources morphologiques spécialisées pour traiter d'un corpus stable du point de vue des langues.

En revanche, pour la couverture d'un événement particulier ou d'un jour J, comme l'a réalisé manuellement une équipe d'analystes [van Dijk 1988], il est intéressant de proposer une analyse thématique informatisée de la presse avec une couverture Unicode, sans restriction due au lexique et à la langue. Le suivi de l'actualité sur la toile est une application naturelle pour le modèle de Yamada, bien adapté au genre journalistique. L'efficacité de cette approche dans d'autres genres est à l'étude.

Le revers de la médaille est que le pré-traitement qui consiste à analyser automatiquement la mise en forme matérielle nécessite un soin méticuleux. Les documents sont formatés de façon très variable, or l'analyse stylistique s'appuie sur la reconnaissance des unités typographiques et de leur rang dans la hiérarchie textuelle.

5. Discussion

La voie d'approche stylistique est explorée notamment par Karlgren dans l'analyse de la presse [Karlgren 2000, Karlgren & Järvinen 2002]. Les principes informatiques de base retenus dans notre approche sont hérités de la recherche d'information sur la toile, une situation où la langue des sites est inconnue, les règles de formatage également [Stienne & Lucas 2003]. On a donc recours à une alternance de déduction et d'induction à partir d'indices endogènes, dont on cherche à établir le statut informationnel [Mukerjea 1998 ; Muslea & *al.* 2002a, 2002b]. Il est tentant de relier le traitement des sites et le traitement du contenu en utilisant les mêmes principes discriminatoires à partir de ressources endogènes. La source exogène de connaissance permettant de faire des choix est alors le type de sortie requis par l'utilisateur (choix de la grille d'analyse, du niveau de détail requis, etc.).

L'analyse de documents et la fouille de textes présentent aussi des convergences. La caractérisation différentielle de sites (classés en deux catégories qualitatives d'après leur contenu) a été menée en France suivant des principes structuralistes exploitant des faisceaux de traits [Valette 2004]. La valeur accordée aux indices matériels (ponctuations, couleurs) est dépendante du contexte et des différentiels observables. Pour notre objectif, l'affectation de valeur est dirigée par le modèle psycho-linguistique de Yamada. On pourrait dire que la classification des

segments de textes est faite par UniTHEM en autant de classes que le modèle en requiert (au moins 2), et se raffine en autant de classes que le modèle le permet (24 si on se limite à trois niveaux d'inclusion).

Pour traiter une autre problématique, par exemple la détection de l'argumentation ou celle des explications, il serait nécessaire d'exploiter un autre modèle de référence et donc d'écrire un algorithme différent. Nous jugeons préférable d'exploiter un seul modèle de référence à la fois.

L'expérience de traitement thématique à couverture multilingue forme un jalon dans une recherche émergente, permettant l'interprétation de données textuelles indépendamment des langues et des écritures (ou presque). L'homogénéité de la norme ISO-CEI 10646, qui utilise un jeu unifié de caractères, sert de socle technologique. Unicode permet de mener des traitements robustes sur un corpus multilingue ou indifférencié en langues, en s'appuyant sur quelques données internes au code ; les tables d'équivalence de glyphes, notamment les ponctuations, deviennent la « pierre de Rosette » du traitement pour UniTHEM. La généricité de l'approche tient à l'exploitation des balises et des graphies qui ont un usage plus large que le lexique d'une langue particulière. Nous souhaitons suivre à l'avenir cette voie d'exploration prometteuse.

6. Références bibliographiques

[Andries 2002] "Introduction à Unicode et à l'ISO 10646", P. Andries, *Document numérique*, vol. 6, n° 3-4. (2002), p. 51-88.

[Chen 1997] "Chinese Text Retrieval without using a Dictionary", A. Chen & J. He, *SIGIR*, 1997.

[Collins et Singer 1999] "Unsupervised models for named entity classification", M. Collins and Y. Singer, *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[Déjean 2000] "ALLiS: a Machine Learning System for Natural Language Learning", H. Déjean, *Conference on Natural Language Learning*, Lisbon, 2000.

[Déjean 2002] "Learning Rules and Their Exceptions", H. Déjean, *Journal of Machine Learning Research*, 2: 669-693, (2002).

[Giguet 2005] "Modélisation de l'activité expérimentale du chercheur en traitement des langues sur corpus multilingues", E. Giguet, *Journée de l'ATALA*, Articuler les traitements sur corpus, 12 février 2005.

[Giguet & al. 2000] "Document structure identification illustrated on news dispatches", E. Giguet, N. Lucas & G. Cousin, *CicLing-2000*, A. Gelbukh (Ed), Mexico, Instituto politécnico nacional, 2000, p. 415-428.

[Grefenstette 1998] Cross-Language Information Retrieval, G. Grefenstette (Ed), Kluwer, 1998.

[He 2002] Finding the Better Indexing units for Chinese Information Retrieval., H. He, P. He, J. Gao, & C. Huang, *First SigHAN Workshop on Chinese Language Processing*, 2002.

[Hearst 1994] "Multi-Paragraph Segmentation of Expository Text", M. Hearst, *32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM, June 1994.

[Hearst 1997] "TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages", M. Hearst, *Computational Linguistics*, 23:1, (1997), p. 33-64.

[Hernandez & Grau 2002] "Analyse thématique du discours : segmentation, structuration, description et représentation", N. Hernandez & B. Grau, In *Cide 5*, Tunis (2002).

[Kando 1999] "Text Structure Analysis as a Tool to Make Retrieved Documents Usable", N. Kando, *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages*, Taipei, Taiwan, nov. 11-12, 1999, p. 126-132.

[Karlgren 2000] Stylistic Experiments for Information Retrieval., J. Karlgren, PhD thesis, Stockholm, Université de Stockholm, 2000.

[Karlgren & Järvinen 2002] "Foreground and background text in retrieval", J. Karlgren & T. Järvinen, In Karlgren, J., Gambäck, B. Kanerva, P. (Eds), *Acquiring (and Using] Linguistic (and World) Knowledge for Information Access*, Menlo Park, Spring 2002, The American Association of Artificial Intelligence.

[Kiss & Strunk 2002] "Viewing sentence boundary detection as collocation identification", *Proceedings of KONVENS 2002*, Saarbrücken, p. 75-82.

[Kiss & Strunk 2003] "Multilingual Least Effort Sentence Boundary Detection", Tibor Kiss, Jan Strunk Under review.

[Kushmerick 2000] "Wrapper induction: Efficiency and Expressiveness", N. Kushmerick, *Artificial Intelligence 118*, 2000.

[Mikheev 2002] "Periods, Capitalized Words, etc.", Andrei Mikheev, *Computation Linguistics*, 28:3, (2002), p. 289-318.

[Mukherjea 2000] "WTMS: A System for Collecting and Analysing Topic-Specific Web Information", S. Mukherjea, *Proceedings of the 9th International World Wide Web Conference*, Amsterdam, Netherlands, May 15-19, 2000. http://www9.org/w9cdrom/293/293.html

[Muslea *et al.* 2002 a] "Adaptive view validation: a case study on wrapper induction", I. Muslea, S. Minton, C. Knoblock, *Proceedings 19th ICML*, 2002.

[Muslea *et al.* 2002 b] "Active+ Semi-Supervised Learning = Robust Multi-View Learning", I. Muslea, S. Minton, C. Knoblock, *Proceedings 19th ICML*, 2002.

[Oard 1997] "Alternative Approaches for Cross-Language Text Retrieval," D. W. Oard, in Cross-Language Text and Speech Retrieval, AAAI Technical Report SS-97-05. http://www.clis.umd.edu/dlrg/filter/sss/papers

[Ogawa 1995] "A new characterbased indexing organization using frequency data for Japanese documents", Y. Ogawa, *Conference on Research and Development in Information Retrieval*, ACMSIGIR, Seattle, 1995, p. 121-129.

[Pinatel, 2003] *Coloriage thématique à l'intérieur d'un document : approche contextuelle*, P. Pinatel, Rapport de projet DESS RADI, Université de Caen, 2003.

[Sakamoto & al. 2002] "Knowledge Discovery from Semistructured Texts", H. Sakamoto, H. Arimura, & S. Arikawa, in Arikawa & Shinohara (eds), *Progress in Discovery Science (LNAI 2281)*, Springer, 2002, p. 586-599.

[Salton 1989] *Automatic text processing*, G. Salton, Addison-Wesley, 1989.

[Segond 2002] *Multilinguisme et traitement de l'information*, F. Segond (Éd.), Paris, Hermès Lavoisier, 2002.

[Stienne & Lucas 2003] "Exploitation d'information disponible sur Internet et génération d'un portail multilingue sur la thématique cinéma", N. Stienne & N. Lucas, *Cide 6*, Faure et Madelaine (éd.), *Document électronique dynamique*, Paris, Europia, 2003, p. 239-255.

[Toussaint 2004] "Extraction de connaissances à partir de textes structurés", Y. Toussaint, *Document numérique*, vol. 8:3, (2004), p. 11-34.

[Valette 2004] « Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet », M. Valette, *CIDE.7*, P. Enjalbert, M. Gaio (éd.), p. 215-230.

[Van Dijk 1988] *News analysis: case studies of international and national news in the press*, T. A. van Dijk Hillsdale, N. J., L. Erlbaum, 1988.

[Vergne 2001] Analyse syntaxique automatique de langue: du combinatoire au calculatoire, J. Vergne, *TALN 2001*, Tours, vol. 1, (2001), p. 15-29.

[Virbel & Pascual 1996] "Semantic and Layout Properties of Text Punctuation", J. Virbel & E. Pascual, In *Proceedings of the workshop on Punctuation in Computational Linguistics*, ACL Conference, Santa Cruz, 1996.

[Yamada 1936] Nihon bunpôgaku gairon [*Somme sur la grammaire japonaise*], Y. Yamada, Tôkyô, Hôbunkan, 1936 (ré-imp. 1989).

Le document dans son agir organisationnel : le modèle de l'organisation dans l'interaction usager système

Maryvonne Holzem[1], Dominique Dionisi[2], Jacques Labiche[2], Eric Trupin[2]

[1] *Dyalang – FRE 2787 CNRS – Université de Rouen – France*
Maryvonne.Holzem@univ-rouen.fr

2 PSI – FRE 2645 CNRS – Université de Rouen – France
Dominique.Dionisi@insa-rouen.fr
{Jacques.Labiche,Eric.Trupin}@univ-rouen.fr

Résumé :

Cet article poursuit, à partir d'un ancrage en sciences du langage et sciences pour l'ingénieur, un travail de recherche pluridisciplinaire mené dans le cadre d'une action spécifique Document et Organisation sous l'égide du RTP33 sur le document. Il se donne pour objectif de démontrer l'apport du contexte organisationnel pour le traitement du document. Apport qui jette, selon nous, les bases de la formalisation de nouveaux usages dans le contexte du document numérique.

Mots-clés : Document, acteurs situés, acculturation, intertextualité, cycle de vie, agir organisationnel., modélisation de l'expérience.

Abstract:

From an anchor position in the sciences of linguistics and engineering, this article continues with a pluridisciplinary research work led in the framework of a specific Document and Organisation action under the ægis of the RTP33 on the document. It aims at demonstrating the contribution of the organisational context to document processing, contribution which, we

believe, lays the basis of the formalisation of new practices within the context of a numerical document.

Keywords: Document, set actors, scientific integration, intertextuality, life cycle, organisational acting, experience modelisation.

Préambule

Nous forgeons le terme d'agir organisationnel sur le modèle de l'*agir communicationnel* emprunté à Jürgen Harbermas [HAB87] pour rappeler, à l'ère du *village planétaire* des réseaux de communication et du mouvement général d'ontologisation qui l'accompagne, que toute représentation l'est toujours de quelque chose qui lui préexiste et qu'il est alors nécessaire de prendre en compte le construit social et l'intersubjectivité : autrement dit l'expérience anthropologique comme éléments constitutifs du sens et condition d'une acculturation[1] des connaissances.

1. Introduction

Présentation de l'article et de sa finalité

Cet article s'inscrit dans la poursuite, d'un travail ayant réuni Sciences pour l'Ingénieur et Sciences Humaines et Sociales dans le cadre d'une action spécifique du CNRS intitulée Document et Organisation (cf. ci-dessous). Notre ancrage linguistique s'ouvre aux pistes de recherches pluridisciplinaires à approfondir avec la sociologie des organisations, les sciences du document[2] et celles pour l'ingénieur. En premier lieu, il rendra compte de la démarche adoptée de l'analyse des modes d'écriture re-écriture de documents dans deux contextes différents. Puis, en démontrant l'intérêt d'étudier du document au sein d'un collectif de travail, il se focalisera sur les perspectives que semble offrir une approche par le biais de l'organisation. En guise de conclusion, nous pourrons, à notre tour, dans la lignée du débat mené sous l'égide de Roger T. Pedauque[3], esquisser quelques perspectives pour des recherches ultérieures, permettant de mieux appréhender la notion de

[1] Le terme acculturation qui signifie : processus par lequel un groupe humain assimile tout ou partie des valeurs culturelles d'un autre groupe, recouvre, selon nous, les termes de circulation des connaissances, d'altérité, de points de vue, de reformulation et d'appropriation des savoirs.

[2] Science à définir ou plutôt à redéfinir à l'heure du numérique. Cet article a pour but d'y apporter notre contribution.

[3] Texte élaboré collectivement dans le cadre du réseau thématique pluridisciplinaire du CNRS (cf. : http://rtp-doc.enssib.fr/pedauque/historique.htm).

document numérique par le biais de ces cycles de vie et des contraintes spécifiques qui l'accompagnent dans des espaces et à des moments donnés.

2. L'apport du contexte organisationnel pour le traitement du document : compte rendu d'une démarche empirique

2.1 Une investigation pluridisciplinaire : apport du contexte organisationnel pour le traitement du document

Sans vouloir, nous appesantir sur les difficultés qui ont amené les chercheurs travaillant sur la reconnaissance de formes à contraindre leur approche de l'interprétation assistée du document, nous retracerons le chemin qui a conduit à une collaboration pluridisciplinaire. Dans la suite de la problématique de la rétroconversion du document « papier » à partir de sa numérisation, qui a permis aux chercheurs de mettre au point et tester des outils de traitement d'images et de reconnaissance de formes pour extraire tous les objets présents sur le document puis les archiver dans des bases de données, il est apparu nécessaire d'utiliser des modèles de documents et d'interfaces homme machine capables de guider intelligemment la reconnaissance de formes, l'extraction puis l'interprétation d'information. Afin d'affiner, préciser et opérationnaliser ces modèles et interfaces, il est devenu indispensable aux Sciences pour l'Ingénieur (*SPI*) (laboratoire PSI de Rouen) de coopérer, avec les Sciences Humaines et Sociales (*SHS*) la sociolinguistique (Dyalang) la sociologie du travail (CRIS) et les sciences de l'information (CRESI-ENSSIB)). Ces domaines des SHS analysent avec leurs focales spécifiques les pratiques professionnelles et les organisations en contexte.

La sociologie du travail et des techniques s'est donnée pour tâche d'analyser les pratiques professionnelles implicites et explicites en identifiant, si besoin, des groupes professionnels grâce à l'analyse de l'organisation de l'entreprise. Elle vise à montrer comment les professionnels vont donner sens aux objets techniques, en tant que construit social, dans le cadre de leurs pratiques professionnelles.

La sociolinguistique se focalisant sur les pratiques langagières au travail à partir d'un corpus d'écrits professionnels, postule que l'organisation est un contexte au sens d'une situation de production, de circulation et de consommation de document. Elle fait l'hypothèse que le contexte construit une interprétation variable selon différentes sphères d'activité[4]. En repérant les marqueurs de sa position énonciative, le sociolinguiste appréhende un auteur lambda en acteur situé.

[4] Nous préférons parler de "sphère d'activité" plutôt que de domaine qui sous-entend l'idée d'une communauté langagière a priori (celle des chercheurs, des ingénieurs, des techniciens, des gestionnaires, etc.) là où les pratiques quotidiennes dans le monde de la

Les sciences de l'information travaillent sur les opérations et les stratégies de mise en forme et de mise en circulation des documents, ainsi que sur les dispositifs mis en place par les individus et les organisations. Elles reprennent par la question de gestion des flux la notion d'information située et de modalités d'usage.

La confrontation de ces différents champs nous permet d'avancer dans la prise en compte d'acteurs situés, des méthodes de travail, des ressources et normes en vigueur au sein de sphères d'activité. Elle offre une meilleure identification des rôles et des stratégies mises en œuvre. Elle conforte les évolutions des modèles en prenant en compte, par le biais de la nécessaire adaptation au contexte, le caractère fondamentalement dynamique des systèmes à venir et des contextes organisationnels comme cadre de l'interaction usager système.

Malgré le handicap des présupposés et vocabulaires différents selon les champs disciplinaires en présence, l'objectif partagé était de pouvoir modéliser l'organisation et ses interactions avec le document pour cerner le statut d'un type de document. Ce document dénommé au départ : d*ocument organisation* puis *métier*, est devenu au cours d'études de terrains, *document activité*. Quoiqu'il en soit, ce document appréhendé dans son agir organisationnel contribue, selon nous, à mieux cerner le statut social du document, notamment numérique, comme nous allons tenter de la démontrer ci-dessous.

2.2 Brève description des études menées et démarche empirique adoptée

Confrontés aux difficultés inhérentes à la pluridisciplinarité, nous avons choisi une démarche empirique qui, en contraignant les équipes disciplinaires à étudier avec leurs propres méthodes et leur propres présupposés scientifiques, les mêmes objets dans les mêmes situations, était susceptible de faire émerger une véritable interdisciplinarité. Ce sont donc deux mêmes études de terrain qui fédèrent notre démarche.

Nous avons décidé d'explorer deux sphères organisationnelles distinctes (collectifs de travail) : celle des travailleurs sociaux au sein d'un CIAD (Centre d'Information et d'Accueil Départemental) situé en banlieue rouennaise et celle des ingénieurs brevet dans un cabinet conseil en brevets d'invention (région parisienne).

2.2.1 Le formulaire RMI

Dans le premier cas, l'étude porte sur le formulaire de demande d'un revenu minimum d'insertion (RMI). L'enquête sur le terrain a consisté, d'une part, à

---

recherche comme celui de l'industrie font état d'un multi-partenariat autour de tâches à réaliser.

enregistrer les entretiens entre une Assistante Sociale et un demandeur (entretien durant lequel est instruit le document), puis d'autre part, à enquêter auprès des mêmes Assistantes Sociales à propos de leur pratique professionnelle relative à cette tâche. Le corpus étudié a donc été fourni par la retranscription de ces deux types d'entretien. L'approche de terrain, menée depuis la sociologie du travail (Martine Blanc-Merigot, Catherine Peyrard), les sciences de l'information (Marie-France Peyrelong) et les sciences du langage (Valérie Delavigne) a tenté d'articuler les dispositifs scripturaux (par l'étude du formulaire), les pratiques professionnelles (par l'observation in situ des entretiens et l'interview des professionnels) et les pratiques discursives (par l'analyse de la retranscription d'entretiens) [BLA04]. Cette étude de la mise en mots puis de la mise en document, saisie au niveau du formulaire, met en évidence tout le travail de traduction et de réduction opéré par les assistantes sociales. Il révèle que le document lui-même (le formulaire RMI rempli lors de l'entretien) n'est qu'un élément d'objectivation de la demande et de la personne et qu'il est profondément *encapsulé* dans les activités de trame (inscrites dans le statut professionnel) et de chaîne (ou d'accompagnement, de traitement de la demande). Il faut donc chercher dans l'interaction, dans l'entour, le contexte, ce qui va donner sens, permettre les décisions (actions), légitimer une situation personnelle en un demandeur puis enfin, en un ayant droit.

2.2.2 La demande brevet

Dans le second cas, le corpus écrit nous a été fourni d'emblée, puisqu'il s'agissait d'un mémoire technique rédigé par une équipe de chercheurs travaillant en reconnaissance de formes et de sa demande brevet correspondante. Celle-ci a été rédigée par un ingénieur brevet qui a été interviewé ainsi que deux de ses collègues, tous rédacteurs confirmés en brevet d'invention. Ces interviews nous ont alors permis d'affiner notre approche du document brevet dans son contexte organisationnel. Brigitte Guyot, pour la sociologie de l'information, et Sylvie Normand, pour les sciences du langage, ont étudié la place du document dans le secteur de la propriété industrielle en s'attachant à la description de la transformation éditoriale d'un mémoire technique en brevet d'invention [GUY04]. Leur approche disciplinaire se croise au point de passage du monde scientifique au monde juridique avec la prise en compte des marques énonciatives et de la stratégie argumentative, d'une part, de la dynamique sociale et d'une analyse sociologique des pratiques des acteurs aux prises avec l'information, d'autre part. Leur étude du document, doublée d'entretiens avec des ingénieurs brevet, montre à quel point celui-ci est indissociable de l'organisation au sein de laquelle il est inséré et contraint pour être transformé en formes opérationnelles juridiquement incontestables.

2.2.3 Modélisation résultante

Un examen de la pratique autour des demandes RMI peut être retracé à l'aide du diagramme d'état UML de la figure 1 :

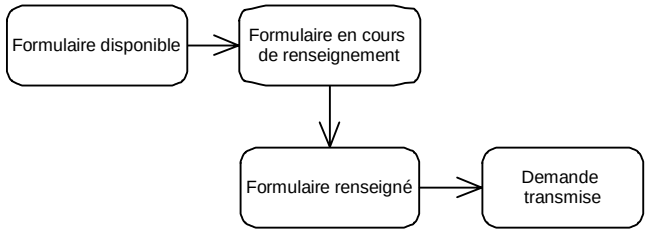


Figure 1

Le même exercice sur les documents brevet produit le diagramme d'état de la figure ci-dessous.

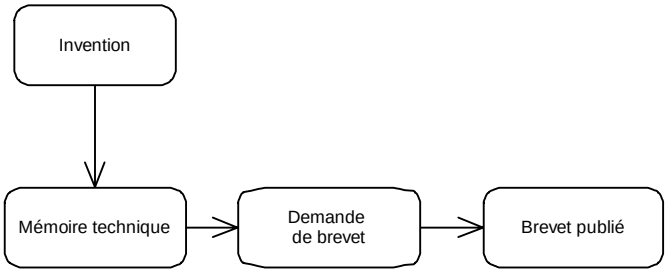


Figure 2

Il s'agit ici d'une représentation très superficielle destinée à exemplariser le propos. Il va de soi qu'une étude réelle irait à un niveau de détails et de granularité beaucoup plus avancé. De ces diagrammes d'états, nous pouvons tirer les représentations d'activités suivantes pour le Rmi (figure 3) et pour le brevet (figure 4).

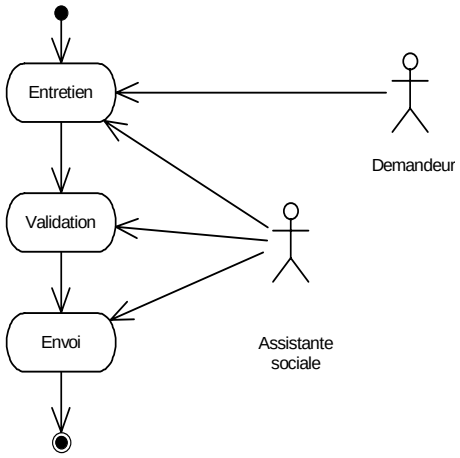


Figure 3

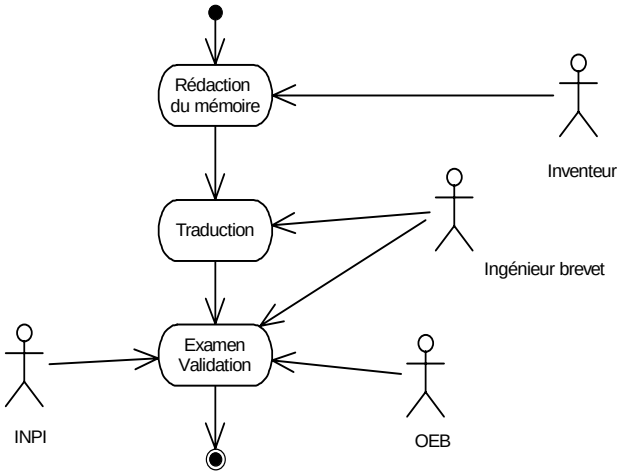


Figure 4

La question soulevée est celle des collaborations entre objets et acteurs permettant la réalisation de ces activités. Ces collaborations permettent entre autres de s'assurer d'un partage consensuel de sens sur les objets partagés par l'ensemble de la communauté concernée.

3. Un premier bilan pour mieux cerner l'intérêt de la mise en situation du document dans son contexte organisationnel

3.1 Spécificités des terrains

Outre les difficultés d'approche, bien connues, des sphères organisationnelles et des documents qui y circulent « en interne », les premiers éléments de conclusion auxquels nous sommes parvenus ont permis de dégager certes des particularités liées à des situations organisationnelles disparates, mais aussi, et surtout des points de convergences[5] nous permettant de mieux cerner la spécificité et l'intérêt d'une approche du document par le biais des sphères organisationnelles (collectifs de travail).

Au niveau des spécificités des terrains, en ce qui concerne la demande RMI, le bilan établi montre en particulier que la rédaction médiée d'un formulaire comme celui du RMI peut n'être qu'un prétexte à rencontre entre acteurs sociaux et usagers. Le statut de ce formulaire et les actions que son contenu peut déclencher ne peuvent plus s'analyser comme une simple réaction réflexe, mais doivent au contraire être examinés dans le cadre complexe d'une institution et de ses différents services : des actions seront déclenchées à des niveaux différents (qui s'ignorent) pour statuer sur le devenir d'un usager en difficulté. Ceci amène à réfléchir sur les modèles du document car les éléments du contenu ne peuvent plus être envisagés comme de simples éléments déclenchant des actions simples.

Par ailleurs les contraintes sociales ici mises en évidence, loin de permettre de catégoriser les intentions possibles, révèlent des intentions cachées des acteurs de l'organisation (ou du moins des intentions ignorées des usagers).

Cela signifie que dans nos modèles, actions déclenchées et intentions initiales sont plus complexes que nous l'avions imaginé.

En ce qui concerne le terrain de la propriété industrielle, c'est l'activité éditoriale qui a été particulièrement prise en compte, car elle consiste à transformer un texte scientifique original et à le rendre compatible avec le fonctionnement du secteur de la propriété industrielle. Nous avons approché cette activité en suivant à la fois le processus qu'elle engage, ce qui aide à repérer les interventions de chaque acteur et leurs interactions, et ses composantes ; les principales opérations étant centrées sur l'écriture et la lecture (pour comparer, compulser, faire des liens et de recherche d'information, repérer) et, surtout, la négociation sur l'axe de l'intertextualité. Les actions menées en contexte organisationnel ont été appréhendées comme des rôles, ce qui permet de préserver la notion de pluralité des

[5] Holzem, Maryvonne, Labiche, Jacques éd., 2004 Document et Organisation : forum document et organisation, semaine du document numérique, La Rochelle, Juin 2004, Paris, Europia, 101 p.

activités (un acteur pouvant avoir plusieurs rôles), mais aussi de mieux cerner la notion d'activité en fonction de l'identification d'un certain nombre de rôles.

3.2 Une analyse sociolinguistique des deux terrains

En ce qui concerne les points de convergences nous avons pu, dans les deux cas, rendre compte d'une montée en généricité, perceptible sur le plan lexical du point de vue linguistique, comme accompagnement d'un processus de légitimation sociale.

Le document, tout en étant le dépositaire de stratégies mises en œuvre en fonction de la sphère visée, n'est est pas le seul témoin. *Il faut pouvoir*, comme le remarque B. Guyot et S. Normand, à propos d'une étude sur le brevet d'invention, *se déplacer vers une analyse sociologique des pratiques des acteurs aux prises avec l'information* [GUY 2004].

Du point de vue de l'analyse linguistique, le document situé dans une pratique sociale offre toutes les caractéristiques du document technique et du discours procédural telles que l'ont analysé linguistes et psycho-sociologues. [HEU01, FAY02, ADA01]. Ces discours, à forte valeur illocutoire du point de vue des stratégies énonciatrices, sont à la fois positivés (absence de vrai négation[6]) et fortement répétitifs. Ils n'ont pas pour finalité d'être lus mais actés. Ils visent à régler un comportement social et de ce point de vue le respect des consignes structurelles et scripturales constitue l'un des enjeux sociaux de la lecture. Enfin, et surtout, ces textes sont à la jonction entre l'action verbale et l'action dans le monde [Hol 2004]. Au regard de la théorie de l'agir communicationnel d'Habermas [opus cité] leur spécificité sera sans doute d'être au carrefour entre l'agir stratégique, qui vise l'efficacité (justifiant sur l'axe spatio-temporel, de la place et du rôle de l'organisation via ses agents) et de l'agir communicationnel orienté vers l'intercompréhension et la recherche du consensus. Sur cet axe de l'agir stratégique, les acteurs situés agissent en quelque sorte *es qualité* : leur activité sera alors fortement monologique (et monologale au sens d'une même voix dans le cadre professionnel). Sur celui de l'agir communicationnel, orienté vers l'entour du document mais aussi de l'organisation, celle-ci sera dialogique au sens de la prise en compte de l'intersubjectivité (points de vue) et de la dimension intertextuelle de tout document [BAK84, RAS01].

---

[6] Si la forme négative intervient au cours de l'entretien RMI : vous n'avez pas travaillé, comme le souligne l'assistante sociale, la négation disparaît obligatoirement du formulaire étant donné que le but visé par l'entretien est d'objectiver le demandeur en lui trouvant, par déduction, une catégorie correspondant à sa situation : s'il n'a pas travaillé, c'est qu'il est encore étudiant, par exemple. Dans le processus de transformation d'un mémoire technique en demande brevet, la marque de la négation apparaît mais de façon positivée, afin de démontrer que l'invention ne présente pas ces inconvénients.

3.3 Des pistes de recherches à approfondir

Ce premier bilan, nous amène aujourd'hui à dégager des pistes de recherche à approfondir. Faut-il considérer le modèle de l'organisation comme intermédiaire dans l'interaction usager-système ? Autrement dit que pouvons nous attendre d'une approche du document par le biais organisationnel. Nous détaillons ci-dessous quelques points méritant un approfondissement.

1. N'est-il pas nécessaire d'évoquer le document sous l'angle d'une dynamique textuelle (transformations, lignée de réécriture) en intégrant une ou plusieurs collections dans le réseau intertextuel d'une communauté d'intérêt. Cette question de l'intertextualité [KRI69] est étroitement liée au dialogisme bakhtinien au sens où tout texte est une combinatoire, un lieu d'échange constant entre fragments à partir de textes antérieurs plus ou moins reconnaissables.
2. Les systèmes de représentation de connaissances ancrés dans les formalismes de représentation sémantique par graphes (ontologie) gagneraient à ne plus atomiser les entités représentées sur l'axe essentialiste de la permanence (domaine de l'être) mais à les réinsérer, comme le souligne François Rastier, *dans une perspective praxéologique*[7] *qui convienne aux textes et permette de les relier aux pratiques où ils sont produits et interprétés*[8] [RAS01]. C'est-à-dire quitter le domaine de l'être de l'univocité de l'intemporalité et de l'invariance pour celui du faire, de la multiplicité des points de vue, de la temporalité et de la variation[9] [RAS04].
3. L'objet de la modélisation est-il le document atomisé ou bien la collection de ses états actuels ou potentiels. De même, ne faudrait-il pas travailler sur les liens entre ces documents ou portions de document, comme autant de traces, au document de départ pour saisir en diachronie ses transformations
4. En ce qui concerne la modélisation des systèmes, il n'est pas suffisant de spécifier un format pivot des données. Nous avons proposé de fédérer ce lieu d'échange en les normalisant par des spécifications de protocoles au niveau des interfaces, à l'instar des réseaux de communication ouverts. Ces spécifications reprennent le principe OSI en s'organisant en couches de représentation des données allant de leurs instances physiques à leurs instances applicatives ou sémantiques. Les Infosphères, du point de vue des sciences pour l'ingénieur, représenteraient quant à elles, l'ensemble des situations possibles de communication pour un utilisateur. Chaque utilisateur professionnel du système interagit selon une intentionnalité, un

---

[7] Théorie de l'action par et dans le langage, étude des formes sémantiques sous l'angle de moments stabilisés de processus productifs et interprétatifs.

[8] [RAS01] Rastier, François. L'action et le sens pour une sémiotique des cultures Journal des anthropologues, n° 85-86, p. 183-219 (2001).

[9] [RSA04] Rastier, François. Ontologies, dans Revue d'intelligence artificielle, vol. 18, n° 1 : techniques informatiques et structuration de terminologies, p. 15-40 (2004).

point de vue actualisé sur son projet en cours (rôle), qui lui permet un certain nombre d'actions (décisions) choisies dans un ensemble possible et qui vont pouvoir s'exprimer par des requêtes. Une connaissance n'existe alors que sous la forme d'un champ qui sera instancié par un contexte (analogie avec la physique quantique). La modélisation doit alors s'attaquer à ce champ avant de catégoriser les actions possibles. Il faudrait alors catégoriser les points de vues professionnels possibles (rôles) avant de tenter de catégoriser les intentions et les actions qui peuvent en résulter.

5. Peut-on caractériser l'interlocuteur professionnel par les actions qu'il peut déclencher (ou les décisions qu'il peut prendre) et par les structures physiques et argumentatives des documents qu'il produit ou consomme ? Les rôles pouvant devenir des éléments pertinents de modélisation de posture professionnelle (nous entendons le terme de posture comme élément d'actualisation de rôles).

4. Propositions pour une approche par le biais de l'organisation

C'est dans le but de préparer l'approfondissement de ces pistes que nous nous proposons maintenant de définir l'intérêt d'une approche du document dans son agir organisationnel comme nous l'évoquions en introduction.

- Les organisations ont des traits communs essentiels, qu'elles partagent avec les langues.
  - ✓ Elles ont des structures variables et ne reflètent pas le même monde.
  - ✓ Elles convoquent la notion de cultures au pluriel, c'est-à-dire non comme une totalité (vision de sens commun) mais en fonction de points de vue, de visées sociales.

- Dans le cadre d'une organisation, les auteurs cessent d'être indéfinis pour devenir acteurs ou agents situés (cf. articles sur le brevet d'invention et sur le formulaire RMI : GUY04 et BLA04)[10]. Les interlocuteurs en son sein (zone identitaire)[11] mais aussi dans leur entour (zone proximale) se

[10] Dans Holzem et Labiche éd. (opus cité).

[11] Nous renvoyons le lecteur à l'excellent article de François Rastier écrit en 2001 (voir réf. ci-dessus) où l'auteur distingue les trois zones identitaire, proximale et distale (p. 191).
En ce qui concerne notre problématique nous pourrions dire que la zone identitaire exprime la personne ou le groupe de personnes (je, nous) qui pourrait être celui d'une organisation, située sur l'axe spatio-temporel de l'ici et du maintenant, munie d'une structure modale exprimant le certain. La zone proximale exprimant son entour sur le plan des personnes (tu, vous), autrement dit les interlocuteurs attendus de l'organisation dans le cadre de son travail, une temporalité exprimant également la proximité par rapport au maintenant

reconnaissent des obligations réciproques en fonction d'un lien social déjà existant et un désir de faire « acte » au sens juridique du terme dans la zone distale (zone de légitimation sanctionnée).

- L'organisation, par ses buts, objectifs, visées, impose une vision non immanente mais contingente de l'action de ses agents dans un temps et un espace donné : actions alors prédictibles et réitérables. Cette dimension représente l'aspect fonctionnel de l'organisation émergeant lors de l'activité autour du document, cette particularité devient alors modélisable et peut faire l'objet d'une expérimentation in *silico* destinée à reproduire l'observation d'origine.

- Elle appelle la notion de genre au sens baktinien qui considère un document, au fil de ces reprises successives dans un cadre spatio-temporel, comme une réponse à des énoncés antérieurs. Le document est alors le maillon d'un échange verbal préexistant et, à ce niveau, les pratiques culturelles et sociales d'un collectif de travail conditionnent largement la production comme l'interprétation de son contenu. Cette perspective offre une *valeur de régulation au document au-delà de la simple relation auteur/lecteur*, pour reprendre les propos de JM Salaün [SAL04] et elle peut ouvrir la voie d'une approche théorique du document *comme objet social, fondé sur un statut*, ce qui, toujours selon Salaün, instaurera un vrai dialogue entre sciences de l'ingénieur et sciences humaines et sociales [sic]. Dans la suite de notre action spécifique pluridisciplinaire « Document et Organisation » qui nous a conduit à étudier le document en contexte organisationnel, nous entrouvrons aujourd'hui cette porte.

- Le contexte organisationnel en tant que construit culturel fait partie de l'entour et non du monde physique (domaine de la phénoménologie). Cet entour, qui conditionne notre perception des données sensibles et donc notre expérience, est le lieu d'expériences partagées [SEA 83]. De ce point de vue, la modélisation et la formalisation des connaissances concernant le document (relié à une collection), le flux (relié à l'organisation) le producteur et l'usager (reliés à une communauté d'intérêt), intègre les notions d'intention et d'arrière plan comme ensemble des capacités

(naguère, bientôt). Elle rend compte des documents et de secteurs d'activité dont le travail est antérieur et justifie le présent du traitement par l'organisation ainsi que des documents et secteurs situés dans un futur proche, qui justifie de l'agir stratégique (sanction de l'efficacité d'une organisation). L'espace est défini (là) et la modalité exprime le probable. Quant à la zone distale, elle sert à exprimer les absents (il, on, là-bas, ailleurs) un passé ou un futur plus éloigné (non en contact direct avec l'organisation). Elle est la zone, nous dit Rastier, de la transcendance, des théories philosophiques, scientifiques, religieuses, des codes juridiques. Autrement dit de ce qui sert de fondement général à l'organisation sociale. C'est une zone où les transformations s'effectuent à l'échelle des décennies (voir beaucoup plus selon les domaines concernés).

mentales (schèmes, pratiques, compétences, routines, etc.) qui conditionnent notre expérience. Dans le cadre professionnel, cette expérience est, d'une part fondamentalement partagée (je perçois l'objet, le document, comme quelque chose que je sais devoir percevoir avec et comme d'autres) et elle est d'autre part conditionnée par des faits institutionnels (présupposition d'une norme dont ces faits sont une instance actualisée en contexte).

- L'approche par l'organisation invite donc à ne pas tant regarder le document que l'usage qui en est fait. Cette question en appelle trois autres selon nous.
  - ✓ Elle invite à considérer la valeur d'usage du document qui déterminera de son statut social mais également celui de l'organisation qui l'aura transformé. D'un point de vue qui va bien au-delà des échanges économiques au sein des sociétés humaines, le document acquiert par l'usage qui en est fait, une valeur marchande et un statut social. La notion de valeur d'usage va de pair avec la circulation du document (l'échange) quant à son statut social, il est à relier avec une typologie des lieux (collectif de travail) et des périodes (axe diachronique).
  - ✓ Elle invite également à cerner les pratiques de lecture et à constater qu'à une lecture scolastique dite intensive focalisée sur un corpus limité de textes (dont l'extrême est la relecture presque obsessionnelle de « classiques » à l'image des héros du roman de Bradbury[12]), le document dans son agir organisationnel invite à une lecture dite extensive à partir de textes fractionnés puis recomposés, à dessein, pour une durée plus ou moins brève. Il est de ce point de vue emblématique d'une lecture « moderne » qui s'inscrit dans une philologie numérique[13].
  - ✓ Dans la suite de l'article de François Rastier [opus cité] nous pourrions dire que l'organisation vise par ses actions (activité autour d'un but) à faire acte dans un domaine ciblé et s'exposer ainsi à des critères d'appréciation (positifs /négatifs) par les acteurs de ce domaine. Elle vise par là même la stabilisation d'une forme en fonction d'un but (propice à une modélisation des connaissances).

- La généralisation, dans les organisations, du document numérique pose le problème de sa modélisation et de celui de son usage. Concevoir le document comme une simple structure de communication, réceptacle d'informations à partager, autoriserait à le modéliser dans un format rigide, totalement prévisible et déterminé. Or, nous venons de voir que l'approche

[12] Bradbury, Ray. Farenheit 451. Paris : Denoël (1953).
[13] La philologie en tant que discipline qui établit et étudie les textes à tous leurs niveaux d'analyse, la philologie numérique traitant, quant à elle, des documents numériques, y compris des textes multimédia.

par l'organisation mettait en exergue le rôle prédominant de l'usage, lequel étant dès lors à considérer comme une véritable expérience au sens phénoménologique du terme. Comme toute expérience de constitution de sens, l'usage du document au sein de l'organisation peut se décliner en termes de :

- conditions de l'expérience ;
- objet visé par l'expérience pré-existant à celle-ci ;
- intentionnalité du sujet ;
- objet intentionnalisé (habitus) résultat de l'expérience.

L'expérience elle-même (l'usage) ne peut être rapportée à une simple interaction médiée par une interface, la plus sophistiquée soit-elle. La notion même d'interface devra à terme s'effacer devant celle beaucoup plus riche de processus partagé entre l'usager et le système de traitement du document numérique. Ce processus, porteur des aspects contraignants et structurants exigés de l'organisation (conditions de l'expérience) ne peut être piloté par un modèle rigide de nature déterminée, mais offrir les propriétés holistiques lui permettant de s'auto-modéliser au cours de l'expérience.

Cette problématique de la modélisation dynamique de l'expérience est en débat au sein de la communauté de la pensée complexe [LER04] et pourrait être en partie illustrée par les applications de traitement des documents numériques et de leur usage.

Les propositions de Husserl [HUS30] sur l'expérience du temps et la dynamique des attentes nous semblent particulièrement appropriés pour rendre opératoire ce processus émergeant et commun aux acteurs logiciels et humains.

L'expérience du temps permet d'identifier la donation originaire (à rapprocher de la zone identitaire citée au paragraphe 4), et correspondant au domaine des impressions ; la rétention correspondant à ce qui est tout juste passé, puis la protention, tournée vers l'avenir immédiat, et représentant en fait, des attentes qui seront, ne seront pas ou seront mal remplies par l'expérience en cours. Du résultat obtenu et de son ajustement aux contraintes précitées dépendra la suite du processus et son aboutissement.

Ce modèle dynamique, pour sa mise en œuvre, demande de s'écarter des voies classiques de l'intelligence artificielle basée sur le langage et les traitements, pour se rapprocher de la dialogie perception/action.

C'est à ces conditions que l'on peut espérer inverser le cours de cet appauvrissement relevé par Bachimont à propos du passage au numérique [BAC04].

5. Essai de schématisation

La figure suivante propose une représentation de la prise en compte du contexte organisationnel, venant se substituer, en la dépassant, à la notion d'usage à proprement parler.

Cette nouvelle dimension privilégie l'émergence d'un processus expérimental, partagé et interactionnel. Les interactions s'y situent à tous les niveaux où apparaît une relation entre les éléments structurels appartenant aux processus mis en communication : usager, système, expert, etc. Une organisation peut être vue comme une structure à laquelle s'ajoutent des fonctions.

Nous voyons sur les schémas s'étendre cette structure, en fonction des éléments impliqués, faisant apparaître les processus fonctionnels englobant les structures d'origine.

Lors des études terrains, l'analyse des processus émergeant, hors informatique, lors de l'usage attaché au document, a fait clairement apparaître l'existence de ces fonctions s'appuyant sur les relations entre les éléments en présence dans les structures concernées. Ce sont précisément ces fonctions que le document numérique, vu comme action, est à même d'inclure dans le processus interactionnel décrit ci-dessus.

Puisque l'usager se voit inclus dans le contexte organisationnel, il devient acteur impermanent mais impactant l'histoire du système.

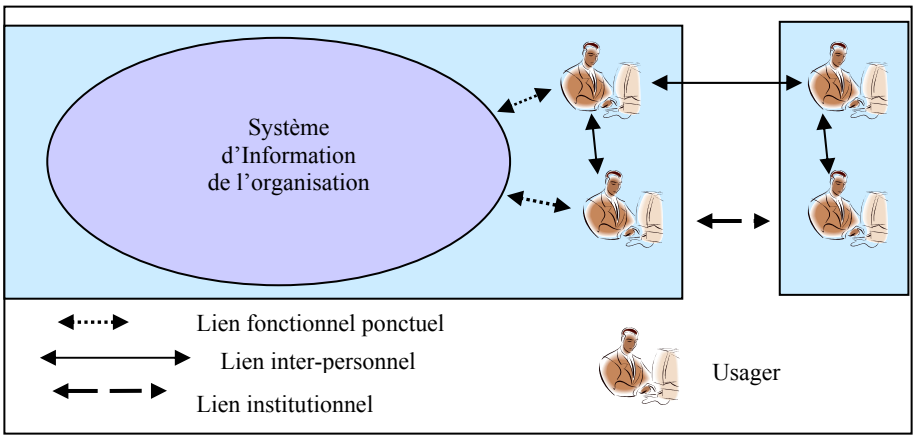


Figure 5 : Deux contextes organisationnels co-habitent

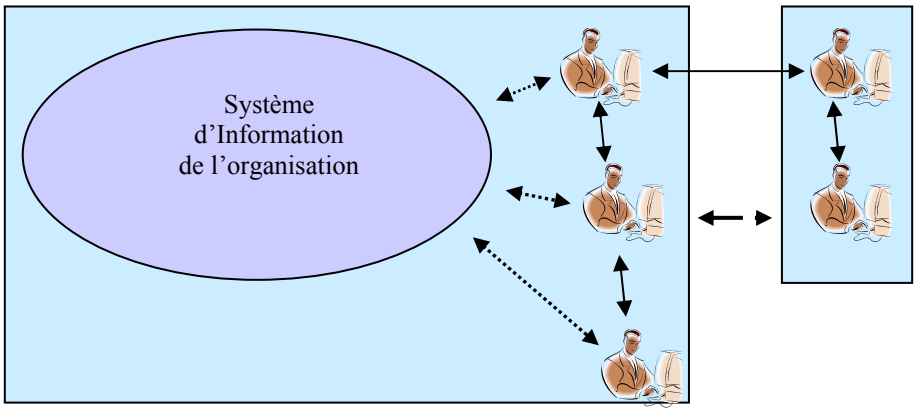


Figure 6 : L'usager, entrant en relation avec le système acteur situé, est englobé dans le processus partagé

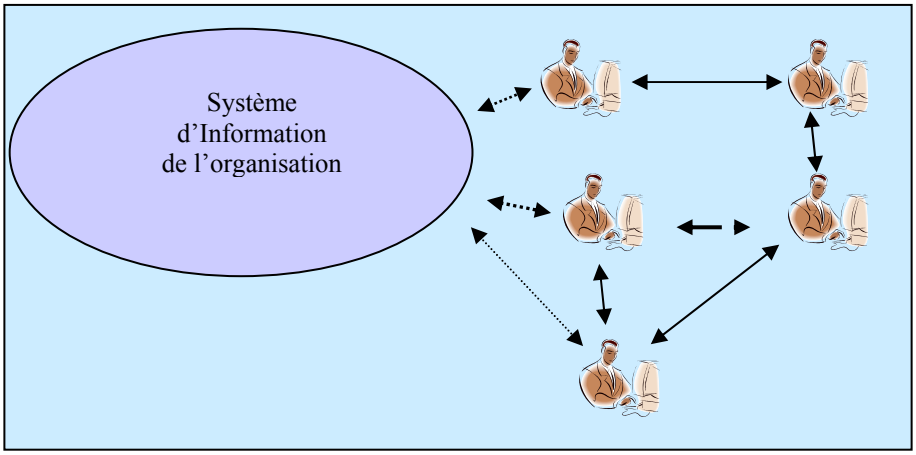


Figure 7 : Tout système externe devient système participant, dès que mis en relation avec le processus, parce que partageant les flux traités par le dit processus.

6. Conclusion : dans le contexte du document numérique

Le document saisi dans son agir organisationnel pourrait peut-être permettre de formaliser de nouveaux usages dans le contexte du document numérique. La perte matérielle occasionnée par le passage du document au document numérique, comme l'illustre McKitterick [KIT96][14] ne peut certes être comblée. Mais si, l'entité stabilisée dans une forme que nous appelons *document*, pouvait garder trace de la relation institutionnelle et sociale au sein de laquelle il s'est stabilisé, son interprétation, hors de l'espace et du temps qui ont présidé à sa production, serait facilitée. Bruno Bachimont et Stéphane Crozat ont poursuivit cette réflexion sur la perte occasionnée par son passage au numérique qui « *conduit à ruiner les conditions d'intégrité, d'identité et d'authenticité du document* » [BAC04]. Nous sommes alors tentés de poursuivre cette étude sur « *l'individualisation induite par le numérique* » qui a pour conséquence, toujours selon ces auteurs, « *d'empêcher l'appropriation en annulant l'objectivisation* ». Il nous semble en effet fort intéressant de relier l'objet document, son autonomie, sa cohésion, sa reconnaissance (notion à considérer avec celle de genre au sein d'une communauté de travail), aux questions de son appropriation et de la prise en compte du construit social qui la conditionne.

Or l'appropriation du contenu d'un document ne peut se réaliser sans une acculturation des connaissances qui offre les conditions d'accès à son contenu. La notion d'acculturation, que nous évoquions en préambule de notre article, est intéressante dans la mesure où elle révoque celle de *sens commun*, qui voudrait, toujours selon François Rastier, *que tout homme puisse comprendre les actions de tout autre, en inférant ses intentions* [RAS01]. Elle nous invite alors à considérer tout le construit social (institutions, modèles de conduites, éducation) variable selon les types d'organisation (*diversité anthropologique*) [DES02] dont se sont dotées les sociétés humaines et qui conditionnent l'interprétation de chacun de ses membres. Ce principe du *sens commun* qui est lié à celui de *charité* a des prétentions hégémoniques puisqu'il s'applique universellement selon Quine et Davidson [DEL01]. Il masque par là même son caractère ethnocentrique ancré dans une culture et une langue donnée (par laquelle s'exprime l'immense majorité des échanges numériques).

Notre approche du document par le biais de l'organisation n'est cependant, ni une solution miracle, ni une parade absolue contre les risques d'un consensus « naturel » autour du partage universel des connaissances. Mais elle invite à définir des zones d'échanges à partir d'acteurs situés dans des communautés de travail et dont les buts, règles de fonctionnement et points de vue peuvent être explicités.

---

[14] « *Pour le spécialiste de codicologie, de paléographie ou de bibliographie, une image digitalisée n'est pas suffisante. Pour les historiens de la lecture, l'image d'un libre ancien, digitalisée sur écran, n'est qu'une représentation partielle, voire trompeuse, et ainsi non historique* », Mc Kitterik, 1996 [McK96].

7. Bibliographie

[ADA 01] Adam, Jean-Michel. Types de textes ou genres de discours ? Comment classer les textes qui disent de et comment faire ?, *Langages*, n° 141, p. 10-27 (2001).

[BAC04] Bachimont, Bruno, Crozat, Stéphane. Instrumentation numérique des documents : pour une séparation fonds/forme, *Information-Interaction-Intelligence*, vol 4, n° 1, p. 95-103 (2004).

[BLA04] Blanc-Merigot, Martine, Delavigne, Valérie, Peyrard, Catherine, Peyrelong, Marie-France. Trois regards disciplinaires sur le RMI et les documents, dans, Holzem Maryvonne, Labiche Jacques éd. (2004)

[DEL01] Delpha, Isabelle. *Quine Dadidson : le principe de charité*, Paris, PUF (2001).

[DES02] Descombes, Vincent. L'idée d'un sens commun, *Philosophia Scientiae*, vol 6, Cahier 2, *L'usage anthropologique du principe de charité*, sous la dir. d'Isabelle Delpla, Editions Kimé, p. 147-161 (2002).

[FAY 02] Fayol, M. Les documents techniques : bilan et perspectives, *Psychologie Française*, n° 47-1, p. 9-18 (2002).

[GUY04] Guyot Brigitte, Normand Sylvie. Le document brevet, un passage entre plusieurs mondes, dans, Holzem Maryvonne, Labiche Jacques éd. (2004).

[HAB73] Habermas Jürgen. La technique et la science comme idéologie, Paris, Gallimard (1973).

[HUS30] Husserl, E. *Leçons pour une phénoménologie de la conscience intime du temps*, PUF, 1930 (édition de 1991).

[LER04] Lerbet-Sereni F. *et al. Expérience de la modélisation et modélisation de l'expérience*, L'Harmattan (2004).

[HAB87] Habermas, Jürgen. *Théorie de l'agir communicationnel*, 2 Vol., Paris, Fayard (1987).

[HEU 01] Heurley, Laurent. « Du langage à l'action : le fonctionnement des textes procéduraux », *Langages*, n° 141, p. 64-79 (2001).

[HOL04] Holzem, Maryvonne, Labiche, Jacques éd. *Document et Organisation : forum document et organisation, semaine du document numérique*, La Rochelle, Juin 2004, Paris, Europia, 101 p. (2004).

[KRI 69] Kristeva, Julia Séméiotikè. *Recherches pour une sémanalyse*, Paris, Le Seuil (1969).

[MCK96] Mc Kitterik. La bibliothèque comme interaction : la lecture et le langage de la bibliographie dans Baratin, Marc, éd., *Le pouvoir des bibliothèques : la mémoire des livres en occident*, Hachette, p. 107-121 (1996).

[RAS01] Rastier, François. L'action et le sens pour une sémiotique des cultures, *Journal des anthropologues*, n° 85-86, p. 183-219 (2001).

[RSA04] Rastier, François. Ontologies, dans *Revue d'intelligence artificielle*, vol 18, n° 1, Techniques informatiques et structuration de terminologies, p. 15-40 (2004).

[SAL04] Salaün, Jean-Michel. « Introduction : un dialogue pluridisciplinaire pour penser le document numérique », *Information-Interaction-Intelligence*, vol 4, n°1, p. 7-17 (2004).

[SEA83] Searle, John. *L'intentionnalité : essai philosophique sur les états mentaux*, éd. de Minuit, 340 p. (1983).

Session 4

Catégorisation et indexation

Une méthode indépendante des langues pour indexer les documents de l'Internet par extraction de termes de structure contrôlée

Jacques Vergne

GREYC – UMR 6072 CNRS – Université de Caen
14032 Caen Cedex - France

Jacques.Vergne@info.unicaen.fr
http://www.info.unicaen.fr/~jvergne

Résumé :

Nous présentons dans cet article une méthode d'indexation automatique de documents de l'Internet, fondée sur l'extraction de termes de structure contrôlée, et qui ne nécessite aucun traitement linguistique, ni stop-list, ni connaissance de la(les) langue(s) du document. Cette méthode s'appuie sur la récurrence de suites de mots, et sur le contrôle de la structure de ces suites. Ce contrôle de structure est basé sur un étiquetage du texte à indexer avec un jeu de deux étiquettes : mots informatifs ou non informatifs. Les mots informatifs sont définis comme étant plus longs et moins fréquents que leurs voisins. On exploite ainsi des propriétés très générales des langues, découvertes par Zipf et par Saussure.

Mots-clés : indexation automatique, termes de structure contrôlée, méthode d'indexation indépendante des langues.

Abstract:

In this paper, we present an automatic indexing method of web documents, based on structure controlled terms extraction, and which does not require any linguistic processing, neither stop-list, nor knowing the document language(s). This method relies on the words sequences recurrence, and on the structure control of these sequences. This structure control is based on tagging the text to index with a two label tagset: informative words or not. Informative words

are defined longer and less frequent than their neighbours. Very general linguistic properties, discovered by Zipf and by Saussure are thus exploited.

Keywords: automatic indexing, structure controlled terms, language independent indexing method.

1. Introduction

On peut caractériser les méthodes d'indexation automatique actuelles entre deux pôles : d'une part, l'indexation par des termes extraits du document, monolingue, qui nécessite des traitements et des ressources linguistiques (une indexation «en profondeur», issue des pratiques des bibliothécaires), d'autre part, l'indexation par tous les mots du document, dite « full text »utilisée par les moteurs de recherche sur l'Internet, consistant en des traitements superficiels de masses textuelles énormes, considérées comme des chaînes de caractères (une indexation « en largeur »), et non pas comme du matériau linguistique.

Ces deux pôles illustrent la divergence croissante des deux problématiques, qui ont paru un temps être analogues : extraction terminologique en profondeur en espace clos et figé, et indexation « full text » en vaste espace ouvert et évolutif. Certains concepts issus de la recherche d'information dans des bases de données textuelles (donc des petits espaces clos, souvent monothématiques et monolingues) ne semblent plus être adéquats à de vastes espaces ouverts. Le silence, par exemple, n'est pas mesurable sur l'Internet, car on ne peut pas compter le nombre total de documents pertinents non récupérés. De manière analogue, l'idf (Inverse Document Frequency) garde-t-il son intérêt sur l'Internet ?

Nous allons caractériser ces deux pôles, pour introduire ensuite notre méthode.

1.1 L'indexation « full text »

L'indexation « full text » a les caractéristiques suivantes (cf. "text indexing" dans [Salton 83], et dans [Brin & Page 98], l'article fondamental sur Google par leurs concepteurs, écrit avant la création de leur société) :

- Tout mot du document est terme et l'indexera, quel qu'il soit (mot plein ou mot vide), quel que soit son effectif (y compris les hapax);
- Il n'y a pas de contrôle de structure du terme, car 1 terme = 1 mot ;
- Le grain traité est le mot (ou le caractère pour les langues non alphabétiques) dans le document (le mot sera une clé d'accès au document) ;
- Les cadres d'utilisation sont les systèmes d'indexation des moteurs de recherche généralistes sur l'Internet ;

- Le corpus à indexer est constitué d'un très grand nombre de petits documents ;
- Le traitement est indépendant des langues (aucun traitement linguistique, seulement reconnaissance de l'écriture alphabétique ou non) ;
- Les recherches par terme (expression entre guillemets) se font grâce à l'offset des mots dans les documents (cf. [Brin & Page 98]);
- Avantages : aucun traitement linguistique, traitement superficiel, indépendant des langues, permettant d'indexer d'énormes masses textuelles ;
- Inconvénients : traitement trop superficiel ? Trop grand nombre de documents récupérés, dont seulement les premiers classés sont visibles par l'utilisateur; la taille des index devient prohibitive, et nécessitera bientôt une évolution de stratégie d'indexation : pour chaque mot de chaque document indexé (8.109 selon Google[1]), sa graphie est présente une fois dans le dictionnaire de l'index, mais les informations propres à l'occurrence (offset, taille relative de fonte, relation à la graphie, relation à l'identifiant du document) sont présentes autant de fois que d'occurrences.

1.2 L'extraction terminologique de corpus clos

L'extraction terminologique de corpus clos se situe à l'opposé (cf. [Bourigault 02]) :

- Un terme est constitué de certains groupes de mots contigus (1 ou plus) du document ;
- La structure syntaxique du terme est contrôlée (ce sont surtout des syntagmes nominaux) ; ce contrôle de structure syntaxique utilise une analyse morphosyntaxique, un dictionnaire, une grammaire monolingues ;
- Le cadre d'utilisation est l'extraction de termes structurés (les différentes expansions d'une même tête) d'un corpus clos, monolingue (langue unique identifiée), souvent monothématique ;
- Avantage : contrôle très fin de la structure des termes ;
- Inconvénient : traitement linguistique lourd, monolingue, nécessitant de reconnaître la langue.

Dans certains systèmes, on exploite uniquement la récurrence en excluant les mots vides à l'aide de stop-lists (cf. [Salem 87], [Salton 93], [Ahonen 99]) ; si l'on a dans un document : *président de la république* et *président de la société X*, le segment répété est : *président de la* qui n'est pas un terme correct ; ce terme est alors corrigé en en supprimant les mots de la stop-list situés en début ou fin du segment.

[1] Mais voir "Web : Le mystère des pages manquantes de Google résolu ?" sur le blog de Jean Véronis : http://aixtal.blogspot.com/2005/02/web-le-mystre-des-pages-manquantes-de.html

1.3 Notre proposition : allier des avantages des deux stratégies

Nous proposons de calculer des termes de structure contrôlée pour indexer des documents (ou des sites) en très grand nombre, de langue(s) inconnue(s), c'est-à-dire d'importer dans l'indexation sur l'Internet la finesse du contrôle de structure des termes de l'extraction terminologique de corpus clos.

Le traitement doit être indépendant des langues des documents et donc ne pas utiliser de stop-list.

Définissons les termes comme étant certains groupes de mots contigus, non hapax, et de structure contrôlée : les termes sont centrés sur des mots « informatifs ».

2. Principes et algorithmes

On suppose que l'on est dans le cadre de l'indexation d'un document du web (ou d'un site web). Le texte à indexer est extrait de la source html du document (ou de certains documents du site, par exemple jusqu'à une certaine profondeur à partir d'un point d'entrée). Pour localiser la partie informative du document (c'est la partie à indexer), nous utilisons la propriété suivante : la partie informative du document est présente une seul fois sur le site[2]. Le texte est découpé en paragraphes (unités physiques marquées dans le balisage), qui constituent l'unité traitée. Dans un document html, les balises de fin de paragraphe sont considérées comme une macro-ponctuation : `title`, `div`, `br`, `p`, `td`.

L'algorithme comporte ensuite 4 étapes :

1. Étiquetage du document avec un jeu de deux étiquettes : mots informatifs ou non informatifs (en première approche, les mots non informatifs sont les mots vides) ;
2. Génération de candidats termes de structure contrôlée ; la structure d'un candidat terme est contrôlée par des motifs fondés sur l'étiquetage mots **I**nformatifs (**`I`**) ou **n**on (**`n`**) : un candidat terme est présent au moins 2 fois dans le grain à indexer (document ou site), il commence et se termine par un mot informatif, et ne contient pas de ponctuation ; les termes hapax sont supprimés au fur et à mesure du déroulement de l'algorithme ;
3. Élagage de l'ensemble des termes : les termes inclus dans des termes de même effectif sont supprimés ;

[2] Les algorithmes de localisation de la partie informative d'un document html seront publiés dans un prochain article : relevé de la mise en forme matérielle (MFM), segmentation en paragraphes (unités physiques marquées dans le balisage), calcul des classes d'équivalence de MFM, et calcul de la structure du document autour du corps de texte, tous ces calculs étant conçus pour fonctionner sur de l'html mal formé.

4. Pondération des termes dans le document, par une estimation de la place occupée dans le rendu du document, soit par exemple pour un terme : effectif * longueur.

2.1 Étiquetage du document avec un jeu de deux étiquettes

Comment différencier les mots informatifs des mots non informatifs ?

Deux définitions sont possibles :

- La définition classique : les mots informatifs sont les mots pleins ou lexicaux (content words), et les mots non informatifs sont les mots vides, ou grammaticaux (function words) ; cette distinction a été introduite par [Tesnière 59], page 53 ; cela conduit alors à l'utilisation de stop-lists, qui ont l'avantage d'être facilement disponibles, mais qui ont deux désavantages : la(les) langue(s) du document doivent être reconnues pour choisir la(les) bonne(s) stop-list(s), et les mots de la stop-list ne sont jamais indexés, par définition, ce qui entraîne qu'un document sur *la vente de stocks d'or* sera inaccessible par la requête « *or* » ;
- Nous proposons la nouvelle définition suivante : un mot informatif est plus long et moins fréquent que ses voisins.

Cette définition allie les propriétés linguistiques très générales mise en évidence par Zipf et par Saussure :

- **Zipf** : ce qui est d'usage fréquent est court : c'est la loi de l'économie d'effort dans l'usage d'un code, caractérisée par Zipf [Zipf 49], et observable aussi dans les langages de programmation (remarquons que la « loi de Zipf », toujours très présente dans la littérature, est une loi sur les effectifs des mots, et que les propriétés statistiques des longueurs des mots sont plus rarement invoquées) ;
- **Saussure** : *dans la langue, il n'y a que des différences* (cf. [Saussure 22], éd. 1974, p.166) ; cette importante propriété du matériau linguistique nous conduit à fonder nos calculs sur des différences entre unités, plutôt que sur des valeurs absolues des unités, comme celles que l'on stocke dans les dictionnaires.

Nous proposons d'allier ces deux propriétés dans un calcul complètement local, fondé sur les différences de longueur et d'effectif de deux mots contigus. Dans des travaux antérieurs (cf. [Vergne 03] et [Vergne 04]), à partir des mêmes propriétés, nous avions proposé un algorithme différent, fondé sur le pavage déterministe du segment, avec des motifs de 5 à 2 mots, testés successivement dans un ordre heuristique imposé (`InnI`, `InI`, par exemple). Le nouvel algorithme présenté ici est plus épuré, car il ne fait aucune hypothèse de motif avant l'étiquetage, mais il calcule l'étiquetage uniquement à partir des différences d'effectif et de longueur entre 2 mots contigus.

Dans un paragraphe, soit deux mots contigus (numérotés 1 et 2) d'**e**ffectifs f1 et f2 (dans le document ou dans le site, selon que l'on veut indexer chaque document ou le site entier) et de **l**ongueurs l1 et l2.

La différence entre deux mots contigus est définie comme un objet à 2 attributs :

- ✓ Le type de la différence (la différence est orientée) :

  si f1>f2 & l1<l2, alors type_diff = nI (**n**on informatif - **I**nformatif)
  sinon si f1<f2 & l1>l2, alors type_diff = In (**I**nform. - **n**on informatif)
  sinon type_diff = contradictoire (contradiction entre les deux critères)

- ✓ La mesure de la différence :

  mesure = (max(f1, f2)/min(f1, f2)) * (max(l1, l2)/min(l1, l2))

  Cette mesure permet d'évaluer si la différence est suffisante pour provoquer un changement d'étiquette. Elle est égale au produit des rapports des longueurs et des effectifs. Elle est donc indépendante de la taille du document.

Règle d'affectation d'une étiquette à un mot à partir des différences avec son voisin précédent : si la différence entre les mots 1 et 2 est de type non contradictoire et de mesure suffisante (supérieure à un seuil calculé), alors le mot 2 prend l'étiquette 2 du type de la différence, sinon le mot 2 prend l'étiquette du mot 1, ce qui donne :

```
si type_diff ≠ contradictoire & mesure > seuil,
alors étiq_mot2 = étiquette 2 du type_diff
        si étiq_mot1 indéterminée,
              alors étiq_mot1 = étiquette 1 du type_diff
sinon étiq_mot2 = étiq_mot1
```

Par exemple, voici le déroulement de l'étiquetage de quelques mots :

	f	l	(f = effectif, l = longueur)
1	649	3	*une*
2	32	8	*nouvelle*
3	1	10	*résolution*
4	3673	2	*de*
5	1500	2	*l'*
6	9	3	*ONU*

Pour expliquer l'algorithme, prenons arbitrairement un seuil de 4 (soit un rapport de 2 sur chaque critère) :

- Différence entre les mots 1 et 2 (*une nouvelle*) :

  f1>f2 & l1<l2 => type_diff = **nI**
  mesure_diff = (649/32) * (8/3) = 54,1
  54,1 > 4 => étiq_mot1 = étiq1_type_diff = **n**
  étiq_mot2 = étiq2_type_diff = *I*

- Différence entre les mots 2 et 3 (*nouvelle résolution*) :

  f2>f3 & l2<l3 => type_diff = n**I**
  mesure_diff = (32/1) * (10/8) = 40
  40 > 4 => étiq_mot3 = étiq2_type_diff = **I**

- Différence entre les mots 3 et 4 (*résolution de*) :

  f3<f4 & l3>l4 => type_diff = I**n**
  mesure_diff = (3673/1) * (10/2) = 18365
  18365 > 4 => étiq_mot4 = étiq2_type_diff = **n**

- Différence entre les mots 4 et 5 (*de l'*) :

  f4>f5 & l4≤l5 => type_diff = nI (égalité possible sur un des 2 critères)
  mesure_diff = (3673/1500) * (2/2) = 2,45
  2,45 < 4 => étiq_mot5 = étiq_mot4 = **n** (car différence insuffisante)

- Différence entre les mots 5 et 6 (*l'ONU*) :

  f5>f6 & l5<l6 => type_diff = n**I**
  mesure_diff = (1500/9) * (3/2) = 250
  250 > 4 => étiq_mot5 = étiq2_type_diff = **I**

- D'où l'étiquetage :

  <n>*une* <I>***nouvelle*** <I>***résolution*** <n>de <n>l'<I>***ONU***

Comment choisir la valeur du seuil de mesure de différence ? Nous proposons que ce seuil soit calculé pour chaque paragraphe.

Dans l'exemple précédent, l'étiquette d'un mot a été affectée en fonction de sa différence avec le mot précédent. On peut aussi calculer son étiquette en fonction de sa différence avec le mot suivant. D'où une mesure de la qualité d'un étiquetage en fonction d'une valeur de seuil : le nombre de désaccords d'étiquetage entre les étiquetages selon la différence avec le mot précédent ou bien avec le mot suivant.

On observe que la fonction : nombre de désaccords d'étiquetage = f(seuil) a un minimum, qui permet de définir la meilleure valeur du seuil.

D'où l'algorithme de calcul du seuil qui consiste à rechercher un minimum approché de cette fonction entre 2 bornes, de valeurs initiales :

borne basse = 1 (égalité sur les 2 critères pour les 2 mots)
borne haute = racine carrée (f_max * l_max) / 2

(les 2 mots les plus différents seraient un hapax de longueur l_max et un mot de longueur 1 et d'effectif f_max).

Pour approcher un minimum de la fonction à partir de faibles valeurs du seuil, le seuil initial est choisi légèrement supérieur à la borne inférieure (1). L'algorithme est classique, et procède par dichotomie, en rapprochant les bornes haute et basse tant que leur différence est supérieure à 1 et que la valeur de la fonction est supérieure à 0, avec une limite de 5 cycles.

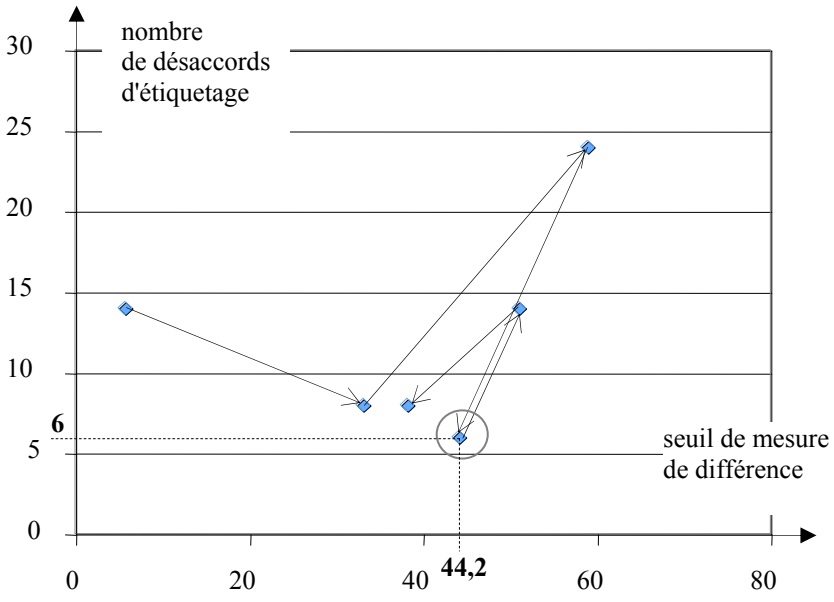
Voici un exemple d'exécution sur un segment extrait du Monde du 15 avril 2004 :

M. Bush évoque une nouvelle résolution de l'ONU qui permettrait à d'autres pays de se joindre à ceux qui participent à une coalition aujourd'hui de plus en plus déstabilisée.

La borne haute est : racine carrée (3673 * 12) / 2 = 104,9

(3673 = effectif de "*de*", 12 = longueur de "*déstabilisée*")

Les flèches indiquent la chronologie de la recherche du minimum de la fonction :



Dans ce cas, un nombre minimal de désaccords d'étiquetage, égal à 6, a été obtenu pour un seuil de mesure de différence égal à 44,2.

Voici le résultat de l'étiquetage (les mots informatifs sont en gras) :

M. Bush évoque *une* ***nouvelle résolution*** *de l'****ONU*** *qui* ***permettrait*** *à d'****autres pays*** *de se* ***joindre*** *à ceux qui* ***participent*** *à une* ***coalition aujourd'hui*** *de plus en plus* ***déstabilisée***.

La complexité pratique de cet algorithme d'étiquetage est linéaire en temps selon le nombre de mots du document, car le nombre de cycles de la répétition « tant que » de la recherche de minimum de fonction est borné (on ne recherche pas la meilleure solution en un nombre inconnu de cycles, mais une bonne solution en un nombre de cycles donné).

2.2 Génération de candidats termes de structure contrôlée

La structure d'un candidat terme est contrôlée par des motifs fondés sur l'étiquetage mots **I**nformatifs (**I**) ou **n**on (**n**).

En utilisant la syntaxe des expressions régulières, on génère d'abord les candidats simples, de motif `I+` , en ne prenant pas les candidats simples hapax (ce sont des mots, dont on connaît l'effectif), pour ne garder que les candidats simples répétés. Par exemple, à partir de ***nouvelle résolution***, de motif `I+`, on génère la combinatoire des termes possibles : ***nouvelle***, ***résolution***, ***nouvelle résolution***.

On génère ensuite les candidats doubles, de motif `I+n+I+`, à partir des candidats simples répétés, car les candidats doubles répétés sont composés de candidats simples répétés. Les candidats doubles hapax sont alors supprimés. Par exemple, à partir de ***nouvelle résolution de l'ONU***, de motif `I+n+I+`, on génère la combinatoire des termes possibles : ***résolution de l'ONU***, ***nouvelle résolution*** de l'***ONU***.

On génère ensuite les candidats triples, de motif `I+n+I+n+I+`, à partir des candidats doubles répétés. Les candidats triples hapax sont alors supprimés. Par exemple, à partir de ***or*** de la ***Banque*** de ***France***, de motif `I+n+I+n+I+`, on génère la combinatoire des termes possibles : ***or*** de la ***Banque***, ***Banque*** de ***France***, ***or*** de la ***Banque*** de ***France***.

On continue ainsi tant qu'on trouve des candidats non hapax, comme dans les algorithmes gloutons.

Cet algorithme étant combinatoire, sa complexité pratique est polynomiale de degré supérieur à 1, mais cette combinatoire reste faible, car les suites de mots informatifs contigus font 1 à 3 mots, et les suites longues et répétées sont rares.

2.3 Élagage de l'ensemble des termes

L'élagage a été déjà réalisé pour les candidats hapax au fur et à mesure du déroulement de l'algorithme de génération des candidats.

Il reste à supprimer les termes inclus dans des termes de même effectif, car ils sont moins informatifs. Par exemple, si ***résolution*** et ***résolution*** de ***l'ONU*** ont chacun 3 occurrences, cela implique que ***résolution*** ne se trouve que dans ***résolution*** de ***l'ONU*** et est moins informatif que le terme double qui le contient.

2.4 Pondération des termes

Une indexation doit s'accompagner d'une pondération, qui servira dans le calcul du « ranking » des réponses à une requête. Nous proposons que cette pondération soit fondée sur une estimation de la place occupée dans le rendu du document, soit pour chaque terme : effectif * longueur.

Dans un document html dont on a calculé la structure (voir note 2), nous faisons aussi intervenir la position du paragraphe par rapport au corps de texte, avec un coefficient défini comme étant égal à 1 pour le corps de texte, à des nombres supérieurs à 1 pour les paragraphes singletons dans leur classe de MFM, et situés avant le corps. Dans ce cas, le poids d'un terme devient : ∑ effectif * longueur * coefficient, en sommant sur chaque occurrence du terme.

3. Résultats

Dans son édition du 15 avril 2004, les 94 documents de profondeur 0 et 1 du site du journal Le Monde (on est ici dans le cadre d'une indexation de site) contenaient 26 occurrences de la graphie « or » (en minuscules), toutes référant au métal jaune, ce qui nous permet de montrer la robustesse de l'algorithme sur les mots informatifs courts et sur l'intérêt d'éviter une stop-list. Par exemple :

	f	l		(f = effectif, l = longueur)
0	8	5	I	**Bercy**
1	8	7	I	**cherche**
2	1353	1	n	à
3	4	8	I	**utiliser**
4	1500	2	n	l'
5	26	2	I	**or**
6	3673	2	n	de
7	2000	2	n	la
8	19	6	I	**Banque**
9	3673	2	n	de
10	120	6	I	**France**

Sur 26 occurrences de « or », 21 ont été étiquetées informatives, 3 non informatives, 2 sont restées indéterminées (pour le seuil du minimum de la fonction, les étiquettes selon les différences avec les mots précédent et suivant étaient divergentes).

Les termes obtenus à partir de la graphie « or » sont :

f*l	f	l	(f = effectif, l = longueur)	motif
275	11	25	**or** de la **Banque** de **France**	InnInI
68	4	17	**rentabiliser** l'**or**	InI
44	4	11	**tonnes** d'**or**	InI
42	21	2	**or**	I
30	2	15	**500 tonnes d'or**	IInI
26	2	13	**utiliser l'or**	InI
16	2	8	**or c'est**	III
8	2	4	**L'or**	II

Voici un exemple tiré de l'International Herald Tribune du 15 octobre 2004 (102 documents de profondeur 0 et 1) sur la graphie « war », mot informatif court en anglais :

0	1	8	I	**SARAJEVO**
1	97	2	n	In
2	1875	3	n	the
3	48	5	I	**years**
4	26	6	I	**before**
5	1875	3	n	the
6	5	7	I	**Bosnian**
7	33	3	I	**war**
8	1013	2	n	of
9	1875	3	n	the
10	7	5	I	**early**
11	7	5	I	**1990s**

Sur 33 occurrences de « war », 21 ont été étiquetées informatives, 10 non informatives, 2 indéterminées.

Voici les termes obtenus à partir de la graphie « war » :

f*l	f	l		motif
63	21	3	**war**	I
46	2	23	**effective war** on **terror**	IInI
22	2	11	**war** in **Iraq**	InI
18	2	9	civil **war**	II

En allemand, la graphie « war » est une forme de l'auxiliaire « sein » (être), et devait être étiquetée non informative. Voici un exemple tiré du Spiegel du 15 avril 2004 (130 documents de profondeur 0 et 1) :

0	144	3	n	Der
1	2	13	I	**Softwareriese**
2	83	3	n	war
3	4	7	I	**zuletzt**
4	12	5	I	**wegen**
5	1213	3	n	der
6	50	5	n	immer
7	1	8	I	**häufiger**
8	1	12	I	**auftretenden**
9	1	17	I	**Sicherheitsmängel**
10	603	2	n	n
11	1055	3	n	die
12	12	6	I	**Kritik**
13	5	7	I	**geraten**

Sur 83 occurrences de « war », 8 ont été étiquetées informatives, 72 non informatives, 3 indéterminées.

Voici un exemple en italien, tiré de La Stampa du 15 avril 2004, sur la graphie « Iraq » (107 documents de profondeur 0 et 1) :

0	4	2	n	Ci
1	1	6	I	**stiamo**
2	1	8	I	**muovendo**
3	140	3	n	con
4	9	6	I	**quelle**
5	5	8	I	**autorità**
6	302	3	n	che
7	33	3	-	all
8	3	7	I	**interno**
9	54	4	n	dell
10	11	4	I	**Iraq**
11	525	1	n	e
12	33	3	-	all
13	3	7	I	**interno**
14	70	5	n	delle
15	8	10	I	**principali**
16	3	8	I	**comunità**
17	1	9	I	**religiose**

18	54	4	n	dell
19	11	4	I	**Iraq**
20	1	9	I	**riteniamo**
21	1	7	I	**abbiano**
22	1	13	I	**autorevolezza**
23	525	1	n	e
24	2	8	I	**capacità**
25	852	2	n	di
26	1	7	I	**indurre**

Sur 11 occurrences de « Iraq », 10 ont été étiquetées informatives, aucune non informative, 1 indéterminée.

4. Conclusion et perspectives

Nous avons proposé une méthode d'indexation de documents de internent, fondée sur l'extraction de termes répétés et de structure contrôlée. Le contrôle de structure se base sur un étiquetage des mots à 2 étiquettes : mot informatif (pouvant être un index) et mot non informatif (ne pouvant pas être index seul). Cet étiquetage n'utilise que des propriétés de longueur et d'effectif des mots, et est donc possible sans connaître la langue du texte. Nous pensons qu'une telle méthode est applicable dans le cadre de l'indexation des moteurs de recherche sur internent, par sa légèreté calculatoire, son indépendance des langues, par l'obtention d'index moins volumineux qu'en indexation full-text, et par l'amélioration de la précision des réponses, mais ce dernier point reste à valider.

La question de l'évaluation d'une méthode d'indexation reste ouverte : peut-on évaluer seul un maillon de la chaîne des traitements d'un moteur de recherche, ou faut-il plutôt évaluer la chaîne entière : crawling + indexation + traitement des requêtes + ranking des réponses ? Cette évaluation devra se faire en collaboration avec des utilisateurs.

D'autre part, nous explorons actuellement la voie d'étiqueter non pas les mots, mais les différences entre mots, et d'en déduire un étiquetage des mots.

Au sujet des langues alphabétiques plus ou moins agglutinantes (par exemple : finnois, mais aussi allemand, espagnol dans une certaine mesure), des mots ne sont plus délimités par des espaces; mais la méthode est transposable en opérant un découpage en morphèmes avant étiquetage (travaux en cours). Au sujet des langues non alphabétiques (par exemple : chinois, japonais, coréen), la méthode est aussi transposable (travaux en cours), le grain atome étant non plus le mot, mais le groupe de caractères répété (cf. [He *et al.* 02]). Dans l'optique d'indexer toutes les écritures existantes, il nous faut ainsi abandonner le concept de « mot », trop centré sur les écritures des langues européennes, au profit d'une « molécule physique de texte »

définie pour toutes les écritures, molécule à segmenter pour obtenir des «atomes», atomes dont les séquences répétées et de structure contrôlée sont à indexer.

Enfin, il faut mettre au point un calcul des termes concernés par une requête donnée, par exemple, les termes incluant tout ou partie de la requête, l'adéquation requête - termes devant participer au ranking des réponses.

5. Références bibliographiques

[Ahonen 99] Ahonen-Myka Helena. Discovery of frequent word sequences in text. *The ESF Exploratory Workshop on Pattern Detection and Discovery in Data Mining*, Imperial College, London, 2002. www.cs.helsinki.fi/u/hahonen/ ahonenmyka_ patws 02.ps

[Bourigault 02] Bourigault Didier. Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002)*, Nancy, 2002, p. 75-84. www.univ-tlse2.fr/erss/textes/pagespersos/ bourigault/TALN02-Bourigault.doc

[Brin & Page 98] Brin Sergey & Page Lawrence, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Computer Networks and ISDN Systems, vol. 30, n° 1-7, p. 107-117, 1998. citeseer.ist.psu.edu/brin98anatomy.html.

[He *et al.* 02] Hongzhao He, Jianfeng Gao, Pilian He, and Changning Huang, Finding the Better Indexing Units for Chinese Information Retrieval. In: *SIGHAN 2002*. Taipei, Taiwan. acl.ldc.upenn.edu/W/W02/W02-1804.pdf.

[Salem 87] Salem André. *Pratique des segments répétés*. Publications de l'INaLF, collection "St Cloud", Klincksieck, Paris, 1987.

Salton 83] Salton, G. and McGill, M.J. *Introduction to modern information retrieval*. New York: McGraw Hill, 1983.

[Salton 93] Salton Gerard and Allan James. Selective Text Utilization and Text Traversal. In *UK Conference on Hypertext*, 1993, p. 131-144.

[Saussure 22] Saussure F. de. *Cours de Linguistique Générale*. Payot, Paris, (éd. 1974), 1922.

[Tesnière 59] Tesnière Lucien. *Éléments de syntaxe structurale*. Klincksieck (Paris), 1959.

[Vergne 03] Vergne Jacques. Un outil d'extraction terminologique endogène et multilingue. *Actes de TALN 2003*, tome 2, 2003, p. 139-148. www.info.unicaen.fr/~jvergne/ TALN2003/ JVergne-TAL2003multV23.pdf.

[Vergne 04] Vergne Jacques. Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource. *Actes des JADT 2004*, vol. 2, 2004, p. 1158-1164.

[Zipf 35] Zipf George Kingsley. *The Psychobiology of Language, an Introduction to Dynamic Philology*. Houghton Mifflin, Boston, 1935.

[Zipf 49] Zipf George Kingsley. *Human Behavior and the Principle of Least Effort*. Harper, New York, 1949.

Application de plusieurs stratégies pour trouver des réponses en anglais à des questions posées en français

Brigitte Grau[1], Gabriel Illouz[1], Laura Monceaux[2], Isabelle Robba[1], Anne Vilnat[1], Olivier Ferret[3], Faiza El Kateb[1]

[1] *LIMSI, BP 133, 91403 Orsay Cedex - France*
Prénom.Nom@limsi.fr

[2] *LINA, Université de Nantes - France*
Laura.Monceaux@univ-nantes.fr

[1] *LIUM-LIST, CEA - France*
Olivier.Ferret@cea.fr

Résumé :

Notre système de question-réponse MUSCLEF, qui a participé à l'évaluation CLEF en 2004, a été conçu pour fournir des réponses en anglais à des questions posées en français. Il est fondé sur notre système pour l'anglais, QALC, qui a participé à TREC, et y a obtenu de bons résultats quand nous avons combiné plusieurs stratégies. QALC recherchait des réponses dans la collection donnée et sur le WEB. Nous avons gardé ces deux stratégies pour CLEF, à partir des questions traduites. Nous avons aussi géré le multilinguisme en traduisant les termes significatifs tirés des questions et en adaptant QALC pour construire le système MUSQAT. Nous avons combiné les résultats de ces trois recherches pour produire le résultat final et nous montrons l'apport de cette combinaison par rapport aux résultats de chacune des stratégies seules.

Mots-clés : système de question-réponse, bilinguisme, termes de recherche.

1. Introduction

La recherche de réponses précises à des questions de types factuels (c'est-à-dire des questions amenant une réponse exprimable par une formulation courte) est un champ de recherche attirant l'intérêt d'un nombre croissant de chercheurs. Les spécifications varient d'un système à l'autre, ou d'une campagne d'évaluation à l'autre. Néanmoins, le but est toujours de fournir un court extrait du ou des documents contenant la réponse, allant de la réponse exacte uniquement à un ou plusieurs passages. Un défi supplémentaire, présent dans la campagne d'évaluation CLEF[1], consiste à pouvoir passer d'une langue à l'autre. L'apport de cette fonctionnalité réside dans l'augmentation de l'espace de recherche, tout en permettant de garder sa langue maternelle pour l'interrogation. Cette possibilité est d'autant plus intéressante si on travaille sur le Web et que l'on recherche des réponses en anglais, comme le montre la figure 1.

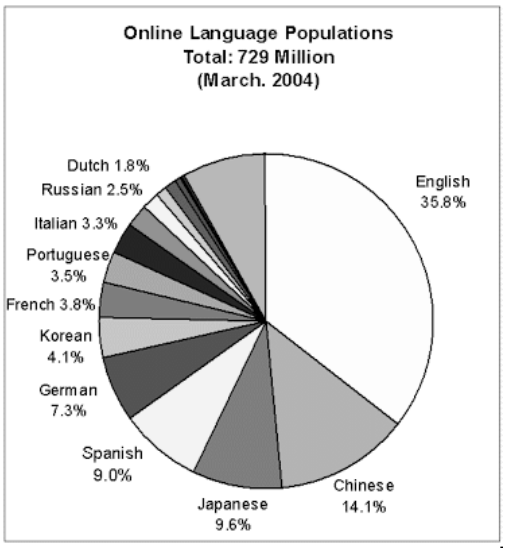


Figure 1 : Répartition des langues sur le Web[2]

Aussi, nous nous sommes intéressés à la recherche de réponses en anglais à partir de questions posées en français. De plus, notre but est de fournir une et une seule réponse à chaque question, sans que des informations non pertinentes soient présentées à l'utilisateur.

---

[1] Cross Language Evaluation Forum.

[2] Cette figure provient du « Centre for Public Policy of the University of Melbourne » : http://www.public-policy.unimelb.edu.au/egovernance/papers/33_Skidmore.pdf.

Outre le traitement du multilinguisme, le problème consiste à estimer la confiance que le système porte à ses propositions. Lors de TREC11[3], nous avions traité ce problème en recherchant la réponse dans la collection d'une part et sur le Web d'autre part, tout en gardant la même stratégie de résolution, et en combinant les propositions obtenues par le système. Lorsque la même réponse figurait dans les 5 premières propositions provenant de chacune des recherches, cette réponse était fortement privilégiée, considérant que ramener la même réponse de deux sources de connaissances différentes lui confère un fort degré de pertinence, venant supplanter le poids que le système attribue en fonction des processus appliqués. Cette évaluation de la pertinence d'une réponse s'est montrée très efficace, et le système QALC a ainsi réorganisé ses propositions de telle manière que la bonne réponse est remontée en première position dans 21 % des cas par rapport aux résultats obtenus à partir d'une seule source de connaissances. QALC était ainsi le système qui ordonnait le mieux ses bonnes réponses [CHA03]. D'autres systèmes ont également montré l'intérêt de combiner de multiples stratégies de résolution [JIJ04], [CHU02], ou de recherches combinées dans différentes sources [BRI01], [CLA01], [HER02] , [MAG02a], [MAG02b].

Le traitement du multilinguisme dans les systèmes de question-réponse relève actuellement de deux approches : la traduction des questions par un traducteur [PER04], [JIJ04], [AHN04], ou la traduction de termes sélectionnés [NEG03], [TAN04], [SUT04]. Comme chacune de ces approches conduit à un nombre de réponses plus restreint que celui obtenu par les systèmes monolingues, nous avons choisi de continuer à combiner les résultats issus de l'application de stratégies différentes, dans la mesure où une même réponse apparaissant dans plusieurs listes a plus de chance d'être correcte, quelle que soit la méthode qui l'a produite. Aussi, notre système MUSCLEF (Multilingual System for CLEF), évalué à CLEF 2004, utilise trois stratégies : la traduction des questions en anglais par la version professionnelle de Systran existant au CEA puis application de QALC sur la collection CLEF d'une part et sur le Web d'autre part, et la traduction des termes de la question, après analyse des questions en français, permettant la recherche dans la collection CLEF, ce qui a conduit au système MUSQAT.

Après une présentation des travaux existant en question-réponse multilingue section 2, nous donnons une vision générale de notre système section 3, pour détailler ensuite les aspects traduction en section 4, la combinaison des listes résultats en section 5 et les résultats obtenus en section 6 avant de conclure.

2. Multilinguisme et question-réponse

Différentes solutions existent pour gérer le multi-linguisme, dans les systèmes de recherche d'information en général et dans les systèmes de question-réponse en particulier. La première consiste à utiliser un traducteur automatique afin de traduire

[3] Text Retrieval Evaluation Conference.

les questions et appliquer un système monolingue par la suite. C'est le choix effectué par [PER04], [JIJ04], [NEU03], [NEU04], [AHN04]. [PER04] et [JIJ04] ont aussi appliqué leur système en version monolingue. Le premier, dans sa version anglais-hollandais, obtient une baisse de 10,5 % sur ses résultats : de 91 (45,5 %) à 70 (35 %) réponses correctes, et les résultats du second, dans sa version anglais-français, voit son pourcentage de bonnes réponses diminuer de 13,5 % : de 49 (24,5 %) à 22 (11 %) réponses. Pour leur système BiQue, Neumann et Sacaleanu [NEU03], [NEU04] ont eu recours à plusieurs outils de traduction pour obtenir une bonne couverture : 3 en 2003 et 8 en 2004. Les lemmes issus de la fusion des différentes traductions sont réunis et constituent un « sac-de-mots » (bag-of-object) qui est utilisé lors de l'expansion de la requête. L'expansion consiste à compléter le sac-de-mots avec des synonymes mais un module de désambiguïsation est alors nécessaire. Ce module utilise EuroWordNet pour connaître les correspondances entre termes dans les 2 langues (anglais et allemand) et, pour chaque mot ambigu, il regarde lesquels de ses sens sont exprimés à la fois dans la question d'origine (en allemand) et dans ses traductions (en anglais).

Les deux problèmes majeurs dans l'utilisation de traducteurs pour les questions résident dans la non (ou la mauvaise) résolution de l'ambiguïté des mots des questions et des traductions syntaxiquement fausses, comme nous le verrons section 4.1. Si un mot pertinent pour la recherche de la réponse est mal traduit, cette erreur ne peut en général être rattrapée par la suite par seule compensation des autres mots de la question. En effet les questions sont souvent assez courtes, et la mauvaise traduction d'un mot en change le sens. Si la question produite est agrammaticale, elle est alors mal analysée, les systèmes de question réponse appliquant fréquemment des analyseurs syntaxiques pour extraire les caractéristiques utiles. C'est pour cette raison que Ahn *et al.* [AHN04] ont choisi de développer en amont et en aval de la traduction de la question produite par Babelfish, un ensemble de règles de transformation, de façon à prévenir ou rectifier des erreurs de traduction assez systématiques sur les formulations syntaxiques des questions[4]. Ainsi, les questions en français « *À quel moment…* » ont été transformées en « *Quand…* » avant d'être soumises au traducteur. De même, les questions avec inversion et reprise pronominale du sujet ont été corrigées après traduction, pour éviter que « *Où X travaille-t-il* ? » ne devienne « *Where X does it work ?* ». Des ensembles de règles similaires ont été développés pour leur système allemand-anglais.

Une deuxième approche consiste à traduire les documents. Dans ce cas, le contexte de traduction est plus grand et donc plus fiable pour gérer l'ambiguïté. Un inconvénient majeur est que la collection résultante est *n* fois plus grande après traduction en *n* langues. Et il n'est pas question de traduire le Web avant interrogation !

La dernière solution consiste à analyser la question dans la langue source et en extraire toutes les caractéristiques utiles à l'extraction des réponses, c'est-à-dire le

[4] Ils ont constaté que sur la traduction des 200 questions de CLEF03 de l'allemand vers l'anglais par Babelfish, seules 29 % étaient jugées acceptables par un juge linguiste.

type de la réponse attendue, les termes importants, les groupes de la phrases (nominaux, verbaux et prépositionnels) ainsi que la structure syntaxique complète (liens de dépendance entre groupes et étiquetage des fonctions grammaticales). La nature de ces informations est la même quelle que soit la langue, hormis les fonctions grammaticales. Aussi, si ces dernières ne sont pas utilisées, ce qui est le cas dans beaucoup de systèmes car elles ne sont pas très fiables, seuls les termes peuvent être traduits, indépendamment de la structure syntaxique complète de la question. Cela ramène au seul problème de la gestion de la polysémie des mots. Nous verrons nos résultats section 4.3 quant aux performances de notre traduction. Cette solution a été choisie par [SUT 04] et [NEG03], [TAN04] dans leur système Diogène qui a participé aux évaluations CLEF avec une tâche monolingue (italien) et 2 tâches bilingues (bulgare/italien vers l'anglais). Tanev ***et al.*** ont jugé que les résultats de la traduction automatique, particulièrement pour les questions, n'étaient pas assez encourageants. Ils ont donc eu recours à une traduction des mots clefs de la question : après une étape d'élimination des mots non pertinents, ils traduisent les mots clefs de la question à l'aide d'un dictionnaire bilingue et de MultiWordNet. Puis, afin d'éliminer le bruit inhérent à un tel procédé, ils ne retiennent que les combinaisons de traduction les plus plausibles, i.e. celles qui apparaissent le plus fréquemment dans 2 corpus de référence (AQUAINT et TIPSTER). L'étape suivante est l'expansion des mots clefs à l'aide de dérivations morphologiques et sémantiques. Ils obtiennent un score de 45 (22,5 %) bonnes réponses en version bilingue contre 56 (28 %) en monolingue, donc avec une perte de 6 % de bonnes réponses seulement. Sutcliffe ***et al.*** [SUT04] ont choisi pour leur part de traduire tous les syntagmes issus de l'analyse des questions par un analyseur de surface selon trois méthodes : deux traducteurs (Reverso et WorldLingo) et un dictionnaire (GDT, Grand Dictionnaire Terminologique). Les résultats obtenus sont ensuite combinés, en donnant la préférence au GDT quand le syntagme y figure. L'ensemble des syntagmes traduits sert ultérieurement à la constitution des requêtes.

Les systèmes de question-réponse comportent tous les mêmes grands modules : analyse des questions, traitement des documents ou passages sélectionnés et extraction des réponses. Ils diffèrent dans leur architecture et dans la nature des processus mis en œuvre. On peut classer les systèmes en trois grandes catégories : les systèmes qui opèrent une analyse syntaxico-sémantique des questions et des réponses afin de les apparier [MOL02], [HAR04], [AHN04], les systèmes qui utilisent des processus plus robustes reposant sur des mesures de similarité plus statistiques, même si il y a utilisation de grammaires locales pour extraire les réponses seules, dont notre système fait partie, et les systèmes multi-stratégies [CHA03], [JIJ04], [CHU02] combinant ensuite les différents résultats. [JIJ04] combinent 8 résultats obtenus par des stratégies de résolution différentes (4 au total) ou l'utilisation de sources de connaissances différentes (collection anglaise, collection hollandaise, Web et encyclopédie hollandaise). Les stratégies appliquées reprennent des types de solutions proposées par beaucoup de systèmes, mais leur application en est différente. Par exemple, les patrons d'extraction de la réponse dérivés des questions sont appliqués sur toute la collection et pas seulement sur des passages sélectionnés.

3. Présentation générale de MUSCLEF

L'architecture globale de MUSCLEF est illustrée Figure 2. Elle regroupe l'application de QALC à partir des questions traduites sur la collection CLEF et le Web et la version « bilingue » MUSQAT utilisant les résultats de l'analyse des phrases en français. Ensuite les mêmes modules s'appliquent sur les documents trouvés en anglais.

Le but du module d'analyse de la question est de déduire les caractéristiques qui peuvent contribuer à trouver les réponses possibles dans les passages retenus et de reformuler les questions sous forme déclarative pour le moteur de recherche sur le Web (Google). Ces caractéristiques sont le focus de la question, le verbe principal et les fonctions syntaxiques des modifieurs. Nous avons concentré nos efforts de traduction sur ces éléments, comme cela sera précisé dans la section suivante. Pour la campagne CLEF 04, nous avons développé une nouvelle version de ce module pour analyser les questions en français. La version française de l'analyseur syntaxique XIP [AIT02] sert de base à ce module. Pour l'analyse des questions traduites, nous utilisons IFSP [AIT97].

Les requêtes ne sont pas identiques pour la recherche sur le Web et pour la recherche dans la collection CLEF. Dans le second cas, nous utilisons MG pour retrouver les passages. Pour interroger sur le Web, nous envoyons une reformulation presque exacte de la réponse, en faisant l'hypothèse que la redondance du Web permettra toujours de sélectionner des documents.

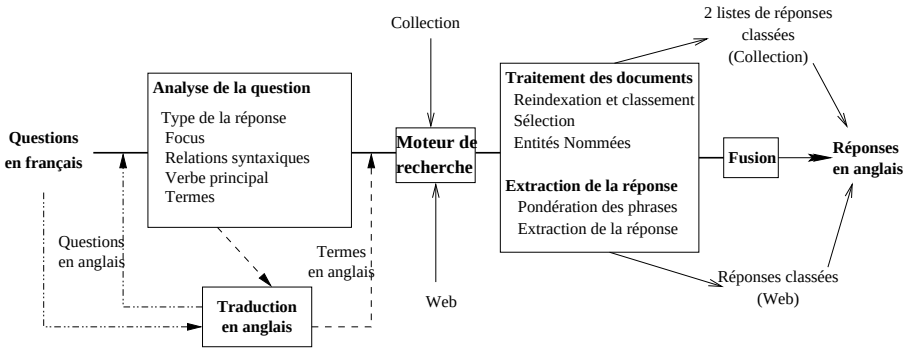


Figure 2 : Architecture de MUSCLEF

Le traitement des documents renvoyés par le moteur commence ensuite. Ils sont ré-indexés par les termes de la question et leurs variantes linguistiques, ré-ordonnés en fonction du nombre et du type des termes qu'ils contiennent, afin de n'en retenir qu'un sous-ensemble. La reconnaissance des différents types d'entités nommées est alors effectuée. L'extraction de la réponse consiste à d'abord attribuer des poids aux phrases avant d'extraire les réponses elles-mêmes. Différents processus sont appliqués suivant le type de réponse attendu, chacun d'eux ayant pour

résultat des réponses pondérées. La dernière étape consiste à combiner les réponses trouvées dans la collection, les réponses issues du Web et les résultats issus du système MUSQAT. Un score final est alors calculé ; son principe est de privilégier une réponse qui a été classée dans les cinq premières possibilités par deux chaînes au moins, même si ses scores individuels sont moindres.

4. Analyse des questions

Comme nous l'avons indiqué plus haut (voir figure 2), deux solutions ont été testées pour représenter les informations contenues dans les questions afin de permettre leur comparaison avec les documents. Dans la première, nous faisons appel à un traducteur automatique pour passer la question du français à l'anglais et ensuite procéder à l'analyse des questions traduites. La seconde solution consiste à analyser les questions dans la langue source (le français dans notre cas) et à traduire en anglais (notre langue cible), ceux des termes qui ont été considérés comme les plus importants, ce qui entraîne aussi la traduction des différentes caractéristiques telles le focus et le verbe principal.

4.1 Traduction automatique des questions

La première solution que nous avons donc testée pour résoudre la différence de langues entre les questions et les documents consiste à faire appel à un traducteur automatique sur les questions. Dans notre cas, cette traduction automatique a été effectuée par l'interface en ligne SYSTRANLinks fournie par Systran (nous tenons à remercier Systran qui nous a donné accès à ce service dans le cadre du projet ALMA). Nous n'avons fait appel à aucun dictionnaire complémentaire. Comme la plupart des questions de l'évaluation CLEF ne présentait pas une grande complexité syntaxique et concernait des sujets généraux, leur traduction peut souvent être considérée comme fiable, comme l'illustre la figure 3.

0009 - Quand est apparu pour la première fois le virus Ebola ?
0009 - When did the Ebola virus appear for the first time?

0166 - Où se trouve Halifax ?
0166 - Where is Halifax?

Figure 3 : Exemples de questions correctement traduites

Cependant, les erreurs de traduction peuvent aussi se produire pour des questions simples, comme l'illustre la figure 4. Ces erreurs peuvent concerner la syntaxe. Dans la question 175 par exemple, « *Quel est* » devrait être traduit par « *Who is* » et non pas par « *Which is* ». De même, dans la question 165, l'expression « *Qu'est-ce que* », spécifique aux questions, n'est que partiellement traduite. En fait cette observation montre que, tout comme pour les étiqueteurs morpho-syntaxiques et les analyseurs syntaxiques, les questions devraient être spécifiquement considérées par les systèmes de traduction, ce qui n'est généralement pas le cas. Les erreurs sont aussi d'ordre sémantique. Dans la question 175 à nouveau, « *réalisateur* » est traduit par « *realizer* » alors qu'il y a plus de chance de trouver une réponse avec une traduction comme « *director* » ou « *film director* ». Enfin la question 165 montre aussi le problème posé par le fait que les dictionnaires sont incomplets, ce qui est inévitable dans un système en domaine ouvert, spécialement pour les acronymes. Ainsi, « *OMC* » (Organisation Mondiale du Commerce) devrait être traduit en « *WTO* » (World Trade Organization), tout comme « *OTAN* » est traduit en « *NATO* » dans la question 143.

0175 - Quel est le réalisateur de "Nikita" ?
0175 - Which is the realizer of "Nikita"?

0165 - Qu'est-ce que L'OMC ?
0165 - What OMC?

0143 - En quelle année a été créée l'OTAN ?
0143 - In which year was created NATO?

Figure 4 : Exemples d'erreurs dans les traductions des questions

4.2 Evaluation des analyses

119 questions sur 200 attendent une entité nommée en réponse. Les types d'entités nommées que nous avons définis sont les types classiques légèrement raffinés : personne, organisation, lieu-endroit, ville, nom propre, nombre, pourcentage, montant financier, quantité physique (surface), vitesse, poids, volume, longueur, température, age, heure, date (différents types de dates tels que jour, jour et mois, etc.), durée et période. Le typage de la réponse sur le français a une précision de 99 % et un rappel de 95 %. Sur l'anglais, la précision est la même, mais le rappel est de 83 %. Il y a 14 types de réponses non reconnus supplémentaires, dus à des formes syntaxiques inhabituelles ou fausses en anglais.

4.3 Traduction de termes

Différentes méthodes peuvent être utilisées pour traduire les termes. Les résultats peuvent être obtenus par une traduction fondée sur des ontologies bilingues ; mais comme nous venons de le voir dans la section précédente, celles-ci n'existent pas en domaine ouvert ou ne sont pas suffisamment complètes. Parmi les autres possibilités de traduction, nous nous sommes intéressés à la plus simple, consistant à utiliser un dictionnaire bilingue pour traduire les termes de la langue source vers la langue cible. Cette méthode présente deux inconvénients : d'une part, il est impossible de lever directement les ambiguïtés sur les différents sens des mots à traduire, d'autre part la richesse lexicale des deux langues doit être comparable. Comme cette dernière contrainte est vérifiée pour le couple français/anglais, nous avons décidé d'utiliser quand même cette méthode. Toutefois, pour avoir une estimation des ambiguïtés que nous risquions de rencontrer dans le contexte Question-Réponse, nous avons étudié le corpus des 1 893 questions en anglais de TREC. Après analyse, nous avons conservé 9 000 des 15 624 mots utilisés dans ce corpus. La moyenne du nombre de leurs sens est de 7,35 dans WordNet. Les valeurs extrêmes sont 1 (pour *neurological* par exemple) et 59 (pour *break* par exemple). Autour de la valeur moyenne, on trouve des mots communs, tels que *prize*, *blood*, *organization*. De ce fait, nous ne pouvons pas considérer un dictionnaire donnant seulement un sens par mot.

Connaissant ces contraintes, nous avons étudié les différents dictionnaires que nous pouvions utiliser : d'une part les dictionnaires en ligne, tels que Reverso[5], Systran[6], Google[7], Dictionnaire Terminologique[8] ou FreeTranslation[9], et d'autre part les dictionnaires sous licence GPL, tels que Magic-Dic[10] ou Unidic. Les dictionnaires en ligne sont généralement complets. Mais ils résolvent les ambiguïtés (ou tentent de le faire) et ne fournissent qu'une traduction par mot. Un autre inconvénient que nous avons constaté a été l'impossibilité de modifier ces dictionnaires et l'obligation de tenir compte de quelques contraintes techniques telles que le nombre limité de requêtes que nous pouvions faire et les temps de réponse. En ce qui concerne les dictionnaires GPL, ils sont notoirement moins complets, mais ils peuvent être augmentés. De plus, ils sont très rapides et pour la plupart, ils donnent plusieurs traductions pour une requête. Parmi les dictionnaires GPL, nous avons choisi Magic-dic, du fait de ses capacités d'évolution : des termes peuvent être ajoutés par n'importe quel utilisateur, mais ils sont vérifiés avant d'être intégrés, ce qui n'est pas le cas d'Unidic. Par exemple, le mot porte donne les résultats suivants (nous n'en donnons qu'un extrait) :

- porte bagages : luggagerack, luggage rack,

---

5 http://translation2.paralink.com.
6 http://babel.altavista/translate.dyn.
7 http://www.google.com/language.tools.
8 http://granddictionnaire.com.
9 http://www.freetranslation.com.
10 http://magic-dic.homeunix.net/.

- porte cigarettes : cigarette holder,
- porte clefs : key-ring,
- porte plume : fountain pen,
- porte parole, locuteur : spokesman,
- porte : door, gate.

4.4 Module multilingue

Nous allons illustrer la stratégie de MUSQAT sur l'exemple suivant : *Quel est le nom de la principale compagnie aérienne allemande ?*, qui est traduite en anglais par : *What is the name of the main German airline company?*.

La première étape est l'analyse de la question en français, qui fournit une liste de tous les mono-termes et de tous les bi-termes (tels que adjectif/nom commun) présents dans la question, et élimine les mots vides. Les bi-termes sont utiles, parce qu'ils permettent de désambiguïser en donnant un (petit) contexte à un mot. Dans notre exemple, les bi-termes (sous leur forme lemmatisée) sont : *principal compagnie*, *compagnie aérien*, *aérien allemand* et les mono-termes : *nom*, *principal*, *compagnie*, *aérien*, *allemand*.

En s'aidant du dictionnaire Magic-dic, nous avons essayé de traduire les bi-termes (quand ils existent), et les mono-termes. Toutes les traductions proposées sont conservées. Tous les termes sont étiquetés grammaticalement. Si un bi-terme ne peut pas être directement traduit, il est reconstitué à partir des traductions des mono-termes qui le composent, en suivant la syntaxe anglaise. Pour notre exemple, nous avons obtenu pour les bi-termes : *principal compagny/main compagny*, *air compagny*, *air german* ; et pour les mono-termes : *name/appellation*, *principal/main*, *compagny*, *german*. Quand un mot est absent du dictionnaire, nous le conservons à l'identique en supprimant les signes diacritiques.

Ces termes, avec leurs catégories remplacent alors les mots d'origine en entrée des autres modules de MUSQAT. Le module de traduction n'essaie pas de résoudre les ambiguïtés entre les différentes traductions : la requête envoyée au moteur MG est constituée de l'union de toutes les traductions et la levée d'ambiguïtés a lieu éventuellement lors de la sélection des documents par le moteur ou après ré-indexation. Si les différents termes sont synonymes, des documents pertinents sont trouvés avec ces synonymes, donnant ainsi lieu à une recherche plus large. Si le mot est incohérent dans le contexte du document, nous faisons l'hypothèse que son influence n'est pas suffisante pour créer du bruit.

Nous avons évalué la traduction produite par MUSQAT. Les 200 questions en français contenaient 731 mots, correspondant à 1 091 mots anglais, et 932 termes (mono-termes + bi-termes) correspondant à 1 464 termes en anglais. En étudiant ces traductions, nous avons observé que :

- 59 % des termes traduits étaient corrects (même si pour 12,63 % des termes la traduction pourrait être améliorée) ;
- 8 % des termes traduits étaient corrects mais identiques aux termes dans la langue source ;
- 33 % des termes traduits étaient incorrects.

Il est évident que le dictionnaire n'était pas assez complet pour cette campagne. Nous devrions obtenir une meilleure couverture en le complétant manuellement avec les traductions manquantes (il n'y a pas par exemple de traduction de verbe français *jouer* dans son sens *to play*). Pour pallier son incomplétude, et parce qu'il a été prouvé qu'utiliser plusieurs dictionnaires donne de meilleurs résultats que d'en utiliser un seul, nous avons l'intention de l'enrichir aussi avec des dictionnaires en ligne.

Une autre évaluation concerne les bi-termes, que nous avons présentés comme très important dans la perspective de désambiguïser les mono-termes ambigus. Pour cela, nous avons déterminé la fréquence documentaire de chaque traduction des différents bi-termes dans le corpus CLEF. Si la fréquence est élevée, alors le bi-terme peut être considéré comme une traduction adéquate. Suivant cette étude, 47,5 % des bi-termes ont été trouvés dans le corpus. Une approche prometteuse pourra être de valider les traductions, en les notant en fonction de leurs fréquences, à la fois dans un corpus bilingue aligné, et dans un corpus monolingue (dans la langue cible).

Nous avons également noté qu'un travail important restait à faire sur les noms propres, spécialement les noms géographiques, les noms d'organisation et les acronymes. Il faudra pour cela développer des listes bilingues des noms les plus fréquents.

5. Combinaison des listes résultats

Comme cela a été dit dans la section 3, nos résultats sont obtenus en comparant 3 ensembles de réponses : le premier de ces ensembles est retourné par MUSQAT, le second par QALC utilisant le Web comme collection et le troisième par QALC appliqué à la collection CLEF. Le Web constitue une source de connaissances beaucoup plus large que la collection CLEF, le fait d'utiliser une telle source nous permet de confirmer les réponses trouvées dans la collection et donc de renforcer leur score de confiance. Notons que les réponses provenant du Web doivent aussi appartenir à la collection, sinon elles ne sont pas acceptées comme réponses correctes : en effet toutes les réponses doivent être accompagnées d'un document les justifiant.

Chacun de ces 3 ensembles contient pour chaque question un ensemble de réponses ordonnées selon leur score de confiance, score mis à jour tout au long du processus d'extraction de la réponse. Avant de décrire l'algorithme écrit pour la

sélection finale, nous décrivons la façon dont ce score est attribué à chaque réponse candidate.

5.1 Pondération des réponses par chaque système

Toutes les phrases issues du traitement des documents sont examinées dans le but de leur attribuer un poids reflétant à la fois la possibilité qu'elle contienne la réponse et la possibilité que le système y localise la réponse. Les critères retenus pour calculer ce poids sont liés aux informations contenues dans la question. Les traits suivants sont ainsi recherchés dans la phrase candidate :

- Les lemmes de la question, chacun possédant un poids qui représente son degré de spécificité[11] ;
- Les variantes de ces lemmes ;
- Les mots exacts de la question (seulement dans le cas de la version « tout en anglais ») ;
- La proximité mutuelle des mots de la question ;
- La présence des entités nommées attendues.

Un premier poids est calculé tout d'abord en fonction de la présence des lemmes et de leurs variantes (les 2 premiers critères). Ensuite, à ce poids est ajouté un poids additionnel pour chaque autre critère satisfait, les poids additionnels ne pouvant dépasser 10 % du poids d'origine.

Au cours de l'extraction de la réponse, un poids est attribué à chaque réponse potentielle. Pour ordonner les réponses de type Entité Nommée, MUSCLEF (resp. QALC) calcule donc des poids additionnels prenant en compte :

- L'entité nommée précise ou généralisée de la réponse ;
- La place de la réponse par rapport à celles des mots de la question dans la phrase ;
- La redondance de la réponse dans les 10 premières phrases candidates.

Quand le type attendu de la réponse n'est pas une entité nommée, nous utilisons des patrons d'extraction. Chaque phrase candidate est analysée en utilisant les patrons d'extraction associés au type de la question (déterminé lors de l'analyse de la question). Ces patrons d'extraction sont écrits avec des expressions régulières qui utilisent le focus comme pivot et sont pondérés en fonction de leur spécificité. Davantage de détails seront trouvés dans [FER02]. A la fin de l'extraction, les 5 réponses de meilleur score sont retenues pour la sélection finale.

[11] Le degré de spécificité d'un lemme est calculé en fonction de l'inverse de sa fréquence relative calculée sur un grand corpus.

5.2 Algorithme de sélection de la réponse

L'idée ici est de comparer des résultats provenant de différentes sources de connaissance afin de renforcer le score des réponses qui appartiennent à plusieurs ensembles, permettant ainsi à un certain nombre de bonnes réponse d'atteindre le premier rang. Le tableau 1 contient un exemple de ces ensembles de réponses, pour la question « *En quelle année Thomas Mann a-t-il obtenu le Prix Nobel ?* ».

Les 3 ensembles de réponses sont comparés 2 à 2 en utilisant un algorithme écrit pour les évaluations TREC. Cet algorithme examine chaque couple ($réponse_i$, $réponse_j$), i et j sont compris entre 0 et 4 et représentent la position de la réponse dans son ensemble. Quand les 2 réponses sont égales ou incluses l'une dans l'autre, l'algorithme attribue un bonus au meilleur score du couple. Ce bonus a été choisi afin de permettre aux réponses confirmées de passer devant les réponses non confirmées ; il est calculé en fonction des positions i et j : (10 – (i + j)) * 100. De cette façon, l'algorithme construit un ensemble de couples ordonnés en fonction de leur nouveau score. Comme dans CLEF il y avait non pas 2 mais 3 ensembles de réponses, nous avons appliqué cet algorithme sur les trois ensembles de couples de réponses c'est-à-dire 3 fois, la réponse finalement retournée étant celle qui obtient le meilleur score.

En consultant le tableau 1, on voit que 2 dates sont présentes dans les 3 ensembles : « *en 1929* » et « *en 1976* ». Le couple qui apparaît en gras est celui qui obtient le meilleur score final : il a reçu un bonus de 900 points pour un score d'origine de 1 082 points. La réponse « e*n 1929* » est donc retournée avec un score final de 1 982. C'est effectivement la bonne réponse.

	QALC + Web		**MUSQAT**		**QALC + Collection**	
Rang	Réponse	Score	Réponse	Score	Réponse	Score
0	**en 1929**	**1 082**	en 1976	721	11 Octobre 1994	878
1	1875-1955	1 005	en 1976	721	**en 1929**	**853**
2	8 Mars 1879	903	en 1929	664	en 1976	798
3	en 1903	877	2	640	12 Octobre 1994	703
4	en 1929	849	1964	561	en 1979	696

Tableau 1 : Un exemple des 3 ensembles de réponses

Comme on vient de le constater, cet algorithme qui effectue les comparaisons sur 2 ensembles est facilement applicable à plus de 2 ensembles, puisqu'il suffit de répéter les comparaisons autant que nécessaire. Néanmoins, nous avons observé qu'une comparaison menée d'emblée sur les 3 ensembles aurait donné des résultats

différents. En effet, en menant la comparaison 2 à 2, nous ne prenons pas en compte de la même façon les réponses présentes dans les 3 ensembles.

6. Résultats

Le tableau 2 présente une évaluation comparative de MUSQAT et de QALC. Cette évaluation a été faite de façon automatique en recherchant dans les phrases réponses les expressions régulières correspondant aux patrons de réponses. Ces résultats ont été calculés pour les 178 questions pour lesquelles nous avions un patron de réponse[12].

La première ligne indique le nombre de réponses correctes trouvées dans les 5 premières phrases retournées par MUSQAT et par les 2 applications de QALC (sur le Web et sur la collection). La deuxième ligne « Réponses à EN » donne le nombre de réponses correctes pour les questions attendant une EN, et la troisième ligne concerne les autres questions. Les résultats sont donnés pour la réponse au rang 1 et pour les 5 premières réponses. La dernière colonne indique le meilleur résultat de notre système, obtenu en utilisant l'algorithme de fusion décrit section 5.2. Le score officiel de MUSQAT étant de 22 bonnes réponses (11 %), nous observons qu'en fusionnant les réponses des différentes stratégies nous avons un gain de 17 réponses (77 %). Ce dernier résultat nous place à égalité avec les systèmes français-anglais de CLEF04.

		MUSQAT	**QALC + Collection**	**QALC + Web**	**Fusion** 200 questions
		178 patrons (200 questions)			
Phrases	5 premiers rangs	56	65	61	
Réponses à EN	rang 1 5 premiers rangs	17 32	26 37	24 43	
Réponses non EN	rang 1 5 premiers rangs	7 12	3 8	0 0	
Total	rang 1 5 premiers rangs	24 (12%) 44	29 (14,5%) 45	24 (12%) 43	39 (19,5 %)

Tableau 2 : Evaluation comparative des différentes stratégies

[12] Les autres sont supposées être des questions n'ayant pas de réponse dans le corpus. Comme nous n'en avons reconnu aucune dans nos propositions, nous supposerons que les 178 patrons sont les réponses aux 200 questions afin de ramener le nombre de réponses au même total.

Pour l'évaluation TREC 2002, le système devait aussi produire pour chaque question une unique réponse. Le tableau 3 donne le nombre de bonnes réponses obtenues. On peut constater que la fusion apporte un gain de 29 réponses, soit 21 % seulement en comparaison du précédent. Cela peut s'expliquer par le meilleur classement intrinsèque des bonnes réponses par le système en monolingue. Lorsque l'on évalue QALC sur les 5 premières réponses, 30 % environ des bonnes réponses se situent après le premier rang, alors que dans MUSQAT ou QALC-CLEF, il y en a plutôt 35,5 %. L'autre argument réside dans le fait que l'on fusionne trois sources de résultats au lieu de deux ; de plus, les résultats sont issus de stratégies différentes et pas seulement de sources de connaissances différentes.

Résultats TREC 11 500 questions	QALC (Collection Trec)	Fusion Collection + Web
Résultats officiels		139 (27,8 %)
Evaluation automatique	136 (27,2 %)	165 (33 %)

Tableau 3 : Résultats obtenus à TREC 11 par QALC

Les résultats de la dernière ligne du tableau 3 peuvent être comparés à ceux de la dernière ligne du tableau 2 pour les systèmes seuls : en ce qui concerne les bonnes réponses au rang 1, on constate que le problème du multilinguisme a entraîné une baisse de 13 dans les pourcentages de bonnes réponses, ce qui est comparable aux systèmes présentés dans la section 2.

On peut aussi remarquer que les 3 stratégies sont équivalentes en nombre de réponses. Mais si on compare l'ensemble des 5 premières réponses données pour chaque question par MUSQAT et par QALC appliqué à la collection, on voit qu'il y a seulement 21 réponses communes, et donc 22 trouvées uniquement par MUSQAT et 24 par QALC. Nous obtenons les mêmes chiffres en comparant les autres résultats deux à deux, avec toutefois un peu plus de réponses en commun entre les deux applications de QALC, ce qui s'explique par le fait de partir de la même formulation des questions. Aussi, même si certaines réponses sont proposées car de meilleur poids, il y a lieu de chercher à améliorer le choix de la réponse lorsqu'elle est présente dans une seule liste, puisque nous disposons au total de 65 bonnes réponses parmi les 5 meilleures par couple de résultats. Un point faible de notre système reste l'extraction des réponses dans le cas des questions n'attendant pas une entité nommée comme en témoigne la 3ème ligne du tableau 2.

7. Conclusion

Même si ces résultats sont encourageants, MUSCLEF, notre premier système multilingue peut encore être amélioré. Son architecture, organisée en plusieurs modules indépendants, a été choisie de façon à permettre aisément ces améliorations. En outre, nous avons vu que les 2 stratégies adoptées : la traduction des termes importants et la traduction des questions, étaient pertinentes et devaient être maintenues dans des expériences futures. Elles sont assez complémentaires et une amélioration consisterait à fusionner les propositions des deux traductions dans chaque système, afin de fournir des synonymes au moment de la recherche des documents. Bien entendu de meilleures ressources sémantiques seront indispensables, mais comme de telles ressources ne sont pas facilement disponibles, utiliser des traductions attestées en corpus pour contrôler les traductions pourrait être une piste intéressante à étudier.

8. Références bibliographiques

[AHN04] K. Ahn, B. Alex, J. Bos, T. Dalmas, J.L. Leidner et M.B. Smillie, 2004, Cross-lingual Question Answering with QED, Working Notes, CLEF Cross-Language Evaluation Forum, Bath UK, p. 335-342.

[AIT97] S. Aït-Mokhtar et J.-P. Chanod, 1997, Incremental finite-state parsing. In Proceedings of the 5th Conference on Applied Natural Language, Processing (ANLP-97), Washington, DC, USA , p. 72-79.

[AIT02] S. Aït -Mokhtar, J.-P. Chanod et C. Roux, 2002, Robustness beyond shallowness: incremental deep parsing. Natural Language Engineering, vol. 8 (2/3), p. 121-144.

[BRI01] E. Brill, J. Lin, M. Banko, S. Dumais et A. Ng, 2001. Data-Intensive Question Answering. TREC 10 Notebook, Gaithersburg, USA.

[CHA03] G. de Chalendar, F. El Kateb, O. Ferret, B. Grau, M. Hurault-Plantet, L. Monceaux, I. Robba, A. Vilnat, 2003, Confronter des sources de connaissances différentes pour obtenir une réponse plus fiable, TALN 03, Batz sur Mer, p. 105-114.

[CHU02] J. Chu-Carroll, J. Prager, C. Welty, K. Czuba et D. Ferruci, 2002. A Multi-Strategy and multi-source Approach to Question Answering. TREC 11 Notebook, Gaithersburg, USA, p. 124-133.

[CLA01] C. L. Clarke, G. V. Cormack, T. R. Lynam, C. M. Li et G. L. McLearn, 2001, Web Reinforced Question Answering (MultiText Experiments for Trec 2001), TREC 10 Notebook, Gaithersburg, USA.

[FER02] O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, C. Jacquemin, L. Monceaux, I. Robba et A.Vilnat, 2002, How NLP Can Improve Question Answering, Knowledge Organization, vol. 29, n° 3-4, p. 135-155.

[HAR04] S. Hartrumpf, 2004, Question answering using sentence Parsing and Semantic Network Matching, Working Notes, CLEF Cross-Language Evaluation Forum, Bath UK, p. 385-393.

[HER02] U. Hermjakob, A. Echihabi et D. Marcu, 2002, Natural Language Based Reformulation Resource and Web Exploitation for Question Answering, TREC 11 Notebook, Gaithersburg, USA.

[JIJ04] V. Jijkoun, G. Mishne, M. de Rijke, S. Schlobach, D. Ahn et K. Muller, 2004, The University of Amsterdam at QA@CLEF2004, Working Notes, CLEF Cross-Language Evaluation Forum, Bath UK, p. 321-325.

[MAG02a] B. Magnini, M. Negri, R. Prevete et H. Tanev, 2002, Is It the Right Answer? Exploiting Web redundancy for Answer Validation, Proceedings of the 40th ACL, p. 425-432.

[MAG02b] B. Magnini, M. Negri, R. Prevete et H. Tanev, 2002, Mining Knowledge from Repeated Co-occurrences: DIOGENE at TREC-2002, TREC 11 Notebook, Gaithersburg, USA.

[MOL02] D. Moldovan, S. Harabagiu, R. Girju, P. Morarescu, F. Lacatusu, A. Novischi, A. Badalescu et O. Bolohan, 2002, LCC Tools for Question Answering, TREC 11 Notebook, Gaithersburg, USA.

[NEG03] M. Negri, H. Tanev et B. Magnini, 2003, Bridging Languages for Question_Answering/ DIOGENE at CLEF2003, Working Notes, CLEF Cross-Language Evaluation Forum, Trondheim, Norvège.

[NEU03] G. Neumann et B. Sacaleanu, 2003, A Cross-Language Question / Answering-System for German and English, Working Notes, CLEF Cross-Language Evaluation Forum, Trondheim, Norvège.

[NEU04] G. Neumann et B. Sacaleanu, 2004, Experiments on Robust NL Question Interpretation and Multi-layered Document Annotation for a Cross-Language Question / Answering System, Working Notes, CLEF Cross-Language Evaluation Forum, Bath UK, p. 311-320.

[PER04] L. Perret, 2004, Question Answering System for the French Language, Working Notes, CLEF Cross-Language Evaluation Forum, Bath UK, p. 295-305.

[SUT04] R. Sutcliffe, I. Gabbay, M. Mulcahy et A. O'Gorman, 2004, Cross-Language French-English Question Answering using the DLT System at CLEF-2004, Working Notes de CLEF Cross-Language Evaluation Forum, Bath UK, p. 305-309.

[TAN04] H. Tanev, M. Negri, B. Magnini et M. Kouylekov, 2004, The DIOGENE Question Answering System at CLEF-2004, Working Notes de CLEF Cross-Language Evaluation Forum, Bath UK, p. 325-333.

Personnalisation des services Web : évaluation des sites fédérateurs (SFQC)

Omar Larouk, Salah Dalhoumi

Systèmes d'Information et Interfaces (SII – URSIDOC, ENSSIB)
Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques
ENSSIB Lyon, 17 bd du 11 Novembre 1918, 69623 Lyon-Villeurbanne Cedex - France

{larouk,dalhoumi}@enssib.fr

Résumé :

Il existe un désordre dans les ressources numériques en ligne. Ces ressources sont très diversifiées et couvrent aussi bien des fichiers des systèmes de bases de données que des pages Web contenant des publications électroniques, et des méta-données. La plupart des sites fédérateurs utilisent des méthodes traditionnellement appliquées dans les bibliothèques. Les projets ont pour but de cataloguer, indexer, classifier les contenus électroniques, mais aussi de normaliser les formats de description et d'échanges. Les premiers, qui ont proposé ce type de sites, sont les spécialistes de l'informatique documentaire et les bibliothécaires. Leurs tâches consistent à structurer des pages web en offrant des *liens validés* vers les sites les plus intéressants d'un domaine précis. De création récente, ce type de sites centralise des accès par disciplines, par zones géographiques et/ou par langues, etc. Ces sites «passerelles» permettent d'orienter les usagers vers des sites, selon leurs besoins documentaires. Cette approche entre dans le cadre des travaux sur le filtrage des informations du web. Nous présentons une évaluation de ces sites fédérateurs de qualité contrôlée et l'organisation des connaissances du WEB.

Mots-clés : Website Gateways, Information Retrieval, Filtering, Metadata, Summarization, Knowledge organisation, Classification, Evaluation, Criteria, Personalisation. SFQC.

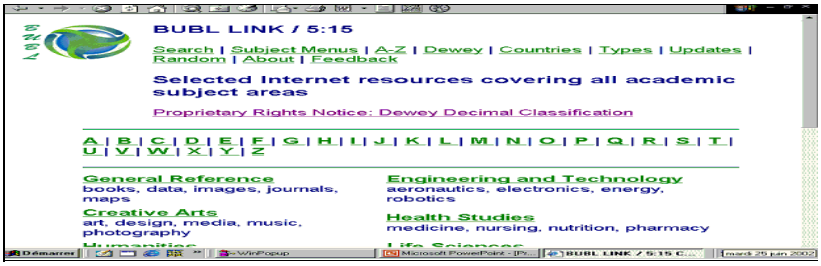
1. Organisation des connaissances du Web

Le Web permet de publier des informations accessibles à plusieurs millions d'utilisateurs mais la difficulté réside dans le repérage des informations pertinentes [LARO01]. Le contenu du Web ne cesse de croître, y compris par les sites Web actualisés ou non, voire volatiles. Les ressources documentaires accessibles par le biais de l'Internet, outre qu'elles sont de qualité et de stabilité variables, ne sont pas organisées. Il n'existe pas de base documentaire centralisée affectée à la recherche de ces ressources. A l'opposé des moteurs de recherche, - qui *prétendent* couvrir tous les domaines de connaissances -, il existe des sites web de regroupements dits "fédérateurs" produits par des spécialistes d'un domaine « précis ».

Les premiers, qui ont proposé ce type de sites, sont les bibliothécaires et les documentalistes. Leurs tâches consistent à structurer des pages web en offrant des *liens validés* vers les sites les plus intéressants d'un domaine. De création récente, ce type de site centralise des accès par disciplines, par zones géographiques et/ou par langues, etc. Ces sites « passerelles » permettent d'orienter les usagers vers des sites, selon leurs besoins documentaires. Cette approche entre dans le cadre des travaux sur le web sémantique[1] et sur le filtrage des informations qui essayent d'optimiser les systèmes de recherche d'information [LARO01].

2. Personnalisation des services Web : Passerelles, Annuaires spécialisés ou SBIG ?

Il existe donc des sites fédérateurs assimilés à des passerelles d'information spécialisées par sujets ou SBIGs (*Subject Based Information Gateways*). Leur lancement est l'œuvre de bibliothécaires et/ou informaticiens-documentalistes qui souhaitent "corriger" certaines insuffisances des services WEB. Un SBIG représentatif des sites fédérateurs est celui de BUBL (htpp://www.bubl.ac.uk)



[1] Voir aussi le site sur le « *web sémantique* » du W3C. http://www.w3.org/2001/sw/.

La principale caractéristique des concepteurs est d'y appliquer les méthodes issues de la bibliothéconomie et des sciences de l'information. Ces démarches nécessitent des indexeurs ou modérateurs de la documentation pour valider le contenu informationnel. [EDWA98].

Les sites SBIG [BRAN02], [AUER98] fournissent des accès par sujet à des ressources d'informations d'une certaine qualité. Les contenus sélectionnés sont évalués et doivent répondre à des critères bien précis. Les ressources retenues sont décrites par des experts qui, en plus du catalogage des documents, fournissent un résumé et des mots clés. Ces *classificateurs* [TILL02] attribuent à chaque document un code de classification, qui est extrait à partir des notices en vue d'une organisation de l'espace du site. Cependant, dans notre observation des différents sites, on peut dire que les sites SBIG peuvent être considérés comme des *sites spécialisés enrichis* par les compétences *de professionnels de l'informatique documentaire et des bibliothécaires* [LARO01].

3. Évaluation des services documentaires Web : démarche et usages

3.1 Usagers et accès au Web

La question principale qui sous-tend chaque évaluation de site est : "*l'usager va-t-il trouver sur ce site l'information, le service ou l'orientation qu'il recherche ?*". Un site peut s'avérer très insatisfaisant pour l'usager, même s'il est bien structuré. La première connexion d'un visiteur à la page d'accueil d'un site WEB est la plus importante. S'il la trouve « *efficace* », il aura envie de parcourir les pages suivantes. Si, par contre, elle le déçoit, il n'y reviendra plus, et il conservera une mauvaise image de l'institution (bibliothèques, entreprises, agences de presse, etc.). Un site WEB permet de valoriser l'offre informationnelle en transmettant une image positive ou négative de cette institution [HARRI03].

3.2 Utilité de la démarche d'instrumentation : Absences de repères documentaires classiques

La sélection des documents trouvés dans une bibliothèque dépend du temps pour lire ou traiter ces documents. En effet, la possession - ou l'accessibilité - du document n'entraîne pas automatiquement une lecture assidue. La pertinence du document est donc liée au travail demandé. Souvent, on consulte la table des matières ou la table des index pour avoir des pistes avant toute lecture sur un thème précis. La question, qu'on peut se poser, est :

Peut-on appliquer les repères documentaires classiques pour évaluer les documents "numériques" ?

On constate que les repères classiques utilisés pour évaluer le contenu des documents-ouvrages ne sont plus applicables de manière automatique aux documents numériques, compte tenu de la multiplicité des sources hétérogènes. En effet, les indices de qualité et de pertinence existent pour les livres et pour les périodiques en testant les différents champs bibliographiques de l'ouvrage (*thème, titre, auteur, nom de l'éditeur, type de collection, qualité scientifique, date d'édition, etc.*) lors d'une recherche documentaire.

De même, pour les articles de périodiques, la connaissance de la valeur scientifique de la revue dépend des éléments suivants : *membres du comité scientifique, qualité de la recherche, signatures, etc.* Les informations mentionnées sur le "papier" sont contextualisées.

Le contenu d'un ouvrage est généralement fiable, sans erreur, vérifié et comparable à d'autres sources ? Ce qui n'est pas le cas du contenu d'un site Web non-institutionnel. N'importe qui peut publier sur le web et ce sans vérifications. L'exhaustivité peut être couverte en profondeur ou superficiellement sur un domaine "précis" à l'aide de la documentation "papier" alors qu'elle est très difficile à évaluer sur le Web.

Les indicateurs précédents se retrouvent dans les documents numériques publiés sur le Web. En effet, un document classique écrit à l'aide d'un logiciel de bureautique permet de concevoir un manuel structuré avec des objectifs clairs et une justification du contenu. Une page "de traitement de texte" est bien définie par le format du logiciel avec une structure logique du document final, déterminée par des liens entre les chapitres et les sections (*suite logique de paragraphes*). Cette structure n'existe plus dans la technologie du WEB, la page "classique" d'un document est complètement transformée, même si elle est représentée parfois par la "page-écran".

4. Évaluation des services Web : Méta-critères des sites fédérateurs

4.1 Evaluation "usager" ou "concepteur" ?

Face l'abondance d'information sur le Web et face à l'absence des repères classiques de recherche documentaire manuelle, l'usager se trouve dans une situation « délicate » pour trouver la "bonne" référence ou le "bon" document. La prise en main par l'usager des outils informatiques a généré une nouvelle manière d'appréhender la recherche documentaire.

Un usager - *comme un concepteur* - se pose la question de l'utilité du contenu du site. En effet, un site est pertinent par rapport à ce que cherche l'usager et par rapport aux usages générés. Une grille d'évaluation qualitative ou quantitative permet à l'usager de voir les fonctionnalités de l'interface, l'ergonomie, l'habillage de l'application et surtout si le site répond au besoin documentaire et/ou informationnel. Cependant, un concepteur peut utiliser une grille d'évaluation pour mesurer la qualité de son site du côté « prescripteur » ou « fournisseur d'informations ».

Pour un enseignant, par exemple, une grille offerte aux étudiants en vue d'évaluer un site web "ciblé" entre dans le cadre pédagogique, c'est-à-dire celui de donner un moyen d'évaluation et de contrôle d'un processus de production. Généralement, les objectifs de l'usage sont pris en compte (objectifs pédagogiques) pour qu'ils appréhendent le site. C'est un processus "permanent" de prise de décision.

4.2 Méta-critères choisis

La qualité informationnelle du contenu nécessite des mesures pour quantifier cette information. Cette mesure peut être qualitative [KIRK02] et/ou quantitative [SMIT03], [CHU-R02]. Nous avons effectué une évaluation des sites fédérateurs, en se basant sur les méta-critères suivants, selon l'approche "usagers" :

- *Accès et Ergonomie*,
- *Crédibilité, Contenu et Services*,
- *Bibliothéconomie et Science de l'information*.

4.2.1 Méta-critère 1 : Accès et Ergonomie

Méta-critères	*"Critères" choisis*	*Eléments*
Accès	Accès au site WEB	*aspects techniques : Indexation humaine, rapidité, stabilité, adresse, etc.*
Ergonomie	Organisation logique du site et navigation	*plan du site, arborescence, navigation dans le site, fonction d'aide, etc.*
	Habillage, ergonomie, lisibilité	*Esthétique, graphismes, lisibilité, multimédia, interactivité, couleurs, etc.*

L'indexation des sites Web est une opération humaine, intellectuelle contrairement à l'indexation automatique fondée plutôt sur le contenu textuel. Contrairement à l'indexation des robots qui donne des résultats qui ne correspondent pas à la requête, l'indexation manuelle est déterminée par une réflexion sur le contenu du document.

L'aide à l'internaute n'apparaît pas comme le point fort des sites évalués, seule la moitié des sites présente un menu d'aide. Il ne suffit pas de mettre de l'information à disposition, en effet, il faut un plan du site qui permet aux clients de bien naviguer et de se retrouver dans cette masse informationnelle. La page d'accueil est le plus souvent bien structurée et bien documentée car elle constitue la porte d'accès au site.

Les meilleurs sites évalués d'après nos méta-critères (accès et ergonomie) sont : *Bubl, Math Guide, Omni-Biome et EEVL*.

4.2.2 Méta-critère 2 : Crédibilité, Contenu et Services

Méta-critères	**"Critères" choisis**	***Eléments***
Crédibilité	Informations sur le site Informations sur l'auteur Informations sur l'organisme	titre, mission, public, date de création, etc coordonnées, autorité, réputation, crédibilité, etc. statut du site (institutionnel, bibliothèque, etc.),
	Validité du contenu	copyright, droits, réputation, crédibilité, Statut du certificateur (validation)
Contenu	Contenu informationnel	quantité, profondeur, mention des sources, liens, qualité d'écriture
	Fiabilité du contenu	comparaison avec d'autres sources, application de la redondance, sites miroirs
Services	signatures	Comité éditorial (réputation, crédibilité)
	Outils, pertinence	Accès personnalisé, aide à la recherche, liens vers d'autres sites, classement des réponses, Moteur interne, date de création, fréquence de mise à jour, maintien du site, URL etc.
	Format des fichiers proposés	formats de fichiers (HTML, DOC, RTF, PDF, etc.)

Le méta-critère basé sur la crédibilité a été retrouvé dans les 31 sites évalués. Ils sont tous bien mentionnés. Cette observation a permis de mettre en évidence le caractère "professionnel" des sites "passerelles" par rapport aux sites classiques avec des dates de publication et de mise à jour. Une mention spéciale concerne la signature et les services rendus aux publics, très souvent par des liens fiables. 15 des sites évalués possèdent un comité de rédaction et proposent des "*moteurs de recherches internes*" pour faciliter la recherche intra-site.

Les sites spécialisés dans le domaine biomédical ou de la médecine constituent les meilleurs sites validés par des spécialistes de la santé.

Les meilleurs sites évalués d'après nos méta-critères sont : ***CISMeF***, ***OMNI***, *Librarian's Index to Internet, Sosig, EEVL*.

4.2.3 Méta-critère 3 : Bibliothéconomie et Science de l'information

Méta-critères	*"Critères" choisis*	*Eléments*
Bibliothéconomie	Référencement : Organisation analytique de l'information	Meta-données (Dublin Core)
	Classifications normalisées utilisées	Dewey, CDU, LCSH, Thématique, Index, Thesaurus, etc.
	Qualité des liens hypertextuels	Organisation, commentaires, clarté, pertinence (liens morts/vivants), etc.

Le référencement des sites a pour but de faciliter l'accès au document et à son contenu sur les moteurs de recherche grâce à la description du site par l'utilisation de méta-données. 32 % des sites évalués (soit 10/31) utilisent les méta-données, pour les sites fédérateurs, c'est un gage de qualité.

L'indexation est une opération analytique très utilisée par les professionnels de la bibliothéconomie et de la documentation, elle permet de recenser l'information d'un site grâce à des mots clés déterminés de manière rationnelle et ciblée. La démarche qualité générée induite par les sites fédérateurs se retrouve dans la qualité des liens hypertextuels (plus de liens vivants que morts).

La classification des documents est un élément primordial dans le cas des sites fédérateurs. L'offre de liens vers d'autres sites ne peut être proposée sans un minimum d'ordre afin de s'y retrouver.

La majorité des sites fédérateurs proposent au moins une classification. Qu'elle soit systématique (Dewey, CDU, etc.), alphabétique ou thématique, cette démarche classificatoire est un signe de l'investissement des professionnels de l'information sur l'analyse des sites. Ces derniers reçoivent une information faible d'où l'idée de « valeur ajoutée ». En plus, elle est complétée par des liens vers d'autres sites fédérateurs. Cette option est donc très pratique pour un usager qui verra ainsi ses demandes bien traitées.

Les meilleurs sites évalués d'après nos méta-critères sont : *BUBL, Sosig et CISMeF*.

5. Organisation des connaissances dans des sites « SFQC » : Classification et indexation

Dans ce domaine les méthodes de représentation des sites Web et de leurs contenus sélectionnés au sein des sites fédérateurs sont multiples. Nous pouvons énumérer les classifications, l'utilisation de thésaurus, de listes de vedettes-matières, les groupements thématiques ou encore des méta-données. Nous qualifierons tous les sites respectant les critères bibliothéconomiques et la validation de contenu comme des SFQC (Sites Fédérateurs de Qualité Contrôlée). Nous donnons une évaluation de ces sites à qualité ajoutée pour la certification de leurs contenus.

- **Les classifications arborescentes thématiques**

  La navigation par les liens hypertextes est souvent utilisée pour visualiser et consulter le contenu des documents Web. La représentation des sujets de façon hiérarchique favorise la navigation à travers les structures d'information complexes (cf. BUBL). Les informations sont regroupées en catégories organisées sous forme de table des matières. Les informations organisées sous forme *arborescente* agissent en tant que treillis. L'activation d'un point de l'arborescence permet d'en visualiser le point précédent ou suivant permettant à l'utilisateur d'élargir ses choix. Les systèmes de classifications traditionnels sont mis en œuvre dans plusieurs sites pour classer les ressources électroniques.

  Les systèmes présents sont la Classification Décimale Universelle (CDU), la classification Dewey (DDC), la classification de la bibliothèque du Congrès (LCC), les vedettes-matières de la Bibliothèque du Congrès, la Medical Subject Headings (MeSH) et la National Library of Medecine (NLM). Ainsi, BUBL a adopté la classification Dewey pour l'organisation des ressources électroniques sous forme de catégories. Dans chaque cas, un indice est affecté à chaque ressource électronique à partir de la table de classification utilisée. Cependant, les ressources ne sont pas cataloguées intégralement et l'accès par sujet reste limité. Les utilisateurs sont contraints de limiter leurs objectifs de recherche à chaque niveau de la hiérarchie correspondant au niveau de la classification en question. Quant à la classification Dewey, les utilisateurs soulignent sa flexibilité d'où son adoption chez BUBL, ADAM et Bized. Nous retrouvons la classification du MeSH chez OMNI.

- **Thesaurus et listes de vedettes-matières**

  L'usage de ces vocabulaires contrôlés qu'ils soient intégrés ou générés de façon automatique, vise à l'organisation des contenus des différentes ressources tout en évitant les problèmes de polysémie, de synonymie ou autres. Ils permettent une meilleure précision de la recherche. Plusieurs projets ont choisi d'intégrer des thésaurus ou des listes de vedettes-matières

existantes, construits de façon traditionnelle afin d'indexer les documents disponibles sur l'Internet et de faciliter leur accessibilité. Parmi ces outils, nous avons identifié, entre autres, la Library of Congress Subject Headings (LCSH) qui est utilisée pour indexer et donner un accès par sujets aux ressources électroniques universitaires cataloguées dans *INFOMINE*. *ADAM* utilise le Art & Architecture Thesaurus, *EEVL* utilise Engineering Information Inc's El Thesaurus, *OMNI* utilise MESH (Medical Subject Heading), *SOSIG*[2] utilise le thesaurus *HASSET* (Humanities and Social Science Electronic Thesaurus) basé sur le thesaurus de l'UNESCO. Selon leurs domaines de prédilection les sites font usage également du Macrothésaurus de l'OCDE spécialisé en sciences économiques, de Thesaurus ERIC pour l'éducation, d'INFODATA spécialisé en sciences de l'information… L'intégration de ces outils rend possible une recherche et un accès par sujets aux diverses ressources du Web. Ils agissent en tant que liens et remplissent une fonction semblable à celle des groupements thématiques permettant ainsi de répartir les ressources électroniques par catégories.

- **Les classifications par spécialité/domaine**

  Le site OMNI (*Organising Medical Networked Information*) est un portail destiné à la communauté des chercheurs et des enseignants universitaires dans le domaine médical. Il leur facilite l'accès à une information de niveau de qualité bien validée sur les aspects cliniques en bio-médical, et de recherche en santé. Il a été créé comme un catalogue des ressources d'informations disponibles. Une fois les ressources filtrées indexées et décrites, elles sont classées à l'aide de la classification de la *National Library of Medecine* (NLM) utilisée au Royaume Uni et le MeSH. D'autres classifications numériques ou alphanumériques spécialisées par domaine ont été intégrées pour organiser diverses collections : *Danish Veterinary and Agricultural Classification pour NOVAGate, Mathematical Classification thémaique pour le MathGuide (Mathguide), etc*.

- **Indexation des documents à l'aide des méta-données**

  L'usage des méta-données apporte des solutions quant à la description des sites et à leur organisation et une meilleure efficacité dans le repérage de l'information, tout en offrant une alternative au format habituel de catalogage : le format MARC. Les méta-données consistent en des éléments qui servent à codifier le contenu d'une page Web en précisant le titre, les mots-clés, les sujets, le créateur et d'autres éléments quant à la présentation du document. Ainsi, certains sites utilisent des standards de description des données tels Dublin Core (*Geosource, MathGuide, Adam, Infomine*…). Le projet propose une méta-notice divisée en quinze éléments : *le titre,*

[2] http://www.ukoln.ac.uk/metadata/DESIRE/classification/

l'auteur, la description, l'éditeur ou l'hôte, la date, le type, la source, la langue, la couverture, les contributions, le sujet ou les mots-clés, le format... Un autre projet britannique, qui concernant les sites comme : OMNI, SOSIG ou RDN, de description des documents par les méta données est ROADS (*Resource Organisation and Discovery in Subject-Based Services*). Il se propose de chercher comment fournir un accès par sujets. Il permet la capture de notices contenant la description des diverses ressources électroniques. L'enjeu des méta-données est de permettre l'échange de références. La démarche de ROADS va jusqu'à essayer d'impliquer les concepteurs de sites en leur faisant intégrer des méta données.

Le projet le plus avancé est *the Australian Literature Gateway* (Austlist). Les attributs sont stockés dans une base de type relationnel, les données XML sont suffisamment riches pour générer des notices au format **MARC** ou des fichiers HTML pourvus de méta-données **Dublin Core (DC)**. Les conclusions de cette réalisation sont encourageantes. Il ressort qu'il a été possible de forger des outils de conversion de données relativement efficaces.

- **Indexation à l'aide d'un moteur "maison"**

  Certains sites font usage de moteurs de recherche "maison" pour optimiser la recherche locale dans leur base indexée. Ainsi, les bibliothécaires-responsables de ces sites paramètrent ces outils de façon à ce qu'ils permettent de poser une véritable équation de recherche selon des critères booléens. Ces outils permettent de personnaliser la recherche et en autorisent le tri des résultats. On retrouve les fonctionnalités de base des Systèmes de Gestion de Bibliothèques (SGB) comme le regroupement par thème, le classement alphabétique, par mots-clés ou par domaine, en nombre limité, dans un ordre de pertinence tout en évitant les doublons.

  Des moteurs permettent l'indexation des champs : *mots-clés, résumés, date, auteur, URL*. Certains vont jusqu'à indexer les données non textuelles comme le moteur utilisé sur ADAM. Ces possibilités de recherche sont décuplées par le recours aux moteurs de recherche limités donnant la possibilité d'un recensement strict sur un sujet particulier et selon des critères précis.

  L'inconvénient des sites organisés de cette manière est l'absence de précision lors de la recherche en mots-clés libres, d'où l'idée appliquée dans d'autres projets de faire appel à des outils traditionnels comme les vocabulaires contrôlés.

6. Qualité des sites fédérateurs (SFQC)

Par rapport aux moteurs de recherche, mais aussi aux annuaires, l'utilisation de sites fédérateurs procure plusieurs avantages issus principalement de la qualité de leurs outils documentaires. Ils présentent les qualités suivantes :

- **Les sites fédérateurs donnent une valeur ajoutée à l'information**

  Par leurs capacités de description des sites, de validation des contenus et catégorisation des ressources, ces portails facilitent la recherche d'information. L'utilisateur perd beaucoup moins de temps à identifier l'adresse exacte qui le mènera à l'information visée. Les contenus étant décrits, il progresse vers les *liens théoriquement* vérifiés et mis à jour, évitant ainsi la perte de temps liée à la désorientation. Les avantages listés ici sont également ceux que l'on retrouve quand on compare les moteurs de recherche aux annuaires. L'atout essentiel des sites fédérateurs est leur gestion par des professionnels de l'information. L'utilisation d'outils documentaires standardisés permet non seulement une recherche plus efficace, mais également l'accès aux documents pertinents. L'intervention d'un spécialiste est la caractéristique principale de ces sites.

- **Limitation du bruit dans les réponses par rapport aux robots**

  Les moteurs de recherche classiques souffrent d'un excès de réponses pour la plupart des interrogations. C'est d'ailleurs le principal défaut porté par les usagers du web. A l'inverse un site, analysé par un spécialiste, aura beaucoup plus de mal à obtenir une surévaluation.

- **Outils efficaces des requêtes exploratoires sur un sujet**

  Face à un nouveau sujet d'étude ou d'intérêt, le web offre des ressources innombrables et, jusque-là, anarchiques. Les projets de sites fédérateurs ont pour ambition de filtrer et de canaliser des ressources de référence permettant une entrée rapide dans un domaine de connaissance. Un utilisateur ne perd ainsi plus de temps à éliminer les sources non-pertinentes en début de recherche.

- **Veille documentaire**

  Le responsable de sites fédérateurs accomplit une grande partie de ce travail afin de maintenir un niveau efficace attractif de contenu. Un utilisateur a ainsi accès à de nouvelles ressources actualisées très rapidement et validées par les spécialistes du domaine.

- **Fiabilité et gain de temps**

  On peut donc dégager deux avantages liés aux logiques des sites fédérateurs. Le premier est lié à la sélection validée des sites référencés qui

évite ainsi les erreurs d'orientation permanentes survenant lors d'une exploration du web à l'index. Le deuxième est celui du gain de temps non négligeable. Tout le travail de repérage et de sélection de sites, qui est un travail déterminant, se retrouve offert au plus grand nombre. On peut considérer ce type de démarche comme un moyen pour faciliter l'accès à des informations fiables.

- **L'utilisation des classifications profondes**

  Elle peut poser problème pour l'orientation de l'usager. La plupart des sites fédérateurs observent, en général, la règle de « trois clics », soit en dissociant la structure de la classification de sa visualisation sur l'écran, soit en limitant la profondeur des indices utilisés, comme par exemple le fait BUBL. Le contre-exemple vient de Renardus où la combinaison d'un nouvel écran à chaque nouvelle catégorie (le tout en langue anglaise) contribue à une désorientation relativement rapide de l'usager novice.

La plupart des problèmes ci-dessus concernent également les vocabulaires d'indexation. Des listes de termes « faits maison » sont souvent rencontrées. De même, peu nombreux sont les sites qui utilisent les thesaurus multilingues (à l'inverse de NovaGate) même s'ils existent gratuitement en ligne et couvrent les domaines concernés. On ne peut que souligner ici le coût de la création d'un thesaurus et l'abondance de divers vocabulaires d'indexation, notamment en anglais … mais aussi multilingues.

7. Conclusion

On peut dire que la description des sites est très différente d'un site fédérateur à l'autre. Elle peut être très simple comportant des liens hypertextuels ou URL ou un vrai portail d'informations. Le système de classification doit être pertinent par rapport au domaine concerné. La Dewey ou la CDU pourront satisfaire des sites fédérateurs généralistes, alors que pour un site spécialisé en médecine, le MeSH est l'outil employé (incontournable) par les professionnels. La profondeur de l'arbre de la classification doit être adaptée au sujet. Plus un site sera généraliste plus le nombre de niveaux devra être important, sous peine d'inefficacité. A l'inverse, un site spécialisé nécessitera moins de niveaux, les sites à recenser étant moins nombreux. La logique de l'accès aux informations est issue de l'organisation logique des connaissances. En effet, le classement des sites est fait de façon adaptée au domaine analysé. Le résultat issu de cette bonne connaissance de ce domaine permettra de retrouver rapidement une information non-bruitée et/ou sans silence excessif. Cette organisation des connaissances peut réduire la surabondance d'information. L'étendu des niveaux doit également permettre une orientation rapide. Le modèle le plus répandu est une classification décimale (ou proche), mais là encore ce critère doit être adapté au domaine. Une majorité des sites fédérateurs

propose un moteur de recherche intra-site pour permettre une recherche rapide. Enfin, les réductions des classifications profondes et du langage «maison», qui désorientent l'usager, seraient souhaitables.

Par rapport aux moteurs de recherche, mais aussi aux annuaires, l'évaluation des sites fédérateurs a montré plusieurs avantages issus principalement de la qualité de leurs outils documentaires comme la *valeur ajoutée à l'information, la limitation du bruit dans les réponses par rapport aux robots classiques, les requêtes exploratoires sur un sujet, la fiabilité de l'information et enfin, le gain de temps.*

8. Références bibliographiques

[AUER98] Auer Nicole., ''Bibliography on Evaluating Internet Resources'', Emergency Librarian [en ligne], May/June 1998, 25(5), [19 Novembre 2004] Disponible à l'adresse suivante : http://www.lib.vt.edu/research/libinst/evalbiblio.html

[BRAN02] Brandt D. Scott. ''Evaluating Information on the Internet'', [19 Novembre2004]. Disponible à l'adresse suivante : http://www.thorplus.lib.purdue.edu/~techman/evaluate.htm

[CAEL03] Caelen I, Eglin V, Hollard S. « Analyse de documents par oculométrie »' in Gaussier E., Stefanini M.H., eds, « Recherche intelligente d'informations », Hermès, 2003.

[CAEL03] Caelen J., Eglin V., Hollard S., Meillon B. « Mouvements oculaires et évaluation de documents électroniques », CIDE.6, 6ème colloque sur le document électronique, Caen, 24-26 novembre 2003.

[CHU-R02] CHU-Rouen, « Centrale Santé, Netscoring : critères de qualité de l'information de santé sur internent », (versions professionnelles et grand public), http://www.netscoring.com/,http://www.chu-rouen.fr/dsii/publi/critqualv2.html

[EDWA98] Edwards J., ''The good, the bad and the useless: evaluating Internet resources. Ariadne'', juillet 1998, vol 16. [19 Novembre 2004] Disponible à l'adresse suivante : http://www.ariadne.ac.uk/issue16/digital/

[ENGL02] Engle M., ''The seven steps of the research process'', [7 juin 2004] Disponible à l'adresse suivante : http://www.library.cornell.edu/okuref/research/skill.htm

[HARR03] Harris R., ''Evaluating internent research sources'', Disponible à l'adresse suivante : [7 Novembre 2004]. Disponible à l'adresse suivante : http://www.sccu.edu/faculty/R_Harris/evaluate8it.htm

[HINC03] Hinchliffe L.J., ''Resource selection and information evaluation'', [7 Novembre 2004] Disponible à l'adresse suivante : http://www.alexia.lis.uiuc.edu/~janicke/Evaluate.html

[KIRK02] Kirk E., ''Evaluating information found on the Internet'', [7 Novembre 2004] Disponible à l'adresse suivante : http://milton.mse.jhu.edu:8001/research/education/net.html

[LARO03] Larouk O., ''Des grilles pour optimiser /guider la conception du point de vue « concepteur » : Le cas des sites WEB de bibliothèques'', Rapport ENSSIB, 2003, http://isdn.enssib.fr/archives/bilanaxe3.html

[SMIT03] Smith A., ‘‘Criteria for evaluation of internent information resources’’, [19 Novembre2004] Disponible à l’adresse suivante : http://www.vuw.ac.nz/~ agsmith/evaln/evaln.htm

[TILL02] Tillman H., ‘‘Evaluating quality on the net’’, [19 Novembre 2004] Disponible à l’adresse suivante : http://www.hopetillman.com/findqual.html

Annexe : Evaluation des méta-données dans les Sites Fédérateurs de Qualité Contrôlée

Cette partie présente une observation de l'indexation des SFQC. Des méta-données sont encapsulées dans les S*ubject Gateways* ? Nous avons dans un premier temps voulu vérifier que les métadonnées de type 1, c'est-à-dire les meta-tags, utilisées dans les pages d'accueil des sites fédérateurs eux-mêmes. On peut constater que beaucoup sites utilisent les champs bibliographiques Dublin Core (**DC**) pour l'indexation des documents électroniques.

Archivage des méta-données dans le code-source des sites (SFQC) testé le 12 juillet 2004

	HTML	Méta-données : **Dublin Core**		**URL**
Sites Fédérateurs		*12.07.04*	*15.03.05*	
1. Adam	4.0	**10**	**10**	http://adam.ac.uk/
2. Anglistik guide	4.0	**-**	**-**	http://www.AnglistikGuide.de
3. Agrigate	4.0	**-**	**-**	http://www.agrigate.edu.au
4. Argus clearinghouse	4.0	**-**	**3**	http://www.clearinghouse.net/
5. AVEL	4.0	**-**	**19**	http://avel.library.uq.edu.au/
6. Biz/ed	4.0	**10**	**10**	http://www.bized.ac.uk/
7. BNF	4.0	**10**	**10**	http:/ /www.bnf.fr/pages/liens
8. Bubl Link	4.0	**-**	**-**	http://bubl.ac.uk
9. Chemdex	4.0	**-**	**-**	http://www.chemdex.org
10. Cismef	4.0	**13**	**13**	http://www.chu-rouen.fr/cismef/
11. Dainet	4.0	**-**	**-**	http://www.dainet.de/eccdb/vitis/
12. Digital librarian	4.0	**-**	**-**	http://www.digital-librarian.com
13. Dmoz	4.0	**-**	**-**	http://www.dmoz.org
14. Dutchess	4.0	**9**	**18**	http://www.kb.nl/dutchess/
15. Eevl	4.0	**10**	**10**	http://www.eevl.ac.uk
16. Finnish V-library	4.0	**10**	**10**	http://www.jyu.fi/library/
17. Gem	4.0	**-**	**-**	http://www.thegateway.org
18. Geo-guide	4.0	**-**	**-**	http://www.Geo-Guide.de
19. Geo-source	4.0	**10**	**10**	http://www.library.uu.nl/geosource/
20. History guide	4.0	**-**	**6**	http://www.historyguide.de/
21. Infomine	4.0	**10**	**2**	http://infomine.ucr.edu/
22. Library System	4.0	**10**	**2**	http://www.howard.edu/library/
23. Internet Public Lib.	4.0	**-**	**-**	http://www.ipl.org/
24. Librarian II	4.0	**-**	**-**	http://www.lii.org
25. Math-guide	4.0	**10**	**0**	http://www.mathguide.de
26. NISS	4.0		**2**	http://www.niss.ac.uk/
27. Novagate	4.0	**13**	**0**	http://novagate.nova-university.org
28. OMNI	4.0	**-**	**-**	http://www.omni.ac.uk
29. Pinakes	4.0	**-**	**-**	http://www.hw.ac.uk/libwww/irn/pinakes/pinakes.html
30. RDN	4.0	**-**	**3**	http://www.rdn.ac.uk
31. Sosig	4.0	**9**	**9**	http://sosig.ac.uk

Session 5

Formatage et multicodage

GetAMsg, une librairie pour le traitement de messages avec variantes et leur localisation

Christian Boitet[1], Hung Vo-Trung[2]

[1] *GETA-CLIPS-IMAG – Université Joseph Fourier*
BP 53, 385 rue de la Bibliothèque, 38041 Grenoble Cedex 9 - France
Christian.Boitet@imag.fr

[2] *GETA-CLIPS-IMAG – Institut National Polytechnique de Grenoble*
BP 53, 385 rue de la Bibliothèque, 38041 Grenoble Cedex 9 - France
Hung.Vo-Trung@imag.fr

Résumé :

Dans les programmes d'ordinateur, le texte entourant les variables dans les messages varie, ou plutôt devrait varier en fonction de certaines caractéristiques, dépendant des langues, et des valeurs des variables passées lors de l'appel. "$n file(s)" a 2 variantes en anglais, et 3 en russe et en arabe, mais pas dans les mêmes cas. Un autre problème est que certaines variables globales, comme le niveau de politesse (tu/vous en français) peuvent aussi créer des variantes. Les outils comme gettext et catgets sont très utiles pour faciliter la multilinguïsation des programmes, mais traitent seulement des messages sans variantes. Dans notre solution, on représente les messages avec variantes non pas comme des formats avec variables, mais comme des automates d'états finis "contrôlés", ou *automates de messages*, dans une syntaxe très simple. La première version de GetAMsg, implémentée en C, est utilisable en C/C++, Perl, Pascal, et Java. Dans un programme, un appel à GetAMsg rend un format sans variables. Pour traduire un message M d'une langue L1 dans une langue L2, on génère automatiquement (en L1) une instance de M pour chaque variante a priori possible de M en L2, on traduit ces instances, qui sont des messages sans variables, et on compacte l'ensemble des variantes obtenues en un automate de messages pour L2.

Mots-clés : multilinguïsation et localisation de logiciels, automates de messages, GetAMsg.

1. Introduction

Dans les programmes d'ordinateur, la plupart des messages contiennent des variables. Cela provoque souvent des variations "linguistiques" à cause d'accords en nombre, en genre, à cause ou du niveau politesse, de la position des variables, etc. On utilise souvent des écritures "compactées" comme "$n fichier(s) supprimé(s)", pour suggérer au lecteur de lire "1 fichier supprimé" ou "2 fichiers supprimés", etc. Nous dirons que ce message a deux *variantes*, les deux formats usuels "$n fichier supprimé" et "$n fichier(s) supprimé(s)". Mais ce genre d'écriture devient vite très lourd. On voudrait donc produire directement la bonne variante, en fonction de la valeur de la variable n (1, 2, 3…).

En contexte multilingue, le problème est que le nombre des variantes et les conditions de choix dépendent des langues. Par exemple, il y a en russe trois formes du pluriel pour les masculins : les variantes du mot архив (archive) sont архив, архива, архивов : архив s'il finit par 1 et pas 11, архива s'il finit par 2, 3, 4, et pas 12, 13, 14, et sinon архивов. En arabe, il y en a 3 aussi, mais pas dans les mêmes conditions. Dans les langues asiatiques (et d'autres), il n'y a qu'une forme. Pour les ordinaux, le russe et le français ont 2 variantes, et l'anglais 4 (-st, -nd, -rd, -th), etc.

C'est un problème important, car les grands éditeurs de logiciels comme Adobe, HP, IBM, Microsoft, etc., localisent dans plus de 30 langues, et ont l'intention de faire plus. Il y a aussi de gros projets de multilinguïsation de logiciels créés en "[code] source ouvert", comme le projet Mozilla et bien d'autres hébergés par le site www.sourceforge.org.

Les systèmes tels que gettext et catgets permettent de localiser des fichiers de messages, mais ils ne permettent pas de traiter les variantes linguistiques dans les messages, parce que le "modèle de message" qu'ils utilisent impose de traiter les variantes (sauf le pluriel) au niveau du code source, qui dépend alors des langues traitées, ce qu'on veut justement éviter. Par exemple, en C, on écrirait en français :

```
if (n<2)        printf("%d fichier supprimé", n);
else            printf("%d fichiers supprimés", n);
```

et en russe :

```
if (n<2) || ((n%10 =1) && (n>20))
        printf(У%d архив извлекалФ, n);
else if (n>=2 && n<=4) || (n>20 && n%10>2 && n%10<=4)
        printf(У%d архива извлекалиФ, n);
else
        printf(У%d архивов извлекалиФ, n);
```

Dans la suite, nous montrons d'abord que les techniques actuelles bien connues (catgets, gettext, QT, java), calculent un format classique, en fonction de la langue souhaitée, dans lequel le programme appelant instancie les variables par leurs valeurs, et que cela qui empêche de traiter les variantes. Nous proposons ensuite une solution, dans laquelle le gestionnaire de messages instancie lui-même les variables, et renvoie un message "final", i.e. un format sans variables. Pour cela, nous modélisons les variantes d'un message (toujours en nombre fini), dans une langue donnée, en les factorisant dans un automate d'états finis "contrôlé". La librairie getamsg implémente cette idée. Enfin, nous montrons comment traduire un message (avec variantes) M écrit en L1 dans une autre langue L2 : nous générons (en L1) une instance de chaque variante possible de M, *par rapport à L2*, faisons traduire ces instances en L2 de façon classique (il s'agit de formats comme ceux qu'on traduit tous les jours depuis 50 ans), et compactons l'ensemble résultant (des variantes en L2) en un automate de message pour L2.

2. Etat de l'art

Pour comprendre l'approche proposée, il est utile d'étudier brièvement des outils antérieurs destinés à faciliter la localisation des messages.

2.1 Gettext (GNU)

Gettext de GNU (http://www.gnu.org/software/gettext/) est peut-être l'outil le plus connu et le plus utilisé pour localiser les logiciels. Il repose sur un ensemble de conventions sur la façon d'organiser les catalogues de messages et de les gérer, et propose des outils pour la "récupération" des messages à traduire, c'est à dire pour réorganiser le code d'une façon propre, externaliser les messages dans des ressources, et gérer les catalogues et les messages. Une fois le code réorganisé et les messages externalisés, il suffit de changer la valeur de la langue locale pour que les messages changent de langue. Dans du code C, une instruction comme "printf("%d fichier supprimé", n);" est remplacée par "printf(gettext("%d fichier supprimé"), n);". Le format "%d fichier supprimé" sert de clé d'accès dans les catalogues de messages.

Le cas des variantes liées au pluriel étant fréquent, la fonction ngettext a été introduite pour permettre de choisir une variante, initialement parmi deux, puis parmi plusieurs[1].

Voici un exemple avec 3 types de variantes[2] :

[1] http://www.delorie.com/gnu/docs/glibc/libc_135.html

[2] Adapté du code de Taoru TAKAHASHI, voir http://lists.gnu.org/archive/html/bug-gnubg/2004-07/msg00022.html

```
// nplurals est le nombre de variantes pour le pluriel.
// plural varie de 0 à nplurals-1, et msgstr[plural] sera sélectionné.
// Catalogue en_US: une variante pour le pluriel
Plural-Forms: nplurals=2; plural=n != 1;          # en-tête du catalogue
msgid "Match to %d point"                         # clé pour le cas n=1
msgid_plural "Match to %d points"                 # clé pour le cas n≠1
msgstr[0] "Match to %d point"                     # n == 1
msgstr[1] "Match to %d points"                    # n == 0, 2, 3 ...
// Catalogue jp: pas de variantes pour le pluriel en japonais
Plural-Forms: nplurals=1; plural= 0;              // plural=0:
msgid "Match to %d point"
msgid_plural "Match to %d points"
msgstr[0] "%d POINTO MATTI"                       # tout n
// Catalogues russe, tchèque, etc. : 3 variantes pour le pluriel
Plural-Forms: nplurals=3;                         # condition OK ici, fausse sur le site GNU
      plural=(n%10==1 && n%100!=11 ? 0 :
          n%10>=2 && n%10<=4 && (n%100<10 || n%100>=20) ? 1 : 2);
msgid "Match to %d point"
msgid_plural "Match to %d points"
msgstr[0] "..."                                   # n == 1, 21, 31, 41, ..., 91, 121 …
msgstr[1] "..."                                   # n == 2, 3, 4, 22, 23, 24, 32 ...
msgstr[2] "..."                                   # n == 5, …, 20, 25, …, 30 …
/* Appel (le même pour toutes les langues) */
printf (ngettext ("Match to %d point", "Match to %d points", n), n);
```

Gettext offre aussi des fonctions utiles pour la localisation, comme le formatage des valeurs de date, d'heure, des nombres, etc. en fonction de la langue courante. Enfin, gettext peut être utilisé dans de nombreux langages de programmation, comme C/C++, Shell Script, Python, Lisp, etc. [Drepper & al. 2002].

En ce qui concerne le traitement des variantes, on vient de voir qu'il est limité à celui d'une seule variable (entière), et qu'il n'y a aucune factorisation des variantes.

2.2 Catgets (Sun)

Catgets de Sun a été développé avant gettext, qui a repris bon nombre d'idées de catgets. La différence principale est qu'il faut nommer les messages (le format utilisé dans l'appel sert de commentaire et de cvaleur par défaut) et que la structure des catalogues est plus complexe, puisque chacun peut contenir des "ensembles", eux aussi nommés [IBM 2004, Wheeler 2003]. Voici un exemple.

```
nl_catd catd = catopen ("catalog_msg", 0); // ouverture du catalogue de messages
printf (catgets(catd, 1, 23, "original string"), n, m); // setno=1, msgid=MSG23
Catgets utilise donc un accès en 3 étapes : catalog => set => msgid ID => format.
```

Imposer de numéroter les messages est beaucoup plus contraignant que de demander de leur associer des identificateurs, ou d'utiliser directement la valeur par défaut comme clé. D'autre part, rien n'est prévu pour les variantes.

2.3 AG5MSG (Ariane-G5)

Ariane-G5, produit par le GETA (la version dont nous parlons date de 1992), est un générateur de systèmes de traduction automatique (TA) basé sur cinq LSPL (Langages Spécialisés pour la Programmation Linguistique). Le traitement des messages est intéressant, même s'il a été développé en 1985-91, car il permet de traiter n'importe quel type de variation, au contraire de tous les outils actuels.

Dans l'environnement AG5MSG [Guillaume 2002], un message est représenté par un réseau sans boucle. Voici par exemple le message : "C'est fini : $NbRegles règle(s) a(ont) été compilée(s), elle(s) contien(nen)t $NbOctets octet(s)," :

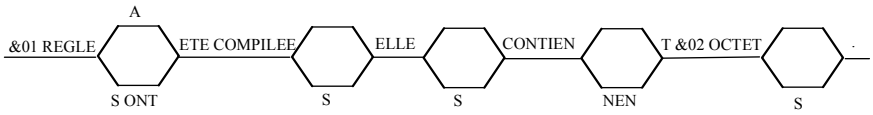


Figure 1 : Automate de messages en AG5MSG

Dans le fichier de messages, on écrit ceci (le ' indique un métacaractère, comme ':) :

```
MESSAGE1 : 'C''EST FINI :'
MESSAGE2 : '&01 REGLE'(A':S ONT') ETE COMPILEE'(':S'), ELLE'(':S')
  CONTIEN'(':NEN')T &02 OCTET'(':S').'
```

AG5MSG y substituera les valeurs de &01 et &02 en caractères. Dans l'exemple, es valeurs correspondant au singulier pour &01, &02 se trouvent dans les branches supérieures : il y en a 4 pour &01, et 1 pour &02. On a 4 variantes qui correspondent aux 4 listes de choix (1 1 1 1 1), (1 1 1 1 2), (2 2 2 2 1) et (2 2 2 2 2).

Pour appeler ce message avec &01=1 et &02=100, le programme appelant passe en arguments à AG5MSG ces valeurs (1 et 100) et la liste des choix (ici : 1 1 1 1 2).

Cela permet de traiter tous les types de variantes en monolingue, mais pas en multilingue s'il y a deux langues trop différentes [Boitet 1982, 1990], puisque les listes de choix sont calculées dans le programme appelant.

2.4 Java

Pour localiser des messages, Java propose la technique suivante. D'abord, comme toujours, on "sort les messages du code" et on les range dans des ressources (packages), avec une entrée pour chaque message. Ensuite, on modifie le code pour calculer le format utilisé en fonction de l'identificateur du message et de la langue (locale). Comme avec gettext, on peut utiliser un ChoiceFormat (ainsi que MessageFormat) pour sélectionner une variante ou une autre, mais de façon plus lourde. Voici par exemple un catalogue appelé SampleResources.properties :

```
none = I have no cars in the garage.
one = I just bought one car.
many = I won the lottery and bought {0} cars.
```

Le code source pour imprimer est par exemple :

```
public class ChoiceFormatExample {
  public static void main (String args[]) {
        ResourceBundle resourceBundle =
        ResourceBundle.getBundle("SampleResources", Locale.US);
        double limits[] = {0,1, 2}; // gammes pour les formats
        String none = resourceBundle.getString("none");
        String one = resourceBundle.getString("one");
        String many = resourceBundle.getString("many");
        String formats[] = {none, one, many}; // formats de la gamme
     ChoiceFormat cf = new ChoiceFormat(limits, formats);
        MessageFormat mf = new MessageFormat("{0}");  // variable
        mf.setFormats(new Format[]{cf});
        for (int i=1; i<5; i++) {
       Object messageArgs[] = {new Integer(i)};
       System.out.println("i: "+i+"/"+mf.format(messageArgs));
}}}
```

Quand on exécute ce programme, on obtient le résultat suivant :

```
i: 0 / I have no cars in the garage.
i: 1 / I just bought one car.
i: 2 / I won the lottery and bought 2 cars.
i: 3 / I won the lottery and bought 3 cars.
i: 4 / I won the lottery and bought 4 cars.
```

Pour localiser dans une autre langue, on traduit que des chaînes dans le fichier SampleResources.properties et on change le lieu (locale) :

```
ResourceBundle resourceBundle =
      ResourceBundle.getBundle("SampleResources", Locale.FR);
```

Cette technique ne permet apparemment pas de gérer les messages avec variantes en contexte multilingue. Par exemple, on ne peut pas utiliser le code source ci-dessus pour le russe (3 variantes avec des conditions différentes 1, 2-4, 5-10…).

3. GetAMsg

3.1 Modélisation des messages avec variables et variantes

Les automates de messages sont des automates finis "contrôlés". Pour chaque langue, l'automate peut-être différent [Boitet 2003]. Par exemple, le message "$n fichier(s) supprimé(s)", peut être représenté en anglais, français et russe par les automates de messages suivants, ici sans factorisation à l'intérieur des mots :

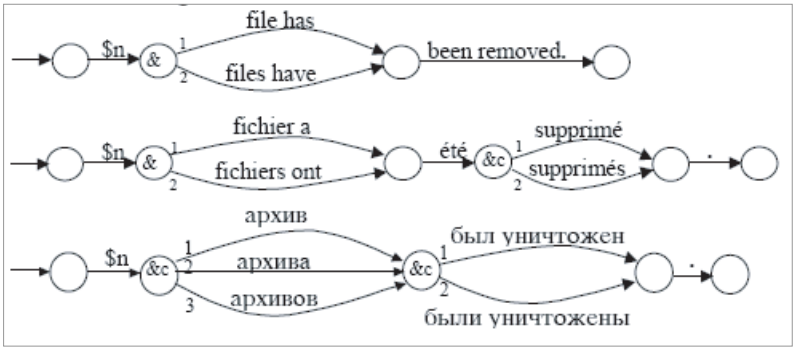


Figure 2 : Automates sans factorisation à l'intérieur des mots

À partir d'un automate de messages, on peut automatiquement construire les variantes qu'il représente, en énumérant des instances de tous les cas possibles.

Soit par exemple le message "n fichier(s) supprimé(s)" en anglais, français et russe. Pour le français et l'anglais, la variable &c est définie par la condition $n<=1. D'où les deux variantes (formats usuels) correspondant aux deux cas possibles :

&c	Condition	Anglais	Français
1	$n <= 1	$n file has been removed	$n fichier a été supprimé
2	Sinon	$n files have been removed	$n fichiers ont été supprimés

Tableau 1 : Variantes pour l'anglais et le français

Pour le russe, la variable &c est définie par les trois conditions ($n<2) || (($n%10 =1) && ($n>20)), ($n>=2 && $n<=4) || ($n>20 && $n%10>=2 && $n%10≤4), et les autres possibilités. D'où les 3 variantes :

&c	Conditions	Russe
1	($n<2) \|\| (($n%10 =1) && ($n>20))	$n архив извлекал
2	($n>=2 && $n<=4) \|\| ($n>20 && $n%10>=2 && $n%10≤4)	$n архива извлекали
3	sinon	$n архивов извлекали

Tableau 2 : Variantes pour le russe

Pour modéliser cela, on définit des variables de contrôle entières &ci commençant toujours à 1, puis s'incrémentant de 1, donc à valeurs dans [1..n_i+1], où n_i est le nombre d'arcs sortant du nœud portant &ci. Si &ci = n_i+1, la trajectoire correspondante s'arrête à ce nœud. À chaque combinaison de valeurs des variables de contrôle correspond une unique "trajectoire" dans l'automate, donc une seule variante du message, dans laquelle les variables seront instanciées [Boitet 2005].

3.2 Spécification de GetAMsg

Il faut d'abord dissocier complètement le code source du programme et les messages. Pour cela, on peut utiliser les outils d'extraction liés à gettext. Un fichier de messages (comme FicMes1_fre.adm) contient des messages, ou plus précisément des automates de messages, pour une langue donnée (ici le français). Plusieurs programmes peuvent partager un même fichier de messages, et un programme peut utiliser plusieurs fichiers de messages, un seul pouvant être actif à chaque instant.

Dans la version actuelle, chaque message est nommé par un identificateur alphanumérique local au fichier de messages. Si cela apparaît indispensable, nous permettrons dans la version suivante d'utiliser une chaîne quelconque, par exemple

une variante ou une chaîne mnémonique dans la première langue d'implémentation, mais cela nous forcera à modifier un peu la syntaxe d'appel de getamsg, en mettant le type des variables du message dans un second argument [Vo-Trung 2004].

La syntaxe générale d'un (automate de) messages est la suivante :

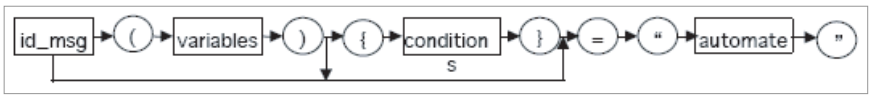


Figure 3 : Syntaxe générale d'un message

Par exemple, le message "$nbfichier fichier(s) a (ont) été compilé(s)." peut être représenté par l'automate suivant, avec &c=1 si $nbfichier<2, &c=2 sinon :

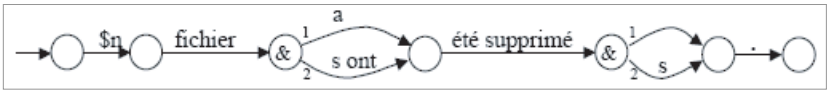


Figure 4 : Exemple d'un automate de messages en GetAMsg

Si son identificateur est M1, on l'écrira par exemple (avec %3d pour le formatage) :

```
M1($n_fichier){&c:$n_fichier<2}= "$n_fichier%3d fichier[&c: a | s ont] été
compilé[&c :  | s]."
```

Dans un programme, nous utilisons la fonction getamsg() avec en paramètres d'abord le nom et les types des variables du messages, puis les valeurs de ces variables. Par exemple, l'appel getamsg("M1 n_", nf) produit un format sans variables dans le langage hôte (C/C++, Pascal, Perl, etc.).

Pour imprimer, on écrira en C : 'printf(getamsg("M1 n_", nf));' et en Pascal 'write(getamsg("M1 n_", nf));'.

Voici les types de variables prédéfinis. Nous utilisons des préfixes comme "$o_" pour éviter les déclarations explicites, cependant toujours possibles (par "nom : type"). Dans la version actuelle, ce tableau est "en dur". Dans la version suivante, il sera modifiable par les utilisateurs, ce qui permettra d'ajouter des types.

Type de variable	Préfixe	Valeur	Description
t_gm_langue	$l_	enum (fra, eng…)	angue (ISO-639-2)
t_gm_cardinal	$n_	int	nombre cardinal
t_gm_ordinal	$o_	int	nombre ordinal
t_gm_string	$s_	string	chaîne de caractères
t_gm_reel	$r_	real	nombre réel
t_gm_genre	$g_	enum (M, F, N)	genre
t_gm_politesse	$p_	[1..4]	politesse
t_gm_titre	$t_	[1..3]	type de titre
t_gm_age	$a_	[0..120]	âge
t_gm_calendrier	$c_	enum (JUL, GRE...)	calendrier
t_gm_heure	$h_	[0..24]	heure

Tableau 3 : Types des variables utilisées dans les automates de messages

Il y a un certain nombre de variables globales prédéfinies, avec des valeurs par défaut dépendant du système d'exploitation et de l'utilisateur courant.

Nom	Description
Variables globales pour le système	
$s_system_name	Nom du système
$s_system_nickname	Surnom du système
$l_system_lang	Langue du système
$p_system_politeness	Niveau de politesse
$c_system_calendar	Calendrier du système
$s_catalogue	Nom du catalogue
Variables globales pour l'utilisateur	
$s_user_nickname	Surnom de l'utilisateur
$s_user_family_name	Nom de famille
$s_user_first_name	Prénom
$s_user_second_name	Deuxième nom
$n_user_age	Age de l'utilisateur
$t_user_civility	Préfixe civil de l'utilisateur
$g_user_gender	Sexe de l'utilisateur
Variables globales pour getamsg	
$l_gm_language	Langue d'interface de GetAMsg
$p_gm_politeness	Niveau de politesse (de GetAMsg)
$t_gm_title	Type de titre civil utilisé

Tableau 4 : Variables globales de GetAMsg

Pour abréger les conditions permettant de déterminer les valeurs des variables de contrôle dans les automates de messages, nous fixons pour chaque type de variable et chaque langue les noms et valeurs (booléennes) des "cas de figure".

Var	**Conditions**			
	Anglais	**Français**	**Russe**	**…**
$g_	HE : $g_=m SHE : $g_=f IT : $g_=n	IL: $g_=m ELLE : $g_=f ÇA : $g_=n	ОН : $g_=m ОНА : $g_=f ЭТО : $g_=n	
$p_	FA : $p_≤ 2 POLI : $p_≥ 3	FA : $p_≤ 1 POLI : $p_≥ 2	БЛИЗКИЙ : $p_=1 ВЕЖЛИВЫЙ : $p_≥ 2	
$t_	NONE: $t_=0 MR_MRS: $t_=1 FUNCTION: $t_=2 RANK: $t_=3	SANS: $t_=0 M_MME: $t_=1 FONCTION: $t_=2 GRADE: $t_=3	БЕЗ : $t_=0 (sans titre) ГОСП : $t_=1 (MmeE) ЧИН: $t_=2 (DirecteurE) ФУНКЦИЯ: $t_=3 (DrE)	
$n_	SINGULAR: $n_≤1 PLURAL: $n_≥2	SINGULIER: $n_≤1 PLURIEL: $n_≤2	ОДИН : ($n_<2) ‖ (($n_%10 ==1) && ($n_>20)) ДВА : (2≤$n_ && $n_≤4) ‖ ($n_>20 && 2≤$n_%10&& $n_%10)≤4) ПЯТЬ : ($n_>0)&&(5≤$n_%10) &&(($n_%10=0)‖ (%≤$n_%10 && $n_%10≤9))	
$a_	BABY: $a_≤5 CHILD: 5<$a_≤13 TEEN: 13≤$a_<20 ADULT: 20≤$a_≤60 SENIOR:60<$a_≤80 OLD: $a_>80	BEBE: $a_≤5 ENFANT: 5<$a_≤13 ADO: 13≤$a_<20 ADULTE: 20≤$a_≤60 SENIOR: 60<$a_≤80 VIEUX: $a_>80	РЕБЁНОК : $a_≤5 ДИТЯ : 5<$a_≤13 ЮНОША: 13≤$a_<20 ВЗРОСЛЫЙ: 20≤$a_≤60 ПОЖИЛОЙ : 60<$a_≤80 СТАРИК : $a_>80	
$c_	JULIAN : $c_=1 GREGORIAN: $c_=2 JAPANESE : $c_=3 BUDDHIST : $c_=4 ISLAM : $c_=5 …	JULIEN : $c_=1 GREGORIEN: $c_=2 JAPONAIS: $c_=3 BOUDDHISTE:$c_=4 ISLAM : $c_=5 …	ДЖУЛИАН : $t_=1 ГРИГОРИАНСКИЙ : $t_=2 ЯПОНСКИЙ : $t_=3 БУДДИСТ : $t_=4 ИСЛАМ : $t_=5 …	
$h_	AM : $h_ ≤ 12 PM : 12 < $h_≤ 24	AM : $h_ ≤ 12 PM : 12 < $h_≤ 2	БЕЗ (hh)	
$o_	FIRST:$o_%10=1 SECOND:$o_%10=2 THIRD:$o_%10=3 OTHERWISE: sinon	PRE: $o_%10=1 AUTRES: sinon	ОЙ : ($o_%10=2 || $o_%10=6 || $o_%10=7 || $o_%10=8) && ($o_<12 || o_<20) ЫЙ : sinon	

Tableau 5 : Cas de figure prédéfinis et nommés

Ces définitions sont rangées dans un tableau contenu dans un fichier texte qui sera modifiable dans la version suivante. L'utilisateur pourra ainsi changer les noms utilisés dans les conditions, changer le nombre de cas, changer les conditions, et rajouter les informations correspondant à de nouvelles langues.

Par exemple, nous utilisons (M, F, N) pour masculin, féminin, neutre, mais on pourra remplacer ces identificateurs par (IL, ELLE, ÇA) en français, par (HE, SHE, IT) en anglais, ou (ОН, ОНА, ЭТО) en russe. On pourra aussi changer des conditions, par exemple, distinguer 3 cas pour les cardinaux en français (0, 1, et >1).

3.3 Utilisation pratique

GetAMsg se présente comme une librairie C. Grâce aux outils javah, h2pas, et h2xs, qui compilent une entête C vers une interface pour Java, Pascal, et Perl, nous pouvons l'utiliser aussi dans ces langages. C'est la technique utilisée par gettext, et il existe des outils analogues pour beaucoup d'autres langages, mais nous ne les avons pas encore testés sur GetAMsg.

Pour utiliser cette fonction, les programmeurs doivent observer quelques règles.

- D'abord, on extrait les messages et on les met dans des fichiers, sous forme d'automates de messages. En parallèle, on modifie bien sûr le code source.
- Ensuite, on utilise la fonction textmsg() pour compiler chaque fichier de message, par exemple, FicMes1_fre.adm vers FicMes1_fre.amg.
- On complète les initialisations du ou des programmes concernés par le choix de la langue de travail de départ et de son codage associé :

```
#include <locale.h>
char *setlocale (int category, char *locale )
```

On initialise aussi si nécessaire les variables globales.

La modification du code d'appel des messages consiste à appeler la fonction getamsg(), dont le prototype est :

```
char *getamsg(char *id_msg, expression-1, expression-2, ..., expression-n)
```

id_msg : identificateur du message et des types des variables locales, donc des expressions passées en paramètres. Par exemple,"001 n_ o_" si l'identificateur du message est 001 et s'il a deux paramètres, le premier de type cardinal et le second de type ordinal (n-ème). Ces déclarations sont nécessaires pour utiliser une liste de paramètres de longueur non connue à l'avance.

expression-i : valeur de la variable locale n° i.

Supposons par exemple que nous avions au départ, dans un programme en C/C++ :

```
printf("Bonjour !");
printf("%d fichier(s) supprimé(s)", nbfichier);
```

Nous aurons créé 3 fichiers de messages pour le français, l'anglais et le russe :

- Fichier de messages en français :

```
msg1= "Bonjour !"
msg2($n_fichier){&c : ($n_fichier<2)} =
"$n_fichier fichier[&c :|s] supprimé[&c :|s]."
```

- En anglais :

```
msg1= "Hello !"
msg2($n_fichier){&c : ($n_fichier<2)} =
"$n_fichier%d file[&c :|s] removed."
```

- En russe :

```
msg1= У Здравствуйте !Ф
msg2($n_fichier){&c: ОДИН($n_fichier), ДВА($n_fichier)}=
   У$n_fichier архив[&c:|а|ов] извлекал[&c:|и|и].Ф
```

Dans le code source unique, nous écrirons simplement :

```
printf(getamsg("msg1"));
        /* msg1 : id de message, sans variable        */
printf(getamsg("msg2 n_", nbfichier));
        /* msg2 : id de message, nbfichier : variable */
```

4. Localisation

Pour localiser un programme dans d'autres langues, il faut traduire les messages associés. *Mais on ne peut pas traduire directement un automate de messages d'une langue dans une autre* ! En effet, aucun traducteur automatique ne traite ce genre de formalisme, et aucun traducteur humain non plus : pour traduire un automate de messages, il faudrait imaginer toutes les variantes possibles en langue cible, et les factoriser en créant les contrôles avec leurs conditions, ce qui est déjà une sorte de programmation. Nous proposons la méthode suivante, en trois étapes :

- On génère d'abord (en langue source) une liste d'instances correspondant aux "cas de figure" possibles dans la langue cible, en gardant les noms de variables (ex: $n_night=2).
- Ensuite, on traduit ces instances, et on obtient un ensemble de messages instanciés, avec pour chacun la condition associée au cas de figure concerné.
- Enfin, on supprime les valeurs effectives utilisées, et on factorise l'ensemble des formats ainsi obtenus pour obtenir un automate de messages en langue cible le plus compact possible.

Nous avons proposé et implémenté deux algorithmes, l'un pour la génération des cas de messages et l'autre pour la factorisation [Vo-Trung, 2004, Boitet, 2005].

Voici un exemple pour illustrer ces étapes lors de la traduction d'un automate de messages d'anglais en russe. Le message anglais est "You reserved $n_rooms room(s) for $n_nights night(s)" et est représenté par l'automate de messages suivant :

```
M003($n_rooms,$n_nights){[&c1:$n_rooms==1],[&c2:$n_nights==1]}
="You reserved $n_rooms [&c1:room|rooms] for $n_nights[&c2:night|nights]."
```

Nous produisons en anglais une instance de chaque "cas" possible en russe, en laissant le nom des variables pour en retrouver la trace après traduction :

No	"Cas de figure" du russe en anglais	Condition
1	You reserved $n_r=1 room for $n_n=1 night.	$n_r=1,$n_n=1, $p_p=1
2	You reserved $n_r=1 room for $n_n=2 nights.	$n_r=1,$n_n=2, $p_p=1
3	You reserved $n_r=1 room for $n_n=5 nights.	$n_r=1,$n_n=5, $p_p=1
4	You reserved $n_r=2 rooms for $n_n=1 night.	$n_r=2,$n_n=1, $p_p=1
5	You reserved $n_r=2 rooms for $n_n=2 nights.	$n_r=2,$n_n=2, $p_p=1
6	You reserved $n_r=2 rooms for $n_n=5 nights.	$n_r=2,$n_n=5, $p_p=1
7	You reserved $n_r=5 rooms for $n_n=1 night.	$n_r=5,$n_n=1, $p_p=1
8	You reserved $n_r=5 rooms for $n_n=2 nights.	$n_r=5,$n_n=2, $p_p=1
9	You reserved $n_r=5 rooms for $n_n=5 nights.	$n_r=5,$n_n=5, $p_p=1
10	You reserved $n_r=1 room for $n_n=1 night.	$n_r=1,$n_n=1, $p_p=2
11	You reserved $n_r=1 room for $n_n=2 nights.	$n_r=1,$n_n=2, $p_p=2
12	You reserved $n_r=1 room for $n_n=5 nights.	$n_r=1,$n_n=5, $p_p=2
13	You reserved $n_r=2 rooms for $n_n=1 night.	$n_r=2,$n_n=1, $p_p=2
14	You reserved $n_r=2 rooms for $n_n=2 nights.	$n_r=2,$n_n=2, $p_p=2
15	You reserved $n_r=2 rooms for $n_n=5 nights.	$n_r=2,$n_n=5, $p_p=2
16	You reserved $n_r=5 rooms for $n_n=1 night.	$n_r=5,$n_n=1, $p_p=2
17	You reserved $n_r=5 rooms for $n_n=2 nights.	$n_r=5,$n_n=2, $p_p=2
18	You reserved $n_r=5 rooms for $n_n=5 nights.	$n_r=5,$n_n=5, $p_p=2

Ici, nous avons supposé que l'utilisateur est masculin. Si nous avions fait intervenir une variable de sexe, nous aurions eu 27 cas de figure au lieu de 18.

Nous traduisons ces (instances de) cas de figure en russe et obtenons :

No	Instances des "cas de figure" traduites en russe	Condition
1	Вы резервировали $n_r=1 комнату на $n_n=1 ночь.	$n_r=1,$n_n=1, $p_p=1
2	Вы резервировали $n_r=1 комнату на $n_n=2 ночи.	$n_r=1,$n_n=2, $p_p=1
3	Вы резервировали $n_r=1 комнату на $n_n=5 ночей.	$n_r=1,$n_n=5, $p_p=1
4	Вы резервировали $n_r=2 комнаты на $n_n=1 ночь.	$n_r=2,$n_n=1, $p_p=1
5	Вы резервировали $n_r=2 комнаты на $n_n=2 ночи.	$n_r=2,$n_n=2, $p_p=1
6	Вы резервировали $n_r=2 комнаты на $n_n=5 ночей.	$n_r=2,$n_n=5, $p_p=1
7	Вы резервировали $n_r=5 комнат на $n_n=1 ночь.	$n_r=5,$n_n=1, $p_p=1
8	Вы резервировали $n_r=5 комнат на $n_n=2 ночи.	$n_r=5,$n_n=2, $p_p=1
9	Вы резервировали $n_r=5 комнат на $n_n=5 ночей.	$n_r=5,$n_n=5, $p_p=1
10	Ты резервировал $n_r=1 комнату на $n_n=1 ночь.	$n_r=1,$n_n=1, $p_p=2
11	Ты резервировал $n_r=1 комнату на $n_n=2 ночи.	$n_r=1,$n_n=2, $p_p=2
12	Ты резервировал $n_r=1 комнату на $n_n=5 ночей.	$n_r=1,$n_n=5, $p_p=2
13	Ты резервировал $n_r=2 комнаты на $n_n=1 ночь.	$n_r=2,$n_n=1, $p_p=2
14	Ты резервировал $n_r=2 комнаты на $n_n=2 ночи.	$n_r=2,$n_n=2, $p_p=2
15	Ты резервировал $n_r=2 комнаты на $n_n=5 ночей.	$n_r=2,$n_n=5, $p_p=2
16	Ты резервировал $n_r=5 комнат на $n_n=1 ночь.	$n_r=5,$n_n=1, $p_p=2
17	Ты резервировал $n_r=5 комнат на $n_n=2 ночи.	$n_r=5,$n_n=2, $p_p=2
18	Ты резервировал $n_r=5 комнат на $n_n=5 ночей.	$n_r=5,$n_n=5, $p_p=2

Enfin, nous effaçons les chaînes "=<valeur>", factorisons les formats correspondants, et obtenons l'automate de messages suivant en russe :

```
M003($n_r, $n_n, $p_p) {[&c1:$p_p==0], [&c2: ОДИН($n_r), ДВА($n_r)],
[&c3: ОДИН($n_n), ДВА($n_n)]}=
"[&c1: Ты резервировал | Вы резервировали ]$n_r комнат[&c2 :у|ы|] на
$n_n ноч[&c3: ь|и|ей]."
```

Ici, il y a deux problèmes à discuter. Le premier concerne la traduction des cas en langue cible. Il est clair que le nombre de variantes possibles peut être assez grand, et on peut se demander si un traducteur accepterait de traduire 18 (ou 27 si on ajoute le féminin) instances du même message, comme ci-dessus. En fait, cela ne pose pas vraiment de problème : s'il s'agit d'un traducteur automatique, le surcoût est minime, et s'il s'agit d'un traducteur humain, l'usage d'une mémoire de traduction rendra également le surcoût minime.

Le deuxième problème est l'apparition des "cas" en langue cible quand ils n'existent pas en langue source. Par exemple, on n'utilise que le pronom "you" en anglais mais, en français, et en russe, il faut distinguer entre deux cas ("vous/tu" et Вы/Ты), selon le niveau de politesse. Si nous utilisons un traducteur automatique pour obtenir le premier jet (il faudra bien sûr le réviser ensuite), nous pouvons utiliser certains de ses paramètres. Par exemple, Systran a deux paramètres pour contrôler le traitement du "you" en anglais-français : traduire par "tu" ou "vous", et mettre au masculin ou au singulier.

S'il s'agit d'un traducteur humain, nous lui envoyons la chaîne à traduire, précédée de l'indication de la valeur souhaitée pour chaque variable n'apparaissant pas dans ladite chaîne. Ce genre de convention ne semble pas être difficile à accepter.

Par exemple, pour obtenir une forme directe (familière) au masculin, on enverra :

```
[DIRECT($p_system_politeness), MASC($g_user_gender)]
= "You reserved $n_r=5 rooms for $n_n=2 nights."
```

5. Conclusion et perspectives

Nous avons présenté une méthode qui permet d'organiser et de traiter des messages dans les logiciels multilingues. Notre solution, implémentée dans l'outil GetAMsg, consiste à utiliser des automates d'états finis contrôlés pour représenter et gérer des messages avec variables et variantes. Cette méthode permet de stocker et de récupérer un message unique correspondant à une "trajectoire" dans un automate de message, pour toute combinaison des valeurs de paramètres du message.

L'avantage de cette méthode par rapport aux autres (gettext, catgets) est la gestion des variantes du message dans plusieurs contextes et plusieurs langues différentes. L'outil GetAMsg est un ensemble de programmes écrits en C. Il est pour l'instant interfacé avec C/C++, Java et Perl.

Dans le futur, nous nous proposons d'améliorer GetAMsg en rendant paramétrables par les développeurs les types de variables traités, ainsi que les conditions associées, pour chaque langue. Si l'usage révèle que c'est nécessaire, nous permettrons aussi d'utiliser des chaînes quelconques, par exemple des formats C, pour identifier les messages, ce qui nécessitera d'introduire une variante syntaxique de la fonction getamsg(), qu'on appellera par exemple getamsgx().

Un axe sur lequel nous avons déjà commencé à travailler est celui de la *mutualisation des traductions*, le but étant d'intégrer GetAMsg dans le mouvement actuel de la localisation en "P2P". Pour cela, nous pensons utiliser l'outil PolyphraZ, actuellement en cours de développement pour la présentation, l'édition et la traduction de corpus de phrases parallèles dans différentes langues sur le web. Une polyphrase est une structure de donnée contenant diverses propositions (humaines ou automatiques) de phrases (ou d'énoncés quelconques comme des titres), en principe dans plusieurs langues. Il nous faudra simplement introduire la notion de "polyphrase prototype", c'est-à-dire capable d'engendrer un nombre fini ou infini d'instances. Les automates de messages, ainsi que d'autres générateurs comme des automates d'états finis classiques, pourront ainsi être stockés avec leurs instances dans un même ensemble de polyphrases.

Enfin, nous avons commencé à explorer la possibilité de représenter un message avec variables par un graphe UNL [Uchida 2001, Tsai 2001] avec variables, indépendant des langues. Une fois un tel graphe obtenu (par "enconversion" à partir d'une instance représentative d'un automate de messages dans une certaine langue), il serait possible, pour chaque langue cible visée, d'instancier ce graphe par des valeurs correspondant aux cas de figure, d'obtenir les instances correspondantes (en langue cible) par "déconversion", et de les factoriser en un automate de messages, de la même façon que dans la méthode actuelle de traduction directe d'instances.

6. Références bibliographiques

[Boitet 1982] Christian Boitet : "*Le point sur ARIANE-78 début 1982*", 3 volumes, Partie 1, GETA-CHAMPOLLION, CAP SOGETI-France, avril 1982.

[Boitet, 1990] Christian Boitet : "*1980 – 90 : TAO du réviseur et TAO du traducteur*", in La TAO à Grenoble en 1990, école d'été à Lannion, www-clips.imag.fr/-geta/christian.boitet/pages_personnelles/ (in English in Proc. ROCling-90, Taipei)

[Boitet, 2002] Christian Boitet : "*A rationale for using UNL as an interlingua and more in various domains*", Proceedings "First International Workshop on UNL, other Interlanguages and their Applications", LREC2002, Las Palmas, Spain, May 2002.

[Boitet, 2003] Christian Boitet : "*Messages avec variantes, automates finis contrôlés, et multilinguisme*", document interne, GETA, laboratoire CLIPS, IMAG, février 2003.

[Boitet, 2005] Christian Boitet : "*Message Automata for Messages with Variants, and Methods for their Translation*", Proc. CICLING 2005, Mexico, Feb. 2005, Springer LNCS 3406, p. 352—371.

[Drepper *et al.* 2002] U. Drepper, J. Meyering, F. Pinard, B. Haible : "*GNU gettext tools, version 0.11.2*", Published by the Free Software Foundation, April 2002.

[Guillaume, 2002] Pierre Guillaume : "*L'interface utilisateur multilingue en Ariane-G5*", Rapport de recherche, Groupe d'Etude pour la Traduction Automatique (GETA), CLIPS, IMAG, avril 2002.

[IBM, 2003] IBM Corporation : "*International Components for Unicode (ICU) – User's Guide*", html:// oss.software.ibm.com/icu/userguide, 2003.

[Vo-Trung, 2004] Hung Vo-Trung : "*Méthodes et outils pour utilisateurs, développeurs et traducteurs de logiciels en contexte multilingue*", thèse d'informatique, Institut national polytechnique de Grenoble, déc. 2004.

[Sun, 2000] Sun Microsystems Inc : "*Building International Applications*", Sun product documentation http://docs.sun.com/db/doc/806-6663-01.

[Tsai, 2001] Wang-Ju Tsai : "*SWIIVRE - a Web Site for the Initiation, Information, Validation, Research and Experimentation on UNL (Universal Networking Language)*", First International UNL Open Conference, Suzhou, China, Nov. 2001.

[Tsai, 2004] Wang-Ju Tsai : "*La coédition langue-UNL pour partager la révision entre langues d'un document multilingue*", thèse d'informatique, Université Joseph Fourier, juillet 2004.

[Uchida, 2001] Uchida Hiroshi, "*The Universal Networking Language beyond Machine Translation*", International symposium on language in cyberspace, Sept. 2001, Seoul, South Korea.

[Wheeler, 2003] David A. Wheeler : "*Secure Programming for Linux and Linux HOWTO*", http://www.dwheeler.com/secure-programs/, March 2003.

Advanced Transformation Rules for Structured document Applications

Règles de transformation avancées pour les applications des documents structurés

Nouhad Amaneddine[1], Jean-Paul Bahsoun[2], Jean-Paul Bodeveix[2]

[1] *GREYC – UMR 6072 CNRS – Université de Caen*
14032 Caen Cedex - France
nouhad.amaneddine@info.unicaen.fr

[2] *IRIT- Université Paul Sabatier – Toulouse III*
118 route de Narbonne, 31062 Toulouse Cedex 4 - France
{bahsoun,bodeveix}@irit.fr

Résumé :

Nous proposons dans ce papier un ensemble de règles pour la transformation de documents structurés. Les règles de transformation présentées constituent le noyau d'un système de transformation de données structurées. L'implantation d'un système de transformation robuste et efficace demande un noyau consistant et durable. Le but de la définition des règles de transformation est de pouvoir construire un système qui peut assurer non seulement les transformations simples et directes, mais aussi celles qui sont plus compliquées, surtout, lorsque des éléments de type récursif peuvent apparaître dans les documents instances. Le problème n'est pas clairement résolu dans les approches de transformation récentes, pour cette raison nous développons notre modèle afin de réaliser des transformations avancées. Les documents XML qui sont conformes à une DTD bien déterminée sont les premiers candidats sur lesquels notre modèle est appliqué. Nous présentons la grammaire de l'ensemble de règles de transformation et nous montrons leurs caractéristiques en appuyant sur des exemples types. Nous concluons par les perspectives sur les axes principaux de notre futur travail.

Mots-clés : XML, DTD, règles de transformation, documents structurés, modèle de transformation.

Abstract:

We propose in this paper a set of transformation rules for structured documents. The rule set we present is the kernel part of a structured document transformation system. Building a robust and efficient transformation system requires vigorous and proficient rules specification. The goal that lies under our rules definition is to build an advanced transformation system that performs not only simple and direct transformations but also complicated ones, especially, where the structure of the document in question allows a recursive appearance of the elements in the document instances. The problem is not clearly resolved in the recent transformation approaches so we build our model in order to realize advanced transformations. XML documents are the first candidates on which we apply our rules set. In particular, we work on XML documents that conform to a predefined document type definition (DTD). We present the grammar of the transformation rules and reveal their main characteristics. We show the rule language adequacy on sample cases. We conclude by some perspectives about the main axis for the future work.

Keywords: Transformation rules, Structured documents, Model transformation, XML, DTD.

1. Introduction

As the Internet applications grow rapidly in the last few years, document engineering is becoming a more and more interesting area in the computer sciences domain. The way the information has been presented in HTML format does not ensure a robust separation between the document content and its structure. Therefore, the world wide web community had launched the extensible markup language XML and it became the first candidate to replace the current formats of the Web documents. Beside XML, many other sub-languages have been defined: CML [CML04] [MR97], MathML [RSD03], VML [BDB04], WML [WML02] and so. Each one is dedicated for a specific domain and respects the rules imposed for their XML parent language. Moreover, the need of managing XML document has increased so the computer society has generated a large number of tools to manage XML documents. Exchanging documents between application has become a need. Therefore, the W3C proposed the XSLT as a transformation language for XML documents [W3C02]. This language has a considerable computation power but it requires detailed and tedious programming to accomplish difficult structured documents transformations. As alternative we construct a transformation system at the structure level that's capable to perform not only simple and direct XML transformation but also complicated one with expressive, descriptive and user friendly transformation rules.

The core part of the transformation system is the transformation rules. The main question that may be asked is what type of information should be selected and

to what destination it will be converted. The process starts by making the associations between nodes in the input document and their corresponding nodes in the output document. Then, filtering may be applied on selected elements. Finally, chosen information should find their way to be placed in the output document. We present in this paper the kernel part of a new structured document transformation system, called **TransM** [ABB04] [Ama04]. The core part we present consists of a set of transformation rules. The handled information are valid XML documents that conform to a predefined document type definition. Both the input and the output document are considered to be valid. The rest of this paper is organized as follows: in section two we present an overview on structured document transformation. Section three shows the **TransM** core, which is the transformation rule. We present the grammar of the rule specification, their semantic, some important sample cases and an example on witch we apply the transformation rules. In section four we show our rule advantages and we present -in a brief way- in section five the general algorithm that hold structured document transformation. Section six present some related work in the literature and we conclude by presenting our perspective and by proposing several future work.

2. Structured document transformation

The main existing components in the transformation procedure are the input element and the output one. Data are represented as structured documents and transformations are made from one structured document to another one. The used method is the bridge that brings us from the source to the destination. The transformation method can be a direct transformation or a specification of the transformation in an abstract level. Direct transformation consists of choosing elements from the source document and transferring those elements to their places in the output document. This procedure is applied without any use of the general structure of both input and output documents. This type of methods works on the lowest level that touches directly the information stored in the document. Those methods can be applied locally on the document and they can not be reused elsewhere [KP96]. On the other hand, the specification of the transformation works on a higher level in order to define how the rules of the transformation process can be used to carry out the transformation. It is a more abstract level than the immediate transformation process and it can be reused each time the user needs to apply some transformations with respect to the pre-defined rules.

The recent research results such as [Wil03] and [PFT02] are focusing on model transformation. We remark that they differ in the literature between document transformations and model transformations, while the two research areas can be placed under one single title which is the structured document transformation. The QVT approach [QVT03] is one of those proposals that works on model transformation. For us, while we can represent any object to be transformed in a structural form, the resolution of the transformation will count on the general axis

defined by the transformation of documents. In the MOF model of OMG community [OMG03], they define four levels for representing models, starting from the lowest one which is the document itself upward the more abstract one which is the meta-meta model. Figure 1 shows the architecture of this classification for models levels.

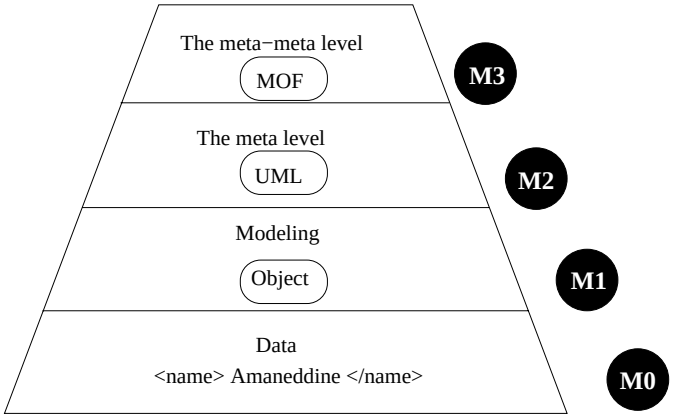


Figure 1: Models Organization

As this general architecture for defining models and their specifications is used in [PFT02] when applying model to model transformations, we should place the source and destination models in their corresponding levels in this architecture. Moreover, the transformation will depend on the position of these two models, so it can be a transformation in the same level or form a model in a particular level to another one in another level. For us, the same problem is imposed but in another form. We consider that even if we specify in the M3 level and we need the transformation to be applied in the M2 level, or the specification are made on the M2 level and the transformation on M1 level, or, the same process is applied between the M1 level and the M0 level, the main procedure persists the same, specification on a higher level and transformation in a lowest one. In the classical document transformation this classification does not exist. We have only the abstract level which is the specification or the grammar of the structure of the document and the second level which is the document itself [KS95].

Another existing difference between model transformation and structured document transformation is that in model transformation models are represented as graphs. Since we can represent graphs in XML by using the ID of the xml element, and we can refer to that element through its ID reference *xmi:idref*, therefore, model transformation can be accomplished using our transformation rules applied on the XML representation of the transformed model. In document transformation, the problem starts when the structure of both input and output documents are quite different and when we need to make changes on the contents of the document in order to produce the required output. In this case, we found either the used method

can hold those transformations but they are too complicated to be understood and applied, and their use does not cover all the complicated situation we could face in such transformation, or, those methods are unable to support advanced transformations. In this case, the eventual question that could be asked is what is the interest of providing such a method and how relevant is its application, especially for a delicate transformation process.

Therefore, an advanced transformation system should be build in order to accomplish all kind of transformations. This system can not achieve its goals if it is not based on robust and advanced transformation rules. **TransM** is a transformation system build on its essential part which is the transformation rules we show in the next section.

3. The kernel model: transformation rules

3.1 Transformation process

The process consists of transforming XML documents conforming to a predefined document type definition to another XML document conforming to a different document type definition. It starts by making the association relations between the input and the output structures as shown in figure 2. Those associations are chosen from a set of pre-defined possible associations. Then the transformations are specified by a set of rules. Those rules are used to hold the required transformations in the concrete level. The program code generated after the transformation rules aims to transform document conforming to the input structure to another document conforming to the output structure.

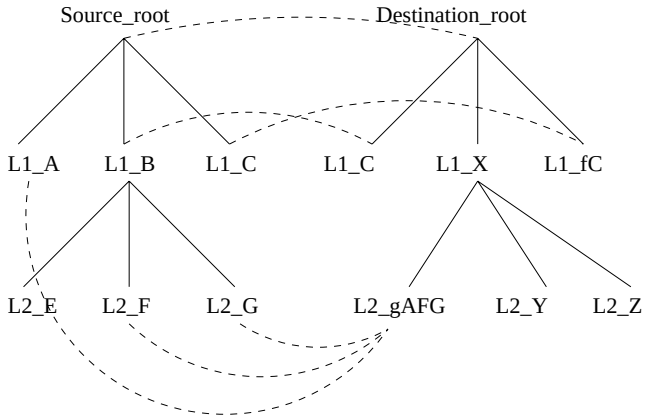


Figure 2: Elements Associations

In this figure we show the input structure and the output structure with a tree forms. Associations are represented by the dashed lines that link nodes from the input structure to nodes of the output one. The association may be simple, which means that it represents a direct copy of the node content, or it may be more complicated and expressed as a function of many variables depending on more than one node in the input structure.

We will not explore the details of all the parts of our model. We present the core part of the transformation system we are working on, which is the specification of the transformation rules that are made according to the chosen user associations.

3.2 Rules definition and their application on sample cases

Our method consists of a compromise solution between relative-path based method and absolute-path based ones. We use an additional strategy that allows us to specify a set of rules for a particular type of document, depending on the type of the transformation. Also, we can control the selection and the writing of the element in any level of deepness.

The transformation is expressed as a set of: ***filter*** → ***expression*** production rules. In order to simplify the understanding of the production rules definition we show by a list of cases how to express specific transformations. Each case represents a general transformation process. Eventually, we can not cover all the possible transformations by some sample cases, we show what we consider interesting for the most frequent transformation process. The rules capacity is much more richer than what the sample cases present.

Filters:

It consists of applying the filters in the left hand side of the rule on the input document. The filter mentions recursively the tag, the tag with a list of attributes (between parenthesis) or the tag with a list of attributes followed by a list of children.

Filters and selection constraints:

This kind of constraints is usually applied to the left hand side of the transformation rule, the filter can be completed by a constraint (between two brackets), here frequently-used sample cases:

1. Constraints on children:

Element_A[(Element_B/Element_C)+, Element_D] → *Target_element*
This rule selects Element_A elements that have a non empty sequence of Element_B children with Element_C child, followed by Element_D child.

2. Constraint on a sequence of elements and on element's occurrence:

Element_D[A+, B*, C] → *Target_element*
This rule identifies Element_D elements that have a list of one or more A child followed by zero or more B child that are followed by a C child.

3. Constraint with absolute path:

Element_A[/C/D@attr] → *Target_element*
This rule selects Element_A elements if the root of the structure that contain A has a grand child named D. Moreover, this grand child possesses an attribute called attr.

4. Left, right sibling and order:

Element_A[.-1=C \&.+1=D] → *Target_element*
This rule identifies Element_A elements that have a C element as left sibling and a D element as right sibling.

Production constraints:

In the selection constraint sample cases, we have considered the selection criteria in the left hand side of the production rule. This part of the rule wraps a filtering conditions or constraints. We represented the right hand side as a simple *Target_element*. The following examples reveal how to manipulate output data in the right hand side of the production rule. The result of the production will eventually depend on the selected information. The left hand side will write down data selected with respect to the left hand side filtering.

Filters, constraints and variable assigning:

Element_A(@att1="value1", @att2=x') (Element_B(@att1=y') → Element_C (@at1="value2", @at2=x'+y')
This rule selects Elements_A elements having two attributes, att1 with value "value1" and att2, and one child, Element_B with attribute att1. The values of att1 and att2 are assigned to the variables x' and y'. The right hand side creates the Element_C element with two attribute, the first one is called at1 and values "value2" the second one is called at2 and values x'+y'.

The selection we use enables us to navigate over all the structure of the document. Form the current node in the hierarchical document, we can select a node using criteria depending on any node of the document. The context dependent filtering mechanism can access ancestors, children sibling or the root of the document. The algorithm processes in a descendent way from the root down to the leaves of the document. Linking the two neighbour elements in the selection part is not enforced, but it can be done according to the needs imposed by the type of the transformation we want. In other words, we can refer to a node directly by the node name or by presenting the parent followed by a slash then the required node, this if we need to add constraints concerning this particular element. Filters and expression

start always by the main pattern which is the element identifier. The other parts of both left and right hand side of the production rule consist of condition criteria and selection tools as same as the writing paradigm in the right hand side.

The most important characteristic of this method is its capability to support recursively structured document, where a grand child can have the same structure as one of its ancestors. When applying the rules, and for the selection of particular element in the source document, it should be clear on what level we are performing the transformation in the hierarchical composition of the document and where to apply the conditions. In the same way in the output document, the rules should specify the exact level where the information will be placed. The general example of a recursive document structure is to consider both the input and the output structures as recursive.

3.3 Example

Let us consider the example of transformation problem given by W3C's XML Query Working Group as Use Case "PARTS" [CFM04]: the input is a flat list of *part* elements, each of which has values for *partid* and *name* attributes. Each *part* may or may not be a component of a larger *part*, indicated by the value of the *partof* attribute. The transformation is to convert the flat representation into an explicit hierarchic representation, based on *part of* attributes. The document type definition DTD, source.dtd of the input XML document is defined as follows:

```
<!ELEMENT partlist (part *)>
<!ELEMENT part (EMPTY)>
<!ATTLIST part
        partid CDATA #REQUIRED
        partof CDATA #IMPLIED
        name CDATA #REQUIRED
```

The DTD of the output XML, is defined as follows:

```
<!ELEMENT parttree (part *)>
<!ELEMENT part (part *)>
<!ATTLIST part
        partid CDATA #REQUIRED
        name CDATA #REQUIRED
```

The corresponding specification rules that carry out this transformation is defined in two rules. We need to apply two rules in a recursive way, the first is called **Rcopy** and the second is called **copy**.

```
Rule Rcopy {
        Partlist → partree;
        part(!@partof) → Copy;
    }
```

```
Rule Copy {
        ANY → part(@partid=./@partid,@name=./@name)
            ((/partlist/part[@partof=./@partid]).Copy);
    }
```

The application of the rules **Rcopy** and **Copy** on following instance:

```
<?xml version="1.0"?>
<!DOCTYPE partlist SYSTEM "source.dtd">
<partlist>
        <part partid="0" name="car"/>
        <part partid="1" partof="0" name="engine"/>
        <part partid="2" partof="0" name="door"/>
        <part partid="3" partof="1" name="piston"/>
        <part partid="4" partof="2" name="window"/>
        <part partid="5" partof="2" name="lock"/>
        <part partid="10" name="skateboard"/>
        <part partid="11" partof="10" name="board"/>
        <part partid="12" partof="10" name="wheel"/>
        <part partid="20" name="canoe"/>
</partlist>
```

returns the following results:

```
<?xml version="1.0"?>
<!DOCTYPE partree SYSTEM "destination.dtd">
<partree>
        <part partid="0" name="car">
            <part partid="1" name="engine">
                <part partid="3" name="piston"/>
            </part>
            <part partid="2" name="door">
                <part partid="4" name="window"/>
                <part partid="5" name="lock"/>
            </part>
        </part>
        <part partid="10" name="skateboard">
                <part partid="11" name="board"/>
                <part partid="12" name="wheel"/>
        </part>
        <part partid="20" name="canoe"/>
</partree>
```

Because of the possible recursive appearance of the elements in the input and the output structures, it is difficult to express the associations only graphically. We need textual forms to convey such associations. Graphical associations are convenient to be utilized in the instance level. In this case, rules could not be used as specifications and they should be applied locally. This is what happens in add-hoc transformation application where transformations are local ones. This is not the case in our model where we define rules on the abstract level in order to be reused everywhere we need the transformations and for any document instance.

4. Advantages of the transformation rules

In order to build a valuable and reusable transformation rules, the developer should respect important requirements in the method he uses. According to many experiences in structured document transformation, and given the problems that persist in the existing transformation approaches, we respected in **TransM** a set of requirements that may be constructive when developing a new transformation for structured documents. Some of those requirements are applied in some research results [QVT03] some others are not. In some cases they are presented indirectly as conditions to be applied as much as the designer can. **TransM** has applied all those requirements especially those related to the design of the abstract model. When the main concepts of new transformation approaches is well built, the other components of the model will be conceived with respect to the main axis fixed in the abstract level. Consequently, the results of the transformation processes achieved by such a model will not be far from the expectations of the constructor. The **TransM** model applies the following requirements:

- Supporting transformations: transformation rules should support not only those simple and direct transformations but also those complicated ones, this is the case when both the input and output structures are not that close.
- Extendable transformation definitions: transformations should be built in a way that the designer might be able to add extensions if he needs future model updating.
- Modeling elements: transformation paradigms should match, filter, copy, write and manipulate elements composing the source document. It should also be capable of working on sets of elements. Moreover, multiple target elements could be defined in a single rule and different rules can provide property values for the same element.
- Handling recursive structures: transformation models should handle recursive input and output structures in any level of depth.
- Rule order dependencies: because of the recursive nature of the transformation process, a rule can apply another rule, therefore, the order of the rule application is important to accomplish the transformation.

- Expressiveness and descriptiveness: the used transformation paradigm should utilize expressive and descriptive language to hold transformations. The used syntax should be clear and user friendly, it should be expressive, easy to be used and to be understood at the same time. Moreover, the translation instructions should be succinct in order to avoid long expressions to state transformations.

5. General transformation algorithm

We present in this section the general algorithm used by the structured document transformation rules. It processes in a depth first traversal and depends on both left and right hand side of the production rules. We define the terms that we use inside the general algorithm. **dr** is the document root, **RS** stands for the rule set, **R** is a given rule, **LHS(R)** is the left hand side of the rule and **RHS(R)** is its right hand side. The comma "," represents the sequence operator, it is associative and the **"Void"** value is its neutral element that represents the empty tree. The **env** variable denotes the environment and contains a list of (variable=value) copies. One of those values is "**current**" and it defines the current node. The element is expressed by its tag followed by a list of its different attributes then a list of its children. The algorithm starts by the **Apply** function and is presented as follows:

Apply(RS, dr)
    Search for the first rule R such that LHS(R) match dr.
    Apply-rule(RS, R, dr).
    If such a rule does not exist returns:
    Apply(RS, c_1), Apply(RS, c_2), ..., Apply(RS, c_n)
    Where (c_1, c_2, ..., c_n)=children(dr).
End of **Apply**.

Apply-rule is defined as follows:

Apply-rule(RS, LHS(R) → RHS(R), A)
    let env=match LHS(R) with A
    Evaluate(RS, RHS(R), env).
End of **Apply-rule**.

The algorithm processes in a top down manner, from the root node down to the end nodes. It searches element nodes filtered by the left hand side of a rule that belongs to the rules set. The result is the juxtaposition of the values of the right hand sides. The evaluation of the right hand side may provoke the recursive application of the rules on the children of the node in question. The recursive transformation of the children of a particular node that is recognized by a rule is implicitly realized if the

right hand side of that rule does not mention the values of the children of the result node.

By example, if in the function **Evaluate**(RS, R, env) we assign the current element as follows:

let tag($@a_1$=u_1, $@a_2$=u_2, ..., $@a_m$=u_m)(c_1, c_2, ..., c_n)= current node

If we match **R** with *tag* it returns *tag*($@a_1$=u_1,..., $@a_m$=u_m)(*Apply*(*RS*, c_1)..., *Apply*(*RS*, c_n)), where (c_1, c_2, ..., c_n)=children(current) and if we match **R** with *tag*(g_1,..., g_n) it returns:

tag($@a_1$=u_1,..., $@a_m$=u_m)(*Evaluate*(*RS*, g_1, *env*), ..., *Evaluate*(*RS*, g_n, *env*))

6. Related work

Recently, different methods have been defined for carrying out document transformations. Some of them work directly on the basic level of the document which is its content. Therefore the transformation is performed on the data level. Some others specify the transformation rules and apply the transformation process with respect to the specifications [BBG01]. Other research groups are working on the same problem but with a different form, like introducing the problem as model transformation. In this case the power of UML to represent and manipulate models is used [Wil03]. On the other hand, we can differ from those textual method and graphical ones, some approaches consider all the needed steps concerning the transformation process can be performed by graphical method, in other words, by using geometric notations to express transformation procedures. Thus, they use a full graphical way to hold the transformation between different applications or different models. We present in a brief way some of the recent transformation approaches. We explain for each approach its relevance and its behaviour, as well as the framework of its application.

6.1 OMG' QVT

Models are the primary artifacts in the OMG's model driven architecture software development approach [OMG03]. MDA made a significant difference from earlier uses of modeling languages such as OMG's UML, in which the primary purpose of models was to aid understanding and communication.

In MDA, transformations play a key role, a standard syntax and execution semantics for transformation is an important enabler for an open MDA tools chain. In [QVT03] the OMG issued a revised Request for Proposal for MOF 2.0 Query, views and Transformations to address a technology part of the OMG Meta Object Facility entering to the main issues in the manipulation of the MOF models. The object management group has issued a Request for Proposal [QVT03] for Query/views/Transformation (QVT) language to exploit the Meta-Object Facility

that share common core concepts with the Unified Modeling Language (UML) [OMG01].

The transformation part of the QVT proposal is the kernel part of the MOF revised submission. Two different layers are used in order to define the transformation approach, the superstructure and the infrastructure. Concepts translation is used to convert concepts which exist in the superstructure but not in the infrastructure. The nature of the use of two different layers allows to extend the transformation model with new features. The QVT proposal provides a dedicated language in order to support transformations, this language is called MTL as model transformation language. MTL is based on pattern matching to perform transformations.

The proposed pattern language is not always the best way for expressing aspects of a particular transformation. The differentiation between relation and mapping cause a wondering confuse. The relations are not clearly the specifications of the mappings. This is obviously shown when some of the mapping rules are kept as they are in their definition in the relations part.

The second problem appears in those complicated transformations, especially when inheritance sub-models appear in the hierarchical architecture of the manipulated model. This problem is difficult to be seen when a graphical representation of the model or the structured document is shown.

6.2 Model transformation

[PFT02] used a transformation-specific language developed at France Telecom. Dedicated languages or domain-specific languages (DSL) are programming languages or executable specification languages that offer, through appropriate notations and abstractions, expressive power focused on, and usually restricted to a particular problem domain.

The use of a domain-specific language may be a solution of some problems related to the software development (reusability, productivity, maintainability). [WE03] demonstrate that dedicated languages can reduce the cost of software maintenance. Other studies [Kru9] present those languages as one of the main solutions to satisfy software reusability.

This DSL expresses how to transform a model compliant to the source meta-model to another model compliant to the destination meta model. The instructions used in the transformation rules represent the way that the entities of the source model are transformed to entities in the destination model. A rule indicates how an instance of the target metamodel is created and how its attributes and roles are assigned in the context of an instance of the source metamodel. Constraints are added to MTrans model transformation to select entities with conditions. Restrictions are based either on attributes or on roles. When a restriction is based on

an attribute the value that must have to be transformed is specified. When it is based on role, the type of the entity that must have to be transformed is specified.

One of the problems that face MTrans is the fact that it transforms only those entities that can appear in a flat meta-model. Flat meta-model is a meta model that contains only concepts, which means that it is a meta model that defines the set of terminal instances [PFT0]. By example, if we have in the source meta-model an inheritance tree structure, only leafs are interesting for transformations. Mtrans does not support transformation of recursive elements. Rules are not named so it is not possible to call a rule in order to apply it from other rules. Another disadvantage of this framework is that the used language seems to be able to do many manipulations on transformations but it is not clear how it does them.

6.3 UMLX

UMLX provides an open source tool to support the OMG's Model Driven Architecture initiative [OMG03]. It describes a primarily graphical transformation language that extends UML through the use of a transformation diagram to define how an input model is to be transformed to an output model.

UMLX uses standard UML class diagrams to define information schema and their instances. It extends the class diagram to define inter schema transformations. Four main extensions have been added to the class diagram to support model transformation [WE03]. The first one is a graphical representation of an invocation which is used when transferring schema syntax to schema syntax. Two other graphical design were added to distinguish the inputmodel from the output one. We think that even if those additional graphical tools may intend to enrich the used graphical language, it is obvious that some of them are irrelevant representations. Since the input model and the output one are clearly known especially when arrows are used in the schema that represent the two models and the transformation. Additional geometrical objects are used to clearly reveal input and output models.

A problem appears in using the UMLX transformation which is the incapability of modeling recursive instances of both source and destination models especially when the depth of the modeled entity is not known. For known depth relations we need to pass an outer context down as an inner context is explored. This can be seen in the resolution of primary key in the UMLX2RDBMS example shown in [Wil03].

According to [Wil03], there should be nothing that can not be expressed graphically that can be done textually, it just requires imagination to identify the appropriate graphical syntax. We believe that it is much more difficult to perform every transformation only graphically than expressing it textually, especially when the transformation plays in a complicated context and manipulations are needed to be performed in the source model in order to produce the target model.

Graphical notations can get cumbersome for strange complicated relationships, but seem simpler for practical simple graphs. The graphics is a less explored area, so

it takes a while to learn the new idioms, and possibly to provide the correct family of helper transformations/syntax extensions.

6.4 Other transformation approaches

Many approaches attempted to perform structured document transformation. The difficulties appear each time the transformation become complicated and advanced manipulation on the selected information are needed to be applied. Aho and Ullman use input-output paired grammars to describe a syntax-directed translation (SDT) [AU72]. Other approaches have added extensions to the SDT method to resolve more critical contexts [KP96] [Lei], some others are based on pattern matching [Ha04] [Cd04]. On the other hand, some transformation models consider as essential the backbone structure of the modelled document: they build formal definitions and they apply the transformation through a syntax that conforms to the predefined specifications. The tree automaton based approach [Mur98] and the filter based approach are such examples. The efficiency of the used methods that carry out the transformation remains a serious problem, as same as the complexity of the adopted algorithms. Some research context explored transformation process approaches in more details. They explained for each one its relevance and its behavior as well as the framework of its application. The problem remains always with non-direct transformations and when the structures of the input model and the output one are not alike. In this case, additional manipulation should be performed in the transformation process and it is not clear how to solve such transformation in those existing approaches.

7. Conclusion and future work

Prior to constructing our transformation system, we examined several approaches for specifying transformation for structured documents. None of them seemed suitable as a specification model for complex transformations. Some are too operational in natures and other can describe local transformations only. The XSLT language is suitable to transform an XML document into another one, but it is hard and error prone to manually write XSLT programs. We presented in this paper the kernel of the *TransM* system. We have shown its core part which is the transformation specifications in term of rules in more details. Our formalism allows expressing transformation rules in a more succinct and readable way.

This model can be used not only for transforming structured document, but also for model transformation. The main requirement will be the representation of the input and output meta model as DTD and their model instances as XML documents conforming to the defined DTD. The difference with the model transformation is that we are working on two levels: the abstract level and the concrete level. In the model transformation, the OMG community has proposed the UML modelling language as model representation. It intends to affront the

incompatibility of having a variety of meta models the OMG has projected a general structure for meta model integration. This organization conducted to a four level architecture: the meta-meta model level, the meta model level, the model level and the data level. In this architecture, each level preserves an instantiation relation with its superior model.

Object Caml is the programming language we used in the implementation of TransM. An additional module is used, it consists of an execution module. XSLT style sheet is generated in order to accomplish the transformation at the instance level. The code generation is based on rules priorities. The template possesses the name of the rule and a priority number is assigned to the rule, more constraint rules have a higher priority on less constraint ones. We aim to study possible extensions on our rule definition. Therefore, and as the architecture of our model permits, our ongoing research will continue to study various potential optimizations to incorporate into our rule specification and the template generating algorithm. Possible updates may take place on the choice of the target language. For now, the target execution language consists of XSLT stylesheet. Following the specification we proposed, the rules' set has achieved required transformation. For additional powerful and more advanced transformation, a language other than the XSLT may be chosen as a target language for the execution model.

8. References

[ABB04] Nouhad Amaneddine, Jean-Paul Bahsoun, Jean-Paul Bodeveix. TransM: A structured document transformation model. In proceeding of the third international conference on information systems and its applications, ISTA, Salt lake city, Utah, USA, p. 53-66, July (2004).

[Ama04] Nouhad Amaneddine. Un modèle de spécification de haut niveau pour la transformation de données structurées. Ph.D thesis, Paul Sabatier University, December (2004).

[AU72] A.V. Aho, J.D. Ullman. The theory of parsing, Translation, and Compiling, vol. 1: Parsing, chapter 3.1 and 3.2. Prentice-Hall, (1972).

[BBG01] A. Banerji, C. Bartolini, D. Ger and al. Web Services Conversation Language WSCL 1.0, http://www.e-speak.hp.com/media/wscl_5_16_01.pdf, (2001).

[BDB04] B. Mathews, D. Lee, B. Dister. Vector Markup Language. http://www.w3.org/TR/NOTE-VML/, (2004).

[Cd04] http://www.cduce.org/, (2004).

[CFM04] D. Chamberlin, P. Fankhauser, M. Marchiori. XML Query Use Cases. W3C Working draft, http://www.xml-cml.org/, (2004).

[CML04] Chemical Markup Language. http://www.xml-cml.org/, (2004).

[Ha04] Haruo Hosoya. Regular expression filters for XML. In Programming Languages Technologies for XML (PLAN-X), Venice, Italy, January 2004.

[KP96] E. Kuikka, M. Penttonen. Transformation of Structured Documents. Processing of Structured Documents Using a Syntax-directed Approach. University of Kuopio, Finland (1996).

[Kru92] C.W. Krueger, Software reuse, in ACM computing survey, (1992).

[KS95] E. Kuikka, A. Salminen. Two-dimensional filters for structured text. Technical report, University of Waterloo, Department of Computer Science, (1995).

[Lei03] P. Leionen. Automating XML Document Structure Transformations. Proceeding of the 2003 ACM Symposium on Document Engineering, Grenoble, France, (2003).

[MR97] P. Murray-Rust. Chemical Markup Language, World Wide Web journal, p. 135-147, http://xml-cml.org, (1997).

[Mur98] M. Murata. Data model for document transformation and assembly (extended abstract). In Principle on Digital Document Processing, p. 140-152, (1998).

[OMG01] Unified Modelling Language, OMG document formal 03-03-01, http://www.omg.org/cgi-bin/doc?formal/, (2001).

[OMG03] The Object Management Group, www.omg.org, (2003).

[PFT02] M. Peltier,France Télécom R&D. Transformation entre un profil UML et un méta-modèle MOF. In Langage et modèles à objets LMO. Montpellier, ISBN: 2-7261-1131-9, (2002).

[QVT03] Revised submission for the MOF 2.0 Query/views/Transformation RFP, August 2003, http://qvtp.org/.

[RSD03] R. Ausbrooks, S. Buswell, D. Carlisle. Mathematical Markup Language, http://www.w3.org/TR/2003/REC-MathML2-20031021/, (2004).

[W3C02] The World Wide Web Consortium, http://www.w3.org/Style/XSL/, (2002).

[WE03] E. Willink, A concrete UML-base graphical transformation syntax- the UML to RDBMS example in UMLX. Thales Research and Technology, England (2003).

[Wil03] E. Willink. UMLX: A graphical transformation language for MDA. In MDAFA03, Workshop on model driven architecture foundation and application. Enschede, The Netherlands, (2003).

[WML02] Wireless Markup Language, http://www.oasis-open.org/cover/wap-wml.html, (2002).

Vers une exploitation structurelle et sémantique de documents

Kaïs Khrouf[1,2], Mohamed Mbarki[1], Chantal Soulé-Dupuy[1,2]

[1] *IRIT – Equipe SIG/D2S2 – Université Paul Sabatier*
118 route de Narbonne, 31062 Toulouse Cedex - France

[2] *Université Toulouse 1*
Place Anatole France, 31042 Toulouse Cedex - France

{khrouf,mbarki,soule}@irit.fr

Résumé :

Face à l'augmentation considérable du nombre de documents numériques, généralement issus de sources disséminées et hétérogènes, le besoin d'outils pour traiter automatiquement les informations contenues s'est rapidement fait sentir. A cette fin, nous proposons le concept d'entrepôt de documents permettant d'intégrer et d'organiser des informations hétérogènes. Ces informations peuvent alors être analysées selon plusieurs dimensions non-prédéfinies (analyse multidimensionnelle) afin d'en déduire de nouvelles informations ou connaissances.

Mots-clés : Entrepôt, méta-modèle, description structurelle, description sémantique, métadonnées, analyse multidimensionnelle.

1. Introduction

De nos jours, il apparaît que l'acquis du personnel d'une entreprise, en terme de connaissances et du savoir-faire, doit être préservé de sorte à ce qu'il puisse être exploité au mieux des intérêts de l'entreprise. Préserver cet acquis revient à constituer une *mémoire d'entreprise*. Cette mémoire consiste à définir les modalités de sauvegarde et de diffusion de tout ce qui a été appris par le personnel de l'entreprise et qui contribue à sa bonne marche et ses succès. Elle consiste aussi à

sauvegarder, organiser et exploiter les documents qui constituent, en plus des données internes de l'entreprise, une mine de connaissances à ne pas négliger.

La définition de normes, telles que Structured Generalized Markup Language (SGML) puis eXtensible Markup Language (XML), et les travaux du World Wide Web Consortium (W3C) ont permis d'initier de nouvelles perspectives concernant l'exploitation des documents électroniques et en particulier des informations semi-structurées ; un document électronique devient alors un ensemble d'objets plus ou moins complexes et non plus seulement une chaîne de caractères. Ceci permet d'envisager de nouvelles perspectives en matière de recherche, d'interrogation et d'analyse des informations semi-structurées.

C'est dans ce contexte de perspectives que nous proposons la constitution d'entrepôts de documents permettant le stockage de documents hétérogènes ainsi que leur classification selon des structures logiques génériques (structures, représentant un découpage de l'information d'un point de vue hiérarchique, communes à un ensemble de documents). Nous distinguons principalement trois types d'hétérogénéité : structurelle (documents ayant des structures logiques différentes), thématique (abordant des domaines très variés) et linguistique (écrits dans des langues différentes).

Une telle organisation des entrepôts permet de faciliter l'exploitation des informations documentaires au travers de plusieurs techniques complémentaires :

1. La recherche d'information consiste à restituer des granules de documents en réponse à une requête utilisateur formulée par des mots-clés,
2. L'interrogation des données consiste à récupérer des données factuelles en utilisant un langage déclaratif,
3. Et l'analyse multidimensionnelle consiste à manipuler les informations contenues dans l'entrepôt selon des dimensions non-prédéfinies.

Ce papier se décompose comme suit. La section 2 décrit les travaux existants pour la manipulation des documents. Dans la section 3, nous présentons le méta-modèle d'entrepôts de documents proposé. Ensuite, nous détaillons notre approche d'analyse multidimensionnelle appliquée aux informations documentaires de l'entrepôt. Enfin, nous décrivons l'outil réalisé DocWare (Document Warehouse).

2. Travaux existants

De nos jours, les informations documentaires constituent un facteur de savoir-faire, de mémoire et de connaissances. La manipulation et l'exploitation des documents électroniques représentent donc un nouveau challenge pour toute entreprise.

Pour assurer le stockage ainsi que l'exploitation des documents semi-structurés, le plus souvent au format XML, deux types de SGBD ont été utilisés : les SGBD natifs et les SGBD middlewares. Les SGBD *natifs* sont développés

spécifiquement pour XML, ils stockent des documents complets ou des parties de documents dans des fichiers et ne réalisent pas de transformations (c'est à dire « mapping ») en tables. Plusieurs travaux ont été proposés tels que Natix [KANN00], InfonyteDB [HUCK00] et Xylème [ABIT02]. Les SGBD *middlewares* réalisent l'intégration des documents XML selon un modèle relationnel/objet, ils transforment les documents XML en tables et vice-versa, tels que XRel [YOSH01], e-XMLRepository [GARD02]. Ces SGBD sont plus appropriés à la définition de méta-schéma permettant ainsi l'instanciation d'entrepôts de données semi-structurées hétérogènes.

Pour stocker et manipuler les documents multimédia, plusieurs modèles ont été présentés qui peuvent être classés en deux catégories. La *première catégorie* rassemble des travaux tels que [LOIS02] et [MOEN04] qui visent à modéliser séparément chaque type de média. Cette séparation permet une description détaillée orientée médium mais pas une modélisation générique et globale d'un document multimédia. La *deuxième catégorie* correspond aux modèles tels que ceux proposés dans [DARM02] et [AMOU02] qui visent à traduire la totalité de chaque document en se basant à la fois sur leur structure et leur contenu. Cette approche assure ainsi une représentation plus fidèle des documents multimédia. Néanmoins, elle n'offre pas une séparation claire des informations relatives au contenu de celles relatives à la structure, ce qui induit un manque de flexibilité dans l'exploitation et la manipulation des documents.

Par rapport aux travaux présentés précédemment, nous proposons un méta-modèle d'entrepôts de documents permettant :

- D'intégrer tout type de documents (structurés, semi-structurés et non structurés),
- D'accepter tout type d'information (texte, image, vidéo, audio).

L'entrepôt doit constituer un référentiel centralisé et pérenne des différents documents utilisés et manipulés par une entreprise.

D'un autre point de vue, le méta-modèle proposé permet de :

- Regrouper les documents par des structures logiques communes,
- Représenter le contenu des documents par des ensembles de métadonnées homogènes et non-prédéfinies.

L'entrepôt permet ainsi de faciliter, en utilisant les éléments de structure et de contenu, l'exploitation des informations documentaires au travers des techniques de recherche, d'interrogation voire d'analyse.

3. Méta-modèle d'entrepôts de documents

Les nouveaux moyens de traitement et de communication de l'information ont favorisé via Internet, les Intranets, les workflows, les bibliothèques numériques… la mise à disposition et l'accès à une masse toujours croissante et quasi-illimitée d'informations hétérogènes. Afin de gérer cette hétérogénéité (structurelle, thématique et linguistique), nous proposons un méta-modèle d'entrepôts de documents qui comprend deux parties :

1. La *description structurelle* qui reflète l'organisation hiérarchique des documents de l'entrepôt ;
2. La *description sémantique* qui consiste à organiser et à homogénéiser les métadonnées pouvant être extraites à partir des éléments de la structure logique.

3.1 Description structurelle

La figure 1 présente la description structurelle du méta-modèle proposé. Cette description consiste à regrouper et à classifier les documents de l'entrepôt selon des structures communes ce qui permet par la suite de faciliter leur exploitation.

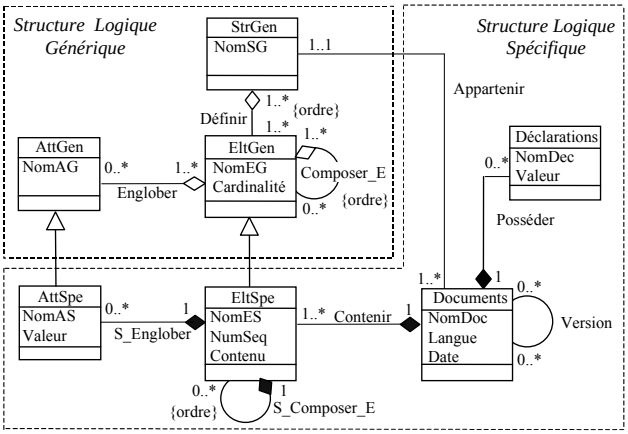


Figure 1 : Partie structurelle du méta-modèle d'entrepôts de documents

La partie structurelle du méta-modèle proposé comporte deux types de structures :

1. La structure logique générique dont relève toute une classe de documents. Elle est définie par un ensemble d'éléments génériques pouvant être composés d'autres éléments génériques et/ou décrits par des attributs génériques ;
2. La structure logique spécifique qui correspond à un document. Elle est définie par un ensemble d'éléments spécifiques « informations » pouvant englober des attributs spécifiques.

Exemple 1 :

Soit la structure logique générique "CV" définie par une "Entête", une "Photo", un ensemble de "Diplômes" et une "Motivation". Les documents qui sont conformes à cette structure deviennent alors des instances de cette dernière.

L'intérêt de la description structurelle du méta-modèle réside dans :

- L'intégration de tous types de documents caractérisés par une absence totale ou partielle de structure [KHRO03a]. Des exemples d'instanciation du méta-modèle par différents types de documents ont été proposés dans [KHRO04],
- Le regroupement des documents selon des structures communes similaires. Dans [KHRO03b], nous avons défini un algorithme de comparaison d'arborescences ordonnées et étiquetées basé sur un degré de ressemblance,
- La persistance des documents jugés pertinents ainsi que la mise en évidence de la granularité des documents.

3.2 Description sémantique

La partie structurelle du méta-modèle reflète l'organisation hiérarchique du document. Cependant, elle ne décrit pas le contenu des différents éléments structurels (textuels ou multimédia). Ceci est expliqué par le fait que ce type de description tient compte seulement des éléments extraits à partir d'une description générale de document telle que Document Type Definition (DTD). Cette description ne propose pas une décomposition plus fine des éléments de structures des documents (si nous reprenons l'exemple 1, la description structurelle de "CV" n'offre pas une description détaillée des éléments "Photo" et "Motivation").

Pour obtenir une description plus détaillée du contenu des documents, nous proposons d'étendre le méta-modèle proposé par une description sémantique (de contenu) pour les différents fragments d'un document (cf. figure 2).

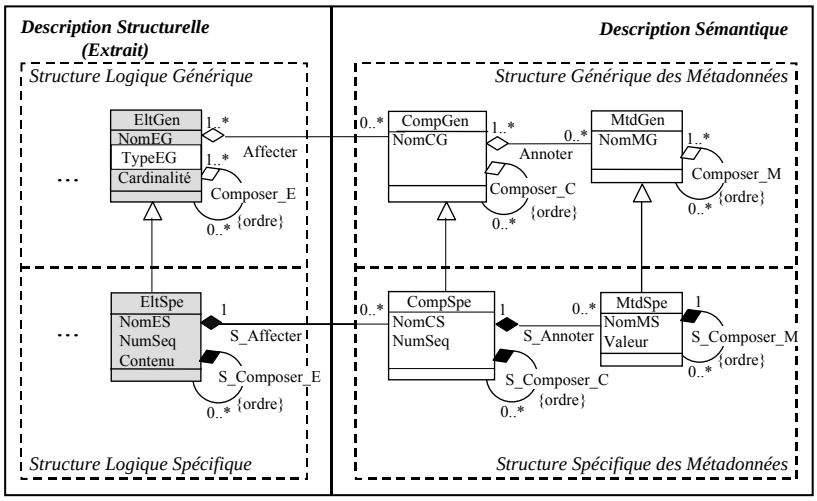


Figure 2 : Partie sémantique du méta-modèle d'entrepôts de documents

La partie sémantique du méta-modèle inclut aussi deux types de structures :

1. La structure générique des métadonnées qui permet de décrire l'ensemble des composants et des métadonnées pouvant caractériser un élément générique ;
2. La structure spécifique des métadonnées qui décrit la composition spécifique d'une information selon la description définie pour l'élément générique correspondant.

Exemple 2 :

Nous proposons d'étendre la description des éléments "Photo" et "Motivation" de l'exemple 1 comme suit :

- L'élément "Photo" (image) peut être défini par les deux composants "Tête" et "Yeux". Ces deux composants peuvent être annotés respectivement par les métadonnées "Forme" et "Couleur". Ces informations peuvent être déduites en utilisant les outils de [DOWD03],
- L'élément "Motivation" (bande sonore) est formé par un ensemble de segments "parole et musique". Chaque segment est caractérisé par un couple temporel qui marque son début et sa fin et la liste des "Langues" utilisées. Les outils de [PARL03] permettent d'extraire de telles informations.

L'intérêt de la partie sémantique du méta-modèle réside dans :

- Une description plus détaillée du contenu des documents au-delà de la transcription fidèle de la description structurelle,

- Une annotation des documents appartenant à une même structure logique générique par des métadonnées homogènes et non-prédéfinies [MBAR04],
- Une gestion plus flexible de documents en offrant la possibilité de manipulations multi-critères en combinant les deux descriptions (structurelle et sémantique).

Pour conclure, l'originalité de notre approche basée sur les entrepôts de documents se situe principalement au niveau du méta-modèle proposé. En effet, une structure générique (logique et/ou métadonnées) peut être considérée comme le schéma de tous les documents appartenant à cette structure alors que les structures spécifiques de ces documents constituent les différentes instances du schéma. Ce méta-modèle permet de représenter n'importe quel document numérique dématérialisé sans imposer de structure à priori. Les niveaux de granularité (structures logiques hétérogènes) et de description des informations (métadonnées et annotation) sont des atouts majeurs pour une exploitation flexible du contenu d'un entrepôt de documents selon des points de vue non-prédéfinis. Chaque point de vue constitue dans ce cas un axe d'analyse en profondeur de l'entrepôt basé sur n'importe quel constituant du méta-modèle (élément, composant ou métadonnée).

Remarque : Lors de l'intégration de documents dans l'entrepôt, les problèmes de la synonymie et de l'unicité des étiquettes des arborescences sont traités [KHRO03b].

4. Analyse multidimensionnelle

La modélisation multidimensionnelle consiste à considérer un sujet analysé comme un point dans un espace à plusieurs dimensions [KIMB02]. A cette fin, nous proposons de dériver des magasins de documents (extraits de l'entrepôt) supportant les processus d'analyses décisionnelles. Un magasin de documents est dédié à un type d'utilisateurs et il doit répondre à un objectif décisionnel précis ou un besoin spécifique. Le processus proposé, pour analyser d'une manière multidimensionnelle les informations contenues dans l'entrepôt, se compose de trois phases (cf. figure 3).

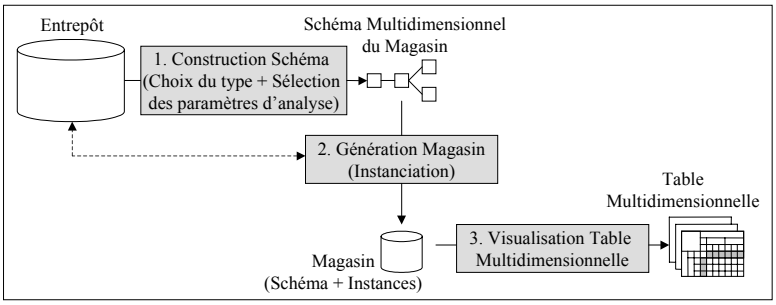


Figure 3 : Processus d'analyse multidimensionnelle

4.1 Construction des schémas des magasins

La première phase du processus d'analyse multidimensionnelle consiste à générer, à partir de l'entrepôt, le schéma du magasin de documents désiré. Cette phase se compose de quatre étapes.

⇨ Etape 1 : Choix du type d'analyse

Cette étape consiste à choisir un type d'analyse par rapport à l'organisation des informations documentaires dans l'entrepôt. Nous distinguons trois types d'analyse, à savoir :

1. Analyse structurelle : l'analyse se base sur les éléments structurels de l'entrepôt,
2. Analyse sémantique : l'analyse se base sur les différentes métadonnées extraites,
3. Analyse mixte : l'analyse se base conjointement sur les éléments de structures et les métadonnées.

Cette étape consiste aussi à préciser le domaine d'application, à savoir :

- par collection : il s'agit d'analyser les documents appartenant à la même structure générique,
- par paramètres : il s'agit d'analyser des documents appartenant à plusieurs structures génériques.

⇨ Etape 2 : Sélection des paramètres d'analyse

Cette étape consiste à préciser les paramètres d'analyse, à savoir :

1. Un fait qui modélise un sujet de l'analyse. Il est caractérisé par un ensemble de mesures de l'activité à analyser,
2. Ses dimensions qui modélisent des perspectives de l'analyse. Ce sont les critères sur lesquels nous souhaitons évaluer, quantifier et qualifier les faits.

Il s'agit aussi d'indiquer l'ordre des dimensions et la fonction d'agrégation pour la mesure du fait (Compte, Somme, Maximum, Minimum, Moyenne).

⇨ Etape 3 : Filtrage

Cette étape consiste à sélectionner des valeurs précises afin d'affiner les analyses. Nous distinguons deux types de filtrage :

1. Pour une dimension, nous choisissons, parmi toutes ses valeurs, celles que nous voulons intégrer dans le magasin,
2. Pour le fait qui est toujours sous forme numérique, nous proposons un filtrage plus fin qui nous permet de fixer des critères de sélection, en utilisant des opérateurs classiques de comparaison (<, >, =, <>, <=, >=).

⇨ Etape 4 : Visualisation du schéma

Cette étape consiste à visualiser le schéma du magasin pour faire des éventuelles modifications. La visualisation se fait selon une représentation graphique facilitant les analyses décisionnelles.

Exemple 3 :

Nous souhaitons analyser les livres édités lors de ces dernières années (2002, 2003, 2004) par nationalité des auteurs et par éditeur. Supposons que l'entrepôt de documents contient la structure générique "Liste_livres" (cf. figure 4) décrivant tous les livres de l'entrepôt.

1. Le type d'analyse à choisir pour ce cas de figure est : l'*analyse mixte par collection* :
   - analyse mixte puisque les paramètres d'analyse sont des éléments et aussi des métadonnées,
   - par collection parce que tous les livres de l'entrepôt sont décrits par la même structure générique.
2. Les paramètres d'analyse sont : la première dimension est "Nationalité", la deuxième dimension est "Auteur" et la troisième dimension est "Editeur", alors que la mesure du fait est le nombre de "Titre".
3. Le filtrage consiste à sélectionner les valeurs 2002, 2003 et 2004 pour le paramètre "Année".
4. La visualisation du schéma multidimensionnel du magasin se fait comme l'indique la figure 4.

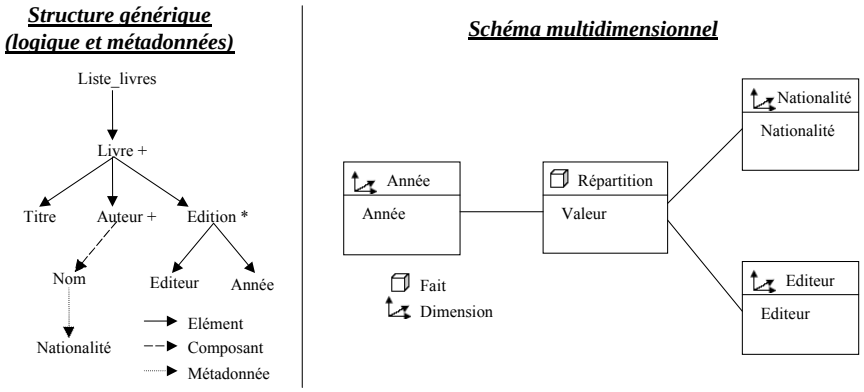


Figure 4 : Structure générique et schéma multidimensionnel

4.2 Génération automatique des magasins

La deuxième phase du processus d'analyse multidimensionnelle consiste à générer le magasin d'une manière automatique afin de récupérer les informations de l'entrepôt. Cette génération se fait en deux étapes.

⇨ Etape 1 : Génération d'une vue pour chaque paramètre d'analyse

Pour chaque objet (élément de structure ou métadonnée) constituant un paramètre d'analyse, le système doit générer une vue englobant trois attributs :

- Le premier attribut "Doc" de la vue correspond aux numéros des documents extraits de l'entrepôt concernant tous les éléments ou métadonnées spécifiques qui héritent respectivement de l'élément ou de la métadonnée générique jouant le rôle d'un paramètre d'analyse (fait ou dimension),
- Le deuxième attribut "Anc" de la vue correspond aux numéros des éléments ou des métadonnées spécifiques qui héritent du premier ancêtre commun de tous les paramètres d'analyse,
- Le dernier attribut "Inf" de la vue correspond à l'information contenue dans l'élément ou la métadonnée spécifique qui hérite respectivement de l'élément ou de la métadonnée générique correspondant.

La vue d'une dimension aura la forme suivante.

```
CREATE VIEW Dim_n (Doc, Anc, Inf) AS
SELECT
   e.[s_composer_m…].[s_composer_c…].[s_composer_e…].sondoc.numdoc,
   e.[s_composer_m…].[s_composer_c…].[s_composer_e…].numes,
   e.contenu
FROM   EltSpec e       (le cas d'un élément)
       CompSpec e      (le cas d'un composant)
       MdtSpec e       (le cas d'une métadonnée)
WHERE  e.herite.nomeg = "Inf"
AND    [s_composer_m…].[s_composer_c…].[s_composer_e…].sondoc.
        appartient.doctype = "SLG"
AND    (e.contenu = "V1" OR … OR e.contenu = "Vn");
```

La figure suivante présente la vue à générer et son résultat pour le paramètre d'analyse "Editeur".

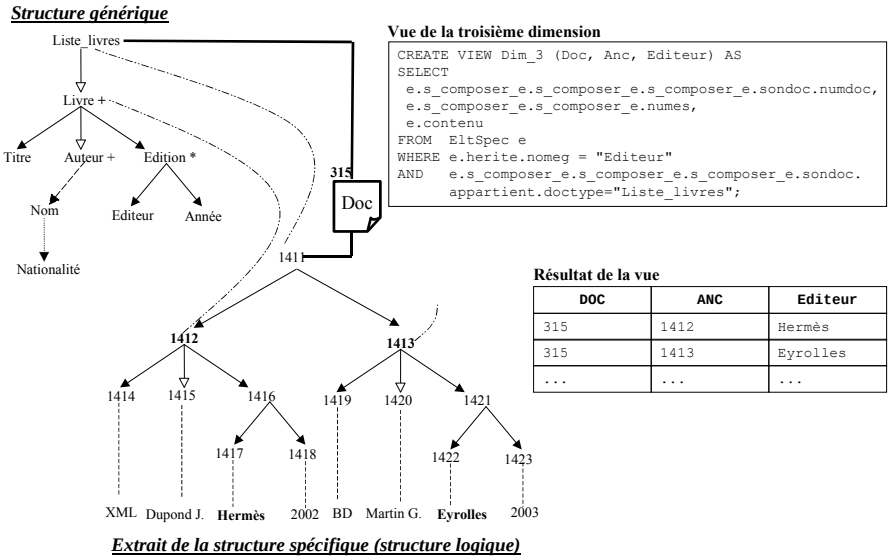


Figure 5 : Exemple de génération de vue

⇨ Etape 2 : Jointure et groupement des différentes vues générées

Le système doit ensuite, à partir des vues générées, établir une nouvelle vue par une jointure sur les deux premiers attributs "Doc" et "Anc" de toutes les vues. La nouvelle vue aura alors la forme suivante.

```
CREATE VIEW Jointure (Inf_1, Inf_2… , Inf_n, Inf) AS
SELECT d1.Inf, d2.Inf… , dn.Inf, f.Inf
FROM     Dim_1 d1, Dim_2 d2… , Dim_n dn, Fact f
WHERE    d1.Doc = d2.Doc       AND     d2.Doc = d3.Doc
…
AND      dn-1.Doc = dn.Doc     AND     dn.Doc = f.Doc
AND      d1.Anc = d2.Anc       AND     d2.Anc = d2.Anc
…
AND      dn-1.Anc = dn.Anc     AND     dn.Anc = f.Anc ;
```

Pour notre exemple, le système doit générer la vue suivante.

```
CREATE VIEW Jointure(d1_Année,d2_Nationalité,d3_Editeur,f_Titre) AS
SELECT  d1.année, d2.nationalité, d3.éditeur, f1.titre
FROM           Dim_1 d1, Dim_2 d2, Dim_3 d3, Fact f
WHERE          d1.anc = d2.anc       AND     d2.anc = d3.anc
AND            d3.anc = f.anc        AND     d1.doc = d2.doc
AND            d2.doc = d3.doc       AND     d3.doc = f.doc ;
```

Le résultat de cette vue est une table ayant la structure suivante.

D1_ANNEE	D2_Nationalité	D3_EDITEUR	F_TITRE
2002	Française	Hermès	XML
2003	Française	Eyrolles	BD
2003	Canadienne	Hermès	JAVA
2004	Américaine	Idea group	PHP
…	…	…	…

Le système doit effectuer à ce niveau une opération de groupement en utilisant la fonction d'agrégation choisie par l'utilisateur, pour générer une dernière vue qui représente le contenu du magasin. Cette vue aura la forme suivante.

```
CREATE VIEW Vue (j.Inf_1, j.Inf_2… , j.Inf_n) AS
SELECT      j.Inf_1, j.Inf_2… , j.Inf_n, Function(j.Inf)
FROM        Jointure j
GROUP BY    j.Inf_1, j.Inf_2… , j.Inf_n ;
```

La dernière vue générée de notre exemple est la suivante.

```
CREATE VIEW Répartition (d1_Année,d2_Nationalité,d3_Editeur, Nb) AS
SELECT j.d1_année, j.d2_nationalité, j.
```


Le résultat de cette vue est une table ayant la structure suivante.

D1_ANNEE	D2_NATIONALITE	D3_EDITEUR	NB
2002	Française	Hermès	5
2003	Française	Eyrolles	1
2003	Canadienne	Hermès	3
2004	Américaine	Idea group	6
…	…	…	…

4.3 Visualisation

La dernière phase consiste à restituer le contenu de la dernière vue générée par le système sous forme de table multidimensionnelle. Plusieurs techniques de visualisation de données ont été proposées dans la littérature telles que [AHLB94], [HOLS00], [KOLS01] et [MOTH02]. Dans nos travaux, nous avons opté pour les tables multidimensionnelles parce qu'elles sont assez simples à manipuler et à interpréter et elles permettent de mieux apprécier le contenu des magasins de documents. Ces tables organisent les données en les classant suivant les dimensions choisies par l'utilisateur. Ainsi, les colonnes représentent la première dimension, les lignes représentent la deuxième dimension et les plans représentent la troisième dimension. Alors que les valeurs des mesures des faits sont représentées à l'intérieur des tables sous formes d'interrelation entre les valeurs des dimensions.

Si le nombre de dimensions est supérieur à trois, chacune des dimensions suivantes : dim_4, dim_5… , dim_n sera affectée par une valeur selon un critère de sélection. Ces valeurs peuvent être modifiées par l'utilisateur. Cependant, les trois premières dimensions seront toujours affichées sous forme de colonnes, lignes et plans, en tenant compte des valeurs affectées aux autres dimensions. Dans le cadre de nos travaux, nous supposons que le nombre maximum de dimensions est égal à trois.

Le passage de la dernière vue générée par le système en une table multidimensionnelle se fait de la manière suivante : Etant donné que chaque plan de la table multidimensionnelle correspond à une seule valeur de la troisième dimension, le système génère une vue en effectuant une sélection sur une valeur précise. Cette nouvelle vue contient trois colonnes :

1. La première dimension,
2. La deuxième dimension,
3. Le fait.

A partir de cette vue, le système doit :

- Récupérer toutes les valeurs possibles de la première dimension, ces valeurs seront affichées dans les colonnes du plan correspondant,
- Récupérer toutes les valeurs possibles de la deuxième dimension, ces valeurs seront affichées dans les lignes du plan correspondant,
- Restituer pour chaque couple (une colonne i et une ligne j) la mesure correspondante à partir de la troisième colonne de la vue (le fait). Cette mesure sera affichée dans la case correspondante (intersection entre i et j).

La dernière vue générée par le système, dans la section précédente, doit être alors visualisée selon la table multidimensionnelle comme l'indique la figure 6.

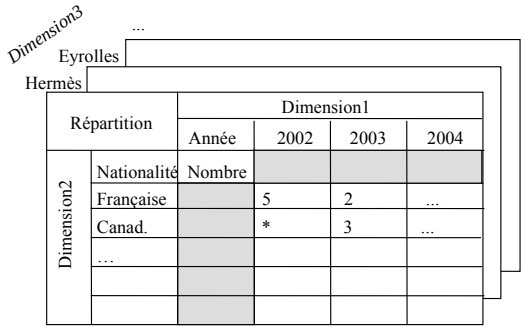


Figure 6 : Table multidimensionnelle

5. Implantation et expérimentations

Afin de valider les propositions présentées, nous avons réalisé un outil d'aide à l'intégration et à l'analyse de documents, intitulé DocWare (Document Warehouse). Les documents ayant servi à nos expérimentations ont été extraits de sources diverses : documents XML issus de sites Web et de CD-ROM fournis dans le cadre de benchmarks (Reuters…) ou de bases de tests (TREC…). Ces documents ne sont pas associés à un domaine particulier, et sont soit en français, soit en anglais. L'instanciation de la partie structurelle du méta-modèle a été réalisée d'une manière automatique au travers d'un parseur PERL [KHRO03b] que nous avons réalisé. Cependant, l'instanciation de la partie sémantique a été réalisée d'une manière manuelle. Les procédures d'automatisation de cette instanciation sont en cours de développement.

Nous présentons, dans cette section, une validation de la démarche présentée concernant le module d'analyse multidimensionnelle.

5.1 Cas d'une analyse structurelle

Nous souhaitons analyser les mois les plus importants de l'année où se déroulent un nombre important de conférences par type (national ou international) et thème de recherche. Le nombre de conférences doit être supérieur à 10.

En observant le contenu de l'entrepôt, nous déduirons que les documents décrivant les publications dans des conférences sont regroupés selon une seule structure logique générique, intitulée "Pub-conf". Chaque publication est décrite par ses "Auteurs", son "Titre", des informations concernant la "Conférence" ("Nom", "Lieu", "Type" et "Date" qui se compose de "Mois" et "Année"), le "Thème" et un "Résumé". Cette structure logique générique contient alors tous les éléments nécessaires pour réaliser cette analyse.

⇨ Etape 1 : Choix du type d'analyse

La première étape de la construction du magasin consiste à sélectionner le type d'analyse (dans notre exemple, c'est une *analyse structurelle par collection*). Ainsi, le système affiche la liste de toutes les structures existantes dans l'entrepôt. Parmi ces structures, nous choisissons la structure logique générique "Pub-conf" qui sera par la suite visualisée d'une manière automatique sous forme d'une arborescence.

⇨ Etape 2 : Sélection des paramètres d'analyse

A ce niveau, nous sélectionnons et définissons les paramètres d'analyse, il s'agit de préciser les dimensions et le fait au travers des menus contextuels. Dans notre exemple, la première dimension est le "Thème", la deuxième dimension est le "Mois" et la troisième dimension est le "Type". La mesure du fait est le calcul du nombre de conférences (l'élément "Nom").

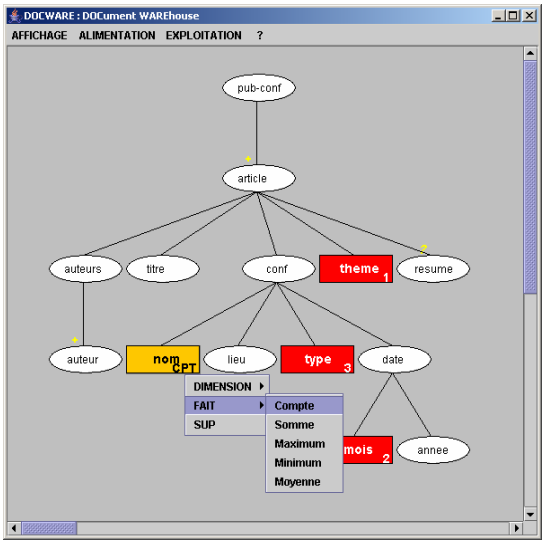


Figure 7 : Sélection des paramètres d'analyse (Analyse structurelle)

⇨ Etape 3 : Filtrage

Nous souhaitons analyser les mois pour les publications nationales et internationales. Nous appliquons ainsi un filtre sur la troisième dimension. Pour cela, le système affiche toutes les valeurs de l'élément "Type". Parmi ces valeurs, nous choisissons les types correspondants. Une deuxième contrainte indique que le nombre de conférences par thème, mois et type doit être supérieur à 10. Nous appliquons alors un filtre sur la mesure du fait. Le

système affiche une boite de dialogue pour nous permettre de spécifier le critère de sélection.

⇨ Etape 4 : Visualisation du schéma

A ce niveau, nous pouvons consulter le schéma multidimensionnel du magasin d'une manière graphique (cf. figure 8).

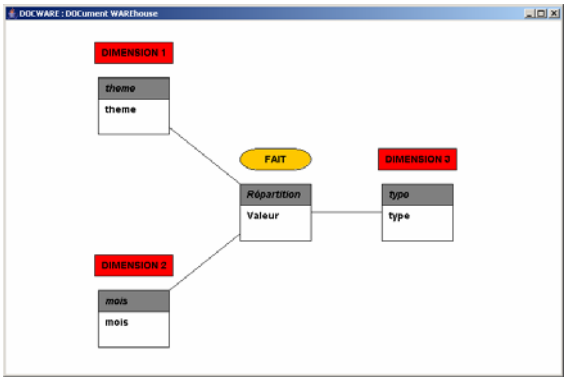
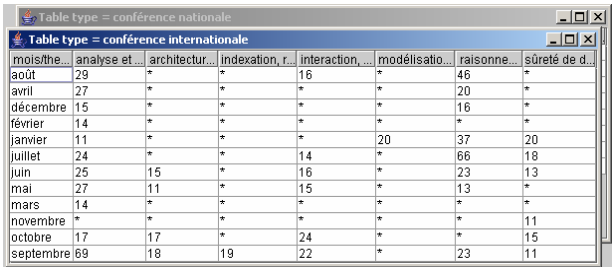


Figure 8 : Visualisation du schéma multidimensionnel (Analyse structurelle)

Pour visualiser le résultat, le système crée les vues selon la démarche décrite dans la section 4.2 et affiche le résultat sous forme de table multidimensionnelle.



mois/the...	analyse et ...	architectur...	indexation, r...	interaction, ...	modélisatio...	raisonne...	sûreté de d...
août	29	*	*	16	*	46	*
avril	27	*	*	*	*	20	*
décembre	15	*	*	*	*	16	*
février	14	*	*	*	*	*	*
janvier	11	*	*	*	20	37	20
juillet	24	*	*	14	*	66	18
juin	25	15	*	16	*	23	13
mai	27	11	*	15	*	13	*
mars	14	*	*	*	*	*	*
novembre	*	*	*	*	*	*	11
octobre	17	17	*	24	*	*	15
septembre	69	18	19	22	*	23	11

Figure 9 : Table multidimensionnelle (Analyse structurelle)

5.2 Cas d'une analyse mixte

Un réalisateur souhaite interroger le contenu de l'entrepôt pour effectuer le « Casting » d'un film. Il peut ainsi utiliser l'analyse multidimensionnelle pour connaître les noms ainsi que les intitulés des diplômes des acteurs qui ont des yeux bleus ou bien noirs et qui parlent au moins deux langues différentes.

Les documents décrivant les CV des artistes au niveau de l'entrepôt sont regroupés selon une seule structure générique, intitulée "Cv-artiste". Chaque CV est décrit par une "Entête" ("Nom", "Prénom" et "Adresse"), une "Photo", un ou plusieurs "Diplômes" ("Intitulé", "Lieu" et "Année") et une "Motivation" :

- Une photo est formée par les composants "Tête" et "Yeux" qui sont décrits par un ensemble de métadonnées ("Forme_T", "Forme_Y" et "Couleur_Y"),
- Une motivation est formée par un ensemble des segments de type parole et musique "Parole_M" dont chacun est caractérisé par un couple temporel qui marque son début et sa fin "Deb_Fin_P" et la liste des "Langues" utilisées.

⇨ Etape 1 : Choix du type d'analyse

Dans cet exemple, nous allons utiliser une *analyse mixte par collection* puisque la structure générique (logique + métadonnées) "Cv-artiste" contient tous les paramètres nécessaires pour réaliser cette analyse.

⇨ Etape 2 : Sélection des paramètres d'analyse

Dans cet exemple, la première dimension est la "Couleur_Y", la deuxième dimension est le "Nom" et la troisième dimension est l'"Intitulé". La mesure du fait est le calcul du nombre de langues (l'élément "Langue").

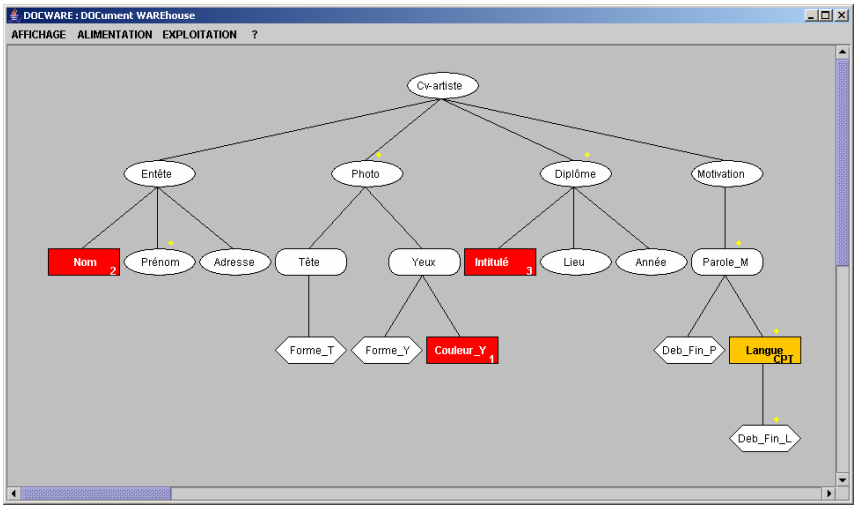


Figure 10 : Sélection des paramètres d'analyse (Analyse mixte)

⇨ Etape 3 : Filtrage

Nous souhaitons trouver les artistes ayant des yeux bleus ou noirs. Nous appliquons alors un filtre sur la première dimension. La deuxième contrainte indique que la motivation doit être exprimée dans au moins deux langues différentes, nous appliquons alors un filtre sur la mesure du fait.

Le résultat de cette analyse peut être visualisé sous forme de table multidimensionnelle comme l'indique la figure suivante.

Figure 11 : Table multidimensionnelle (Analyse mixte)

6. Conclusion

Le document électronique devient un moyen d'échange d'informations prisé pour des raisons évidentes liées en grande partie à des possibilités quasi-infinies d'organisation et d'exploitation. Ainsi, le contenu de l'information diffusée de nos jours a évolué, elle est devenue de plus en plus riche et complexe en intégrant des structures et des contenus hétérogènes. L'approche que nous avons choisie pour la gestion de cette hétérogénéité se base sur un méta-modèle intégrant plusieurs niveaux de description imbriqués : une couche générique, une couche spécifique, une description structurelle ainsi qu'une description sémantique (métadonnées). Les entrepôts de documents basés sur un tel méta-modèle doivent ainsi permettre une exploitation aisée et multi-vues des informations documentaires.

Les perspectives envisagées à ces travaux concernent :

1. L'intégration des techniques du « text mining » dans le processus d'analyse multidimensionnelle afin de déduire des connaissances communes entre les différents granules,
2. L'adaptation du méta-modèle et des techniques d'intégration de documents dans l'entrepôt pour gérer les aspects « multi-structures » et « temps » ; ces deux aspects concernent respectivement l'affectation à un même document plusieurs structures différentes ainsi que la détection et la gestion des différences structurelles et de contenus entre plusieurs versions d'un même document,

3. Une expérimentation « à plus grande échelle » afin de procéder à des évaluations quantitative et qualitative plus approfondies de l'outil réalisé DocWare.

7. Références bibliographiques

[ABIT02] Abiteboul S., Cluet S., Ferran G., Rousset M.C., "The Xyleme Project", Computer Networks, (3): 225-238, 2002.

[AHLB94] Ahlberg C., Shneiderman B., "Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays", ACM Conference on Human Factors in Computing Systems, p. 313-317, Boston, MA, USA, 1994.

[AMOU02] Amous I., Jedidi A., "Modélisation des métadonnées pour une recomposition dynamique des documents", Congrès de l'Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID'02), p. 243-258, Nantes, France, 2002.

[DARM02] Darmont J., Boussaid O., Bentayeb F., "Warehousing Web Data", International Conference on Information Integration and Web-based Applications and Services (iiWAS 02), p. 148-152, Bandung, Indonesia, 2002.

[DOWD03] Dowdalla J.B., Pavlidisa I., Bebisb, G., "Face Detection in the Near-IR Spectrum", Image and Vision Computing, vol. 21, n° 7, p. 565-578, 2003.

[GARD02] Gardarin G., Mensch A., Tomasic A., "An Introduction to the e-XML Data Integration Suite", Conference on Extending Database Technology (EDBT'02), p. 297-306, Prague, Czech Republic, 2002.

[HOLS00] Hölscher C., Strube G., "Web Search Bahavior of Internet Experts and Newbies", International Conference on the World Wide Web (WWW'00), Amsterdam, Pays-Bas, 2000.

[HUCK00] Huck G., Macherius I., Fankhauser P., "PDOM: Lightweigt Persistency Support for the Document Object Model", Succeeding with Object Databases, p. 107-118, John Wiley, 2000.

[KANN00] Kanne C.C., Moerkotte G., "Efficient Storage of XML Data", International Conference on Data Engineering, San Diego, California, USA, 2000.

[KIMB02] Kimball R., Ross M., "The Data Warehouse Toolkit", John Wiley & Sons, New York, USA, second edition, 2002.

[KHRO03a] Khrouf K., Soulé-Dupuy C., "Vers une mémoire d'entreprise via les entrepôts de documents : Extraction de structures logiques", Extraction et Gestion des Connaissances (EGC'03), p. 201-206, Lyon, France, 2003.

[KHRO03b] Khrouf K., Ravat F., Soulé-Dupuy C., "Comparaison et fusion de structures logiques de documents semi-structurés", Ingénierie des Systèmes d'Information (ISI), Edition Hermès, vol. 8, n° 5-6/2003, p. 127-151, 2003.

[KHRO04] Khrouf K., "Entrepôts de documents : De l'alimentation à l'exploitation", Thèse de doctorat, Université Paul Sabatier, Toulouse III, Juillet 2004.

[KOLS01] Kolski C., "Systèmes d'information et interactions Homme-Machine", Environnements évolués et évaluation de l'interaction Homme-Machine, Edition Hermès, vol. 2, p. 80-114, 2001.

[LOIS02] Loisant E., Ishikawa H., Martinez J., "Designing a Model Independent Multimedia Database", Days of Science and Technology, Tokyo, Japan, 2002.

[MBAR04] Mbarki M., Soulé-Dupuy C., "A Conceptual Modeling of Multimedia Documents", IADIS International Conference WWW/Internet 2004, p. 1051-1056, Madrid, Espagne, 2004.

[MOEN04] Moënne-Loccoz N., Janvier B., Marchand-Maillet S., Bruno E., "Managing Video Collections at Large", International Workshop on Computer Vision meets Databases (CVDB'2004), Paris, 2004.

[MOTH02] Mothe J., Chrisment C., Alaux J., "Visualisation globale de collections de documents sous forme d'hypercube", Extraction et Gestion des Connaissances (EGC'02), p. 131-142, Montpellier, France, 2002.

[PARL03] Parlangeau-Vallès N., Farinas J., Fohr D., Illina I., Magrin-Chagnolleau I., Mella O., Pinquier J., Rouas J-L., Sénac C., "Audio Indexing on the Web: A Preliminary Study of Some Audio Descriptors", SCI 2003, Orlando, Florida, USA, 2003.

[YOSH01] Yoshikawa M., Amagasa T., Shimura T., Uemura S., "XRel: A Path-Based Approach to Storage and Retrieval of XML Documents Using Relational Databases", ACM Transactions on Internet Technology, 1(1):110-141, 2001.

Conférence

Invitée

Systèmes multilingue recherche interlingue

Christian Fluhr

CEA/LIST – Laboratoire d'ingénierie de la connaissance multimédia multilingue
BP 6, Route du Panorama, 92265 Fontenay aux Roses Cedex - France

Christian.fluhr@cea.fr

Résumé :

Cet article présente les difficultés de la prise en compte d'une grande généralité de langues dans les systèmes de recherche d'information textuelle. Après avoir précisé les raisons pour lesquelles la dernière décade a vu une accélération de l'intérêt pour le multilinguisme, on donne une définition des systèmes multilingues et de l'interrogation interlingue.

Les différents problèmes qui peuvent se poser pour l'analyse de différentes langues sont abordés. On présente ensuite les différentes approches de la recherche interlingue et les campagnes de tests qui ont été élaborées pour évaluer la qualité des systèmes.

On donne enfin un exemple d'interrogation interlingue sur une base trilingue français, anglais et arabe.

Mots-clés : information multilingue, systèmes de recherche interlingues.

1. Importance croissante de l'information bilingue

Deux idées qui avaient encore cours récemment sont aujourd'hui largement battues en brèche. La première consiste à penser que tout ce qui est intéressant (en particulier scientifiquement) l'est de toute manière en anglais. L'autre que sur le web une large majorité de l'information disponible est en anglais.

La réalité est tout autre, la généralisation de l'usage de l'Internet dans le monde a amené une multiplication des sites dont tout ou partie sont dans la langue du pays. Certains grand pays comme la Chine ont vu leur part devenir très importante. De plus, pour les sites qui ont des pages en anglais donnants des informations de nature

générale, on s'aperçoit que si l'on explore le site en profondeur jusqu'à accéder à des documents entiers (donc très informationnels), ces documents sont souvent seulement dans la langue locale. Les informations les plus utiles sont en fait souvent trouvées dans ces documents.

Une étude menée en 1998 sur des publications dans des revues russes non traduites, auprès de chercheurs russes émigrés, sur leur sujet de prédilection, a montré qu'un tiers seulement des informations étaient connues d'eux, pour un autre tiers, il connaissaient les équipes mais pas les travaux et pour le dernier tiers ils ignoraient complètement les équipes et les travaux. Dans l'ensemble les travaux non connus se sont révélés très intéressants.

Les européens se sont très tôt intéressés à l'information exprimée en d'autre langues que l'anglais du fait que les systèmes essentiellement statistiques n'étaient pas bien adaptés à la recherche d'information dans les bases en texte intégral pour leur propres langues ce qui a amené à réaliser très tôt l'introduction de traitement de la langue naturelle dans les systèmes. Bien entendu, certaines langues posent plus de problèmes que d'autres et l'introduction de traitements linguistiques s'est avérée plus ou moins indispensable. Le premier projet européen ayant pour but l'interrogation en une langue de documents écrits dans plusieurs langues (EMIR European Multilingual Information retrieval ESPRIT 5312) date de 1990 mais il s'appuyait sur une technologie de traitement linguistique qui était capable depuis de nombreuses années de traiter français, anglais et arabe.

L'intérêt pour des langues plus éloignées des langues européennes comme l'arabe, le chinois, le malais ou le japonais a largement justifié la démarche qui consiste à réaliser des systèmes linguistiques ouverts vers des mécanismes linguistiques très diversifiés.

Aux Etats-Unis, la prise de conscience de l'intérêt du multilinguisme date du milieu des années 1990. Un rapport de DARPA souligne le fait que la barrière linguistique "puts the United States at a distinct STRATEGIC DISADVANTAGE in the international competition for Information Dominance, because technical and military personnel in other nations have far better skills in English than we have in their languages".

Cette constatation a amené la politique américaine à largement financer depuis cette date la recherche sur l'information multilingue et donc l'intérêt pour les technologies linguistiques dans la recherche d'information.

Aujourd'hui la veille stratégique, commerciale, scientifique, technique et de sécurité est un consommateur important de ces technologies. De plus la mondialisation qui permet à un producteur de vendre au travers Internet dans le monde entier demande de plus en plus aux systèmes d'être capables de s'adresser au client dans sa propre langue sous peine d'un manque d'efficacité.

Enfin, la multiplication de sociétés multinationales due aux regroupements d'acteurs de différents pays nécessite la disponibilité de systèmes capables de partager la connaissance quelle que soit la langue dans laquelle elle est exprimée.

2. Système multilingue – système interlingue

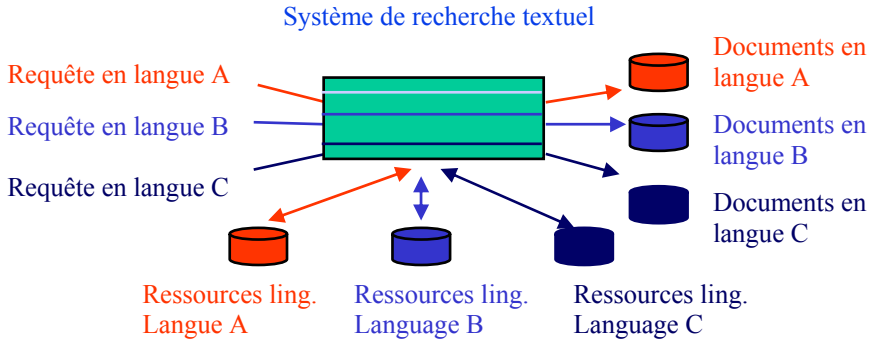
Avant d'entrer dans une description des difficultés et des techniques, il convient de clarifier quelques points de vocabulaires.

Un système multilingue est un système capable de traiter plusieurs langues avec le même programme mais en changeant de ressources linguistiques à chaque changement de langue. Les ressources linguistiques peuvent être des automates de tokenisation, des dictionnaires, des grammaires, des règles de reconnaissance d'entités nommées.

Cela demande pour être réalisé de manière optimale, de prendre en compte cet aspect dès la conception du système car il est toujours très difficile de faire évoluer un système conçu pour une langue pour le faire fonctionner sur une autre surtout si elle possède des caractéristiques très différentes.

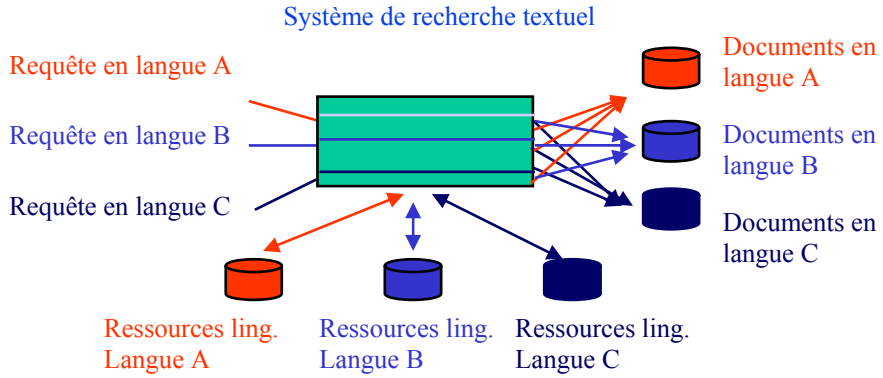
Cela amène aussi quelquefois à repenser la manière dont la langue est décrite par les linguistes. En effet, pour des raisons économiques de facilité de maintenance, il est intéressant de minimiser les algorithmes et donc de traiter de manière homogène des phénomènes linguistiques qui semblent sans liens mais qui d'un point de vue algorithmique peuvent être traités de la même manière. Je citerai comme exemple le fait que nous traitons avec le même algorithme la correction automatique d'erreurs d'accents en français, l'ajout des voyelles en arabe et la correction de mots russes qui contiennent des lettres latines de même forme que les lettres cyrilliques.

Dans le contexte d'une recherche d'information un système multilingue correspond au schéma suivant :



Un système de recherche interlingue (crosslingual en anglais) est un système multilingue qui possède en plus la faculté, à partir d'une requête en une seule langue, de fournir des réponses trouvées dans des documents qui sont dans n'importe laquelle des langues prises en compte dans le système.

Cela peut être visualisé par le schéma suivant :



3. Difficultés de réalisation des systèmes multilingues

Même si le système de recherche d'information est un système statistique, il est important pour que cette statistique ait un sens que les unités sur lesquelles portent les comptages soient significatives du point de vue du sens.

Il est donc important que le texte soit bien segmenté en mots et en particulier que les mots significatifs soient bien isolés.

De plus, il ne s'agit pas de compter des chaînes de caractères mais des concepts et donc :

- Identifier que différentes chaînes peuvent identifier les mêmes concepts.
- Déterminer en fonction du contexte le sens de mots polysémiques.

Il faut aussi pouvoir reconnaître les mots composés qu'ils soient des expressions figées comme "chemin de fer" ou des concepts exprimés par plusieurs termes comme "traitement du langage naturel".

3.1 Segmentation des textes en candidats mots (tokenisation)

Cette opération est réalisée au moyen d'automates qui examinent la syntaxe locale de caractères.

Par exemple le caractère espace permet de délimiter des mots dans un certain nombre de langues. Toutefois si on considère l'expression figée "chemin de fer" on s'aperçoit que le mot contient des espaces. Dans un premier temps ses éléments seront considérés comme des mots indépendants.

Le tiret pose aussi des problèmes du fait qu'un mot comme "abat-jour" contient un tiret mais que "prenez-le" doit être considéré comme deux mots.

Les choses deviennent plus compliquées dans des langues où certains mots peuvent être concaténés.

Cela peut être le cas des mots de plein sens comme les mots composés en allemand.

⇨ Exemple en allemand

Gunstbezeigung

Doit être décomposé en "Gunst" (faveur) et "Bezeigung" (marque, témoignage) si l'on veut par exemple traiter une requête portant sur "Gunst" (faveur).

Cela peut être le cas pour un mot de plein sens auquel sont concaténés des mots outils (préposition, articles, pronom, conjonction, etc.).

C'est le cas de langues dites agglutinantes.

⇨Exemple en turc

Izmirde où Izmir est la ville d'Anatolie (Smyrne) qui est concaténé à la préposition "de" qui signifie "à").

⇨Exemple en arabe

وبرامجهم (Wa Baramijouhoum) <=> et leurs programmes

qui doit être décomposé de la manière suivante :

وبرامجهم

------|----------------|--

Le premier mot (de la droite vers la gauche) est la conjonction (wa = et).
Le mot central qui est le seul mot significatif est (Baramij = programmes).
Le troisième mot est le possessif (houm = leurs).

On comprend bien que dans ce cas, même un palliatif comme l'opérateur de troncature n'aurait pas pu permettre de trouver le concept de programme.

Certaines langues ne mettent pas de blanc entre les mots. Seule la ponctuation peut permettre de limiter la chaîne à traiter. Cela signifie que le découpage du texte donnera non des mots mais en général toute une phrase.

Quelquefois dans ce type de langue, la présence de plusieurs jeux de caractères peut permettre de limiter la longueur des chaînes à traiter.

C'est le cas par exemple en japonais où le passage de Kana (caractère syllabique) au caractère Kanji (chinois) permet d'obtenir une segmentation mais qui conserve en général plusieurs mots dans le segment.

La segmentation des mots concaténés peut être réalisée à l'aide du dictionnaire de formes de la langues (cas des mots composés en allemand) ou à l'aide du dictionnaire de formes de la langues et de dictionnaires de particules (proclitiques et enclitiques) pouvant être concaténées devant ou derrière le mot significatif.

Les découpages sont souvent ambigus. Il est donc nécessaire de disposer de règles d'incompatibilité qui limitent les propositions de découpage. Les ambiguïtés restantes doivent être traitées par l'analyse générale de la phrase.

⇨ Exemple en arabe

العملة

----|----------|---

Le premier mot est l'article défini, le deuxième est le mot travail, le troisième est un pronom possessif. On peut éliminer l'interprétation pronom possessif par le fait que le proclitique est une article défini (et le mot central un nom).

العملة

----|----------|--|--

Le découpage correct donne une particule interrogative, une préposition, le mot "travail" et un possessif

⇨ Exemple en chinois

Le cas du chinois est plus délicat car la tokenisation par syntaxe locale de caractères fournit en général une phrase entière. Le dictionnaire de la langue permet de découper la phrase en mots mais cela donne quelquefois plusieurs milliers de découpages possibles. En effet, un mot chinois peux être représenté par 1, 2 ou 3 idéogrammes mais pour une mot de trois idéogrammes, le premier idéogramme est en général lui même un mot de la langue et les deux premiers aussi. L'analyse syntaxique permet d'en éliminer une grande part mais il faut ajouter ensuite des heuristiques pour encore éliminer des hypothèses.

□ □ □

□ / □ /□ / □□ / □□ /□□/ □

tsunami/ catastrophe / par / sous-marin /tremblement de terre/causer/.

La catastrophe du tsunami a pour origine un tremblement de terre sous-marin.

3.2 Identification des chaînes qui peuvent représenter le même concept

3.2.1 Flexions

Les langues à traiter ont une morphologie plus ou moins complexe. L'anglais est de ce point de vue très pauvre. Il existe peu de variations de formes en fonction des féminins, pluriels, conjugaisons ou rôles syntaxiques (cas).

Le français présente déjà plus de difficultés du fait des variations en genre et en nombre des noms et adjectifs et des nombreux temps et modes des verbes.

Les langues à déclinaison comme l'allemand ou le russe ont encore une plus grande variabilité des mots significatifs.

Dans beaucoup de langues européennes (groupe des langues indo-européennes), la flexions se fait en ajoutant une terminaison avec éventuellement une modification de la fin de la racine. Toutefois certains cas n'obéissent pas à cette règle. On peut citer le cas de l'allemand où certaines flexions du nom modifient la voyelle de la racine (Buch → Bücher). Ou encore l'utilisation d'une particule devant le mot en plus de la terminaison modifiée (le verbe "machen" donne au participe passé "gemacht")

Tout cela milite pour une prise en compte de ces phénomènes une fois pour toute par la constitution de dictionnaires de formes. En effet, même si la langue produit un grand nombre de formes pour un même mot, le nombre de ces formes reste compatible avec les volumes de stockages possibles par l'informatique d'aujourd'hui sur disque mais aussi en mémoire centrale.

Il est donc très intéressant de limiter l'analyse morphologique lors du traitement des textes à une simple consultation de dictionnaire. Cela signifie que, bien entendu, on aura réalisé auparavant une flexion automatique des tous les mots d'un dictionnaire de lemmes pour réaliser un dictionnaire de formes. Ce dernier associe la forme construite automatiquement avec les propriétés linguistiques (partie du discours, genre nombre, temps, mode, personne, cas) et le lemme qui a servi à le produire.

Exemple : Peignes

- verbe peindre deuxième personne du singulier présent du subjonctif
- verbe peigner deuxième personne du singulier présent de l'indicatif
- verbe peigner deuxième personne du singulier présent du subjonctif
- nom masculin pluriel

3.2.2 Variations orthographiques

Dans beaucoup de langues, certains mots ont plusieurs orthographes acceptées par exemple en français on peut écrire "heuristique" ou "euristique" ou en anglais "maharaja" ou " maharajah".

On peut ajouter à cela les expressions complètement équivalentes provenant de la même langue (vélo, bicyclette), de régions différentes (truck, lorry), ou de langues différentes (baladeur, walkman), (chien, klebs).

Il existe aussi des problèmes liés à une mauvaise connaissance des règles d'écriture. Doit-on écrire "nondestructif" ou "non-destructif" ou "non destructif". L'expérience montre que dans les corpus toutes les possibilités sont rencontrées. On doit rappeler qu'un système basé sur de l'ingénierie linguistique doit être robuste. Il doit donc dans la mesure du possible prendre en compte de manière correcte les situations qui ne respectent pas complètement des règles d'écriture.

3.3 Les polysèmes

La polysémie pose des problèmes plus difficiles à résoudre. Il s'agit d'un phénomène très courant dans la grande majorité des langues ou une même graphie représente des concepts différents.

Dans le cadre de la recherche d'information nous nous intéresseront aux différences de sens importantes car il est toujours possible de déterminer pour un mot des usages différents pour lesquels les sens sont légèrement différents. Ces petites différences n'ont pas aujourd'hui une assez grande influence sur les résultats des systèmes qui ont beaucoup de problème plus importants à traiter.

En revanche, on doit essayer de traiter des ambiguïtés qui portent sur des sens qui n'ont aucun rapport immédiat. On peut donner comme exemple en français : pièce (de théâtre) (de monnaie), inspirer (de l'air) (une personne), vol (à la tire) (d'un oiseau).

La résolution de ces ambiguïtés lors de l'indexation permettrait de mieux représenter le sens des documents traités. C'est un problème difficile et qui ne donne pas aujourd'hui dans le cas général un pourcentage suffisant de bonne résolution. C'est un sujet de recherche actuel et on peut espérer une évolution dans un avenir assez proche.

Cela n'empêche pas que le problème soit traité d'une certaine manière lors de l'interrogation par les systèmes qui pratiquent une comparaison pondérée. En effet, il n'est pas absolument nécessaire de déterminer le sens des mots polysémiques. Il suffit d'assurer que le sens du mot polysémique dans la question et dans le document sont les mêmes. Cela est possible dans le contexte de questions qui comportent plusieurs mots et pour les documents jugés les plus pertinents (ceux qui ont le plus de mots en commun avec la question).

En effet, la simple cooccurrence des mots permet d'assurer dans la plupart des cas que les mots polysémiques sont pris dans le même sens dans la question et dans le document. On trouvera dans le chapitre 4 un exemple dans le cadre d'une interrogation interlingue pour lequel ce mécanisme permet de choisir les bonnes traductions.

3.4 Les mots composés

L'intérêt des mots composés pour la recherche d'information n'est plus à démontrer. La cooccurrence dans un même texte des mots "indexation" et "automatique" ne veut pas forcément dire qu'il s'agit d' "indexation automatique".

Parmi les expressions composées de plusieurs mots simples, certaines n'ont de signification que globalement. Ces expressions idiomatiques peuvent être aussi bien des mots outils ("à concurrence de") que des mots de plein sens ("hot dog"). Il est clair que le système de recherche ne doit pas prendre en compte le mot "concurrence" dans ce contexte ni le mot "dog".

Les expressions peuvent être discontinues mais leur reconnaissance globale est indispensable en particulier dans un contexte d'interrogation interlingue car ces expressions ne peuvent se traduire mot à mot.

Un mot comme "switch on" devra être reconnu malgré la discontinuité dans la phrase "I switch the light on." pour comprendre que le concept "éteindre" est présent dans la phrase.

Les autres expressions composées de plusieurs mots simples ont leurs éléments qui conservent leur sens et peuvent souvent être traduits mots à mots.

Toutefois, il est important qu'à concept égal, la représentation interne issue de l'analyse linguistique soit la même. On a tout intérêt à avoir une représentation la plus indépendante de la structure de surface de façon à permettre une traduction mot à mot plus facile et une bonne prise en compte des variantes dans la même langue.

"Management of water resources" et "water resources management" doivent pouvoir être représentés par le même codage interne.

4. Interrogations interlingues – les approches

Les travaux, visant à mettre en place une interrogation interlingue, se sont regroupés autour de quatre grandes voies de solutions.

La première, promue par Tom Landauer et Susan Dumais [LAN 90] [LIT 98], consiste à utiliser des modèles statistiques sur des bases multilingues dont certains documents ont plusieurs versions linguistiques. L’approche la plus utilisée est celle des LSI (Latent Semantic Indexing).

L'approche LSI consiste à diminuer la dimension des matrices par une suppression des valeurs propres les plus petites a pour conséquence en monolingue

le rapprochement de documents portant sur des sujets semblables ce qui permet grâce à la présence de documents traduits de rapprocher des documents portant sur des sujets proches mais exprimés dans des langues différentes.

Bien entendu, il n'est pas toujours possible de disposer de documents traduits. Dans ce cas, certains auteurs les produisent par traduction automatique. Les erreurs de traduction que réalisent les systèmes actuels semblent ne pas trop perturber l'approche par LSI.

Un inconvénient de cette approche par LSI est la difficulté d'expliquer à un utilisateur les raisons qui ont permis au système de considérer un document comme pertinent du fait que les reformulations réalisées par les systèmes ne sont pas explicites.

La deuxième approche consiste à traduire les documents ou les questions à l'aide d'un système de traduction automatique [EST97]. En général, on traduit plutôt les questions et on traduit à la demande les seuls documents jugés pertinents par l'utilisateur.

On se ramène dans ce cas à une interrogation monolingue. Cette approche est efficace tant que le système de traduction fournit la bonne traduction de chaque concept. Mais dans le cas de polysèmes par exemple « pièce », « assurance », « avocat », si le système se trompe dans la traduction d'un mot important de la question la recherche sera inefficace (silence).

> **Exemple** de question pour une base de données image traduite par le système SYSTRAN de traduction automatique (version démonstration en ligne) :
>
> *Un avocat sur une table : a lawyer on a table*
>
> On a évidemment peu de chance de trouver la photo même si l'interprétation du traducteur automatique n'est pas à proprement parler une erreur.

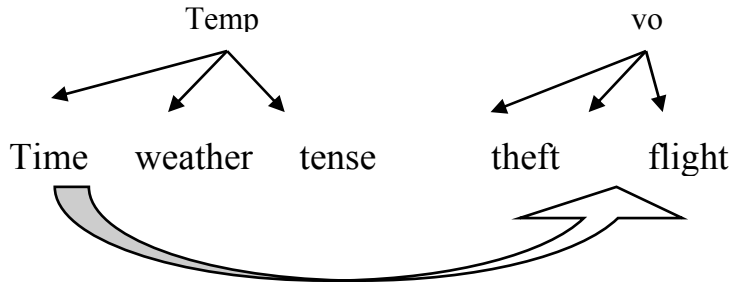
La troisième approche consiste non pas à traduire toute la requête, mais à proposer toutes les traductions possibles de chacun des concepts de la requête.

En effet la production d'une traduction complète de la requête comme le ferait un système de traduction automatique n'est d'aucune utilité pour le programme de recherche.

En revanche, le fait de tenter toutes les traductions possibles au lieu d'une seule qui peut se révéler fausse permet de diminuer le silence mais peut provoquer du bruit. En fait, cette approche permet dans le cas de questions qui portent sur plusieurs mots de lever les ambiguïtés en utilisant la base comme filtrage sémantique à la condition que la base contienne une réponse à la question.

En effet la cooccurrence d'une traduction de chacun des concepts de la question éventuellement liée par les mêmes relations syntaxique est un gage de bonne traduction.

Par exemple :



Si on interroge par "temps de vol" et qu'on traduit mot à mots on obtiens 9 solutions possibles. En fait, la base qui porte sur la physique n'a trouvé que "time of flight" mots composé validant ainsi la traduction de "temps" par "time" et "vol" par "flight".

Une quatrième approche consiste à traduire les documents et la requête dans une représentation indépendante de la langue. Il s'agit ici d'une reprise du vieux rêve des acteurs de la traduction automatique qui permettrait de résoudre un problème matériel très important. En effet, il n'est pas envisageable d'un point de vue économique de constituer des ressources permettrant un transfert direct entre chaque langue de la terre. L'idée du langage pivot consiste donc à se limiter à élaborer des ressources et des traitements pour chaque langue. Cela permettrait de convertir le langage source en langage pivot et ensuite de faire une génération en langage cible.

Cette approche n'a pu aboutir jusqu'ici mais des approches partielles peuvent être aujourd'hui envisagées.

Une première solution utilisée pour la recherche d'information consiste à attribuer pour chaque texte un ou plusieurs mots-clés (ou thème) pris dans une liste fermée. Ces mots clés ont été choisis pour représenter des concepts inambigus. Bien entendu, on perd la possibilité de poser des questions très précises et avoir des réponses très précises. Mais cela permet tout de même de trouver des documents pertinents avec un peu de bruit.

Le traitement consiste donc à comparer le vocabulaire des requêtes avec une représentation du sens de chaque mot-clé et de choisir les mots-clés les plus pertinents. Ce sont des méthodes couramment utilisées en catégorisation. Il faut, bien entendu, disposer d'une description de chaque mot-clé dans chaque langue.

Une tentative qui essaie de traiter les langues de manière plus générale est apportée par l'Université de l'ONU à Tokyo. Il s'agit de représenter le contenu des documents dans un langage de représentation unique quelle que soit la langue. Ce langage UNL (Universal networking language) représente les concepts de manière

unique par des chaînes de caractère en anglais (Universal Words) qui sont rendues inambiguës par un jeu de contraintes. Ces concepts sont liés entre eux par des relations qui expriment là aussi de manière indépendante de la langue les relations que les concepts possèdent dans le texte. On pourra en savoir plus en lisant [HIR 01].

Actuellement, alors que la génération semble fonctionner de manière entièrement automatique, la production de l'UNL à partir du langage source demande encore une part de résolution d'ambiguïtés de manière interactive. Ce dernier point rend cette technologie difficilement utilisable dans le cadre d'une recherche interlingue.

Toutefois, elle constitue un moyen économique de produire des documents dans un grand nombre de langues car une fois le texte traduit en UNL la génération en de nombreuses langues semble possible ce qui rentabilise l'effort fait lors de la seule analyse du texte source

5. Campagnes d'évaluation

En 1992 le NIST (*National Institute of Standard*) sur des crédits de DARPA (*Defense Advanced Research Program Agency*) a mis en place la campagne d'évaluation TREC (*Text REtrieval Conference*). Cette campagne qui se poursuit annuellement permet à toutes les équipes du monde qui désirent se comparer aux autres de traiter les mêmes données et de rendre des résultats à une date déterminée. Les résultats des évaluations sont ensuite discutés dans une conférence dans les locaux du NIST où, seules, les équipes ayant rendu des résultats sont autorisées à participer.

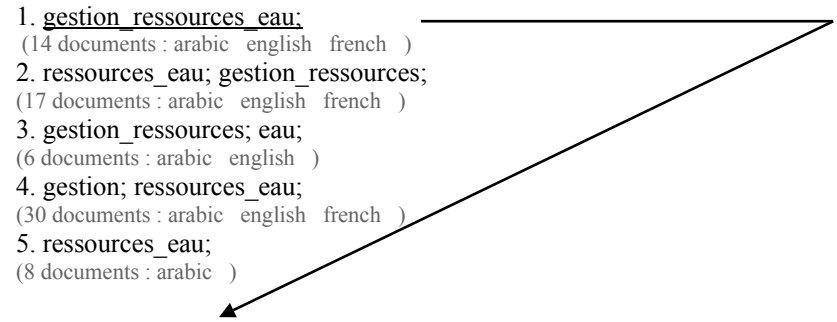
Le NIST au cours du temps a diversifié les types d'évaluations sous formes de pistes (tracks). En 1998 une piste est crée pour l'évaluation des systèmes de recherche interlingue pour les langues français, anglais, allemand et italien. Le paradigme d'évaluation est identique à celui pratiqué en monolingue. Toutefois les volumes mis en jeux sont moins importants.

En 2000 une initiative européenne appuyée par le NIST devient un projet européen sous le nom de CLEF (*Cross-language Evaluation Forum*). Ces campagnes basées sur le paradigme d'évaluation de TREC se poursuivent aujourd'hui et intègrent un nombre croissant de langues.

6. Exemples d'interrogation

Nous illustrerons ce qui précède par un exemple d'interrogation sur une base trilingue (français, anglais et arabe) réalisée dans le cadre du projet européen ALMA (Arabic Language Multilingual Application). La base porte sur l'eau, l'écotourisme et l'environnement.

Requête : gestion des ressources en eau



Classe de pertinence n° 1 : gestion_ressources_eau;

1. القمة العالمية المعنية بالتنمية المستدامة
language : arabic ;
... الحاجة إلى إقامة آلية حكومية دولية إجراء الحوار من أجل السياسات المتعلقة بالتحديات التي تواجه إدارة موارد المياه في أفريقيا. ...

2. منتدى إدارة الطلب على المياه
language : arabic ;
... أبو شرار وبطيخي؛ إدارة موارد المياه في ظل الطلب التنافسي من القطاعات، دراسة حالة من الأردن؛ مجلة "ووتر إنترناشيونال"، المجلد 27، رقم ...

3. قضية المياه في عملية السلام الفلسطينية الإسرائيلية
language : arabic ;
... الفلسطينية في الضفة الغربية بحصة من مياه نهر الاردن . ويضيف أن الطرفين اتفقا على التعاون في ادارة موارد المياه وانظمة الصرف في الضفة الغربية ، مع الالتزام المتبادل للسلطات والمسئوليات في مناطقهما كل على حدة. ...

4. TECHNICAL SYNTHESIS INTERNATIONAL WATERSHED AGREEMENTS PRINCIPLES AND APPLICATIONS
language: english;
... 114 organizations or councils have been created for allocation and management of water resources [3]. ...

5. TECHNICAL SYNTHESIS WATER RIGHTS MARKETS: PRINCIPLES AND RELEVANCE
language: english;
... In Globalization and Water Resources Management: The Changing Value of Water. ...

The Southern African Development Community (SADC) is also integrating the water resource policy at the regional level and in 1995 signed a Protocol on a shared watercourse system. [10] INTER-STATES AGREEMENTS More than 286 treaties are in force, but they only concern 61 watercourses [2]. 114

organizations or councils have been created for allocation and management of water resources [3]. Hammer et Wolf collected 145 treaties signed after 1870 excluding those that are only for navigation, borders and fishing rights in the Transboundary Freshwater Dispute Database [12][13].

Le traitement de la requête précédente est le suivant :

1. Analyse de la requête dans la langue spécifiée (ici le français). L'analyse produit des concepts normalisés sous forme de mots simples ou de mots composés. La normalisation des mots simple porte au moins sur une lématisation mais prend en compte aussi les variantes orthographiques. La représentation des mots composés est sous forme profonde et utilise les mots simples normalisés.
2. Les concepts ainsi identifiés par les mots normalisés sont soumis à une reformulation monolingue en français et une comparaison avec les index des documents français est réalisée. De même, ces concepts sont soumis à une reformulation bilingue français-anglais et comparée aux documents en anglais. Les concepts sont soumis à une reformulation bilingue français-arabe et comparée aux documents en arabe.
3. Chaque interrogation (vers le français, vers l'anglais, vers l'arabe) donne un résultat monolingue sous forme de documents regroupés par classes de pertinence, classes qui sont triées par ordre de pertinence décroissante.
4. Les trois réponses sont fusionnées pour créer un nouvel ensemble de classes. Le poids informationnel des concepts est recalculé dynamiquement comme si la base n'avait qu'une seule langue et que les polysémies avaient été levées.
5. La réponse est fournie sous forme d'une liste de classes unique exprimée dans la langue d'interrogation (ici le français) avec mention du nombre de documents dans chaque langue.
6. Pour chaque classe, on peut visualiser certaines informations comme le titre et une partie informationnelle du document.
7. En sélectionnant un document on va directement à la partie la plus informationnelle et on peut naviguer de partie informationnelle en partie informationnelle.

7. Conclusion

La prise en compte d'un nombre important de langues est possible à condition de disposer d'une bonne vision d'ensemble des différentes difficultés qui seront rencontrées et cela dès la conception du système.

Pour ce qui concerne la recherche interlingue, l'approche par dictionnaire bilingue semble la plus efficace. Toutefois, elle présente l'inconvénient de nécessiter

autant de dictionnaires bilingues que de couples de langues. Cela est impossible dans la pratique. En fait, on considérera une ou deux langues pivots (pour nous le français et l'anglais) et lors d'une interrogation pour un couple de langue inexistant, on pratique une double traduction. Cela donne un peu de bruit mais fonctionne. Une amélioration possible en cas de deux langues pivots est de faire l'intersection des résultats obtenus par les deux doubles traductions pour réduire les erreurs.

8. Références bibliographiques

[AND02] André J. *et al.*, « Unicode, écriture du monde ? », éditeurs scientifiques André J. et Hudrisier H., Collection « document numérique », vol. 6, n° 3-4/2002, Hermès/Lavoisier, Paris, 2002.

[BRA02] Braschler M., Peters C., "CLEF 2002 Methodology and Metrics", Advances in Cross-Language Information retrieval, Third Workshop of the cross-Language Evaluation forum, CLEF 2002, Rome, Italy, 19-20 septembre 2002, Articles révisés, Lecture notes in Computer Science 2785, Springer 2003.

[BRA99] Braschler M., Peters C., Schauble P., "Cross-Language Information retrieval (CLIR) Track Overview", Proceeding of the Eigth Text Retrieval Conference, Gaithersburg, Maryland, USA, 17-19 November 1999.

[EMI94] EMIR Consortium, "Final report of the EMIR project", Commission de l'Union Européenne, rapport final du projet EMIR ESPRIT 5312, Luxembourg, 1994.

[EST97] Estival D., "Machine translation and Multi-Lingual Text Processing", Cross-language Text and Speech Retrieval,Papers from the AAAI Spring Symposium, David Hull and Oard Co-Chairs, Technical report SS-97-05, 219 p., may 1997, Stanford University, California.

[FLU98] Fluhr C. *et al.*, "Distributed Cross-langual Information Retrieval", chapter in "Cross-language Information Retrieval", Editeur scientifique Grefenstette G., Kluwer Academic Publisher, Boston, 1998.

[GRE98] Grefenstette G. *et al.*, "Cross-language Information Retrieval", Editeur scientifique Grefenstette G., Kluwer Academic Publisher, Boston, 1998.

[HIR01] Hiroshi U., "The Universal Networking Language Beyond machine translation", International Symposium on Language in Cyberspace, 26-27 September 2001, Seoul, Korea.

[LAN90] Landauer T.K., Littman M.L.., "fully automatic cross-language document retrieval using latent semantic indexing", Proceeding of the 6th Annual Conference of UW Centre for the New Oxford English Dictionary and Text Research, Center for the New OED and Text Research, Waterloo, Ontario, 1990.

[LIT98] Littman M.L., Dumais S.T., Landauer T.K., "Automatic cross-language information retrieval using latent semantic indexing", in G. Grefenstette, editor, Cross language Information retrieval, chapter 5, Kluwer Academic Publishers, Boston, 1998.

[PET03] Peters C., Borri F. (sous la dir. de), "Cross Language Evaluation Forum, Results of the CLEF 2003 Cross-Language System Evaluation Campaign", Working notes for the CLEF 2003 Workshop, Trondheim, Norway, 21-22 August 2003.

[SEM98] Semenova V., Fluhr C., Golubchik S., Sushkova G., Veselago V., “What treasures are hidden in Russian scientific magazines and how to find them?”, (Linguistic technologies improve scientific cooperation) / Conference "ELSNET in Wonderland", Soesterberg, Netherlands, 25-27 Mars 1998.