

Discours et TAL

Annotation de structures multi-échelle dans le projet ANNODIS

M-P. Péry-Woodley, C. Fabre,
J. Rebeyrolle, L. Tanguy
3 novembre 2009

Plan

- I. Introduction
- II. Principes de l'annotation macro
- III. Cadre expérimental

I. Discours et TAL: introduction

1. Un constat : le « discourse-NLP gap »

Cf. Synthèse de Moore & Wiemer-Hastings 2003: “Discourse in Computational Linguistics and Artificial Intelligence” (analysée dans introduction à TAL 47/2):

- Section 1: Computational theories of discourse structure and semantics (DRT, Grosz and Sidner’s Theory, RST, SDRT) : 18 p.
- Section 2: Generating coherent discourse: 15 p.
- Section 3: Current discourse applications: 4 p.

Section 3 entièrement consacrée à des techniques surfaciques et statistiques, strictement aucun lien avec les modèles présentés précédemment...

Cf. Marcu (2006) sur « automatic discourse parsing »:

“a young research area that finds itself in a somewhat awkward position”

due to:

- a) lack of maturity of text/discourse linguistics
- b) there are many simpler problems which are perceived as more likely to have immediate impact on system performance

Cf. Nazarenko 2005: *Sur quelle sémantique reposent les méthodes automatiques d'accès au contenu textuel ?*

- Accéder au *contenu* vs. accéder au *sens* des textes
- « de quoi parle le document? » vs. « que dit le document? »
- Méthodes diverses combinant différents niveaux d'analyse
- « Il n'existe pas encore de modèle permettant de donner une représentation unifiée de ces différents niveaux d'analyse, de comprendre le rôle et la contribution de chacun dans le résultat d'analyse global, et partant, de définir une architecture d'analyse sémantique. »

Discours et TAL: introduction (fin)

A partir de ce constat... nécessité de rapprocher les modèles des données

- 2. Une démarche** : annotation systématique de structures discursives dans un corpus diversifié
 - pour décrire
 - pour expérimenter

Préambule: approche ascendante vs. descendante

Analyse ascendante

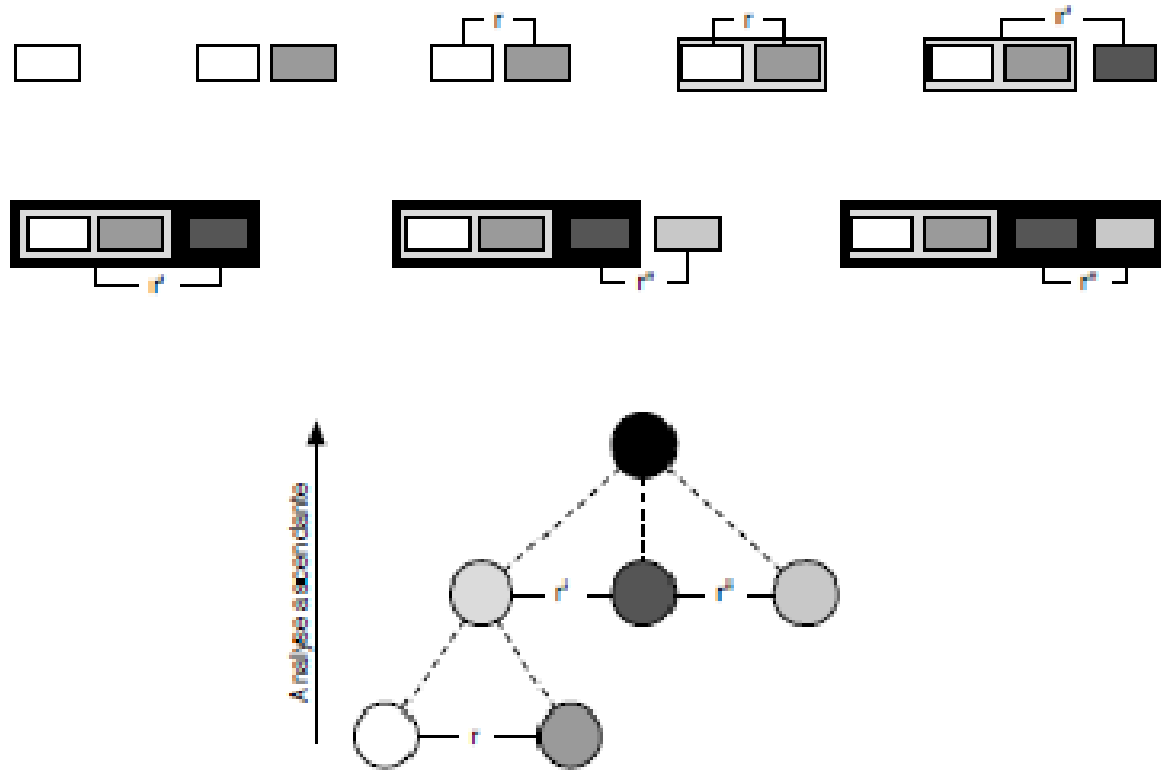


FIG. 5.19 – Analyse ascendante

Analyse descendante

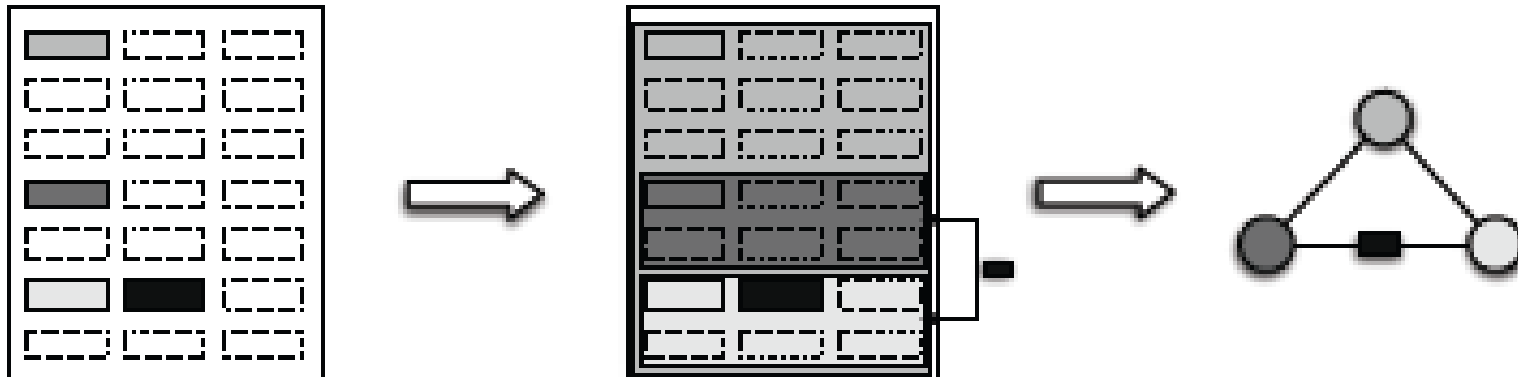


FIG. 5.20 – Analyse descendante et sporadique

Annotation discursive: approches

- Approches descendantes
 - Structure de document
 - Cadres
 - Chaînes
 - Structures énumératives
- Approche ascendante
 - Relations de discours

Sur un exemple...

Structure de document

Structure 1:
Objets textuels, MFM

Titre 1^{er} niveau

Section

Paragraphes

Titre 2nd niveau

Structure énumérative

Cf. MAT, Luc & Virbel

4. La «communale»

De la fin du siècle dernier jusqu'aux années 1950, l'école primaire a été le pilier du système scolaire français. Elle inculquait les connaissances de base, lire, écrire et compter, qui serviraient toute la vie. Elle avait aussi pour mission de former les citoyens de la République. Elle délivrait le certificat d'études qui, pour le plus grand nombre, attestait de la réussite des études et marquait l'entrée dans le monde du travail.

Les sessions du certificat d'études n'ont plus lieu. Nombre d'écoles communales de campagne ont été fermées, ou vont l'être, faute d'enfants à accueillir. Et l'école primaire n'est plus que le premier degré de scolarités ayant maintenant pour objectif le collège puis le lycée. La loi d'orientation de 1989 l'organise en cycles, depuis la maternelle jusqu'à la dernière année des études élémentaires; et les instituteurs font dorénavant partie du corps des professeurs.

La «communale» de Jules Ferry et de la Troisième République appartient au passé.

4.1. Des effectifs en diminution

À la rentrée 1992-93, l'enseignement public et privé du premier degré a accueilli 6 610 000 élèves:

- 2 550 000 dans l'enseignement préélémentaire;
- 3 985 000 dans l'enseignement élémentaire;
- 75 000 dans les classes d'initiation, d'adaptation et d'enseignement spécial.

Après avoir beaucoup augmenté dans les années 1950 et 1960, les effectifs du premier degré diminuent lentement depuis une vingtaine d'années. Les raisons en sont d'abord démographiques. La chute de la natalité a été telle que, malgré l'augmentation de la population, le nombre absolu des naissances a fortement diminué, et par voie de conséquence les effectifs de la population scolarisable en primaire. C'est à la rentrée 1977 que les écoles élémentaires ont eu à faire face à l'arrivée à l'âge de la scolarité obligatoire de la cohorte d'enfants la plus nombreuse que le pays ait jamais connue jusqu'alors (le

**P 79 de
l'Atlas de la
France
Scolaire,
un extrait de
document**

Cadres de discours

4. La «communale»

Structure 2 :

cadre de discours

Cf. Charolles

De la fin du siècle dernier jusqu'aux années 1950, l'école primaire a été le pilier du système scolaire français. Elle inculquait les connaissances de base, lire, écrire et compter, qui serviraient toute la vie. Elle avait aussi pour mission de former les citoyens de la République. Elle délivrait le certificat d'études qui, pour le plus grand nombre, attestait de la réussite des études et marquait l'entrée dans le monde du travail.

Les sessions du certificat d'études n'ont plus lieu. Nombre d'écoles communales de campagne ont été fermées, ou vont l'être, faute d'enfants à accueillir. Et l'école primaire n'est plus que le premier degré de scolarités ayant maintenant pour objectif le collège puis le lycée. La loi d'orientation de 1989 l'organise en cycles, depuis la maternelle jusqu'à la dernière année des études élémentaires; et les instituteurs font dorénavant partie du corps des professeurs.

La «communale» de Jules Ferry et de la Troisième République appartient au passé.

4.1. Des effectifs en diminution

Chaînes

4. La «communale»

De la fin du siècle dernier jusqu'aux années 1950, l'école primaire a été le pilier du système scolaire français. Elle inculquait les connaissances de base, lire, écrire et compter, qui serviraient toute la vie. Elle avait aussi pour mission de former les citoyens de la République. Elle délivrait le certificat d'études qui, pour le plus grand nombre, attestait de la réussite des études et marquait l'entrée dans le monde du travail.

Les sessions du certificat d'études n'ont plus lieu. Nombre d'écoles communales de campagne ont été fermées, ou vont l'être, faute d'enfants à accueillir. Et l'école primaire n'est plus que le premier degré de scolarités ayant maintenant pour objectif le collège puis le lycée. La loi d'orientation de 1989 l'organise en cycles, depuis la maternelle jusqu'à la dernière année des études élémentaires; et les instituteurs font dorénavant partie du corps des professeurs.

La «communale» de Jules Ferry et de la Troisième République appartient au passé.

4.1. Des effectifs en diminution

Structure 3 :

Chaînes

Cf. Cornish

**Cohésion
(reference,
lexical
cohesion)**

**Cf. Halliday &
Hasan**

Relations de discours

Structure 4 : relations de discours (Cf. Rhetorical Structure Theory)

4. La «communale»

De la fin du siècle dernier jusqu'aux années 1950, l'école primaire a été le pilier du système scolaire français. Elle inculquait les connaissances de base, lire, écrire et compter, qui serviraient toute la vie. Elle avait aussi pour mission de former les citoyens de la République. Elle délivrait le certificat d'études qui, pour le plus grand nombre, attestait de la réussite des études et marquait l'entrée dans le monde du travail.

Les sessions du certificat d'études n'ont plus lieu. Nombre d'écoles communales de campagne ont été fermées, ou vont l'être, faute d'enfants à accueillir. Et l'école primaire n'est plus que le premier degré de scolarités ayant maintenant pour objectif le collège puis le lycée. La loi d'orientation de 1989 l'organise en cycles, depuis la maternelle jusqu'à la dernière année des études élémentaires; et les instituteurs font dorénavant partie du corps des professeurs.

La «communale» de Jules Ferry et de la Troisième République appartient au passé.

4.1. Des effectifs en diminution

Relation d'élaboration

Relation de contraste

Relation de résumé

Relation d'élaboration avec le reste de la section

Annotation discursive : des corpus annotés discursivement

- Penn Discourse TreeBank: 1808 articles étiquetés du Wall Street Journal extraits du Penn TreeBank
- RST TreeBank: 385 articles, même origine, modèle du discours différent

Annotation de relations du discours dans le PDTB

Les connecteurs et leurs arguments:

- **Connecteur explicite**: connecteur explicitement présent dans le texte, qui établit une relation de discours entre deux arguments.
- **Connecteur implicite**: établit une relation qui n'est pas réalisée explicitement dans le texte, et qui doit être inférée par le lecteur.
 - Dans ce cas, adjonction au nœud correspondant au connecteur d'un trait contenant le connecteur qui semble le mieux exprimer la relation + sa classe sémantique.
- Argument d'un connecteur = expression d'un objet abstrait en relation (implicite ou explicite) avec une autre expression du même type.
 - cas les plus simples: argument = proposition.
 - étiquetage des arguments: Arg1 et Arg2.

Annotation de relations de discours dans le PDTB (connecteurs explicites et implicites)

*She hasn't played any music **since** the earthquake hit.*

*Third-quarter sales in Europe were exceptionally strong, boosted by promotional programs and new products – **although** weaker foreign currencies reduced the company's earnings.*

The small, wiry Mr. Morishita comes across as an outspoken man of the world.

IMPLICIT=WHEN IMPLICIT=FOR

EXAMPLE (Stretching his arms in his silky white shirt and squeaking his black shoes) **he lectures a visitor about the way to sell American real estate and boasts about his friendship with Margaret Thatcher's son.**

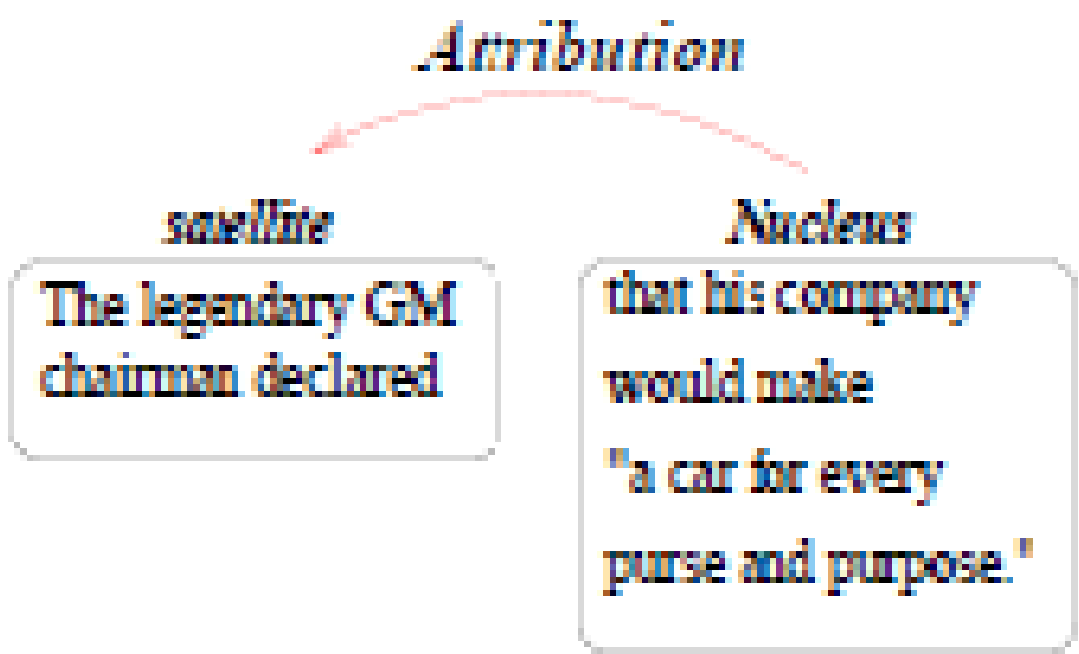
Annotation de l'attribution dans le PDTB

Les attributions

- Qui dit quoi?: « ... relation d'appartenance entre des objets abstraits et des individus ou agents ».
 - Attribuer des croyances/assertions exprimées dans le texte à l'agent (aux agents) qui les exprime(nt).
 - But: classification de la factualité des objets abstraits associés aux arguments des relations de discours et à la relation elle-même.
- Les traits de l'attribution
 - Source: rédacteur (**Writer**) vs. autre (**Other**)
 - Factualité: fait (**Fact**) vs. non-fait (**Non-fact**)
 - Polarité: positive (**Positive**) vs. négative (**Negative**)

RST Treebank

Attribution



satellite

The legendary GM
chairman declared

Nucleus

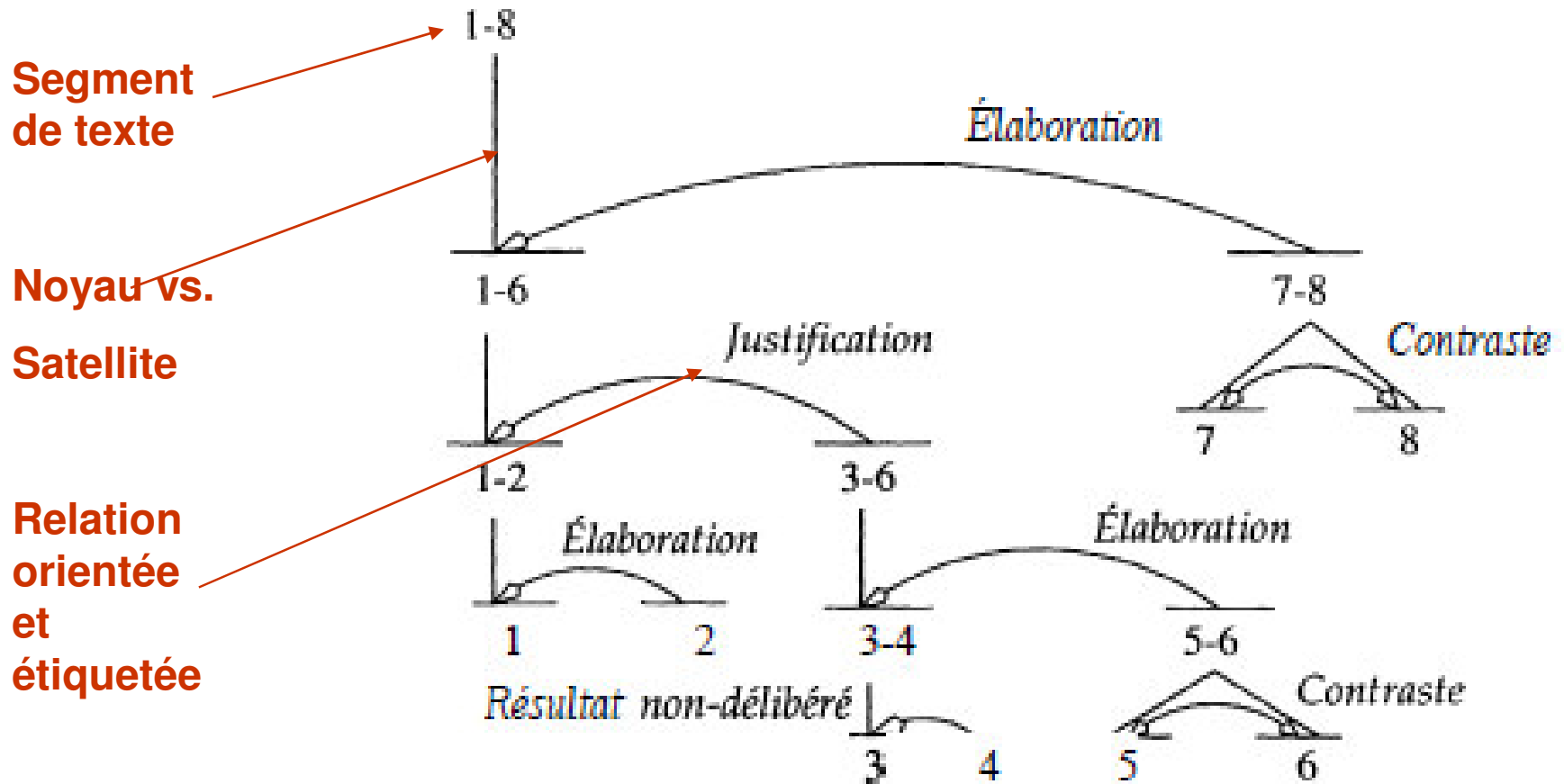
that his company
would make
"a car for every
purse and purpose."

RST Treebank

- Annotations selon Rhetorical Structure Theory (Mann & Thompson 1988):
 - Relations asymétriques noyau-satellite
 - 2 types de relations
 - Subject matter (cause, condition, élaboration,...)
 - Presentation (antithèse, arrière-plan, motivation,...)
 - Annotation fondée sur marqueurs de surface

Voir <http://www.isi.edu/~marcu/discourse/>

Représentation RST d'un texte



Annoter des corpus pour quoi faire?

- Entraînement d'analyseurs
- Construction de grammaires
- Matériau de base pour la construction de lexiques
- Evaluation de divers traitements, notamment analyse syntaxique
- Base de données pour la recherche linguistique

Annoter *discursivement* pour quoi faire?

- Indexation en lien avec structures discursives pour EI, QR
- Identification de segments homogènes par rapport à un critère donné (navigation)
- Hiérarchisation de segments; repérage de zones fonctionnelles pour la synthèse automatique
- ...
- Annoter pour expérimenter

Annoter pour expérimenter

- mettre à l'épreuve
 - les modèles (structures et relations)
 - les hypothèses sur le marquage
- approche « data-driven »
 - au-delà de la validation d'hypothèses, faire émerger des motifs, des marqueurs

II. Principes de l'annotation macro

1. L'approche descendante au sein du projet ANNODIS
2. Organisation multi-dimensionnelle du discours
3. Choix d'annotation de deux structures multi-échelles
 - Structure énumérative (SE)
 - Structure à unité référentielle (SUR)
4. Dispositif d'annotation : du pré-marquage à la mise au jour de marqueurs

Le projet ANNODIS : une approche outillée de l'annotation de structures discursives

- Projet financé pour 3 ans par l'ANR
CLLE-ERSS, Toulouse - IRIT, Toulouse - GREYC, Caen
<http://w3.erss.univ-tlse2.fr/annodis>
- Constitution d'un corpus de français écrit enrichi d'annotations concernant le niveau discursif
 - Annotation de structures à différents niveaux de grain
 - Corpus diversifié permettant la prise en compte de réalisations discursives variées
 - Prétraitements automatiques pour guider les annotations
- Exploitation du corpus annoté
 - Une ressource pour les recherches linguistiques sur le discours
 - Une ressource pour le développement d'applications en TAL

Deux approches complémentaires du discours

- Perspective ascendante
 - unités de discours minimales pour construire des structures complexes *via un jeu de* relations de discours
 - Compositionnalité sémantique
- Perspective descendante
 - Texte est considéré dans son entier
 - Recherches sur les indices de discontinuité
 - Identification de configurations d'indices signalant des segments textuels de haut niveau

Perspective descendante : un point de vue global sur le texte

- L'interprétation du contenu propositionnel d'un texte est influencée par le repérage immédiat de structures qui organisent le discours à un haut niveau de grain
- Les textes sont envisagés dans leur dimension de document, comme formant un tout
- La mise en texte est abordée à partir d'une caractéristique incontournable : la séquentialité
 - Deux stratégies :
 - continuation avec ce qui précède
 - rupture/transition (*shift*)

Rapport Avicenne¹

Synthèse

A la veille de l'élection présidentielle, un groupe de réflexion a réuni chercheurs et diplomates ayant été impliqués depuis de nombreuses années, dans les relations entre la France et les pays d'Afrique du Nord et du Moyen-Orient et appartenant à des sensibilités politiques différentes. Leur point commun est leur conviction que la France se doit d'entretenir des relations politiques, économiques et culturelles fortes avec les pays d'une zone particulièrement sensible où elle a des intérêts vitaux. Il leur est apparu opportun de faire connaître leur analyse sur la situation actuelle de cette région et de proposer quelques pistes de réflexion qui pourraient être utiles à nos décideurs. Tel est l'objet du rapport joint à cette note de présentation.

I) Le constat.

Une zone de hautes turbulences vitale pour les intérêts de la France

Le Moyen-Orient et, de façon plus générale, la zone qui s'étend du Maroc à l'Afghanistan a toujours été une région sensible où se sont affrontés l'Orient et l'Occident, mais également les grandes puissances. De 1945 à 1989, elle fut le champ clos de la guerre froide sur fond de conflits israélo-arabes, ce à quoi s'est ajoutée la vision missionnaire de la révolution iranienne. Aujourd'hui, après le 11 septembre et les affrontements militaires qui se sont déroulés en Afghanistan, en Irak puis, tout récemment, au Liban, cette zone se trouve au cœur de turbulences d'une ampleur inconnue jusqu'alors qui représentent non seulement pour les pouvoirs en place mais également pour notre sécurité, au sens large, un défi majeur.

L'Irak est un pays en plein naufrage politique, économique et social ; le Liban est paralysé par le jeu d'influences contradictoires et le spectre de la guerre civile réapparaît ; la sécurité d'Israël a été gravement affectée par son intervention contre le Hezbollah ; la question palestinienne est dans l'impasse ; l'Iran émerge, à travers les communautés chiïtes et l'affirmation de sa souveraineté nucléaire, comme une menace qui affecte la région et au-delà ; des acteurs non étatiques qui recourent parfois à la violence se sont multipliés ; des Etats se sont écroulés ou se délitent dans une sorte de processus de « somalisation » ; la menace terroriste s'est globalement amplifiée.

Face à cette situation, les Etats-Unis, piégés en Irak, sont sur la défensive ; l'Europe est politiquement marginalisée même si son appui à la souveraineté du Liban peut être jugé

¹ Abu Ali Husayn Ibn Sine, connu sous le nom d'Avicenne. Médecin, philosophe et mystique arabo-islamique d'origine iranienne (980 – 1037) son *Canon de la médecine* fut longtemps la base des études médicales tant en Orient qu'en Occident.

positivement ; la Russie est de retour dans une zone traditionnelle d'influence ; l'Iran s'affirme comme une puissance régionale et un acteur incontournable.

Il en résulte que les perspectives d'avenir, si les tendances actuelles devaient se confirmer, sont préoccupantes. Après l'échec de la Pax Americana, on peut craindre des évolutions plus radicales : un ordre islamiste, tout au moins dans certains pays, un chaos généralisé. Au mieux se vérifierait une évolution moins extrême comme un processus de dégradation progressif et modulé selon les pays.

De telles évolutions ne sont pas une fatalité. Pour arrêter l'engrenage de violences dans lequel tombe cette région ou la menace d'une telle évolution, il convient que, du côté des pays occidentaux, se développe une politique plus déterminée et soutenue. La France peut et se doit d'y apporter sa contribution.

II) Des orientations d'action

Les orientations proposées peuvent être regroupées autour de quatre thèmes.

1 – Mieux organiser notre politique étrangère dans la région ce qui passe, notamment, par la mise en place de structures permettant de mieux appréhender les problèmes afin d'aider à une meilleure définition des politiques : création d'une cellule de réflexion et de contacts sur les questions de l'Islam, pilotage interministériel de notre politique en direction du Maghreb, apport aux pays de la région d'une offre médiatique plus dense et mieux adaptée, présence dans les médias locaux, notamment les chaînes satellitaires panarabes.

2 – Accentuer notre coopération avec des partenaires d'influence, notamment en établissant une coopération renforcée avec certains de nos partenaires européens, en poursuivant la normalisation post irakienne de notre relation avec les Etats-Unis, en nous concertant avec la Russie et la Chine, en dialoguant davantage avec le monde arabe et musulman et, tout spécialement, avec l'Arabie Saoudite dont il convient d'appuyer les efforts en vue de réduire les sources de tension.

3 – Manifester notre souci de voir émerger des systèmes démocratiques dans la région en développant une politique d'influence auprès des « forces vives » de la région, en particulier auprès des sociétés civiles et des mouvements islamistes intégrés dans la vie politique locale et s'engageant à renoncer à la violence.

4 – Contribuer plus efficacement à la solution des principales crises régionales, ce qui comporterait les actions suivantes :

- une médiation active entre les parties au conflit israélo-arabe sur la base croisée de garanties données à Israël pour sa sécurité et de l'établissement d'un Etat palestinien ; à cet effet, notre appui au plan de la Ligue Arabe doit s'accompagner de démarches diplomatiques effectives ;

Trois propriétés fondamentales d'un texte

- L'organisation d'un texte est multi-dimensionnelle, le texte est le résultat d'un "*struggle between (...) different forces*" (Enkvist 1985)

La linguistique systémique fonctionnelle conceptualise cette multidimensionalité au moyen de trois fonctions : idéationnelle, textuelle, interpersonnelle

- Les principes d'organisation des textes sont généralement récursifs, et donnent lieu à des structures enchâssées
- La plupart des structures sont multi-échelles, c'est-à-dire qu'on peut les rencontrer à différents niveaux de grain

Exemple d'organisation multi-dimensionnelle

Depuis la fin de la guerre froide, le débat entre spécialistes des relations transatlantiques s'est trop souvent contenté d'osciller entre les bons sentiments et la simplification. Il ne s'est pas suffisamment porté sur l'ampleur des changements de fond rendus inévitables par le changement de système international produit par l'effondrement du régime soviétique. La première tendance, parfois marquée par une frilosité nourrie par la crainte de remettre en cause l'édifice institutionnel issu de la guerre froide, s'est exprimée le plus souvent sous la forme de satisfecits donnés à l'Alliance atlantique pour ses progrès supposés en matière d'adaptation aux conditions de l'après-guerre froide, et parfois sous la forme plus dynamique de projets d'élargissement, géographique et fonctionnel de l'OTAN et de l'Union européenne. Les travaux de la Rand Corporation, et en particulier ceux de Larabee, Asmus, Gompert et Kugler, avaient ainsi contribué en leur temps à lancer le débat sur l'élargissement de l'OTAN à trois pays qui a finalement abouti en 1999.

Plus récemment, la discussion s'était portée sur un éloignement supposé des valeurs sociales entre les deux rives de l'Atlantique, auquel les événements du 11 septembre 2001 ont au moins provisoirement mis fin. Ce débat se poursuit, mais il est maintenant limité à la sphère de l'analyse sociale. En termes de politique étrangère, cette discussion sur la dérive des continents a pris la forme d'une opposition entre l'unilatéralisme de la politique américaine et le multilatéralisme de leurs partenaires européens.

DES VISIONS DIVERGENTES [titre]

Le moindre mérite de l'article de Robert Kagan, dont Commentaire a publié la version française, n'est pas de sortir le débat de cette ornière....

Le débat entre spécialistes des relations transatlantiques

Introduction

Depuis la fin de la guerre froide,

Plus récemment,

Des visions divergentes

Deux structures multi-échelles

1) Les Structures Énumératives (SE)

- Une amorce = un segment qui annonce une énumération et qui peut inclure un énumérathème (élément prospectif spécifiant le critère de co-énumérabilité des items de l'énumération)
- L'énumération = une chaîne d'au moins deux items
- Une clôture = un segment qui conclut l'énumération et qui peut comporter un énumérathème (élément qui condense le contenu des items de l'énumération)

II) Des orientations d'action

Les orientations proposées peuvent être regroupées autour de quatre thèmes **TRIGGER**

1. Mieux organiser notre politique étrangère dans la région ce qui passe, notamment, par la mise en place de structures permettant de mieux appréhender les problèmes afin d'aider à une meilleure définition des politiques : création d'une cellule de **ITEM 1** et de contacts sur les questions de l'Islam, pilotage interministériel de notre politique en direction du Maghreb, apport aux pays de la région d'une offre médiatique plus dense et mieux adaptée, présence dans les médias locaux, notamment les chaînes satellitaires panarabes.

2. Accentuer notre coopération avec des partenaires d'influence, notamment en établissant une coopération renforcée avec certains de nos partenaires européens. **ITEM 2** suivant la normalisation post irakienne de notre relation avec les États-Unis, en nous concertant avec la Russie et la Chine, en dialoguant davantage avec le monde arabe et musulman et, tout spécialement, avec l'Arabie Saoudite dont il convient d'appuyer les efforts en vue de réduire les sources de tension.

3. Manifester notre souci de voir émerger des systèmes démocratiques dans **ITEM 3** en développant une politique d'influence auprès des "forces vives" de la région, en particulier auprès des sociétés civiles et des mouvements islamistes intégrés dans la vie politique locale et s'engageant à renoncer à la violence.

4. Contribuer plus efficacement à la solution des principales crises régionales, ce qui comporterait les actions suivantes . [...]

En conclusion, les turbulences qui affectent le moyen orient ont atteint un niveau de haute intensité qui représente, pour les pays occidentaux et, plus spécialement, **CLOSURE** de grands risques, notamment dans le domaine de la sécurité au sens large du terme : accroissement du terrorisme, perturbations dans notre approvisionnement en hydrocarbures,

Deux structures multi-échelles

2) Les Segments ayant une Unité Référentielle (SUR)

- Segments se caractérisant par le fait qu'une majorité des propositions qui les composent ont pour objet un seul et même référent

5.1. La montée des eaux

Une des conséquences du réchauffement planétaire sur lesquelles s'accordent les scientifiques est une montée du niveau des océans. Deux phénomènes engendrent cette élévation :

- l'augmentation du volume de l'eau due à son réchauffement (dilatation thermique) ;
- l'apport d'eau supplémentaire provenant de la fonte des calottes glaciaires continentales.

Selon le troisième rapport du GIEC, le niveau de la mer s'est élevé de 0,1 à 0,2 m au XX^e siècle. La montée du niveau des eaux est due principalement au réchauffement des eaux océaniques et à leur dilatation thermique. L'effet de la fonte des glaciers ne se ferait sentir qu'à beaucoup plus long terme, et celle des calottes polaires à l'échelle de plusieurs siècles ou millénaires. De même que pour les températures, les incertitudes concernant le niveau de la mer sont liées aux modèles, d'une part, et aux émissions futures de gaz à effet de serre, d'autre part.

L'élévation entre 1993 et 2003 est estimée à 3,1 mm par an (plus ou moins 0,7 mm). L'élévation prévue du niveau de la mer en 2100 est de 18 à 59 cm, selon le 4^e rapport du GIEC. Elle pourrait être de 2 mètres en 2300.

Une montée des eaux de quelques centimètres n'a pas d'impact très visible sur les côtes rocheuses, mais peut avoir des effets très importants sur la dynamique sédimentaire des côtes plates : dans ces régions, qui sont en équilibre dynamique, la montée des eaux renforce les capacités érosives de la mer, et déplace donc globalement l'équilibre vers une reprise de l'érosion qui fait reculer les côtes. La montée du niveau moyen de la mer a ainsi des effets beaucoup plus importants que la simple translation de la ligne de côte jusqu'aux courbes de niveau correspondantes.

Deux structures qui incarnent la complexité de l'organisation discursive

- Elles sont toutes deux multi-échelles
- Elles peuvent s'imbriquer
 - SE : structure récursive
 - Une SUR peut être incluse dans une SE
- Elles organisent toutes les dimensions de la présentation du discours :
 - SUR : structures topicales
 - SE : structures topicales, temporelles, rhétoriques
- Elles sont très fréquentes dans les textes non narratifs

Ex de SUR à grande échelle

3.2. Vie sociale

Einstein a rencontré un grand nombre de personnalités majeures de son époque, dans les domaines scientifique, politique et artistique, et il a laissé une correspondance très riche. Il entretient par exemple longtemps une relation amicale avec la reine Élisabeth de Belgique, avec qui il joue du violon. Il s'est toujours étonné de sa célébrité et de ses effets.

Sa première épouse, Mileva Maric est atteinte de coxalgie, qui la rend boiteuse. C'est aussi une jeune femme brillante, élève du Polytechnicum. Elle tombe enceinte alors qu'ils ne sont pas encore mariés, et elle accouche chez ses parents en Serbie d'une fille, Lieserl. Einstein se montra très dur avec sa compagne suivante, Elsa. Ils faisaient chambre à part et il lui arrivait de lui interdire son bureau, se faisant presque servir : « Je traitais ma femme comme une employée, mais une employée que je ne pouvais pas congédier[réf. nécessaire]. »

Il voit peu son fils Hans-Albert qui, à l'âge adulte, travaille en Californie. La santé mentale de son autre fils, Eduard, se détériore brutalement alors qu'il est âgé de vingt ans, et il doit être interné une première fois à Zurich en 1930. Son père lui rend une dernière visite en 1933. D'abord critique envers la psychanalyse, il refuse que son fils Eduard suive un nouveau traitement psychanalytique[réf. nécessaire], mais il finit par accepter l'essentiel des idées de Sigmund Freud. En 1933, il choisit Sigmund Freud pour publier un échange de lettres intitulé Pourquoi la guerre ?

3.3. Einstein et la religion

Einstein écrit plusieurs textes traitant des relations entre science et religion. Dans son article paru en 1930¹³, Einstein distingue trois formes de religion : la première est due à la crainte et à une incompréhension de la causalité des phénomènes naturels, d'où l'invention d'êtres surnaturels. La deuxième est sociale et morale. La troisième qu'Einstein appelle "religiosité cosmique" est une contemplation de la structure de l'univers. Elle est compatible avec la science et n'est associée à aucun dogme, ni croyance. Einstein déclare être religieux, mais seulement dans ce troisième sens qu'il voit dans le mot religion.

[...]

3.4. Einstein et la philosophie

La philosophie n'est pas l'un de ses domaines de prédilection, mais Albert Einstein marque son intérêt pour la vision de l'humanité que propose Friedrich Nietzsche, et certaines idées présentes dans les réflexions de Spinoza. Néanmoins, il apporte une nouvelle vision du monde moderne par ses travaux scientifiques comme par ses ouvrages non scientifiques. Ainsi, dans son ouvrage Comment je vois le monde publié en 1934, un an après son installation aux États-Unis, Albert Einstein présente sa vision de l'humanité, et pose la question de la place de la science vis-à-vis de l'humanité. Ces travaux ont pu avoir une certaine influence sur des philosophes comme Martin Heidegger ou Jean-Paul Sartre.

Exemple d'imbrication SE / SUR

3.2. L'économie de guerre explication de la sortie de crise ?

L'Allemagne suivit dès le début des années 30 une politique différente des recettes de l'orthodoxie libérale dominante à l'époque. Sous la responsabilité financière de Herr Schacht **elle** se lance dans une politique d'investissement massif, au départ principalement avec des objectifs civils. Galbraith écrira dans son livre sur "la monnaie" que la politique allemande fut à cette époque une politique keynésienne complète avant l'heure. La doctrine de Keynes est en effet qu'il faut rétablir par une politique d'investissement public l'équilibre perdu entre épargne et investissement. C'est de cette époque que date le réseau d'autoroutes allemand (dont l'équivalent en France ne sera construit que trente ans plus tard). Cette politique est menée sans aucune inflation, ce qui vaudra une réputation durable au ministre des finances, malgré son rôle dans l'appareil nazi. Le plein emploi est quasiment revenu avant même qu'Hitler oriente l'économie allemande vers la production militaire, qui d'ailleurs, est largement réalisée ... en Union Soviétique pour contourner les traités. Le Pacte Germano-Soviétique a été précédé par une longue et secrète coopération militaire. Il va de soi que **l'Allemagne sortira de la guerre** ruinée.

La situation est différente **en Italie** où l'exemple allemand n'est suivi que très partiellement et où les aventures coloniales extérieures absorbent une partie importante de l'énergie nationale. **Elle sortira également de la guerre** ruinée.

La France, principalement agricole, subit la crise de plein fouet, les exportations étant pratiquement arrêtées. **Elle** se replie sur son Empire. Les troubles sociaux et politiques qui aboutissent au Front populaire ne permettent pas l'élaboration d'une politique constante. Alfred Sauvy dans son "histoire économique de la France entre les deux guerres" constate que les "quarante heures" bloque la reprise qui commençait à se manifester. L'effort de production militaire est tardif et n'a qu'une influence marginale sur l'activité. **La France sortira de la guerre** pillée et ruinée.

La situation est peu ou prou la même **au Royaume Uni** qui a tenté de revenir à un taux de change en or intenable pour la Livre avant même 1929 et qui a connu une stagnation plus longue que les autres. La politique d'armement ne commence vraiment que très peu de temps avant la guerre et ne peut être considérée comme la méthode qui a permis de sortir de la crise. **Elle sortira de la guerre** victorieuse mais ruinée.

Le Japon connaît une période d'avant guerre très différente des démocraties du fait de son expansionnisme militaire et de l'encadrement rigoureux de la population. **Elle** manque de pétrole pour ses entreprises. La guerre avec les Etats-Unis sera largement provoquée par l'embargo décidé par ce pays sur les exportations pétrolières vers le Japon. Le pays **sortira ruiné par la guerre**.

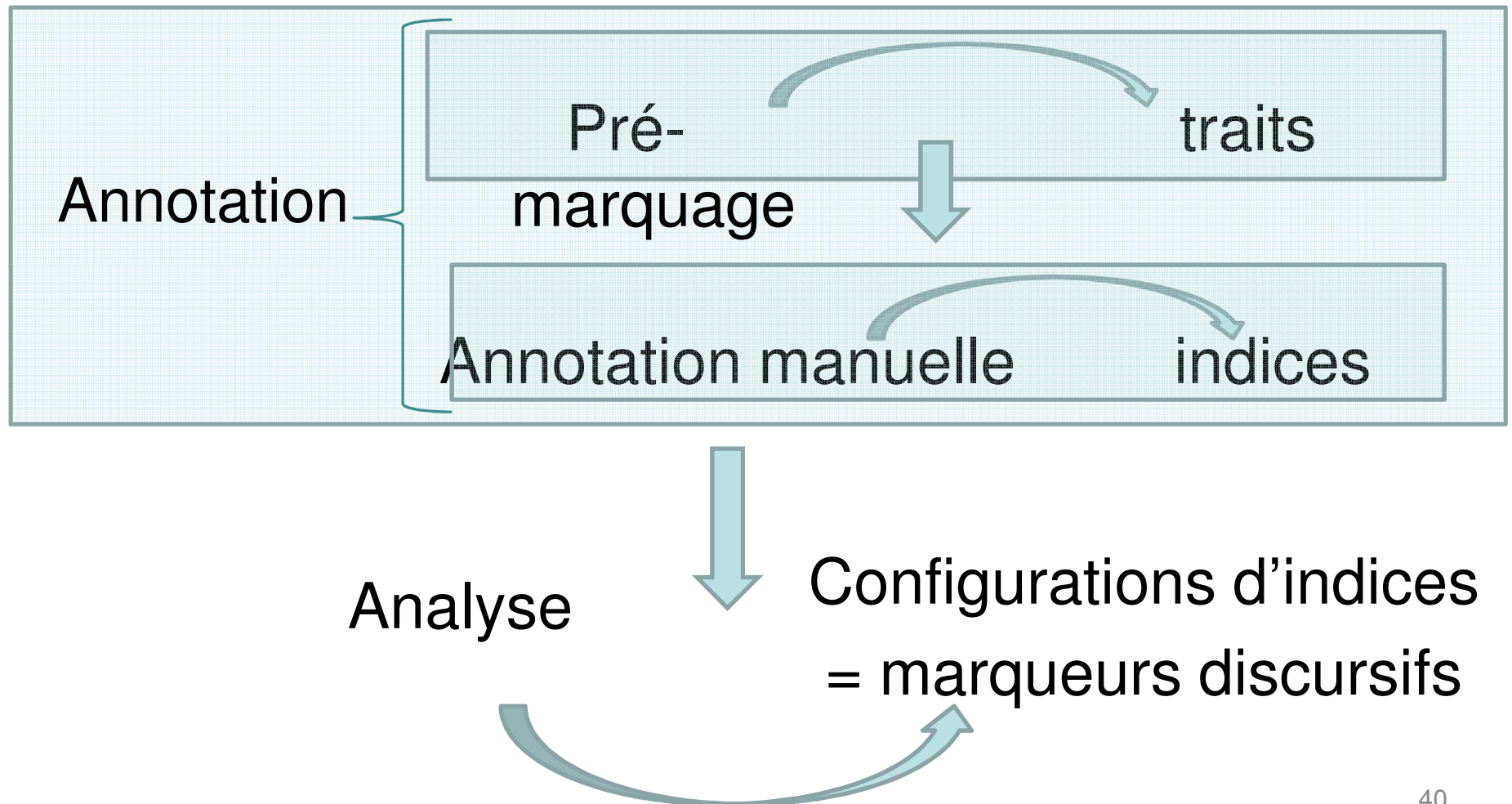
Les États-Unis connurent une période de forte activité pendant la guerre de quarante avec le retour au plein emploi, la mobilisation des hommes jeunes étant compensée par le recours massif à la main d'œuvre féminine dans les usines d'armement. D'énormes investissements furent faits dans beaucoup de domaines qui donnèrent un avantage technologique au pays après guerre. Lorsque la guerre arriva à son terme, le retour des millions de soldats dans leurs foyers imposa une période de réajustement de l'économie. C'est cette transition qu'était censée faciliter le G.I. Bill. Au total ce fut le seul pays important **à ne pas sortir ruiné de la guerre**. La guerre avait également permis à des économistes keynésiens, sous l'influence de Hansen, de peupler l'administration qui se dote pendant la période des moyens en hommes, en idées et en droit, de son action. La paix retrouvée ils mirent en place une politique de dépense publique qui ne se relâchera plus.

Ces exemples montrent que la montée vers la guerre ne sera nulle part le secret de la fin de la crise de 1929. La guerre marquera une rupture dans les mentalités, provoquera un besoin de reconstruction intense pendant une dizaine d'année, provoquera une concentration du pouvoir économique dans l'Etat qui est désormais partout chargé du droit au travail et à la sécurité sociale. Tous les gouvernements deviennent "keynésiens". L'orthodoxie d'avant 1929 est morte.

Dispositif d'analyse des structures

- Pour obtenir des résultats généralisables :
 - Analyse de grands volumes de textes
 - Annotation d'un grand nombre de structures
- L'annotation manuelle est guidée, assistée, par le biais d'une phase de pré-marquage des textes
- Cette phase de pré-marquage est basée sur des techniques de TAL

Analyse des marqueurs en 3 étapes



Traits pertinents pour le prémarquage

Trait	Description du segment marqué
Traits textuels	
Patrons typographiques	Le premier mot de chaque item dans une liste formattée, le dernier mot d'un paragraphe se terminant par “:”
Patrons ponctuationnels	Le dernier mot avant “:”, les ponctuations et les coordinations dans une phrase qui inclue ‘: . , ; et ou”
Titres	
Paragrapes	
Traits lexicaux, détachés à l'initiale	
Adverbes spatiaux, temporaux et notionnels	<i>aujourd'hui, hier, prochainement, ici, là-bas, selon, etc.</i>
MIL (marqueurs d'intégration linéaire)	<i>premièrement, parallèlement, enfin, etc.</i>
Connecteurs	Conjonctions de coordination et adverbes au début d'une phrase
Éléments prospectifs	GP incluant un item lexical susceptible de signaler un élément prospectif (<i>les suivants...</i>)
Encapsulations	GP démonstratif incluant un numéral et un item lexical susceptible de signaler une encapsulation (<i>ces trois scenarios...</i>)
Dans la zone préverbale, différents types d'expressions référentielles en position de sujet	<i>pronoms, GP démonstratifs, GP possessifs, etc.</i>

III. Cadre expérimental

1. Corpus
2. Plate-forme d'annotation
3. Campagne d'annotation
4. Exploitation des annotations

Le corpus

- Des textes non-narratifs et longs
- Une mise en page élaborée
- Diversité des genres de textes :
 - Textes encyclopédiques (Wikipedia)
 - Articles de recherche (CMLF)
 - Textes argumentatifs longs (textes de l'IFRI)
- Traitement des textes :
 - textes XMLisés et convertis au format TEI-P5
 - Étiquetage et analyse syntaxique (Syntex)
 - Marquage automatique des traits

Format TEI-P5 - *header*

```
<TEI>
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Sélection naturelle</title>
        <editor>Wikipedia.org</editor>
        <principal>Wikipedia.org</principal>
        (...)
      <publicationStmt>
        <distributor>
          <name>CLLE-ERSS</name>
          <address>
            <addrLine>Maison de la recherche</addrLine>
            <addrLine>5 allées Antonio Machado</addrLine>
            <addrLine>F 31058 TOULOUSE CEDEX</addrLine>
          </address>
        </distributor>
        <availability>
          <p>Contenu soumis à la GFDL</p>
          <p> exploitable et diffusable pour la communauté
scientifique </p>
          <p>ne peut être utilisé à des fins commerciales</p>
        </availability>
      </publicationStmt>
```

Format TEI-P5 – *marquage de la structure logique*

```
<div type="level2">
```

```
<head>II.4. Une histoire imaginée par Richard Dawkins</head>
```

```
<p> Cette histoire amusante n'a d'autre but que de bien fixer un point important de la théorie darwinienne. </p>
```

```
<p> Deux brontosaures voient un T-Rex avancer dans leur direction et se mettent à courir aussi vite qu'ils le peuvent. Puis l'un des deux dit à l'autre : </p>
```

```
<cit>
```

```
<quote>
```

```
« Pourquoi nous fatiguons-nous au juste ? Nous n'avons de toute façon pas la moindre chance d'arriver à courir plus vite qu'un T-Rex ! »
```

```
</quote>
```

```
</cit>
```

```
<p> Et l'autre lui répond cyniquement : </p>
```

```
<cit>
```

```
<quote> « Je ne cherche pas à courir plus vite que le T-Rex. Je cherche juste à courir plus vite que toi ! »
```

```
</quote>
```

```
</cit>
```

```
<p> L'idée est de rappeler que le processus concerne moins une compétition entre espèces, qu'une compétition à l'intérieur de chaque espèce. (...) </p>
```

```
</div>
```

Plate-forme d'annotation

- **Glozz** (Widlöcher & Mathet, GREYC, U. de Caen)
 - Visualisation de la mise en page et du marquage
 - Coloration des marques
 - 2 vues complémentaires sur le texte :
 - Vue principale pour l'annotation
 - Vue synoptique pour un regard à large échelle sur le texte



Parallèlement, les échanges économiques internationaux subissent de plein fouet d'abord le ralentissement qui commence aux USA, ensuite l'effet négatif des réactions protectionnistes, d'abord des USA, puis de tous les autres pays quand ils sont touchés à leur tour, la France et le royaume uni tentent de se replier sur leurs colonies, mettant au point la « préférence impériale », interdite lors de la Conférence de Berlin (1885) mais largement pratiquée après 1914

Les relations économiques étant à l'époque bien moindre qu'aujourd'hui, ces répercussions mettront du temps à se diffuser : par exemple la France sera touchée à partir du second semestre de 1930, soit six mois plus tard. L'Italie est touchée à partir de 1931

Les réactions gouvernementales en Europe ne seront pas plus adéquates que celles aux USA. La France a crise sera aggravée par les mesures déflationnistes (baisse des prix et des salaires) des gouvernements Tardieu et Laval, bien que ceux-ci tenteront, de façon limitée, quelques grands travaux (dont l'électrification des campagnes). En Allemagne, le taux de chômage atteint des sommets (plus de 25 % de la population active en 1932), alimentant la déshérence et la colère de la population, et c'est en promettant de régler le problème de la crise qu'Adolf Hitler parvint au pouvoir le 30 janvier 1933

En Amérique du Sud, en Asie et en Afrique, se produit la « crise des produits de dessert » liée à la forte baisse du pouvoir d'achat en Europe et en Amérique du Nord. Au Brésil, pour limiter la névente et faire grimper le cours, le café est brûlé dans les locomotives.

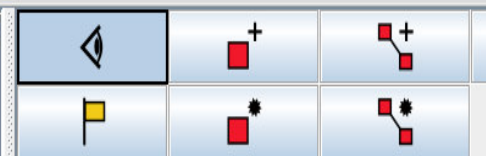
Le monde entier est touché excepté l'Union soviétique de Staline, protégée par son système économique autarcique.

II. Les causes de la crise économique

La réflexion économique a posteriori a porté d'une part sur les causes immédiates de la crise et d'autres part sur les raisons de la transformation de la récession en dépression.

II.1. La survenue de la crise

Les innombrables écrits sur la crise de 1929 n'ont pas permis de dessiner une explication généralement admise de sa survenue. Les ouvrages qui lui sont consacrés sont le plus souvent normatifs ou descriptifs et très peu explicatifs. Chaque école a tiré un tant soit peu la couverture en faveur de ses thèses générales. Charles Kindleberger distingue les explications qui reposent sur une cause ou une origine unique (unicausales) de celles, telles la sienne, reposant sur la conjonction de plusieurs facteurs



Min Depth :

Max Depth :

Search text :

Unit type :

Search Unit :

Attribute constr. :

Value constr. :

Units	Relations	Schemas
default	default	default

Feature name	Feature v

Sort/Type	Sort/Date	Show sel.

u_enumerateTheme(28076,28084) ID=232
 s_SE(231,286,277,230,287,201,282,309,279,229,288,280,276,234,26
 u_UR(22163,22681) ID=284

Campagne d'annotation

- L'annotation porte sur :
 - les structures (schémas) et leurs unités
 - les indices : traits prémarqués validés, ou introduction de nouveaux indices
- Principales étapes :
 - Elaboration des spécifications => guide d'annotation
 - Phase exploratoire (printemps 2009)
 - évaluation des spécifications par les membres du projet sur une sélection de 4 textes prémarqués
 - => modification du guide, du prémarquage
 - Phase opérationnelle : annotateurs neutres (setp-oct)

Phase d'annotation opérationnelle

- 3 annotateurs
- 88 textes
- attendu (sur la base des premières annotations réalisées) :
 - environ 1000 SE, 700 SUR
- Evaluation du jugement inter-annotateurs :
 - 6 textes ont été annotés par les 3
 - F-mesure de 0.7 à 0.8 en moyenne

Exploitation des annotations

- Les annotations vont être exploitées :
 - Manuellement, par le biais d'une interface d'exploration
 - Automatiquement, par des techniques de fouilles de données

Perspectives de recherche

– Corrélation indices / structure :

- Quelles configurations d'indices émergent ?
- Quelle est l'efficacité des indices existants pour détecter une structure ?
- Possibilité d'enrichir le marquage (ex : parallélisme syntaxique, temps verbaux)

=> Amélioration de la phase de prémarquage pour une nouvelle phase d'annotation

Perspectives de recherche

- Découvrir la variété des structures :
 - Peut-on faire émerger des clusters de structures (en fonction de différentes caractéristiques : échelle, position, indices...)
 - Les caractéristiques et la répartition des structures varient-elles selon les types de textes ?
- Liens SE / SUR :
 - Un mode de structuration complémentaire ?
- Fonctionnement lexical :
 - articulation entre segmentation par les structures / segmentation lexicale, thématique

Démo : texte annoté sous
Glozz

Références bibliographiques

Introduction:

- Marcu, D., Discourse Parsing, Automatic, in K. Brown (Ed.), *Encyclopedia of Language and Linguistics*, Elsevier, 2005, vol.3.
- Moore, J. D., Wiemer-Hastings, P., Discourse in Computational Linguistics and Artificial Intelligence, in A. C. Graesser, M. A. Gernsbacher, S. R. Goldman (Eds.), *Handbook of Discourse Processes*, London, Lawrence Erlbaum, 2003, p. 439-485.
- Nazarenko, A., Sur quelle sémantique reposent les méthodes automatiques d'accès au contenu textuel ? , in A. Condamines (Ed.), *Sémantique et Corpus*, Paris, Hermès, 2005, p. 211-244.
- Péry-Woodley, M.-P., & Scott, D. (2006). Computational Approaches to Discourse and Document Processing. *TAL* 47 2, 7-19

ANNODIS:

Consultables à partir du site du projet ANNODIS: <http://w3erss.univ-tlse2.fr/annodis>

- Péry-Woodley, M.-P., *et al.* (2009). "ANNODIS: une approche outillée de l'annotation de structures discursives", *TALN 2009*, Senlis (France) (2009, 24-26 juin 2009).
- Ho-Dac, M., Fabre, C., Péry-Woodley, M.-P., & Rebeyrolle, J. (2009). "Des indices aux marqueurs : méthodes de découverte de marqueurs discursifs complexes", *Linguistic and Psycholinguistic Approaches to Text Structuring*, Paris (2009, 21-23 septembre 2009).