Accès au (contenu du) document

Nicolas.Hernandez@limsi.fr Université Paris-Sud XI
LIMSI/CNRS – LIR

Deux significations

- Compréhension de texte (année 80)
  1. Acquisition/extraction de connaissances
     - Indépendamment de l'objet du document
     - Approche corpus
     - E.g. Svetlan…
  2. Analyse d'un contenu
     - Intérêt sur le document (but, thème, propos, etc.)
     - Le 1er pouvant intervenir pour l'analyse du 2nd
     - E.g. réseaux de co-occurrences pour la segmentation

Le 14 décembre 2004 ?

- Annonce de Google de scanner et numériser 15 millions de livres de 5 bibliothèques anglo-saxones parmi les plus riches et les plus célèbres (New York Public Library, University of Michigan, Stanford, Harvard (USA), Oxford (GB)) afin de les mettre à disposition en ligne

- Soutenance de ma thèse

  « Description et détection automatique de structures de texte »

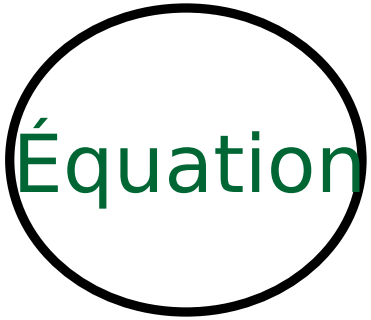
L’accès au contenu d’un document

- Types d’accès souhaités
  - Lecture rapide
  - Exploration libre
    - besoin d’information vague ou mal spécifié
  - Fouille de certaines parties
    - éléments réponses à plusieurs endroits

- Problèmes
  - Écrans de taille limitée (Ordinateur, Téléphone, PDA)
  - Peu de moyens de navigation offerts exceptés ascenseur et recherche de mots clefs
  - Difficulté de représenter du texte autrement que par du texte

Le Système Idéal

- Interface
  - Propriétés physiques (e.g. un écran qui restreint le champ de vision)
  - Et logicielles…



- Utilisateur
  - Tâche
  - Besoin d'information
  - Connaissances du domaine
  - Expériences avec ce type de documents…

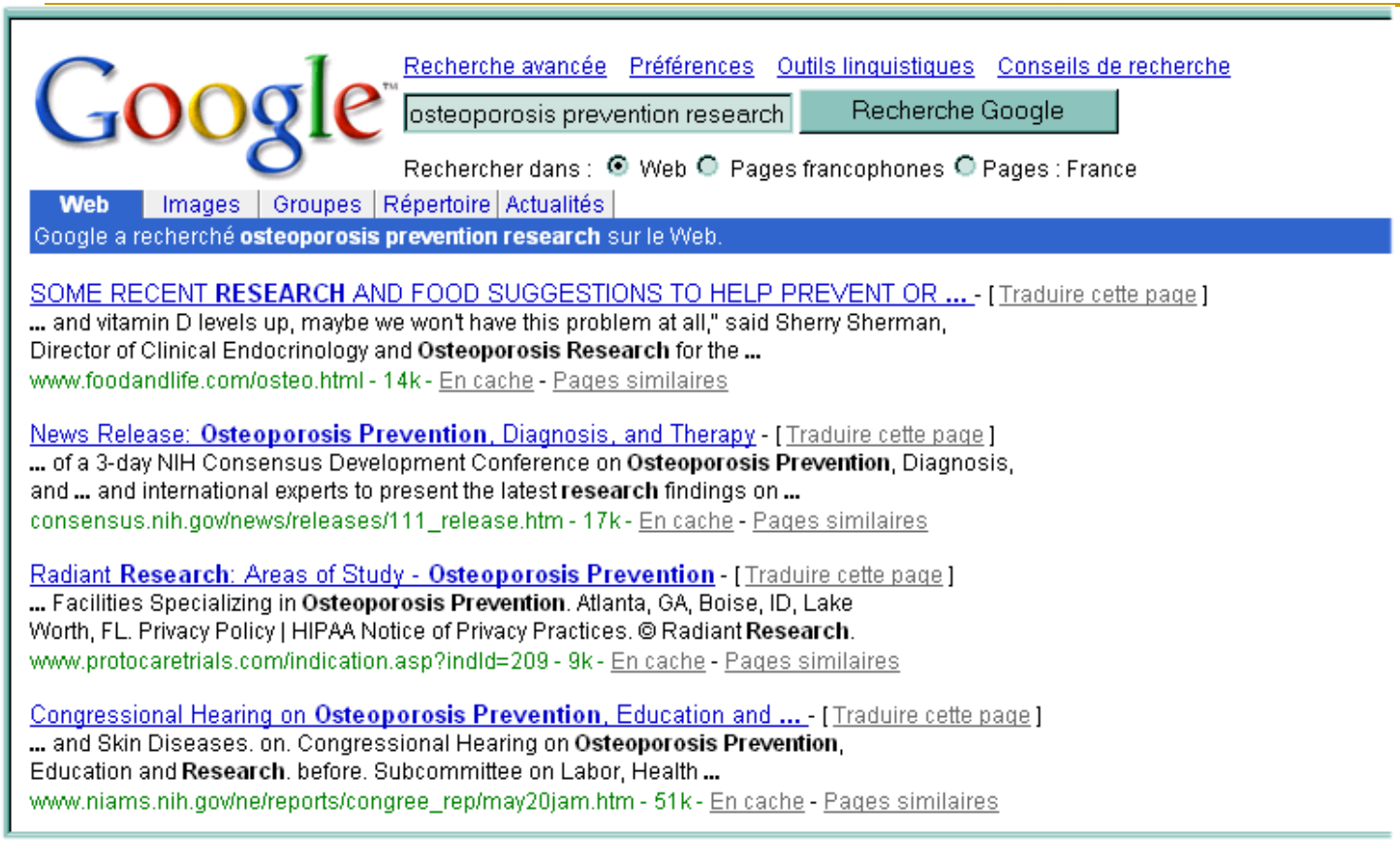
- Document
  - Thème
  - Genre
  - Taille…

Cadre de recherche

- Composante textuelle du document
  - (indépendamment de tout balisage de mise en forme ou autres)
- Expositifs
  - scientifique et technique
- Français et anglais
  - selon la disponibilité des ressources

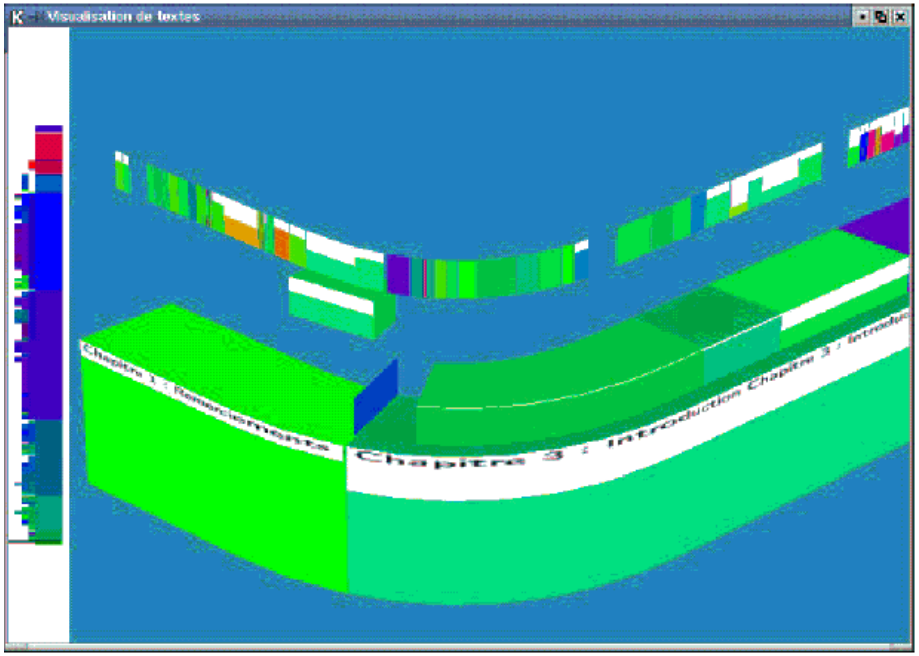
Systèmes d'accès au contenu de documents simples

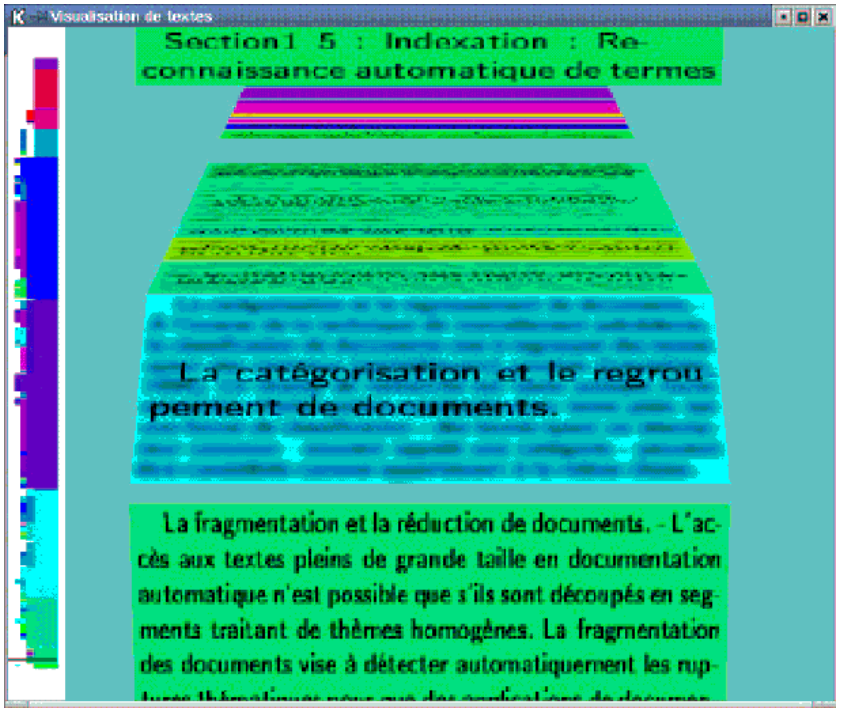
- Donne une brève description (à partir de requêtes)
  - Google, Tilebars
- Bien adapté pour visualiser et naviguer dans de larges documents structurés
  - 3D-XV
- Une description thématique d'une structure globale plate
  - ViewTool
- Structuration thématique au niveau de la phrase
  - CH00
- La dimension sémantico-rhétorique pour décrire
  - Argumentative Zoning, SumUM



- Présente le titre des documents et graisse les mots de la requête dans un passage extrait (KWIC)
- N'exonère pas de consulter le document pour juger sa pertinence
- Ne fournit pas de moyens d'exploration (exceptés recherche de mots clefs et défilement linéaire)

3D-XV [Jacquemin & Jardino 02]

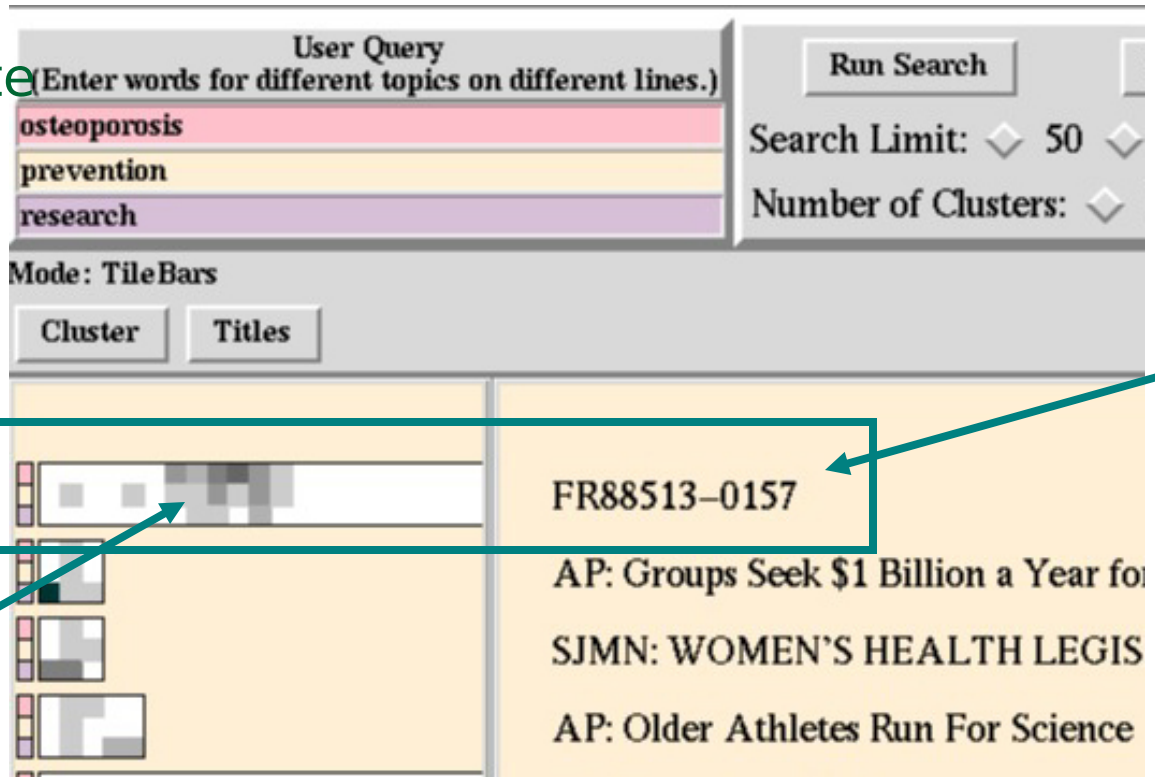




- Offre une visualisation 3D dynamique et interactive des documents structurés (XML) de large taille
- La couleur exprime la cohérence thématique entre diverses parties logiques
- Bien adapté pour les documents de large taille
- Requiert la présence d'une structure
- ~~Illustre la difficulté de représenter du texte autrement que par du~~ texte

Retour

Tilebars [Hearst 96]



- Ajoute au document une barre graphique qui indiquent les segments du texte qui contiennent les termes de la requête
- Un début d'informations sur la structure (taille et couverture thématique)
- Mêmes limites que précédemment

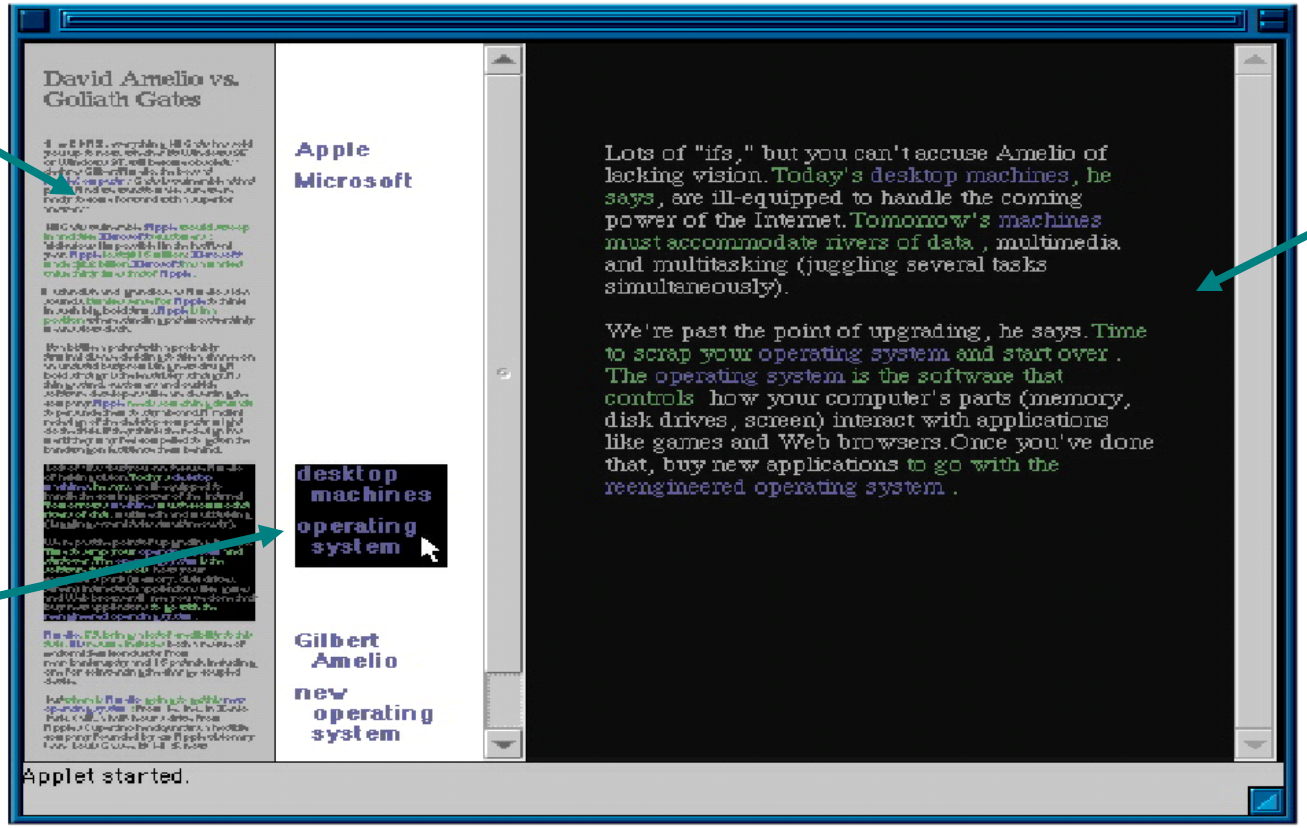
Retour

ViewTool [Boguraev & Kennedy 97]

Le document

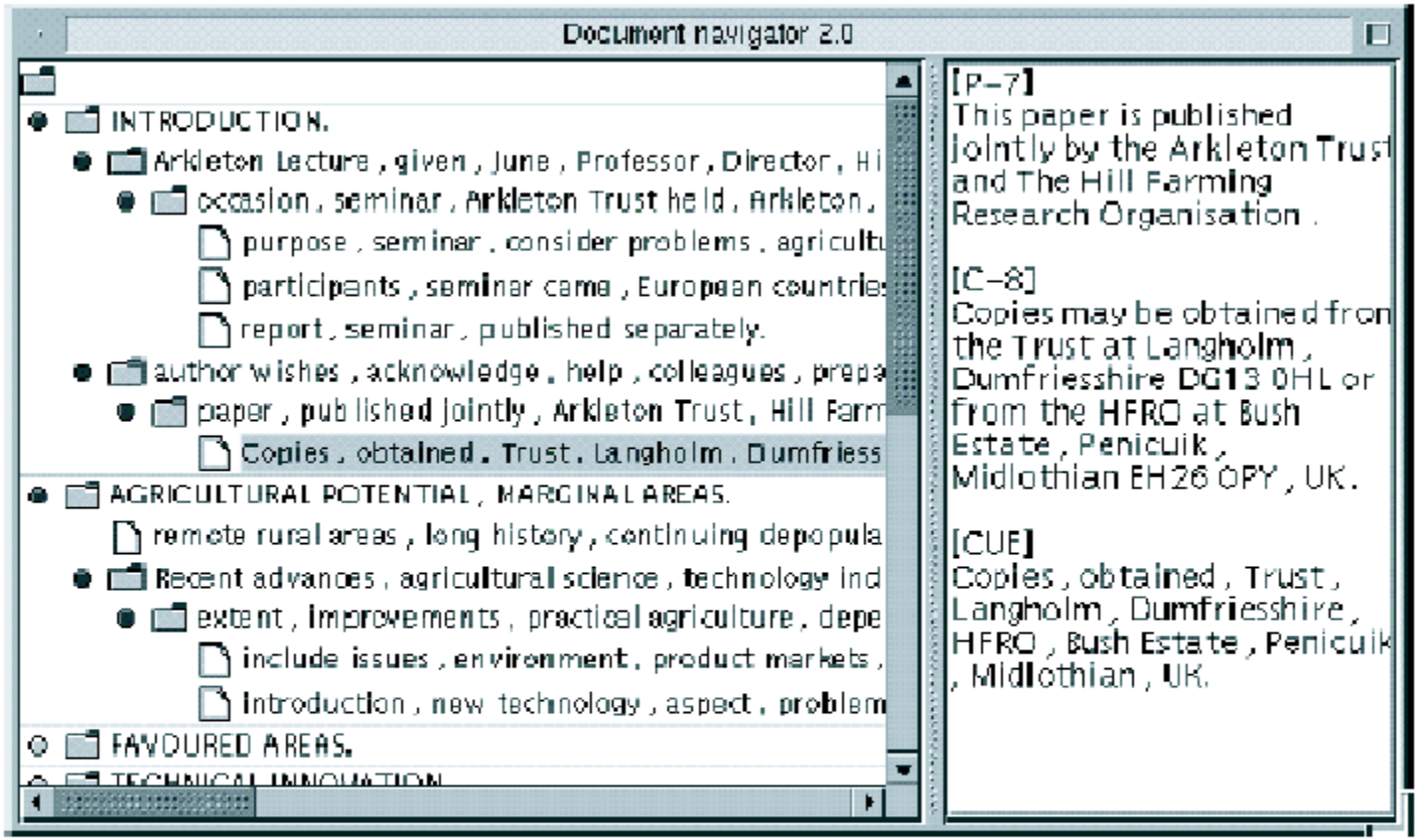
Un segment particulier

Etiquette thématique (*Topic stamps*)



- Structuration thématique globale plate
- Descripteurs thématiques des segments (résolution d'anaphores)

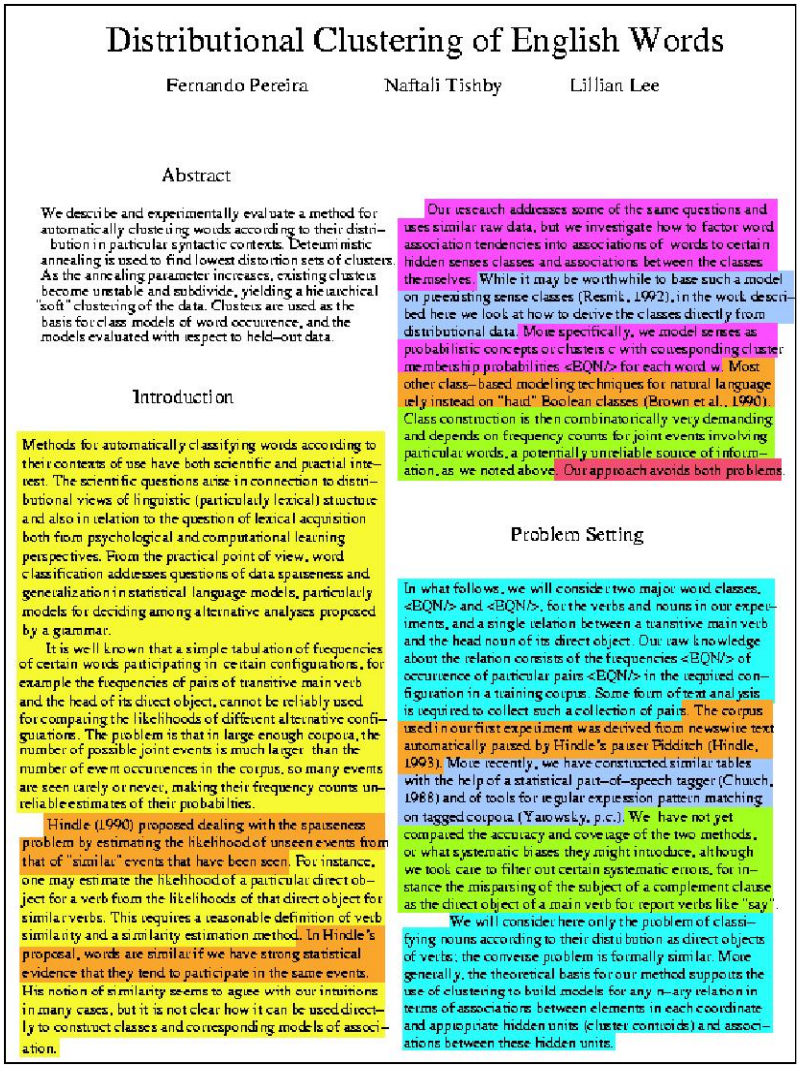
CH00 [Choi 02]



- Structuration thématique au niveau de la phrase
- Phrases décrites par les termes ayant un fort tf.idf

Argumentative zoning [Teufel & Moens 99]

- Coloration des phrases selon le type d'information qu'elles contiennent
  - **Objectif de recherche du papier**
  - **Travaux de l'auteur (méthodes, résultats, etc.)**
  - **Comparaisons avec d'autres travaux**
  - **Structure logique**



Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word coocurrence, and the models evaluated with respect to held-out data.

Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practial interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilties.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agiee with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorically very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information, as we noted above. Our approach avoids both problems.

Problem Setting

In what follows, we will consider two major word classes, <EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs <EQN/> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster centroids) and associations between these hidden units.

Résumé automatique - techniques

Sélection d'information

Instanciation de patrons

Genre de texte défini

Construction d'une représentation

Inférences, Omission, généralisation

Production

Extraction de phrases ou de termes « saillants »

Position dans le document,

Fréquence

Mots clefs

Questions

- Quelles unités d'informations manipuler ?
  - Pour survoler globalement ou s'attarder sur une partie
- Quelles informations présenter ?
  - Pour juger la pertinence d'un passage
- Quelles structures de texte adopter ?
  - Pour soutenir une navigation à différents niveaux de granularité
  - Des rapprochements entre informations
  - Des mises en contexte

Sommaire

- Systèmes existants
- **Différents aspects d'un texte**
- Propositions
  1. Identification des descripteurs thématiques propre à un segment
  2. Identification de descripteurs de type d'information (sémantico-rhétorique) contenu dans une phrase
  3. Organisation thématique du discours à la fois aux niveaux global et local

La cohérence d'un texte

Un lecteur comprend un texte
s'il en reconnaît la cohérence

Cohérence :

- Identifier des unités (thématique, intention, avec une mise en forme visuelle, etc.)

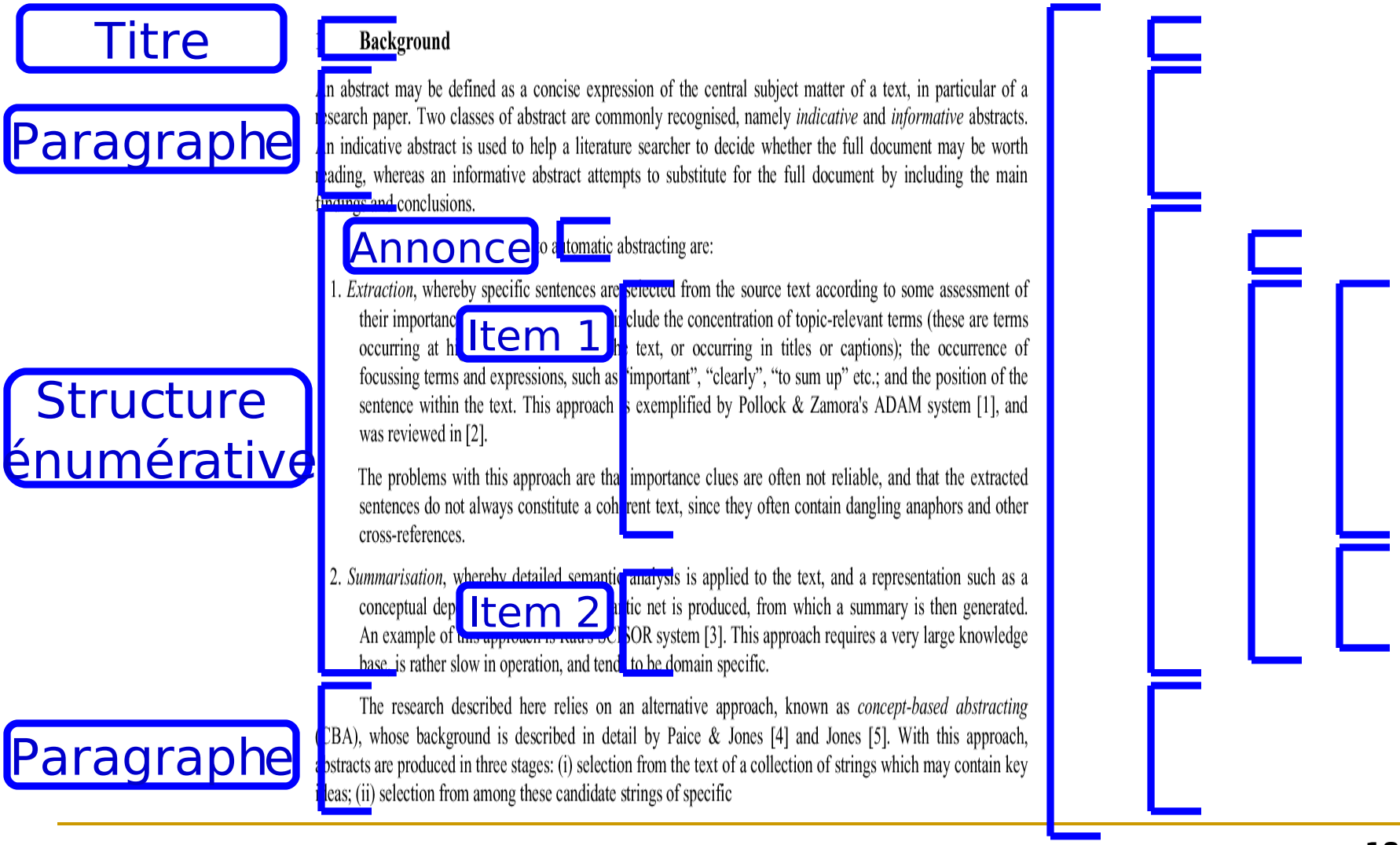
  et reconnaître des relations entre ces unités

  [Mann & Thompson 87; Grosz & Sidner 86; Virbel 89; Danes 74; …]

- Etablir les conditions de vérité sémantico-rhétorique d'un contenu

  [Kamp 81] et [Hobbs 85; Asher & Lascarides 94]

Structure visuelle (typo-dispositionnelle) [Virbel 87; Luc 00]



Différentes vues descriptives

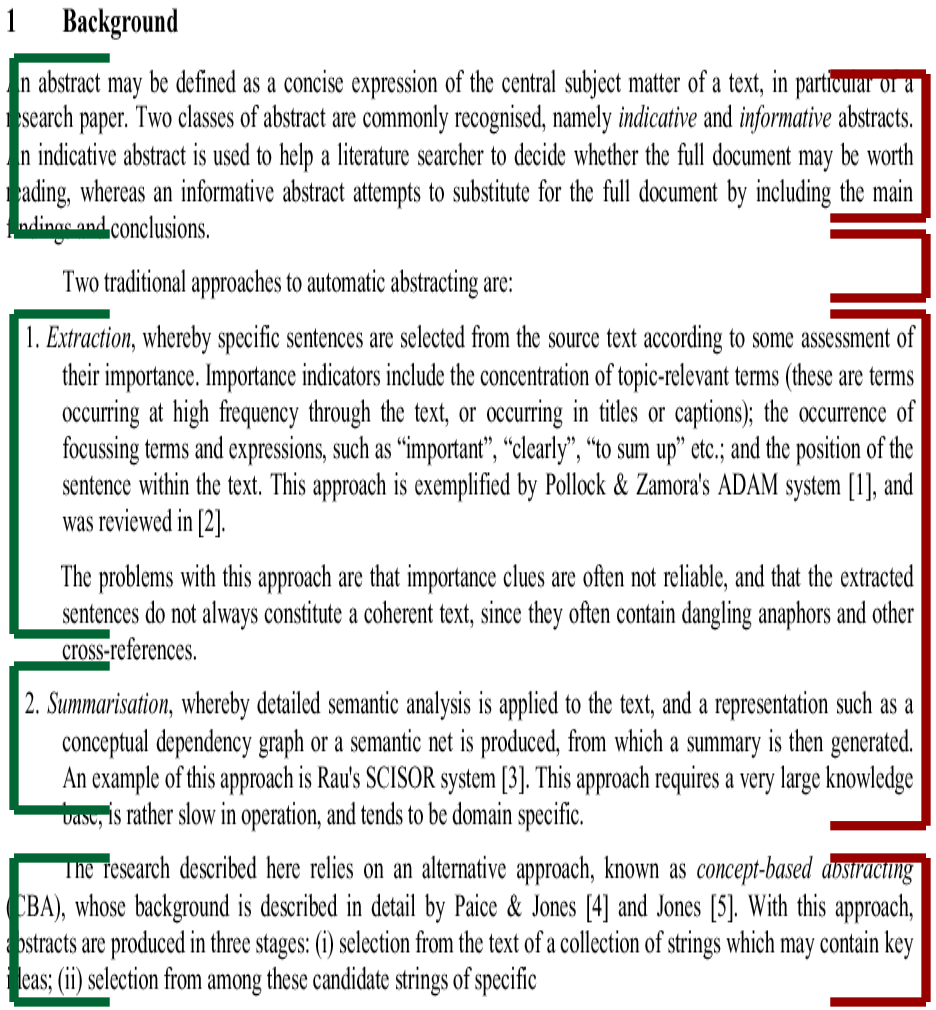
Thématique

Résumé

Par extraction

Par abstraction

Par sélection et génération



1 Background

An abstract may be defined as a concise expression of the central subject matter of a text, in particular of a research paper. Two classes of abstract are commonly recognised, namely *indicative* and *informative* abstracts. An indicative abstract is used to help a literature searcher to decide whether the full document may be worth reading, whereas an informative abstract attempts to substitute for the full document by including the main findings and conclusions.

Two traditional approaches to automatic abstracting are:

1. *Extraction*, whereby specific sentences are selected from the source text according to some assessment of their importance. Importance indicators include the concentration of topic-relevant terms (these are terms occurring at high frequency through the text, or occurring in titles or captions); the occurrence of focussing terms and expressions, such as "important", "clearly", "to sum up" etc.; and the position of the sentence within the text. This approach is exemplified by Pollock & Zamora's ADAM system [1], and was reviewed in [2].

   The problems with this approach are that importance clues are often not reliable, and that the extracted sentences do not always constitute a coherent text, since they often contain dangling anaphors and other cross-references.

2. *Summarisation*, whereby detailed semantic analysis is applied to the text, and a representation such as a conceptual dependency graph or a semantic net is produced, from which a summary is then generated. An example of this approach is Rau's SCISOR system [3]. This approach requires a very large knowledge base, is rather slow in operation, and tends to be domain specific.

The research described here relies on an alternative approach, known as *concept-based abstracting* (CBA), whose background is described in detail by Paice & Jones [4] and Jones [5]. With this approach, abstracts are produced in three stages: (i) selection from the text of a collection of strings which may contain key ideas; (ii) selection from among these candidate strings of specific

Sémantico-rhétorique

Définition

Organisation logique

Méthodes traditionnelles

Méthode alternative

Différentes vues descriptives

Thématique

Résumé

Par extraction

Par abstraction

Par sélection et génération

1 **Background**

An abstract may be defined as a concise expression of the central subject matter of a text, in particular of a research paper. Two classes of abstract are commonly recognised, namely *indicative* and *informative* abstracts. An indicative abstract is used to help a literature searcher to decide whether the full document may be worth reading, whereas an informative abstract attempts to substitute for the full document by including the main findings and conclusions.

Two traditional approaches to automatic abstracting are:

1. *Extraction*, whereby specific sentences are selected from the source text according to some assessment of their importance. Importance indicators include the concentration of topic-relevant terms (these are terms occurring at high frequency through the text, or occurring in titles or captions); the occurrence of focussing terms and expressions, such as "important", "clearly", "to sum up" etc.; and the position of the sentence within the text. This approach is exemplified by Pollock & Zamora's ADAM system [1], and was reviewed in [2].

   The problems with this approach are that importance clues are often not reliable, and that the extracted sentences do not always constitute a coherent text, since they often contain dangling anaphors and other cross-references.

2. *Summarisation*, whereby detailed semantic analysis is applied to the text, and a representation such as a conceptual dependency graph or a semantic net is produced, from which a summary is then generated. An example of this approach is Rau's SCISOR system [3]. This approach requires a very large knowledge base, is rather slow in operation, and tends to be domain specific.

The research described here relies on an alternative approach, known as *concept-based abstracting* (CBA), whose background is described in detail by Paice & Jones [4] and Jones [5]. With this approach, abstracts are produced in three stages: (i) selection from the text of a collection of strings which may contain key ideas; (ii) selection from among these candidate strings of specific

Différentes vues descriptives

Thématique

Résumé

Par extraction

Par abstraction

Par sélection et génération

1 Background

An abstract may be defined as a concise expression of the central subject matter of a text, in particular of a research paper. Two classes of abstract are commonly recognised, namely *indicative* and *informative* abstracts. An indicative abstract is used to help a literature searcher to decide whether the full document may be worth reading, whereas an informative abstract attempts to substitute for the full document by including the main findings and conclusions.

Two traditional approaches to automatic abstracting are:

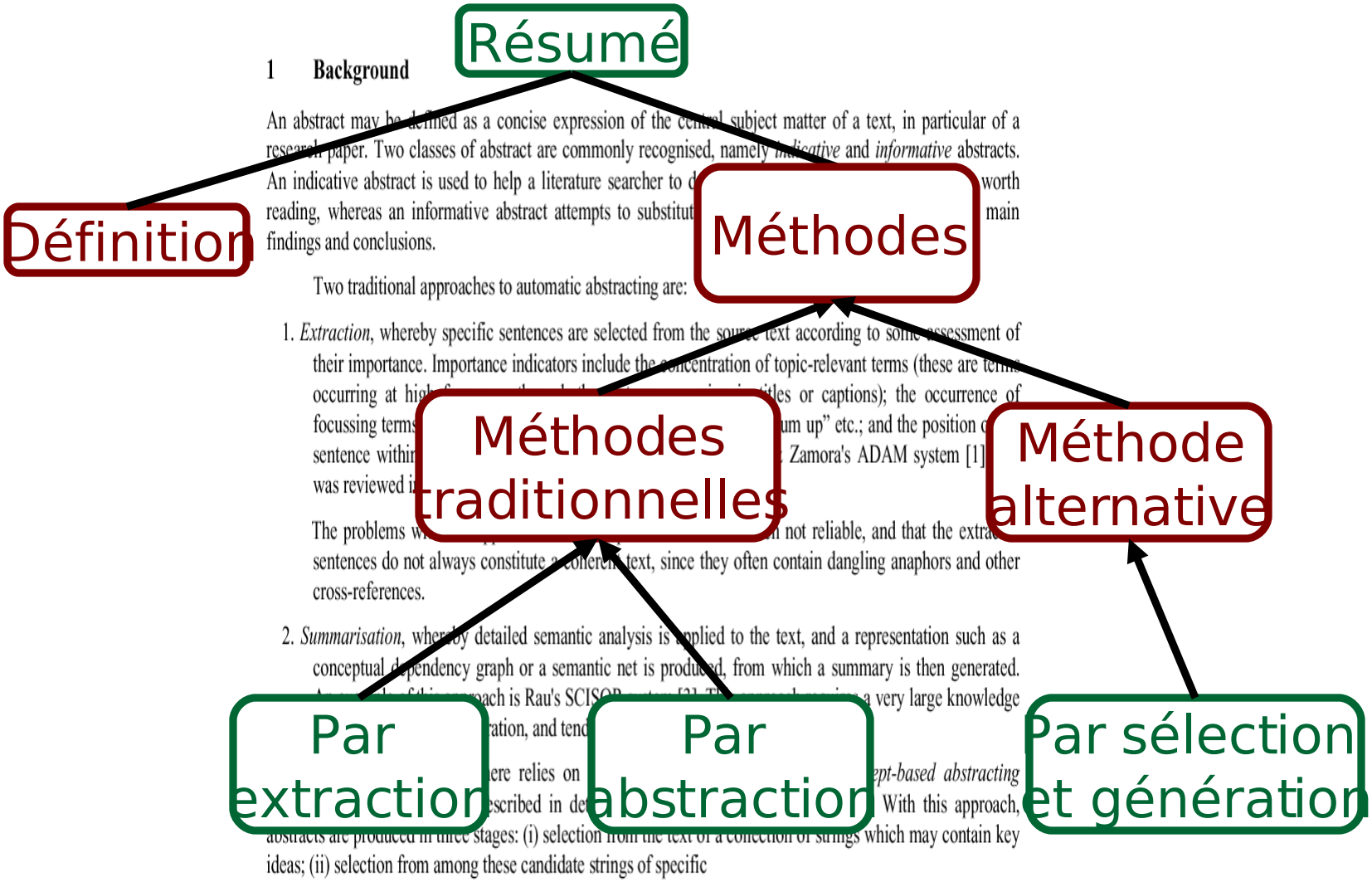
1. *Extraction*, whereby specific sentences are selected from the source text according to some assessment of their importance. Importance indicators include the concentration of topic-relevant terms (these are terms ugh the text, or occurring in titles or captions); the occurrence of ch as "important", "clearly", "to sum up" etc.; and the position of the ach is exemplified by Pollock & Zamora's ADAM system [1], and

The problems with this approach are that importance clues are often not reliable, and that the extracted sentences do not always constitute a coherent text, since they often contain dangling anaphors and other cross-references.

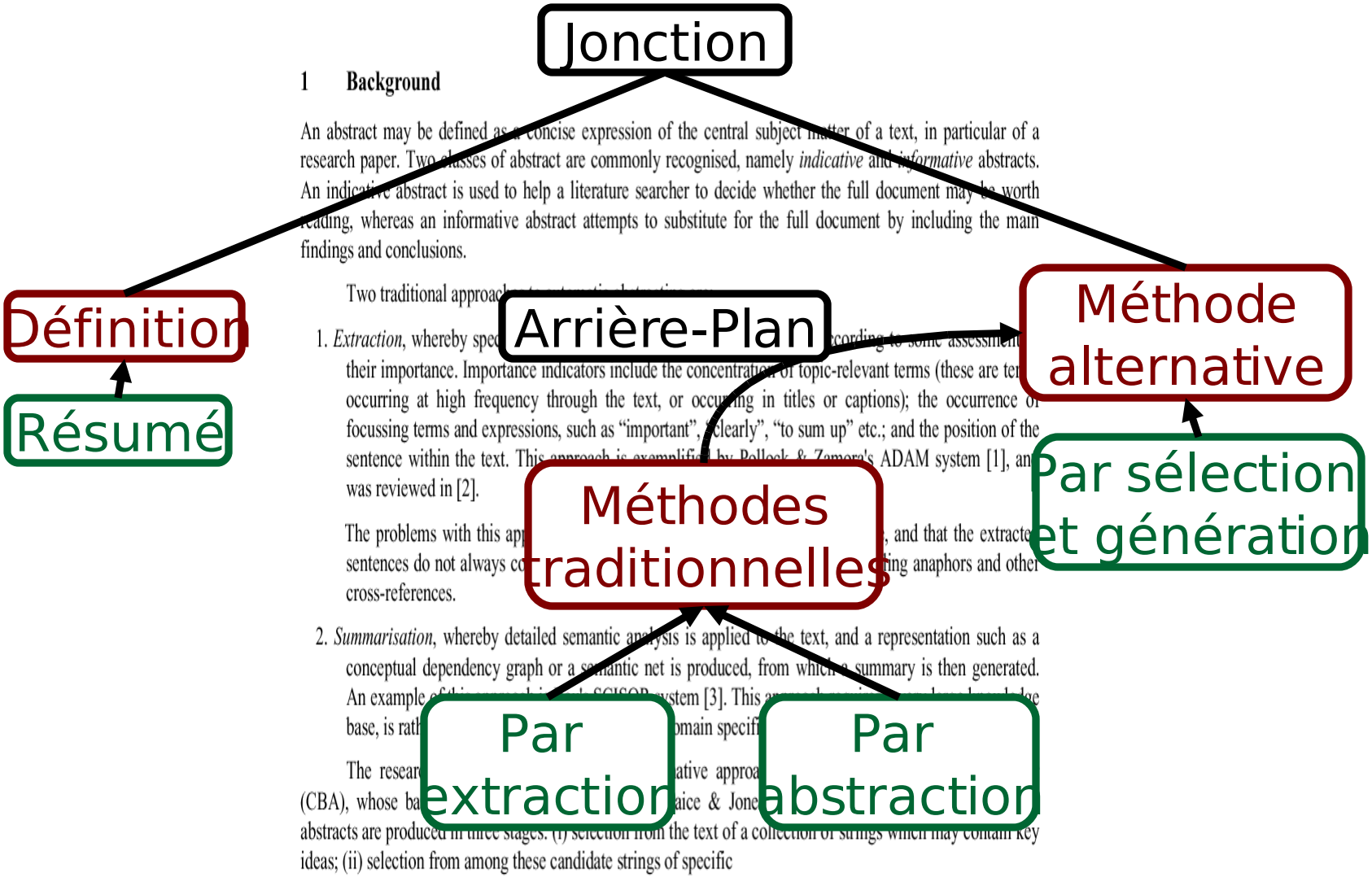
ntic analysis is applied to the text, and a representation such as a mantic net is produced, from which a summary is then generated. SCISOR system [3]. This approach requires a very large knowledge ends to be domain specific.

lies on an alternative approach, known as *concept-based abstracting* d in detail by Paice & Jones [4] and Jones [5]. With this approach, i) selection from the text of a collection of strings which may contain key andidate strings of specific

Organisation informationnelle



Organisation rhétorique



Différentes vues descriptives

Two traditional approaches to automatic abstracting are:

1. Extraction, where specific sentences are selected from the source text, based on some assessment of their importance.

Résumé par extraction

Définition

Importance indicators include terms with high frequency, cue phrases, such as "important", "to sum up" etc., and the position of the sentence within the text.

Critères d'importance

This approach is exemplified by Pollock and Zamora's CRYSTAM system [1].

Exemple

Another example was reviewed in [2].

Exemple

The problems with this approach is that the extracted sentences do not always constitute a coherent text, since they often contain dangling anaphors and other cross-references.

Problème

2. Abstraction, where detailed semantic analysis is applied to the text, and a representation such as a semantic net is produced, from which a summary is then generated.

Résumé par abstraction

Définition

Organisation informationnelle

Two traditional approaches to automatic abstracting are:

1. Extraction, whereby specific sentences are selected from the source text according to some assessment of their importance.

Importance indicators include terms with high frequency, occurrence of expressions, such as "important", "to sum up" etc., and the position of the sentence within the text.

This approach is exemplified by Pollock and Zamora's ADAM system [1].

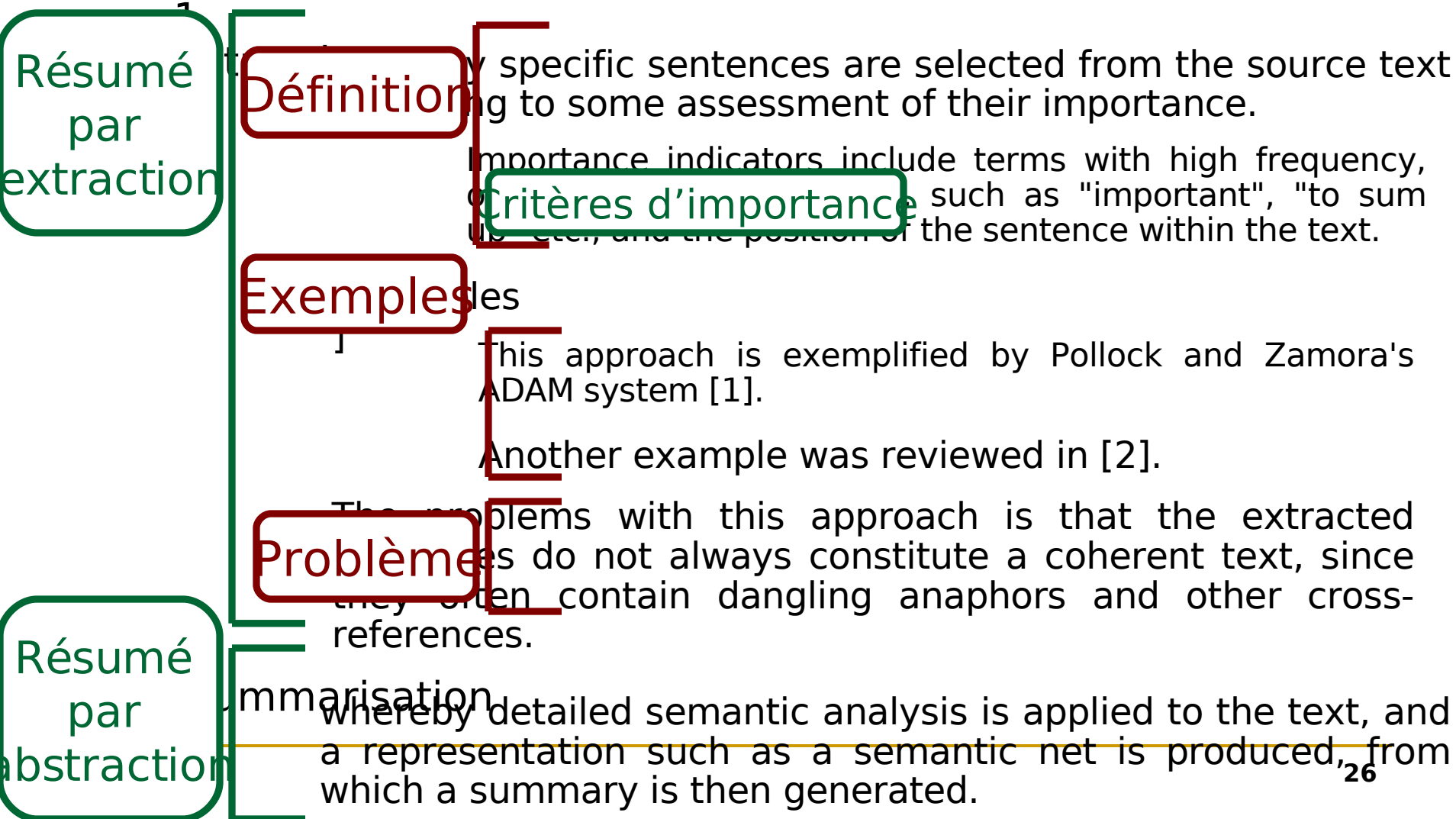
Another example was reviewed in [2].

The problems with this approach is that the extracted sentences do not always constitute a coherent text, since they often contain dangling anaphors and other cross-references.

2. Summarisation, whereby detailed semantic analysis is applied to the text, and a representation such as a semantic net is produced, from which a summary is then generated.

Organisation informationnelle

Two traditional approaches to automatic abstracting are:



Propriétés structurelles communes

- Théories
  - Orientées « relation »

    [Mann & Thompson 87; Polanyi 88]
  - Orientées « segment »

    [Hearst 97; Daneš 74; Minel Al 01; Teufel & Moens 02; Saggion & Lapalme 02]
- Relations structurelles
  - Subordination : dépendance
    - Sémantique (exemple, explication),
    - Rhétorique (justification)
    - Intentionnelle (la satisfaction d'un but de l'auteur requiert l'accomplissement d'un sous-but)
    - Thématique (un thème se décompose en sous-thèmes)
  - Coordination : même importance
    - Informationnelle (définition, problème et solution d'un sujet)
    - Rhétorique (Items d'une liste d'arguments)

Modélisation de la structure d'un texte

- Plusieurs plans d'organisation concomitants
  - Visuel, Thématique, Rhétorique
- Organisation
  - Niveau propositionnel
  - Macro (le segment : unité discursive homogène selon un critère)
- Intertextualité
  - Régularités textuelles propres à un genre (i.e. type d'information) qui guident la production et l'analyse
- Cohésion
  - Indices discursifs qui expriment des relations entre les propositions

Lu au dos d’une tablette de chocolat… L’auteur est-il suisse ou français ?

(1) [En général, les gens se serrent la main droite quand ils se rencontrent ou se séparent, ou bien ils s'embrassent.]

(2) [Hello, bonjour, namaste !]

(3) [Chez nous, un baiser est surtout une preuve d'amour et de tendresse à l'égard de quelqu'un de cher,]

(4) [mais chez certains peuples c'est un salut courant.]

(5) [En Inde, les gens se saluent mains jointes sur la poitrine, comme s'ils priaient.]

(6) [Au Japon, les gens s'inclinent à plusieurs reprises, face à face, en joignant les mains.]

(7) [En France, les hommes faisaient le baisemain aux femmes mariées en signe de respect, et les jeunes filles la révérence,]

(8) [mais cette coutume se perd de plus en plus.]

Interprétation de l'intention de l'auteur n°1

(1) [En général, les gens se serrent la main droite quand ils se rencontrent ou se séparent, ou bien ils s'embrassent.]

(2) [Hello, bonjour, namaste !]

Lien par une reprise sémantique

Suivi thématique : l'objet -> sujet

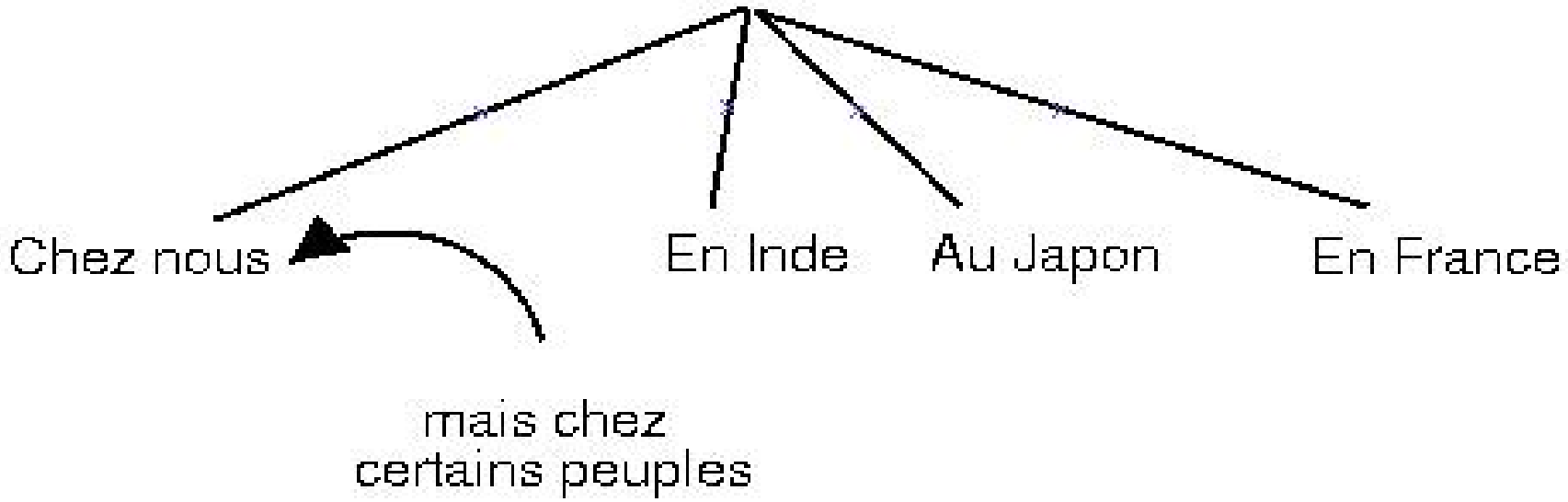
(3) [Chez nous, un baiser est surtout une preuve d'amour et de tendresse à l'égard de quelqu'un de cher.]

Valeur de vérité spécifiée dans un cadre précis marqu

(4) [mais chez certains peuples c'est un salut courant.]

(4) Commentaire à la proposition (3)

1ère interprétation



- Les autres propositions marqués par un introducteur peuvent se coordonner au premier cadre
- Ambiguïté co-référentielle : suggère que « chez nous » est différent de « en France »

Interprétation de l'intention de l'auteur n°2

lecture de (5)-(7), (3) utilisée pour mettre en avant (4

Hyperonyme
des classifieurs

(3) [Chez nous, un baiser est surtout une preuve d'amour et de tendresse à l'égard de quelqu'un de cher,]

(4) [mais chez certains peuples c'est un salut courant.]

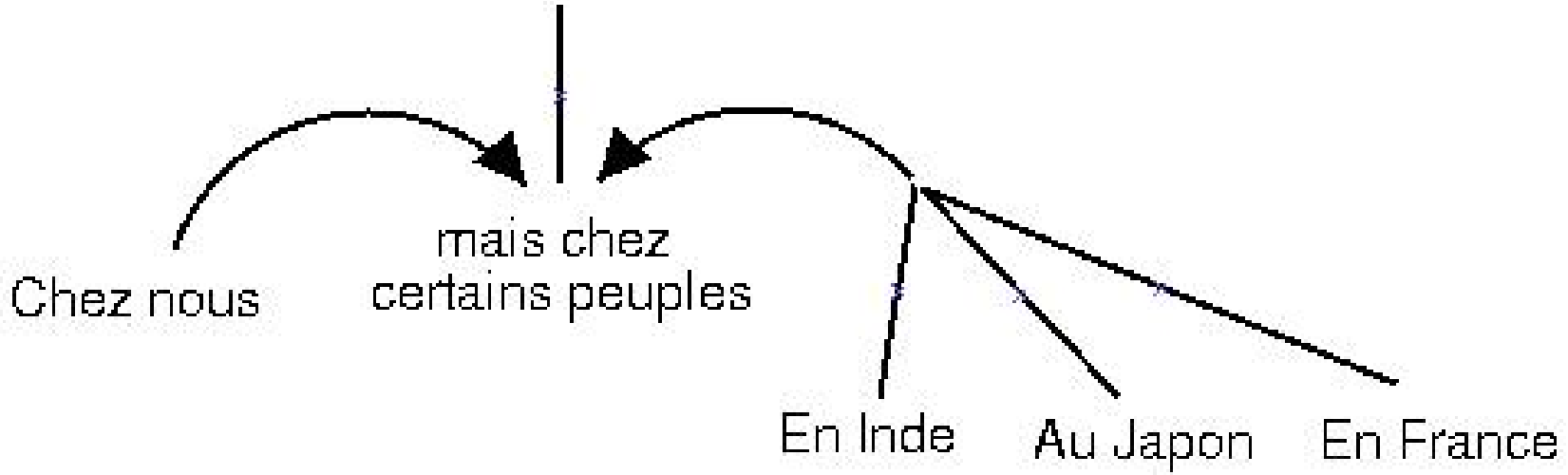
Reprise sémantiques

(5) [En Inde les gens se saluent mains jointes sur la poitrine, comme s'ils priaient.]

(6) [Au Japon, les gens s'inclinent à plusieurs reprises, face à face, en joignant les mains.]

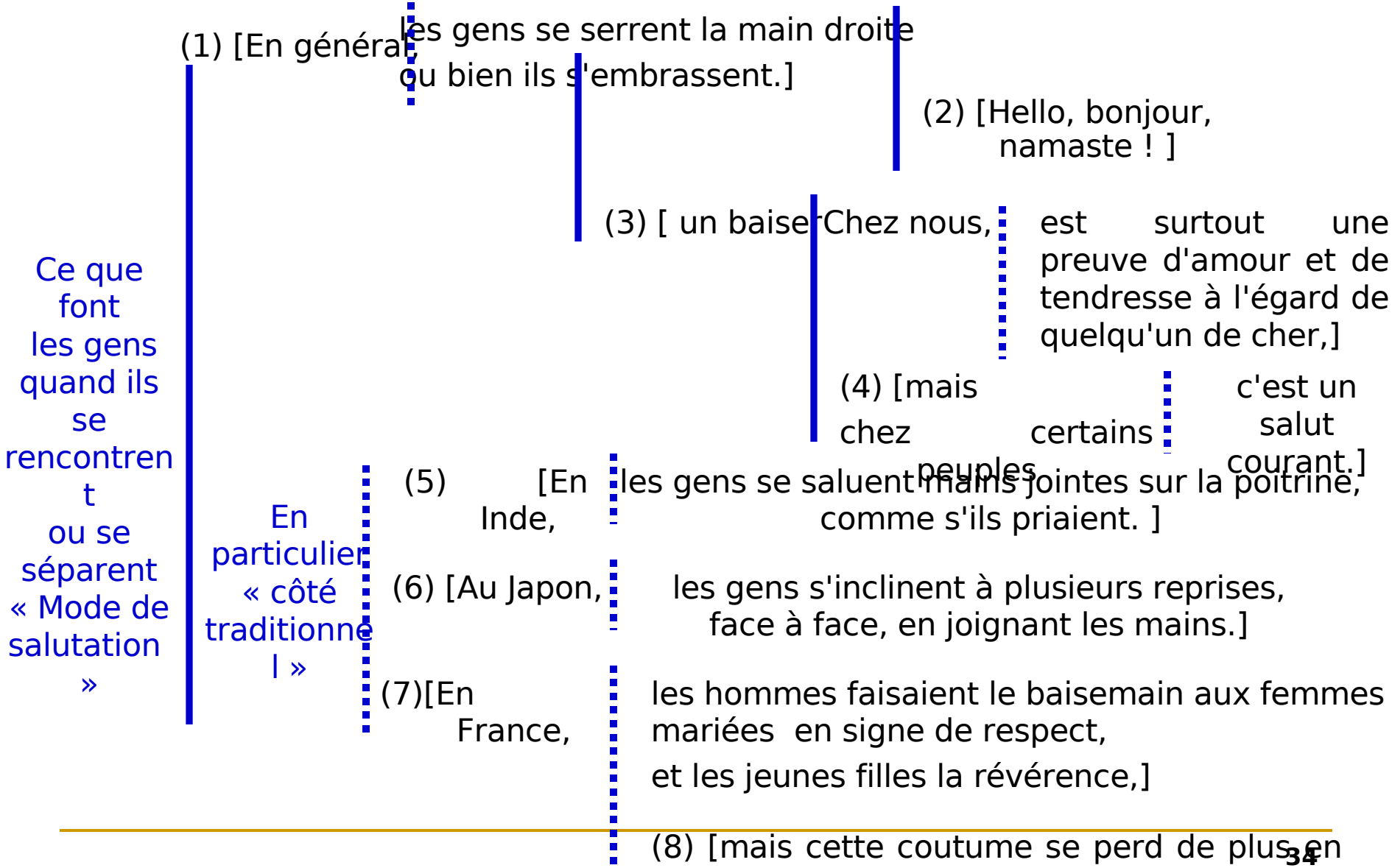
(7) [En France, les hommes faisaient le baisemain aux femmes mariées en signe de respect, et les jeunes filles la révérence,mais cette coutume se perd de plus en plus.]

2eme interprétation



- Toujours ambiguë : « chez nous » est différent de « en France »

Abstraction thématique



Abstraction thématique

- Élément fédérateur à la majorité des phrases
- Présenté dans la première phrase
- Combinée avec un cadre se qui rend difficile son interprétation individuelle
- (2) Commentaire n'existant pas sans (1)
- (3) Reprise d'un thème introduit dans (1) qui est spécifié selon un cadre « chez nous »
- (4) Même thème reprise avec « c'est », marque de coordination « et », autre cadre, correspond à un 2eme propos sur même thème
- (5) À (7) coordination : hyperonymie des classifieur « dans un pays », mais surtout parallélisme syntaxico-sémantique
- « IntroDeCadre, les gens se VERBE» aussi présent dans (7) avec gens hyperonyme de « homme » et de « jeunes filles »
- Malgré liens sémantiques et ambiguïté co-référentielle aucun lien direct entre (5-7) et (3) ou (4)
- (3) Et (4) traitent du baiser et (5-7) du mode de salutation ; le propos de (3-4) ne s'applique pas au thème de (5-7) par transitivité
- (5-7) implicitement en opposition au cadre général

Conclusion

- De multiples indices dans le discours
- Requiert des connaissances sémantiques et pragmatiques
- Requiert analyse sub-phrastique et processus d’abstraction
- De multiples organisations i.e. de multiples interprétations
- Si un humain éprouve des difficultés que peut on attendre des résultats d’une machine ?

Indices de cohésion et mécanismes de structuration

- Mise en Forme Matérielle
  - Information typo-dispositionnelle
- Les connecteurs et les expressions indicatives
  - « car » et « la raison en est que »
- Les chaînes lexicales (sémantique, anaphorique)
- Introducteurs de cadres
- Suivi thématique (e.g. rhème en thème)
- Parallélisme syntaxico-sémantique

Encadrement du discours [Charolles 97]

- Analyse du discours à travers l'étude d'un type de marques linguistiques
  - Les introducteurs de cadres

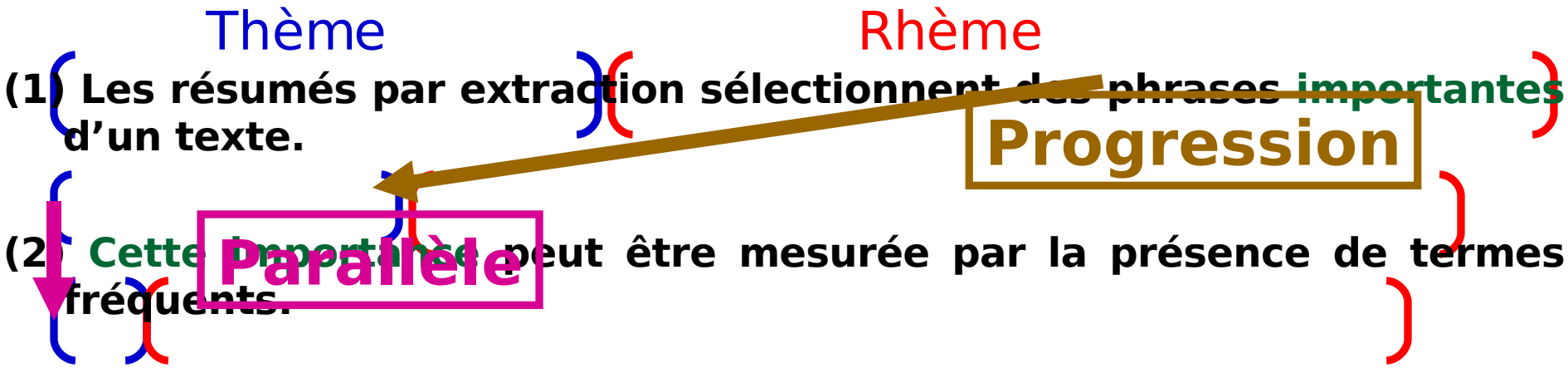
- Exemple
  - « En Corée du Sud », « Selon Pierre », « En ce qui concerne la méthode », « d'une part... d'autre part... »

- Propriétés
  - Situés généralement en début de phrases
  - Réunion d'une ou plusieurs propositions consécutives partageant une même relation sémantique
    - Temps/espace, thème, logique, etc.

Suivi thématique

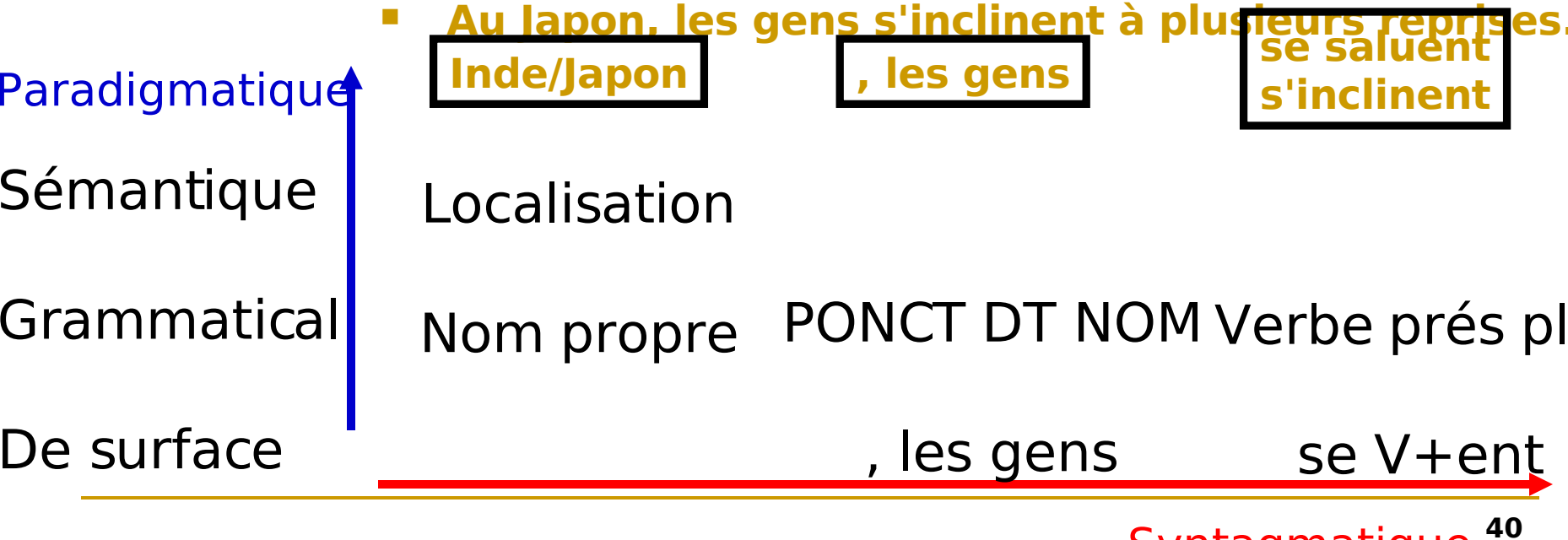
- Partition sémantique binaire des phrases
  - Thème : décrit le sujet et est lié au contexte
  - Rhème : information nouvelle associée à ce thème
- Type de suivi thématique

  [Daneš 74; Kruijff-Korbayová & Kruijff 96; Komagata 00; Steedman 00]
  - Progression : rhème(1) = thème(2)
  - Parallèle : thème(2) = thème(3)



Parallélisme syntaxico-sémantique

- Hypothèse
  - Des similarités syntaxico-sémantiques entre des énoncés signalent un même plan d'égalité entre ces énoncés (information, intention, temporel, etc.)
    - **En Inde, les gens se saluent mains jointes sur la poitrine.**
    - **Au Japon, les gens s'inclinent à plusieurs reprises.**



Mes propositions

- Analyse thématique
- Identification des descripteurs thématiques propres à un segment
- Identification de descripteurs de type d'information (sémantico-rhétorique) contenus dans une phrase
- Organisation thématique du discours à la fois aux niveaux global et local

Sommaire

- Systèmes existants
- Différents aspects d'un texte
- **Propositions**
  - **Identification des descripteurs thématiques propre à un segment**
  - Identification de descripteurs de type d'information (sémantico-rhétorique) contenu dans une phrase
  - Organisation thématique du discours à la fois aux niveaux global et local

Identification des descripteurs thématiques

- Thème : entité localement saillante
  - Entité du discours : ce dont on parle, ce à quoi on se réfère
- Problèmes
  - Variantes linguistiques
    - Morpho-syntaxiques, sémantiques et anaphoriques
    - E.g. « assister, assistance, assistent, aider, cette aide, elle, etc.»
  - Pertinence locale
    - Au niveau global : la fréquence est suffisante
    - Au niveau local : requiert davantage de critères

Identification des descripteurs thématiques

- Système de Résolution d'Anaphores
  - [Boguraev et Kennedy 97, Mitkov 98]

- Construction de chaînes lexicales
  - [Barzilay et Elhadad 97;Hirst et St-Onge 98]

Système de résolution d'anaphores

- Thèmes
  - Groupes Nominaux (GN) simples et pronoms personnels
- Principe
  - Calculer l'antécédent le plus probable
    - Proéminence syntaxique : sujet > verbe > objet ; GN défini > GN indéfini
    - Contraintes : même tête sémantique, genre/nombre, distance, etc.
- Poids d'un thème
  - Selon la « saillance » locale de toutes ses occurrences
- Particularité de l'approche
  - Les contraintes ne filtrent pas mais ordonnent
  - Anaphores les plus susceptibles d'être résolues automatiquement
    - GN avec démonstratif, pronoms personnels non indéfinis (filtrés)

Texte exemple pour l'évaluation

Le texte parle de :

Le texte traite des réseaux bayesiens et de leur expansion à travers l'entreprise Microsoft.

Les réseaux sont présentés comme aidant les utilisateurs dans diverses activités (diagnostic médical, résolution de problèmes de matériels informatiques, etc.).

Le texte rapporte comment Microsoft travaille à la maîtrise de cette technologie pour investir dans le marché des services de divertissement sur internet

SRA – intérêts

(1) When [Microsoft Vice President Steve Ballmer]^{i} first heard [his company]$^{i}_{m}$ was…

(2) After all, [Ballmer]$_{j}$ has billions of dollars of his own money^{j} in Microsoft stock, and entertainment isn't exactly the company^{m}'s strong point.

(3) But [Gates]$_{k}$ dismissed such reservations.

(4) Microsoft's [competitive advantage]$_{l}$, he^{k} responded, was its expertise^{l} in

Anaphore	FORTE <- - Antécédents ordonnés par préférence - - >					Validité
	Même tête et genre et nombre	Même tête	Même genre et nombre	Le plus fort poids abs et relatif	Le plus fort poids abs	
his own money^{j}	0	0	[Ballmer]$_{j}$	[Ballmer]$_{j}$	MVSP	V
the company^{m}	[his company]$^{i}_{m}$	[his company]^{i}	[Ballmer]$_{j}$	MSVSP	MVSP	V
he^{k}	0	0	[Gates]	such reservations	MVSP	V
its expertise^{l}	0	0	[competitive advantage]	[competitive advantage]	MVSP	V

SRA – évaluation

- Existant
  - Précision 75% à 90% (manuel d'utilisation et anaphore « it »)
- Évaluation manuelle

	9 anaphores « he, this, his »	18 en comptant « the »
Approche de base : le GN le plus récent	33%	16%
Mon système SRA	100%	66%

- Performances satisfaisantes

Construction de chaînes lexicales

- Existant
  - Chaînes de noms liés sémantiquement
- Thèmes
  - mots pleins (nom, verbe, adjectif)
- Rapprochement sémantique et grammatical
  - Wordnet et Celex
- Principe
  - Recherche des associations au niveau local
  - Sens conservé : le plus sollicité dans les associations
- L'élément le plus fréquent d'une chaîne sert de descripteur dans les segments où la chaîne se trouve présente

CCL évaluation

Rang	Poids	Chaînes lexicales
1	9	*Bayesian net, bayesian networks, bayesian system*
2	9	*Help, exploit, boost*
3	8	*Service, company*
4	7	*Support, activity, use, tool, work, user, answer, turn*

- Hors contexte, cohérence conceptuelle
- Pertinence de la considération de diverses formes grammaticales et de variations morpho-syntaxiques

Identification des descripteurs thématiques – Conclusion

- Amélioration des descripteurs thématiques

- Variations et extensions de techniques existantes

Sommaire

- Systèmes existants
- Différents aspects d'un texte
- **Propositions**
  - Identification des descripteurs thématiques propre à un segment
  - **Identification de descripteurs de type d'information (sémantico-rhétorique) contenu dans une phrase**
  - Organisation thématique du discours à la fois aux niveaux global et local

Identification du type d'information contenue dans une phrase

- Pertinence des [illegible]iques pour le typage et la sél[illegible]on

  [Paice 90, Minel Al 02, Teufel & Moens 02, Saggion & Lapalme 02]

  - Description du type d'information sémantico-rhétorique

« This approach is exemplified by Pollock and Zamora's ADAM system [1]. »

  - Indication d'éléments saillants

« The problems with this approach is that the extracted sentences do not always constitute a coherent text »

Méta-descripteurs - objectif

- Existant

  [Paice 90, Minel Al 01, Teufel & Moens 02; Saggion & Lapalme 02]

  - Listes orientées pour la sélection de certains types d'information
  - Pas de méthodes d'acquisition ni de classification automatique

➡ Proposer une méthode automatique d'acquisition

  - Portabilité

Caractéristiques des méta-descripteurs

- Nature changeante
  - Toutes les catégories grammaticales sont candidates
    - **“Nous montrons que”, “En conclusion”,**
  - Patrons syntaxiques non prévisibles
  - Variations morpho-syntaxiques et sémantiques
    - **[*focus* ADVERB? *on*] → [*focus* here *on*] | [*focus on*]**
    - **« This is exemplified by » , « An example of this is»**
    - **“This is illustrated by”**
  - Possibilités de combinaison
    - **[*described in*] + [AUTHOR REF] ou [*section* CREF]**

- Difficultés
  - Définir des patrons d’extraction
  - Capturer le caractère “méta”

Principe d'extraction

- Proposition :
  - N-grams avec des éléments génériques
    - Déterminant, Adverbe, Modal, Ponctuation
  - Utiliser les régularités textuelles du genre scientifique
    - Fréquence Inter-Documents (FID) d'un n-gram candidat avec un seuil (fixé à 7)
- Données
  - Corpus de genre et de domaine fixés
    - Computational-Language (Anglais)
    - Actes TALN et RECITAL (Français)
  - 100 documents chacun

Taille des “N-gram”

- Comment délimiter la taille de la forme valide ?
  - « *point of view* »
  - « *point of* », « *of view* »
  - « *the point of view of this* »
  - « *of view of this* »
- Sélection en fonction des fréquences des sur- et sous-séquences
  - Si deux expressions ont des fréquences proches
  - Et si l’une est une sous-séquence de l’autre
    - Alors la sur-quence est retenue comme la plus significative

Taille des “N-gram”

- Comment délimiter la taille de la forme valide ?
  - « *point of view* »
  - « *point of* », « *of view* »
  - « *the point of view of this* »
  - « *of view of this* »

Candidat	FID
…	
point of	35
point of view	35
…	

- Sélection en fonction des fréquences des sur- et sous-séquences
  - Si deux expressions ont des fréquences proches
  - Et si l’une est une sous-séquence de l’autre
    - Alors la sur-quence est retenue comme la plus significative

Taille des “N-gram”

- Comment délimiter la taille de la forme valide ?
  - « *point of view* »
  - « *point of* », « *of view* »
  - « *the point of view of this* »
  - « *of view of this* »

Candidat	FID
…	
point of	35
point of view	35
	…

- Sélection en fonction des fréquences des sur- et sous-séquences
  - Si deux expressions ont des fréquences proches
  - Et si l’une est une sous-séquence de l’autre
    - Alors la sur-quence est retenue comme la plus significative

Quelques exemples de

be show in table CD
be list in table CD
be give in table CD
on DT other hand
be DT result of
as show in fig
in DT study
in DT same
in DT paper
in DT literature

in DT context
in accordance with
important role in
serve as
series of
role of
study be

structure of
source of
solution of
similar to
property of
proceeding of
procedure be
problem of
problem be
present work

In the study
PREP DT NOM

Important role in
ADJ NOM PREP

Problem be
NOM VERB

Show that
VERB CONJ

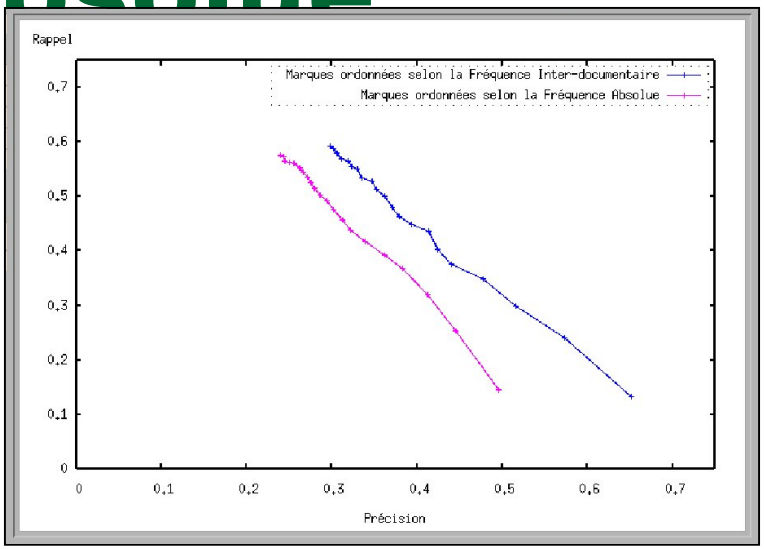
Évaluation

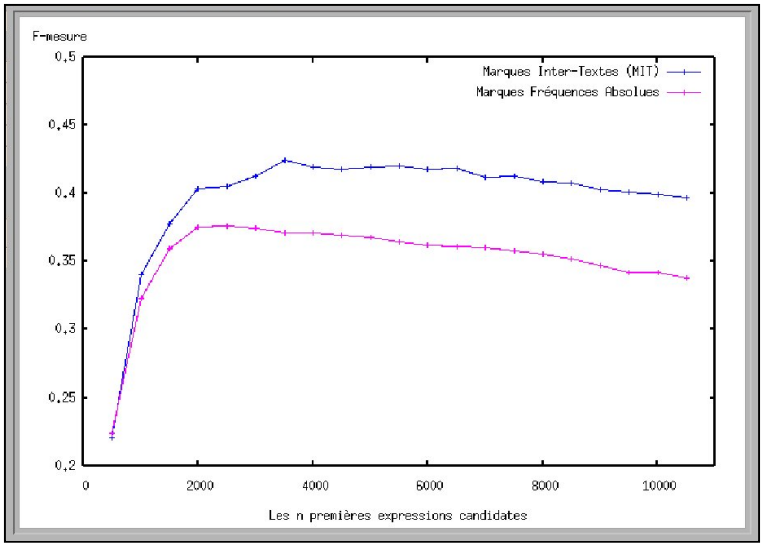
- Difficulté de comparer des méta-descripteurs entre elles
  - Représentations

    formes dérivées vs. lemmatisées, généralisations
    - maListe : « give RB » ;
    - Teufel : « GIVEN here » où GIVEN = {given, shown, noted…}
  - Délimitations
    - E.g. « be shown » « be shown in » « show that » : des marques ≠ !?

- Comparaison par rapport aux mots pleins
  - Essentiellement avec liste de [Teufel 99]

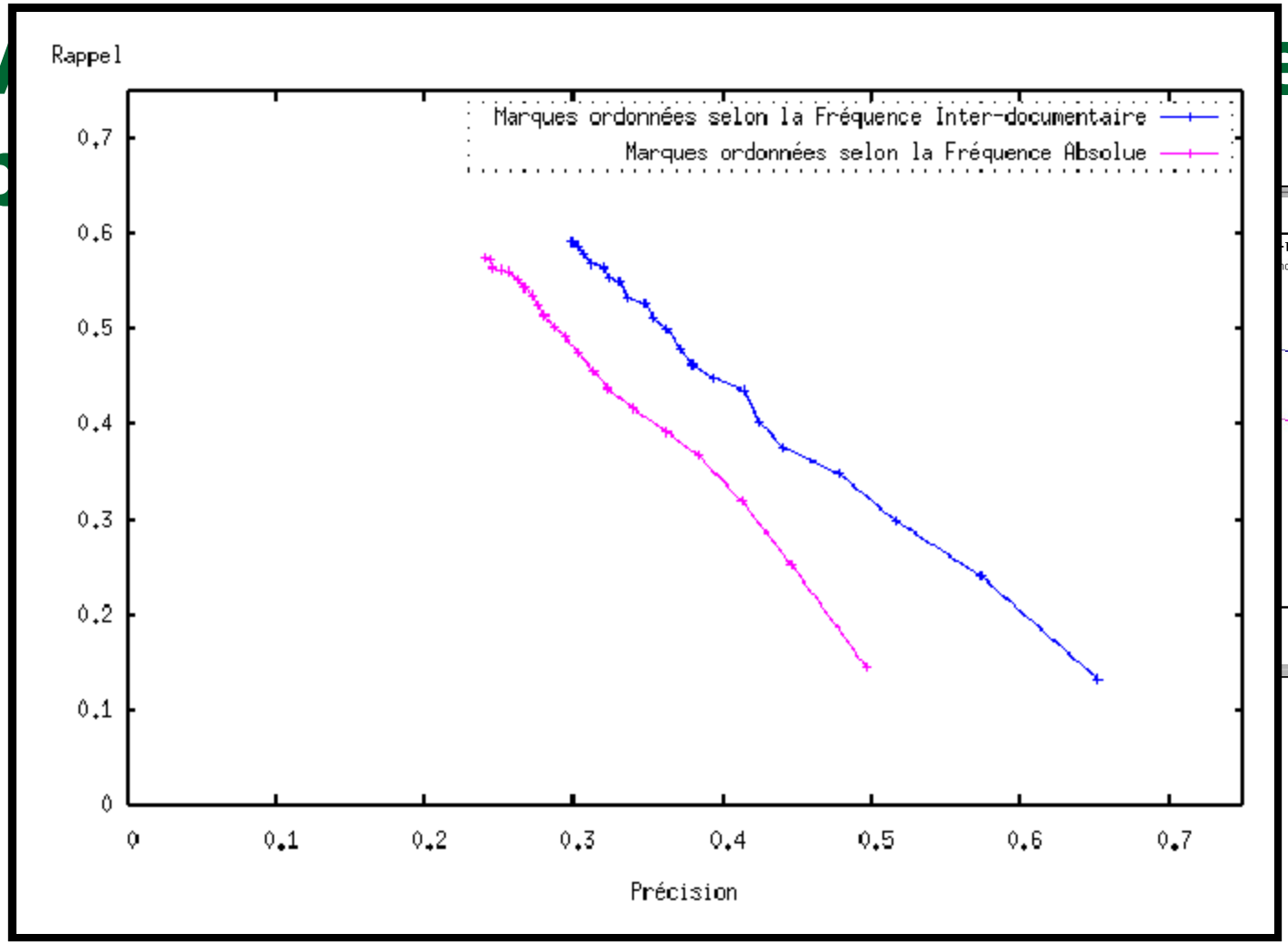
Évaluation - FID vs. Fréquence absolue

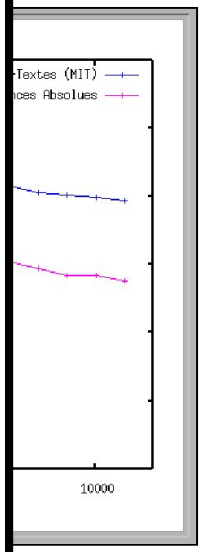




- Rappel équivalent mais précision toujours supérieure
- Filtrage effectif du lexique spécifique au domaine :
  - **« sentence, language, Grosz »**

Év ence ab





- Rappel équivalent mais précision toujours supérieure
- Filtrage effectif du lexique spécifique au domaine :
  - **« sentence, language, Grosz »**

Évaluation - Instanciation des patrons de Teufel

- Corpus de même nature issu d'une même source
- Marques de [Teufel 99]
  - Classes de termes

    e.g. GIVEN : given, shown, noted…
  - Regroupées en patrons

    e.g. WORK_N GIVEN {here | below | in this PRESENTATION_N}
- 53422 marques possibles
- 1229 présentes dans mon corpus
- 357 avec FID > 7 (seuil de sélection des n-grams)

La généralisation n'est pas forcément la solution

En faveur de courts n-grams peu généralisés

Évaluation - Comparaison mots pleins

	# mots pleins	maListe	Teufel	Saggion	Knott	Présents dans corpus	Corpus et FID >7	Corpus et FID >7 et maListe
maListe	760		36%	18%	8%			
Teufel	800	35%		14%	4%	84%	54%	63%
Saggion	194	72%	57%		10%	94%	85,05 %	84,84%
Knott	90	71%	36%	22%		95%	84,44 %	84,21%

- Saggion liste incomplète ; Knott connecteurs
- Meilleur rappel pour maListe
- Faible recoupement avec connecteurs
- Détectables / détectés
  - 46% des marques de Teufel : FID < 7, pertinence ?

Évaluation - Comparaison de marques à marques

- Problème :
  - Alignement des marques 2 à 2
- Solution :
  - Normalisation manuelle
  - Simplification des marques :
    - Pour chaque mot plein : un mot avant/après
- Exemple

show dans « *be show in figure CREF* » donne « *be show in* »

dans « *show that DT* » donne « *show that* »

Évaluation - Comparaison de marques à marques : exemple

- Comparaison des marques « give » et « show »
- Résultat
  - 2 *identiques*
  - 6 de Teufel sont *proches* de maListe (*par effacement)*

    E.g. NN GIVEN RB -> GIVEN RB
  - 5 de maListe sont *proches* de Teufel (*par ajout*)
  - 14 formes *nouvelles* dans maListe (1 ou 2 éléments nouveaux) ;

    E.g. be give in, show DT, RB give, PUNCT give DT, be shown in, CREF show, PP show, show that
- Conclusion
  - maListe permet le rappel des marques de Teufel
  - Formes non relevées manuellement

Sommaire

- Systèmes existants
- Différents aspects d'un texte
- **Propositions**
  - Identification des descripteurs thématiques propre à un segment
  - Identification de descripteurs de type d'information (sémantico-rhétorique) contenu dans une phrase
  - **Organisation thématique du discours à la fois aux niveaux global et local**

Organisation du contenu

- RÉsumé Guidé par les Attentes du Lecteur (REGAL)
  - Projet Cognitique ; LIMSI, LALLIC, LATTICE, CEA
  - Structure macroscopique
  - « Emboîtement » de segments thématiques

- Détection de Structures de Texte (DST)
  - Analyse au niveau microscopique
  - Recherche du point d'attache de la phrase entrante

REGAL - Hypothèse

Les approches linguistiques et numériques sont complémentaires et repèrent des segments de même nature

- Analyse fine : introducteurs de cadres
  - [Charolles 97]
  - E.g. « En ce qui concerne X, … » « Premièrement, … »

Ouvertures de cadres : débuts de segments

- Analyse globale : segmentation par cohésion lexicale
  - [Ferret al 97, Ferret 02]
  - Un segment = un regroupement de paragraphes contigus ayant une similarité lexicale suffisante

Marques de débuts et fins de segments

REGAL - Architecture du système

Pré-traitement

Etiquetage morpho-syntaxique + repérage des phrases et des paragraphes

Repérage des introducteurs par ConteXtO

Segmentation par cohésion lexicale : Anathème

Intégration

Structuration

En 1991, à la Station INRA de Dijon, Patrick Étiévant et Bruno Martin commençaient l'analyse du vin jaune, produit seulement dans le Jura. Le goût spécifique de ces vins résulte de leur technique d'élevage: on laisse le vin vieillir en tonneau pendant plusieurs années, sous un voile épais de levures Saccharomyces cerevisiae. Ce type de vin est également fabriqué en Alsace, en Bourgogne et à Gaillac sous le nom de vin de fleur ou vin de voile; il n'a d'équivalent à l'étranger que dans le xérès, les sherrys ou le tokay de Hongrie. Quelles molécules sont responsables de son goût caractéristique?

Les vins contiennent des centaines de composés volatils, dont un dixième sont aromatiques, de sorte que la détection des molécules responsables d'un arôme particulier est rarement difficile: chercher le coupable, parmi 300 suspects... Au début des années 1970, certains avaient cru que la sotolone (la 4 acétyl gamma butyrolactone) était l'arôme principal du vin jaune, mais, en 1982, Pierre Dubois, à Dijon, retrouva la sotolone dans des vins rouges: la molécule avait un alibi.

On soupçonna alors le 4,5 diméthyl 3 hydroxy 2(5H) furanone, ou sotolon, molécule construite autour d'un cycle de quatre atomes de carbone et d'un atome d'oxygène. Comme le sotolon et la sotolone sont, en concentrations minimes dans les vins de voile et, de surcroît, chimiquement instables, les chimistes dijonnais ont cherché à optimiser leur extraction afin de déterminer la molécule responsable du goût de jaune.

L'analyse la plus directe d'extrait de vin est la chromatographie: on injecte un échantillon dans un solvant que l'on vaporise et on fait traverser au mélange une colonne métallique intérieurement d'un polymère, qui retient les divers composés du mélange à des degrés divers; au bout de la colonne, on détecte la sortie des composés séparés. La première travail des chimistes fut la mise au point d'une variante de cette technique pour identifier les composés présents en quantités minimes dans des mélanges complexes.

Les chromatogrammes d'échantillons de vin furent alors comparés à ceux de solutions pures de sotolon et de sotolone de synthèse: le sotolon est ainsi présent entre 40 et 150 parties par milliard dans les sherrys; la sotolone semble moins spécifique, et ses concentrations sont supérieures dans les sherrys, ce qui explique pourquoi on l'a d'abord trouvée dans ces vins.

Enfin les dosages, complétés de tests sensoriels des fractions séparées, montrèrent que la sotolone, aux concentrations trouvées dans du savagnin (le cépage à partir duquel on fabrique le vin jaune), n'était perçue par les consommateurs ni dans les vins, ni dans des solutions modèles: la sotolone n'était pas la molécule caractéristique; le jugement était sans appel.

En 1992, les chimistes se consacrèrent alors complètement au sotolon, qui avait été observé dans des mélasses de canne à sucre, dans des graines de fenugrec, dans de la sauce de soja, dans du saké... Il était également présent dans certains vins botrytisés, c'est-à-dire faits à partir de raisins surmaturés et atteints par la pourriture noble: ce champignon, Botrytis cinerea, fait, par exemple, les sauternes ou les vins dits de vendanges tardives. Le sotolon n'a pas été trouvé dans les vins rouges ni dans les vins oxydés et, surtout, il fut déterminé que son seuil de perception était de 15 parties par milliard seulement.

Mieux encore, des tests de consommation montrèrent que les vins de voile étaient jugés typiques, avec une note de noix, quand la concentration en sotolon était forte dans ces vins. A plus forte concentration, les jurys de dégustation décrivirent une note de curry.

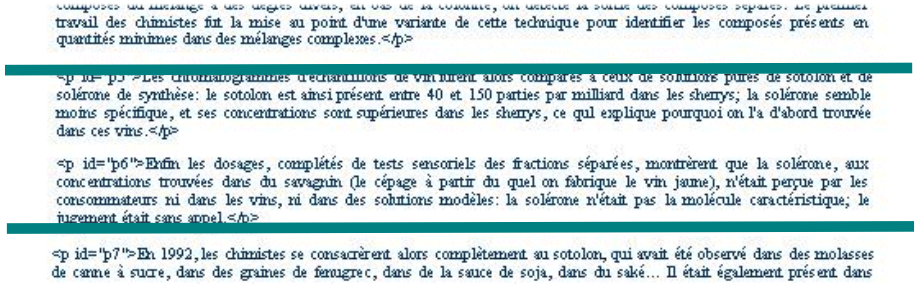
La piste du sotolon est aujourd'hui suivie par Élisabeth Guichard, qui a mis au point une méthode rapide de dosage; la concentration en sotolon dans le vin de paille (un vin préparé à partir de baies séchées sur des claies), qui n'avait pas été observée, est comprise entre 6 et 15 parties par milliard; le sotolon du vin jaune est synthétisé à la fin de la phase de croissance exponentielle des levures. Dans des vins vieillis respectivement un an, deux ans, trois ans, quatre ans, cinq ans et six ans, la quantité de sotolon est faible dans les débuts de la maturation et augmente notablement après quatre ans d'élevage, surtout dans les caves pas trop fraîches.

Des prélèvements à différentes profondeurs, sous le voile, dans les tonneaux, ont révélé que le sotolon est deux fois plus concentré au milieu et au fond des tonneaux que juste sous le voile. On suppose que le sotolon est indirectement produit par les levures du voile, quand le degré alcoolique est élevé: celles-ci transformeraient un acide aminé du vin en un cétoacide, qui serait libéré à la mort des levures, tomberait au fond du tonneau, puis une réaction chimique transformerait le cétoacide en sotolon, enrichissant d'abord le fond, puis le milieu, puis les couches supérieures du vin.

Puisque le sotolon est bien la molécule du goût de jaune, on cherche aujourd'hui des souches de levures qui ont la capacité d'en produire beaucoup; on cherche aussi les conditions qui favorisent la production de ce goût.

REGAL - Principes d'intégration

- Par un jeu de règles au cas par cas
- Exemple :
  - Marques d'intégration linéaires (« d'abord, ensuite, enfin) à cheval sur deux segments


travail des chimistes fut la mise au point d'une variante de cette technique pour identifier les composés présents en quantités minimes dans des mélanges complexes.

Les chromatogrammes d'échantillons de vin furent alors comparés à ceux de solutions pures de sotolon et de solérone de synthèse: le sotolon est ainsi présent entre 40 et 150 parties par milliard dans les sherrys; la solérone semble moins spécifique, et ses concentrations sont supérieures dans les sherrys, ce qui explique pourquoi on l'a d'abord trouvée dans ces vins.

Enfin les dosages, complétés de tests sensoriels des fractions séparées, montrèrent que la solérone, aux concentrations trouvées dans du savagnin (le cépage à partir du quel on fabrique le vin jaune), n'était perçue par les consommateurs ni dans les vins, ni dans des solutions modèles: la solérone n'était pas la molécule caractéristique; le jugement était sans appel.

En 1992, les chimistes se consacrèrent alors complètement au sotolon, qui avait été observé dans des molasses de canne à sucre, dans des graines de fenugrec, dans de la sauce de soja, dans du saké... Il était également présent dans

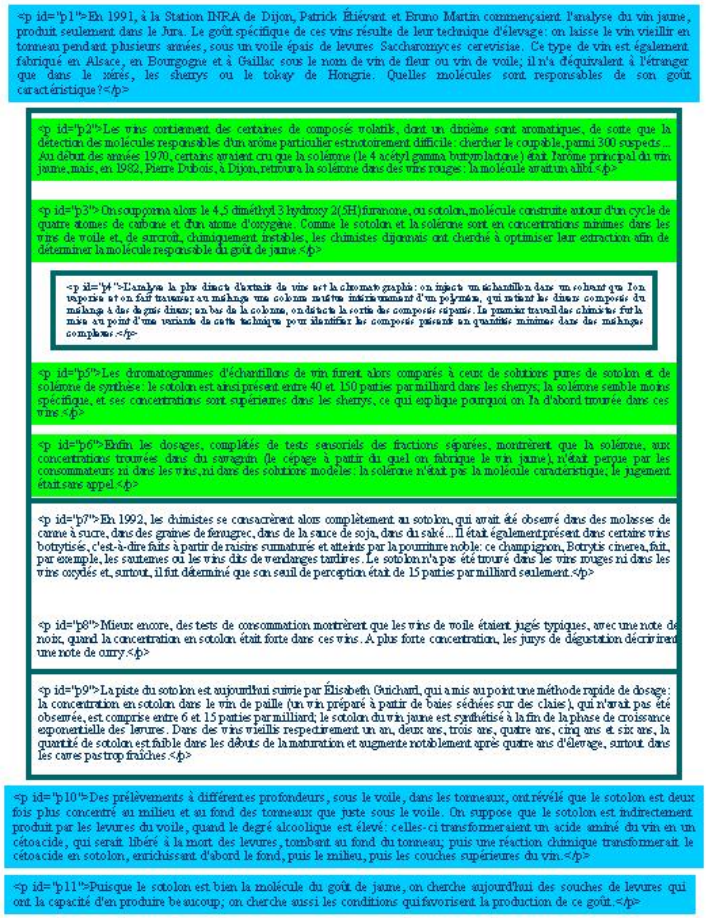
Réalignement par intégration de tous les indices dans un même segment

REGAL - Principe de structuration

- Repérage de structures emboîtées [Masson, 1998]
  - Digressions, développements d'aspects particuliers
  - Fréquent dans textes expositifs

Algorithme

- Recherche des 2 segments les plus liés et non-consécutifs
- Ré-appliquer récursivement pour les segments englobés ou non englobés restant



Évaluation - Concordance des ruptures

- Étude menée avec Denis Vigier (linguiste, LATTICE)

- Observation de paragraphes saturés en cadres (éventuellement avec phrase introductive)
  - I.e. $\forall$ phrases $\in$ cadre

- Corpus : 10 textes ; 15 paragraphes
  - Monde diplomatique, la Recherche, Atlas français scolaire, Web
- Détection de changement lexical [Hearst 97]
  - Fenêtre 20 mots pleins (2 phrases) ; décalage 6

Évaluation - Concordance des ruptures

- Concomitance rupture de cohésion lexicale et cadre
  - 70% concordance
  - 25% cadres seuls
  - 5% intra-cadre

- Limites de segmentation par cohésion lexicale
  - Suivant taille des unités comparées ⇒ agrégation/rupture abusive $\Rightarrow\!\!\!\!\!\neq$
  - Cohésion lexicale $\neq\!\!\!\!\!\Rightarrow$ cohérence (thématique)
  - Absence de cohésion lexicale non-cohérence

- Conclusion
  - Cohésion lexicale des cadres du discours
  - Besoin d'autres indices et d'une analyse fine pour détecter phénomènes discursifs locaux

Sommaire

- Systèmes existants
- Différents aspects d'un texte
- **Propositions**
  - Identification des descripteurs thématiques propre à un segment
  - Identification de descripteurs de type d'information (sémantico-rhétorique) contenu dans une phrase
  - **Organisation thématique du discours à la fois aux niveaux global et local**

DST - Modélisation du discours

- Hypothèse
  - Information entrante est une spécification de ce qui précède ou un aspect complémentaire

- Une phrase entrante est rattachée au discours
  - Soit par une relation de subordination
  - Soit par une coordination

- Des indices permettent de les repérer

DST – type de relations

(1) Les résumés par extraction sélectionnent des phrases d'un texte source selon leur importance.

(2) Les critères d'importance incluent la présence de termes fréquents, des mots clefs tels que « en résumé », « meilleur », et la position de la phrase dans le texte.

(3) Cette approche est illustrée par le système ADAM.

(4) Un autre exemple est donné par [2].

(5) Le problème de cette approche est que les phrases extraites ne constituent pas toujours un texte cohérent du fait d'anaphores ambiguës.

DST – type de relations

(1) Les résumés par extraction sélectionnent des phrases d'un texte source selon leur importance.

(2) Les critères d'importance incluent la présence de termes fréquents, des mots clefs tels que « en résumé », « meilleur », et la position de la phrase dans le texte.

Subordination

(3) Cette approche est illustrée par le système ADAM.

(4) Un autre exemple est donné par [2].

(5) Le problème de cette approche est que les phrases extraites ne constituent pas toujours un texte cohérent du fait d'anaphores ambiguës.

DST – type de relations

(1) Les résumés par extraction sélectionnent des phrases d'un texte source selon leur importance.

(2) Les critères d'importance incluent la présence de termes fréquents, des mots clefs tels que « en résumé », « meilleur », et la position de la phrase dans le texte.

Subordination

(3) Cette approche est illustrée par le système ADAM.

(4) Un autre exemple est donné par [2]

Coordination

(5) Le problème de cette approche est que les phrases extraites ne constituent pas toujours un texte cohérent du fait d'anaphores ambiguës.

Construction d'un modèle de prédiction de relations

- Annotation manuelle de relations de coordination et de subordination
  - 5 articles scientifiques de 8 à 10 pages
  - 1190 couples de phrases mises en relation
- Caractérisation automatique de ces couples
  - Progression thématique et cohésion lexicale
  - Présence de classes de connecteurs
  - Parallélisme syntaxico-sémantique
- Application d'un algorithme d'apprentissage
  - Arbre de décision C4.5

Évaluation du modèle de prédiction

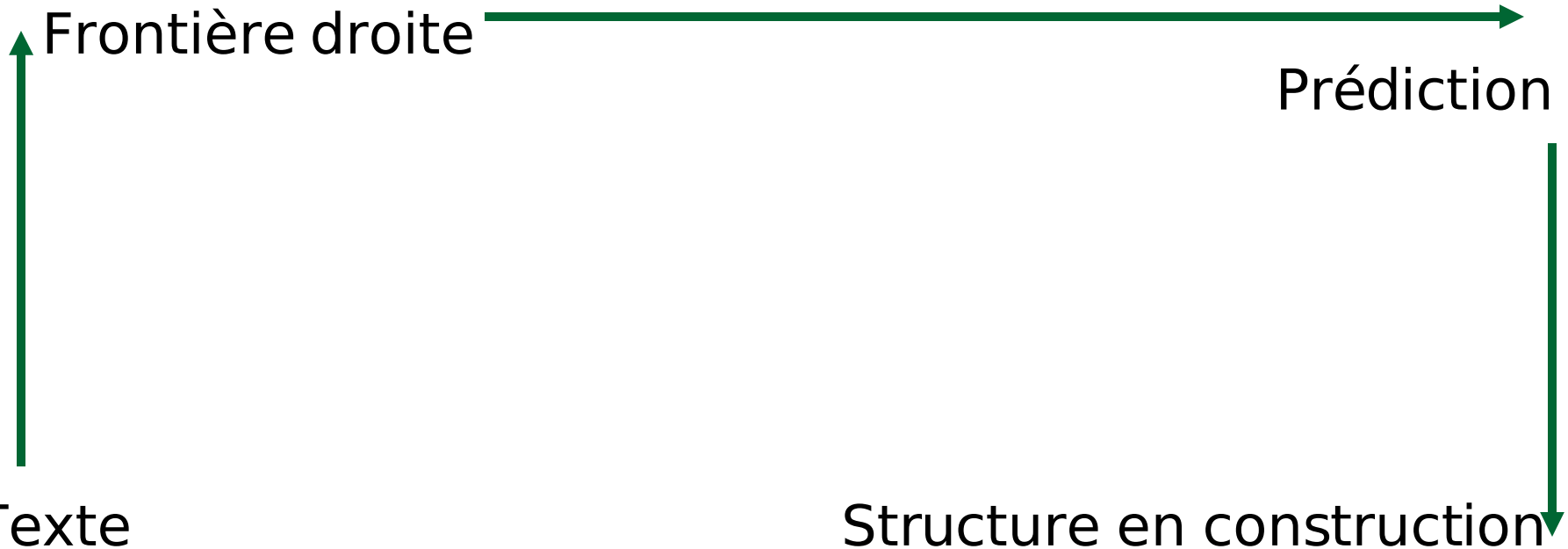
- Validation croisée entre 10 partitions
- Système de base : probabilité de la relation la plus fréquente

Coordination et subordination	
Algorithme	Précision
Système de Base	53,10%
DST	**58,57%**

Seulement la subordination	
Algorithme	Précision
Système de Base	69,83%
DST	**76,35%**
Choi 02	73,61%

➡ Performances > aux systèmes de base et à l'existant

DST – Détection de l'organisation



(1) Les résumés par extraction sélectionnent des phrases d'un texte source selon leur importance.

(2) Les critères d'importance incluent la présence de termes fréquents, des mots clefs tels que « en résumé », « meilleur », et la position de la phrase dans le texte.

(3) Cette approche est illustrée par le système ADAM.

(4) Un autre exemple est donné par [2].

(5) Le problème de cette approche est que les phrases extraites ne constituent pas toujours un texte cohérent du fait d'anaphores ambigües.

DST – Détection de l'organisation

Frontière droite

Prédiction

1

(1) Les résumés par extraction sélectionnent des phrases d'un texte source selon leur importance.

On empile

Texte

Structure en construction

(1)

(2) Les critères d'importance incluent la présence de termes fréquents, des mots clefs tels que « en résumé », « meilleur », et la position de la phrase dans le texte.

(3) Cette approche est illustrée par le système ADAM.

(4) Un autre exemple est donné par [2].

(5) Le problème de cette approche est que les phrases extraites ne constituent pas toujours un texte cohérent du fait d'anaphores ambiguës.

DST – Détection de l'organisation

Frontière droite

Prédiction

(1) Les résumés par extraction sélectionnent des phrases d'un texte source selon leur importance.

On défile

(2) Les critères d'importance incluent la présence de termes fréquents, des mots clefs tels que « en résumé », « meilleur », et la position de la phrase dans le texte.

Texte

Structure en construction

(1)

(3) Cette approche est illustrée par le système ADAM.

(4) Un autre exemple est donné par [2].

(5) Le problème de cette approche est que les phrases extraites ne constituent pas toujours un texte cohérent du fait d'anaphores ambiguës.

DST – Détection de l'organisation

Frontière droite

Prédiction

(1) Les résumés par extraction sélectionnent des phrases d'un texte source selon leur importance.

3 Application du modèle de prédiction

(2) Les critères d'importance incluent la présence de termes fréquents, des mots clefs tels que « en résumé », « meilleur », et la position de la phrase dans le texte.

Texte

Structure en construction

(1)

(3) Cette approche est illustrée par le système ADAM.

(4) Un autre exemple est donné par [2].

(5) Le problème de cette approche est que les phrases extraites ne constituent pas toujours un texte cohérent du fait d'anaphores ambiguës.

DST – Détection de l'organisation

Frontière droite

Prédiction

(1) Les résumés par extraction sélectionnent des phrases d'un texte source selon leur importance.

Application du modèle de prédiction

(2) Les critères d'importance incluent la présence de termes fréquents, des mots clefs tels que « en résumé », « meilleur », et la position de la phrase dans le texte.

4

Relation de subordination

Texte

Structure en construction

(1)

(3) Cette approche est illustrée par le système ADAM.

(4) Un autre exemple est donné par [2].

(5) Le problème de cette approche est que les phrases extraites ne constituent pas toujours un texte cohérent du fait d'anaphores ambiguës.

DST – Détection de l'organisation

Frontière droite

Prédiction

(1) Les résumés par extraction sélectionnent des phrases d'un texte source selon leur importance.

Application du modèle de prédiction

(2) Les critères d'importance incluent la présence de termes fréquents, des mots clefs tels que « en résumé », « meilleur », et la position de la phrase dans le texte.

Relation de subordination

Texte

Structure en construction

(1) ← (2)

5

Construction de la structure

(3) Cette approche est illustrée par le système ADAM.

(4) Un autre exemple est donné par [2].

(5) Le problème de cette approche est que les phrases extraites ne constituent pas toujours un texte cohérent du fait d'anaphores ambiguës.

DST – Détection de l'organisation

Frontière droite

Prédiction

(1) Les résumés par extraction sélectionnent des phrases d'un texte source selon leur importance.

(2) Les critères d'importance incluent la présence de termes fréquents, des mots clefs tels que « en résumé », « meilleur », et la position de la phrase dans le texte.

Application du modèle de prédiction

(3) Cette approche est illustrée par le système ADAM.

4

Aucune relation

Texte

Structure en construction

(1) ← (2)

(4) Un autre exemple est donné par [2].

(5) Le problème de cette approche est que les phrases extraites ne constituent pas toujours un texte cohérent du fait d'anaphores ambiguës.

DST – Détection de l'organisation

Frontière droite

Prédiction

(1) Les résumés par extraction sélectionnent des phrases d'un texte source selon leur importance.

Application du modèle de prédiction

(3) Cette approche est illustrée par le système ADAM.

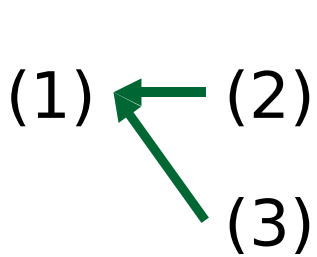
Relation de subordination

Texte

Structure en construction

(4) Un autre exemple est donné par [2].

(5) Le problème de cette approche est que les phrases extraites ne constituent pas toujours un texte cohérent du fait d'anaphores ambiguës.



Construction de la structure

DST – Détection de l'organisation

Frontière droite

Prédiction

(1) Les résumés par extraction sélectionnent des phrases d'un texte source selon leur importance.

(3) Cette approche est illustrée par le système ADAM.

Application du modèle de prédiction

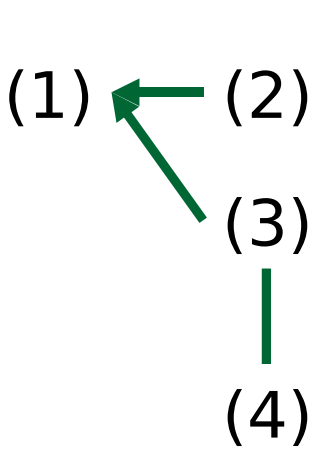
(4) Un autre exemple est donné par [2].

Relation de coordination

Texte

Structure en construction

(5) Le problème de cette approche est que les phrases extraites ne constituent pas toujours un texte cohérent du fait d'anaphores ambiguës.



Construction de la structure

DST - conclusion

- Détection de l’organisation informationnelle locale
- Modèle de prédiction permet de détecter des relations ; il est indépendant des hypothèses de structuration (i.e. de l’algorithme)

Conclusion

- Annotation de la structure thématique du discours au niveau global et local

  ➡ Offre la possibilité de naviguer dans le document

- Indexation de phrases et de segments de texte avec des descripteurs thématiques et méta-

  ➡ Permet une abstraction du texte et par là en facilite la visualisation

Étude globale de différents aspects d'analyse d'un texte

Perspectives

- Acquisition et structuration de connaissances
- Analyse du document notamment par connaissances acquises

- Approche : combiner méthodes statistiques et analyses linguistiques

Perspectives

- Description
  - Identifier plus précisément les entités thématiques
  - Classer les méta-descripteurs selon le type d'information
  - Patrons de reconnaissance plus précis et descriptifs
    - Combiner des thèmes avec des méta-
- Structuration
  - Combiner structure locale et globale
  - Enrichir le modèle en intégrant le plan rhétorique et visuel
- Évaluer l'apport de nos informations pour l'utilisateur
- Applications

Dans 20 ans…

Document + Utilisateur + Système

=

Système idéal

Merci