Annotation de structures discursives

L'expérience ANNODIS

L.-M. Ho-Dac

UE TAL – 18 novembre 2013

Plan

1. Éléments de définition
2. le projet ANNODIS
3. Analyse de Annotations
4. Annoter du "contenu" ou des documents ?
5. Conclusion

Structures discursives et TAL

- Accès au contenu (extraction de connaissances, compréhension de texte, navigation, résumé automatique, etc.)
- Génération de documents (résumé automatique, bulletin météo, aide à la rédaction, amélioration de la lisibilité, etc.)

"Discourse structures are the patterns that one sees in multi-sentence (multi-clausal) texts. Recognizing these pattern(s) in terms of the elements that compose them is essential to correctly deriving and interpreting information in the text." (Webber et al. 2012)

Force est néanmoins de constater que l'impact de ces travaux d'un point de vue applicatif reste globalement limité et en tout cas, moins étendu qu'il ne pourrait l'être (O. Ferret & N. Hernandez, introduction à la journée ATALA discours et TAL – mai 2012)

Structures discursives et TAL

Force est néanmoins de constater que l'impact de ces travaux d'un point de vue applicatif reste globalement limité et en tout cas, moins étendu qu'il ne pourrait l'être (O. Ferret & N. Hernandez, introduction à la journée ATALA discours et TAL – mai 2012)

- Faible stabilisation des modèles
- Peu de validation empirique des modèles
- Des phénomènes linguistiques complexes

Complexité des structures discursives

source : Webber et al. (2012) [Webber et al., 2012]

4 types d'éléments structurant

- les structures topicales (chaînes de références, segmentation thématique, extraction de topique)
- les structures "intentionnelles" ([Grosz and Sidner, 1986], les *moves* [Swales, 1981], le zonage rhétorique [Teufel and Moens, 2002])
- les structures événementielles (temporalité et définition des EDUs)
- les relations de discours (relations de contraste, de causalité, représentations en arbre (RST, SDRT))

qui se combinent au sein de mêmes structures signalées (plus ou moins explicitement) par des traits linguistiques "emmêlés".

Complexité des structures discursives

blue source : Webber et al. (2012) [Webber et al., 2012]

Un signalisation multi-fonctionnelle

Ces traits ont un rôle "instructionnel" en signalant au lecteur (plus ou moins explicitement) :

- une (dis)continuité avec le discours précédent (topicale, intentionnelle, événementielle)
- une relation de discours entre des segments

source : conférence MAD 2010 : Signalling text organisation (CFP)

> *Signalling text organisation refers to the observation that within texts, certain features or elements seem to have a special instructional role with regard to text organisation.*

Appoches top-down et bottom-up

source : Biber et al. (2007) [Biber et al., 2007]

Étapes pour l'analyse en corpus de structures discursives

- Définir l'unité minimale
- Segmenter le texte en unités minimales
- Catégoriser les unités minimales
- Analyser et décrire les caractéristiques linguistiques des unités catégorisées
- Décrire les structures qui regroupent les unités décrites
- Décrire les patrons linguistiques qui signalent ces structures

Annotation de structures discursives

source : Sophie Rosset – séminaire CLLE-ERSS, mars 2013

Annoter : poser une marque interprétative [Leech, 1997]

> *[corpus annotation] can be defined as the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data. "Annotation" can also refer to the end-product of this process.*

Annoter : ajouter une information [Habert, 2005]

> *L'annotation consiste à ajouter de l'information (une interprétation stabilisée) aux données langagières*

Annotation de structures discursives

source : Sophie Rosset – séminaire CLLE-ERSS, mars 2013

Pourquoi annoter des corpus ?

- Développer des systèmes d'analyse automatique et les évaluer (RefGen) : annotation de chaînes de référence et apprentissage automatique
- Évaluer un modèle et son applicabilité (ANNODIS_rr) : les annotations permettent de valider empiriquement un modèle théorique préétabli
- Observer des phénomènes ciblés (ANNODIS_me) : les annotations permettent de rassembler les réalisations linguistiques d'un objet potentiellement flou pour stabiliser sa définition

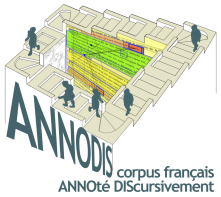
Annoter équivaut à

- catégoriser (chunk = expression coref ou pas)
- segmenter et catégoriser (ANNODIS_rr)
- localiser et catégoriser (ANNODIS_me)

Plan

1. Éléments de définition
2. le projet ANNODIS
   - Besoins propres aux deux approches
   - Post-Traitement des Annotations
   - Annotations diffusées
3. Analyse de Annotations
4. Annoter du "contenu" ou des documents ?
5. Conclusion

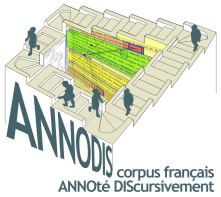
ANNODIS : une expérience d'annotation discursive



Constitution d'un corpus diversifié de français écrit
Enrichi d'annotations concernant le niveau discursif
Ressource ANNODIS
+
développement d'outils d'aide à l'annotation et à la navigation
GLOZZ et outils de prémarquage

- Projet ANR , Programme SHS : "Corpus et outils pour la recherche en SHS" - 2008-2010
- **Participants** : CLLE-ERSS (Toulouse), IRIT (Toulouse) et le GREYC (Caen)
- **Responsable** : M.-P. Péry-Woodley (CLLE-ERSS)
- [Péry-Woodley et al., 2011]

ANNODIS : une expérience d'annotation discursive



Constitution d'un corpus diversifié de français écrit
Enrichi d'annotations concernant le niveau discursif
Ressource ANNODIS
+
développement d'outils d'aide à l'annotation et à la navigation
GLOZZ et outils de prémarquage

- Projet ANR , Programme SHS : "Corpus et outils pour la recherche en SHS" - 2008-2010
- **Participants** : CLLE-ERSS (Toulouse), IRIT (Toulouse) et le GREYC (Caen)
- **Responsable** : M.-P. Péry-Woodley (CLLE-ERSS)
- [Péry-Woodley et al., 2011]

ANNODIS : une expérience d'annotation discursive

Originalité : aborder l'organisation discursive selon deux approches complémentaires :

Approche "descendante" (ANNODIS_me) : texte = unité fonctionnelle et sémantique

Approche "ascendante" (ANNODIS_rr) : texte = assemblage "cohérent" d'unités élémentaires

ANNODIS : une expérience d'annotation discursive

Originalité : aborder l'organisation discursive selon deux approches complémentaires :

Approche "descendante" (ANNODIS_me) : texte = unité fonctionnelle et sémantique

- localiser et délimiter des structures multi-échelles* (objets flous, pas d'unité minimale)* *2 phrases < X < ++ sections*
- annoter les structures énumératives et leurs composants ainsi que les chaînes topicales
- décrire les patrons signalant ces structures <=> un accès aux discontinuités en discours

Approche "ascendante" (ANNODIS_rr) : texte = assemblage "cohérent" d'unités élémentaires

Structures multi-échelles

Une structure énumérative (**SE** en abrégé) est un segment de texte défini par une structuration interne mettant en jeu les sous-segments suivants :

1. Une amorce (optionnelle) : segment qui introduit l'énumération ;
2. Plusieurs items : segments qui constituent l'énumération (au moins deux items doivent être identifiés pour qu'une structure soit annotée) ;
3. Une clôture (optionnelle) : segment qui résume ou clôt l'énumération.

Les chaînes topicales (**CT** en abrégé), quant à elles, sont définies comme des segments de texte regroupant des phrases reliées par un référent commun. Ce référent doit être exprimé en position pré-verbale (i.e. potentiellement topicale) dans plusieurs des phrases du segment.

En revanche, *le régime* a patronné **trois** **formations** importantes.	CT1	SE	amorce
Bien qu'*il* ait réduit de moitié les effectifs de la Garde républicaine, passée de 150 000 à 70 000 hommes, *il* a veillé à en reconstituer les précieuses unités mécanisées et blindées. Pour ce faire *il* a eu recours, outre quelques importations illégales, à la cannibalisation des matériels rescapés du pilonnage, souvent au détriment de l'armée.			item 1
Le régime s'est **aussi** détourné de son aviation au profit d'un Corps aérien plus opérationnel. *Il* en a consolidé les escadrons habitués à opérer en coordination étroite avec la Garde républicaine.			item 2
L'importation de pièces de rechange s'est d'ailleurs révélée plus facile pour les hélicoptères, qui bénéficient d'un double statut civil et militaire.			
Enfin, les incursions quasi quotidiennes des avions anglo-saxons dans les zones d'exclusion aérienne et les "frappes" régulières de missiles de croisière ont stimulé l'intérêt porté par Saddam Hussein à la Défense aérienne, rénovée et amadouée par des privilèges semblables à ceux dont bénéficie la Garde républicaine. On ne saurait souligner assez que c'est là la principale disposition militaire classique prise par l'Irak contre un adversaire étranger.			item 3
En somme, *le régime* a remodelé et réorienté ses forces armées pour aller vers un système plus sûr et plus compact, au caractère répressif et défensif.	CT2		clôture
Dans cette configuration, *il* ne représente plus guère, en dépit des			

ANNODIS : une expérience d'annotation discursive

Originalité : aborder l'organisation discursive selon deux approches complémentaires :

Approche "descendante" (ANNODIS_me) : texte = unité fonctionnelle et sémantique

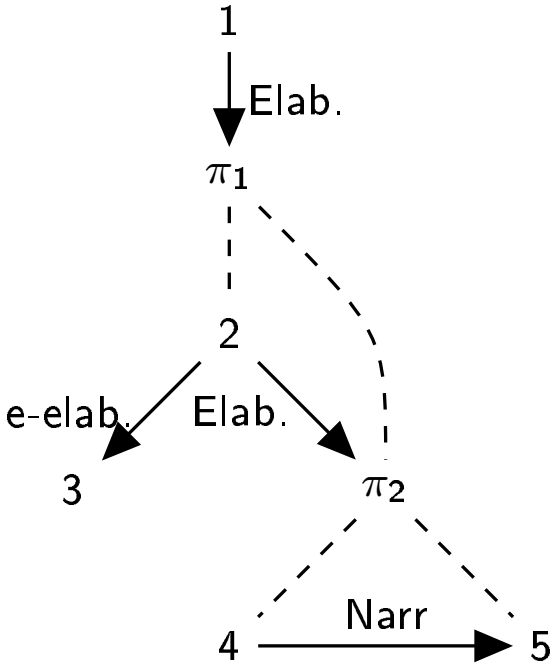
Approche "ascendante" (ANNODIS_rr) : texte = assemblage "cohérent" d'unités élémentaires

- segmenter le contenu des textes en unités élémentaires (EDU)
- annoter exhaustivement les relations de discours entre unités (EDU et CDU) selon un modèle préétabli (SDRT*)
- valider le modèle et décrire les marqueurs des différentes relations de discours

Relations de discours

Structure complète d'un discours : ensemble d'unités discursives reliées par des relations de cohérence (dites rhétoriques). Une structure discursive peut donc être représentée par un graphe, où les nœuds sont des unités discursives et où les arcs étiquetés représentent les relations rhétoriques entre ces unités.
Les unités discursives peuvent être simples, les EDU ou *Elementary Discourse Units*, ou complexes, les CDU ou *Complex Discourse Units*, qui regroupent plusieurs EDU en une structure discursive (un graphe).

[Milutinovic devant le TPI.]_1 [L'ancien président de Serbie Milan Milutinovic, [inculpé aux côtés de l'ex-chef d'Etat yougoslave Slobodan Milosevic pour crimes de guerre au Kosovo,]_3 s'est rendu volontairement hier au Tribunal pénal international pour l'ex-Yougoslavie de La Haye.]_2 [Arrivé aux Pays-Bas dans un avion du gouvernement yougoslave,]_4 [M. Milutinovic a été incarcéré au centre de détention du Tribunal pénal international en début d'après-midi.]_5



Les nœuds correspondent aux unités discursives : les EDU représentées par leur numérotation et les CDU par un nœud étiqueté π_n. Les arêtes avec flèches représentent les relations rhétoriques, les arêtes en pointillé sans flèches représentent l'inclusion d'EDU dans un CDU. Elab. = Élaboration, e-elab = Élaboration d'entité, Narr. = Narration.

Besoins propres aux deux approches

Différents besoins en terme de données à annoter

Relations rhétoriques

- Textes courts (complexité de la tâche, exigence d'exhaustivité)
- Textes nus présentés sans mise en forme (sauts de paragraphe, contrastes typographiques etc. éliminés)

Structures multi-échelles

- Textes longs et structurés (structures de haut niveau, prise en compte de la structure de document)
- Textes pré-marqués (traits supposés pertinents) et présentés avec mise en forme

Le corpus ANNODIS

1. textes libres de droit
2. un corpus diversifié => études comparatives et pertinence multi-genre du modèle d'annotation
3. des textes expositifs, davantage enclins à des organisations complexes et des relations variées
4. courts pour permettre une annotation exhaustive des relations de discours
5. longs pour permettre une annotation des structures multi-échelle
   => organisations + complexes
   => signalisation via éléments variés (structure visuelle)

Le corpus ANNODIS

id	source (S), genre (G), type majeur (T)	Struct.Doc. de document	nombre de	
NEWS	S = *Est Républicain* G = brèves T = narratif	faible	articles mots mots/texte	39 10 000 250
WIK1	S = *extraits Wikipedia* G = art. encyclopédique T = expositif	faible	extraits mots mots/texte	30 11 000 412
WIK2	S = *articles Wikipedia entiers* G = art. encyclopédique T = expositif	forte	articles mots mots/texte	30 231 000 7 700
LING	S = *CMLF, colloque de linguistique* G = articles de recherche T = expositif	moyenne	articles mots mots/texte	25 169 000 6 760
GEOP	S = *IFRI, institut de géopolitique* G = rapports et articles T = argumentatif	moyenne	articles mots mots/texte	32 266 000 8 325
			TOTAL : 156 textes/extraits, 687 000 mots	

Besoins propres aux deux approches

Différents besoins en terme d'outils pour annoter

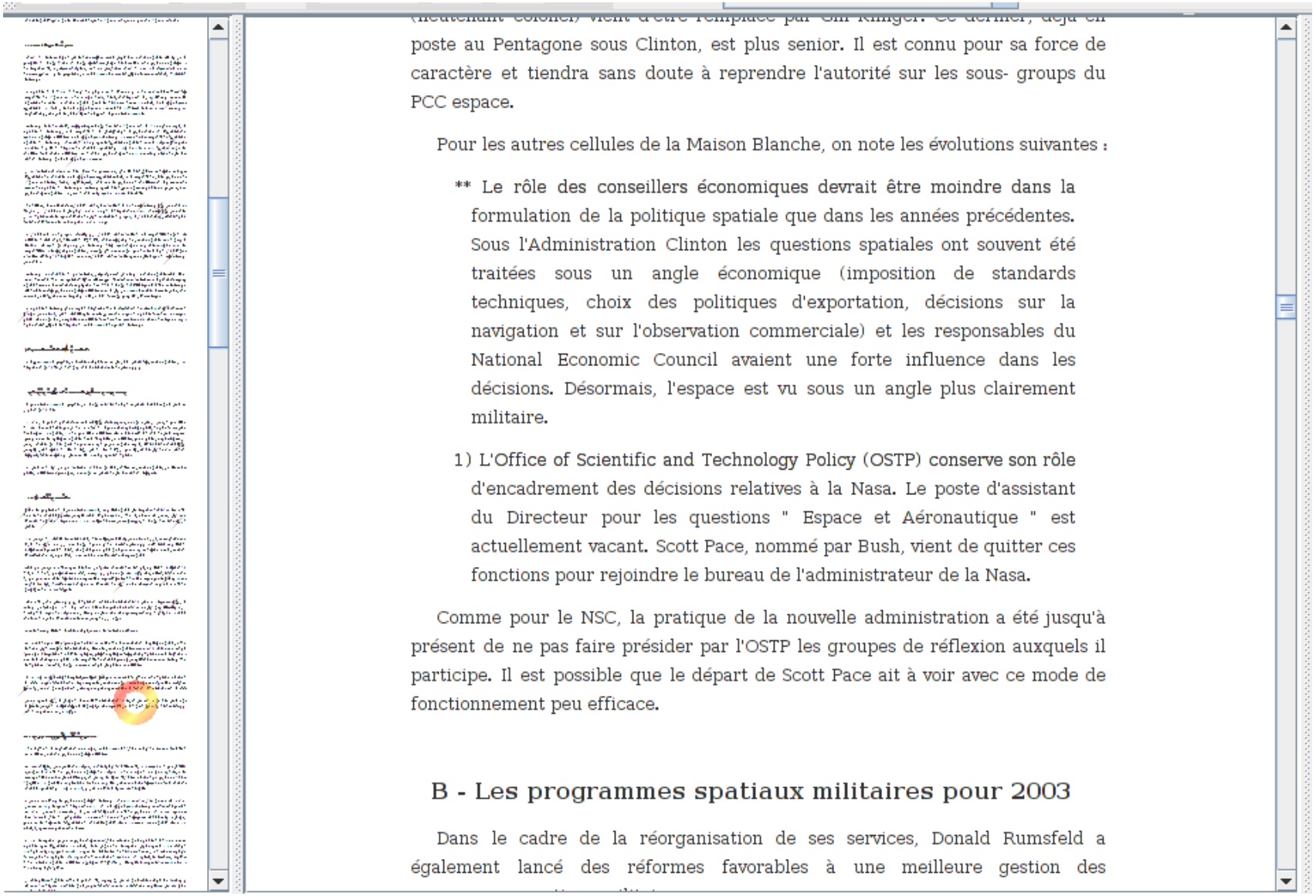
- Avoir une vue de haut du texte => ruban
- Avoir une vue de haut de zones à annoter => prémarquage
- Avoir une vue graphique des annotations

Besoins propres aux deux approches

Différents besoins en terme d'outils pour annoter

- **Avoir une vue de haut du texte => ruban**
- Avoir une vue de haut de zones à annoter => prémarquage
- Avoir une vue graphique des annotations

Vue de haut du texte => ruban



poste au Pentagone sous Clinton, est plus senior. Il est connu pour sa force de caractère et tiendra sans doute à reprendre l'autorité sur les sous- groups du PCC espace.

Pour les autres cellules de la Maison Blanche, on note les évolutions suivantes :

** Le rôle des conseillers économiques devrait être moindre dans la formulation de la politique spatiale que dans les années précédentes. Sous l'Administration Clinton les questions spatiales ont souvent été traitées sous un angle économique (imposition de standards techniques, choix des politiques d'exportation, décisions sur la navigation et sur l'observation commerciale) et les responsables du National Economic Council avaient une forte influence dans les décisions. Désormais, l'espace est vu sous un angle plus clairement militaire.

1) L'Office of Scientific and Technology Policy (OSTP) conserve son rôle d'encadrement des décisions relatives à la Nasa. Le poste d'assistant du Directeur pour les questions " Espace et Aéronautique " est actuellement vacant. Scott Pace, nommé par Bush, vient de quitter ces fonctions pour rejoindre le bureau de l'administrateur de la Nasa.

Comme pour le NSC, la pratique de la nouvelle administration a été jusqu'à présent de ne pas faire présider par l'OSTP les groupes de réflexion auxquels il participe. Il est possible que le départ de Scott Pace ait à voir avec ce mode de fonctionnement peu efficace.

B - Les programmes spatiaux militaires pour 2003

Dans le cadre de la réorganisation de ses services, Donald Rumsfeld a également lancé des réformes favorables à une meilleure gestion des

Besoins propres aux deux approches

Différents besoins en terme d'outils pour annoter

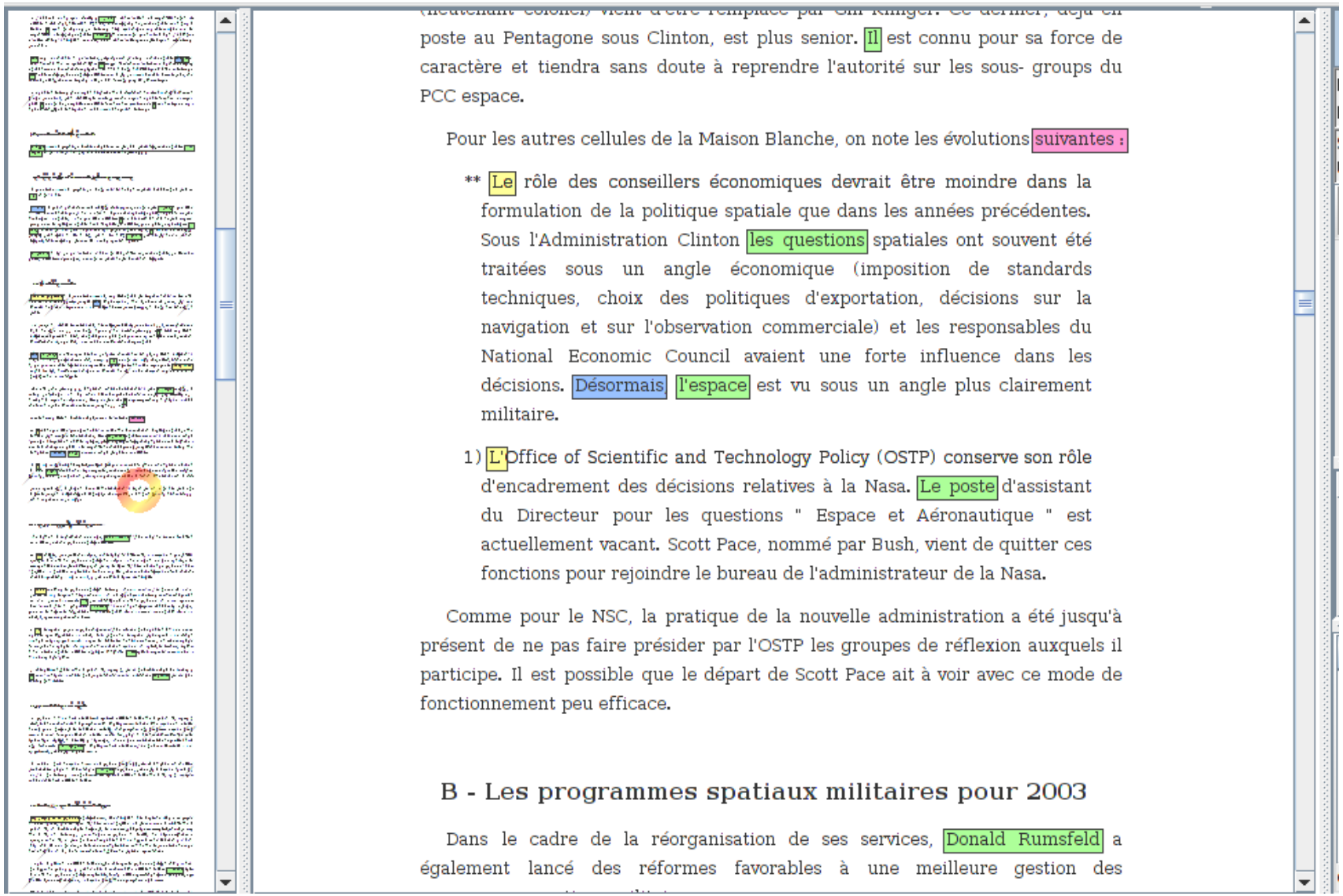
- Avoir une vue de haut du texte => ruban
- Avoir une vue de haut de zones à annoter => prémarquage
- Avoir une vue graphique des annotations

Besoins propres aux deux approches

Différents besoins en terme d'outils pour annoter

- Avoir une vue de haut du texte => ruban
- **Avoir une vue de haut de zones à annoter => prémarquage**
- Avoir une vue graphique des annotations

Vue de haut de zones à annoter => prémarquage



poste au Pentagone sous Clinton, est plus senior. Il est connu pour sa force de caractère et tiendra sans doute à reprendre l'autorité sur les sous- groups du PCC espace.

Pour les autres cellules de la Maison Blanche, on note les évolutions suivantes :

** Le rôle des conseillers économiques devrait être moindre dans la formulation de la politique spatiale que dans les années précédentes. Sous l'Administration Clinton les questions spatiales ont souvent été traitées sous un angle économique (imposition de standards techniques, choix des politiques d'exportation, décisions sur la navigation et sur l'observation commerciale) et les responsables du National Economic Council avaient une forte influence dans les décisions. Désormais, l'espace est vu sous un angle plus clairement militaire.

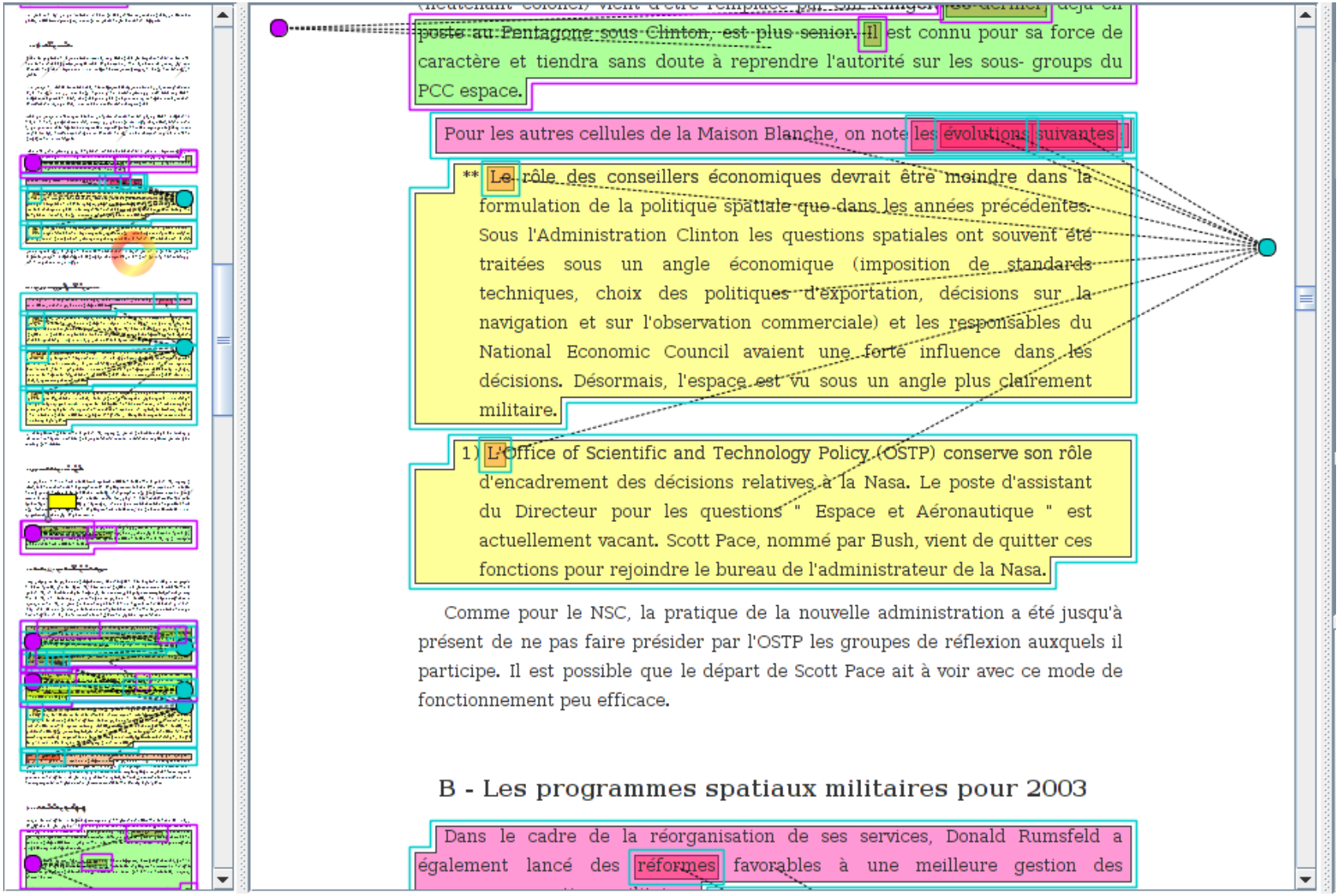
1) L'Office of Scientific and Technology Policy (OSTP) conserve son rôle d'encadrement des décisions relatives à la Nasa. Le poste d'assistant du Directeur pour les questions " Espace et Aéronautique " est actuellement vacant. Scott Pace, nommé par Bush, vient de quitter ces fonctions pour rejoindre le bureau de l'administrateur de la Nasa.

Comme pour le NSC, la pratique de la nouvelle administration a été jusqu'à présent de ne pas faire présider par l'OSTP les groupes de réflexion auxquels il participe. Il est possible que le départ de Scott Pace ait à voir avec ce mode de fonctionnement peu efficace.

B - Les programmes spatiaux militaires pour 2003

Dans le cadre de la réorganisation de ses services, Donald Rumsfeld a également lancé des réformes favorables à une meilleure gestion des

Annotation des structures et des indices



Post-Traitement des structures multi-échelles annotées

1. Caractérisation des indices non typés lors de l'annotation (Traits prémarqués non typés automatiquement et validés par les annotateurs sans remplissage du champ « nature de l'indice » ; indices ajoutés manuellement sans remplissage du champ « nature de l'indice »)
2. Homogénéisation manuelle des annotations des énumérathèmes : exclusion des déterminants et des modifieurs pour ne conserver que la tête lexicale. Par exemple, l'énumérathème *deux types de lectures antinomiques* a été réduit à *lectures*.
3. Correction des annotations des énumérathèmes mal délimités suite à une mauvaise manipulation de l'interface (e.g. "ibertés collectives" devient "libertés collectives").
4. Suppression des unités (traits prémarqués) non rattachées à un schéma.
5. Normalisation de la structure de traits de l'annotation des indices, qu'il s'agisse de traits prémarqués validés par les annotateurs ou d'indices ajoutés : toutes ces unités font maintenant l'objet de la même caractérisation en termes de type (INDICE) et de nature (CIRCtps, MIL, etc.).
6. production d'un Gold pour les cas de multi-annotation

La ressource ANNODIS

Relations rhétoriques et structures multi-échelles dans ANNODIS.

corpus	Annotated objects						
			bottom-up approach			top-down approach	
	words	texts	EDU	Rh.Rel.	CDU	ES	CT
NEWS	9,768	39	1159	1203	510		
WIK1	17,330	42	1949	2034	829		
WIK2	231,000	30	53	65	38	401	266
LING	169,000	25	12	14	9	297	88
GEOP	266,000	32	15	19	9	293	234
ANNODIS	687,000		3188	3355	1395	991	588

- EDU = Unité de discours élémentaire ; Rh.Rel. = Relation rhétoriques ; CDU = Unité de discours complexe
- SE = Structure Énumérative ; CT = Chaîne Topicale

Relations de discours annotées (version experte*)

	Nb	(%)	News %	Wik1 %
alternation	18	0.5	0.3	0.6
attribution	75	2.2	3.0	1.7
background	155	4.6	5.2	4.8
comment	78	2.3	3.6	1.3
continuation	681	20.3	20.1	21.1
contrast	144	4.3	3.7	4.6
E-elab	527	15.7	14.1	16.4
elaboration	625	18.6	16.3	19.4
explanation	130	3.9	4.4	3.3
flashback	27	0.8	1.4	0.6
frame	211	6.3	6.2	5.7
goal	95	2.8	3.1	2.4
narration	349	10.4	11.1	10.4
parralel	59	1.8	2.2	1.8
result	163	4.9	4.7	5.4
temploc	18	0.5	0.5	0.5
totRel(nb)	3355			

Structures multi-échelles annotées (gold inclus*)

corpus	ES	item	trigger	closure	TC	topical expr.
WIK2	401	1653	300	36	266	1853
LING	297	850	230	46	88	478
GEOP	293	863	209	49	234	1125
Total	991	3366	740	131	588	3456

Plan

1. Éléments de définition
2. le projet ANNODIS
3. Analyse de Annotations
   - Accord inter-annotateur
   - Analyse de l'insertion des structures dans la MFM
4. Annoter du "contenu" ou des documents ?
5. Conclusion

Accord inter-annotateur

Structures multi-échelles – F-mesure

$$F = \frac{(2 \times \textit{nombre de structures communes})}{\textit{structures annotées par } A + \textit{structures annotées par } B}$$

Moyenne de toutes les F-mesures calculées par comparaison de l'annotateur A et de l'annotateur B pour chaque texte.

- deux SE se trouvent au même endroit dans le texte ET les items repérés sont les mêmes : 0,67
- deux CT recouvrent une même zone de texte (même partiellement) ET les indices repérés sont les mêmes : 0,65.

Relations rhétoriques – Kappa de Cohen

- rattachement entre EDU et CDU (comparer si 2 segments sont communément reliés ou pas) : 0,6
- type de relation entre deux segments rattachés $< 0{,}6$

Accord inter-annotateur – observations

La mesure de l'accord inter-annotateur permet d'observer des phénomènes +/- interprétables :

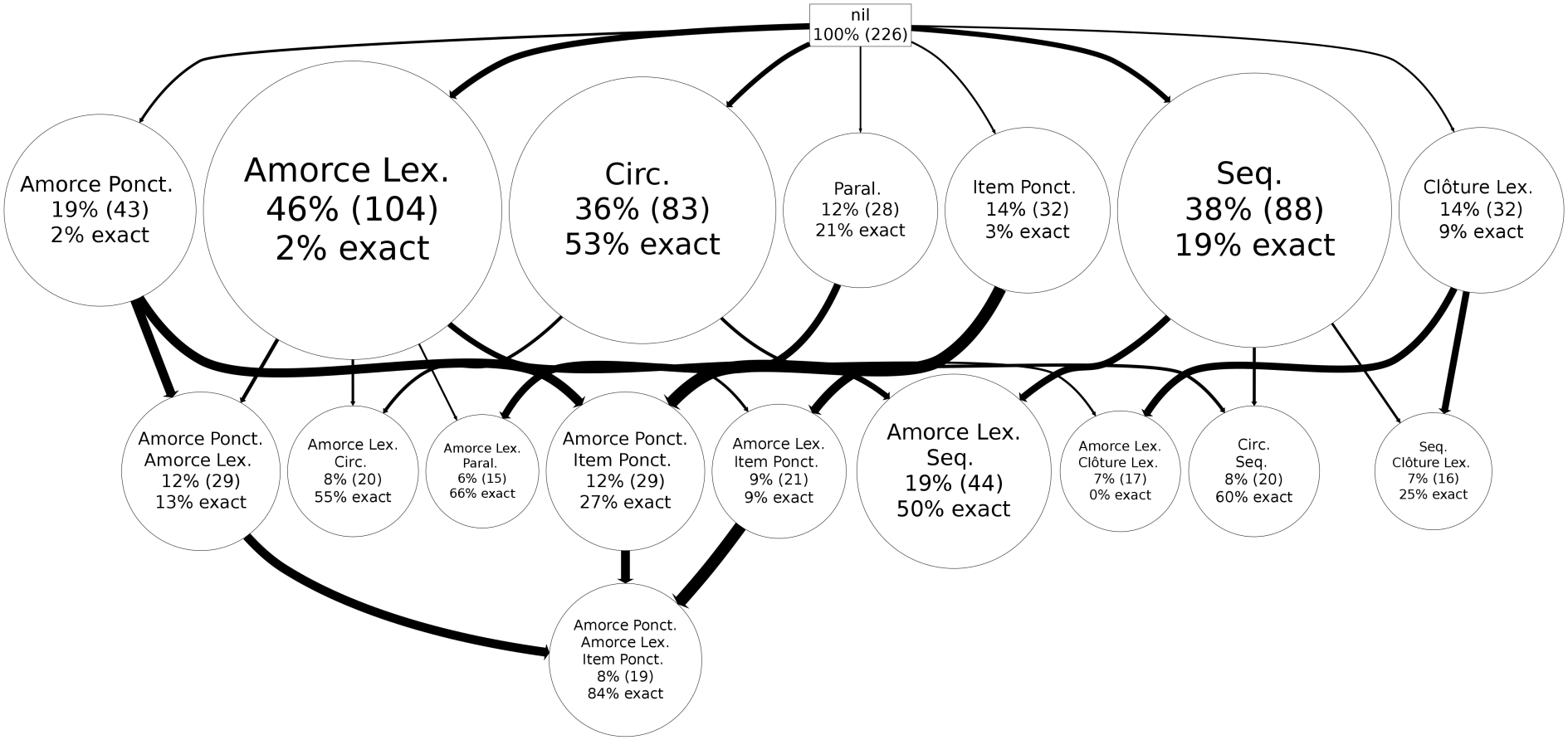
- moins il y a de SE dans un texte, moins l'accord est bon ;
- les SE de faible empan sont difficiles à arbitrer ;
- certains types de SE sont plus propices aux désaccords : argumentations, énumération chronologique/temporelle.
- meilleur taux d'accord entre les CT dont la majorité des indices est un pronom personnel et entre les CT comprenant des titres de sections.
- des relations fréquemment confondues ou inversées (e.g. explication - résultat)

Prise en compte de la mise en forme et de l'échelle dans l'analyse des annotations

Typologie de l'interaction avec la MFM – pour les SE uniquement

Corpus	Niveaux de grain			
	sections titrées	liste formatée	multi-paragraphique	intra-paragraphique
WIK1	19,3%	39,1%	20,8%	20,8%
LING	9,1%	23,2%	26,6%	41,1%
GEOP	6,8%	10,3%	20,9%	62%

Treillis : SE de type 3 (multiparagraphiques)



Questions soulevées pour l'annotation de haut-niveau

- Intégration de la MFM dans l'étude de la signalisation des structures
- Besoin d'expérimentations psycholinguistiques – données primaires = documents construits faisant varier la MFM pour en mesurer l'influence
- Prise en compte de la dimension "documents"
  - pour l'annotation
  - pour l'analyse
- Difficulté d'analyse de structures multi-échelles annotées (sans unité minimale). Des besoin en terme d'exploration des données (navigateur ANNODIS).

Plan

1. Éléments de définition
2. le projet ANNODIS
3. Analyse de Annotations
4. Annoter du "contenu" ou des documents ?
   - Original vs. "contenu propositionnel"
   - Original vs. document préparé
5. Conclusion

Annoter du "contenu" ou des documents ? Quelles "données primaires"

On annote toujours des données préparées, transformées selon des choix.

Ces choix reflètent ce que l'on considère comme étant "linguistique" ou pas (ce dont on garde la trace)

Annotation de haut-niveau implique structure de document (Power et al. 2003) et de la mise en forme matérielle – MFM (Luc & Virbel 2001)

Question de la mise en forme et de son influence sur l'interprétation (annoter = interpréter)

- ANNODIS_me – *ANNODIS multi-échelle*
  données primaires = "documents préparés"

Annoter du "contenu" ou des documents ? Quelles "données primaires"

On annote toujours des données préparées, transformées selon des choix.

Ces choix reflètent ce que l'on considère comme étant "linguistique" ou pas (ce dont on garde la trace)

Annotation de haut-niveau implique structure de document (Power et al. 2003) et de la mise en forme matérielle – MFM (Luc & Virbel 2001)

Question de la mise en forme et de son influence sur l'interprétation (annoter = interpréter)

- ANNODIS_me – *ANNODIS multi-échelle*
  données primaires = "documents préparés"
- ANNODIS_rr – *ANNODIS relations rhétoriques*
  données primaires = "contenu propositionnel"

Annoter du "contenu" ou des documents ? Quelles "données primaires"

On annote toujours des données préparées, transformées selon des choix.

Ces choix reflètent ce que l'on considère comme étant "linguistique" ou pas (ce dont on garde la trace)

Annotation de haut-niveau implique structure de document (Power et al. 2003) et de la mise en forme matérielle – MFM (Luc & Virbel 2001)

Question de la mise en forme et de son influence sur l'interprétation (annoter = interpréter)

- ANNODIS_me – *ANNODIS multi-échelle*
  données primaires = "documents préparés"
- ANNODIS_rr – *ANNODIS relations rhétoriques*
  données primaires = "contenu propositionnel"

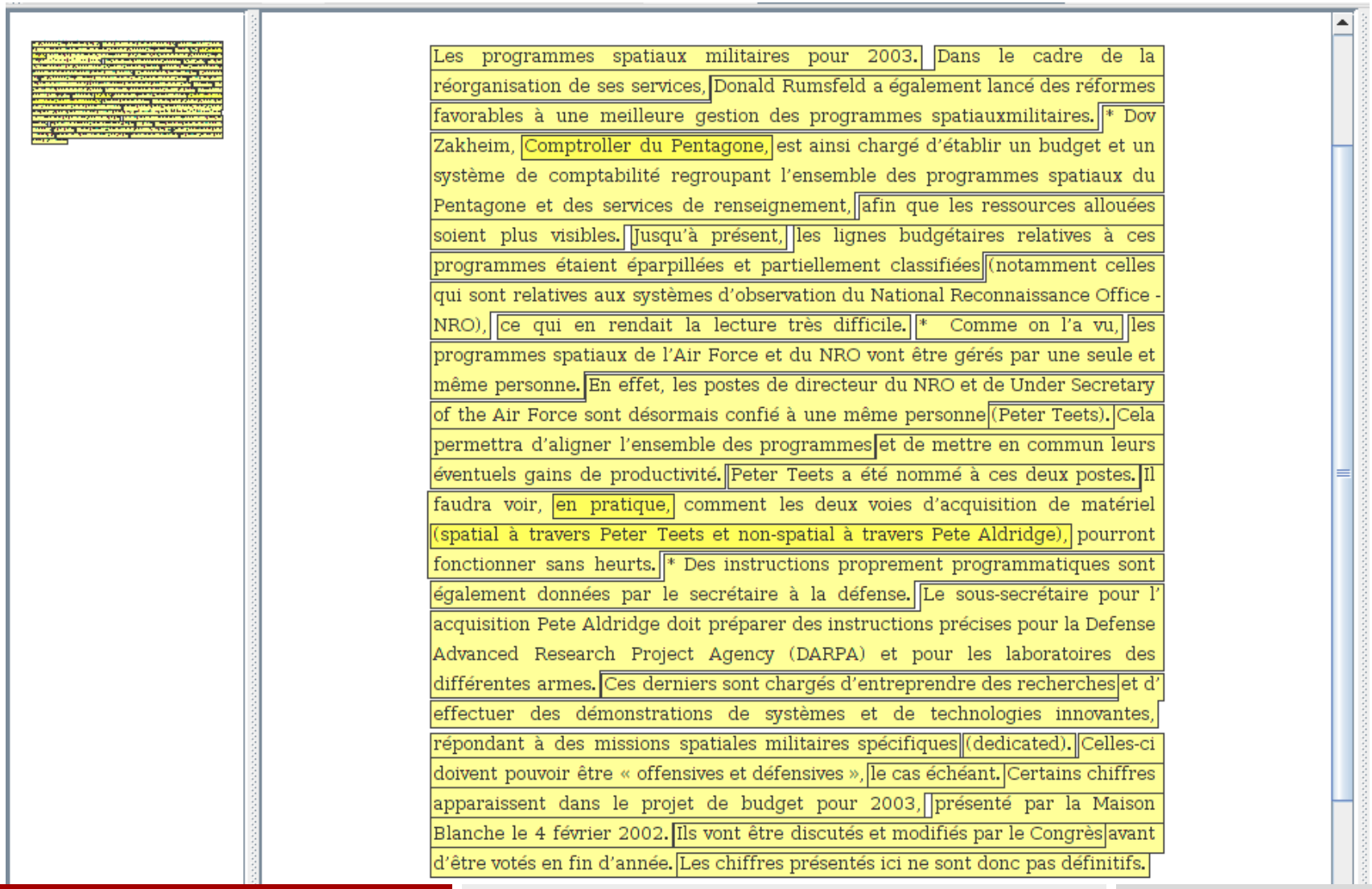
Annoter du "contenu" ou des documents ? – illustration

Données primaires = "contenu propositionnel"

Les programmes spatiaux militaires pour 2003. Dans le cadre de la réorganisation de ses services, Donald Rumsfeld a également lancé des réformes favorables à une meilleure gestion des programmes spatiauxmilitaires. * Dov Zakheim, Comptroller du Pentagone, est ainsi chargé d'établir un budget et un système de comptabilité regroupant l'ensemble des programmes spatiaux du Pentagone et des services de renseignement, afin que les ressources allouées soient plus visibles. Jusqu'à présent, les lignes budgétaires relatives à ces programmes étaient éparpillées et partiellement classifiées (notamment celles qui sont relatives aux systèmes d'observation du National Reconnaissance Office - NRO), ce qui en rendait la lecture très difficile. * Comme on l'a vu, les programmes spatiaux de l'Air Force et du NRO vont être gérés par une seule et même personne. En effet, les postes de directeur du NRO et de Under Secretary of the Air Force sont désormais confié à une même personne (Peter Teets). Cela permettra d'aligner l'ensemble des programmes et de mettre en commun leurs éventuels gains de productivité. Peter Teets a été nommé à ces deux postes. Il faudra voir, en pratique, comment les deux voies d'acquisition de matériel (spatial à travers Peter Teets et non-spatial à travers Pete Aldridge), pourront fonctionner sans heurts. * Des instructions proprement programmatiques sont également données par le secrétaire à la défense. Le sous-secrétaire pour l' acquisition Pete Aldridge doit préparer des instructions précises pour la Defense Advanced Research Project Agency (DARPA) et pour les laboratoires des différentes armes. Ces derniers sont chargés d'entreprendre des recherches et d' effectuer des démonstrations de systèmes et de technologies innovantes, répondant à des missions spatiales militaires spécifiques (dedicated). Celles-ci doivent pouvoir être « offensives et défensives », le cas échéant. Certains chiffres apparaissent dans le projet de budget pour 2003, présenté par la Maison Blanche le 4 février 2002. Ils vont être discutés et modifiés par le Congrès avant d'être votés en fin d'année. Les chiffres présentés ici ne sont donc pas définitifs.

Annoter du "contenu" ou des documents ? – illustration

Données primaires = "contenu propositionnel" segmenté



Les programmes spatiaux militaires pour 2003. Dans le cadre de la réorganisation de ses services, Donald Rumsfeld a également lancé des réformes favorables à une meilleure gestion des programmes spatiauxmilitaires. * Dov Zakheim, Comptroller du Pentagone, est ainsi chargé d'établir un budget et un système de comptabilité regroupant l'ensemble des programmes spatiaux du Pentagone et des services de renseignement, afin que les ressources allouées soient plus visibles. Jusqu'à présent, les lignes budgétaires relatives à ces programmes étaient éparpillées et partiellement classifiées (notamment celles qui sont relatives aux systèmes d'observation du National Reconnaissance Office - NRO), ce qui en rendait la lecture très difficile. * Comme on l'a vu, les programmes spatiaux de l'Air Force et du NRO vont être gérés par une seule et même personne. En effet, les postes de directeur du NRO et de Under Secretary of the Air Force sont désormais confié à une même personne (Peter Teets). Cela permettra d'aligner l'ensemble des programmes et de mettre en commun leurs éventuels gains de productivité. Peter Teets a été nommé à ces deux postes. Il faudra voir, en pratique, comment les deux voies d'acquisition de matériel (spatial à travers Peter Teets et non-spatial à travers Pete Aldridge), pourront fonctionner sans heurts. * Des instructions proprement programmatiques sont également données par le secrétaire à la défense. Le sous-secrétaire pour l' acquisition Pete Aldridge doit préparer des instructions précises pour la Defense Advanced Research Project Agency (DARPA) et pour les laboratoires des différentes armes. Ces derniers sont chargés d'entreprendre des recherches et d' effectuer des démonstrations de systèmes et de technologies innovantes, répondant à des missions spatiales militaires spécifiques (dedicated). Celles-ci doivent pouvoir être « offensives et défensives », le cas échéant. Certains chiffres apparaissent dans le projet de budget pour 2003, présenté par la Maison Blanche le 4 février 2002. Ils vont être discutés et modifiés par le Congrès avant d'être votés en fin d'année. Les chiffres présentés ici ne sont donc pas définitifs.

Annoter du "contenu" ou des documents ? Quelles "données primaires"

On annote toujours des données préparées, transformées selon des choix.

Ces choix reflètent ce que l'on considère comme étant "linguistique" ou pas (ce dont on garde la trace)

Annotation de haut-niveau implique structure de document (Power et al.) et de la mise en forme matérielle (MFM) (Virbel)

Question de la mise en forme et de son influence sur l'interprétation (annoter = interpréter)

- ANNODIS_me – *ANNODIS multi-échelle*
  données primaires = "documents préparés"
- ANNODIS_rr – *ANNODIS relations rhétoriques*
  données primaires = "contenu propositionnel"

Annoter du "contenu" ou des documents ? Quelles "données primaires"

On annote toujours des données préparées, transformées selon des choix.

Ces choix reflètent ce que l'on considère comme étant "linguistique" ou pas (ce dont on garde la trace)

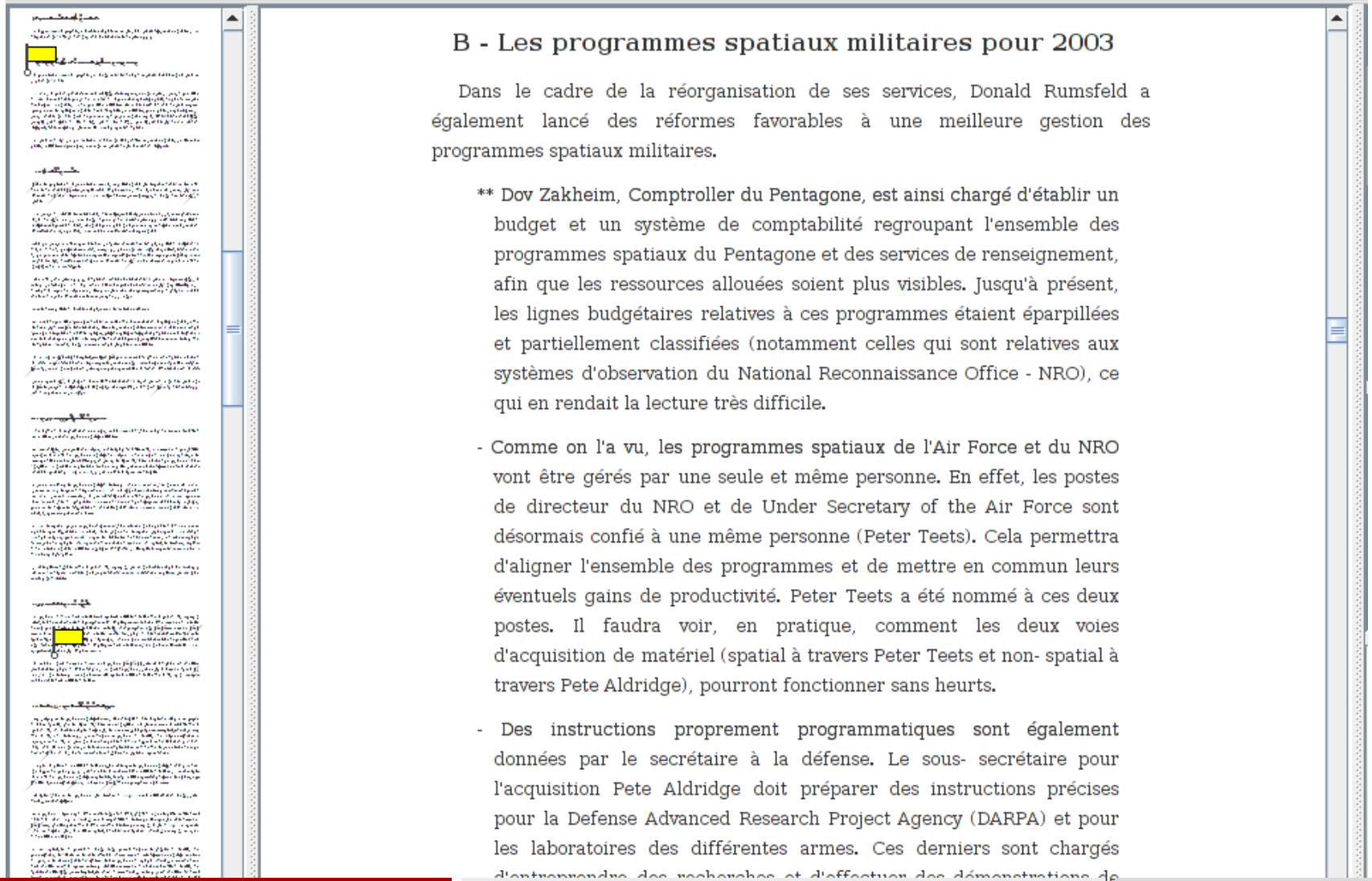
Annotation de haut-niveau implique structure de document (Power et al.) et de la mise en forme matérielle (MFM) (Virbel)

Question de la mise en forme et de son influence sur l'interprétation (annoter = interpréter)

- ANNODIS_me – *ANNODIS multi-échelle*
  données primaires = "documents préparés"
- ANNODIS_rr – *ANNODIS relations rhétoriques*
  données primaires = "contenu propositionnel"

Annoter du "contenu" ou des documents ? – illustration

Données primaires = documents préparés



B - Les programmes spatiaux militaires pour 2003

Dans le cadre de la réorganisation de ses services, Donald Rumsfeld a également lancé des réformes favorables à une meilleure gestion des programmes spatiaux militaires.

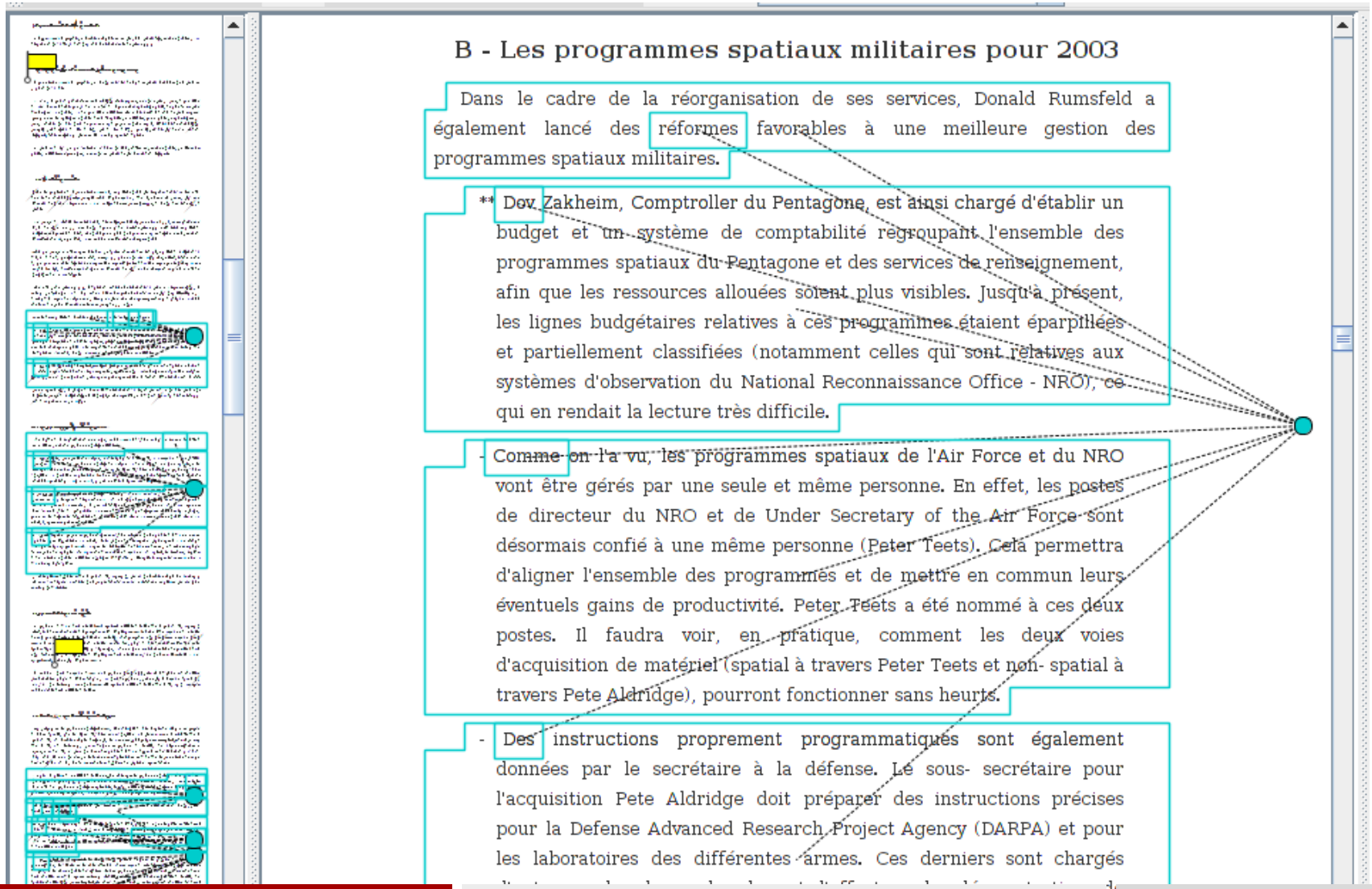
** Dov Zakheim, Comptroller du Pentagone, est ainsi chargé d'établir un budget et un système de comptabilité regroupant l'ensemble des programmes spatiaux du Pentagone et des services de renseignement, afin que les ressources allouées soient plus visibles. Jusqu'à présent, les lignes budgétaires relatives à ces programmes étaient éparpillées et partiellement classifiées (notamment celles qui sont relatives aux systèmes d'observation du National Reconnaissance Office - NRO), ce qui en rendait la lecture très difficile.

- Comme on l'a vu, les programmes spatiaux de l'Air Force et du NRO vont être gérés par une seule et même personne. En effet, les postes de directeur du NRO et de Under Secretary of the Air Force sont désormais confié à une même personne (Peter Teets). Cela permettra d'aligner l'ensemble des programmes et de mettre en commun leurs éventuels gains de productivité. Peter Teets a été nommé à ces deux postes. Il faudra voir, en pratique, comment les deux voies d'acquisition de matériel (spatial à travers Peter Teets et non- spatial à travers Pete Aldridge), pourront fonctionner sans heurts.

- Des instructions proprement programmatiques sont également données par le secrétaire à la défense. Le sous- secrétaire pour l'acquisition Pete Aldridge doit préparer des instructions précises pour la Defense Advanced Research Project Agency (DARPA) et pour les laboratoires des différentes armes. Ces derniers sont chargés

Annoter du "contenu" ou des documents ? – illustration

Données primaires = documents préparés



B - Les programmes spatiaux militaires pour 2003

Dans le cadre de la réorganisation de ses services, Donald Rumsfeld a également lancé des réformes favorables à une meilleure gestion des programmes spatiaux militaires.

** Dov Zakheim, Comptroller du Pentagone, est ainsi chargé d'établir un budget et un système de comptabilité regroupant l'ensemble des programmes spatiaux du Pentagone et des services de renseignement, afin que les ressources allouées soient plus visibles. Jusqu'à présent, les lignes budgétaires relatives à ces programmes étaient éparpillées et partiellement classifiées (notamment celles qui sont relatives aux systèmes d'observation du National Reconnaissance Office - NRO), ce qui en rendait la lecture très difficile.

- Comme on l'a vu, les programmes spatiaux de l'Air Force et du NRO vont être gérés par une seule et même personne. En effet, les postes de directeur du NRO et de Under Secretary of the Air Force sont désormais confié à une même personne (Peter Teets). Cela permettra d'aligner l'ensemble des programmes et de mettre en commun leurs éventuels gains de productivité. Peter Teets a été nommé à ces deux postes. Il faudra voir, en pratique, comment les deux voies d'acquisition de matériel (spatial à travers Peter Teets et non- spatial à travers Pete Aldridge), pourront fonctionner sans heurts.

- Des instructions proprement programmatiques sont également données par le secrétaire à la défense. Le sous- secrétaire pour l'acquisition Pete Aldridge doit préparer des instructions précises pour la Defense Advanced Research Project Agency (DARPA) et pour les laboratoires des différentes armes. Ces derniers sont chargés

Annoter du "contenu" ou des documents ? Quelles "données primaires"

On annote toujours des données préparées, transformées selon des choix.

Ces choix reflètent ce que l'on considère comme étant "linguistique" ou pas (ce dont on garde la trace)

Annotation de haut-niveau implique structure de document (Power et al.) et de la mise en forme matérielle (MFM) (Virbel)

Question de la mise en forme et de son influence sur l'interprétation (annoter = interpréter)

- ANNODIS_me – *ANNODIS multi-échelle*
  données primaires = "documents préparés"
- ANNODIS_rr – *ANNODIS relations rhétoriques*
  données primaires = "contenu propositionnel"

Annoter du "contenu" ou des documents ? Quelles "données primaires"

On annote toujours des données préparées, transformées selon des choix.

Ces choix reflètent ce que l'on considère comme étant "linguistique" ou pas (ce dont on garde la trace)

Annotation de haut-niveau implique structure de document (Power et al.) et de la mise en forme matérielle (MFM) (Virbel)

Question de la mise en forme et de son influence sur l'interprétation (annoter = interpréter)

- ANNODIS_me – *ANNODIS multi-échelle*
  données primaires = "documents préparés"
- ANNODIS_rr – *ANNODIS relations rhétoriques*
  données primaires = "contenu propositionnel"

Annoter du "contenu" ou des documents ? – illustration

Données primaires = le document original

Thumbnails

B- Les programmes spatiaux militaires pour 2003

Dans le cadre de la réorganisation de ses services, Donald Rumsfeld a également lancé des réformes favorables à une meilleure gestion des programmes spatiaux militaires.

- Dov Zakheim, *Comptroller* du Pentagone, est ainsi chargé d'établir un budget et un système de comptabilité regroupant l'ensemble des programmes spatiaux du Pentagone et des services de renseignement, afin que les ressources allouées soient plus visibles. Jusqu'à présent, les lignes budgétaires relatives à ces programmes étaient éparpillées et partiellement classifiées (notamment celles qui sont relatives aux systèmes d'observation du *National Reconnaissance Office* -NRO), ce qui en rendait la lecture très difficile.

- Comme on l'a vu, les programmes spatiaux de l'*Air Force* et du NRO vont être gérés par une seule et même personne. En effet, les postes de directeur du NRO et de *Under Secretary of the Air Force* sont désormais confié à une même personne (Peter Teets). Cela permettra d'aligner l'ensemble des programmes et de mettre en commun leurs éventuels gains de productivité. Peter Teets a été nommé à ces deux postes. Il

[4] Scott Pace a longtemps été chercheur au *Science and Technology Policy Institute* de la Rand Corporation, un think-tank qui travaille principalement pour l'Air Force et le Pentagone. Il a écrit plusieurs ouvrages sur les questions d'observation spatiale et de militarisation de l'espace. Il est à l'origine de la décision présidentielle d'ouverture totale du système GPS aux civils, adoptée en 1996.

L'administration Bush et l'espace 8

faudra voir, en pratique, comment les deux voies d'acquisition de matériel (spatial à travers Peter Teets et non-spatial à travers Pete Aldridge), pourront fonctionner sans

Annoter du "contenu" ou des documents ? Quelles "données primaires"

On annote toujours des données préparées, transformées selon des choix.

Question de la mise en forme et de son influence sur l'interprétation (annoter = interpréter)

- Différence d'interprétation entre Original et "contenu propositionnel"
- Différence d'interprétation entre Original et document préparé

Annoter du "contenu" ou des documents ? Quelles "données primaires"

On annote toujours des données préparées, transformées selon des choix.

Question de la mise en forme et de son influence sur l'interprétation (annoter = interpréter)

- Différence d'interprétation entre Original et "contenu propositionnel"
- Différence d'interprétation entre Original et document préparé

Annotation experte sur des "textes" nus à la MFM effacée

[Une pluie d'étoiles**]_1 [**Non !]_2 [Il ne s'agit pas d'un phénomène météorologique accompagnant le solstice d'été.]_3 [Plus simplement,]_4 [les hasards du calendrier du Comité départemental d'action touristique a fait coïncider la promotion de l'Office de tourisme avec le nouveau classement de l'hôtel-restaurant Le Relais à Arc-et-Senans.**]_5**

elaboration(1/[2-5])
elaboration (2/3)
elaboration(4/5)
contrast ([2,3]/[4,5])

MFM cachée

```
<head> Une pluie d'étoiles </head>

<p> Non ! Il ne s'agit pas d'un phénomène météorologique accompagnant le solstice
d'été. Plus simplement, les hasards du calendrier du Comité départemental d'action
touristique a fait coïncider la promotion de l'Office de tourisme avec le nouveau
classement de l'hôtel-restaurant Le Relais à Arc-et-Senans. Le double événement
était fêté comme il se doit et conjointement par les deux établissements, mercredi
23 juin, à l'occasion de l'assemblée générale décentralisée des Offices de
tourisme et Syndicats d'initiative du Doubs qui se tenait le même jour, sous la
présidence d'Edmond Maire, à la Saline Royale. </p>

<p> Deux étoiles pour l'hôtel Le Relais, situé sur la place de l'Eglise ; une étoile
pour le restaurant du même nom, bien connu des habitants et des touristes. Une
juste récompense pour les efforts de qualité entrepris par Jeanine et Willy
Schaer, jeune couple d'hôteliers originaires de Haute-Saône et installés à
Arc-et-Senans depuis maintenant treize ans. </p>

<p> Deux étoiles aussi, c'est le nouveau classement de l'Office de tourisme
d'Arc-et-Senans qui prend ainsi sa place dans la cour des grands, juste derrière
les offices de Besançon et Montbéliard. Sur des critères très stricts de
compétence et d'accueil, c'est le premier établissement en France qui reçoit cette
promotion selon la toute nouvelle loi du 16 décembre 1998 sur les Offices de
tourisme. </p>
```

Annoter du "contenu" ou des documents ? Quelles "données primaires"

On annote toujours des données préparées, transformées selon des choix.

Question de la mise en forme et de son influence sur l'interprétation (annoter = interpréter)

- Différence d'interprétation entre Original et "contenu propositionnel"
- Différence d'interprétation entre Original et document préparé

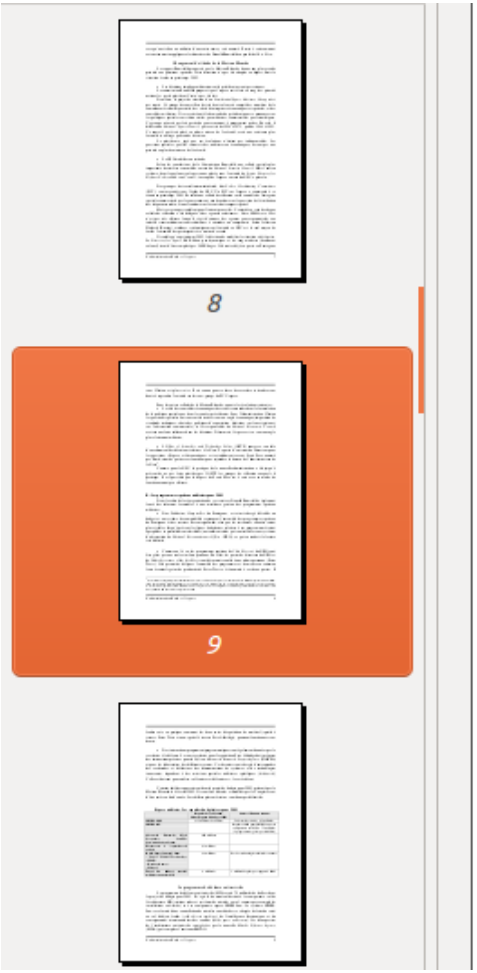
Annoter du "contenu" ou des documents ? Quelles "données primaires"

On annote toujours des données préparées, transformées selon des choix.

Question de la mise en forme et de son influence sur l'interprétation (annoter = interpréter)

- Différence d'interprétation entre Original et "contenu propositionnel"
- Différence d'interprétation entre Original et document préparé

Annotation experte sur document original imprimé puis reporté sous Glozz



sous Clinton, est plus *senior*. Il est connu pour sa force de caractère et tiendra sans doute à reprendre l'autorité sur les sous-groups du PCC-espace.

Pour les autres cellules de la Maison Blanche, on note les évolutions suivantes :

- Le rôle des conseillers économiques devrait être moindre dans la formulation de la politique spatiale que dans les années précédentes. Sous l'Administration Clinton les questions spatiales ont souvent été traitées sous un angle économique (imposition de standards techniques, choix des politiques d'exportation, décisions sur la navigation et sur l'observation commerciale) et les responsables du *National Economic Council* avaient une forte influence dans les décisions. Désormais, l'espace est vu sous un angle plus clairement militaire.

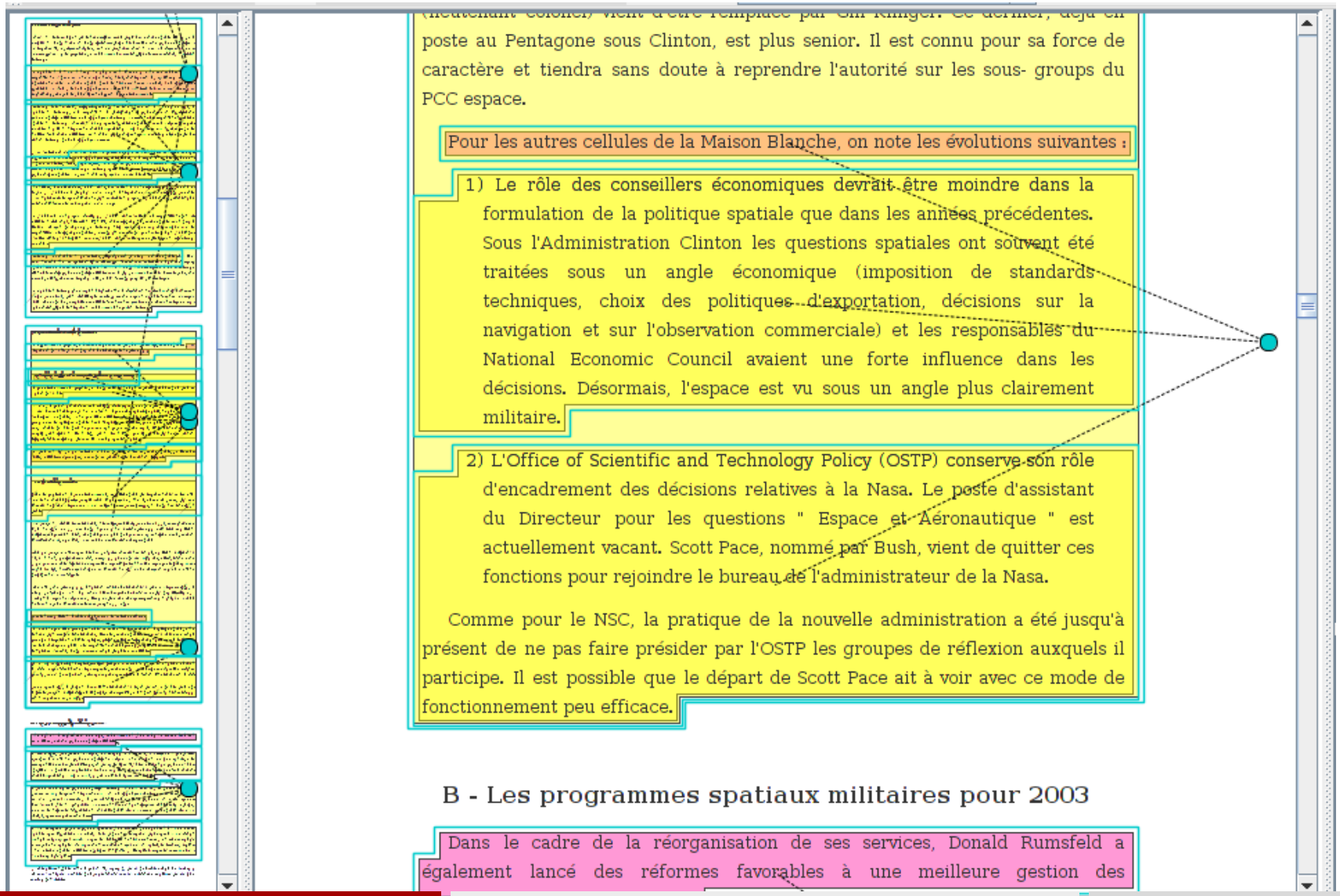
- L'*Office of Scientific and Technology Policy* (OSTP) conserve son rôle d'encadrement des décisions relatives à la Nasa. Le poste d'assistant du Directeur pour les questions « Espace et Aéronautique » est actuellement vacant. Scott Pace, nommé par Bush, vient de quitter ces fonctions pour rejoindre le bureau de l'administrateur de la Nasa[4].

Comme pour le NSC, la pratique de la nouvelle administration a été jusqu'à présent de ne pas faire présider par l'OSTP les groupes de réflexion auxquels il participe. Il est possible que le départ de Scott Pace ait à voir avec ce mode de fonctionnement peu efficace.

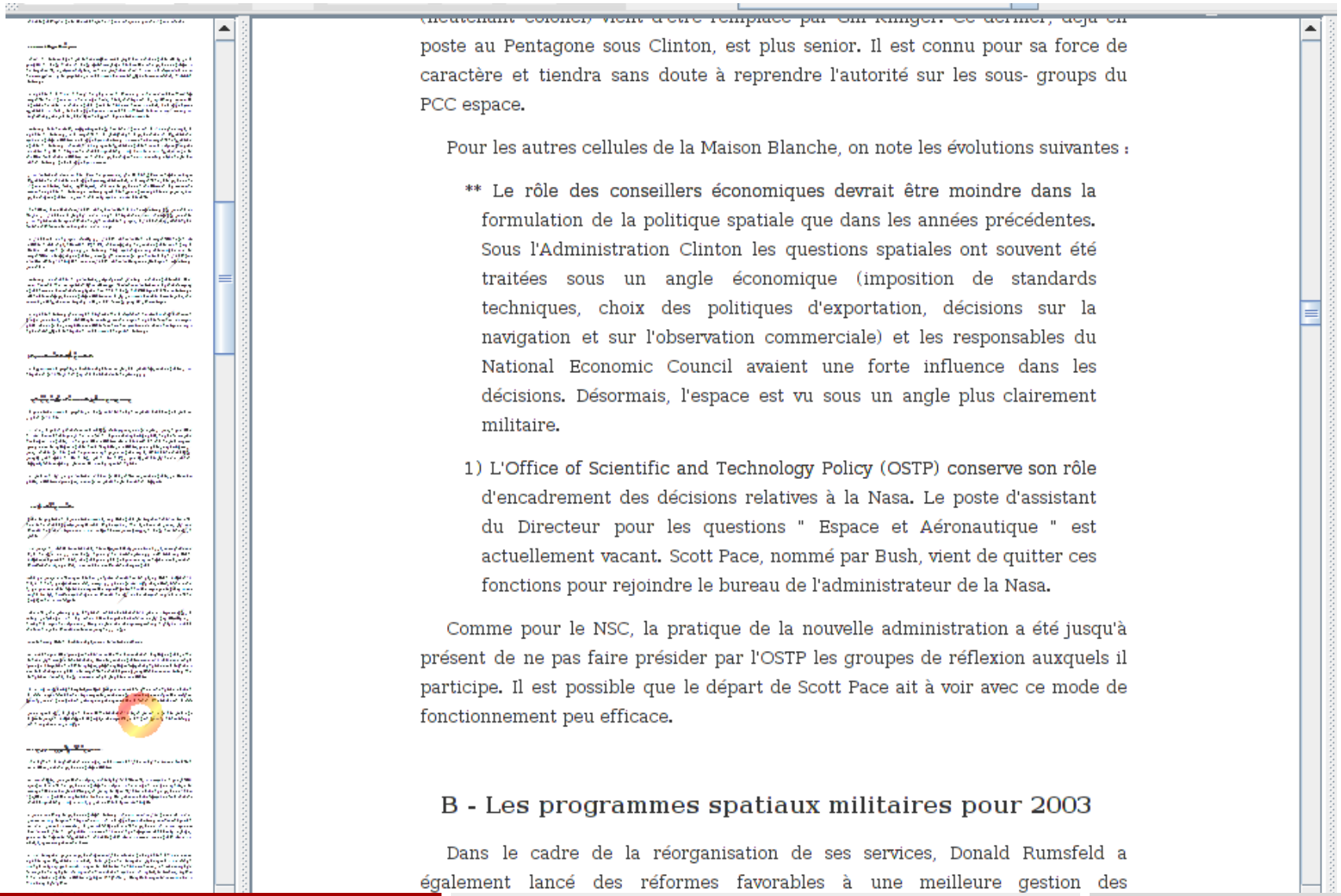
B- Les programmes spatiaux militaires pour 2003

Dans le cadre de la réorganisation de ses services, Donald Rumsfeld a également lancé des réformes favorables à une meilleure gestion des programmes spatiaux militaires.

Annotation experte sur document original imprimé puis reporté sous Glozz



Annotation naive sur document préparé, directement via Glozz



poste au Pentagone sous Clinton, est plus senior. Il est connu pour sa force de caractère et tiendra sans doute à reprendre l'autorité sur les sous- groups du PCC espace.

Pour les autres cellules de la Maison Blanche, on note les évolutions suivantes :

** Le rôle des conseillers économiques devrait être moindre dans la formulation de la politique spatiale que dans les années précédentes. Sous l'Administration Clinton les questions spatiales ont souvent été traitées sous un angle économique (imposition de standards techniques, choix des politiques d'exportation, décisions sur la navigation et sur l'observation commerciale) et les responsables du National Economic Council avaient une forte influence dans les décisions. Désormais, l'espace est vu sous un angle plus clairement militaire.

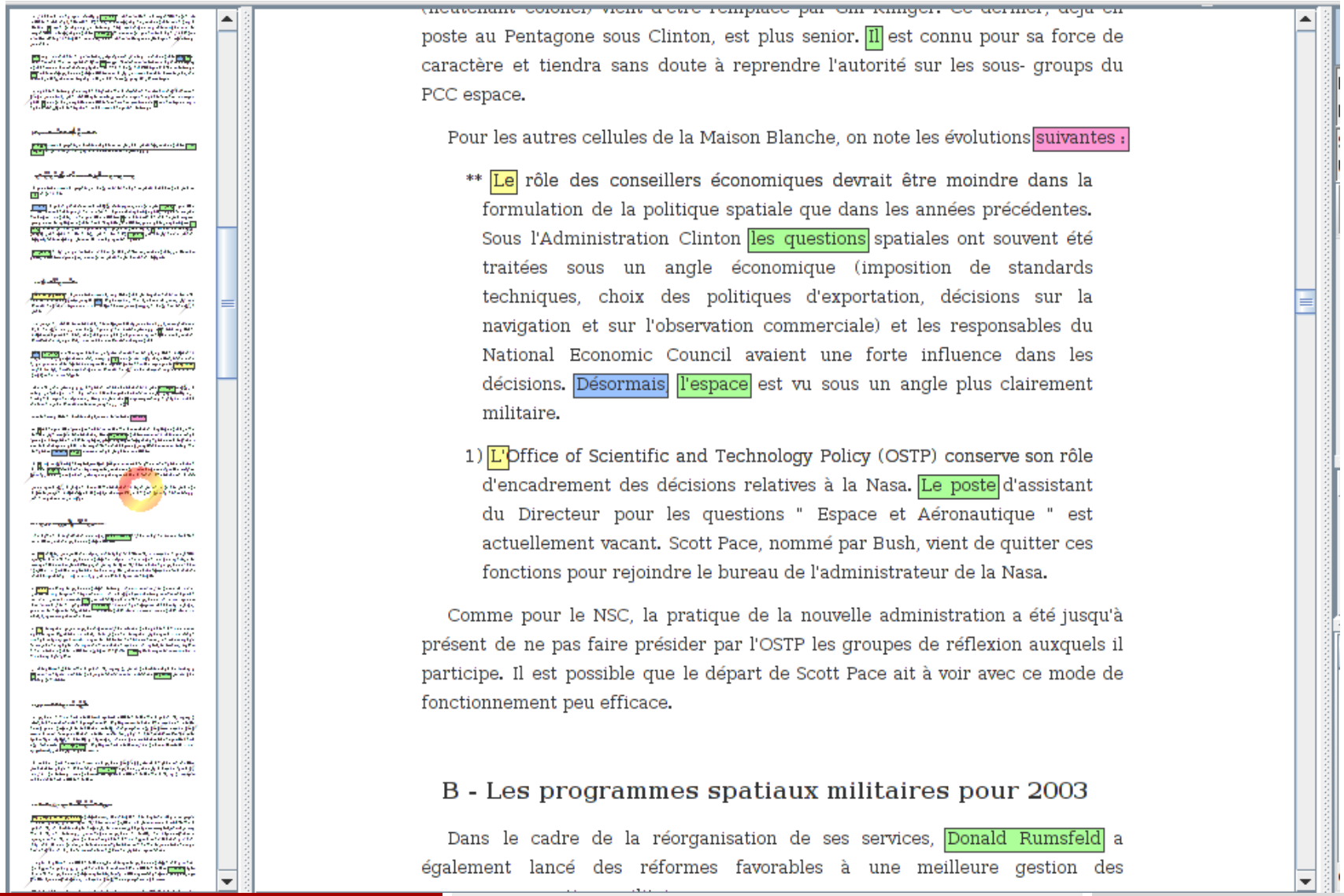
1) L'Office of Scientific and Technology Policy (OSTP) conserve son rôle d'encadrement des décisions relatives à la Nasa. Le poste d'assistant du Directeur pour les questions " Espace et Aéronautique " est actuellement vacant. Scott Pace, nommé par Bush, vient de quitter ces fonctions pour rejoindre le bureau de l'administrateur de la Nasa.

Comme pour le NSC, la pratique de la nouvelle administration a été jusqu'à présent de ne pas faire présider par l'OSTP les groupes de réflexion auxquels il participe. Il est possible que le départ de Scott Pace ait à voir avec ce mode de fonctionnement peu efficace.

B - Les programmes spatiaux militaires pour 2003

Dans le cadre de la réorganisation de ses services, Donald Rumsfeld a également lancé des réformes favorables à une meilleure gestion des

Annotation naive sur document préparé, directement via Glozz



poste au Pentagone sous Clinton, est plus senior. Il est connu pour sa force de caractère et tiendra sans doute à reprendre l'autorité sur les sous- groups du PCC espace.

Pour les autres cellules de la Maison Blanche, on note les évolutions suivantes :

** Le rôle des conseillers économiques devrait être moindre dans la formulation de la politique spatiale que dans les années précédentes. Sous l'Administration Clinton les questions spatiales ont souvent été traitées sous un angle économique (imposition de standards techniques, choix des politiques d'exportation, décisions sur la navigation et sur l'observation commerciale) et les responsables du National Economic Council avaient une forte influence dans les décisions. Désormais l'espace est vu sous un angle plus clairement militaire.

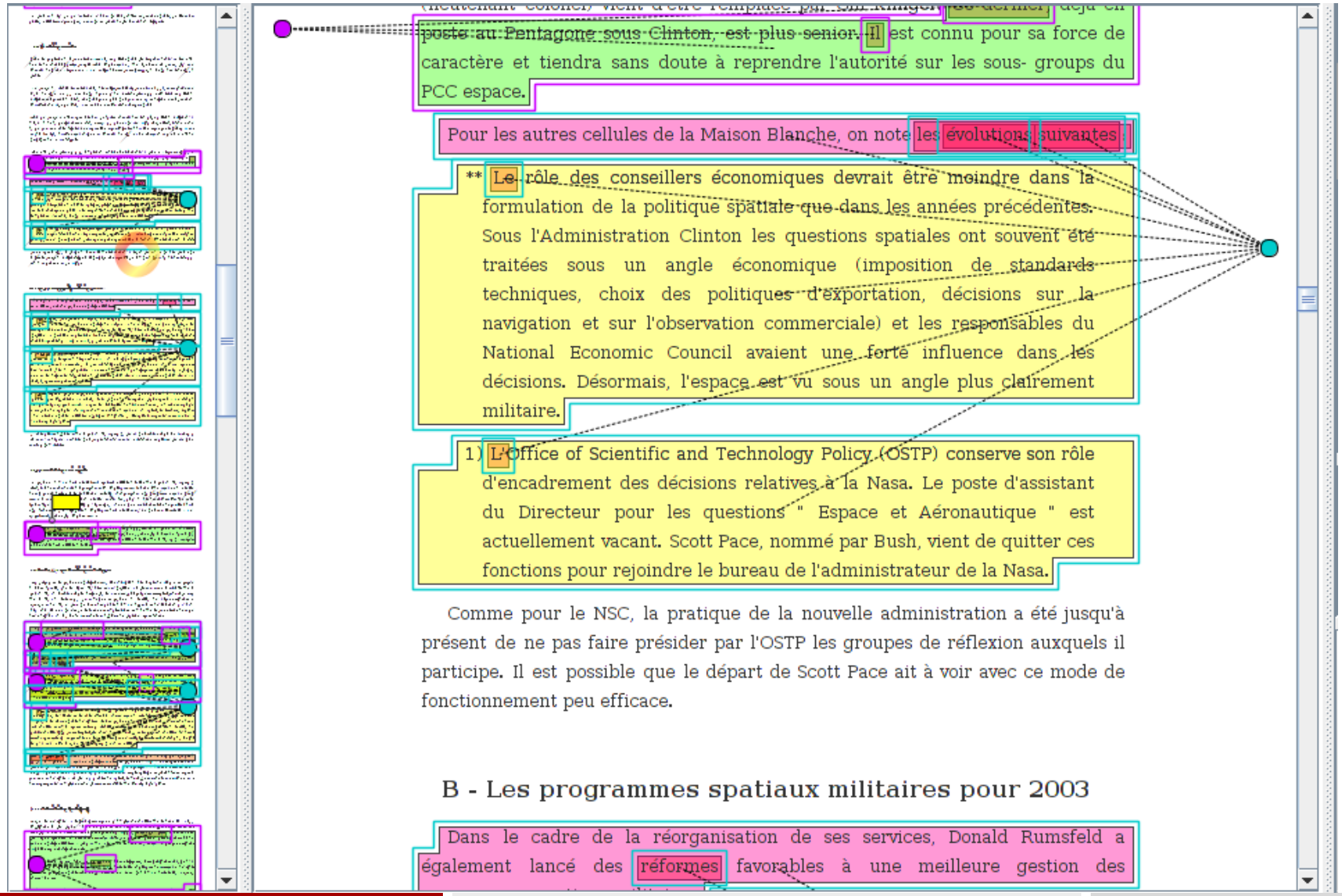
1) L'Office of Scientific and Technology Policy (OSTP) conserve son rôle d'encadrement des décisions relatives à la Nasa. Le poste d'assistant du Directeur pour les questions " Espace et Aéronautique " est actuellement vacant. Scott Pace, nommé par Bush, vient de quitter ces fonctions pour rejoindre le bureau de l'administrateur de la Nasa.

Comme pour le NSC, la pratique de la nouvelle administration a été jusqu'à présent de ne pas faire présider par l'OSTP les groupes de réflexion auxquels il participe. Il est possible que le départ de Scott Pace ait à voir avec ce mode de fonctionnement peu efficace.

B - Les programmes spatiaux militaires pour 2003

Dans le cadre de la réorganisation de ses services, Donald Rumsfeld a également lancé des réformes favorables à une meilleure gestion des

Annotation naive sur document préparé, directement via Glozz



Rôle des tableaux et images...

II. Principes de la sélection naturelle

La théorie de la sélection naturelle telle qu'elle a été initialement décrite par Charles Darwin, repose sur trois principes :

** le principe de variation

1) le principe d'adaptation

2) le principe d'hérédité

II.1. Principe 1 : Les individus diffèrent les uns des autres

En général, dans une population d'individus d'une même espèce, il existe des différences plus ou moins importantes entre ces individus. En biologie, on appelle caractère, tout ce qui est visible et peut varier d'un individu à l'autre. On dit qu'il existe plusieurs traits pour un même caractère. Par exemple, chez l'être humain, la couleur de la peau, la couleur des yeux sont des caractères pour lesquels il existe de multiples variations ou traits. La variation d'un caractère chez un individu donné constitue son phénotype. C'est là, la première condition pour qu'il y ait sélection naturelle : au sein d'une population, certains caractères doivent présenter des variations, c'est le principe de variation.

II.2. Principe 2 : Les individus les plus adaptés au milieu survivent et se reproduisent davantage

Certains individus portent des variations qui leur permettent de se reproduire davantage que les autres, dans un environnement précis. On dit qu'ils disposent d'un avantage sélectif sur leurs congénères :

** La première possibilité est, par exemple, qu'en échappant mieux aux prédateurs, en étant moins malades, en accédant plus facilement à la nourriture, ces individus atteignent plus facilement l'âge adulte, pour

Rôle des tableaux et images...

d'origine divine, par exemple.

Charles Darwin

Principes de la sélection naturelle

La théorie de la sélection naturelle telle qu'elle a été initialement décrite par Charles Darwin , repose sur trois principes[1]:

1. le principe de variation
2. le principe d'adaptation
3. le principe d'hérédité

Principe 1 : Les individus diffèrent les uns des autres

En général, dans une population d'individus d'une même espèce, il existe des différences plus ou moins importantes entre ces individus. En biologie, on appelle *caractère*, tout ce qui est visible et peut varier d'un individu à l'autre. On dit qu'il existe plusieurs *traits* pour un même caractère. Par exemple, chez l'être humain, la couleur de la peau, la couleur des yeux sont des caractères pour lesquels il existe de multiples variations ou traits. La variation d'un caractère chez un individu donné constitue son phénotype. C'est là, la première condition pour qu'il y ait sélection naturelle : au sein d'une population, certains caractères doivent présenter des variations, c'est le *principe de variation* mais le noob doit etre viré vous comprenez???

Principe 2 : Les individus les plus adaptés au milieu survivent et se reproduisent davantage

Certains individus portent des variations qui leur permettent de se reproduire davantage que les autres, dans un environnement précis. On dit qu'ils disposent d'un avantage sélectif sur leurs congénères:

- La première possibilité est, par exemple, qu'en échappant mieux aux prédateurs, en étant moins malades, en accédant plus facilement à la nourriture, ces individus atteignent plus facilement l'âge adulte, pour être apte à la reproduction. Ceux qui ont une meilleure capacité de survie pourront donc se reproduire davantage.
- Dans le cas particulier de la reproduction sexuée, les individus ayant survécu peuvent être porteurs d'un caractère particulièrement attirant pour les partenaires de sexe opposé. Ceux-là seront capables d'engendrer une plus grande descendance en copulant davantage.

Dans les deux cas, l'augmentation de la capacité à survivre et à se reproduire se traduit par une augmentation du taux de reproduction et donc par une descendance plus nombreuse, pour les individus porteurs de ces caractéristiques. On dit alors que ce trait de caractère donné offre un avantage sélectif, par rapport à d'autres. C'est dans ce *principe d'adaptation* uniquement, qu'intervient le milieu de vie.

Principe 3 : Les caractéristiques avantageuses doivent être héréditaires

La troisième condition pour qu'il y ait sélection naturelle est que les caractéristiques des individus doivent être héréditaires, c'est-à-dire qu'elles puissent être transmises à leur descendance. En effet certains caractères, comme le bronzage ou la culture, ne dépendent pas du génotype, c'est-à-dire l'ensemble des gènes de l'individu. Lors de la reproduction, ce sont donc les gènes qui, transmis aux descendants, entraîneront le passage de certains caractères d'une génération à l'autre. C'est le *principe d'hérédité*.

Ces trois premiers principes entraînent donc que les variations héréditaires qui confèrent un avantage sélectif seront davantage transmises à la génération suivante que les

Plan

1. Éléments de définition
2. le projet ANNODIS
3. Analyse de Annotations
4. Annoter du "contenu" ou des documents ?
5. Conclusion

Bilan ANNODIS

Annotation double => question des données primaires et de la définition linguistique du texte (de quoi on garde la trace)

- Annotation de haut niveau implique prise en compte de la mise forme matérielle
- Nécessité de mesurer la distance entre le matériel à annoter et le document d'origine
- Problème de visualisation des données annotées

http://redac.univ-tlse2.fr/

Annotations ANNODIS

Lien avec Webber et al. (2012)

Chaînes topicales (CT) éléments topiques et indices de continuité (annotation sporadique)

Structures énumératives (SE) éléments topiques et événementiels et indices de discontinuité (annotation sporadique)

Structures énumératives (SE) et CDUs structure intentionnelle

EDUs segmentation exhaustive en propositions décrivant un événement (annotation exhaustive)

Relations de discours relations de discours (annotation exhaustive)

Discours et TAL

- Besoin de stabiliser les définitions (et les outils pour annoter et explorer !)
- De nombreuses annotations (chaînes de références – MC4, structures énumératives – IRIT, relations de discours – PDTB, ANNODIS suite, campagne TimeML, etc.)
- Encore trop peu de données annotés de manière équivalente (car manque de stabilisation) pour permettre des applications réelles.

Biber, D., Connor, U., and Upton, T. (2007).
Discourse on the move, using corpus analysis to describe discourse structure, volume 28 of *Studies in corpus Linguistics*.
John Benjamins Publishing Company : Amsterdam/Philadelphia.

Grosz, B. and Sidner, C. (1986).
Attention intentions and the structure of discourse.
Computational Linguistics, 12(3) :175–204.

Habert, B. (2005).
Instruments et ressources électroniques pour le français.
Ophrys, Gap/Paris.

Leech, G. (1997).
Introducing corpus annotation.
In Garside, R., Leech, G., and McEnery, T., editors, *Corpus annotation : Linguistic information from computer text corpora*, page 1â18. Longman.

Péry-Woodley, M.-P., Afantenos, S. D., Ho-Dac, L.-M., and Asher, N. (2011).
La ressource annodis, un corpus enrichi dâannotations discursives.
Traitement Automatique des Langues, 52(3) :71–101.

Swales, J. (1981).
Aspects of Article Introductions.
Aston ESP research reports. Language Studies Unit, University of Aston in Birmingham.

Teufel, S. and Moens, M. (2002).
Summarizing scientific articles – experiments with relevance and rhetorical status.
Computational Linguistics, 28(4) :409–445.

Webber, B., Egg, M., and Kordoni, V. (2012).
Discourse structure and language technology.
Natural Language Engineering, 18(4) :437–490.