# Les corpus ANNODIS, une ressource multi-annotée : comment, pour quoi et pour qui ?

Marie-Paule Péry-Woodley

*Master ECIL – 7 novembre 2011*

# Organisation

1. Des questions sur l'annotation:
    1. Annoter, c'est quoi ?
    2. Annoter pour quoi ?
    3. Annoter quoi ?
    4. Et puis, annoter dans quoi ?
    5. Enfin, qui annote ?
2. Des réponses dans le projet ANNODIS:
    1. Présentation de la double annotation
        - Annotation ascendante
        - Annotation macro
    2. Les réponses ANNODIS avec 3 zooms:
        - Les guides d'annotation
        - La préparation du corpus
        - L'accord inter-annotateur

# 1. Les questions

# 1.1. Annoter, c'est quoi?

- ✓ [annotation is] the practice of adding interpretative linguistic information to an electronic corpus of spoken and/or written data (Garside et al, 1997)

- ✓ L'annotation consiste à ajouter de l'information (une interprétation *stabilisée*) aux données langagières (Habert, 2005)

# Ajouter de l'information

1. Segmenter pour déterminer/délimiter des *éléments* et leur attribuer une catégorie
   - Pavage complet (thématique, rhétorique) ou pointage sporadique (anaphores, structures particulières)
2. Éventuellement regrouper ces éléments
   - Identifier des *relations* : Ex. relations anaphoriques (expr. réf.) ; relations syntaxiques (syntagmes) ; relations de discours (segments de discours élémentaires)
   - Identifier des *structures* ou *schémas* englobants : Ex. ANNODIS : unités (amorce + items) + indices = structure énumérative

# 1.2. Annoter pour quoi?

- Recueillir des données pour tester des intuitions, des hypothèses
- Valider un modèle, le faire évoluer
- Produire un corpus annoté (corpus de référence?)
  - Pour des analyses linguistiques (manuelles ou outillées)
  - Pour de l'apprentissage automatique
  - Pour des applications spécifiques (TAL,didactique)

# 1.3. Annoter quoi?

## Cf. Leech (in Wynne 2005)

- *syntactic annotation*: how a given sentence is parsed, in terms of syntactic analysis into such units such phrases and clauses
- *semantic annotation* : e.g. semantic category of words
- *pragmatic annotation* : e.g. kinds of speech act (or dialogue act)
- *discourse annotation* : e.g. anaphoric links in a text
- *stylistic annotation*: e.g. adding information about speech and thought presentation (direct speech, indirect speech, etc.)
- *lexical annotation* : adding the identity of the lemma of each word form in a text — i.e. the base form of the word, such as would occur as its headword in a dictionary

# …mais aussi

- Des erreurs dans des corpus d'apprenants
- Des entités nommées, des expressions d'opinion
- Et des objets divers et variés:
  - fonctions discursives (Teufel et al. 1999, 2006 ; Biber et al. 2006)
  - qui fait quoi où? dans des transcriptions de commentaires de football (→résumé auto.) cf. Fort & Nazarenko 2011
  - segments obsolescents dans des encyclopédies (→ repérage auto.) cf. Laignelet et al. 2010

# 1.4. Annoter dans quoi?

- Constitution du corpus
  - Quel corpus pour quel projet d'annotation?
  - Corpus homogène ou diversifié?
  - Textes complets ou échantillons?
- Préparation du corpus
  - Texte nu ou enrichi ?
  - Métadonnées
  - Impact du nettoyage des textes: formatage, images, notes, tableaux etc.

# 1.5. Qui annote?

- Annotation automatique ou humaine ?
- Validation manuelle si annot. automatique ?
- Annotateurs naïfs ou experts ?
- Annotations par auteurs? Par utilisateurs? (contextes applicatifs)
- Questions autour des annotateurs
  - Formation des annotateurs (un naïf formé est-il encore naïf ?)
  - Fonctions du guide d'annotation
  - Accord inter-annotateur

# 2. Des réponses dans ANNODIS

- Projet financé par l'ANR 2007-10, 3 partenaires : CLLE-ERSS et IRIT, Toulouse ; GREYC, Caen
  http://w3.erss.univ-tlse2.fr/ANNODIS
- Double objectif :
  - Constitution d'un corpus de français écrit enrichi d'annotations concernant le niveau discursif
    - Une ressource pour les recherches le discours
    - Une ressource pour le développement d'applications en TAL
  - Exploitation du corpus annoté

# 2.1. ANNODIS : une double annotation

- Approche ascendante
  - des unités élémentaires vers des unités plus complexes par relations de discours
  - approche incrémentale et récursive
- Approche macro
  - motifs textuels multi-échelle jusqu'à très hauts niveaux d'organisation
    - interaction avec la structure de document
    - hypothèse de l'influence du niveau « macro » sur l'interprétation (niveau propositionnel et interpropositionnel)
  - zoning rhétorique (cf. atelier Y. Mathet & A. Widlöcher)
- Rencontre sur des points spécifiques (e.g. structures énumératives et élaboration)

# ANNODIS ascendante: principes d'annotation

- Segmentation en Unités de Discours Elémentaires (EDU)
- Construction récursive d'Unités Complexes (CDU) en reliant les EDU par des relations de discours (cadre théorique = SDRT)
  - Recherche d'un point d'attachement
  - Attribution d'une relation de discours (parmi les 16 décrites dans le guide d'annotation)

Annotation exhaustive des textes (ou extraits): pavage complet – 3 188 EDU & 3 355 relations

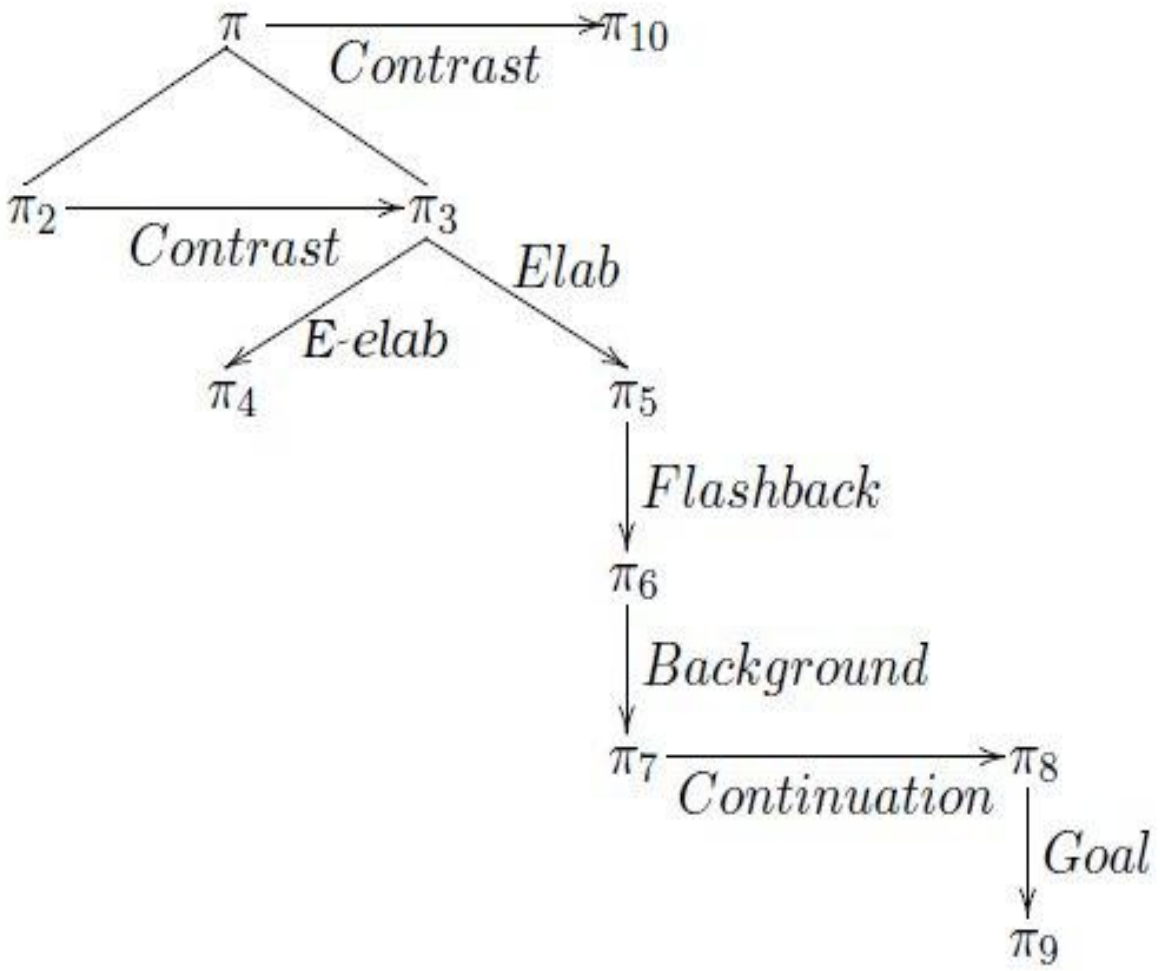
# ANNODIS ascendant : Ex. de segmentation

[Principes de la sélection naturelle.]_1 [La théorie de la sélection naturelle]_2 [telle qu'elle a été initialement décrite par Charles Darwin,]_3 [repose sur trois principes :]_4 [1.]_5 [le principe de variation]_6 [2.]_7 [le principe d'adaptation]_8 [3.]_9 [le principe d'hérédité]_10
[Principe 1 :]_11 [Les individus diffèrent les uns des autres.]_12 [En général,]_13 [dans une population d'individus d'une même espèce,]_14 [il existe des différences plus ou moins importantes entre ces individus.]_15 [En biologie,]_16 [on appelle caractère,]_17 [tout ce qui est visible et peut varier d'un individu à l'autre.]_18 [On dit]_19 [qu'il existe plusieurs traits pour un même caractère.]_20 [Par exemple,]_21 [chez l'être humain,]_22 [la couleur de la peau, la couleur des yeux sont des caractères pour lesquels il existe de multiples variations ou traits.]_23 [La variation d'un caractère chez un individu donné constitue son phénotype.]_24 [C'est là,]_25 [la première condition]_26 [pour qu'il y ait sélection naturelle :]_27 [au sein d'une population, certains caractères doivent présenter des variations,]_28 [c'est le principe de variation.]_29
[Principe 2 :]_30 ... [xxxx]_57
[Principe 3 :]_58 ... [xxxx]_70
[Ces trois premiers principes entraînent ...]_71 [xxxx]_80

# ANNODIS ascendant : relations annotées

elaboration(1/[2-80]) fusion(2/4) e-elab(2/3) elaboration([2,4]/[11-70]) elaboration([5-10]/[11-70]) elaboration([2,4]/[5-10]) continuation([5,6]/[7,8]) continuation([7,8]/[9,10]) fusion(5/6) fusion(7/8) fusion(9/10) frame([11,29]/[12-28]) elaboration(12/[13-28])
fusion(13/14) fusion(14/15) elaboration(15/[16-28]) frame(16/[18-20])
fusion(17/18) fusion(19/20) continuation([17,18]/[19,20]) elaboration([19,20]/22) fusion(21/23) frame(22/[21,23]) continuation(23/24)
continuation(24/[25-27]) fusion(25/26) fusion(26/27) elaboration([25-27]/28) continuation(11/29) continuation([11-29]/30)

# ANNODIS ascendant : graphe SDRT



# ANNODIS macro (motifs multi-échelle): principes d'annotation

- Constituer des données pour aborder trois caractéristiques de l'organisation du discours
  - multi-échelle : niveaux de grain, structure de document
  - multi-dimensionnelle : dimensions idéationnelle et textuelle
  - récursivité
- … et pour étudier la signalisation de ces structures
  - marqueurs = non seulement expressions lexicales dédiées, mais faisceaux de traits (à faire émerger)
- 2 stratégies discursives, 2 structures annotées
  - chaînage (unités connectées): chaînes topicales (SUR) –487 SUR annotées
  - empaquetage (unités rassemblées par critère commun) : structures énumératives (SE) – 829 SE annotées

# ANNODIS « macro »: annoter deux structures multi-échelles

## 1) Les Structures Énumératives (SE)

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

## 2) Les Segments ayant une Unité Référentielle (SUR)

- Segments caractérisés par le fait qu'une majorité des propositions qui les composent ont pour sujet le même référent

# Structure énumérative: exemple « classique »

**II ) Des orientations d'action**

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

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

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

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

4. Contribuer plus efficacement à la solution des principales crises régionales, ce [illegible] les actions suivantes : [...] **ITEM 4**

En conclusion, les turbulences qui affectent le moyen orient ont atteint un niveau de haute intensité qui représente, pour les pays occidentaux et, plus spécialement, pour l'Europe, de grands risques, notamment dans le domaine de la sécurité au sens large du terme : accroissement [illegible], perturbations dans notre approvisionnement en hydrocarbures, attaques contre nos forces au Liban, dislocation des États. **CLOSURE**

33

# Segment ayant une Unité Référentielle (SUR)

## Modèles climatiques [modifier]

Les projections par les scientifiques de l'évolution future du climat est possible par l'utilisation de modèles mathématiques traités informatiquement sur des superordinateurs[82]. Ces modèles, dits de circulation générale, reposent sur les lois générales de la thermodynamique et simulent les déplacements et les températures des masses atmosphériques et océaniques. Les plus récents prennent aussi en considération d'autres phénomènes, comme le cycle du carbone.

Ces modèles sont considérés comme valides par la communauté scientifique lorsqu'ils sont capables de simuler des variatio connues du climat, comme les variations saisonnières, le phénomène El Niño, ou l'oscillation nord-atlantique. Les modèles le plus récents simulent de façon satisfaisante les variations de température au cours du xx$^{e}$ siècle. En particulier, les simulatio menées sur le climat du xx$^{e}$ siècle sans intégrer l'influence humaine ne rendent pas compte du réchauffement climatique, tandis que celles incluant cette influence sont en accord avec les observations[11].

Les modèles informatiques simulant le climat sont alors utilisés par les scientifiques pour établir des scenarios d'évolution future du climat, mais aussi pour cerner les causes du réchauffement climatique actuel, en comparant les changements climatiques observés avec les changements induits dans ces modèles par différentes causes, naturelles ou humaines.

Ces modèles sont l'objet d'incertitudes de nature mathématique, informatique, physique, etc. Les trois principales sources d'incertitude mentionnées par les climatologues sont :

# 2.2. Les réponses ANNODIS

# ANNODIS: annoter pour quoi ?

## Objectifs ascendants

- Tester/valider un modèle pré-existant à la tâche
- Produire un corpus « de référence » utilisable pour des techniques d'apprentissage automatique, et aussi pour des études linguistiques du discours

## Objectifs macro

- Tester des hypothèses

→ modèle construit pour la tâche

- Produire un corpus enrichi de plusieurs niveaux d'annotation, utilisable pour des analyses de corpus (dont type *data-driven*)

# ANNODIS: annoter quoi ?

## Objets ascendants

- Décrire comment les EDU s'articulent en discours:
  - Point de rattachement
  - Type de relation rhétorique

## Objets macro

- Identifier deux structures
  - SE: délimiter puis regrouper des unités (amorce, items, clôture)
  - SUR: délimiter une portion de texte
- Indiquer comment elles sont signalées

# ANNODIS: annoter, c'est quoi ?

## Tâche ascendante

- Segmenter : pavage complet, 1 type d'unité (EDU)
- Regrouper par connexion: indiquer et typer exhaustivement les relations entre unités

## Tâche macro

- Segmenter : pointage sporadique
- Techniques de skimming assistées par pré-marquage
- Regrouper par empaquetage: identifier des schémas regroupant des unités et des indices

→Besoins différents

- choix et préparation des textes du corpus
- outils d'annotation

Pour les deux tâches, les annotations seront débarquées (stand-off annotation)

# Zoom 1: les guides d'annotation

- Fonction déclarative
  - Expliciter les objets (définitions, exemples)
  - Du modèle linguistique au modèle d'annotation (conception de l'interface)
- Fonction procédurale
  - Expliciter la tâche
  - Fournir une procédure action par action (apprentissage de l'interface)
  - La question des indices/marqueurs
  - Fournir des tests, prévoir les difficultés, les confusions possibles

# Les guides d'annotation (1): l'annotation des relations rhétoriques (ascendante)

- Relations rhétoriques
  - Principe
  - Attachement
  - Segments composés ou complexes
  - Gestion de l'incertitude
  - Relations
  - Équivalences
- Mode opératoire
  - Annotation quasi-manuelle
    - Noms des relations
  - Annotation avec outil spécifique
- Appendices
  - Table des marqueurs de relations
- Exemple complet

# [annodis:elaboration_annot]]





## Elaboration

Table des matières



**Nom pour l'annotation: elaboration**

La relation d'Elaboration relie deux propositions si la seconde proposition décrit un sous-état ou sous-événement de l'état ou l'événement décrit dans la première proposition.

La relation d'Elaboration inclut également les cas d'exemplification, de reformulation et paraphrase.

Éditer

## Exemples

### Elaboration entre événements

```
[Cette année-là vit de nombreux changements dans la vie de nos héros.]_1 [Jean épousa Adèle,]_2 [Marie s'acheta une maison à la
[et Paul partit pour le Brésil.]_4
```

elaboration(1,[2-4])

```
[La Lausitz, {une région pauvre de l'est de l'Allemagne,}_1 {réputée pour ses mines de charbon à ciel ouvert,}_2 a été le théât
première mondiale, mardi 9 septembre.]_3
[Le groupe suédois Vattenfall a inauguré, dans la petite ville de Spremberg, une centrale électrique à charbon expérimentale qu
toute la chaîne des techniques de captage et de stockage du carbone (CCS).]_4
```

e-elab(3,[1-2]) % la lausitz
elaboration(3,4)

Éditer

## Marqueurs d'Elaboration

Le plus souvent, Elaboration apparait sans marqueur.
Seuls les cas d'Elaboration qui sont des exemplifications ou des reformulations apparaissent parfois avec les marqueurs suivants :
par exemple, notamment, c'est à dire, à savoir

Éditer

## Confusions possibles

### Arrière-Plan

Arrière-Plan relie un état à un événement, et l'état ne fait pas partie de cet événement, il décrit un aspect de la scène dans laquelle l'évenement se déroule, même si cette scène concerne des participants de l'événement (cf. Marie, portant un chapeau, est entrée dans le bar). A l'inverse, Elaboration relie soit deux événements, soit deux états, dont l'un fait partie de l'autre. La distinction état / évenement suffit donc à séparer Elaboration d'Arrière-Plan.

### Elaboration d'entité

Elaboration introduit des précisions sur un événement ou un état, mais ces précisions sont vues en termes de sous-événements ou sous-états. Lorsque la précision porte sur un participant de l'événement ou l'état, il s'agit plutôt d'une Elaboration d'entité. Là encore, la distinction état / événement peut parfois aider, puisque la description d'une entité est un état, alors que le premier argument peut décrire un événement.

Retour au manuel

# Les guides d'annotation (2): annoter une amorce de SE (macro)

- 2.1 Amorce
  - Définition
  - Illustration
  - Indices
  - Tests

# ANNODIS : annotation des structures

| | | |
|---|---|---|
| intro | annotation | procédures |
| SE | SUR | exemples | imprimer |
| amorce | item | clôture | |

Texte d'illustration : le rapport Avicenne

## 2.1 Amorce

[ex][ind][test]

### Définition

L'***amorce*** est un segment qui annonce une énumération. Elle peut comporter un prospect contenant ce que nous appelons un ***énumérathème***. Il s'agit d'un lexème qui a pour fonction de spécifier le critère de co-énumérabilité des items de l'énumération, autrement dit d'expliciter ce qui justifie la réunion des items autour d'un même ***thème énumératif***

### Illustration

L'amorce apparaît surlignée et le prospect en ***italique-gras***. On désigne ainsi le groupe nominal composé généralement d'un déterminant numéral et d'un nom. Ce nom appelé ici énumérathème (dans l'exemple ci-dessous, il s'agit du nom ***avantage***), désigne le critère de réunion des items de l'énumération.

**Exemple d'amorce avec énumérathème**

Placer l'accent sur l'occupation et la nécessité d'y mettre fin, aurait *trois avantages* : repositionner le débat autour du problème de la terre et non des identités religieuses pour redonner ainsi force au courant

pédagogique utile. Placer l'accent sur l'occupation et la nécessité d'y mettre fin, aurait trois avantages : repositionner le débat autour du problème de la terre et non des identités religieuses pour redonner ainsi force au courant nationaliste que les pragmatiques de la mouvance islamiste sont prêts à suivre ; découpler l'enjeu de la lutte contre l'occupation de celui du droit à l'existence d'Israël en réaffirmant les droits des deux peuples à vivre chacun dans un Etat viable et à l'intérieur de frontières sûres ; désamorcer le débat qui lie l'opposition à la politique israélienne à la question de l'antisémitisme.

Pour ne pas se laisser bloquer par les membres réticents de l'Union, il convient d'utiliser le clavier de la diplomatie française pour tester des idées ou options à l'échelle nationale, puis européaniser celles -ci quand elles s'avèrent réussies ou prometteuses. Elles auront alors de plus grandes chances de vaincre les réticences, d'autant qu'elles peuvent aussi s'appuyer sur des acteurs non européens : Russie, Chine, monde arabe, etc

Dans cette même perspective dynamique, il nous faut aller au-delà d'une simple observation des tentatives régionales de recherche de compromis pour les accompagner. Les Palestiniens en période de crise se tournent naturellement vers le giron arabe. Le plan de paix arabe offre une porte d'entrée au Hamas pour intégrer un processus diplomatique par étapes. Les conditions du Quartet ne seront pas acceptées par un gouvernement d'union nationale palestinien sans une reformulation des ces exigences ou un rééquilibrage dans le sens d'exigences équivalentes à l'égard d'Israël. Leur maintien figé est de nature à perpétuer l'impasse.

La France pourrait, conjointement avec d'autres États européens, prendre l'initiative d'un tel exercice de reformulation, à mi-chemin entre les textes des accords de la Mecque, du plan de paix arabe et

# ANNODIS : annotation des structures

| | | | |
|---|---|---|---|
| intro | annotation | procédures | |
| SE | SUR | exemples | imprimer |
| amorce | item | clôture | |

Texte d'illustration : le rapport Avicenne

## Indices

Les indices participant au signalement d'une amorce peuvent être :

- les prospects : ces syntagmes nominaux à valeur cataphorique sont généralement au pluriel et comportent un déterminant numéral, un indéfini (*quelques*, *plusieurs*, etc) ou un collectif (*une foule de*, *une grande variété de*, *un grand nombre de*, etc). La tête lexicale de ces syntagmes indique le critère de réunion des éléments énumérés (*conséquences*, *enjeux*, *avantages*, etc) :c'est l'énumérathè:me.
- les titres de section peuvent également constituer un indice d'amorce : des sections (titres compris) de niveau inférieur : exemple de SE à travers la titraille, de la section titrée : exemple de SE amorcée par le titre de section
- les deux-points indiquent fréquemment la fin d'une amoorce.

Les indice d'amorce qui ont été repérés automatiquement apparaissent colorés en rose dans le texte comme le montre l'exemple ci-contre.

## Tests

Pour repérer l'énumérathème d'une amorce, vous pouvez tenter d'insérer ***tel(le)s que énuméré(e)s ci-dessous*** et/ou ***tel(le)(s) que décrit(e)(s) ci-dessous*** immédiatement après l'expression présumée en être un. La possibilité d'une telle insertion confirme sa présence.

Le dialogue doit donc être modulé avec pragmatisme, c'est-à-dire en fonction du mouvement concerné, une grande variété de formules s'offrant autour des suivantes :

- un dialogue à caractère technique pour la mise en oeuvre de coopérations ; il pourrait impliquer des collectivités locales, voire des responsables syndicaux ou d'ONG ; il s'agit d'une approche essentiellement pratique n'allant pas, sur le plan politique, au-delà d'une sorte de signal ;
- un dialogue informel à travers des rencontres et séminaires associant des personnalités d'origine diverse. Le contenu politique serait plus fort mais ne lierait pas les autorités ; il pourrait donc inclure des mouvements répondant aux exigences déjà mentionnées sans être formellement reconnues par le pouvoir en place ( ainsi les Frères musulmans en Egypte ) ;
- un dialogue politique lui -même modulable : à Paris, ou dans la capitale concernée ou dans un lieu tiers ; à un niveau subalterne ou responsable ; direct ou via des intermédiaires ; bilatéral ou à l'occasion d'une réunion plus large etc

L'important doit être une disposition au dialogue pour autant que l'interlocuteur respecte, lui aussi, ce que nous sommes.

### 4 Contribuer plus efficacement à la solution des crises régionales.

*4.1 La question palestinienne*

Depuis les accords d'Oslo, la voie tracée pour le règlement du conflit israélo-palestinien résidait dans un consensus international de soutien

# Les guides d'annotation: procédure d'annotation macro (avec Glozz)

- 4.2.1 Charger les textes à annoter
- 4.2.2 Distinguer plusieurs étapes d'annotation et jouer avec les styles
- 4.2.3 Repérer les structures discursives d'un texte (SE/SUR)
- 4.2.4 Valider, supprimer, créer les indices
- 4.2.5 Regrouper les éléments composant une structure discursive (SE/SUR)
- 4.2.6 Modifier et supprimer une annotation
- 4.2.7 Enregistrer les annotations
- 4.2.8 Gestion de l'incertitude

# Les guides d'annotation: éléments de bilan d'expérience

- La rédaction itérative du guide : un travail important
  - d'explicitation collective du modèle linguistique qui sous-tend le modèle d'annotation (f. déclarative)
  - de définition des fonctionnalités de l'interface (f. procédurale)
- Il faut tester le guide pour évaluer
  - la cohérence des définitions
  - la faisabilité de la tâche
  - l'utilité et l'utilisabilité de l'interface
- Le guide a deux fonctions:
  1. guider les annotateurs pendant l'annotation du corpus
  2. guider les utilisateurs dans l'exploitation du corpus annoté

# Les guides d'annotation: tentative de chronologie idéale

1. Rédaction d'une version préliminaire du guide
   - accord entre experts sur les définitions des objets, sur la tâche, son découpage en actions
2. Annotation exploratoire
   - Observation des problèmes rencontrés, analyse des désaccords
3. Révision collective du guide
4. Phase 1 de l'annotation opérationnelle (sur corpus de test)
   - Calcul de l'accord inter-annotateurs. Révision itérative tant que le taux d'accord est insuffisant.
5. Finalisation du guide et annotation
6. Après l'annotation et le dépouillement des données annotées, ajout de notes pour l'utilisateur

# ANNODIS : annoter dans quoi ?

## Exigences ascendants

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

## Exigences macro

- 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

• Corpus différents, mais avec une section commune
• Corpus diversifié
• Besoins différents en terme de visualisation des textes pour l'annotation et pour l'exploitation

# Zoom 2 : la préparation du corpus ANNODIS macro

- Un corpus *diversifié*
  - pour tester la variabilité des fonctionnements observés
  - 73 textes (557 047 mots) en 3 sous-corpus
- Une approche *outillée*
  - Prétraitements
    - Textes encodés en XML aux normes TEI-P5
    - Documentation du corpus : métadonnées
    - Prétraitements spécifiques en lien avec nos hypothèses sur fonctionnements discursifs
  - Fonctionnalités dédiées de l'interface Glozz

Encodage : texte --> XML TEIP5

- métadonnées
- propriétés visuelles (structure de document, mise en forme)

Prémarquage sur corpus analysé syntaxiquement

- traits associés à l'organisation discursive
- traits spécifiquement associés aux structures à l'étude

Annotation manuelle

- structures énumératives (SE) : [amorce? + items + clôture? +indices --> schéma SE]
- chaînes topicales (CT) : [zone + indices --> schéma CT]

# Corpus ANNODIS macro : prétraitements (1)

- Encodage spécifique:
  - Encodage de la structure du document (découpage en sections titrées hiérarchisées) et de la mise en forme matérielle
    - → représentation visuellement proche de la mise en forme originale (pour l'annotateur)
    - traits de structure et de mise en forme – découpage, hiérarchie, indentation etc. – exploitables pour le prémarquage
- Cf.hypothèses sur interaction entre structures à l'étude (surtout SE) et structure de document

# Corpus ANNODIS macro : prétraitements (2)

- Prémarquage automatique de traits associés (de + ou – près) aux structures à l'étude
  - Objectif 1: assister la phase d'annotation en autorisant des procédures de "text scanning"
  - Objectif 2: permettre l'identification de marqueurs complexes (faisceaux d'indices) par des techniques de fouille (phase d'exploitation)
- Le prémarquage est réalisé par le biais de grammaires locales (programmes perl) projetées sur les textes étiquetés (TreeTagger) et analysés syntaxiquement (Syntex)
- En lien avec hypothèses sur la signalisation des SE, leur classification formelle et fonctionnelle

# Le prémarquage: traits généralement associés à la signalisation des SE et SUR

| Traits | Description. . . |
|---|---|
| Listes formatées | puces, indentations. . . |
| Patrons ponctuationnels | *: [...] ; [...] et/ou [...]. . .* |
| Séquenceurs (MIL) | *Premièrement, Un second X, Parallèlement, Enfin, etc.. . .* |
| Prospections | SN pluriels + selon, suivants, etc... ou SN dont la tête = nom générique *(selon trois points, les éléments suivants. . .). . .* |
| Encapsulations | SN démonstratifs dont la tête est un nom générique et/ou avec adjectif numéral *(ces trois scénarios. . .). . .* |
| Expressions co-référentielles | pronoms, réitérations, etc. . . . |

# Le prémarquage: traits associés à la signalisation de l'organisation du discours

| Traits | Description. . . |
| --- | --- |
| Typographie et disposition | titres de section, sauts de paragraphe. . . |
| Position textuelle | éléments détachés en initiale, position sujet... |
| Éléments lexico-syntaxiques | adverbiaux circonstanciels (spatial, temporel, notionnel), connecteurs. . . |

# Le prémarquage: visualisation sous Glozz pour l'annotation

## 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 :

1) le principe de variation

2) le principe d'adaptation

3) le principe d'hérédité

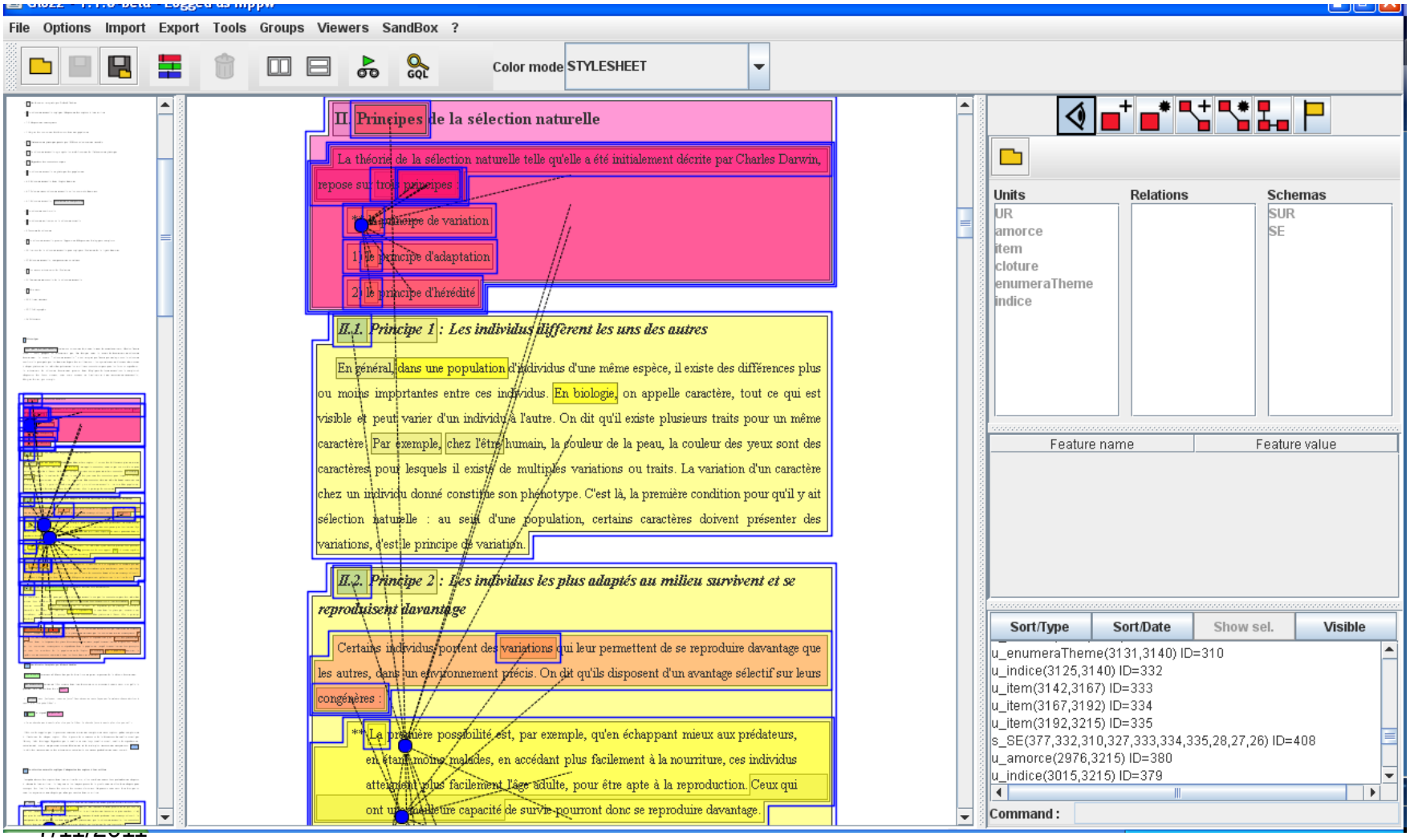
### *II.1. Principe 1 : Les individus diffèrent les uns des autres*

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

# ANNODIS ascendant : Ex. de segmentation

[Principes de la sélection naturelle.]_1 [La théorie de la sélection naturelle]_2 [telle qu'elle a été initialement décrite par Charles Darwin,]_3 [repose sur trois principes :]_4 [1.]_5 [le principe de variation]_6 [2.]_7 [le principe d'adaptation]_8 [3.]_9 [le principe d'hérédité]_10
[Principe 1 :]_11 [Les individus diffèrent les uns des autres.]_12 [En général,]_13 [dans une population d'individus d'une même espèce,]_14 [il existe des différences plus ou moins importantes entre ces individus.]_15 [En biologie,]_16 [on appelle caractère,]_17 [tout ce qui est visible et peut varier d'un individu à l'autre.]_18 [On dit]_19 [qu'il existe plusieurs traits pour un même caractère.]_20 [Par exemple,]_21 [chez l'être humain,]_22 [la couleur de la peau, la couleur des yeux sont des caractères pour lesquels il existe de multiples variations ou traits.]_23 [La variation d'un caractère chez un individu donné constitue son phénotype.]_24 [C'est là,]_25 [la première condition]_26 [pour qu'il y ait sélection naturelle :]_27 [au sein d'une population, certains caractères doivent présenter des variations,]_28 [c'est le principe de variation.]_29
[Principe 2 :]_30 ... [xxxx]_57
[Principe 3 :]_58 ... [xxxx]_70
[Ces trois premiers principes entraînent ...]_71 [xxxx]_80

# Au terme de l'annotation: des traits prémarqués aux indices validés



# Corpus ANNODIS macro (fin): des traits aux marqueurs

Prémarquage
[automatique]

*Traits* prémarqués sur texte étiqueté et analysé syntaxiquement

= indices candidats

→

Annotation
[manuelle]

- Traits prémarqués validés = *indices*
- Nouveaux indices annotés = *indices*

Les *indices* (et éventuellement certains traits) permettront d'identifier des *marqueurs* complexes (faisceaux d'indices) pour la caractérisation et le typage des structures

# ANNODIS: qui annote ?

## Annotation ascendante

- Annotation naïve : 3 étudiants L3 avec guide, 47 textes annotés x 2
  - Accord faible : 65% pour segmentation, kappa =0.45 pour relations
- Annotation experte : 7 experts, arbitrage des 47 « naïfs » et annotation de 40 textes, production d'une annotation de référence

## Annotation macro

- Annotation naïve: 3 étudiants L3 avec guide
  - phase A = 3 textes x 3 annotateurs
  - phase B = 6 textes x 3 annotateurs. F mesure = 0.7
  - phase C = 73 textes x 1 annotateur chacun
- Annotation experte de validation

# Zoom 3 : l'accord inter-annotateur
## comparaison des annotations ascendantes

- Définir l'identité entre annotations
  - segments identiques ? Relations identiques?
- Contours à déterminer à partir de 2 regards:
  - rapprochement entre segments pour une même RD
  - rapprochement entre segments pour une RD différente

# Comparaison des doubles annotations ascendantes : annotations « identiques »

- **Inversion Attachement :**

  Result (20/21) par ANN1 Result (21/20) par ANN2

  → Result (20/21) par ANN2

- **Attachement partiel :**

  Explanation ([19,20,21,23,24,**25**]/26) par ANN1

  Explanation (**25**/26) par ANN2

  *Autres cas possibles (autres textes) :*

  Explanation (10, [11-12]) et Explanation (10,11)

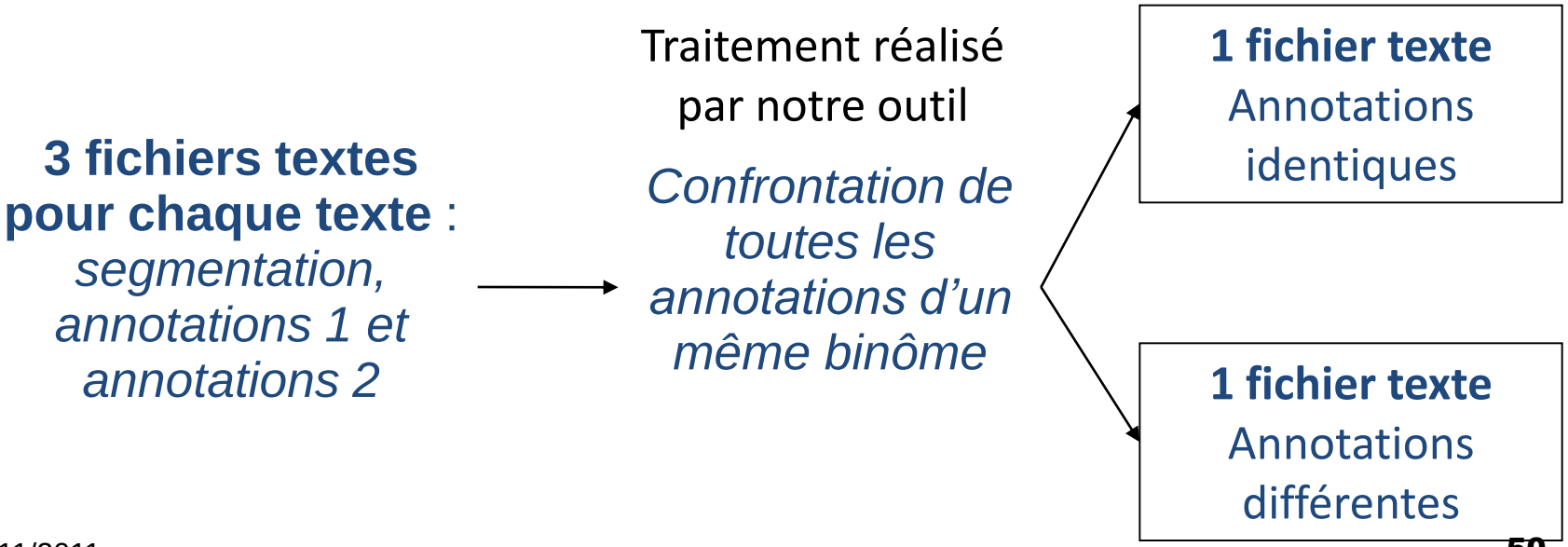
  Result ([21,22,23],[24,25]) et Result (23,24)

# Comparaison des doubles annotations ascendantes : annotations différentes

- **Annotations appariables :**
  - Parallel([9,10]/11) par ANN1
  - Explanation(11/[8,9,10] par ANN2
- **Annotations non appariables :**
  - 2 segments reliés par 1 RD chez ANN1
  - pas reliés chez ANN2

# Comparaison des doubles annotations ascendantes : méthode de confrontation

- Annotations identiques *vs* annotations différentes
- Annotations appariables *vs* annotations non appariables



# Comparaison des doubles annotations ascendantes : résultats des confrontations

| | *Explication* | *Résultat* | Total |
|---|---|---|---|
| Annotations identiques | 25 | 22 | 47 |
| Annotations différentes appariables | 72 | 68 | 140 |
| Annotations différentes non appariables | 31 | 31 | 62 |

- 47 cas d'annotations identiques (94 annotations)
- 202 cas d'annotations différentes
- Au total, 296 annotations concernant les relations d'*Explication* et de *Résultat*

# Comparaison des doubles annotations ascendantes : observations

- Correction (ou validation) de chaque annotation
  - Désaccord sur les arguments de la relation
    - Problème de la formation des segments complexes
  - Désaccord sur la nature de la relation
    - Aucune relation pragmatique repérée
    - Confusion entre des relations :

    [*C'est probablement l'explication de l'expression française.* ]_26
- Pb des « marqueurs (possibles) » du guide : tendance au keywording chez les annotateurs

# Annoter des marqueurs

## ANNODIS ascendant: marqueurs possibles

| Résultat : | Contraste : | Explication : |
|---|---|---|
| *du coup, donc, par conséquent, en conséquence, par suite, à la suite de quoi* | *mais, cependant, toutefois, par contre, bien que, néanmoins, si* | *car, parce que, a cause de, du fait de, par la faute de, grâce à, si 1 c'est parce que 2, depuis (si causalité évidente)* |

- ANNODIS ascendant : annoter arguments & relation en utilisant les définitions (avec listes de marqueurs) du guide
- PDTB : 1) annoter arguments & relation depuis marqueurs explicites pré-annotés (liste); 2) autres relations identifiées: insérer marqueur « implicite »
- ANNODIS macro : annoter des indices = valider traits prémarqués et annoter nouveaux indices

# Remarques conclusives (1)

- Corpus annotés dits « de référence »
  →recherche cumulative → avancées scientifiques
- Mais la qualité et fiabilité des annotations est fonction de l'état de la science…
- Il convient peut-être de différencier
  - annotation de référence : modèle stabilisé et consensuel
  - annotation de recherche : expérimentale
- Un corpus de référence annoté discursivement: une gageure dans l'état actuel des recherches

# Remarques conclusives (2)

- Ce qu'on aurait pu faire et qu'on n'a pas fait :
  - L'annotation comme processus : la procédure d'annotation manuelle naïve comme expérimentation (cf. psycho. cog.)
    - temps d'annotation, confusions entre relations = information utilisable
  - Retour d'expérience des annotateurs
  - Carnet de bord d'annotation

# Au final, la ressource ANNODIS

- Corpus ANNODIS-ascendant:
  - 87 textes/extraits, 28 Kmots, 3188 EDU, 3355 relations
- Corpus ANNODIS-macro :
  - 82 documents, 666 Kmots, 1316 structures, 7873 indices
- Corpus ANNODIS-duo :
  - 18 extraits annotés, 7 Kmots

# Bibliographie


- Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 3  4, 555-596.
- Biber, D., Connor, U., & Upton, T. A.  (2007). *Discourse on the move: using discourse analysis to describe discourse structure. Amsterdam, Philadelphia: John Benjamins.*
- Habert, B. (2005). Portrait de linguiste(s) à l'instrument. *Texto, X/4.*
- Garside, R., Leech, G., & McEnery, A. (Eds.).  (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora . London, Addison Wesley.*
- Laignelet, M., Péry-Woodley, M.-P., & Tanguy, L. (2010). Découverte de configurations de traits textuels pour la caractérisation des segments d'obsolescence. *Document Numérique 13  3, 41-68.*
- Power R., Scott, D. & Bouayad-Agha, N. 2003. "Document structure." *Computational Linguistics* 29 (2): 211-260.

- Prasad, R., Miltsakaki, E., Joshi, A., & Webber, B. (2004). Annotation and Data Mining of the Penn Discourse TreeBank, *ACL 2004 Workshop on Discourse Annotation*, Barcelona, Spain p. 88-95.
- Prévot , L., Vieu, L. & Asher, N. (2009). Une formalisation plus précise pour une annotation moins confuse: la relation d'élaboration d'entité*. Journal of French Language Studies*, 19, pp. 207-228.
- Spooren, W., & Degand, L. (2010). Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistics Theory* 6 /2, 241-266.
- Stede, M., Wiebe, J., Hajicova, E., Reese, B., Teufel, S., Webber, B., & Wilson, T. (2007). Discourse annotation working group report, Proc. of the Linguistic Annotation Workshop at ACL 07, Prague.
- Teufel, S., Carletta, J., & Moens, M. (1999). An annotation scheme for discourse-level argumentation in research articles" *EACL 1999*
- Teufel, S., Siddharthan, A., & Tidhar, D. (2006). An annotation scheme for citation function, *Sigdial-06, Sydney, Australia.*

- Webber, B., Joshi, A., Miltsakaki, E., Prasad, R., Dinesh, N., Lee, A., & Forbes, K. (2005). A Short Introduction to the Penn Discourse TreeBank. *Copenhagen Working Papers in Language and Speech Processing* .
- Wilcock, G. (2009). *Introduction to Linguistic Annotation and Text Analytics*: Morgan & Claypool Publishers. Companion website http://sites.morganclaypool.com/wilcock.
- Wynne, M. (Ed.). (2005). *Developing Linguistic Corpora: a Guide to Good Practice . Oxford, Oxbow Books.*


Bibliographie et webographie ANNODIS sur:

http://w3.erss.univ-tlse2.fr/annodis/