Carnets de Grammaire

Rapports internes de CLLE-ERSS

Rapport n°20 – décembre 2012

La ressource ANNODIS multi-échelle : guide d'annotation et "bonus"

Maud Colléter[*], Cécile Fabre[*],
Lydia-Mai Ho-Dac[*], Marie-Paule Péry-Woodley[*],
Josette Rebeyrolle[*], Ludovic Tanguy[*]

[*] Cognition, Langues, Langage, Ergonomie (CLLE-ERSS, UMR 5263), CNRS & Université de Toulouse-Le Mirail
Courriel: cecile.fabre,hodac,pery,josette.rebeyrolle,ludovic.tanguy@univ-tlse2.fr, maud.colleter@gmail.com

Carnets de Grammaire est le nom d'une série de rapports internes édités par CLLE-ERSS. Cette série de rapports prédiffuse des travaux que leur degré d'aboutissement, leur nature ou leur longueur ne permettent pas de publier rapidement par les canaux habituels.

Comité de rédaction

Michel Aurnague, Hélène Giraudo, Frédéric Lambert, Fabio Montermini, Patrick Sauzet, Ludovic Tanguy

ISSN : 1965-0019

CLLE-ERSS – Maison de la Recherche – Université de Toulouse-Le Mirail
5, allées Antonio Machado
31058 Toulouse Cedex 9

Table des matières

Pourquoi ce carnet de grammaire ? **2**

1 Avant la campagne : préparation des textes et rédaction du guide **4**
1.1 Choix du corpus 4
1.2 Des documents originaux aux textes annotables 5
1.3 Annotation exploratoire 6

2 Guide d'annotation livré aux annotateurs **8**

3 Déroulement de la campagne d'annotation **34**

4 Accord inter-annotateur et production d'une version Gold **35**
4.1 Accord inter-annotateur 35
4.1.1 Calcul des SE annotées 35
4.1.2 Calcul entre CT annotées 36
4.2 Méthode d'arbitrage et constitution du "gold" 36

5 Post-traitements : des données brutes à des données exploitables **38**

6 Postface : retours sur la campagne **39**

7 Travaux publiés **52**

Annexe **54**

Pourquoi ce carnet de grammaire ?

La mise en ligne de la ressource ANNODIS, corpus de français écrit enrichi d'annotations discursives, s'accompagne de la parution de deux Carnets de Grammaire, correspondant chacun à un des deux volets du programme d'annotation : relations rhétoriques (ANNODIS_rr) et structures multi-échelles (ANNODIS_me). Dans ce Carnet de Grammaire, nous présentons en le contextualisant le guide d'annotation rédigé pour l'annotation manuelle des structures multi-échelles – structures énumératives et chaînes topicales – de la ressource ANNODIS_me. Cette contextualisation vise un double objectif et deux types de lecteurs : pour les utilisateurs de la ressource, l'ensemble de la démarche d'annotation doit être rendu compréhensible ; pour des chercheurs s'engageant à leur tour dans une entreprise d'annotation, il nous a paru utile de conserver la trace des questionnements, tâtonnements et éventuels changements de direction à différents moments du projet. Autour du guide original sont donc agencés les éléments suivants, ordonnés selon la chronologie de la campagne d'annotation : la section 1 est consacrée à la construction et préparation du corpus à annoter et à l'annotation exploratoire sur laquelle se fonde la rédaction du guide. Le guide tel que l'ont utilisé les annotateurs est reproduit en section 2. La section suivante (3) présente l'organisation du déroulement de la campagne en trois phases. C'est cette organisation qui a permis de calculer l'accord inter-annotateur et de produire une "version Gold" à partir des textes multi-annotés, aspects qui sont présentés dans la section 4. La section 5 dresse la liste des post-traitements grâce auxquels la ressource a été nettoyée des inévitables scories. Enfin, en postface (6), nous articulons plusieurs types de retours sur la campagne : une reformulation a posteriori des définitions du modèle d'annotation, illustrée par de "beaux" exemples sélectionnés parmi les structures annotées, et accompagnée de remarques des annotateurs.

Avant d'entrer dans le vif du sujet, nous situons brièvement la ressource dans son contexte institutionnel ; nous portons ensuite un regard en arrière pour formuler certaines questions qui se posent dans un tel projet et rappeler les grands axes des réponses qui ont été apportées dans le cadre d'ANNODIS.

La ressource ANNODIS_me, distribuée ainsi que la ressource ANNODIS _rr sur le site REDAC (Ressources développées à CLLE-ERSS)[1], est issue du projet ANNODIS, projet financé par l'Agence Nationale pour la Recherche

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

(appel Corpus 2007) [2], dont un objectif était la constitution d'un corpus diversifié de français écrit enrichi d'annotations discursives, le second étant le développement de ressources logicielles pour l'annotation [3]. Elle a été élaborée par des membres du laboratoire CLLE-ERSS avec la contribution d'étudiants du département de Sciences du langage de l'Université de Toulouse 2.

Nous ne reviendrons pas sur les arguments linguistiques concernant le choix des objets à annoter, qui ont été déployés ailleurs (cf. Péry-Woodley et al., 2012 ; Ho-Dac et al., 2012), notre objet ici étant le processus de constitution de la ressource annotée. L'expérience ANNODIS a ceci de particulier qu'elle est sous- tendue par deux approches des structures discursives. Cette dualité d'approche a fait qu'aucun aspect de l'expérience ne pouvait être considéré comme acquis : depuis la constitution du corpus jusqu'aux post-traitements, en passant par les différentes phases de l'annotation, tout a été matière à discussion. Nous évoquerons dans ce qui suit de nombreuses questions qui se sont posées au fil de l'expérience, parmi lesquelles :

- L'annotation implique-t-elle une couverture complète des textes ou s'agit-il d'une annotation partielle correspondant à un pointage sélectif de segments possédant des propriétés spécifiques ?
- Dans les deux cas, quelles sont les unités pertinentes pour la segmentation préalable ?
- Si annoter consiste à ajouter aux données langagières de l'information correspondant à une interprétation stabilisée (Habert, 2005), une tâche, et non des moindres dans un domaine aussi labile que l'étude des structures discursives, sera de stabiliser un modèle le temps de l'expérience (modèle qui prend corps dans le guide d'annotation).
- Sur la base de ce modèle, le guide d'annotation présente une tâche d'annotation. Il y a lieu d'examiner avec soin la faisabilité de la tâche pour les annotateurs choisis.
- Quels sont précisément les critères qui vont intervenir dans le choix des annotateurs ? Pourquoi choisir des annotateurs naïfs ou experts ? Peut-on encore parler d'annotateurs naïfs s'ils doivent faire l'objet d'une formation et se référer à un guide complexe ? Peut-on alors encore considérer qu'on accède à leurs intuitions de locuteurs du français ? Les lecteurs naïfs sont-ils à même de formuler des intuitions dans des

2. Partenaires : CLLE-ERSS (UMR 5263 Toulouse), IRIT (UMR 5505 Toulouse), GREYC (UMR 6072 Caen). Coordination : M-P. Péry-Woodley (CLLE-ERSS). Voir Péry-Woodley et al (2012) pour une présentation détaillée des objectifs du projet et de la ressource dans son entier, ANNODIS_me n'en constituant qu'une partie.

3. La plate-forme Glozz (`www.glozz.org/`) a été conçue et développée dans le cadre de ce projet.

tâches aussi complexes que l'interprétation des discours ?
- Le processus d'annotation va inévitablement mettre en cause certains aspects du guide. Faut-il le faire évoluer pour l'améliorer au risque de mettre en danger la cohérence des différentes phases de la campagne ?
- Ces questions sont liées aux objectifs de l'annotation, et à l'articulation entre théorisation et démarche empirique : s'agit-il de valider un modèle pré-existant, ou plutôt de produire des données autorisant une approche exploratoire ?

Contributions
L'équipe qui a mené à bien la campagne d'annotation ANNODIS-me était composée de Cécile Fabre, Marie-Paule Péry-Woodley, Josette Rebeyrolle et Ludovic Tanguy (chercheurs) et Mai Ho-Dac (post-doctorante). Une des annotatrices a été chargée du rôle d'annotateur référent, devenant ainsi la mémoire de la campagne d'annotation. Il s'agit de Maud Colléter, grâce à qui ce Carnet de Grammaire s'est enrichi de nombreux éléments du "vécu" de l'annotation. Les post-traitements sur les annotations ont été effectués avec l'aide de Basilio Calderone et Franck Sajous (ITA), de Nikola Tulechki (étudiant) et d'Anny Soubeille (vacataire), qui a également contribué à la conception de ce document.

1 Avant la campagne : préparation des textes et rédaction du guide

1.1 Choix du corpus

Le choix des textes composant la ressource ANNODIS_me reflète les objectifs et les hypothèses de départ de la campagne d'annotation :

1. Nous avons souhaité intégrer d'entrée de jeu l'hypothèse de la variation dans les réalisations discursives en fonction de variations extralinguistiques, ce qui distingue notre ressource des corpus annotés en relations et structures de discours pour l'anglais, composés uniquement de textes issus de la presse écrite. La diversification des textes du corpus n'a pas été envisagée dans l'optique de fournir un corpus de référence des genres écrits du français, mais dans l'idée de constituer des données autorisant des comparaisons intergenres.
2. Notre intérêt pour les structures multi-échelles dans leur interaction avec la structure de document nous a amenés à rechercher des textes longs non-narratifs, dont la cohérence ne peut dépendre de la seule

cohésion thématique, et où différents modes de structuration sont susceptibles d'être sollicités et signalés (structuration en sections et sous-sections, usage de titres et de sous-titres). Ces caractéristiques distinguent à nouveau la ressource ANNODIS_me des corpus annotés existants, qui sont composés de textes brefs et peu structurés (brèves et dépêches).

3. Outre ces deux critères, le choix des textes a également été déterminé par les possibilités de diffusion des corpus annotés.

La composition du corpus est décrite dans le tableau 1 ci-dessous :

identifiant	source (S), genre (G), type majeur (T)	structure de document	nombre de
WIK2	S = *articles Wikipedia entiers* G = art. encyclopédique T = expositif	forte	articles 30 mots 231 000 mots/texte 7 700
LING	S = *CMLF, colloque de linguistique* G = articles de recherche T = expositif	moyenne	articles 25 mots 169 000 mots/texte 6 760
GEOP	S = *IFRI, institut de géopolitique* G = rapports et articles T = argumentatif	moyenne	articles 32 mots 266 000 mots/texte 8 325
			TOTAL : 87 textes, 666 000 mots

Tableau 1 – Composition du corpus ANNODIS_me

1.2 Des documents originaux aux textes annotables

Selon les sources (S), le format des documents d'origine varie : format HTML pour les articles Wikipedia (WIK2), format MSword (.doc) pour les articles de linguistiques (LING) et format PDF pour les rapports de géopolitique (GEOP). Afin de permettre une annotation de ces documents via l'interface Glozz, plusieurs traitements ont été appliqués [4] :

1. Un ensemble de traitements semi-automatiques ont été réalisés afin d'homogénéiser ces différents formats selon un balisage XML suivant la norme TEI-P5. Ce format a permis l'encodage des méta-données et des éléments de la structure du document (niveaux de section, paragraphes, citations, listes formatées, etc.).

4. Ces traitements ont été conçus par Lydia-Mai Ho-Dac et Nikola Tulechki

2. Les documents au format XML ont ensuite été analysés syntaxiquement par l'outil Syntex.
3. Cet étiquetage syntaxique a permis le prémarquage automatique de l'ensemble des traits retenus (prospections, encapsulations, expressions co-référentielles, connecteurs en initiale de phrase, circonstants, séquenceurs, éléments détachés en initiale tel que appositions, constructions détachées, modalisations).
4. Des configurations ponctuationnelles et typodispositionnelles ont également été prémarquées (titres de section, deux points en fin de paragraphe, puces et numérotations en début de segment, schémas ponctuationnels particuliers).
5. Les textes ont ensuite été préparés pour l'annotation via la plate-forme Glozz : séparation du texte (extension *.ac*) et des informations de mise en forme et des traits prémarqués. Ces informations sont alors "débarquées" dans un fichier d'extension *.aa*.

1.3 Annotation exploratoire

Afin de clarifier le modèle, l'interface d'annotation et le guide, l'annotation exploratoire de trois textes a été mise en place. Les annotateurs étaient des membres impliqués dans d'autres volets du projet ANNODIS : annotation des relations rhétoriques et conception de l'interface. La phase exploratoire nous a amené à revoir le prémarquage et à distinguer deux types d'éléments :

- les **traits**, qui correspondent aux éléments prémarqués automatiquement ;
- les **indices**, qui correspondent à des éléments annotés.

Lors de l'annotation exploratoire, plusieurs aspects de la visualisation des traits prémarqués se sont révélés problématiques. Le fait que ces traits apparaissent avec l'étiquette "indice" ("indiceAmorce", "indiceItem", "indiceCIRC", etc.) perturbait l'annotation : il fallait clarifier leur statut d'indices candidats repérés automatiquement, et par conséquent susceptibles de bruits comme de silences (traits prémarqués non pertinents et "vrais" indices non prémarqués). Par ailleurs, le fait que certains traits prémarqués couvraient des segments de taille importante (longues subordonnées temporelles détachées en initiale de phrase par exemple) gênait l'appréhension du texte, si bien que certains annotateurs avaient préféré masquer la coloration du prémarquage et lire le texte sans phase manuelle d'écrémage. L'analyse de ces problèmes a conduit aux modifications suivantes :

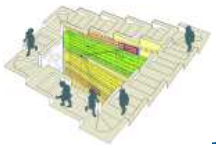
Taille des traits prémarqués
La taille des traits prémarqués a été limitée à un maximum de trois mots et un minimum d'un mot, ce qui a amené à prémarquer des mots aux alentours des éléments ponctuationnels (mot précédent pour les deux points et mot suivant pour les puces, tirets, points-virgules)

Étiquettes des traits prémarqués

- L'étiquetage des traits prémarqués a été revu de manière à désigner non plus leur statut d'indices candidats mais leur catégorie : les Marqueurs d'Intégration Linéaire sont ainsi étiquetés "CIRC" et non "indiceItem" ; de même, les expressions coréférentielles ont fait l'objet d'un étiquetage reflétant une caractérisation linguistique (COREFredeno, COREFproposs, etc.) au lieu de "indiceCT".
- Pour certains traits prémarqués, une étiquette spécifique indique la position en initiale de phrase (zone préverbale) : l'étiquette "CIRCtps_INIT" marque ainsi un circonstanciel de temps (CIRCtps) initial. Cette précision autorise une visualisation sélective des éléments prémarqués en position initiale.

Interface Glozz
Ces modifications ont entraîné une refonte du modèle d'annotation défini pour la plate-forme Glozz (ANNODIS_me.aam). Par ailleurs, l'interface a été modifiée pour permettre la visualisation de traits qui ne font pas partie du modèle (traits prémarqués).

2 Guide d'annotation livré aux annotateurs



ANNODIS : annotation des structures macro

Sommaire

1. Introduction
   1.1 Objets à identifier et caractériser
   1.2 Utilisation du guide d'annotation des structures discursives
2. Structures énumératives (SE)
   2.1 Amorce
   2.2 Items
   2.3 Clôture
3. Chaînes topicales (CT, anciens SUR)
4. Procédure d'annotation
   4.1 Interface d'annotation
   4.2 Etapes de l'annotation dans l'interface
5. Annexe : liste des exemples

1. Introduction

Ce guide s'inscrit dans un projet ANR (Agence Nationale de la Recherche) : le projet ANNODIS, qui vise la constitution d'un corpus de texte français annoté discursivement. Il s'agit de fournir à la communauté scientifique une ressource de qualité pour travailler sur l'organisation discursive en français. En effectuant la phase que nous appelons : "annotation macro", vous participez à ce projet.
Plus précisément, votre tâche va consister à identifier et caractériser des structures discursives qui jouent un rôle à différents niveaux d'organisation. Il s'agit de structures qui ont la capacité d'organiser des portions de textes de taille variable et qui impliquent des phénomènes linguistiques variés (continuité et segmentation thématiques, organisation spatio-temporelle, articulation rhétorique, etc.) et font appel à des modes de signalisation variés : mise en forme matérielle - i.e. découpage en paragraphes et sections, mise en titre, mise en liste, etc. -, réitération lexicale, parallélismes structurels, usage de connecteurs et d'expressions détachées en initiale de phrase, etc.

Les textes que vous allez annoter sont des textes dits expositifs, autrement dit des textes ayant pour but principal d'exposer un thème, un fait, un argumentaire, etc. Ces textes sont relativement longs (il s'agit généralement d'articles d'une dizaine de pages).
Dans ces textes, vous allez annoter deux types de structures discursives : les structures énumératives et les structures ayant une unité référentielle. Les premières sont des structures qui présentent un thème, un fait, un argument, etc. en le découpant en sous-thèmes, événements, arguments. Les secondes forment des zones de textes caractérisées par une unité référentielle. L'annotation de chacune de ces structures est explicitée dans les sections 2 et 3.

1.1. Objets à identifier et caractériser

L'annotation des structures macro se fait en naviguant dans les textes. Cette navigation n'est pas forcément linéaire. Elle peut nécessiter des *zooms* qui permettent à l'annotateur de se représenter le contexte textuel dans lequel il se situe. Par exemple, il peut être utile de savoir s'il y a plusieurs paragraphes dans la section en cours d'annotation, s'il y a des titres de section alentour, de même niveau, de niveau supérieur/inférieur.
L'interface d'annotation que vous allez utiliser facilite ces zooms grâce à un ruban qui représente le texte à annoter vu *de haut* (voir la section 2). De plus, dans le texte, certains éléments sont colorés. Il s'agit de marques de surface repérées automatiquement et pouvant être des indices participant au signalement des structures discursives macro. Ces marques, que l'on appelle des indices prémarqués, constituent des points d'accès au texte. Elles permettent de repérer d'un simple coup d'œil des zones où vous avez des chances de trouver des structures à annoter. Plus précisément, elles permettent d'échapper à une lecture linéaire du texte en rendant possible des stratégies d'*écrémage*.

Remarque 1.

Vos annotations ne concernent pas nécessairement l'ensemble du texte. Une fois l'annotation achevée, certaines zones de textes peuvent rester non annotées.

Remarque 2.

Nous vous demandons d'être attentifs aux structures de très haut niveau. L'identification de structures qui s'étendent sur plusieurs paragraphes étant particulièrement délicate, nous vous demandons de **porter une attention toute particulière aux titres de section et aux éléments qui se trouvent à l'initiale des paragraphes** et d'utiliser la version papier du texte que vous êtes en train d'annoter (version html imprimée que l'on vous aura fournie).

Remarque 3.

Toute structure annotée doit nécessairement couvrir plus d'une phrase (phrase entendue au sens de chaîne de caractères située entre deux signes de ponctuation appartenant à la liste suivante : point de suspension, point d'exclamation, point d'interrogation, point-virgule, puce et numérotation, changement de paragraphe).

1.2. Utilisation du guide d'annotation des structures discursives

Les sections 2 et 3 présentent les objets à annoter. Cette présentation est organisée en quatre volets :

Définition Le volet *définition* fournit une description générale de l'objet à annoter.

Illustration Le volet *illustration* propose plusieurs exemples allant du cas prototypique aux cas marginaux ainsi qu'une liste de liens vers d'autres exemples annotés.

Indices Le volet *indices* recouvre l'ensemble des unités qui signalent des structures ou des éléments de ces structures. Certains éléments linguistiques ont été prémarqués automatiquement. Ce marquage automatique produit ce qu'on appelle des "indices prémarqués". Cependant, d'un côté, tous les indices participant au signalement d'une structure ne sont pas repérés automatiquement, de l'autre, certains indices prémarqués automatiquement ne sont pas pertinents. Au moment où vous ferez votre annotation, vous devrez donc :

- indiquer quels indices non repérés automatiquement participent au **signalement** d'une structure donnée ;
- décider quels indices prémarqués constituent des indices pertinents.

Tests Quand cela est possible, des manipulations sont proposées. Elles doivent vous permettre notamment de confirmer que vous êtes bien face à un indice.

La section 4 explique les procédures d'annotation à suivre ainsi que les modalités d'utilisation de l'interface d'annotation.

Remarque 4.
Tous les exemples contenus dans ce guide proviennent d'un seul et même texte intitulé *Rapport Avicenne* que vous pouvez voir dans son entier dans le cadre droit de cette page [a].

a. La version en ligne proposait une vision simultanée du guide et de ce texte, ce guide en ligne est disponible sur Redac : guide original en ligne

Remarque 5.
Vous pouvez également avoir recours à la liste complète des exemples présents en fin de guide.

Remarque 6.
En parcourant le guide, il se peut que vous rencontriez des termes abrégés ou inconnus. Certains d'entre eux apparaissent soulignés, ce qui signifie qu'un texte info-bulle leur est associé. Ce texte apparaît en plaçant quelques secondes le curseur de la souris sur les mots soulignés, comme, par exemple, l'abréviation SE qui apparaît dans le titre de la section suivante (2, ci-dessous) [a].

a. Cette remarque n'est valable que pour la version en ligne

2. Les structures énumératives SE

La structure énumérative SE est un mode d'organisation fondamental des textes expositifs. Elle consiste à agencer un contenu sous la forme de segments successifs.

Exemple 1 : exemple prototypique de SE.

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 œuvre 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.

voir l'exemple en contexte sur redac

Dans cet exemple, on trouve les trois principaux éléments de la structure énumérative :

- amorce,
- énumération composée d'une série d'items,
- clôture.

SE minimales Une série d'items suffit pour qu'on puisse parler de structure énumérative (amorce et clôture ne sont pas obligatoires).

Annoter une SE Annoter une structure énumérative consiste à identifier son **amorce** (s'il y a amorce), ses **items**, et sa **clôture** (s'il y a clôture). A l'intérieur de ces trois éléments, d'autres objets seront également à repérer : PROSPECT dans l'amorce et ENCAPS dans la clôture. Voir leur définition dans les sections concernées.

2.1. Amorce

Définition

L'**amorce** est un segment qui annonce une énumération. Elle peut comporter

ce que nous appelons un **ÉNUMÉRATHÈME**, 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**. Comme illustré ci-dessous (dans les sections Illustration et Indices), l'ÉNUMÉRATHÈME est souvent signalé par une expression de type PROSPECT, par exemple un groupe nominal composé d'un déterminant numéral et d'un nom.

Illustration

L'amorce apparaît surlignée et l'ÉNUMÉRATHÈME (**avantage**) en italique-gras. Il s'inscrit dans le groupe nominal de type PROSPECT (**trois avantages**). Le second exemple illustre une amorce sans ÉNUMÉRATHÈME.

Exemple 2 : 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 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 état 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.

voir l'exemple en contexte sur redac

Exemple 3 : amorce sans ÉNUMÉRATHÈME.

Il est important de ne pas remettre en cause cette évolution et de poursuivre le rapprochement entre les sociétés à la condition, toutefois, que ce rapprochement ne se fasse pas au détriment de :

- l'expression publique des positions françaises sur le conflit israélo-arabe,
- notre capacité d'action dans la région, fondée certes sur la sécurité d'Israël mais aussi sur le refus de l'occupation et la nécessité d'une évacuation totale des territoires occupés en 1967 et de la création d'un État palestinien indépendant. La persistance depuis bientôt quarante ans de cette occupation est au cœur de l'instabilité dans la région.

voir l'exemple en contexte sur redac

Indices

Les indices d'amorce prémarqués apparaissent colorés en rose dans le texte à annoter.
Les indices d'amorce sont des signes de ponctuation (:) et/ou des éléments lexicaux. Dans le cas d'éléments lexicaux, les expressions prémarquées en rose ont de fortes chances d'englober ou de correspondre à un PROSPECT.
Les titres de section constituent un autre type de marque d'amorce. Un titre de section peut être l'amorce :

- des sections (titres y compris) de niveau inférieur : exemple de SE à travers la titraille,
- de la section titrée : exemple de SE amorcée par le titre de section.

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.

Exemple 4 : test d'identification d'un ÉNUMÉRATHÈME en amorce.
Le dialogue doit donc être modulé avec pragmatisme, c'est-à-dire en fonction du mouvement concerné, une grande variété de formules ***, telles que énumérées ci-dessous,*** s'offrant autour des suivantes :

voir l'exemple en contexte sur redac

2.2. Items

Définition

On distingue deux types d'agencement des items :

- dans les SE dites **verticales**, les items sont séparés par un saut de ligne et signalés par un titre ou un signe placé en début de ligne (numérotation, tiret, puce) ;
- dans les SE dites **horizontales**, les items ne sont pas visuellement signalés par la disposition (pas de saut de ligne, pas de titre, aucun signe typographique).

Un item peut comporter lui-même une structure énumérative, voir l'exemple de ci-dessous (d'autres exemples dans la liste des exemples).

Illustration

Exemple 5 : SE verticale.

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 œuvre de coopérations ; il pourrait impliquer des collectivités locales, voire [...] ;
- un dialogue informel à travers des rencontres et séminaires associant des personnalités d'origine diverse. Le contenu politique serait plus fort mais [...] ;
- 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.

voir l'exemple en contexte sur redac

Dans cet exemple, le dernier item est lui-même composé d'une SE horizontale composée d'une amorce et de 4 items :

Exemple 6 : SE horizontale dans la SE verticale - ex5.

- 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.

Exemple 7 : SE horizontale.

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 état 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.

voir l'exemple en contexte sur redac

Indices

Les indices d'items prémarqués apparaissent colorés en jaune dans le texte à annoter.

Parmi les indices signalant les items, on peut citer :

- les éléments issus de la mise en forme matérielle du texte : titres, tirets, puces
- les signes de ponctuation : virgules, points-virgules
- les marqueurs d'intégration linéaire : *premièrement*, *en second lieu*, *le troisième*, *l'autre*, *tout d'abord*, *ensuite*, *enfin*, etc.
- les circonstants temporels et/ou spatiaux (*en 1976*, *depuis 2009*, *à Toulouse*, *près de Caen...*) (voir ci-dessous un exemple de SE avec circonstant spatial)

Exemple 8 : SE avec série de circonstants spatiaux.

Les relations nouées depuis des siècles dans la région nous valent assurément estime et considération. Elles suscitent aussi des attentes et des déceptions.
Au Maghreb, les gouvernements attendent de nous concours et, pour chacun d'entre eux, soutien exclusif. Les populations sont [...]
Au proche orient, nos prises de parole sont scrutées et analysées dans le détail. Nous y sommes [...]
L'approche est différente dans le Golfe où nous sommes [...]

voir l'exemple en contexte sur redac

- les circonstants notionnels : ce sont des syntagmes prépositionnels détachés en tête de phrase qui précisent un domaine d'activité ou de connaissance (*dans le domaine de la biologie*), une thématique particulière (*concernant la France*), un point de vue (*en général*), ou encore un ensemble de concepts particuliers à un domaine précis (*en hôtellerie homologuée*)
- les parallélismes syntaxiques, comme cette suite d'infinitives dans l'exemple ci-dessous :

Exemple 9 : SE avec parallélisme syntaxique.

Placer l'accent sur l'occupation et la nécessité d'y mettre fin, aurait trois avantages : **repositionner le débat** autour du problème de [...] **découpler l'enjeu de la lutte** contre l'occupation de [...] **désamorcer le débat** qui lie l'opposition à la politique israélienne à la question de l'antisémitisme.

voir l'exemple en contexte sur redac

Tests

Pour repérer un item, vous pouvez tenter d'insérer ***tout d'abord***, ***ensuite*** ou ***enfin*** au début ou à l'intérieur de l'item.

Exemple 10 : test d'identification des items d'une SE.

Placer l'accent sur l'occupation et la nécessité d'y mettre fin, aurait trois avantages : ***tout d'abord*** 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 ; ***ensuite*** 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 État viable et à l'intérieur de frontières sûres ; ***enfin*** désamorcer le débat qui lie l'opposition à la politique israélienne à la question de l'antisémitisme.

2.3. Clôture

Définition

La clôture est un segment qui conclut la structure énumérative. Comme l'amorce, la clôture peut comporter l'expression de l'ÉNUMÉRATHÈME. Il s'inscrit généralement dans un syntagme nominal appelé encapsulation (ENCAPS), correspondant auPROSPECT de l'amorce.
Il peut arriver qu'il y ait une ENCAPS sans pour autant qu'il y ait un segment de clôture de l'énumération, à proprement parler. Dans ce cas, seule l'ENCAPS sera annotée (voir ci-dessous l'exemple de SE avec ENCAPS (*trois directions*, où *directions* est l'ÉNUMÉRATHÈME) mais sans clôture).

Illustration

La clôture apparaît surlignée et l'ÉNUMÉRATHÈME (**scénarios**) en **italique-gras**. Il s'inscrit dans le groupe nominal de type ENCAPS **ces scénarios**.

Exemple 11 : SE avec clôture et ÉNUMÉRATHÈME.

3. Perspectives : quatre scénarios

Tout exercice d'anticipation sur une zone aussi sensible que le moyen orient est à l'évidence très risqué. Il ne peut être abordé qu'avec prudence et humilité. Sur la base de la situation particulièrement préoccupante qui prévaut au moyen orient et des tendances actuelles, plusieurs scénarios peuvent être théoriquement envisagés.

3.1 La Pax Americana

...

3.2 L' ordre islamiste

...

3.3 Le chaos

...

3.4 Un processus de dégradation lent et modulé

...

Dans les faits, il est probable qu'aucun de ***ces scénarios*** ne se réalisera, même s'ils ont leur propre cohérence. L'hypothèse la plus probable sera sans doute composite, ...

voir l'exemple en contexte sur redac

Exemple 12 : SE avec ÉNUMÉRATHÈME mais sans clôture (ni amorce).

Depuis septembre 2004, la France a pris la direction d'un mouvement diplomatique qui a conduit à l'adoption par le Conseil de sécurité de la résolution 1559 appelant au retrait des forces syriennes du Liban. Après l'assassinat de l'ancien Premier Ministre Rafic Hariri, elle a pris clairement position pour la coalition des forces politiques du 14 mars, un bloc dont le principal ciment et l'objectif commun étaient de mettre fin à l'influence syrienne au Liban. Au lendemain de la guerre d'Israël contre le Hezbollah à l'été 2006, elle a su mobiliser un large soutien international pour la mise en place d'une FINUL renforcée et pour la reconstruction du pays dévasté lors de la conférence de Paris 3. Ces trois directions , engagées au cours des trois dernières années, méritent un examen critique, au niveau des objectifs d'une part, du cadre dans lequel la France déploie son activité et des partenaires qu'elle choisit d'autre part, pour envisager les options politiques à venir.

voir l'exemple en contexte sur redac

Indices

Les indices de clôture prémarqués apparaissent colorés en orange dans le texte à annoter.
La clôture peut être annoncée par des syntagmes comme *en conclusion*, *en résumé*, *pour conclure* et/ou signalée par des ENCAPS (i.e. encapsulation) qui condensent en une expression référentielle les différents items énumérés.
Les ENCAPS ont généralement la forme d'un syntagme nominal au pluriel, dont le déterminant est souvent un démonstratif, accompagné ou non d'un numéral. Par exemple, *ces sénarios*, *ces trois options*, *ces différents points*.
La tête lexicale de ce syntagme indique le type des éléments énumérés, c'est l'ÉNUMÉRATHÈME.
Ces éléments sont particulièrement utiles pour identifier une SE. Notamment, une ENCAPS peut suggérer qu'en amont se trouve une énumération des référents qu'elle condense. Naturellement, tout syntagme nominal démonstratif n'est pas nécessairement une encapsulation. C'est à vous de vérifier si la relation suggérée est ou non motivée.

Tests

Pour repérer l'ÉNUMÉRATHÈME d'une clôture, vous pouvez tenter d'insérer ***tel(le)s que énuméré(e)s ci-dessus*** et/ou ***tel(le)(s) que décrit(e)(s) ci-dessus***, immédiatement après l'expression présumée en être un. La possibilité d'une telle insertion en confirme la présence. Il faut noter en revanche que son impossibilité ne peut pas l'infirmer.

Exemple 13 : test d'identification de l'ÉNUMÉRATHÈME en clôture.
De telles évolutions ***, telles que nous venons de les énumérer ci-dessus,*** ne sont pas une fatalité. Pour arrêter l'engrenage de violences [...]

voir l'exemple en contexte sur redac

3. Les Segments ayant une Unité Référentielle CT

Remarque 7 : Changement de label : des SUR aux CT.
Les Segments ayant une Unité Référentielle, abrégés SUR lors de l'annotation, ont été renommés **Chaînes Topicales – CT** après annotation. Ce terme est celui qui est désormais utilisé dans la ressource, dans la documentation, et dans les publications. Voir Guide Si. Par soucis de cohérence, nous avons remplacé le terme SUR par CT dans cette version du guide original.

Définition

Une CT est un segment qui se caractérise par le fait que la majorité des propositions qui le composent ont pour objet (parlent de, sont à propos de, apportent des informations au sujet de) **un seul et même référent**. Selon cette définition, l'expression de ce référent commun doit passer nécessairement par le sujet grammatical.
Veuillez noter tout de même qu'une CT n'est pas nécessairement composé uniquement de propositions portant sur le référent qui fait l'unité du segment. En effet, des commentaires ou illustrations, par exemple, peuvent être insérés à l'intérieur d'une CT.

Annoter une CT

Annoter une CT consiste à identifier ce qui fait son **unité référentielle** ainsi que les **indices** vous ayant permis de le repérer.

Illustration

Exemple 14 : CT autour du topique "notre politique/position".
Les indices des CT apparaissent en vert

Notre position doit prendre en considération la pérennité du régime islamique : malgré ses échecs économique et politique et ses tensions internes, on ne voit pas comment le régime des ayatollahs pourrait s'écrouler dans un avenir prévisible. Elle doit également tenir compte du fait que, plus par un effet d'aubaine que par une volonté expansionniste, l'Iran est devenu un acteur incontournable au moyen orient : les états-Unis, en débarrassant l'Iran des ses deux principaux ennemis, les Talibans et Saddam Hussein, et Israël, en déclenchant imprudemment une guerre contre le Hezbollah, ont renforcé sa capacité d'influence et de nuisance. Aussi notre politique doit-elle se garder de s'associer à toute tentative de "regime change" et doit-elle considérer que l'Iran est une puissance régionale avec laquelle il faut compter et dialoguer. Elle doit également tenir compte du fait que, du côté américain, une intervention militaire est une option qui est non seulement "sur la table", mais aussi sérieusement envisagée. Il serait difficile au Président Bush qui, à maintes reprises a dénoncé le caractère inacceptable des ambitions nucléaires de l'Iran de se déjuger et de ne rien faire, d'autant plus qu'il est soumis à la pression d'Israël qui qualifie la menace iranienne d'existentielle. Une telle intervention, hasardeuse sur le plan technique, ne pourrait avoir que des effets désastreux au moyen orient comme dans l'ensemble du monde musulman.

voir l'exemple en contexte sur redac

La section 3.2. fournit une autre illustration d'une CT ayant pour Unité Référentielle *les partis islamistes*.

Indices

Les indices de CT prémarqués apparaissent colorés en vert dans le texte à annoter.
Dans les textes à annoter, les expressions coréférentielles en position sujet sont toutes prémarquées en tant qu'indice-candidats de ce segment. Les expressions coréférentielles sujets peuvent être :

- des pronoms personnels de 3e personne
- des pronoms démonstratifs
- des syntagmes nominaux possessifs
- des syntagmes démonstratifs
- des syntagmes dont la tête lexicale réitère un nom déjà mentionnée dans la section en cours
- des reprises lexicales des noms présents dans le titre de la section en cours.

Des circonstants notionnels d'un type particulier (*Quant à Euronews* – voir dans le texte ; *S'agissant de la Russie* – voir dans le texte) peuvent également indiquer le début ou la continuation d'une CT.

Tests

Pour vous assurer que les expressions sont bien coréférentielles, vous pouvez tenter de les substituer par l'expression référentielle complète. L'impossibilité d'une telle substitution amène à conclure à la non coréférentialité.

4. Procédure d'annotation

L'objectif de l'annotation que vous allez réaliser est double. Il s'agit :

1. d'identifier et délimiter les **éléments qui composent une structure (SE ou CT)**, voir section 4.2.3 ;
2. d'indiquer les indices qui assurent la signalisation de la structure et/ou de ses composants, voir section 4.2.4 ;
3. d'associer ces éléments et ces indices à une même structure (section 4.2.5) ;

L'interface utilise deux types d'objets principaux : les schémas et les unités.

Les structures (SE et CT) forment ce qu'on appelle des **SCHÉMAS** *i.e.* des objets complexes composés d'unités pouvant entretenir entre elles certaines relations

Ces schémas sont constitués de deux types d'**UNITÉS** : les composants de la structure et les indices qui signalent cette structure et/ou ses composants.

Les éléments **UNITÉS** composant les structures sont les suivants :

- **pour une SE** :
  - l'amorce,
  - les items,
  - la clôture,
  - le(s) énumérathème(s)
- **pour une CT,** une seule unité est à délimiter : l'UR (unité référentielle) dont l'unique différence avec la CT est d'être une unité et non un schéma. En l'absence de délimitation d'une quelconque unité, le schéma CT n'aurait aucune substance.

Les **UNITÉS** indices signalant les structures ou leurs composants correspondent soit aux indices prémarqués automatiquement soit à toute autre forme identifiée comme indice par l'annotateur.

4.1. Interface d'annotation

Toutes les procédures d'annotation se font avec l'interface d'annotation dans laquelle nous distinguons 7 éléments :



4.2. Etapes de l'annotation dans l'interface

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 une structure discursive en délimitant les unités qui la composent (SE/CT)
4.2.4 Valider, supprimer, créer les indices
4.2.5 Regrouper les éléments composant une structure discursive (SE/CT)
4.2.6 Modifier et supprimer une annotation

4.2.7 Enregistrer les annotations
4.2.8 Gestion de l'incertitude

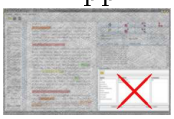
4.2.1. Charger les textes à annoter

Une fois l'interface ouverte, voici les procédures à effectuer pour charger les fichiers nécessaires à l'annotation :

1. Charger le document à annoter en cliquant sur le bouton (Open corpus) situé dans la barre d'outils. Deux éléments doivent être chargés :
   - le texte (fichier avec extension **.ac** comme ***a**nnodis **c**orpus*), normalement situé dans le dossier /data/corpus.
   - ses annotations (fichier avec extension **.aa** comme ***a**nnodis **a**nnotation*), normalement situées dans le dossier /data/annotations.

   Remarque 8.
   Vous pouvez également charger un fichier contenant des annotations que vous avez réalisées et sauvegardées (voir la section 4.2.7.).

   Bien entendu, les deux éléments doivent porter le même nom, hormis leur extension (*e.g.* avicenne_TEIP5.ac et avicenne_TEIP5.aa).
2. Charger la feuille de style qui permet de colorer dans le ruban et la zone texte les indices prémarqués et les annotations associées au document. Pour ce faire, cliquer sur le bouton (Style editor) situé dans la barre d'outil, puis sur le bouton (Open style) dans la fenêtre concernée. Le fichier de base pour le marquage macro se trouve dans le fichier **data/styles/macro.as**
3. Charger le modèle d'annotation en cliquant sur le bouton (LAM) dans la zone modèle (à droite de la zone texte). Pour l'annotation macro, charger le modèle **data/annotationModels/macro.aam**. Les différents éléments du modèle apparaissent alors dans la zone modèle.



4.2.2. Distinguer plusieurs étapes d'annotation et jouer avec les styles

Il est fortement recommandé de distinguer trois étapes d'annotation :

- Annotation des SE de plus haut niveau en masquant les indices prémarqués qui n'apparaissent pas en position initiale, ainsi que les indices

prémarqués de CT, et en effectuant une lecture complète de la titraille (lecture en 'sautant' de titre de section en titre de section). Ces SE de haut niveau peuvent également être amorcées par une petite SE en fin de section dont les différents items sont repris par les titres de section.
- Annotation des SE de niveau paragraphique. Pour cette étape, il peut être nécessaire de faire apparaître tous les indices (même les expressions co-référentielles). À vous de jouer !
- Annotation des CT en masquant les indices prémarquées non concernés (circonstants spatiaux, MIL, etc.).

À la fin de chaque étape, vous devrez vérifier qu'il ne reste pas de zones inexplorées présentant une certaine concentration d'indices prémarqués. Pour ce faire, le ruban s'avère vraiment pratique, parce qu'il donne une vision générale du texte et de ses annotations.

Pour chaque étape, vous pouvez décider de masquer vos annotations précédentes en masquant le style concerné (pour masquer les CT lors de l'annotation des SE et inversement) ou en masquant au cas par cas les schémas annotés via l'outil d'exploration **'Annotation as text'** (voir explication en remarque 15)

4.2.3. Repérer une structure discursive en délimitant les unités qui la compose (SE/ CT)

Maintenant que le texte est ouvert dans l'interface d'annotation, voici comment procéder pour l'annoter, c'est-à-dire pour délimiter et caractériser les SE et les CT en commençant par délimiter les éléments qui les composent.

1. Dans la zone édition, sélectionnez le bouton  (Create a new simple Unit) qui permet de poser les bornes de début et de fin des unités à annoter : amorce, item, clôture, énumérathème, indices.

2. Dans le ruban, cherchez une zone présentant des indices prémarqués de SE ou de CT.

3. En cliquant sur la zone sélectionnée, le texte correspondant s'affiche dans la zone texte.

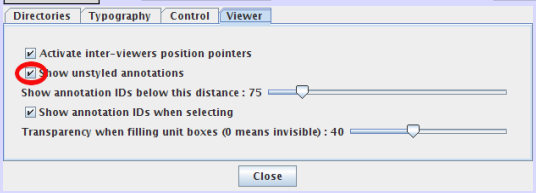


4. En vous appuyant sur les indices prémarqués (colorés selon le jeu de style défini dans la fenêtre **Style editor**, voir 2 ci-dessous), vous devez repérer si la zone contient ou non une SE (ou l'un de ses éléments) ou une CT.
   La **délimitation des éléments des structures** peut se faire de deux manières distinctes :
   - **soit en deux temps** :
     (a) positionnez d'abord votre souris sur le début de l'unité et cliquez pour ancrer la borne start
     

     .
     (b) positionnez ensuite votre souris à la fin de l'unité et cliquez pour ancrer la borne end
     

     .
   - **soit en un seul mouvement** : positionnez votre souris sur le début de l'unité, cliquez et maintenez le clic pour tracer la délimitation de l'unité (un cadre en pointillé apparaît). Glissez le curseur jusqu'à la fin de l'unité et alors seulement lâchez le clic.

Remarque 9.
Si **les délimitations n'apparaissent pas à l'écran** cela signifie qu'aucun style n'est associé à l'unité. Vous devez alors vérifiez que l'objet que vous annotez (dont le nom est inscrit et sélectionné dans la zone modèle) a bien un style associé dans la fenêtre **Style editor**. Si vous ne voulez pas associer de style à l'objet en question mais uniquement le visualiser, choisissez alors d'afficher les annotations sans style (*unstyled annotations*) : Barre d'outils : Options > Préférences puis sur l'onglet Viewer :



5. Toute unité doit être associée à un type (amorce, item, clôture, énumérathème, indice, UR). Par défaut, toute nouvelle unité est associée au type de l'unité précédemment annotée. Si aucune unité n'a encore été délimitée, la nouvelle unité sera associée au type *u_ default*.

Pour associer un type différent à une nouvelle unité, assurez-vous que celle-ci est bien sélectionnée et cliquez sur le type adéquat dans le modèle d'annotation affiché dans la zone LAM.

Liste et Codification couleur des indices prémarqués

Vous trouverez ci-dessous la table présentant la liste des indices prémarqués automatiquement. Tous les indices prémarqués sont considérés comme des unités par l'interface, de la même manière que les unités que vous avez délimitées lors de votre annotation.

Chaque indice est caractérisé par un type, associé à un jeu de couleur par le style *macro.as*. La table ci-dessous liste chaque indice en indiquant son type (étiquette apparaissant dans l'interface), sa couleur dans macro.as et une définition. Vous pouvez, à tout moment et selon votre convenance,

Tableau 2 – Code couleur par défaut des éléments prémarqués

PONCT	pattern ponctuationnel d'amorce (plus le mot qui précède pour une meilleure visualisation)
PROSPECT	prospection
PONCTitem	pattern ponctuationnel d'item (plus le mot qui suit pour une meilleure visualisation)
MIL(_init)	marqueur d'intégration linéaire (en initiale de phrase)
CIRCnot(_init)	circonstant notionnel (en initiale de phrase)
CIRCspa(_init)	circonstant spatial (en initiale de phrase)
CIRCtps(_init)	circonstant temporel (en initiale de phrase)
ENCAPS	encapsulation
COREFproposs	forme pronominale ou possessive en position sujet
COREFredeno	SN sujet dont la tête reprend un nom déjà présent dans la section
COREFdemo	SN démonstratif en position sujet
Rtitre	reprise nominale d'un élément du titre en position sujet
HEADING	titre de section
CONNECT	connecteur simple en initiale de phrase

modifier le jeu de couleur ou choisir de ne pas colorer tel ou tel type d'indice. Pour ce faire, ouvrez le style macro.as et cliquez sur l'indice dont la couleur est à modifier, ou cochez la case *Hide* pour ne plus voir son surlignement.

Remarque 10.

ATTENTION !! Il se peut que **plusieurs fenêtres *Style Editor* soient ouvertes simultanément** (si vous avez à chaque fois cliqué sur le bouton ). Du coup, vos modifications peuvent ne pas prendre effet. Pour être sûr de modifier le 'bon' jeu de style, vérifiez dans la barre des tâches qu'un seul 

apparaît. Si plusieurs fenêtres sont ouvertes, fermez les toutes pour n'en laisser qu'une sur laquelle vous ferez vos modifications (que vous pourrez sauvegarder en cliquant, dans la fenêtre **Style Editor**, sur le bouton ).

4.2.4. Valider, supprimer, créer les indices

Cette phase de l'annotation consiste à associer à chaque unité annotée les indices qui ont servi à la repérer. Lors de cette phase, vous serez amenés à effectuer trois types d'opérations : valider, modifier ou créer des indices.

- **validation d'un indice prémarqué :** lorsque l'indice (coloré) est bien un indice sur lequel vous vous êtes appuyé pour identifier un objet, vous devez rattacher cet indice à la structure qu'il signale en l'incluant dans le schéma correspondant (voir section suivante 4.2.5). Si les éléments colorés incluent trop d'éléments ou n'incluent pas l'ensemble des éléments pertinents, il vous faut **redélimiter** l'indice (voir la section 4.2.6).
- **création d'un indice :** tout indice vous paraissant significatif doit être associé au schéma qu'il signale. Pour cela il faut créer une nouvelle unité de type **indice** (voir section 4.2.3). Une fois cette unité indice créée, il faut renseigner la nature de cet indice dans la zone modèle de l'interface (exemple : nom propre répété, changement de temps verbal, parallélisme, etc.) Ensuite, il reste à rattacher cette unité-indice au schéma concerné (voir section suivante 4.2.5).
- **suppression d'un indice :** les indices prémarqués jugés non pertinents sont simplement laissés tels quels. Les indices créés par l'annotateur mais jugés au final non pertinents devront être supprimés (voir section 4.2.6).

4.2.5. Regrouper les éléments composant une même structure discursive (SE/ CT)

Une fois que vous avez annoté les **UNITÉS**  (section 4.2.3) qui composent une SE ou une CT, vous devez les regrouper en créant un **SCHÉMA** .

Pour regrouper chaque unité d'une structure dans un même schéma :

1. Dans la zone édition, cliquez sur le bouton  (Schemas)

2. Créez un nouveau schéma en cliquant sur  (Create a new Schemas)

Remarque 11.
Comme pour les unités, tout schéma doit être associé à un type (selon les cas, choisissez SE ou CT). Par défaut, tout nouveau schéma est associé au type du schéma précédemment annoté. Si aucun schéma n'a encore été créé, le nouveau schéma sera associé au type *s_ default*.
Pour associer un type différent au nouveau schéma, assurez-vous que celui-ci soit bien sélectionné et cliquez sur le type adéquat dans le modèle d'annotation affiché au niveau de la zone LAM

3. Regroupez les différentes unités d'un même schéma en cliquant sur  (Add Unit to schema), puis sur toutes les unités concernées (amorce, items, clôture, énumérathème(s), UR, indices).

Lors de l'identification des différents éléments d'une SE, vous pouvez vous retrouver en présence de **structures enchâssées**, c'est-à-dire de SE dans une SE (voir les différents exemples d'enchâssement). Face à de telles situations vous pouvez :

- **soit** laisser en attente l'annotation de la première SE pour annoter la nouvelle SE et alors insérer une **glue note**[5]  qui rappellera la présence d'une structure dont l'annotation est inachevée
- **soit** indiquer la présence d'une autre SE en insérant une **glue note**  sans en faire davantage afin de continuer l'annotation de la première SE et effectuer dans un second temps l'annotation de la nouvelle SE.

Remarque 12 : Les glue notes.
À tout moment, il est possible d'associer un commentaire à une annotation ou à n'importe quelle position dans le texte par le biais de **glue note** .

Pour cela, cliquer sur l'icône  et remplissez le cadre jaune qui s'affiche. Vous pouvez ensuite éditer ces **glue notes** ou les supprimer, une par une ou toutes ensemble.

4.2.6. Modifier et supprimer une annotation

– Modifier la délimitation d'une unité

5. une glue note permet d'associer un commentaire à une position dans le texte

– Modifier le type d'une unité
– Supprimer une unité
– Modifier la composition d'un schéma
– Supprimer un schéma

À tout moment il est possible modifier ou supprimer une annotation en choisissant le mode adéquat dans la zone édition.



Modifier la délimitation d'une unité

1. cliquez sur  (Edit/Delete Units) ;
2. **sélectionnez l'unité** à modifier en cliquant dessus. Les lignes délimitant l'unité se changent en pointillés rouges et deux petits ronds apparaissent aux bornes initiale et finale.

Remarque 13.

Lorsque le curseur de la souris passe sur une unité annotée, celle-ci change de couleur. En présence d'**unités superposées**, toutes les unités concernées changent de couleur. Pour sélectionner une unité lorsque les unités superposées ont les mêmes limites, il faut cliquer plusieurs fois pour sélectionner l'unité désirée.

3. **positionnez le curseur** sur les ronds de borne initiale et/ou finale et déplacer la ou les borne(s) afin d'obtenir la délimitation correcte.

Remarque 14.

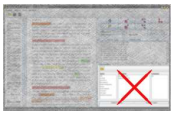
Il peut s'avérer rapidement difficile de distinguer les différents niveaux de structuration. Pour cela, l'interface propose un outil appelé **Depth Selector** qui permet de jouer sur les niveaux d'annotation visibles. Pour activer cette fonction, cliquez dans la barre d'outils sur Tools puis Depth Selector. La boîte de dialogue suivante apparaît alors dans la zone droite de l'interface :



Il suffit ensuite de **manipuler le curseur** pour faire varier l'affichage des différents niveaux d'annotation.

Modifier le type d'une unité

1. cliquez sur le bouton  (Edit/Delete Units) ;
2. sélectionnez l'unité à modifier en cliquant dessus (voir l'item sélectionner l'unité ci-dessus) ;
3. changez l'annotation au niveau de la zone modèle (l'annotation actuelle apparaît surlignée)



Supprimer une unité et son annotation

1. cliquez sur le bouton  (Edit/Delete Units) ;
2. sélectionnez l'objet à modifier en cliquant dessus (voir l'item sélectionner l'objet ci-dessus) ;
3. appuyez sur la touche Suppr du clavier, l'unité et son annotation sont supprimées.

Modifier la composition d'un schéma

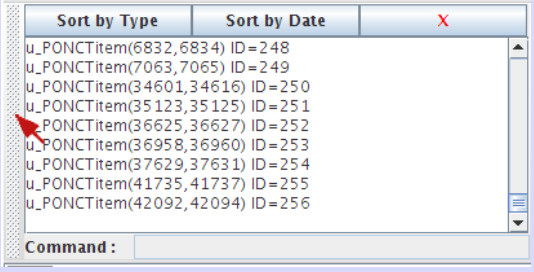
1. Cliquez sur  (Schema). Une boîte de dialogue apparaît dans laquelle vous cliquez sur .
2. **Sélectionnez le schéma** à modifier en cliquant dessus. Les lignes encadrant le schéma sélectionné se changent en pointillés rouges.
3. **Pour enlever une unité** :
   (a) Cliquez sur  dans la boîte à outils dédiée à l'édition des schémas
   (b) Cliquez dans la zone texte sur l'unité à enlever

3' **Pour ajouter une unité** :
   (a) Cliquez sur  dans la boîte de dialogue dédiée à l'édition des schémas
   (b) Cliquez dans la zone texte sur l'unité à ajouter

Remarque 15.

Vous pouvez vérifier votre action en observant ce qui se passe dans la boîte **Annotation as Text** (pour activer cette fonction, cliquez dans la barre d'outils sur Tools puis Annotation as Text). Vous accèderez ainsi à une vision listée de toutes les annotations (schémas, unités, relations). Cette fenêtre vous permet également de naviguer d'annotation en annotation dans la zone texte par un simple clic sur l'annotation désirée.

REMARQUE : Cette fenêtre peut s'avérer gênante parce qu'elle pousse vers le bas la fenêtre du modèle d'annotation par exemple. Vous pouvez alors déplacer les différentes boîtes à outils en effectuant un clic maintenu sur la barre verticale gauche de la boîte et en la déplaçant où vous le souhaitez.



Pour fermer cette boîte, cliquez sur la croix en haut à droite.

Supprimer un schéma

1. cliquez sur le bouton (Edit/Delete Units) ;
2. sélectionnez le schéma à modifier en cliquant dessus (il peut parfois être préférable d'utiliser la boîte à outils **Annotation as Text**, voir ici) ;
3. appuyez sur la touche Suppr du clavier, le schéma et son annotation sont supprimés.

Remarque 16.

Supprimer un schéma n'entraîne aucunement la suppression de ses unités composantes.

4.2.7. Enregistrer les annotations

Pour enregistrer vos annotations, cliquez sur le bouton (Save Annotations). Nommez le fichier d'annotation selon le format suivant :
NomTexte_NomAnnotateur_JJmoisAA.aa
(ex : avicenne_hodac_01janvier09.aa)

4.2.8 Gestion de l'incertitude

Toute annotation dont vous n'êtes pas complètement convaincu peut-être associée à ce caractère incertain. Pour cela, dans la zone modèle, chaque unité est associée par défaut à un degré d'incertitude 0. Pour noter votre sentiment d'incertitude, il vous suffit d'associer la valeur 1.

Liste des exemples

Cette liste contient tous les liens vers les exemples utilisés pour illustrer le guide d'annotation pour l'analyse macro.

- SE complète avec mise en forme matérielle
- SE locale avec amorce et énumérathème
- SE mise en forme avec amorce sans énumérathème
- SE globale sur plusieurs sections
- SE globale à travers la titraille
- SE amorcée par le titre de section
- enchâssement de SE verticales
- enchâssement de SE horizontales (cet extrait fait lui-même partie d'un item d'énumération marquée et indexée par les titres de section)

Exemple 15.

Cette absence de réflexion stratégique sur une position proprement française ou sur une action visant à influencer l'Europe pour en définir une ont eu pour conséquence que la France n'a pas été en mesure de formuler des vues claires sur **les deux développements majeurs des trois dernières années** : la politique unilatéraliste préconisée par le gouvernement de Sharon et, deux ans plus tard, la victoire du Hamas aux élections législatives palestiniennes de janvier 2006.

Devant la première, elle s'est laissé entraîner vers une vision selon laquelle l'unilatéralisme pouvait constituer une approche alternative à la négociation. De même a-t-elle été prise au dépourvu par l'élection du Hamas et a fait le choix de se ranger sur une position européenne qui s'est vite avérée intenable. La politique française a été en somme largement réactive. La diplomatie n'a pas fait usage de la panoplie de moyens disponibles.

- enchâssement de SE mixtes
- SE avec circonstants spatiaux
- SE avec circonstants temporels
- SE sans amorce avec une clôture et énumérathème
- SE sans amorce ni clôture

- SE au premier item non marqué (sans amorce ni clôture)
- CT
- CT avec phrase de commentaire ou d'illustration insérée

Auteurs : Lydia-Mai Ho-Dac, Josette Rebeyrolle, Cécile Fabre, Marie-Paule Péry-Woodley (version : 10 juillet 2009, adaptée pour cette publication)

Ce guide est disponible sur le site ANNODIS : http ://redac.univ-tlse2.fr/corpus/annodis/

Il est disponible sous licence Creative Commons By-NC-SA 3.0 (Paternité, usage non commercial, partage à l'identique). Merci de la lire attentivement.



3 Déroulement de la campagne d'annotation

Phase d'apprentissage
Trois étudiants en Licence et Master Sciences du langage ont été recrutés pour annoter les textes du corpus. À l'un d'entre eux, l'annotateur référent, a été présenté le projet ainsi que la tâche d'annotation en utilisant le guide comme support. Une fois les consignes comprises, l'annotateur référent a été formé à l'utilisation de la plateforme d'annotation Glozz. L'annotateur référent a ensuite transmis ses connaissances aux deux autres annotateurs et tous trois ont annoté un même texte afin de prendre en main la plateforme d'annotation. Cette phase d'annotation a soulevé un certain nombre de questions auxquelles le guide n'apportait pas de réponses. Une réunion entre l'équipe de recherche et les annotateurs a permis de répondre à ces interrogations.

Phase A : mise au point
Pour la Phase A, comme pour la phase d'apprentissage, les annotateurs ont annoté tous les trois les mêmes textes et étaient autorisés à dialoguer ainsi qu'à confronter leurs points de vue lorsqu'une annotation posait une difficulté. Lors de cette phase, un texte extrait de chaque sous corpus a été annoté (soit trois textes) afin de confronter les annotateurs à une grande variété de structures avant qu'ils se lancent dans les annotations individuelles. Après avoir répondu aux nouvelles questions soulevées par ces annotations, l'équipe de recherche a modifié et complété le guide en tenant compte des remarques.

Phase B : répétition générale
Les annotateurs ont, comme pour les phases précédentes, annoté tous les trois les mêmes textes (deux pour chaque sous corpus), mais il leur a été demandé de ne pas communiquer pendant l'annotation. Les six textes ainsi annotés ont permis le calcul d'un accord inter-annotateur portant sur l'identification des deux types de structures (SE et CT).

Phase C : on tourne !
Les annotateurs se sont répartis équitablement les textes du corpus et les ont annotés individuellement.

Lors de cette phase le temps consacré à l'annotation des SE a été chronométré. Le temps d'annotation d'un texte entier va de 10 minutes pour un texte de 7 pages contenant 4 SE (ling_mangenot) à 2h15 pour un texte de 16 pages contenant 10 SE (geop_9), avec des valeurs intermédiaires autour

de 30 minutes pour des articles de wikipédia (plus simples à annoter) longs de 20 pages et contenant 15 SE (e.g. wik2_titanic).

4 Accord inter-annotateur et production d'une version Gold

4.1 Accord inter-annotateur

L'accord inter-annotateur a été mesuré sur le principe de la F-mesure[6] correspondant à la moyenne harmonique de mesures de rappel et de précision obtenues par comparaison de l'annotateur A et de l'annotateur B pour chaque texte annoté. La formule est résumée ci-dessous :

$$F = \frac{2 \times R \times P}{R + P}$$

simplifiée de la façon suivante :

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

Le résultat est compris entre 0 et 1, avec 1 pour un accord parfait et 0 pour un désaccord total.

La moyenne de toutes les F-mesures calculées fournit une appréciation d'ensemble de l'accord inter-annotateur (0,67 pour les SE et 0,65 pour les CT).

4.1.1 Calcul des SE annotées

On considère qu'il y a accord si :
- les deux structures se trouvent au même endroit dans le texte ;
  ET
- les items repérés sont les mêmes.

Dans les cas nécessitant un arbitrage, les critères suivants ont été pris en compte :
- nombre d'items ;
- présence/absence d'amorce et/ou de clôture ;
- empan des différents composants.

Cette confrontation des annotations a permis d'observer les tendances suivantes :

6. Étant donné qu'il ne s'agissait pas d'une simple tâche de classification, le Kappa de Cohen était inadapté.

	geop_2	geop_8	wik2_i	wik2_h	ling_l	ling_p	Moyenne
A2/A1	0,18	0,67	0,8	0,85	0,88	0,76	0,69
A2/A3	0,2	0,5	0,84	0,84	0,69	0,64	0,62
A3/A1	0,67	0,41	0,84	0,87	0,7	0,76	0,71
Moyenne	0,35	0,53	0,83	0,85	0,76	0,72	**0,67**

Tableau 3 – F-mesure pour les SE annotées par les trois annotateurs (A1, A2, A3) durant les phases A et B

- moins il y a de structures dans un texte, moins l'accord est bon ;
- les structures de faible empan sont difficiles à arbitrer ;
- certains types de structures sont plus propices aux désaccords : argumentations, énumération chronologique/temporelle.

4.1.2 Calcul entre CT annotées

On considère qu'il y a accord si :
- les deux structures recouvrent une même zone de texte (même partiellement) ;
  ET
- les indices repérés sont les mêmes.

	geop_2	geop_8	wik2_i	wik2_h	ling_l	ling_p	Moyenne
A2/A1	0,61	0,88	0,63	0,67	0,73	0,55	0,68
A2/A3	0,56	0,95	0,54	0,55	0,72	0,67	0,67
A3/A1	0,42	0,83	0,54	0,71	0,52	0,55	0,6
Moyenne	0,53	0,89	0,57	0,64	0,66	0,59	**0,65**

Tableau 4 – F-mesure pour les CT annotées par les trois annotateurs (A1, A2, A3) durant les phases A et B

Cette confrontation des annotations a permis d'observer les tendances suivantes : un meilleur taux d'accord est observé entre les CT dont la majorité des indices est un pronom personnel et entre les CT comprenant des titres de sections.

4.2 Méthode d'arbitrage et constitution du "gold"

Afin de pouvoir inclure dans la ressource diffusée les textes multi-annotés (phases A et B), l'annotateur référent a fourni une version arbitrée ren-

dant compte des trois annotations effectuées. Cette sous-section présente les grandes lignes de la méthode d'arbitrage. Comme dans tout arbitrage, les décisions sont prises au cas par cas. Des exemples d'arbitrage sont donnés en annexe.

Arbitrage des structures

- Toute structure annotée par au moins deux annotateurs sur les trois de manière strictement identique est ajoutée au gold.
- Les structures repérées par un seul annotateur ont été arbitrées au cas par cas [7].
- Dans le cas où une même partie de texte est annotée par plusieurs annotateurs de manières différentes, l'annotateur référent décide soit de ne garder que la structure la plus pertinente (nombre d'indices, cohérence avec les structures environnantes) ; soit, quand cela est possible, de fusionner les annotations, créant ainsi une nouvelle structure.

Arbitrage des amorces, items et clôtures

- Pour une même structure (i.e. items identiques), les annotateurs ont pu relier ou non une amorce ou une clôture. L'amorce ou clôture est conservée dans le gold si : deux des annotateurs l'ont repérée ou si un seul des annotateurs l'a repérée mais qu'elle est signalée par des indices convaincants.
- En cas de litige sur le nombre ou la taille des items, la majorité l'emporte. S'il n'y a pas de majorité, l'annotateur référent tranche.

Arbitrage des indices

- Les indices ayant été repérés par au moins deux annotateurs sont reportés systématiquement.
- Les indices ayant été repérés par un seul annotateur font l'objet d'un arbitrage par l'annotateur référent.
- Dans le cas où aucun annotateur n'a précisé la nature de l'indice, l'annotateur référent l'indique.

7. S'il s'avère que celui-ci est l'annotateur-référent, toutes les annotations liées à la structure sont systématiquement remises en question

5 Post-traitements : des données brutes à des données exploitables

Les textes annotés ont nécessité certaines opérations de nettoyage et d'harmonisation, en particulier pour mettre en cohérence les noms d'indices ajoutés par les annotateurs. Ces post-traitements, dont l'objectif est d'améliorer la fiabilité des extractions et des analyses à partir du corpus, sont les suivants :

1. Caractérisation des indices non typés lors de l'annotation. On distingue deux cas :
   - les traits prémarqués non typés automatiquement (numéros de titraille, certains connecteurs (*En revanche*), certains circonstants (*Ultérieurement*)), validés par les annotateurs sans remplissage du champ « nature de l'indice »,
   - les indices ajoutés manuellement par les annotateurs sans remplissage du champ « nature de l'indice ».

   Dans ces deux cas, le typage a été ajouté.
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 [8] : toutes ces unités font maintenant l'objet de la même caractérisation en termes de type (INDICE) et de nature (CIRCtps, MIL, etc.).

8. Pour des raisons de feuille de style, les traits prémarqués étaient à l'origine de type CIRCtps, MIL, etc.

6 Postface : retours sur la campagne

En guise de postface, nous mettons à profit le recul acquis au cours de cette expérience d'annotation à grande-échelle pour proposer des définitions clarifiées des objets annotés, accompagnées d'exemples plus parlants sélectionnés par l'annotateur-référent parmi les structures annotées. Ce guide revu et corrigé, qui pourrait s'intituler "guide si..." (comme "si on avait su"), concerne principalement le modèle d'annotation. Nous n'y revenons pas sur l'explication de la tâche d'annotation (présentée dans le guide original), étant donné que la procédure d'annotation est largement indépendante du modèle. Nous l'enrichissons en revanche de témoignages recueillis par l'annotateur-référent auprès des annotateurs au terme de la campagne ANNODIS.

Structures Enumératives (SE)

L'exemple 1 ci-dessous illustre les quatre composants d'une structure énumérative :

- une amorce (surligné en rose dans les exemples) : segment qui introduit l'énumération ;
- des items (surligné en jaune dans les exemples et ici au nombre de quatre) : segments qui constituent l'énumération ;
- une clôture (surlignés en orange dans les exemples) : segment qui résume ou clôt l'énumération ;
- un **ÉNUMÉRATHÈME** (en gras dans les exemples) : lexème présent dans l'amorce et/ou la clôture qui spécifie le critère de co-énumérabilité ou principe d'énumération (dans cet exemple, l'énumérathème « éléments » est présent dans l'amorce et répété dans la clôture).

Témoignage 1.

[une SE selon l'annotateur 2]Une SE est une structure pouvant recouvrir un empan textuel plus ou moins grand et qui établit une liste d'item sur la base d'une caractéristique commune. La structure canonique est constituée d'une amorce, d'items, et d'une clôture mais certains éléments peuvent souvent venir à manquer sans que ce soit pour autant un problème majeur pour repérer la SE.

Exemple 1 : exemple de SE.

L'évolution vers Homo sapiens se caractérise par les **éléments** suivants :

- expansion de la boîte crânienne et du volume du cerveau, en moyenne 1 400 cm3 (plus de deux fois celui des chimpanzés ou des gorilles). Pour certains anthropologues, la modification de la structure du cerveau est plus importante encore que l'augmentation de sa taille ;
- diminution de la taille des canines ;
- locomotion bipède, marche ; toutefois pour certains anthropologues, l'aptitude à courir est plus importante que l'aptitude à marcher.
- descente du larynx, ce qui permet le langage articulé.

Les liens entre ces **éléments**, leur valeur adaptative, et leur rôle dans l'organisation sociale est sujet à débat parmi les anthropologues.

voir l'exemple en contexte sur redac

Témoignage 2.

[annoter les SE selon l'annotateur 1]Il y a clairement des SE plus agréables à annoter que d'autres [...]

Témoignage 3.

[annoter les SE selon l'annotateur 2]Pour ce qui est des SE, je crois que le problème venait principalement du fait que certaines structures étaient intuitivement ressenties comme énumératives sans pour autant qu'il y ait d'indices clairs. Il y avait aussi le problème des items difficilement repérables ou découpables. Et celui des structures avec une magnifique amorce et pas d'items clairs. Ou un seul. La difficulté résidait aussi dans le repérage de l'amorce et de la clôture. Je crois avoir mis très peu de ces dernières, d'ailleurs.

L'amorce

Définition

L'amorce est un segment qui annonce une énumération. La présence d'une amorce n'est pas obligatoire dans une SE.

Témoignage 4.

[remarque de l'annotateur-référent] Pour certaines structures, l'amorce est le composant le plus facilement repérable. Cependant, pour d'autres, ce sont les items qui sont les plus visibles. Dans ce cas, il peut être nécessaire de remonter dans le texte pour rechercher une éventuelle amorce.

Exemple 2 : Exemple d'amorce.

Il est important de ne pas remettre en cause cette évolution et de poursuivre le rapprochement entre les sociétés à la condition, toutefois, que ce rapprochement ne se fasse pas au détriment de :

- l'expression publique des positions françaises sur le conflit israélo-arabe,
- notre capacité d'action dans la région, fondée certes sur la sécurité d'Israël mais aussi sur le refus de l'occupation et la nécessité d'une évacuation totale des territoires occupés en 1967 et de la création d'un état palestinien indépendant. La persistance depuis bientôt quarante ans de cette occupation est au cœur de l'instabilité dans la région.

Témoignage 5.

[l'amorce selon l'annotateur 1] L'amorce est une portion de texte, généralement courte, située directement avant les items. Elle annonce l'énumération de manière plus ou moins explicite. Exemple :

1. Les cinq piliers de l'islam sont les suivants :
2. Les piliers de l'islam sont au nombre de cinq :
3. Les piliers de l'islam sont au nombre de cinq.
4. Les piliers de l'islam sont des règles que tout musulman doit respecter.

Dans ces quatre exemples (construits), l'amorce est de moins en moins explicite.

Exemple 1 : Le prospect *sont les suivants* est l'indice le plus explicite au sens où il rend l'énumération obligatoire. Ne pas énumérer *les piliers* rendrait le texte incohérent.

Exemples 2 et 3 : C'est la ponctuation qui est l'indice le plus important. Dans l'exemple 2 comme dans l'exemple 1, ne pas énumérer après l'amorce serait incohérent à cause des deux points. En revanche, dans l'exemple 3, l'énumération n'est pas obligatoire mais le prospect sont au nombre de cinq crée l'attente.

Exemple 4 : Cette phrase pourrait ne pas être suivie d'une énumération et aucun élément ne crée l'attente. Si énumération il y a, c'est l'énumérathème *des règles* qui permet de la relier aux items et qui lui confère le statut d'amorce.

Indices pouvant signaler une amorce

Parmi les indices signalant les amorces, on peut citer :

- les titres de section : un titre peut être l'amorce soit des sections de niveau inférieur (titres compris), soit de la section titrée ;

– des indices ponctuationnels : les deux-points en fin d'amorce ;
– des indices lexicaux : toute portion de texte jugée par l'annotateur comme étant un indice d'amorce doit être annotée (il en va de même pour les indices d'item et de clôture). Les indices lexicaux d'amorce seront typés "annonce". Ils peuvent correspondre plus ou moins à des syntagmes nominaux à valeur cataphorique, généralement au pluriel, comportant un déterminant numéral, indéfini (*quelques*, *plusieurs*, etc.) ou collectif (*une foule de*, *une grande variété de*, *un grand nombre de*, *une série de*, etc.) et potentiellement entourés d'expressions telles que (*selon*, *suivant*, *à savoir*, etc.).

Les items

Définition

Les items sont les éléments obligatoires d'une SE, contrairement à l'amorce, la clôture et l'énumérathème qui sont facultatifs. Une SE est au minimum constituée de deux items.

Témoignage 6.

[les items selon l'annotateur 1] Les items sont les seuls éléments indispensables dans une SE. Ils vont au minimum par deux et peuvent avoir n'importe quelle taille (selon le grain de la structure). Les items d'une même SE sont cependant généralement tous à peu près de la même taille.

On distingue deux types d'agencements des items :
– dans les SE dites horizontales, les items ne sont pas visuellement signalés par la disposition (pas de saut de ligne, pas de titre, aucun signe typographique en dehors de la ponctuation).
– dans les SE dites verticales, les items sont séparés par un saut de ligne et signalés par un titre ou un signe typographique placé en début de ligne (numérotation, tiret, puce).

Exemple 3 : SE horizontale.

Les théoriciens de l'intégration considèrent depuis longtemps les groupes d'intérêts comme un indicateur fondamental du caractère des institutions européennes : soit leur influence est faible, et le cadre institutionnel européen peut être considéré comme intergouvernemental ; soit elle est importante, et ce cadre se rapproche alors du modèle fédéraliste ou fonctionnaliste.

Témoignage 7.

[les SE « binaires » selon l'annotateur 1] Les SE « binaires » : ces SE ne comportent que deux items et ont rarement une amorce ou une clôture. Elles sont typiquement marquées par les indices *d'une part*/*d'autre part*, *dans un premier temps*/*dans un second temps*.

Exemple 4 : SE verticale.

Le troisième rapport du GIEC insiste en particulier sur les **points** suivants :

- certains gaz à effet de serre, ont une espérance de vie longue, et influent donc sur l'effet de serre longtemps après leur émission (durée supérieure à 1 000 ans pour le CO2 selon le quatrième rapport) ;
- de par l'inertie du système climatique, le réchauffement planétaire se poursuivra après la stabilisation de la concentration des gaz à effet de serre. Ce réchauffement devrait cependant être plus lent ;
- l'inertie, plus grande encore, de la masse océanique fait que l'élévation du niveau des mers se poursuivra même après la stabilisation de la température moyenne du globe. La fonte de calottes glaciaires, comme celle du Groënland, sont des phénomènes se déroulant sur des centaines voire des milliers d'années.

Indices pouvant signaler des items

Parmi les indices signalant les items, on peut citer :

- les éléments issus de la mise en forme matérielle du texte : titres, tirets, puces, retraits
- les signes de ponctuation : virgules, points-virgules
- les marqueurs d'intégration linéaire (MIL) : *premièrement*, *en second lieu*, *le troisième*, *l'autre*, *tout d'abord*, *ensuite*, *enfin*, etc.)
- les circonstants temporels et/ou spatiaux (*en 1976*, *depuis 2009*, *à Toulouse*, *près de Caen*, etc.) (voir ci-dessous un exemple de SE avec circonstants spatiaux et de SE avec circonstants temporels)
- les circonstants notionnels : ce sont des syntagmes prépositionnels détachés en tête de phrase qui précisent un domaine d'activités et de connaissances spécifiques (*dans le domaine de la biologie*), une thématique particulière (*concernant la France*), un point de vue particulier (*en général*), ou encore un ensemble de concepts particuliers à un domaine précis (*en hôtellerie homologuée*)
- les parallélismes syntaxiques, comme cette suite d'infinitif dans l'exemple 7 ci-dessous
- tout autre indice lexical jugé par l'annotateur comme un indice d'item.

Exemple 5 : SE avec série de circonstants spatiaux**.**

Les relations nouées depuis des siècles dans la région nous valent assurément estime et considération. Elles suscitent aussi des attentes et des déceptions.

Au Maghreb, les gouvernements attendent de nous concours et, pour chacun d'entre eux, soutien exclusif. Les populations sont plus attentives à la coopération, à la liberté de circulation et à la situation des immigrés chez nous.

Au Proche-Orient, nos prises de parole sont scrut'ees et analys'ees dans le détail. Nous y sommes attendus, sollicités et espérés tant l'image d'une France compagnon de route des grandes causes arabes demeure encore enracinée.

L'approche est différente **dans le Golfe** où nous sommes vus comme un partenaire privilégié pour se soustraire à un tête-à-tête trop exclusif avec les États-Unis.

Les perspectives pour la France dans tous les domaines y sont remarquables. En témoignent tout récemment les opérations du Louvre et de la Sorbonne à Abou Dhabi.

Témoignage 8.

[les SE spatiales selon l'annotateur 1] Les SE spatiales, très présentes dans les textes issus du corpus GEOP, sont également facilement repérables. Le plus souvent, chaque item comporte un nom de pays ou de région du monde.

Exemple 6 : SE avec série de circonstants temporels.

Depuis septembre 2004, la France a pris la direction d'un mouvement diplomatique qui a conduit à l'adoption par le Conseil de sécurité de la résolution 1559 appelant au retrait des forces syriennes du Liban. **Après l'assassinat de l'ancien Premier Ministre Rafic Hariri,** elle a pris clairement position pour la coalition des forces politiques du 14 mars, un bloc dont le principal ciment et l'objectif commun étaient de mettre fin à l'influence syrienne au Liban. **Au lendemain de la guerre d'Israël contre le Hezbollah à l'été 2006,** elle a su mobiliser un large soutien international pour la mise en place d'une FINUL renforcée et pour la reconstruction du pays dévasté lors de la conférence de Paris 3.
Ces trois directions, engagées au cours des trois dernières années, méritent un examen critique, au niveau des objectifs d'une part, du cadre dans lequel la France déploie son activité et des partenaires qu'elle choisit d'autre part, pour envisager les options politiques à venir.

Témoignage 9.

[les SE temporelles selon l'annotateur 1] Les SE temporelles sont marquées par une énumération de dates. Elles sont faciles à repérer et comportent généralement un grand nombre d'items. Dans ces structures, les items sont souvent de tailles inégales, selon l'importance de l'événement relaté.

Exemple 7 : SE avec parallélisme syntaxique.

Placer l'accent sur l'occupation et la nécessité d'y mettre fin, aurait trois avantages : **repositionner le débat** autour du problème de 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 état 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.

Témoignage 10.

[Histoire de parallélismes par l'annotateur-référent] Le cas des parallélismes syntaxiques illustre bien l'avantage de ne pas donner aux annotateurs une liste fermée d'indices. En effet, la première version du guide n'en faisait pas mention. Dès l'annotation des premiers textes, les annotateurs ont remarqué qu'il était fréquent que les items d'une SE soient construits sur un même modèle syntaxique. Il a alors été décidé que, pour avoir une annotation homogène, ce phénomène serait annoté comme "parallélisme syntaxique" par tous les annotateurs. Les parallélismes syntaxiques ont ainsi été ajoutés à la liste des indices d'items.

Test

Pour repérer un item, vous pouvez tenter d'insérer *tout d'abord*, *ensuite* ou *enfin* au début ou à l'intérieur de l'item.

Exemple 8 : Exemple d'application du test sur l'exemple SEpara.

Placer l'accent sur l'occupation et la nécessité d'y mettre fin, aurait trois avantages : **tout d'abord** 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 ; **ensuite** 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 état viable et à l'intérieur de frontières sûres ; **enfin** désamorcer le débat qui lie l'opposition à la politique israélienne à la question de l'antisémitisme.

Témoignage 11.

En plus des indices de ponctuation, il y a des MIL et les parallélismes qui sont selon moi les trois types d'indices les plus révélateurs de la présence d'une SE.

La clôture

Définition

La clôture est un segment qui conclut la structure énumérative. La présence d'une clôture n'est pas obligatoire.

Exemple 9 : Exemple de clôture.

Pour tenter de concilier ces positions antagonistes, un équilibre délicat a été bâti à Doha : les questions de "mise en œuvre" relevant de négociations ouvertes dans le nouveau cycle sont traitées dans le cadre de ces négociations ; les autres questions, dites "en suspens", sont traitées par les "organes pertinents" de l'OMC. Ce découpage correspond au souci des pays du Nord de ne pas rouvrir, même partiellement, les négociations closes en 1994.

Indices pouvant signaler une clôture

Les indices signalant les clôtures sont uniquement des indices lexicaux. Parmi eux, on peut citer :

- des expressions conclusives comme *en conclusion*, *en résumé*, *pour conclure*
- des encapsulations : syntagme nominal fréquemment pluriel, dont le déterminant est généralement un démonstratif ou l'expression *de tel(le)s*, accompagné ou non d'un numéral (*ces sénarios*, *ces trois options*, *de telles avancées*, etc.)
- toute autre expression jugée par l'annotateur comme étant un indice de clôture.

Enumérathème

L'ÉNUMÉRATHÈME est 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 au sein d'une liste. Il peut apparaître en amorce et/ou clôture, notamment à l'intérieur d'un indice lexical d'amorce (e.g. une annonce) et/ou de clôture (e.g. une encapsulation).

L'expression de l'ÉNUMÉRATHÈME doit être limitée au maximum, elle correspond généralement à la tête lexicale d'un syntagme, comme dans l'exemple 1 où seul le lexème "éléments" est annoté comme étant l'ÉNUMÉRATHÈME, sans englober les déterminants ou extensions.

Exemple 10 : ÉNUMÉRATHÈME en amorce, reformulé en clôture.

Deux autres **peintures** semblent dater de cette période à l'atelier, qui sont tous les deux des "Annonciations". L'un est petit, large de cinquante-neuf centimètres pour seulement quatorze de haut. Il s'agit d'une prédelle se plaçant à la base d'une composition plus large, et, dans ce cas, pour un tableau de Lorenzo di Credi duquel il fut séparé. L'autre est un travail beaucoup plus important, de deux cent dix-sept centimètres de large.

Dans ces deux **annonciations**, Léonard a dépeint la Vierge Marie assise ou agenouillée à la droite de l'image, et un ange de profil s'approchant d'elle par la gauche.

Témoignage 12.

Après réflexion, j'ai identifié d'où venait ma difficulté à considérer *points* ou *parties* comme énumérathèmes. Je pense que c'est parce qu'ils peuvent plus ou moins s'appliquer à n'importe quelle énumération. Intuitivement, cela me pose problème de mettre au même niveau deux énumérathèmes comme *les piliers de l'Islam* et *les points importants*.

Test

Que ce soit en amorce ou en clôture, la présence de l'ÉNUMÉRATHÈME peut être confirmé en insérant les expressions suivantes immédiatement après l'expression contenant l'ÉNUMÉRATHÈME (qui en est alors la tête lexicale) :

> **tel(le)s que énuméré(e)s ci-dessous/ci-dessus**
> et/ou **tel(le)(s) que décrit(e)(s) ci-dessous/ci-dessus**

Il faut noter en revanche que son impossibilité ne peut pas l'infirmer.

Exemple 11 : Exemple d'application du test.

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 :

Devient :

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

Exemple 12 : Exemple d'application du test en clôture.

De telles **évolutions** ne sont pas une fatalité. Pour arrêter l'engrenage de violences [...]
Devient :
Les évolutions, ***telles que nous venons de les énumérer ci-dessus,*** ne sont pas une fatalité. Pour arrêter l'engrenage de violences [...]

Chaînes Topicales (CT)

Définition

Une CT est un segment qui se caractérise par le fait que la majorité des propositions qui le composent ont pour objet (parlent de, sont à propos de, apportent des informations au sujet de) **un seul et même référent**. L'expression de ce référent commun doit passer nécessairement par la position sujet grammatical.

Témoignage 13.

Il me semble que les SE étaient au final plus difficile à annoter que les CT. Cela est paradoxal puisque en terme de clarté les SE semblent plus évidentes que les CT, notamment en raison de certains indices forts.

Exemple 13 : CT.

César désigna dans son testament trois héritiers, les petits-fils de ses sœurs, à savoir Octave, Lucius Pinarius Scarpus et Quintus Pedius. **Il** légua les trois quarts de son héritage au premier et le quart restant aux deux autres. Dans la dernière clause de son testament, **César** adopta Octave, le futur empereur Auguste, et lui donna son nom. Enfin, **il** légua au peuple romain ses jardins près du Tibre et trois cents sesterces par tête.

Il faut noter tout de même qu'une CT n'est pas nécessairement composée uniquement de propositions portant sur le référent qui fait l'unité du segment. En effet, des commentaires, illustrations, par exemple, peuvent être insérés à l'intérieur d'une CT.

Indices pouvant signaler une CT

Toute expression thématique et coréférentielle constitue un indice de CT. Des circonstants notionnels d'un type particulier (*quant à Euronews*, *s'agissant de la Russie*) peuvent également indiquer le début ou la continuation d'une CT.

Témoignage 14.
Pour les CT, il me semble que le repérage automatique des expressions potentiellement co-référentielles était bien plus utile : repérer des occurrences multiples est assez aisé.

Exemple 14 : une CT et ses indices**.**
Il a été créé par **un courant socio-politique post- soixante-huitard**, proche de « l'éducation alternative », et qui réfléchissait sur la place de l'enfant dans la société et les relations adultes enfants. On y trouvait des sociologues, des philosophes, des architectes, des écrivains, des éducateurs, des enseignants, des médecins, qui avaient en commun une curiosité pour les organisations sociales qui mettaient l'enfance au centre de leurs préoccupations. Malgré une absence de tabous toute scientifique, **ce courant intellectuel** a voulu se démarquer de la pédérastie, et évacuer la dimension sexuelle des relations adultes enfants. **Il** a donc inventé le mot « pédophilie » qui, comme dit dans la définition étymologique, vient du grec « paidos », « enfant » et « philein », « aimer ». **Ce petit cercle intellectuel** ne pouvait maintenir longtemps le sens sémantique du mot pédophilie dans sa stricte étymologie.

Exemple 15 : CT introduite par un circonstant notionnel**.**
Quant aux Américains, **ils** ne doivent pas devenir prisonniers de leurs moyens militaires : il leur faut se donner les moyens de contribuer à la solution des crises par d'autres moyens. S'**ils** laissent les Européens reconstruire après eux aux quatre coins du globe, **ils** feront de l'Europe ce qu'un observateur américain appelait récemment le centre moral du monde. **Leur leadership** n'en sera que plus contestable, et donc plus fragile car plus contesté.

Test
Pour vérifier que les expressions sont bien coréférentielles, il doit être possible de les substituer par l'expression référentielle complète. L'impossibilité d'une telle substitution amène à conclure à la non coréférentialité.

Témoignage 15.
J'ai parfois annoté des « structures » en n'étant pas persuadée qu'elles en étaient vraiment.

De la difficulté d'annoter « de haut »

Témoignage 16.

Pour annoter un texte, les annotateurs ont tous procédé dans le même ordre : les SE sont annotées en premier et les SUR le sont dans un second temps. Il peut arriver que, lors de l'annotation des SUR, une SE qui n'avait pas été repérée au premier passage soit découverte (la présence de coref prémarquées attire l'attention sur des zones qui n'ont été que survolées lors de l'annotation des SE). Dans ce cas, l'annotateur pose un *glue note* « SE » pour repérer l'endroit et y revenir plus tard pour annoter la structure. On repère ainsi des SE peu visibles, ne comportant généralement aucun indice prémarqué. C'est le cas des SE présentées dans cet exemple (*glue note* présent dans un des .aa). Par expérience, je peux dire que les SE annotées « après coup » le sont rarement par plusieurs annotateurs et sont en règle générale horizontales, petites (peu d'items) et peu fournies en indices. Il est très rare de repérer après coup une SE qui ne soit pas problématique.

Témoignage 17.

L'annotation se faisant « de haut », autrement dit sans une lecture précise du texte, il est parfois difficile de savoir si l'auteur est toujours en train de parler de la même chose ou s'il a changé de sujet. C'est particulièrement le cas pour les textes du sous-corpus GEOP et du sous-corpus LING, qui comportent des termes spécialisés.

7 Travaux publiés

Afantenos, S., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-Dac, M., Le Draoulec, A., Muller, P., Péry-Woodley, M.-P., Prévot, L., Rebeyrolle, J., Tanguy, L., Vergez-Couret, M., & Vieu, L. (2012). An empirical resource for discovering cognitive principles of discourse organisation : the ANNODIS corpus. In Actes *LREC*, Istanbul, Turkey, juillet 2012. URL : `http://www.lrec-conf.org/proceedings/lrec2012/pdf/836_Paper.pdf`)

Ho-Dac, L.-M., Fabre, C., Péry-Woodley, M.-P., & Rebeyrolle, J. (2010). On the signalling of multi-level discourse structures. In Actes *MAD 2010* : Multidisciplinary Perspectives on Signalling Text Organisation, Moissac (France) 17-20 mars 2010, p. 94-105.

Ho-Dac, L.-M., Fabre, C., Péry-Woodley, M.-P., & Rebeyrolle, J. (2009). Des indices aux marqueurs : méthodes de découverte de marqueurs discursifs complexes. In *Linguistic and Psycholinguistic Approaches to Text Structuring (LPTS-09)*, Paris, France, septembre 2009.

Ho-Dac, L.-M., Fabre, C., Péry-Woodley, M.-P., & Rebeyrolle, J. (2009). A top-down approach to discourse-level annotation. *Corpus Linguistics Conference*, Liverpool, Angleterre, 10–23 juillet 2009.

Ho-Dac, L.-M., Fabre, C., Péry-Woodley, M.-P., & Rebeyrolle, J. (2009). Corpus annotation of macro discourse structures. In *1st International conference on corpus linguistics (CILC-09)*, University of Murcia, Espagne, 7–9 mai 2009.

Ho-Dac, L.-M., Fabre, C., Péry-Woodley, M.-P., Rebeyrolle, J., & Tanguy, L. (2012). An empirical approach to the signalling of enumerative structures. *Discours* [En ligne], 10 | 2012, mis en ligne le 16 juillet 2012, consulté le 11 septembre 2012. URL : `http://discours.revues.org/8611` ; DOI : 10.4000/discours.8611

Ho-Dac, L.-M., Péry-Woodley, M.-P., & Tanguy, L. (2010). Anatomie des structures énumératives. In Actes *TALN 2010*, Montréal, Québec, 19-23 juillet 2010. URL : `http://www.iro.umontreal.ca/~felipe/TALN2010/Xml/Papers/all/taln2010_submission_26.pdf`

Péry-Woodley, M.-P., Afantenos, S. D., Ho-Dac, L.-M., & Asher, N. (2012). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *TAL* 52 3, 71-101. URL : `http://www.atala.org/La-ressource-ANNODIS-un-corpus`

Péry-Woodley M.-P., Asher N., Enjalbert P., Benamara F., Bras M., Fabre C., Ferrari S., Ho-Dac L.-M., Le Draoulec A. , Mathet Y., Muller P., Prévot L., Rebeyrolle J., Tanguy L., Vergez-Couret M., Vieu L., & Wildöcher A. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives, In *TALN 2009*, Senlis, France, Juin 2009.

Péry-Woodley, M.-P., Ho-Dac, L.-M., Fabre, C., Rebeyrolle, J., & Tanguy, L. (2011). High-level discourse structures : Topical Chains and Enumerative Structures in a diversified annotated corpus. *Corpus Linguistics Conference*, Birmingham, Angleterre, 19–22 juillet 2011.

Annexe [9]

Exemples d'arbitrage des SE annotées

Exemple 1 : Arbitrage au niveau du schéma.

SE annotée par A1

La taille moyenne des hommes, aujourd'hui, en France, est de 1,75 m, et celle des femmes de 1,62 m, pour des masses respectives moyennes de 75 et 61 kg. Les données individuelles sont très variables autour de ces moyennes, avec une forte influence de facteurs environnementaux, des comportements et des régimes nutritionnels. Les moyennes elles-mêmes varient beaucoup selon les populations et les époques.
|Les jeunes naissent avec une masse autour de 3 kg, et une taille d'environ 50 à 60 cm, après une gestation de neuf mois.| |Totalement dépendants à la naissance, leur croissance dure plusieurs années.| |La maturité sexuelle survient entre 12 et 15 ans.| |La croissance des garçons continue souvent jusque vers 18 ans (la croissance se termine vers 21-25 ans avec la solidification de la clavicule).| |L'espérance de vie est très dépendante des conditions matérielles et de la disponibilité de soins médicaux. L'espérance de vie se situe aujourd'hui autour de 75 ans dans les pays les plus riches, et est inférieure à 40 ans dans les plus pauvres. Des cas isolés de longévité approchent 120 ans, et la personne ayant vécu le plus longtemps sans doute possible sur son âge est la française Jeanne Calment, qui a vécu plus de 122 ans.|
|L'être humain possède 23 paires de chromosomes (contre 32 pour le cheval).|

SE annotée par A2

La taille moyenne des hommes, aujourd'hui, en France, est de 1,75 m, et celle des femmes de 1,62 m, pour des masses respectives moyennes de 75 et 61 kg. Les données individuelles sont très variables autour de ces moyennes, avec une forte influence de facteurs environnementaux, des comportements et des régimes nutritionnels. Les moyennes elles-mêmes varient beaucoup selon les populations et les époques.
Les jeunes naissent avec une masse autour de 3 kg, et une taille d'environ 50 à 60 cm, après une gestation de neuf mois. Totalement dépendants à la naissance, leur croissance dure plusieurs années. |La maturité sexuelle survient entre 12 et 15 ans.| |La croissance des garçons continue souvent jusque vers 18 ans (la croissance se termine vers 21-25 ans avec la solidification de la clavicule).| |L'espérance de vie est très dépendante des conditions matérielles et de la disponibilité de soins médicaux. L'espérance de vie se situe aujourd'hui autour de 75 ans dans les pays les plus riches, et est inférieure à 40 ans dans les plus pauvres.| |Des cas isolés de longévité approchent 120 ans, et la personne ayant vécu le plus longtemps sans doute possible sur son âge est la française Jeanne Calment, qui a vécu plus de 122 ans.|
|L'être humain possède 23 paires de chromosomes (contre 32 pour le cheval).|

9. Cette annexe est une sélection élaborée à partir du compte-rendu rédigé par l'annotateur référent lors du processus d'arbitrage.

L'extrait du texte présenté ici est reproduit tel qu'il a été annoté par les deux annotateurs ayant repéré une SE à cet endroit. A1 n'a relié aucun indice à la SE. A2 a associé à chaque item un indice de type parallélisme syntaxique.

L'absence d'indices clairs, d'amorce et de clôture rend difficile la délimitation d'items. Les deux annotateurs se sont surtout appuyés sur la ponctuation mais ne sont pas pour autant totalement en accord.

Cette portion de texte fournit une liste de différentes caractéristiques de l'espèce humaine mais elle est peu (voire pas) structurée :

Paragraphe 1 : taille moyenne de l'adulte
Paragraphe 2 : poids et taille à la naissance ;
croissance ;
maturité sexuelle ;
croissance des garçons ;
espérance de vie ;
Paragraphe 3 : nombre de chromosomes.

Les paragraphes 1 et 2 présentent des moyennes ou des tranches d'âge concernant le développement et la morphologie de l'humain alors que le paragraphe 3 donne une « constante » génétique. Ces informations ne se situent pas sur le même plan. Cet extrait n'est donc structuré ni par le sens, ni par la typographie, ni par des indices lexicaux. S'il s'agit bien d'une énumération de faits, peut-on pour autant parler de structure ?

Décision : la SE annotée par A1, bien que n'étant reliée à aucun indice, semble plus cohérente que celle annotées par A2 qui débute au milieu d'un paragraphe et divise en trois la partie traitant de l'espérance de vie sur la base d'un parallélisme syntaxique douteux. C'est donc la structure annotée par A1 qui a été reportée dans le gold.

Exemple 2 : Arbitrage au niveau de l'amorce.

La signification des différents éléments de cette dénomination est la suivante :
- Homo est un mot latin [...] qui signifie [...]
- sapiens est un adjectif latin [...] qui signifie [...]
- Linné est le nom du scientifique qui a nommé et décrit l'espèce.
- 1758 est l'année de l'appellation.

Cette SE a été repérée par les trois annotateurs, qui s'accordent sur les items. Ils sont cependant en désaccord pour l'amorce.

Deux annotateurs ont annoté comme suit (indice de type PROSPECT encadré, énumérathème en **gras**) :

Exemple 3.

amorce annotée par A1 et A2
La signification des différents **éléments** de cette dénomination est la suivante :

amorce annotée par A3
La **signification** des différents éléments de cette dénomination est la suivante :

Dans le premier cas, on considère que chaque item correspond à un « élément » dont on donne la signification. Dans le second, on considère que chaque item correspond à la signification d'un des éléments.

Décision : dans le gold, c'est le premier cas (l'annotation de A1 et A2) qui a été retenu, la majorité l'emportant.

Exemple 4 : Arbitrage au niveau des items.

VII. Variantes théologiques

[...] trois branches principales [...]

VII.1. Le sunnisme
[1mm] [...]

VII.2. Le chiisme
[1mm] [...]

VII.3. Le Kharidjisme
[1mm] [...]

VII.4. Autres
[1mm] [...]

Cette SE a été repérée par les trois annotateurs mais ils ne l'ont pas tous annotée de la même manière.

Deux d'entre eux ont repéré le PROSPECT « trois branches principales ».

Dans l'amorce, l'auteur explique pourquoi l'Islam a plusieurs variantes théologiques mais il ne cite que les trois principales sans préciser que les autres seront évoquées par la suite. Les deux annotateurs ont donc considéré que la SE était constituée de l'amorce et des items 1, 2 et 3.

Le troisième annotateur a quant à lui donné priorité aux titres de section. En effet, le titre de plus haut niveau « Variantes théologiques » permet d'englober les quatre items dans la SE.

On peut donc avoir une SE a trois ou quatre items selon qu'on considère que l'indice le plus « important » est le titre ou le PROSPECT.

Décision : la SE à quatre items signalées par les titres de section a été reportée dans le gold.

Exemples d'arbitrage des CT annotées

Exemple 5 : Arbitrage au niveau du schéma.

indices en gras pour A1 et encadrés pour A2 et A3	A1	A2	A3
Le nom Homo sapiens relève de **la terminologie** scientifique introduite par Carl von Linné, élaborée pour sa classification systématique des espèces : **la dénomination binomiale**. En dehors de l'usage qui en est fait pour cette **dénomination** le mot latin « homo » doit porter une minuscule lorsqu'il est utilisé uniquement en tant que mot latin. Lorsqu'il est utilisé en tant que nom biologique de genre (« Homo »), c'est-à-dire le premier terme de la **dénomination**, il doit porter la majuscule. **La dénomination scientifique** complète de l'espèce humaine est, suivant **cette terminologie** : Homo sapiens, Linné 1758.	–		–
	x		x
	x		x
	x		x
	x		x
	x		x
	x		x
	x		x
	x		x
	x		–
La signification des différents éléments de cette dénomination est la suivante :	x		
	x		
– Homo est un mot latin au nominatif (avec majuscule et en italique) qui signifie « homme » en français. Il désigne ici le genre biologique.	x	–	
	x	x	
	x	–	
– sapiens est un adjectif latin (avec minuscule et italique), qui signifie en français : intelligent, sage, raisonnable ou encore prudent. Il désigne ici l'espèce.	x	–	
	x	x	
	x	–	
– Linné est le nom du scientifique qui a nommé et décrit l'espèce.	x		
– 1758 est l'année de l'appellation.	x		
Toutefois, en pratique, en zoologie, le nom et l'année sont rarement précisés.	x		
	x		
Jusqu'en 2003, l'espèce Homo sapiens était subdivisée en deux groupes distincts, considérés comme deux sous-espèces, dont l'une était l'espèce humaine actuelle, et l'autre, une espèce cousine éteinte, celle de l'homme de Néandertal. Comme pour toute sous-espèce la conséquence **terminologique** a été de créer des noms trinomiaux en rajoutant un adjectif, toujours latin (et en italique), après le nom d'espèce. C'est ainsi que l'espèce humaine était appelée Homo sapiens sapiens. Bien que souvent encore entendue, **cette terminologie** n'est plus en vigueur pour la majorité des scientifiques. En effet, n'étant pas **une terminologie** constitutive, mais référentielle, **elle** est le réceptacle évolutif qui reflète l'état des connaissances et la place de l'homme dans la compréhension que celui-ci a du monde : de nouvelles connaissances ou une nouvelle compréhension pourront produire une nouvelle classification, qui pourra conduire à une **nouvelle dénomination**.	x	–	
	x	x	
	x	x	
	x	x	
	x	x	
	x	x	
	x	x	
	x	x	
	x	x	
Le deuxième atout de **cette terminologie** est, depuis Linné, d'avoir offert un langage commun. Par delà les noms vernaculaires propres à chaque langue pour désigner l'espèce humaine ou les membres de celle-ci : Human, Mensch, Ser humano... et parfois multiples au sein d'une même langue : l'espèce humaine, l'homme, l'humain ; Homo sapiens se présente comme un vocable de référence, certes de nature scientifique, mais qui a su par ailleurs acquérir une notoriété dépassant celle du jargon.	x	x	
	x	–	
	x		
	x		
	x		
	x		
	x		
	–		

Les trois annotateurs ont repéré des CT mais ne sont d'accord ni sur leur empan ni sur les indices. Il y a d'une part le terme Homo sapiens et d'autre part la terminologie qui permet de nommer scientifiquement les espèces.

En lisant attentivement le texte, on voit que l'auteur fait alternativement référence à l'un ou l'autre en utilisant des termes très proches.

A1 a repéré deux petites structures (référents : Homo et sapiens). La troisième structure repérée par A1 est englobée par la structure plus large repérée par A3 (référent : la terminologie et/ou la dénomination). A2 a repéré une seule structure ayant pour référent Le nom Homo sapiens située sur une portion de texte repérée par A3 mais avec des indices différents.

Ceci est l'un des rares cas où l'annotateur référent est amené à recréer une structure à partir des trois annotations. En effet, les petites structures repérées par A1 peuvent être rattachées à la CT repérée par A2. La CT annotée par A3 tend à considérer terminologie et dénomination comme faisant référence à la même chose. Or, ce n'est pas le cas même si l'auteur a tendance à employer les deux termes pour désigner l'un ou l'autre des référents.

Décision : dans le gold, on décide d'annoter deux CT ayant les mêmes frontières (i.e. celles délimitées par A3) mais ayant deux référents. Selon le contexte, les termes terminologie et dénomination sont rattachés à l'une ou l'autre structure, la première regroupant largement ce qui se rapporte au nom scientifique donné à l'espèce humaine et la seconde regroupant ce qui se rapporte à la terminologie permettant de nommer les espèces.

L'extrait de texte ci-dessous reproduit les structures telles qu'elles ont été annotées dans le gold. Les indices reliés à la CT le nom Homo sapiens sont colorés en rose et ceux reliés à la CT la terminologie sont colorés en bleu.

Exemple 6 : Annotation décidée pour le gold.

Le nom Homo sapiens relève de la terminologie scientifique introduite par Carl von Linné, élaborée pour sa classification systématique des espèces : la dénomination binomiale. En dehors de l'usage qui en est fait pour cette dénomination le mot latin « homo » doit porter une minuscule lorsqu'il est utilisé uniquement en tant que mot latin. Lorsqu'il est utilisé en tant que nom biologique de genre (« Homo »), c'est-à-dire le premier terme de la dénomination, il doit porter la majuscule. La dénomination scientifique complète de l'espèce humaine est, suivant cette terminologie : Homo sapiens, Linné 1758.
La signification des différents éléments de cette dénomination est la suivante :
– Homo est un mot latin au nominatif (avec majuscule et en italique) qui signifie « homme » en français. Il désigne ici le genre biologique.
– sapiens est un adjectif latin (avec minuscule et italique), qui signifie en français : intelligent, sage, raisonnable ou encore prudent. Il désigne ici l'espèce.
– Linné est le nom du scientifique qui a nommé et décrit l'espèce.
– 1758 est l'année de l'appellation.
Toutefois, en pratique, en zoologie, le nom et l'année sont rarement précisés.
Jusqu'en 2003, l'espèce Homo sapiens était subdivisée en deux groupes distincts, considérés comme deux sous-espèces, dont l'une était l'espèce humaine actuelle, et l'autre, une espèce cousine éteinte, celle de l'homme de Néandertal. Comme pour toute sous-espèce la conséquence terminologique a été de créer des noms trinomiaux en rajoutant un adjectif, toujours latin (et en italique), après le nom d'espèce. [C'est ainsi que l'espèce humaine était appelée Homo sapiens sapiens. Bien que souvent encore entendue, cette terminologie n'est plus en vigueur pour la majorité des scientifiques. En effet, n'étant pas une terminologie constitutive, mais référentielle, elle est le réceptacle évolutif qui reflète l'état des connaissances] et la place de l'homme dans la compréhension que celui-ci a du monde : de nouvelles connaissances ou une nouvelle compréhension pourront produire une nouvelle classification, qui pourra conduire à une nouvelle dénomination.
Le deuxième atout de cette terminologie est, depuis Linné, d'avoir offert un langage commun. Par delà les noms vernaculaires propres à chaque langue pour désigner l'espèce humaine ou les membres de celle-ci : Human, Mensch, Ser humano... et parfois multiples au sein d'une même langue : l'espèce humaine, l'homme, l'humain ; Homo sapiens se présente comme un vocable de référence, certes de nature scientifique, mais qui a su par ailleurs acquérir une notoriété dépassant celle du jargon.

Le rattachement des indices peut être discuté. Ils reposent sur les principes suivants :

- quand l'auteur parle d'une caractéristique relative à la dénomination de toutes les espèces, l'indice est rattaché à la CT terminologie.
- quand il s'agit d'une caractéristique propre à la dénomination de l'espèce humaine, l'indice est rattaché à la CT le nom Homo sapiens.

Dans le passage contenu entre [ ], le premier emploi de terminologie fait référence à l'utilisation du terme Homo sapiens sapiens exclusivement et non

à l'ajout d'un second adjectif pour différencier les sous-espèces. Le second emploi, en revanche, donne une caractéristique générale de la terminologie utilisée pour nommer les espèces. Les deux indices ne sont donc pas reliés à la même structure.

Exemple 7 : Arbitrage au niveau de indices.

indices annotés par A1

Au début de la campagne d'Afghanistan, le gouvernement américain a **acheté en exclusivité toutes les images** de la zone en guerre prise par le satellite à haute résolution Ikonos. **Cette procédure** était plus facile à adopter que l'interdiction de photographier, initialement prévue. Même si les agences de renseignement se sont montrées réticentes à distribuer l'imagerie commerciale auprès des forces armées, l'objectif initial **de contrôle de l'information** a été atteint.
Mais **cette procédure** ne peut fonctionner avec des producteurs d'imagerie plus nombreux. **Contrôler la diffusion** de l'imagerie apparaît comme une tâche de plus en plus illusoire.

indices annotés par A2

Au début de la campagne d'Afghanistan, **le gouvernement américain a acheté en exclusivité toutes les images de la zone en guerre** prise par le satellite à haute résolution Ikonos. **Cette procédure** était plus facile à adopter que l'interdiction de photographier, initialement prévue. Même si les agences de renseignement se sont montrées réticentes à distribuer l'imagerie commerciale auprès des forces armées, l'objectif initial de contrôle de l'information a été atteint.
Mais **cette procédure** ne peut fonctionner avec des producteurs d'imagerie plus nombreux. Contrôler la diffusion de l'imagerie apparaît comme une tâche de plus en plus illusoire.

NB : Ici, le premier indice (qui pose le topique) est une proposition.

Les différences entre les annotateurs concernent :

- l'empan du premier indice : le sujet de la proposition est inclus par A2 mais pas par A1 ;
- le nombre d'indices : A1 a annoté deux indices de plus qu'A2.

Décision : l'annotation de A2 est retenue, les deux indices *de contrôle de l'information* et *Contrôler la diffusion* ayant été considérés comme ne correspondant pas à la procédure mais à son objectif (i.e. les américains achètent en exclusivité les images satellites dans le but de contrôler leur diffusion).