**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Guide d’annotation**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Projet PASTEL (P**erforming **A**utomated **S**peech **T**ranscription for **E**nhancing **L**earning)

Depuis la démocratisation des technologies de l’information et de la communication, le monde de l’enseignement supérieur et de la formation pour adultes est de plus en plus interrogé par la société quant au renouvellement et à l’adaptation des pratiques pédagogiques. D’une part, les frontières entre apprentissage guidé et auto-apprentissage sont de moins en moins marquées, ce qui tend à la redéfinition du rôle de l’enseignant et de l’apprenant, et d’autre part la technologie, de plus en plus accessible, permet de diversifier et de combiner les modes d’interaction enseignant/apprenant et apprenant/apprenant.

Alors l’idée est d l’utiliser des technologies numériques, et plus spécialement des cours en ligne, dans le système éducatif comme dans la formation professionnelle, ou à destination du grand public.

Le projet PASTEL a pour objectif d’explorer le potentiel de la transcription automatique en temps réel pour l’instrumentation de situations pédagogiques mixtes, où les modalités d’interaction sont présentielles ou à distance, synchrones ou asynchrones. Les technologies de reconnaissance de la parole approchent d’un niveau de maturité suffisant qui permet d’envisager de nouvelles possibilités au niveau de l’instrumentation des pratiques pédagogiques et générer de nouveaux usages.

La finalité du projet est de proposer des outils susceptibles d'enrichir les situations pédagogiques en répondant aux besoins exprimés par les enseignants et les étudiants.

**La segmentation thématique**

La segmentation thématique a pour objectif de découper des textes en parties thématiquement homogènes.

La segmentation thématique intervient dans différents domaines en organisation de l’information, tels que créer un résumé automatique de textes pour lequel la segmentation sert à isoler les thématiques et les parties les plus représentatives ou récupérer des segments de texte courts qui répondent directement à une requête au lieu de récupérer de longs documents qui doivent être lus et examinés par l’utilisateur afin de ramener l’information précise.

**Objectifs**

Afin de guider la segmentation thématique, nous devons répondre à la question suivante : Qu’est-ce qu’un thème ?

Nous avons décidé de répondre à cette question en fonction des objectifs que nous nous sommes fixés de l’utilisation de ces segments thématiques.

**Objectif 1: Adaptation du modèle de langage**

Les modèles de langage sont un des nombreux constituants d’un système de reconnaissance de la parole. La modélisation du langage a pour objectif de résumer les connaissances générales liées à un langage naturel. Un modèle de langage est généralement formé à partir d’une large quantité de données représentatives de la tâche pour laquelle le modèle sera utilisé. Cependant, ce ML n’est pas fiable lorsqu’il s’agit de transcrire des documents oraux traitant une autre tâche. Les systèmes dont les modèles de langage sont entraînés à partir de données généralistes ne sont pas performants pour transcrire des données liées à des domaines spécifiques.

**Objectif 2 : Enrichissement de contenu**

L’enrichissement du cours consiste à lier les parties du discours avec des informations et des connaissances contenues dans des documents extérieurs, issus de bases de données spécifiques, d’encyclopédies en ligne comme Wikipédia, ou d’autres sources du web. Ceci permet d’améliorer la compréhension du cours.

**Objectif 3 : Obtenir un découpage de séquences autonomes pour miniMOOC**

Il s’agit de découper le cours en une suite de séquences de quelques minutes homogènes, qui soient suffisamment autonomes et complètes pour constituer l’objet d’un mini-mooc (feuilletonage). En règle générale, la construction même du cours permet d’obtenir ce découpage en feuilletons et peut par exemple se retrouver dans la diapositive qui présente le plan.

**Définition de la tâche d’annotation**

Il s’agit de déterminer les frontières de changement thématique.

**Hypothèse**

Un slide ne contient qu’une seule thématique.

—> La frontière thématique ne peut être qu’au voisinage de changement du slide.

Il existe cependant plusieurs exceptions :

1. l’énoncé du plan : ce slide est multi-thème puisqu’il présente toutes les parties abordées dans le cours
2. les slides de conclusion et de références bibliographiques (à la fin du cours) : ils peuvent être également multi-thèmes car s’appliquent à l’ensemble du cours
3. les apartés : le professeur peut digresser et être « hors sujet » par rapport au slide qu’il présente

**Quoi annoter ?**

Procédure : À chaque changement de slide, on recherche un changement de thème et on annote :

1. l’instant exact de changement thématique. Cet instant doit être défini entre deux mots complets.
2. le niveau de changement thématique par les tags 1 et 2.

Tag « **1** » : on aborde une nouvelle notion tout en restant dans le même thème.

Tag « **2** » : on note un changement de thème plus générale qui permet d’arrêter l’apprentissage à ce moment là et de reprendre plus tard l’apprentissage d’autres ensemble de notions. Les tag « 2 » permettent ainsi de créer les mini-MOOC autonomes et sont constitués de sous- segments de type 1

Vidéos à annoter pour bien fixer le guide d’annotation:

Colin de la Higuera - Réseaux sociaux

<http://coconotes.comin-ocw.org/video/m20131010/>

Béatrice Daille - Langage naturel

<http://coconotes.comin-ocw.org/video/m20131115/>

**Données CoCo:**

Les données CoCo sont issus du projet COCo qui met à disposition un certain nombre de vidéos avec des ressources potentielles.

Les vidéos collectées concernent des cours magistraux de licence donnés par des enseignants, des séminaires de recherche et des webinaires (interview de deux personnes par deux autres). Les vidéos sont alignés temporellement avec les diapositives de la présentation.

**Outil à utiliser pour l’annotation ?**

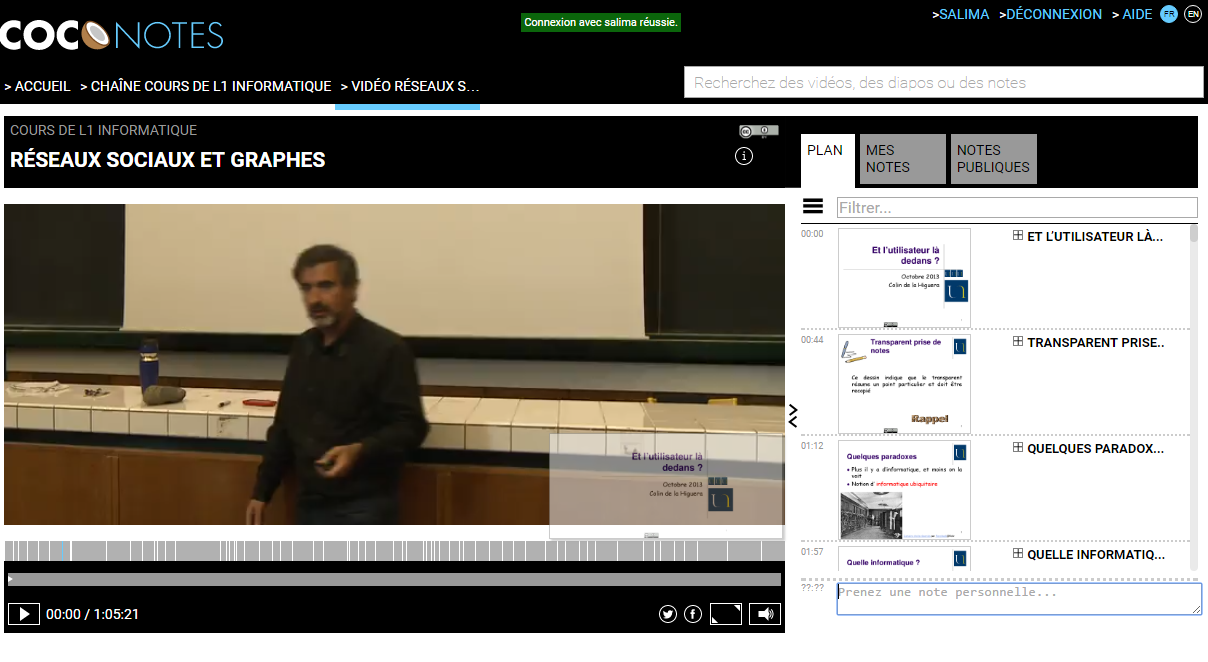
**Plateforme coco**

**Etapes à suivre :**

Créer un compte dans la plateforme coco

Se connecter

Ouvrir le vidéo à annoter : une interface s’affiche (voir la figure ci-dessous)



**1**

**2**

**3**

**4**

Légende :

1. la vidéo et la diapositive plus petite en transparent (si on clique au centre de la diapo, la diapositive prend la place de la vidéo et cette dernière passe en transparent)

2. le lecteur audio et le cours découpé en diapositives. Le découpage en diapositives permet de passer d’une diapositive à l’autre.

3. trois onglets permettant d’accéder : au plan, aux notes personnelles, aux notes publiques.

4. un champ texte pour saisir une note. Lorsque l’on clique dans la zone de texte, si la lecture est en cours, elle est mise en pause. Le timecode de la note sera celui du lecteur audio. Lorsque l'on tape sur la touche "Entrée", cela enregistre la note et relance la vidéo. La note apparaît dans la liste des notes, le temps associé peut ensuite être modifié manuellement. Il est également possible de supprimer la note et de changer sa visibilité (publique, privée).

Pour cette tâche d’annotation :

* se placer aux frontières du changement de slide,
* s’il y a un changement de niveau 1, cliquer dans la zone de texte en mode privé et saisir « 1 ».
* s’il y a un changement de niveau 2, cliquer dans la zone de texte en mode privé et saisir « 2 ».
* A la fin du travail, récupérer le fichier json accessible en ajoutant « /cinelab » à la fin de l’url (exemple : <http://coconotes.cominocw.org/video/m20131010/cinelab)>