

RAPPORT D'ÉTUDE: SÉANCE 7

CAMILLE COBOS, MARS 2012

1. Testez le logiciel "Cordial" (accessible uniquement depuis les postes du LaboC) avec un corpus de votre choix. Quels sont les avantages et les désavantages ?
 - **Avantages:** correction orthographique, grammaticale et typographique ; commentaire sur le texte traité à partir de statistiques et analyses syntaxique et sémantique; statistiques des occurrences du texte ; plusieurs dictionnaires dont Littré, synonymes etc. ; outils de traduction ; personnalisation des paramètres d'étiquetage ; relevé de mots clés ; analyse en arbre de la phrase (cf. capture ci-dessous).

The screenshot shows the Cordial software interface. The main window displays a text document with the sentence: "Trayvon n'avait qu'un paquet de bonbons Skittles à la main." A pop-up window titled "Analyse en arbre de la phrase" is open, showing a syntax tree for this sentence. The tree structure is as follows:

- Prop: indépendante
 - Subj[évent]
 - Groupe nominal
 - Sub Trayvon
 - Verbe avec ségation[évent]
 - Subj[évent]
 - Adv n (non-élisé)(inacceptable)
 - Ver: évent (propre)
 - Com[pl]
 - Complément d'objet direct[évent]
 - Groupe nominal
 - Dét: un
 - Sub paquet (agglutination/contenus/correspondance/métaphore)
 - Groupe nominal prépositionnel
 - prép: de
 - Sub bonbons (bonbonnerie)
 - Apposition (bonbons)
 - Groupe nominal
 - Sub Skittles
 - Complément circonstanciel[évent]
 - Groupe nominal prépositionnel
 - prép: à
 - Dét: la
 - Sub main (polytémique: membre)
 - Subj[évent]

The interface also shows a toolbar at the top and a status bar at the bottom with the text "M.M. Param. | DÉFAUT | 50 % | ** Sans CLASSE **".

- **Désavantages:** Pour ce qui est de la version au LaboC, le logiciel ne parvient pas à reconnaître certains éléments dont noms propres et certains mots en anglais.

2. Testez les deux outils de Traduction automatique avec un corpus de votre choix: [Google Translat](#) et [Systran](#), expliquez de manière très brève pourquoi parle-t-on "Vers un schéma général".

Corpus Test:

Apple TV Isn't Coming Until 2013

Most analysts say Apple will release an Apple TV this year. Not Asian research group CLSA, which said in a note this morning that it thinks it comes out in 2013. We don't have the full note, just a brief of what they said, which is included below.

➤ Résultats Google Translat:

Apple TV ne vient pas jusqu'en 2013

La plupart des analystes estiment que Apple sortira un Apple TV de cette année. Non asiatique groupe de recherche CLSA, qui a déclaré dans une note ce matin qu'il pense qu'il sort en 2013. Nous n'avons pas la note en entier, juste un aperçu de ce qu'ils ont dit, qui est inclus ci-dessous.

➤ Résultats Systran :

Apple TV ne vient pas jusqu'en 2013

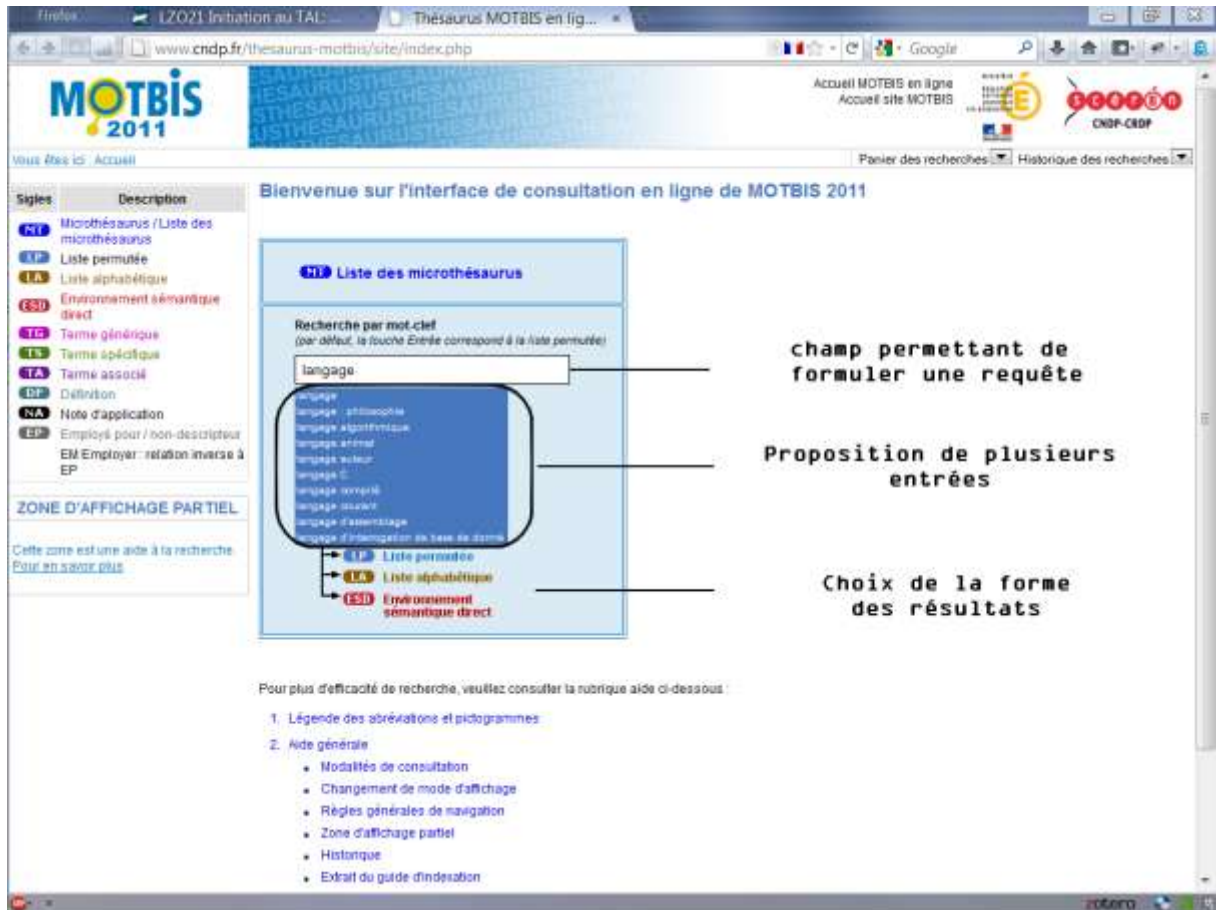
La plupart des analystes disent qu'Apple libérera Apple TV cette année. Le groupe de recherche non asiatique CLSA, qui a indiqué dans une note ce matin qu'elle la pense sort en 2013. Nous n'avons pas la pleine note, juste un dossier de ce qu'ils ont dit, qui est inclus ci-dessous.

→ Pourquoi parle-t-on d'un schéma général ?

Comme on peut le voir dans les deux exemples de traduction ci-dessus, elles comportent quelques failles dans la syntaxe et la structure de la phrase. Si l'on parle d'un schéma général, c'est parce qu'actuellement le traitement automatique des langues se fait à partir d'un système qui traite d'une part la forme et d'autre part le sens grâce à une suite de modules. On cherche désormais à faire dans un premier temps une analyse syntaxique, puis dans un second temps à prendre en compte le sens et les traits linguistiques. Pour traduire, un langage-pivot (ou métalangage) est donc nécessaire pour pouvoir rendre le sens d'une phrase ou d'un texte généré dans une autre langue.

3. Testez MOTBIS, Qu'est-ce c'est un **thésaurus**? **Comment ça marche ?**

MOTBIS est un thésaurus pour l'Éducation. Voici une capture d'écran montrant la page d'accueil, et la manière dont sont traitées les requêtes:



Si l'on clique sur environnement sémantique direct, on trouve un résultat sous cette forme :



- Qu'est ce qu'un thésaurus ? Comment ça marche ?

« Un thésaurus est un ensemble structuré de mots clés qui est utilisé dans un système de documentation pour rendre de manière univoque les notions contenues dans un document. Il s'agit donc d'un instrument qui sert à formaliser la terminologie utilisée par l'auteur d'un texte. Comme les auteurs des documents sont libres dans le choix des mots, leur langage n'est souvent pas univoque (utilisation de synonymes par exemple), si bien qu'en procédant à une recherche classique dans un texte, on n'obtient pas toujours le résultat escompté.

C'est aussi un outil de recherche. Comme l'utilisateur a accès au vocabulaire de l'indexeur, les deux parties peuvent utiliser les mêmes termes pour désigner les mêmes concepts. Les potentialités d'un thésaurus apparaissent de manière encore plus évidente dans un environnement multilingue. L'indexation et la recherche peuvent se faire dans une autre langue grâce au lien que constitue le thésaurus multilingue.

En pratique, un thésaurus se compose de concepts structurés selon une série de catégories (domaine, micro-thésaurus, descripteur, non-descripteur) et de liens entre les concepts. »

Source : http://www.senate.be/zoek/thesaurushelp_fr.html

Note : L'organisation des thésaurus monolingues est définie par des normes (ISO 2788 et AFNOR NF Z47-100.)

4. Rédiger un compte rendu de lecture sur le document "*Une petite introduction au Traitement Automatique des Langues Naturelles*".

Cette introduction au Traitement Automatique des Langues Naturelles par François Yvon redéfinit ce qui se cache derrière l'étiquette « traitement automatique des langues naturelles (TALN) ». Pour cela, il commence par un bref rappel historique et fait état des difficultés majeures : l'ambiguïté et l'implicite.

C'est grâce à de nombreuses applications que ces difficultés peuvent être limitées et être ignorées ou traitées. Delà, plusieurs niveaux de traitement sont nécessaires pour parvenir à une compréhension naturelle du langage. Pour cela différents modules qui permettent dans un premier temps la segmentation des mots, puis le traitement lexical, le traitement syntaxique, le traitement sémantique et enfin le traitement pragmatique. Pour parvenir à de meilleurs résultats, il faut concevoir les niveaux de traitements de façon à ce qu'ils communiquent entre eux, on parlera alors de « bas niveau » et de « haut niveau ». Pour pouvoir obtenir une structure syntaxique correcte, il faut faire appel à des informations sémantiques.