TER 2010 Groupe Cydia

CONTACT PDF DIAPORAMA VIDEO DEMO TELECHARGER

Notre équipe est composée de 4 étudiants en Master 1 Informatique novices dans le domaine du TALN. Notre travail n'aurait jamais abouti sans l'expertise, la disponibilité et le recadrage permanent de Mme la Pr. Violaine Prince.

Nous tenons aussi à remercier Monsieur le Pr. Jacques Chauché de nous avoir aimablement fourni les indispensables ressources SYGFRAN ainsi que Monsieur Alexandre Labadié pour son précieux utilitaire SygServeur.

1) Notre approche

SygAnnot est une **application pour l'annotation automatique des textes en langue française**. Sa vocation est d'expérimenter notre approche de l'**étiquetage thématique**.

Dans notre approche orientée vers le traitement automatique de l'information, nous entendons par « annotation » le fait d'insérer *dans le texte* des balises terminologiques afin de constituer un étiquetage thématique structuré susceptible d'être exploité par d'autres applications comme par exemple des logiciels de recherche d'information, des systèmes de question/réponse ou encore de fouille de texte, d'aide au résumé etc. Cet étiquetage dans le texte, au niveau de la phrase ou du paragraphe précise "*de quoi parle*" la partie annotée.

Pour cela, SygAnnot met en œuvre 3 phases distinctes :

Etape 1) La fabrication du lexique:

pour la **création et la mise à jour des lexiques** dans un thème donné (Histoire de France, Marine, Transports etc.), le menu « *Lexique* » permet d'extraire une liste d'étiquettes (que nous nommons par la suite items lexicaux afin d'éviter la confusion avec les étiquettes grammaticales de SYGFRAN) d'un corpus constitué de textes choisis pour leur cohérence thématique. Les fichiers constituant le corpus sont au préalable étiquetés par l'analyseur SYGFRAN de Jacques CHAUCHÉ qui génère l'association entre les lemmes et leur catégorie grammaticale. Un exemple est donné plus loin.

Etape 2) La validation du lexique:

Le lexique peut être **validé par intervention humaine**. Via le menu « *Lexique* > *Etiquettes*», l'utilisateur (un expert du domaine) peut décider de supprimer ou ajouter des étiquettes. Ce travail peut être effectué via Internet et donne un aspect **collaboratif** à la fabrication du lexique.

Etape 3) L'annotation automatique

Le menu « *Fichier* > *Annoter* » permet de réaliser l'**appariement** des textes à annoter avec les items lexicaux pour décider de l'**étiquetage** (TAG). Il y a deux niveaux d'annotation : phrase et paragraphe. SygAnnot v1.x est limité au niveau phrase. Le niveau paragraphe permettra d'envisager la segmentation automatique des textes qui sera, n'en doutons pas, une évolution de la recherche sur Internet. Dans cette perspective, nous pouvons envisager le résultat d'une recherche non plus comme un ensemble de liens vers des documents mais un *assemblage d'informations qui parlent de ce que l'on cherche*.

En termes de technologie, SygAnnot respecte la syntaxe XML :

- Arbre morpho-syntaxique au format XML : La manipulation de l'arbre parenthèsé SYGMART est délicate. Pour plus de facilité, SygAnnot exploite l'utilitaire JAVA **SYGserver** fournit par Alexandre LABADIE pour transformer l'arbre parenthèsé généré par SYGFRAN en arbre JDOM.

- Lexique au format XML

« Ordinairement, l'annotation est construite sur l'idée du travail collaboratif : les documents d'origine sont mis en concordance avec un public porteur de nouvelles idées (informations, connaissances), avec un vocabulaire commun, sur des thèmes proches et donc avec des habitudes spécifiques. Les annotations apportent des éléments informationnels qui enrichissent et qui valorisent le contenu et le contenant. Et enfin, elles introduisent une valeur ajoutée sur le document par accumulation des interprétations fondées du point de vue de l'utilisateur et des intérêts assignés par rapport aux domaines spécifiques des annotateurs. » [SIDHOM ROBERT DAVID]

Sur le Web 2.0, nous disposons de Wikipedia qui constitue un véritable corpus collaboratif thématique. Pour perpétuer cette dimension collaborative, les lexiques fabriqués par SygAnnot peuvent facilement être contrôlés via Internet.

Validation collaborative des lexiques

Quels items lexicaux ? Annoter quoi ?

Le premier principe fondateur de SygAnnot est donné par le linguiste LE GUERN: « *Le syntagme est « l'unité minimale du discours qui permet de désigner un objet*.» [LE GUERN]

Découper en syntagmes

Le syntagme, c'est-à-dire un mot (*marine*, *naviguer…)* ou un groupe de mot (*marine marchande*, *naviguer au près)* est notre unité. "*Félix le chat*", "*Naviguer au près*", "*Toto*" ne laisse pas de doute sur l'objet qu'ils désignent.

Dans une première version de SygAnnot (SygAnnot v.1.x), nous nous concentrons sur les groupes nominaux, c'est-à-dire des groupes de mots dont le gouverneur *(noyau) est un nom (nom commun, nom propre etc.).*

Nous procédons tout d'abord à une extraction automatique des groupes pour les enregistrés dans un lexique structuré. Nous créons ainsi nos items lexicaux. Ceux-ci sont validés par des utilisateurs qui ont vocation à ne conserver que les items prépondérants du domaine. Cette validation est possible dans le cadre d'un travail collaboratif via Internet grâce à une mise en forme du lexique XML via une feuille de style XSL.

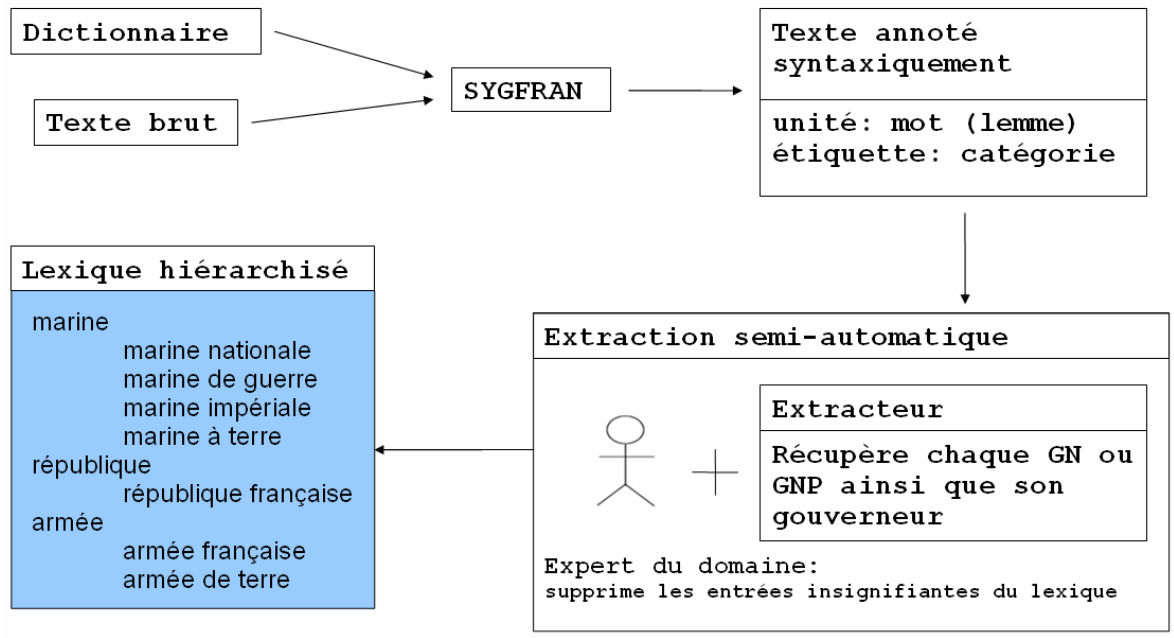
Sur quels critères peut-on définir la prépondérance d'un syntagme **?** *Comment mesurer l'importance qu'il tient dans le discours ?*

SygAnnot comptabilise le nombre d'occurrence de chaque item lexical pour pré filtrer les structures de groupe les plus courantes et les soumettre au contrôle du lexique. En d'autres termes, SygAnnot compte combien de fois apparaît une étiquette dans le corpus et cette valeur constitue un indicateur pour le contrôle du lexique. On pense qu'un objet mentionné fréquemment dans un discours à de forte chance d'y tenir un rôle important.

Dans la constitution du lexique, le mot de la fin revient à l'utilisateur qui en qualité d'expert valide les étiquettes.

2) La mise en oeuvre

2.1) Création du lexique



Pour extraire les items lexicaux, la première règle retenue pour appréhender la constitution des groupes est la suivante (soient X, Y et Z des lemmes de catégorie N (Nom Commun, Nom Propre…) :

Cas (1): Groupe de la forme « X + Préposition + Y » *chat de Mireille* Cas (2): ou groupe de la forme « X + Adjectif » *chien fatigué*
Dans le cas (1), on tient un groupe nominal prépositionnel (appelons-le G1) et on le retient comme étiquette. (Inutile de chercher une relation avec autre groupe de la forme « Préposition + Z »). On vérifie ensuite si G1 est suivi d un adjectif, cela afin de détecter un autre groupe nominal de la forme « G1 + adjectif ».
Dans le cas (2): Soit X est un groupe de la forme « Z + Préposition + T » (X est alors groupe nominal propositionnel) *à l'assemblée* Soit X est un nom. On tient alors un groupe nominal et on le retient comme étiquette.

Suite à cette extraction, la structure du lexique est hiérarchisée. Les étiquettes sont ordonnées selon une superclasse (notion générique) donnée par le gouverneur du groupe et des sous-classes (notions spécialisées) données par les gouvernés. Par exemple, dans le thème « Histoire de France », l'ensemble « *guerre de cent ans* » est plus précis que le mot « *guerre* ».

Pour illustrer la phase de création du lexique, considérons un bref extrait de corpus sur la thématique "système d'informations". Un extrait du texte source qui fonde le corpus est le suivant: "*Les navires marchands sont soumis à la réglementation internationale.*"

L'analyse morpho-syntaxique de SYGFRAN donne l'association lemme/catégorie suivante :

Les[Article défini Masculin Féminin] navires[Nom Commun Masculin] marchands[Adjectif Masculin] sont [Verbe conjugé] soumis[Participe Passé] à[Préposition] la[Article défini Féminin Singulier] réglementation[Nom Commun Féminin Singulier] internationale[Adjectif Féminin Singulier] .[Ponctuation]

Ce résultat est copié dans un fichier que nous appelons "fichier SYG". Avec SygAnnot v.1.x, cette copie est manuelle.

Le résultat de l'extraction par SygAnnot v.1.x est le suivant :

navires
* :navires marchands :1
* :navires de plaisance :1
---
réglementation
* :réglementation internationale :1

Nous constatons la présence d'une superclasse « navires» qui contient 2 sous-classes plus précises : « navires marchands » et « navires de plaisance ».

C'est une structure hiérarchisée qui met en jeu une « superclasse » (notion générique) donnée par le gouverneur du groupe et des « sous-classes » (notions spécialisées) données par l'ensemble du groupe. Cette hiérarchisation est essentielle pour affiner l'annotation.

En pratique, pour vérifier la cohérence du lexique généré, il suffit d'annoter les fichiers qui ont constitués le corpus.

Structure lexicale hiérarchisée

2.2) Annotation

Ce travail se base sur la *théorie du gouvernement* [CHOMSKY & TESNIERE].

On distingue deux niveaux d'annotation :

Annotation au niveau de la phrase :

Les linguistes s'accordent à reconnaître le gouverneur du groupe nominal comme sujet de la phrase. Dans tout groupe nominal, en règle générale, on choisit de retenir le membre de gauche comme gouverneur du groupe nominal.

Soient X et Y des lemmes de catégorie NOM (noms communs, noms propres etc.), l'extraction lexicale est implémentée selon les règles suivantes :

POUR chaque phrase:

SI la phrase possède un groupe sujet

SI le groupe sujet de la phrase est apparié au gouverneur de l'item lexical (superclasse)
ALORS
Chercher à apparier un terme de la phrase avec les gouvernés donnés par le lexique. (sous-classes)
Annoter avec l'item lexical de plus courte distance.
SINON
SI le verbe de la phrase est apparié au gouverneur de l'item lexical (superclasse)
ALORS chercher à apparier un terme de la phrase avec les gouvernés donné par le lexique. (sous-classes)

SINON ne pas étiqueter (pour éviter les erreurs d'annotation).

Dans SygAnnot v.1, le filtre est réduit à la règle suivante :

POUR chaque phrase:

SI la phrase possède un groupe sujet

SI GS pertinent
ALORS évaluer distance de L avec
- tous les gouverneurs des items lexicaux (superclasse) (1)
- toutes les sous-classes (2)

récupérer la meilleure distance de type (1)
SI distance > seuil
SI (1) > (2) Alors apparier phrase avec seulement superclasse
SINON apparier phrase avec la sous-classe et sa superclasse

SINON (c'est une phrase nominale ou un ensemble de phrases nominales)
ALORS évalue distance de L avec
- tous les gouverneurs des items lexicaux (superclasse) (1)
- toutes les sous-classes (2)

Récupérer la meilleure distance de type (1)
SI distance > seuil
SI (1) > (2) Alors apparie PN avec seulement superclasse
SINON apparier PN avec la sous-classe et sa superclasse

Pour chaque phrase du texte à annoter, SygAnnot v.1, récupère le groupe sujet sous la forme d'une chaîne de caractères et calcule la distance de Levenshtein entre :

- ce groupe sujet et tous les groupes et gouverneurs (les termes qui gouvernent leur groupe) des superclasses du lexique.

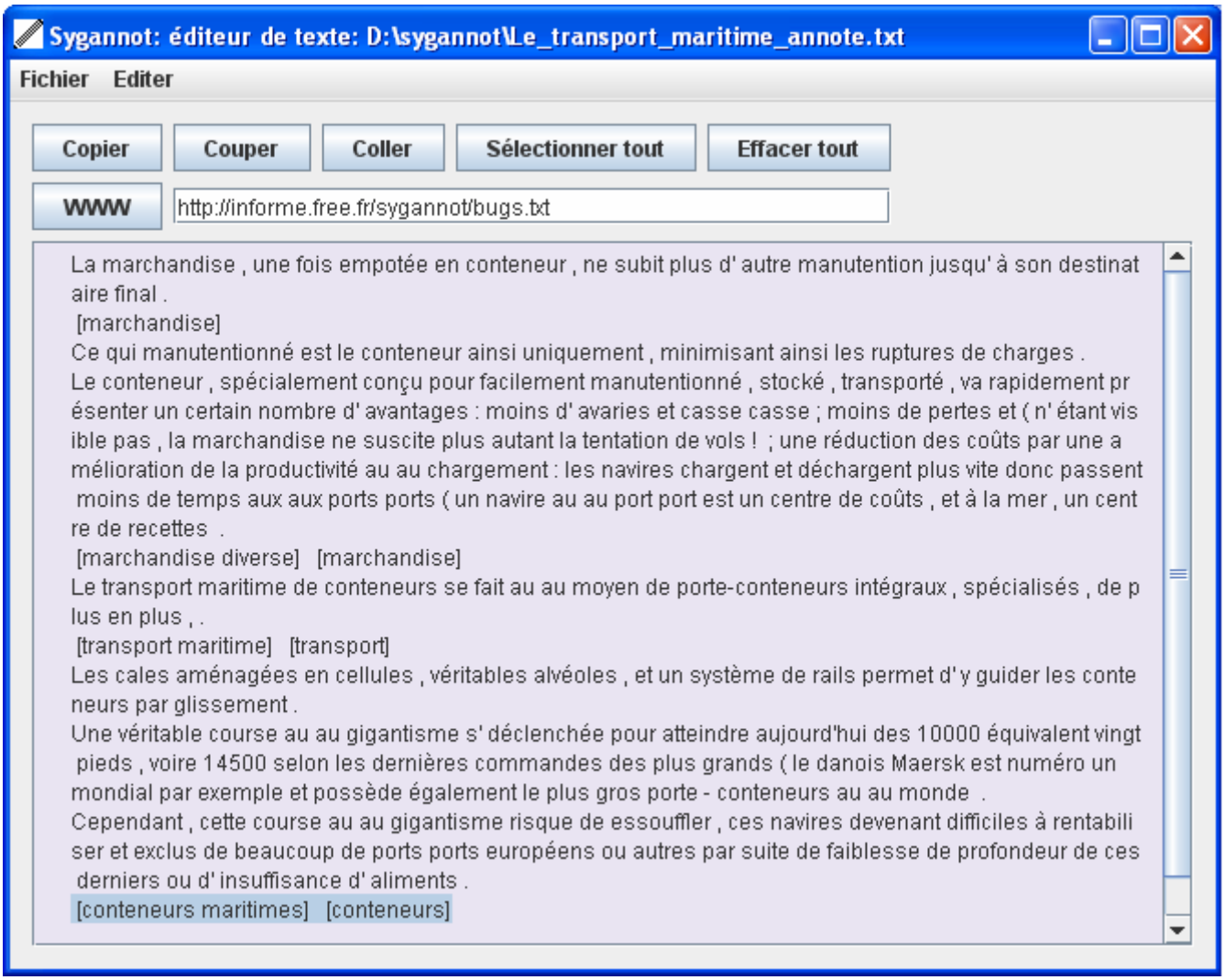
On aura donc deux distances :

- celle entre le sujet et le gouverneur de l'item lexical
- celle entre le sujet et le groupe de l'item lexical

On retient l'étiquette du lexique qui retourne la plus grande distance de Levenshtein en fixant un seuil minimal en dessous duquel l'appariement est jugé trop faible.
L'intention est d'étiqueter avec le maximum d'étiquettes, en les ordonnant par ordre décroissant selon la mesure de Levenshtein relevée. [Marine moderne], [Marine]...
Le seuil pour l'appariement est ajusté empiriquement. Pour notre approche expérimentale, la version 2 de SygAnnot proposera un curseur pour permettre à l'utilisateur de rechercher facilement la valeur qui permet l'appariement le plus pertinent.

Lors de l'appariement, certains termes (déterminants, pronom…) peuvent être considérés comme du bruit. Nous proposons une gestion simple de ces termes via la mise à jour d'un simple fichier texte.

Nous donnons priorité à l'appariement le plus grand. Mais le lexique est une hiérarchie et par conséquent, il faut étiqueter non seulement avec l'étiquette la plus proche, mais également avec les termes qui sont des hyperonymes (des surclasses).



Annotation au niveau du paragraphe:

1) En annotant au niveau du paragraphe, nous recherchons, dans le texte, la cohérence thématique: délimiter les paragraphes thématiques par le regroupement de tiers phrases qui se suivent ayant au moins une étiquette en commun. Sygannot v.1 n'implémente pas cette phase.

Note : pour évaluer la distance de Levenshtein et apparier les syntagmes aux items lexicaux , nous utilisons la librairie Java simmetrics de Sam Chapman.

3) Des sources

Voici les deux documents clés pour notre travail:

Sahbi SIDHOM - Charles ROBERT - Amos DAVID *« Analyse automatique de textes comme point de départ d'un processus d'annotation »*
http://hal.archives-ouvertes.fr/docs/00/03/64/79/PDF/revue_e-TI_2005.pdf
Revue électronique internationale en technologies de l'information (e-TI) (2005)

Seyed Mohammad MAHMOUDI « *Indexation automatique et la Recherche d'information dans les documents »*
http://www.webreview.dz/IMG/pdf/indexation_automatique_de_la_reherche.pdf
RIST Vol16 N°02 Année 2006

Et nos sources d'inspiration sur le Web :

Didier SCHWAB page personnelle
http://getalp.imag.fr/homepages/schwab/

Lylia ABROUK —Abdelkader GOUAICH — Chedy RAISSI « Annotation automatique de documents »
http://hal.archives-ouvertes.fr/docs/00/20/45/14/PDF/inforsid_final.pdf

Nicolas BÉCHET Thèse 8 décembre 2009 « Extraction et regroupement de descripteurs morpho-syntaxiques pour des processus de Fouille de Textes »
http://www.lirmm.fr/~bechet/These/These.pdf

Mehdi YOUSFI-MONOD — Violaine PRINCE *« Compression de phrases par élagage de leur arbre morpho-syntaxique »*
http://hal-lirmm.ccsd.cnrs.fr/docs/00/12/28/42/PDF/Yousfi-Monod_Prince_06.pdf

Anne ABEILLE *« Guide des annotations en constituants »*
http://www.llf.cnrs.fr/Gens/Abeille/guide-annot.pdf

John F. SOWA
http://www.jfsowa.com/ontology/

Céline BENNINGER « Une meute de loups, une brassée de questions : Collections, quantification et métaphore »
http://www.armand-colin.com/download_pdf.php?idd=0&cr=9&idr=7&idart=2324

http://infogrid.org/wiki/Reference/PidcockArticle?story=20030115211223271
http://websemantique.org/Ontologie
http://www.sites.univ-rennes2.fr/urfist/Supports/Indexation/Indexation4IndexAutomatisee.html
http://fr.wikipedia.org/wiki/Syntagme

Sans oublier les précieux supports de notre Pr. encadrant Violaine PRINCE :
http://www.lirmm.fr/~prince/nouveautes/troisieme-cours-tal.pdf

4) Des perspectives

L'avenir de SygAnnot passe avant tout par des ***améliorations*** :

- Pour une meilleure extraction des items lexicaux. Cela se fera par une analyse plus poussée des fichiers SYG en complétant les règles de filtrage pour couvrir l'ensemble de l'étiquetage de SYGFRAN : ex. du[Préposition] du[Article défini Masculin Singulier]

- Par une gestion plus fine du bruit, pour un gain de performance.

- Par une interface plus complète offrant notamment des curseurs pour l'expérimentation (ajustement du seuil d'appariement notamment).

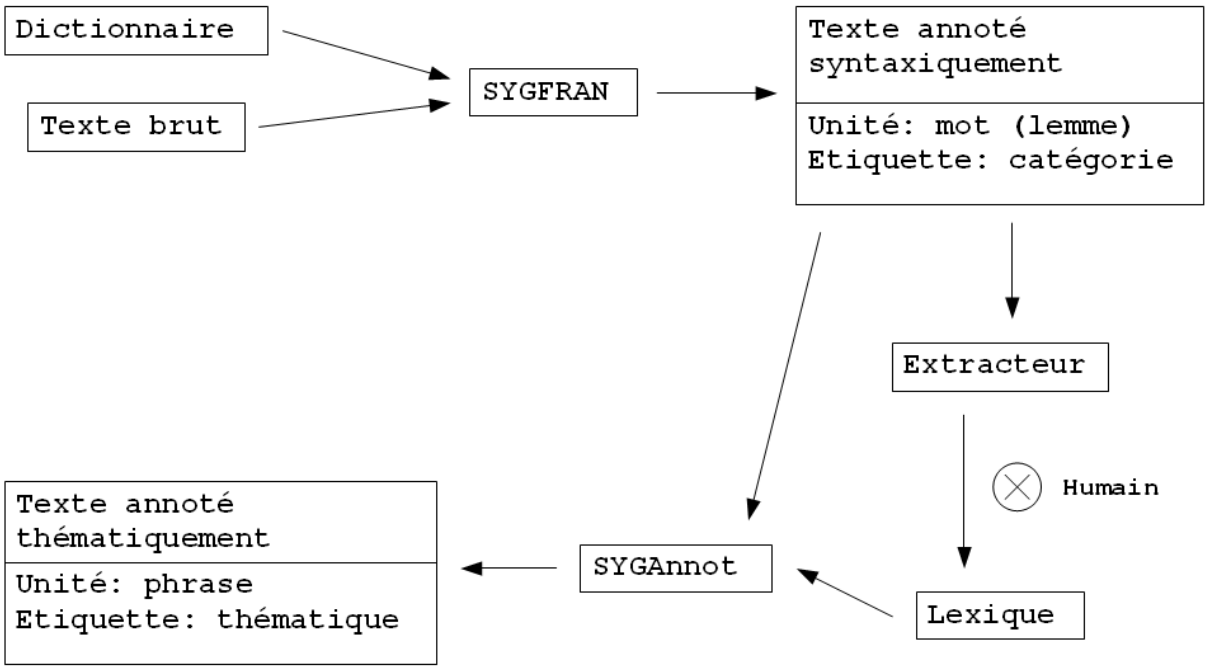
L'avenir de SygAnnot laisse aussi espérer des ***évolutions :***

- L'Annotation au niveau du paragraphe pour autoriser la recherche de la cohérence thématique dans le texte (regroupement de phrases qui se suivent ayant au moins une étiquette en commun.)

- L'élargissement du traitement de SygAnnot aux autres types de syntagmes, notamment les syntagmes verbaux [naviguer] [naviguer au près] pour une portée thématique plus importante.

Ne poussons pas plus loin notre ambition même si l'idée d'interconnecter SygAnnot à un réseau lexical tel que jeuxdemots nous laisse songeur…

5) En résumé



SYGFRAN offre l'arbre parenthèsé morpho-syntaxique (étiquetage grammatical au niveau du lemme) que convertit l'utilitaire SygServeur. L'étiquetage de SYGFRAN est exploité par SygAnnot pour extraire les syntagmes du corpus constitué de textes choisis pour leur forte cohérence thématique et structurelle. La phase d'annotation insert des étiquettes thématiques. La granularité de l'étiquetage de SygAnnot 1.x est la phrase.

SygAnnot, une application qui permet d'expérimenter l'annotation et se projeter vers une recherche documentaire de nouvelle génération.

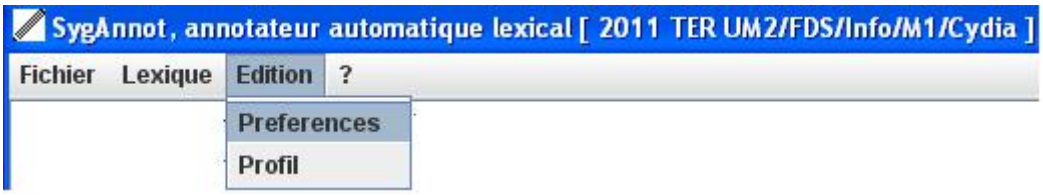
6) Mode d'emploi

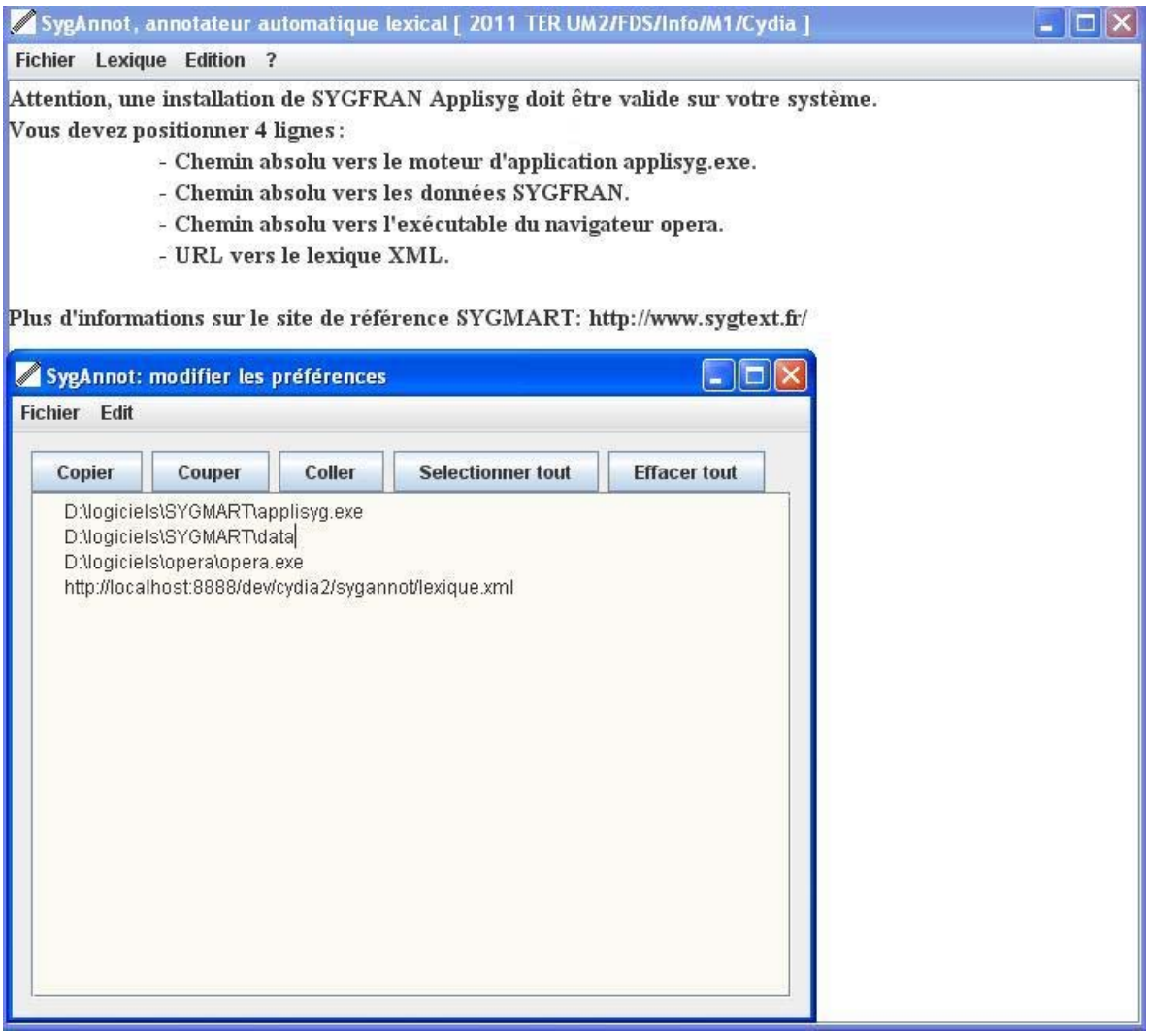
6.1) Installation

Pour installer SygAnnot, rendez-vous sur http://informe.free.fr/sygannot/maj

6.2) Utilisation

Lancez l'application via le fichier sygannot.bat. La fenêtre principale s'affiche. Saisissez votre paramétrage via le menu *Edition* > *Préférences* et suivez les instructions qui s'affichent dans la fenêtre principale :





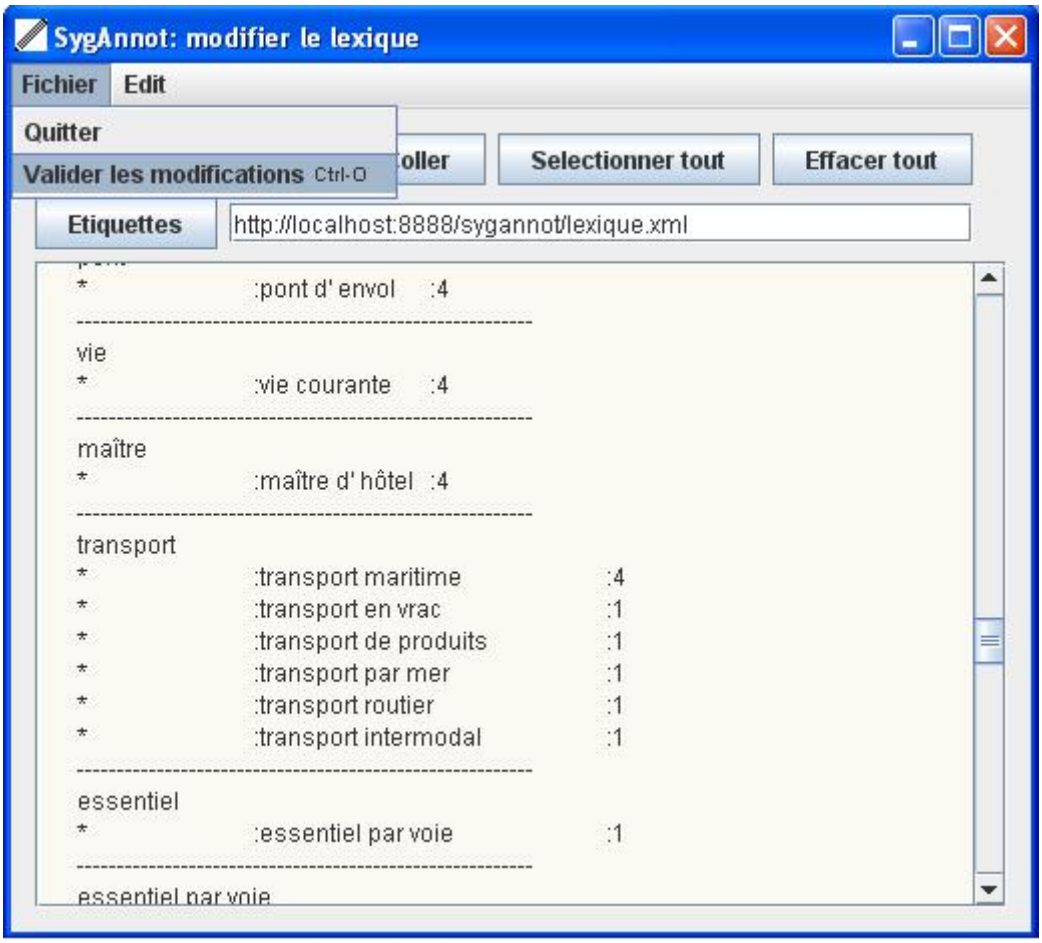
L'utilisation de SygAnnot se décompose en 4 étapes :

1. La **constitution du corpus** via le menu *Fichier > Corpus*

2. La **fabrication du lexique** via le menu *Lexique > Fabriquer*

3. La **validation du lexique** via le menu *Lexique >Modifier* (lexique local)

Si vous êtes expert du domaine, vous pouvez supprimer les étiquettes non pertinentes.



Ou via le menu *Lexique > Etiquettes* (lexique en ligne sur le Web)

4. L'**annotation** via le menu *Fichier* > *Annoter*

A chaque étape les **instructions** sont précisées dans la fenêtre principale.

7) Annexes

Convention pour le codage
Exemple d'arbre parenthèsé généré par AppliSyg
Exemple d'arbre XML généré par SYGserver
SYGtext.fr: Moteurs d'application SYGMART
Analyse SYGFRAN en ligne
Bugs connus
Démo Versions Beta
Version PDF
Dépôt SVN
Résultats

8) Du vocabulaire

Analyse morpho-syntaxique
Annotation
Antidictionnaire
AppliSyg
Catégorie grammaticale
Constituants (d'une phrase)
Contexte
Corpus
Dictionnaire
Distance de Levenshtein
Enoncé
Etiquette
Fichiers SYG
Forme fléchie
Fréquence d'un terme
Fréquence d'occurrences
Gouverneur
Gouverné
Groupe
Item lexical
Lemmatisation
Lemme
Lexique
Morphème
Mot ou lemme
Morphologie d'un item lexical
Mot
Nature
Nombre d'occurrences
Noyau
Ontologie
Partie du discours
Phrase
Segment textuel
SYGMART
SYGFRAN
Synonymie
Syntagme
Syntaxe
SYGserver
Grammaire
TALN
Taxonomie
Texte
Thésaurus
Vocabulaire contrôlé

Analyse morpho-syntaxique

Une analyse morphosyntaxique consiste à donner sur les éléments d'un texte des informations morphologiques (temps, genre, nombre…) et syntaxiques (nature, fonction…).Cette analyse est la base de toute application en traitement automatique des langues, naturelles ou non.

Annotation

- une information graphique ou textuelle attachée à un document et le plus souvent placée dans ce document ". [DESMONTILS et al, 2004] ;
- " bref commentaire ou explication d'un document ou de son contenu, ou même une très brève description, habituellement ajouté(e) en note après la référence bibliographique du document ".[GDT, 1983] ;

Dans notre approche, l'annotation est thématique. Elle consiste à insérer dans le texte des balises terminologiques spécifique à un domaine afin de constituer un étiquetage thématique structuré susceptible d'être exploité par d'autres applications.

Antidictionnaire

- Un Antidictionnaire est une liste de mots qui doivent être ignorés car considérés comme non pertinents dans le cadre d'une certaine application. Ainsi, une telle liste, dans le cadre d'une application visant la sémantique, comprendra les mots vides de sens comme les mots outils (pronoms, articles, …), dans un cadre distributionnel, les mots trop fréquents dans le corpus. (anglais : stop-list)
- Antonymie
- Deux items lexicaux sont en relation d'antonymie si on peut exhiber une symétrie de leurs traits sémantiques par rapport à un axe.
- Anaphore
- Un mot à valeur anaphorique ne peut être interprété que lorsqu'il est mis en relation avec un autre élément de l'énoncé. Par exemple, dans "En ce moment, le second attira de nouveau l'attention du capitaine. Celui-ci suspendit sa promenade et dirigea sa lunette vers le point indiqué, "celui-ci" est une anaphore de "capitaine".

AppliSyg

Moteur application SYGMART. Prend un fichier texte et génère l'arbre morpho-syntaxique parenthèsé selon un fichier de données spécifique à un langage. Les différents moteurs SYGMART sont en ligne sur http://www.sygtext.fr/

Catégorie grammaticale

cf. Nature

Constituants (d'une phrase)

En syntaxe, les constituants de la phrases sont les unités linguistiques qui composent la phrase : les mots et les syntagmes.

Contexte

Au niveau pragmatique, la situation dans laquelle se déroule l'énoncé.

Corpus

Un corpus est un ensemble de documents rassemblés dans une optique précise.

Dictionnaire

Un ouvrage de référence contenant l'ensemble des mots d'une langue ou d'un domaine d'activité généralement présentés par ordre alphabétique et fournissant pour chacun une définition, une explication ou une correspondance (synonyme, antonyme, cooccurrence, traduction, étymologie).

Distance de Levenshtein

Elle mesure la similarité entre deux chaînes de caractères. Elle est égale au nombre minimal de caractères qu'il faut *supprimer*, *insérer* ou *remplacer* pour passer d'une chaîne à l'autre.

Enoncé

Séquence de termes et de phrases en langue naturelle prononcée (appelée alors paroles ou énoncé oral) ou écrite (texte ou énoncé écrit) constituant un tout.

Etiquette

Item lexical.

Fichiers SYG

Ces sont des fichiers qui permettent de constituer un corpus pour SygAnnot. Leur contenu texte encodé en UTF-8 est issu de l'analyse grammaticale de SYGFRAN. L'analyse peut être exécutée depuis l'adresse suivante :
http://www.lirmm.fr/~chauche/ExempleAnl.html

Forme fléchie

Outre le lemme, un mot peut être représenté sous différentes formes. La forme fléchie d'un mot ou flexion est sa représentation "usuelle". Ainsi, il paraît naturel de décrire un corpus avec ce type de descripteurs issus du lemme.
Nous distinguons deux types de flexions, les verbales ou conjugaisons propres aux verbes et les nominales ou déclinaisons propres aux noms mais également aux adjectifs, articles et pronoms. Les formes fléchies sont caractérisées par un certain nombre de traits morphologiques pouvant être le genre, le nombre, le temps, le mode, etc. en fonction du type de flexion. Elles sont également caractérisées par un lemme. Ainsi, en définissant une flexion avec le lemme "faire" et les traits subjonctif présent, première personne du pluriel", nous obtenons la flexion "fassions".
Les mots sous forme fléchie comportent un radical et une ou plusieurs désinences. Les désinences sont les morphèmes porteurs des indications de nombre et de genre pour les noms, adjectifs et déterminants, de personnes, de temps et de mode pour les verbes. Ainsi, "lisions" est constitué du radical lis- issu de l'item 'lire', de la désinence temporelle 'i' et de la désinence personnelle -ons' [Lehmann, 1998] tandis que « rattes » est lui formé par rat (radical) + te (féminin) + s" (pluriel). En aucun cas, la flexion ne modifie donc la catégorie syntaxique.

Fréquence d'un terme

On appelle fréquence d'un terme (term frequency) le nombre de fois où ce terme apparaît, on parle aussi du nombre d'occurrences ou de la fréquence d'occurrence.

Fréquence d'occurrences

cf. fréquence d'un terme.

Gouverneur

Le noyau d'un syntagme. Dans le groupe « *La maison de Marie* », *maison* gouverne *Marie*.

Gouverné

Un satellite d'un syntagme. Dans le groupe *« les poissons de mer* », *mer* est gouverné par *poissons*.

Groupe

Un ensemble de termes.

Item lexical

Un item lexical est une suite de caractères formant une unité sémantique et pouvant constituer une entrée de dictionnaire. Par exemple, 'voiture' tout comme 'pomme de terre', 'moulin à vent' et même des termes techniques comme 'pompe bivalve à échappement central' sont des items lexicaux.

Lemmatisation

La lemmatisation désigne l'analyse lexicale du contenu d'un texte regroupant les mots d'une même famille. Chacun des mots d'un contenu se trouve ainsi réduit en une entité appelée lemme (forme canonique). La lemmatisation regroupe les différentes formes que peut revêtir un mot, soit : le nom, le pluriel, le verbe à l'infinitif, etc. La lemmatisation d'une forme d'un mot consiste à en prendre sa forme canonique. Celle-ci est définie comme suit :

- pour un verbe : ce verbe à l'infinitif
- pour les autres mots : le mot au masculin singulier.

On notera donc que toutes les entrées d'un dictionnaire sont lemmatisées

Lemme

Le lemme (ou lexie, ou item lexical) est l'unité autonome constituante du lexique d'une langue. C'est une suite de caractères formant une unité sémantique et pouvant constituer une entrée de dictionnaire.

Lexique

Un lexique est un ensemble de mots liés à un domaine (le lexique de l'armement), une personne (le lexique de Balzac) ou un ensemble de personnes (le lexique des jeunes). Il faut dans ce cas le comprendre comme une liste de termes. Il sera alors synonyme de vocabulaire, idiolecte, glossaire, dictionnaire, etc. En lexicologie cette confusion des sens courants n'est pas acceptable. L'objet du travail devant être un corpus d'unités attestées et considérées avant toute lemmatisation, on traite du vocabulaire de tel ou tel domaine.

En linguistique, le lexique d'une langue constitue l'ensemble de ses lemmes ou, d'une manière plus courante mais moins précise, « l'ensemble de ses mots ». Toujours dans les usages courants, on utilise, plus facilement le terme vocabulaire. Par métonymie (figure de style), un lexique est un recueil de termes dont le sens est expliqué.

Morphème

La plus petite unité porteuse de sens qu'il soit possible d'isoler dans un énoncé.

Mot ou lemme

Un corpus, écrit en langue naturelle indo-européenne, est constitué de mots afin de le décrire. Nous distinguons alors deux types de mots : les variables et invariables. Ces derniers peuvent être des adverbes, interjections, conjonctions ou prépositions. Nous nous focalisons sur les mots variables pouvant être des adjectifs, substantifs (ou noms), articles, pronoms et verbes. Les mots variables ont la propriété de pouvoir être déclinés ou conjugués (dans le cas de langues indo-européennes). Nous parlons alors de forme fléchie du mot. Notons que le “mot” tel que nous l’avons défini peut également se nommer lemme. Les lemmes sont en d’autres termes les entrées de dictionnaires.

Morphologie d'un item lexical

La morphologie d'un item lexical regroupe les informations concernant sa nature et son genre. Ainsi, 'courir' est un verbe, 'souris' est un nom féminin, 'orgues' est un nom masculin pluriel,…

Mot

Un mot est la forme fléchie d'un item lexical.

Nature

Les termes de même nature se caractérisent par la possibilité de les substituer syntaxiquement. Elle est constituée, entre autres, des verbes, des adjectifs, des noms, des adverbes. La nature est aussi appelée parfois catégorie grammaticale ou partie du discours.

Nombre d'occurrences

cf. Fréquence d'un terme

Noyau

Cf. Gouvernant

Ontologie

L'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations
En fait, un thésaurus ou même une taxonomie sont des formes d'ontologie dont la grammaire n'a pas été formalisée. Lorsque l'on établit une catégorie et une hiérarchisation de cette catégorisation, on établit des dépendances entre ces termes. Ces hiérarchisations ont un sens en dehors du vocabulaire lui-même. Par exemple, quand je dis ce terme est une sous-catégorie de cet autre terme Je viens de donner un sens à cette relation, je viens de dessiner une flèche entre les deux et j'ai qualifié la flèche en affirmant quel type de relation cela signifiait. Une ontologie correspond donc à un vocabulaire contrôlé et organisé et à la formalisation explicite des relations créées entre les différents termes du vocabulaire. Pour réaliser cette formalisation, on peut utiliser un langage particulier. Un des langages utilisés pour décrire les relations entre les différents termes d'un vocabulaire s'appelle RDF. Selon J.F. Sowa, l’inventeur des graphes conceptuels, une ontologie est un catalogue des types de choses supposées exister dans un domaine, du point de vue d'une personne utilisant un langage pour parler du domaine.

Partie du discours

cf. Nature. En Anglais, part of speech (POS).

Phrase

Ensemble autonome, réunissant des unités syntaxiques organisées selon différents réseaux de relations plus ou moins complexes appelés subordination, coordination ou juxtaposition. Le sens d'une phrase ne dépend pas seulement des mots (aspect lexical). L'organisation grammaticale y est aussi très importante : c'est l'aspect syntaxique.
Langage naturel (ou langue naturelle)
Langage tel qu'il est parlé quotidiennement par les êtres humains et qu'ils ont créé de façon émergente (comme le français, l'anglais, le chinois ou le malais) par opposition aux langages artificiels construits de façon consciente par l'être humain et utilisés en logique, mathématiques ou informatique.
fonction des conditions situationnelles et contextuelles dans lesquelles il apparaît. La pragmatique s'occupe en particulier des problèmes d'anaphore, de subjectivité.

Segment textuel

La classe des segments textuels regroupe les portions d'un texte ayant une unité sémantique mots, syntagme, phrase, paragraphe, texte,…

SYGMART

Introduit pour la première fois par Jacques CHAUCHÉ ([Cha84]), SYGMART est un système de transformation d'éléments structurés qui peut servir à diverses opérations sur les chaînes de caractères, entre autres leur analyse syntaxique. Il se présente sous la forme de trois sous-systèmes :
– OPALE réalise le passage entre les chaînes de caractères et les éléments structurés manipulés par le système ;
– TELESI réalise les manipulations d'éléments structurés ;
– AGATE réalise le passage d'un élément structuré au format de sortie souhaité pour le système (chaîne de caractère, fichier texte, fichier XML…)
Les "éléments structurés" manipulés par SYGMART sont des arbres multi étiquetés. Pour effectuer les transformations, SYGMART utilise, après les avoir compilés, un ensemble de règles de grammaire et des dictionnaires, fournis par l'utilisateur, et écrits dans le langage spécifique au sous-système qui les utilisera. Ces langages sont décrits dans le manuel de référence ([Cha01]), qui explique aussi plus finement le fonctionnement même des transformations, sans pour autant spécifier leur implémentation, dont il est plus particulièrement question dans [Cha84].
Une telle architecture en couches a plusieurs avantages, d'abord en permettant à l'utilisateur d'effectuer toutes les opérations qu'il souhaite effectuer avant la sortie finale en faisant simplement appel à TELESI autant de fois que nécessaires avec les grammaires de son choix, mais aussi en permettant de modifier ou d'ajouter une grammaire (et donc le traitement qui lui correspond) sans modifier les autres.

SYGFRAN

http://www.lirmm.fr/~chauche/SourcesAnalyse/
Analyseur d'énoncé français morpho-syntaxique écrit en SYGMART.
SYGFRAN utilise un ensemble de règles de transformations d'éléments structurés, mettant en oeuvre les règles de la grammaire française, qui permettent de transformer une phrase (texte brut) en un arbre syntaxique (élément structuré) enrichi d'informations sur les constituants. Cet analyseur a les avantages suivants :
– la rapidité : la complexité théorique d'analyse est en O(k * n * log2(n)) où k est le nombre de règles (12000 en décembre 2005) et n la taille du texte en nombre de mots. Il s'agit d'une limite supérieure, car l'analyseur étant structuré en plusieurs grammaires ordonnées, le facteur multiplicatif réel est beaucoup plus petit que k (en fait il est de l'ordre de 16). Cela dit, même ainsi, plus le texte est important, plus k est petit devant n. Aujourd'hui SYGFRAN analyse un corpus de 220000 phrases, d'en moyenne 25 mots, en environ 24 heures, sur un ordinateur grand public disposant d'un processeur cadencé à 2,4 Ghz et d'une capacité de mémoire vive de 1 Go ;
– la robustesse : SYGFRAN parvient, en décembre 2005, à obtenir une structure correcte pour au moins 35 % de l'ensemble des différents cas de syntaxe des phrases du français, pour les autres cas, SYGFRAN fournit une analyse partielle mais exploitable ;
– la production d'un arbre syntaxique : la plupart des systèmes actuels d'analyse syntaxique ne réalisent qu'un simple marquage linéaire, ceux qui produisent un arbre n'ont qu'une très faible couverture sur l'ensemble des constructions syntaxiques existantes.
SYGFRAN prend en entrée du texte brut et produit une structure parenthésée, correspondant à l'arbre morpho-syntaxique de chaque phrase du texte, dans laquelle de nombreuses variables sont renseignées sur les différents natures, fonctions syntaxiques, formes canoniques, catégories grammaticales, temps, modes, genres, nombres, etc. des constituants.
Par exemple, l'analyse de cette phrase produit l'arbre syntaxique de la figure 1. Les noms des noeuds internes (rectangles) correspondent aux natures desconstituants : PH pour PHrase, GN pour Groupe Nominal, GV pour Groupe Verbal,GA pour Groupe Adjectival, GNPREP pour Groupe Nominal PRÉPositionnel et GCARD pour Groupe CARDinal. Les noms des feuilles (ellipses) sont les formes canoniques des lexies (masculin, singulier, infinitif). Le numéro de chaque noeud est un pointeur sur les informations des variables SYGFRAN associées au noeud. Par exemple, le noeud 3 possède entre autres les variables et valeurs suivantes :

Lorsque SYGFRAN ne parvient pas à produire l'intégralité de la structure syntaxique d'une phrase ou d'un constituant c, il crée un noeud de nom « ULFRA », qui signifie unité linguistique française de nature indéterminée, auquel il ajoute, pour chaque sous constituant s de c, l'arbre syntaxique de s.

(Unknown Locution) comme nom du noeud père du mot « basé » pour exprimer son incompréhension. La structure de certains constituants reste tout de même correcte et peut donc être exploitée. Par exemple, l'arbre ayant pour racine le noeud 11, contient l'information que « sur cette phrase » est un groupe prépositionnel.

Synonymie

La synonymie est la relation sémantique qu'il existe entre deux items lexicaux qui diffèrent sur leur forme mais expriment le même sens ou un sens très proche.

Syntagme

L'ensemble structuré des termes représentant le sens d'un champ d'informations. Un groupe nominal est constitué d'éléments qui se rattachent tous à un nom noyau.
Groupe d'éléments constituant une unité sémantique. En grammaire moderne, on appelle syntagme (ou groupe), l'unité syntaxique plus ou moins complexe située entre la limite supérieure de la syntaxe, constituée par la phrase, et la limite inférieure, constituée par la catégorie simple (unité de base indissociable, ou élément terminal).
Dans la phrase « *Il a acheté une modeste maison de briques rouges.* », le groupe « *une modeste maison de briques rouges* » inclut dans ses éléments le syntagme inférieur « *de briques rouges* »

Éléments constitutifs du syntagme

Les éléments constitutifs du syntagme sont : d'une part le noyau, d'autre part un ou plusieurs satellites.
Un syntagme réduit à son seul noyau, c'est-à-dire sans aucun satellite, n'est plus vraiment un syntagme mais une catégorie isolée. Il est cependant possible (et parfois, pratique) de considérer qu'une unité isolée (nom, pronom, verbe, adverbe, adjectif qualificatif) est un syntagme à élément unique. En d'autres termes, « tout syntagme peut contenir de zéro à plusieurs satellites ».

Noyau du syntagme.

Le noyau (ou chef de groupe, ou support) est l'élément central d'un syntagme. Le noyau transmet toujours sa catégorie et sa fonction au syntagme dont il est le composant principal.
Le noyau d'un syntagme peut être une catégorie de base (mot simple ou mot composé), mais également un syntagme, c'est-à-dire, un sous-syntagme par rapport au syntagme de référence :
Une chemise en velours déchirée.
Dans ce syntagme nominal, le noyau est « *chemise en velours* », ce dernier étant lui-même un syntagme, ayant pour noyau le nom « *chemise* ».
Lorsque deux noyaux (ou davantage) sont présents dans un même syntagme, appartenant à la même catégorie et ayant la même fonction, ils sont dits parallèles. Deux noyaux parallèles sont, soit juxtaposés (c'est-à-dire, littéralement mis l'un à côté de l'autre), soit coordonnés, c'est-à-dire, reliés par un ou plusieurs coordonnants :
Il porte toujours un pantalon et une chemise parfaitement repassés.
Dans le syntagme nominal « *un pantalon et une chemise parfaitement repassés* », le noyau est double, constitué des syntagmes nominaux « *un pantalon* » et *« une chemise* », syntagmes coordonnés. L'unique satellite de ce syntagme est le syntagme adjectival « *parfaitement repassés* », épithète du double noyau.

Satellites du syntagme

Dans un syntagme, un satellite est un composant dépendant du noyau de ce syntagme.
À la place du mot satellite, certains grammairiens préfèrent parler d'expansion, de subordonné (avec un sens très général), ou encore, de complément (mais ce dernier terme pose un problème, puisqu'il est déjà employé en grammaire traditionnelle avec un sens précis).
Contrairement au noyau, tous les satellites d'un syntagme, sans exception, quelle que soit leur taille et leur aspect, peuvent appartenir à n'importe quelle catégorie (nom, article, conjonction…) :
Quel spectacle émouvant !
Le nom « *spectacle* » est le noyau de ce syntagme nominal. Les satellites en sont : l'adjectif exclamatif « *quel* » et l'adjectif qualificatif « *émouvant* ».
Jacques a aimablement invité Nathalie.
Le verbe *« a invité* » est le noyau de cette proposition (ou groupe verbal). Les satellites en sont : les deux noms propres « *Jacques* » (sujet du verbe) et « *Nathalie* » (C.O.D. du verbe), et l'adverbe « *aimablement* ».
Tout satellite peut donc être une catégorie ordinaire (mot simple ou mot composé), mais également un syntagme, c'est-à-dire, un sous-syntagme par rapport au syntagme de référence :
Une chemise bon marché.
Le nom « chemise » est le noyau de ce syntagme nominal. Les satellites sont : l'article indéfini « une » (mot simple) et l'adjectif qualificatif « bon marché » (locution adjectivale).
La voiture que j'ai achetée.
Le nom « *voiture* » est le noyau de ce syntagme nominal. Les satellites sont : l'article défini « *la* » (mot simple) et le syntagme verbal *« que j'ai achetée* » (proposition subordonnée relative).
Un même mot peut être noyau de plusieurs syntagmes concentriques :
Un gentil petit chat.
Ce syntagme nominal a pour noyau, non pas le nom « *chat* », mais le syntagme nominal « *petit chat* », ce dernier étant à son tour composé d'un noyau (le nom « *chat* ») et d'un satellite (l'adjectif qualificatif épithète « *petit* »).
À l'instar de ce qui ce passe pour le noyau, lorsque deux satellites (ou davantage) d'un même syntagme appartiennent à la même catégorie et ont la même fonction, ils sont dits parallèles. Deux satellites parallèles sont dits coordonnés s'ils sont réunis par un mot-outil (un coordonnant), et juxtaposés (c'est-à-dire, littéralement mis l'un à côté de l'autre), dans le cas contraire :
Tu as mangé de la salade, une pizza achetée au marché, des cerises que mon voisin m'a offertes.

Les trois syntagmes *« de la salade »*, *« une pizza achetée au marché »*, et *« des cerises que mon voisin m'a offertes »*, sont des satellites parallèles juxtaposés ; tous les trois sont des syntagmes nominaux, et leur fonction est C.O.D. du noyau verbal *« as mangé »*.
Le voisin dont je t'ai parlé et qui m'a offert des cerises, aimerait te rencontrer.
Les deux syntagmes verbaux *« dont je t'ai parlé »* et *« qui m'a offert des cerises »*, sont des satellites parallèles coordonnés (reliés par la conjonction de coordination « et »); tous les deux sont des propositions subordonnées relatives, et leur fonction est complément de l'antécédent « le voisin ».

Différents types de syntagmes

Seul un mot plein (ou à la rigueur, un pronom) peut être le noyau d'un syntagme, donc, selon la catégorie du noyau, on pourra distinguer seulement quelques types de syntagmes.

Syntagme nominal

Un syntagme nominal est un syntagme dont le noyau est un nom :
Le petit chien blanc de mon voisin a aboyé toute la nuit.
Le syntagme nominal « *Le petit chien blanc de mon voisin* » a pour noyau le nom chien.
Le courrier électronique est probablement le service le plus utilisé par les internautes.
Le syntagme nominal *« le service le plus utilisé par les internautes »* a pour noyau le nom service.

Syntagme pronominal

Un syntagme pronominal est un syntagme dont le noyau est un pronom :
A midi, nous avons mangé quelque chose de bon.
Le syntagme pronominal « *quelque chose de bon* » a pour noyau le pronom « *quelque chose* ».
Dans une préposition, le gouvernant du syntagme nominal est sujet.

Syntagme adjectival

Un syntagme adjectival est un syntagme dont le noyau est un adjectif qualificatif :
[J'ai un jardin] tout plein de roses odorantes.
Le syntagme adjectival « *tout plein de roses odorantes* » a pour noyau l'adjectif qualificatif « *plein* ».

Syntagme adverbial

Un syntagme adverbial est un syntagme dont le noyau est un adverbe :
Ils ont dû payer une amende conformément à la loi.
Le syntagme adverbial « *conformément à la loi* » a pour noyau l'adverbe « *conformément* ».

Syntagme verbal

Un syntagme verbal est un syntagme dont le noyau est un verbe. En conséquence, le syntagme verbal correspond, selon le cas, à une proposition ou bien à une phrase :
Il a travaillé courageusement toute la fin de semaine.
Le syntagme verbal « *Il a travaillé courageusement toute la fin de semaine* » a pour noyau le verbe « *a travaillé* ».
Lorsque le noyau d'un syntagme verbal est un verbe non conjugué, on peut préciser, selon le mode du verbe noyau : groupe infinitif, groupe participe présent ou groupe participe passé.

Syntagme propositionnel

Par extension on admettra que, conformément à la définition du syntagme, les propositions, parties de phrases représentant les propositions de type sémantique des éléments constitutifs du discours, soient désignées par l'expression syntagme propositionnel. La notion de syntagme propositionnel est nécessaire pour les travaux portant sur la cohérence du langage naturel.

Syntaxe

La syntaxe étudie la manière dont les mots se combinent pour former des syntagmes et les syntagmes se combinent pour former des phrases.

SYGserver

Utilitaire fournit par Alexandre LABADIE qui transforme l'arbre parenthèsé généré par AppliSyg en fichier XML plus simple à parcourir.

Grammaire

La grammaire est l'étude systématique des éléments constitutifs d'une langue. Par extension, on nomme aussi grammaire un manuel ou un ensemble de documents décrivant des règles grammaticales.

TALN

Traitement Automatique du Langage naturel (ou des langues naturelles**)**

Domaine d'étude des techniques d'analyse (compréhension) et de génération (production) automatiques d'énoncés oraux ou écrits.

Taxonomie

Dans une taxonomie, le vocabulaire contrôlé est organisé sous forme hiérarchique simple. Cette hiérarchisation correspond souvent à une spécialisation. Il existe donc un lien précis entre un terme du vocabulaire et ses enfants. Ce lien donne un sens supplémentaire, une signification. D'un vocabulaire contrôlé, on passe à un vocabulaire organisé.
Par exemple, dans une classification animale, nous aurons les vertébrés, invertébrés et puis sous les vertébrés nous auront les mammifères, les ovipares, etc. Tous ces termes nous permettront de classifier les animaux. On pourra donc dire que Les mammifères sont une sous-catégorie (sous-classe) des vertébrés.**)** Le terme taxinomie signifie littéralement, la "loi du rangement". Il désigne une classification systématique d'un ensemble d'éléments dans un domaine précis (taxinomie des êtres vivants, taxinomie de Flynn sur les architectures informatiques,...) ou général. Ce terme désigne aussi la science qui vise à établir de telles classifications.

Texte

Une succession de caractères organisée selon un langage. Il est exprimé par des phrases. La limite habituelle de la phrase est un signe de ponctuation : le point, mais également, le point d'exclamation, le point d'interrogation, les trois points de suspension (parfois, le double point, ou encore, le point-virgule). Par ailleurs, la première lettre de la phrase est obligatoirement une majuscule.

Thésaurus

Un thésaurus est une taxonomie qui fonctionne dans les deux sens. La taxonomie permettait d'obtenir une spécialisation des termes employés. Le thésaurus donnera de l'information sur les sujets connexes également. On pourra donc restreindre ou élargir le champ de connaissance. Cet élargissement se fait en donnant les termes relatifs. Des liens qui permettent la spécialisation, on pourra alors dire : c'est une sous-catégorie (spécialisation) ou est « relatif à » ou « voir également » (élargissement).Par exemple, imaginons une taxonomie qui organise l'information à propos des différentes races de vaches et chacune des sous branches de ces races. Une personne utilisant ce thésaurus voudra peut-être voir pendant sa recherche, explorer les types de fromages faits de lait de vache. Fromage ne fait pas partie de la taxonomie des vaches mais dans un thésaurus, celui-ci peut avoir un intérêt car le fromage est un des produits dérivés fait à partir du lait de vache. On a donc élargi le champ d'étude.

Vocabulaire contrôlé

C'est un ensemble de termes définis par un groupe (une communauté de pratiques) afin de pouvoir labelliser des contenus, écrire un document. La signification des termes n'est pas forcément définie et il n'y a pas nécessairement d'organisations logiques des termes entre eux. Par exemple, le glossaire d'un livre ou encore des catégories dans un système de carnets Web partagés entre différents auteurs.