Analyse automatique de textes comme point de départ d'un processus d'annotation

Sahbi Sidhom, Charles Robert, Amos David

▶ **To cite this version:**

Sahbi Sidhom, Charles Robert, Amos David. Analyse automatique de textes comme point de départ d'un processus d'annotation. Revue électronique internationale en technologies de l'information (e-TI), 2005. inria-00000265v2

HAL Id: inria-00000265

https://inria.hal.science/inria-00000265v2

Submitted on 20 Apr 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Analyse automatique de textes comme point de départ d'un processus d'annotation

Sahbi SIDHOM *MCF & Chercheur de l' équipe SITE du LORIA*

Charles ROBERT *Doctorant de l' équipe SITE du LORIA*

Amos DAVID *Professeur & Responsable de l' équipe SITE du LORIA*

LORIA - Université Nancy2, BP. 239, 54506 Vandoeuvre cedex - France.

Sahbi.Sidhom@loria.fr **Robert@loria.fr** **Amos.David@loria.fr**

Résumé

L'analyse automatique de textes a élargi la perspective de travail sur les contenus en ouvrant les études sur les productions langagières : l'annotation est un cas d'étude. Cette dernière est définie comme une information textuelle, graphique ou sonore, attachée à un document.

Notre contribution se distingue par une différenciation de la *représentation de l'annotation* comme valeur ajoutée à un contenu, le contenu lui-même du document et les *informations obtenues* par un système de recherche d'information. C'est une nouvelle approche dans la conception d'un système d'information dédié à l'intelligence économique. On se donne comme objectif de présenter des modélisations concurrentes afin de faciliter l'appariement entre requêtes d'interrogation et sources documentaires dans un processus de recherche d'information en tenant compte de la pertinence des résultats. La validation de la pertinence des résultats et leur fiabilité sont pondérés aux besoins et aux centres d'intérêts des utilisateurs finaux : les décideurs.

Mots-clés

Processus d'annotation, analyse automatique, gestion de connaissances, spécification de concepts, classification, système de recherche d'information (SRI), concept, connaissance, syntagme nominal (SN), attribut, valeur.

Abstract

Automatic text analysis widened the prospect for work on document contents by opening up the studies on the linguistic productions: annotation is a case of study. The latter is defined as textual, graphic or sound information, attached to a document. Our contribution is characterized by a differentiation of annotation representation as added value to contents, the contents of the document itself and the information obtained from an information retrieval system. It is a new approach in information system design dedicated to "economic intelligence". Our objective is to present concurrent models in order to facilitate matching between requests and document sources in a information retrieval process by taking into account of relevant results. Validation of relevant results and their reliability are weighted with the needs and the interest centers of the targeted end-users: decision makers.

Keywords

Annotation process, automatic analyse, knowledge management, concept specification, classification, information retreival system (IRS), concept, knowledge, noun phrase (NP), attribute, value.

1. Introduction

Les traitements informatiques que l'on peut appliquer aux documents ont élargi le champ de l'analyse sémantiques de contenus. La linguistique computationnelle se doit de reconsidérer la notion d'interprétation sous cette nouvelle perspective. Il importe de proposer une nouvelle approche qui englobe l'ensemble des productions textuelles (textes existants ou à rajouter aux documents) afin de permettre l'élaboration des méthodes d'analyse automatique déterministes.

L'approche de l'analyse automatique de textes a élargi la perspective de travail sur les contenus en ouvrant au champ de cette recherche les études sur les productions langagières et les applications liées au caractère utilitaire de la linguistique computationnelle.

Deux perspectives peuvent être dégagées, qui orientent les recherches dans deux directions apparemment différentes : la première, directement applicative, s'attache à la mise au point d'outils de reconnaissance automatique ; la seconde intéresse plus les chercheurs et ne vise pas tant à classer qu'à extraire, qui pouvait rapprocher les documents multimédia aux éléments sémantiques (unités de connaissance) par des éléments de contenus textuels. Derrière ces deux perspectives, les deux visées s'opposent, mais elles se retrouvent confrontées au même problème, celui de la mise en évidence de traits discriminants permettant d'identifier les unités textuelles. Ces dernières par l'apport de la linguistique computationnelle seront formellement identifiables pour donner accès à une interprétation sémantique, soit en phase d'analyse soit en phase de production de contenus.

La reconnaissance automatique des unités de connaissances [BACHIMONT, 1999] intéresse de près les acteurs du domaine de l'ingénierie documentaire que celui de l'intelligence économique. Tout d'abord, dans le cadre d'une recherche d'information à l'intérieur d'une source de documents, qu'elle soit interne (base locale, plus ou moins enrichie sémantiquement), soit externe (banque de données, enrichie graduellement) ou quasi-insondable (le Web, enrichie continuellement), un utilisateur doit pouvoir disposer d'***indicateurs*** sur le document et son contenu, doit pouvoir ***filtrer*** des textes ou des objets non textuels sur des critère générique ou doit pouvoir disposer du moyen d'***annoter*** ses résultats de recherche. Ces moyens (indicateurs, filtrages informationnels et annotations) seront une valeur ajoutée incontestable aux contenus des sources documentaires. Les outils d'analyse, d'indexation, d'annotation et de recherche documentaire seront incontournables sur le marché stratégique des entreprises. Tout autant que les moteurs de recherche, sur le Web, seront en mesure à prendre en compte le modèle utilisateur et ses interactions en phase de formalisation ou d'exploitation de ses besoins informationnels.

Un regard critique du mot « *annotation »* implique deux connotations. C'est un objet (le contenu d'annotation) aussi bien qu'une action (l'annotateur dans le processus de mise en valeur d'un document). Pour cette étude, premièrement, l'annotation est orientée vers l'acte d'interprétation d'un objet documentaire. Dans ce cas, l'annotateur est le producteur de l'objet et son action consiste à interpréter un document. Deuxièmement, l'annotation est un objet (écrit, sonore ou graphique) attaché au document source.

Notre contribution se distingue par une différenciation de la *représentation de l'annotation* comme valeur ajoutée à un contenu (document, requête ou autre) aux *informations obtenues* par un système de recherche d'information (SRI), qui sont susceptibles d'interagir avec des modèles dédiés à l'IE [MARTINET, 1995], [MARTRE, 1994], en terme d'information pertinentes demandées dans un processus décisionnel. En vertu de la présentation de ces deux problématiques (annotations, recherche d'informations), nous présenterons dans ce qui suit l'importance de l'annotation qui inclut des visées sémantiques pour la recherche d'information.

2. Présentation générale de ce qu'est l'annotation

L'annotation et les outils qui lui sont dédiés deviennent de plus en plus indispensables dans les processus de recherche d'information et principalement dans la validation des résultats obtenus (adéquation aux besoins informationnels, pertinence, etc.). Ce qui permet aux individus d'accomplir des interprétations contextuelles ou hors contextuelles sur les contenus de documents.

Plusieurs significations ont été attribuées à la définition d'une annotation. Les plus complètes accordent les interprétations suivantes :

- " *une annotation est 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] ;

– " *any object that is associated with another object by some relationship*". [W3C-Annotation, 2004]; ie. Tout objet qui est relié à un autre par un certain rapport (ou relation à définir).

Les annotations se caractérisent selon différentes dimensions. Ces dimensions donnent accès aux propriétés de l'annotation à plusieurs niveaux : sa structuration, sa fonction et son rôle dans la communication entres les individus { lecteur + rédacteur, lecteur + (autre) lecteur, rédacteur + (autre) rédacteur }.

2.1. Composition de l'annotation

Une annotation regroupe essentiellement trois éléments principaux, à savoir :

- l'annotateur, la personne qui réalise l'annotation
- le document source concerné par l'annotation
- les objets d'annotation introduits sur le document

Dans cette étude, nos soucis ne portent pas sur la modélisation de l'utilisateur ni sur son profil dans un processus d'annotation, mais plutôt sur le contenu de l'annotation et ses fonctions représentatives comme valeur ajoutée au document (*Fig. 2.1*). Dans sa représentation la plus générale, un document est une trace de l'activité humaine [PRIE, 1999], considération que nous la retenons d'un point de vue de l'effort intellectuel humain pour représenter des faits, des connaissances et des savoir-faire. De ce point de vue, les traces de l'activité humaine peuvent inclure des sources diverses comme des matériaux archéologiques, des édifices, des œuvres cinématographiques, des manuscrits, des gravures, des monuments d'art, etc.

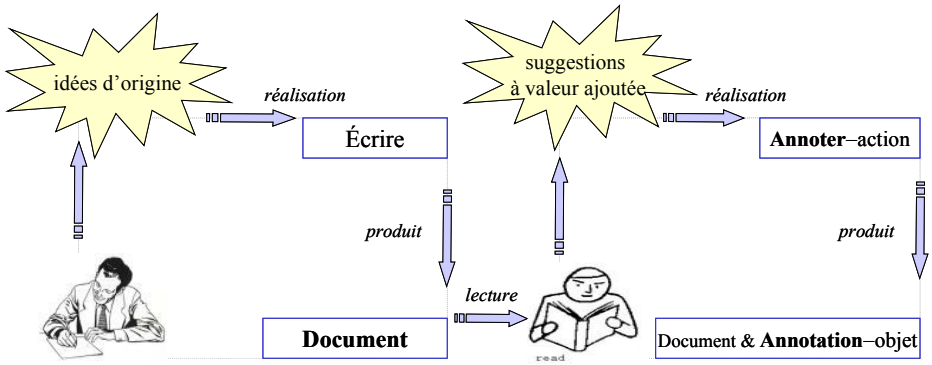


Fig. 2.1: Position de l'annotation dans la chaîne de production d'information.

Sur un document, cette trace contient essentiellement des informations et est conçue dans l'objectif d'être interprétée ou expliquée par les cinq sens humains : le toucher, l'odorat, le goût, l'ouïe et la vue. C'est par ses sens que l'Homme reçoit et transmet la connaissance du monde physique : la perception des concepts contenus dans ce document que le sens de la parole peut le rendre accessible à son public potentiel. C'est ainsi que le document aussi bien l'annotation peuvent prendre une forme diverse : textuelle, orale, graphique, filmique, etc.

A part de la forme de présentation d'un document, les annotations sur ce dernier prendront habituellement une forme complémentaire ou parfois différente de la source. La différence ou la complémentarité parviennent des concepts et de leurs attributs à inclure en terme de valeur ajoutée interprétables : compréhensions aisées par l'Homme et disposition à l'analyse automatique.

Le document à annoter peut associer de nouvelles entités (ou éléments d'annotation) comme les ponctuations, les mots, les images, les éléments terminologiques, les phrases, les passages, les liens, les mises en forme typographiques, les séquences audio ou vidéo,… : une collection d'annotation par des éléments homogènes ou hétérogènes.

2.2. Importance de l'annotation

L'idée de Vannevar Bush dans sa communication "As We May Think", en juillet 1945, dans "The Atlantic Monthly" [BUSH, 1945] était d'un intérêt fondé sur le travail collaboratif et qui reste d'actualité. 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. Mais, l'information pertinente cherchée est fortement distribuée : la source d'information est volumineuse, évolutive, volatile, très "bruitée", très hétérogène et souvent très peu structurée (ie. dynamique du Web).

L'annotation pouvait s'adresser à l'auteur lui-même, à l'annotateur ou au public intéressé à la fois par le document et ses annotations. Nous considérons ainsi que les annotations rapportés au document, dans un processus de recherche d'information, lèvent les ambiguïtés sur des éléments d'information présents dans le contenu. Elles 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.

Les objectifs d'annotation ne sont pas toujours liés aux questions de collaboration, ils peuvent inclure des visées sémantiques pour la recherche d'information (analyse, indexation, filtrage, etc.), à savoir :

- construire une représentation externe du texte source,
- introduire des éléments d'évaluation sur le document : témoignage, apport, constat, démonstration, réfutation, etc. ,
- permettre une prise de vue indépendante de celle de l'auteur,
- fournir une traçabilité d'exploitation du document,
- accumuler des commentaires explicites sur le contenu,
- favoriser le raisonnement critique,
- partager l'information,
- filtrer l'information,
- faciliter la compréhension et la relecture d'un document,
- insérer du marquage sémantique (symbolique ou alphabétique) au contenu du document.

2.3. Annotation dans la recherche d'information

Quelques outils d'annotation ont suscités notre intérêt comme *CritLink* [YEE, 2002] , *YAWAS* [DENOUE, 1999], *Commentor* [OVSIANNIKOV & al, 1999] et Nestor [Zeiliger, 2001]. Ces derniers sont conçus dans l'objectif de promouvoir les travaux collaboratifs sur Internet par marquage ou par ancrage des annotations (couleurs, soulignements, phrases, mots, …) sur le document source [HECK & al., 2003].

Dans notre travail, nous observons l'utilisation des annotations dans un processus de recherche d'information. Principalement, notre objectif s'oriente vers l'exploitation des annotations pour déterminer des sources informationnelles pertinentes.

Le modèle d'annotation que nous proposons, AMIE (ie. « ***A**nnotation **M**odel for **I**nformation **E**xchange* »), est la conjonction de paramètres caractéristiques à l'annotation de documents par rapport à un système de recherche d'information (SRI). Les processus d'annotation mis en œuvre s'associe avec le système SIMBAD (ie. « ***S**ystème d'**I**ndexation du **M**ultimédia **B**asé sur l'**A**nalyse de contenu **D**ocumentaire », cf.* §3.) dans la communication et l'analyse des informations. Ainsi, AMIE est amené à préciser certains paramètres et fonctions, comme le contexte d'annotation, l'annotateur, le document à annoter et les fonctions d'annotation (*Fig.2.3*).

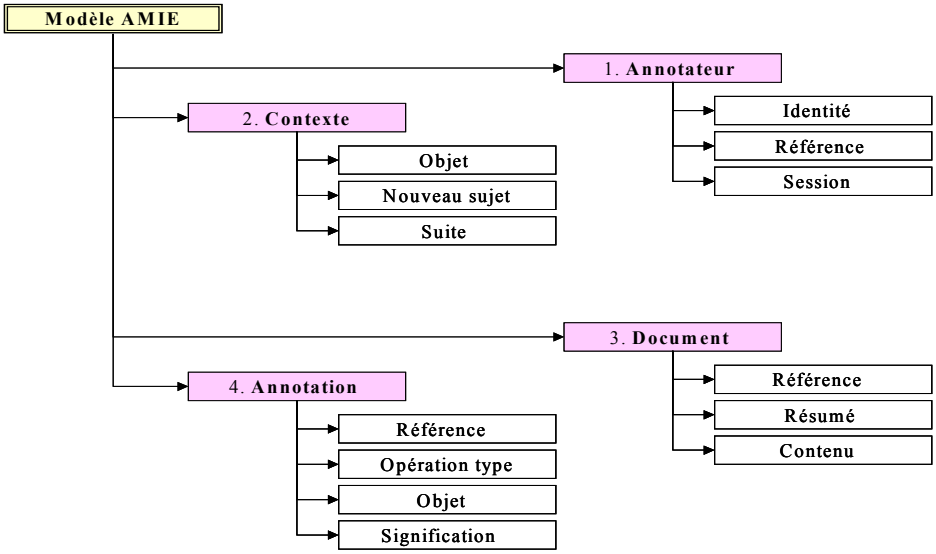


Fig.2.3 : Le modèle AMIE et ses classes types : Contexte, Annotateur, Document, Annotation.

Lors de la conception du modèle AMIE, nous nous sommes posés la question suivante :

- Est-ce qu'une annotation est équivalente à une indexation ?

Nous avons observé que l'annotation est plus qu'un processus d'indexation. Dans le cas d'une indexation automatisée, il n'y a pas d'interprétation (subjective) alors que, dans le cas d'une indexation intellectuelle, souvent ce processus atteint une forme de stabilité interprétative de l'indexeur humain. Ce qui conduit l'indexeur à définir des grilles d'analyse et des règles « normalisées » (par des consensus ou par partage d'expériences avec sa communauté : les indexeurs) dans le but de décrire ou de donner des informations sur le contenu et non de l'interpréter. Ainsi, par extension du processus d'indexation, l'annotation peut s'établir comme une indexation du contenu augmentée d'un facteur lambda (λ). Ce λ en plus représente *le contexte interprétatif de l'annotateur* (intérêt porté sur le document, le sujet, le domaine, etc.) et *le contexte interprétatif de l'annotation* (mise en valeur du contenu par de nouveaux éléments informationnels). Ce qui semble important dans ces deux contextes interprétatifs est que nous nous retrouvons dans des théories duelles : les dernières indications sur annotation et annotateur montrent de manière évidente à quel point la distance est grande entre la *sémantique cognitive* et le *paradigme de la référence directe*. Une idée centrale et commune entre ces deux contextes se détermine dans le fait suivant :

> *la référence de mots, dans une annotation comme '**peugeot**' ou '**voiture**', est déterminée non pas par le concept de peugeot ou de voiture qu'un annotateur aurait ' à l'esprit ', mais par sa position objective dans le monde réel, et par la nature de celle-ci : contexte de la référence et besoins informationnels. La référence de mots dans le contenu du document peut suggérer à l'annotateur des informations à valeur ajoutée, comme : « le conducteur de la **peugeot** 307sw », « la **voiture** d'Amos DAVID », etc.*

Pour conclure, deux propositions concurrentes sont alors définies sur l'annotation, sur ses énoncés et sur ses valeurs sémantiques, à savoir :

- L'intension annotative : c'est une logique d'annotation sans référentiel sur l'usage de l'annotation (sans objectifs) et sans classes d'usager (veilleur, décideur). C'est un plus sur le contenu, des mises en forme, des indications ou signes rajoutés, …
- L'extension annotative : c'est une logique d'annotation avec référentiel où on peut envisager une classe d'objets (mots-clés, concepts, termes, indicateurs, …) et la possibilité de déterminer des attributs et des valeurs dans l'annotation. C'est une association entre le contenu et le contenant, c'est les thèmes et les parties du document, c'est des formes de synthèses ou de résumés, …

Dans ce travail, nous expliciterons nos propositions sur les aspects logiques d'annotation et sur la présentation des fonctions sémantiques de l'annotation (*cf.* §4.), dans le modèle AMIE.

Comme prévision au processus d'annotation, l'analyse morphosyntaxique automatique a été construite pour démontrer des relations syntagmatiques typiques dans des contenus textuels, leurs compositions morphologiques et syntaxiques, afin de traduire les éléments informationnels en éléments conceptuels pour représenter des connaissances à extraire ou à chercher. Ce cadre de recherche fera l'objet de la problématique suivante.

3. Analyse automatique pour l'extraction de concepts

Pour l'extraction de concepts dans un processus d'analyse ou de recherche d'information, nous disposons d'un système complet d'ingénierie linguistique, de gestion et management des connaissances.

Pour l'ingénierie linguistique, le modèle mis en application [SIDHOM & HASSOUN, 2003] permet la reconnaissance et l'extraction automatiques des syntagmes nominaux (SN). Les occurrences SN sont des objets de la réalité extra-linguistique (ou objets de l'univers).

Ce modèle met en évidence les transitions entre les constituants morphologiques (ou prédicats libres) et les relations entre constituants pour former les SN (ou prédicats liés).

Ce processus de transitions s'effectue à travers l'identification de structures syntaxiques dans le texte du document, en repérant les SN. il a été conçu en ayant les objectifs suivants [SIDHOM, 2002] :

- identifier les SN dans les contextes d'analyse : textes des résumés, des annotations ou des requêtes,
- déterminer les structures SN en mettant en évidence les relations entre ses constituants,
- permettre l'entreposage des représentations SN puis augmenter leurs fonctionnalités pour la recherche d'information,

– formaliser le mécanisme de passage de la logique intensionnelle (qui englobe les prédicats libres) à la logique extensionnelle (les occurrences SN ou prédicats liés).

3.1. Approche logico-sémantique du modèle SN

Dans l'approche logico-sémantique du modèle SN, ce dernier est en effet *« l'unité minimale du discours qui permet de désigner un objet* [LE GUERN, 1989]*»*. Nous sommes donc confrontés à deux ordres logiques [LE GUERN, 1991] afin que le SN se définisse comme étant *la plus petite unité d'information porteuse d'une valeur référentielle*, à savoir :

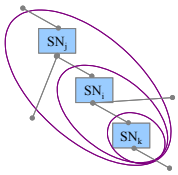
– La logique intensionnelle : une logique sans référentiel et sans classe, qui est constituée de relations entre les mots et de propriétés du mot envisagées indépendamment de quelque objet que ce soit.
– La logique extensionnelle : le mot « prédicat libre » prend ses valeurs sur un univers du discours, là on peut envisager une classe d'objets et la possibilité de déterminer des classes, au moins virtuelles avec la mise en relation des mots et des objets.

Le mécanisme d'analyse automatique des textes s'est concrétisé par la conception d'un noyau d'indexation automatique. Le noyau d'indexation se scinde en trois modules :

– la conception de différents outils automatiques servant à l'analyse morphologique du langage naturel. Il s'agit des ressources linguistiques composées essentiellement par les dictionnaires électroniques et la grammaire de réécriture des éléments syntagmatiques et celle du modèle de la phrase résumé ;
– l'implémentation de l'analyseur morphosyntaxique par la compilation des règles de réécriture, pour l'analyse des corpus textuels ;
– l'extraction des SN à partir des arbres syntagmatiques des phrases analysées. L'architecture de l'analyseur est fondée sur les automates à transitions augmentées et en cascade (ATN et CATN) de W. Woods [WOODS, 1980-1997] : c'est à partir des objets décorés (arbres) que le filtrage automatique des SN s'opère.

3.2. Formalisation de concepts par les syntagmes nominaux

Les syntagmes nominaux ont une organisation naturelle. Dans un sens, ils ont un rapport d'*emboîtement* les uns avec les autres, ce qui permet de les classifier linéairement en des niveaux informationnels distincts :

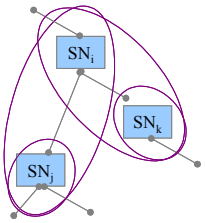
Logique d'emboîtement	**Conditions**	**Schéma : graphe linéaire**
$\forall SN_i, \exists \{ SN_j , SN_k \} /$ $(SN_j \supset SN_i \supset SN_k)$;	*avec :* $SN_i = \{SN : \neq\varnothing \mid \neq saturé\}$; $SN_j = \{SN : \neq\varnothing \mid saturé \mid \neq saturé \}$; $SN_k = \{ SN : \neq\varnothing \mid \neq saturé\}$.	

En exemple :
« M.Clinton a aussi évoqué le résultat du référendum en France ». (réf. FR3, JT. publié le 06/06/05 à 21:27)
= « [M.Clinton]^{SN} a aussi évoqué [le résultat de [le référendum en [France]^{SN_1}] ^{SN_2}] ^{SN_3} ».

Ainsi : $SN_3 \supset SN_2 \supset SN_1$;

avec :
SN_3 = « le résultat du référendum en France », *qui est un SN saturé ;*
SN_2 = « le référendum en France », *qui est un SN non saturé ;*
SN_1 = « France », *qui est un SN non saturé ;*

Et dans l'autre, ils ont un rapport de ramification, dans le cas où le syntagme nominal se retrouve dans un schéma arborescent. Cette dernière propriété permet d'ordonner et de distinguer les classes d'informations (structure d'arbre des classes d'informations) :

Logique d'arborescence	Conditions	Schéma : graphe d'arbre
$\forall SN_i , \exists \{ SN_j , SN_k \} /$ $(SN_i \supset SN_j) \wedge (SN_i \supset SN_k) ;$	*avec :* $SN_i = \{SN : \neq\varnothing \mid saturé \mid \neq saturé\} ;$ $SN_j = \{SN : \neq\varnothing \mid \neq saturé\} ;$ $SN_k = \{SN : \neq\varnothing \mid \neq saturé\}.$	

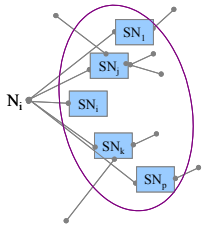
En exemple :
« l'annonce contre la DCA allemande du débarquement allié du 6 juin 1944 ». (réf. ouest-france.fr, 06/06/05)
« [l'annonce contre [la DCA allemande]$^{SN}{}_{1g}$ de [le débarquement allié de [le 6 juin 1944]$^{SN}{}_{1d}$]$^{SN}{}_{2d}$]$^{SN}{}_{max}$ ».

Ainsi:
$(SN_{max} \supset SN_{2d} \supset SN_{1d})$
$\wedge$
$(SN_{max} \supset SN_{1g}) ;$

avec :
SN_{max} = « *l'annonce contre la DCA allemande du débarquement allié du 6 juin 1944* » ;
SN_{1g} = « *la DCA allemande* » ;
SN_{1d} = « *le débarquement allié du 6 juin 1944* » ;
SN_{2d} = « *le 6 juin 1944* » ;

Ces caractéristiques logiques permettent de construire une architecture de gestion des connaissances, qui exploite les informations autour du SN, au moyen de la navigation dans des structures d'arbres ou des emboîtements. Par la superposition de ces deux logiques, la navigation entre les SN s'intègre dans une architecture d'un treillis de connaissances.

A ces deux aspects logiques de navigation, le centre nominal (N) dans un SN est un élément appartenant à la logique intensionnelle. N ne peut construire un objet de discours, mais comme trait d'une classe pour accéder à ses éléments. Ce prédicat libre N va contribuer à la description d'une classe d'objets <SN> : nous le situons comme la clé d'accès à cette classe ayant comme attribut commun le trait <N>.

Logique d'appartenance	Conditions	Schéma : classe de SN
$\forall SN_i , \exists! \ N_i \ /$ $(N_i \in SN_i)$ $\wedge$ $(\exists \{SN_{1..k}\} / \ N_i \in \{ SN_{1..k} \}) ;$	*avec :* $SN_i = \{SN : \neq\varnothing \mid saturé \mid \neq saturé\} ;$ $SN_{1..k} = \{SN : \neq\varnothing \mid saturé \mid \neq saturé \}.$	

En exemple :
« un convoi de troupes américain est attaqué par les escadrilles japonaises ». (réf. INA.archives, 06/06/05)
= « [un convoi de [des troupes] américain] est attaqué par [les escadrilles japonaises] ».
Ainsi : les $\{ N_i \in SN_i \}$

convoi $\in$ un convoi de troupes américain $\supset$ des troupes
troupe $\in$ des troupes
escadrille $\in$ les escadrilles japonaises

3.3. Application

Dans ce contexte, l'indexation d'un document (par sa représentation textuelle) ou d'une requête passent par le même mécanisme d'analyse dans SIMBAD.

> SIMBAD (ie. « ***S**ystème d'**I**ndexation du **M**ultimédia **B**asé sur l'**A**nalyse de contenu **D**ocumentaire* ») pour l'indexation de documents multimédia à partir de leurs représentations textuelles associées (contenus résumés). Il regroupe principalement le *parseur morphosyntaxique* qui est associé à des outils dictionnairiques électroniques, le noyau d'indexation des contenus et le système de recherche d'information pour la gestion des connaissances autour du syntagme nominal (SN).

Ainsi, le schéma d'interrogation (*cf.* Figure 3.2.), dans le système de recherche d'information, consiste à retrouver les SN d'une requête (SN-requête) qui sont présents dans la base de documents (SN-base). Bien entendu, les documents qui répondent le mieux à la requête sont ceux identifiés par des SN saturés, bien moins que par ceux identifiés par les SN non-saturés, et moins par ceux identifiés par les de prédicats intensionnels (N) ou leur synonyme (N_s).

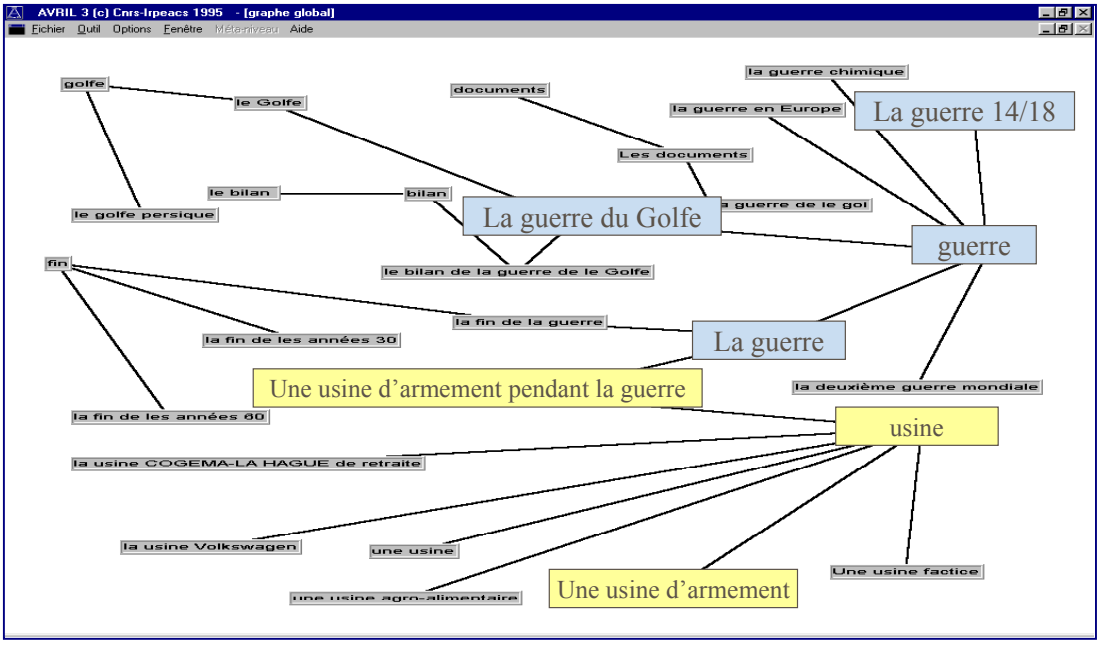


Fig. 3.3 : Réseau de concepts autour d'un thème portant sur la guerre.

L'approche logico-sémantique du modèle SN a fondé les représentations sur les thèmes, les concepts clés et leurs relations dans les contenus. En application, SIMBAD a permis de faciliter l'appariement entre requêtes d'interrogation et sources documentaires lors du processus de recherche d'information. Par le biais de cette approche, nos propositions de modélisation se prolongent au processus d'annotation pour structurer ses éléments pour qu'ils deviennent une source d'alimentation pour l'outil d'analyse automatique. C'est l'objet de la section suivante.

4. Spécification de concepts pour le processus d'annotation

4.1. Problématiques

L'objet de cette partie concerne la structuration de l'annotation dans un processus de recherche d'information et les fonctions sémantiques à associer au contenu de l'annotation.

Les dimensions du problème se heurtent à trois problématiques, à savoir :

- la dimension *formalisation* : les annotations sont amenées à être structurées complètement ou semi-complètement afin d'identifier les attributs et les valeurs d'un problème de recherche d'information (requête d'interrogation) ou valoriser la pertinence des résultats obtenus d'un système de recherche d'information (pertinence de documents) ;
- la dimension *explicitation* : une annotation ne suffit pas à elle-même. Elle est souvent destinée à une ou plusieurs personnes. Donc, elle nécessite des adaptations au profil de son usage avec les

interprétations non ambiguës, comme les conventions adoptées (langue cible, liste de mnémonique, liste de valeurs, liste d'indicateurs, liste de symboles, table de lecture,…) ;

- la dimension *traduction* : l'annotation au niveau de sa structuration doit intégrer les propriétés de son rôle dans la communication, entre l'annotateur (producteur de l'annotation) et le prospecteur (exploiteur de l'annotation). Ce dernier, dans le contexte d'intelligence économique, est un agent humain (veilleur ou décideur), qui s'outille dans son exploitation par un agent logiciel (plate-forme et outils informatiques).

4.2. Application

Nous attribuons à ses dimensions une classe : la classe de la structuration de l'annotation. Des règles de constructions de cette classe peuvent être proposées ou « normalisées » lors de l'implémentation d'outils [DAVID & THIERY, 2002-2003] qui vont se greffer sur le document dès son ouverture afin de faciliter la tâche de l'annotateur et préserver la sémantique des annotations.

Une deuxième classe, que nous explicitions dans ce travail, concerne les fonctions à attribuer dans un processus d'annotation. Elles sont regroupées dans l'annotation manager. Les fonctions ainsi définies (*cf.* Fig.2.3) sont :

- la fonction contexte d'annotation : création d'une nouvelle annotation, le suivi d'une ancienne annotation ou l'objet d'une annotation (requête, interprétation,…) ;
- la fonction document à annoter : spécification portant sur le document à annoter ;
- la fonction annotateur : description du profil de l'annotateur (explicitement ou implicitement) soit par ses recherches dans la base d'annotation soit par les alimentations apportées à la base ;
- la fonction annotation : c'est la principale fonction dans cette classe qui comporte les opérations types, leurs objectifs dans le processus et leurs significations ;

Nous développons dans ce travail des schémas XML pour représenter l'annotation d'un document. Dans ce dernier (*cf.* Fig.4.1), il s'agit de la représentation de l'annotateur qui fait référence au document et aux éléments d'annotation.

L'avantage dans cette modélisation sur l'annotateur réside dans les représentations suivantes :

- un annotateur peut annoter plusieurs documents ;
- un annotateur peut annoter un document plusieurs fois ;

Par cette modélisation, la relation peut s'établir entre les systèmes AMIE et SIMBAD par une alimentation de concepts obtenus par l'analyse automatique de textes (textes de documents ou d'annotations). Ce dernier par son approche sur l'identification des SN dans les contenus textes permet d'alimenter explicitement les classes d'objets (les éléments N dans les SN) et les attributs ou valeurs (les SN et les SN en relation avec d'autres).

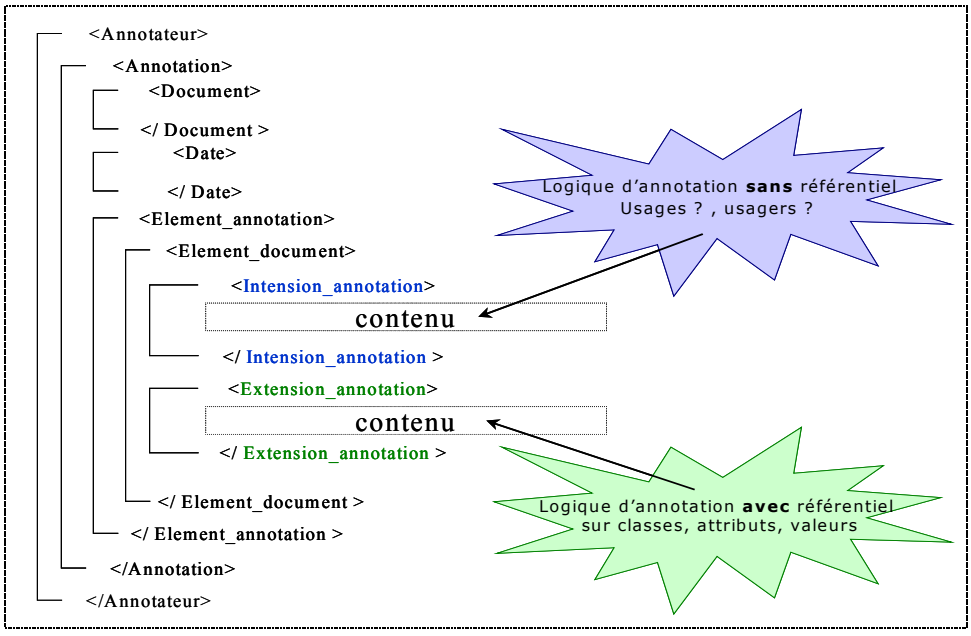


Fig.4.1 : Schéma d'annotation d'un document.

Dans le processus d'annotation, nous avons défini des fonctions sémantiques qui permettent de répondre aux trois problématiques posées : formalisation, explicitation et traduction.

Pour la fonction sémantique « 4. Annotation », nous avons défini un ensemble d'opérations spécifiques (corriger/ajouter/inscrire, souligner/marquer, réponse/question, etc.) pour insérer leurs résultats soit dans le contenu intensionnel de l'annotation soit dans celui extensionnel.

Dans la sous fonction « Opération type », les opérations définies sont attribuées à l'aspect intensionnel de l'annotation, comme « marquer » une section ou une image dans un document puis « ajouter » un commentaire explicatif sur le contenu marqué [MATTHEW & al., 1996] . Cet ajout explicatif de l'annotateur n'impose pas l'usage qui se fera de son commentaire et ni l'usager qui l'utilisera.

Dans la sous fonction « Signification », les opérations définies sont attribuées à l'aspect extensionnel de l'annotation, comme « structurer » les annotations dans la partie intensionnelle pour constituer les éléments en « Classe » ou distinguer les éléments qui peuvent servir d' « Attribut » à ceux qui peuvent l'être comme « Valeur » et ainsi évoluer vers une équation de recherche (*cf.* Fig.4.2).

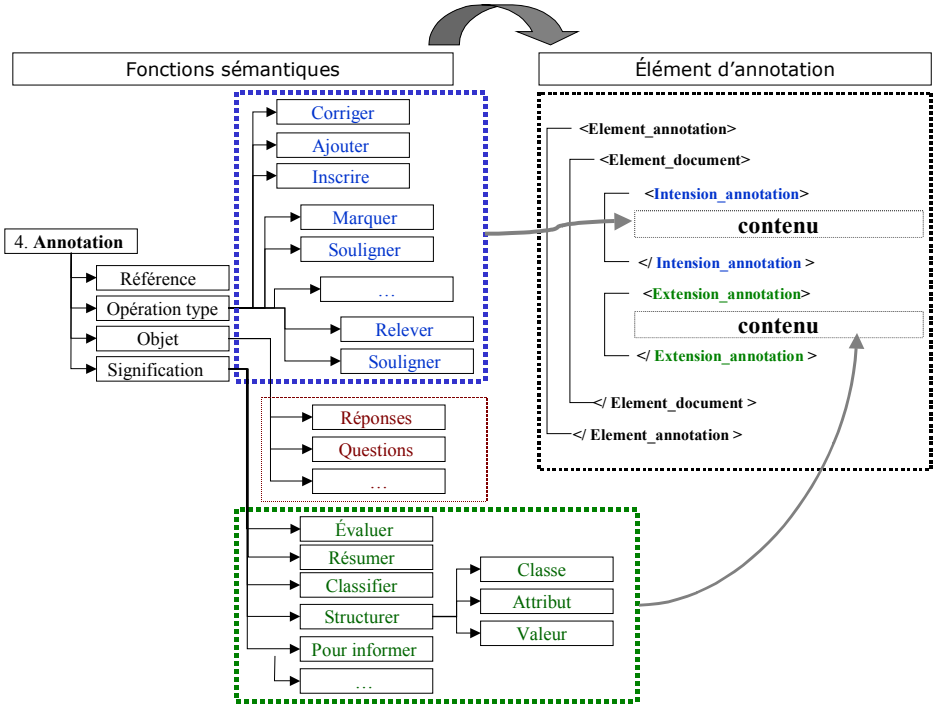


Fig.4.2 : Fonctions sémantiques de l'annotation.

En relation avec les outils d'analyse automatique (SIMBAD), l'affectation d'un SN comme « Attribut » ou son assignation comme « Valeur », reste du domaine de compétence de l'annotateur et son pouvoir interprétatif. Dans un contexte d'IE, un veilleur aura le moyen selon la demande d'un décideur (besoins informationnels) à traduire le problème décisionnel en un problème de recherche d'information [FROISSART, 2001] en s'appuyant sur les deux propositions logiques d'annotation (intensionnel et extensionnel).

L'architecture du modèle AMIE se retrouve entraînée dans la chaîne de traitements de l'information et vers l'information à valeur ajoutée pour la prise de décision.

5. Conclusion

La *recherche d'information* coordonnée avec un processus d'analyse automatique de textes et avec un processus d'annotation de documents, a permis d'intégrer de nouvelles dimensions relatives au mécanisme de représentations des indicateurs. Dans cette démarche, il est question d'affiner un raisonnement aux "frontières" des modèles et des problématiques sur les SRI, les modèles de représentations de concepts et l'IE [COUZINET, 2005]. L'apport de l'annotation dans cette perspective permet de faciliter la traduction d'un problème décisionnel en un problème de recherche d'information en clarifiant les objectifs et en formalisant les indicateurs en une liste d'attributs et de valeurs.

Le thème " *systèmes d'information* " nous a fournit un cadre d'analyse sur des problèmes liées au document et l'évolution de sa gestion : modélisation de l'utilisateur (veilleur, décideur), information à valeur ajoutée dans la

prise de décisions (processus d'annotation) et outils d'analyse automatique de contenus textuels (processus d'analyse et d'indexation).

Dans ce travail, les modèles représentés et leur évolution vers l'expérimentation offrent une nouvelle dimension dans les processus d'IE. Il est question de faire des mises en correspondance automatisées par l'apport de chaque modèle dans l'analyse des ressources informationnelles : documents sources et annotations. Egalement, de contribuer à faire des recoupements d'informations sur les sources et sur les annotations afin d'expliciter les problématiques par un agent humain [KISLIN, 2005] pour un traitement automatisée. Il s'agit de faire remonter les indicateurs sur l'information : ce qui échappe au système et ce qui couvre les besoins informationnels [ALQUIER, 2000]. C'est ainsi que l'information à valeur ajoutée est mise en évidence dans cette nouvelle architecture logicielle.

L'accessibilité aux bases de documents pour retrouver l'information pertinente [GUARINO & WELTY, 2000a-b] et satisfaire une demande stratégique d'un décideur reste un véritable défi. Plusieurs aspects fondamentaux sont à prendre en considération comme les modélisations à mettre en œuvre pour l'analyse et l'extraction. La validation de la pertinence des résultats et leur fiabilité sont pondérés aux besoins et aux centres d'intérêts des utilisateurs finaux : les décideurs.

Ces questions nécessitent encore des efforts d'étude et la continuité dans les propositions pour apporter des solutions tangibles dont les retombées techniques et économiques sont considérables.

6. Bibliographie

[ALQUIER, 2000] : ALQUIER Anne-Marie (2000), « Quelques principes méthodologiques pour la conception de Systèmes d'Information d'Intelligence Economique en fonction des exigences en aide à la décision », Revue d'Intelligence Economique, N° 6-7, Association Française pour le Développement de l'Intelligence Economique, Oct. 2000.

[BACHIMONT, 1999] : Bachimont, B. (1999), " Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances.", In J. Charlet, M. Zacklad & G. Kassel (Eds.), Ingénierie des connaissances, Paris : Eyrolles.

[BUSH, 1945] : Vannevar Bush, 1945, *As We May Think*, The Atlantic Monthly, July 1945, http://www.ps.uni-sb.de/~duchier/pub/vbush/vbush.shtml (visited 06/06/2005).

[COUZINET, 2005] : Couzinet, Viviane. (2005). Intelligence économique et sciences de l'information et de la communication : quelles questions de recherche ? In : Actes du colloque international de ISKO-France. Sous la direction de Amos DAVID. Edition PUN Nancy. pp 13-26.

[DAVID & THIERY, 2002] : David, Amos and Thiery, Odile. Application of "EQuA^{2}te" Architecture in Economic Intelligence. In Information and Communication Technologies applied to Economic Intelligence - ICTEI'2002. (Ibadan, Nigeria). 2002.

[DENOUE, 1999] : DENOUE Laurent, 1999. Adding Metadata to improve retrieval: Yet Another Web Annotation System Syscom Team. Technical Report: University of Savoie. France. http://www.fxpal.com/people/denoue/publications/TR1999-01.pdf (visited 06/06/2005).

[DESMONTILS & AL, 2004] : E Desmontils, C Jacquin, and L Simon (2004). Dinosys : un outil d'annotation pour l'enseignement à distance sur le Web. In: Colloque "Miage et e-mi@ge". e-mi@ge, Marrakech, Maroc. http://e-miage.ups-tlse.fr/colloque/papiers/E.DESMONTILS.pdf (visited 06/06/2005).

[FROISSART, 2001] : Froissart C. « De la communication homme-machine à la recherche d'information dans la documentation technique ». Mémoire pour l'Habilitation à diriger des Recherches, Université Jean Monnet – Saint-Etienne, 2001.

[GDT, 1983] : Le grand dictionnaire terminologique (1983). Domaine : science de l'information sur l'acquisition et traitement des documents. http://www.granddictionnaire.com/ (visited 06/06/2005).

[GUARINO & WELTY, 2000a] : Guarino N., Welty, C. A Formal Ontology of Properties. in Dieng, R., and Corby, O., eds, Proceedings of EKAW-2000: The 12th International Conference on Knowledge Engineering and Knowledge Management. Spring-Verlag LNCS. October, 2000.

[GUARINO & WELTY, 2000b] : Guarino N., Welty, C. Ontological Analysis of Taxonomic Relationships. in Laender, A. and Storey, V. eds, Proceedings of ER-2000: The 19th International Conference on Conceptual Modeling. Springer-Verlag LNCS. October, 2000.

[HECK & al., 2003] : Rachel M. Heck, Sarah M. Luebke, Chad H. Obermark, « A Survey of Web Annotation Systems ». *Work supported by Grinnell College Noyce Science Summer Research Fund*, http://www.math.grin.edu/~rebelsky/Blazers/Annotations/Summer1999/Papers/ (visited 06/06/2005).

[KISLIN, 2005] : Kislin, Philippe. (2005). Les activités de recherche d'information du veilleur dans le contexte d'IE : le modèle WISP. In : Actes du colloque international de ISKO-France. Sous la direction de Amos DAVID. Edition PUN Nancy. pp 97-118.

[LE GUERN, 1989] : Le Guern M. Sur les relations entre terminologie et lexique. in actes du colloque: les terminologies spécialisés - Approches quantitatives et logico-sémantique, et Meta Vol.34, No.3., sept. 89.

[LE GUERN, 1991] : Le Guern M., Un analyseur morphosyntaxique pour l'indexation automatique. Revue de linguistique française : Le Français moderne . n°1, juin 1991.

[MARSHALL, 1998] : Marshall, C. C. (1998). Toward an ecology of hypertext annotation. In ACM Hypertext, pages 40–49. ACM Press.

[MARTINET, 1995] : B. MARTINET. L'intelligence économique. Les Editions d'Organisation, 1995.

[MARTRE, 1994] : MARTRE, Henri, "Intelligence économique et stratégie des entreprises", Rapport du Commissariat Général au Plan, Paris, La Documentation Française, 1994.

[MATTHEW & al., 1996] : Matthew A. Schickler, Murray S. Mazer and Charles Brooks. (*1996*). Pan-Browser Support for Annotations and Other Meta-Information on the World Wide Web. *in Fifth International World Wide Web Conference, 6-10 May, 1996 , Paris (France).* (URL visite: nov.2004) http://www5conf.inria.fr/fich_html/papers/P15/Overview.html.

[OVSIANNIKOV & al, 1999] : OVSIANNIKOV I., ARBIB M.A. and McNEILL T.H., 1999, Annotation Technology. Int. J. Human-Computer Studies, 1999, pp 329 – 362. http://portal.acm.org/ft_gateway.cfm?id=989877&type=pdf (visited 06/06/2005).

[SIDHOM & HASSOUN, 2003] : SIDHOM Sahbi, HASSOUN Mohamed. « Morpho-sytactic Parsing for a Text Mining Environment ». In Official Journal « Knowledge Organization » KO. 29(2002) No. 3-4, Edited by Olson, Hope A. – Saranchuk, Georgina R. Zaharia, (c) 2003 Ergon Verlag.

[SIDHOM, 2002] : SIDHOM, Sahbi. " Plate-forme d'analyse morphosyntaxique pour l'indexation automatique et la recherché d'information: de l'écrit vers la gestion des connaissances.", Thèse de Doctorat à l'Université Claude Bernard Lyon1, France, Mars 2002.

[THIERY & DAVID, 2003] : Thiery, Odile et David, Amos. L'architecture EQuA^{2}te et son application à l'intelligence économique. Conférence "Intelligence Economique : Recherches et Applications" - IERA'2003. (INIST, France). 2003.

[W3C-Annotation, 2004] : W3C Collaboration Working Group: Annotation (2004). Collaboration, Knowledge Representation and Automatability. http://www.w3.org/Collaboration/Overview.html#annotation (visited 06/06/2005).

[WOODS, 1980] : William A. Woods. Cascaded ATN Grammars. in American Journal of Computational Linguistics, January-March 1980, vol.6, n°1.

[WOODS, 1997] : William A. Woods. Conceptual Indexing : a better way to organize knowledge. Technical Report SMLI TR-97-61 : SUN Micosystems, Lab. Mountain View Canada, April 1997.

[YEE, 2002] : YEE Ka-Ping, 2002, CritLink: Advanced Hyperlinks Enable Public Annotation on the Web, Demo to the CSCW 2002 conference, New Orleans, Dec 2002, http://www.zesty.ca/pubs (visited 06/06/2005).

[Zeiliger, 2001] : Zeiliger, R. (2001). Nestor : The web browser and cartographer (Last update April 11, 2005). http://www.gate.cnrs.fr/~zeiliger/nestor/nestor.htm . CNRS. (visited 06/06/2005).