DEA de Linguistique Théorique, Descriptive et automatique

Option linguistique informatique
Université Paris VII

Organisation du lexique pour assister l'accès lexical

Gaëlle Lortal

sous la direction de

Mme Laurence Danlos (Professeur, Paris VII)
et
Mme Brigitte Grau (Maître de conférences, LIMSI-CNRS, IIE)
M. Michaël Zock (Directeur de recherches, LIMSI-CNRS)
groupe LIR

septembre 2003

Remerciements

A mes professeurs, *Laurence Danlos* et *Adil El Ghali*,
mes maîtres de stage*; Michael Zock* et *Brigitte Grau* pour leur bonne humeur et leur patience,

A *Nicolas Hernandez* et *Nicolas Babut*
pour les formations minute (heures ?) programmation et le soutien

A insi qu'à mon voisinage de bureau ou de labo, *Gilles Cotteret* et *François Goasdoué,*
mes collègues de galère, de DEA de Jussieu et du LIMSI, principalement, *Anaïté, Barbara, Goritsa, Salim, Diana*, *Anne-Laure* et *Vincent* pour leur soutien et leur amitié.

SOMMAIRE

Introduction

Présentation

Ce mémoire de DEA est l'union entre mes centres d'intérêt en linguistique, les cours suivis en DEA de Linguistique Théorique, Descriptive et automatique (option linguistique informatique ; Univ. Paris VII) sous la direction de Mme Laurence Danlos et un sujet proposé par M. Michaël Zock et Mme Brigitte Grau au LIMSI (Univ. Paris XI).
« Le mot sur le bout de la langue » est un thème étudié depuis longtemps au LIMSI. Il est fondé sur les théories en psychologie et en linguistique principalement, par rapport à la notion d'accès lexical. Ce sujet prend de l'importance en informatique du moment qu'on souhaite utiliser les résultats et mettre en pratique les découvertes théoriques.

Problématique

La problématique générale est l'organisation du lexique pour assister l'accès lexical. En effet, l'architecture mentale n'est pas quelque chose de parfaitement décrit, et de nombreuses zones d'ombre subsistent et sont très étudiées en psychologie. Ainsi, on considère que différents liens mentaux mènent à un mot.
La plupart des outils développés, principalement pour de la génération de texte, ne développe qu'un certain type de liens qui ne sont pas forcément suffisants. C'est pourquoi M.Zock et J-P.Fournier [ZOCKFOUR] ont mis en place un outil théorique d'aide à l'accès au mot recherché (développement pour des dictionnaires) utilisant différents types de liens.

Fondé sur le problème du TOT, l'application poursuit deux buts principaux ; tout d'abord aider à trouver le mot que l'on a sur le bout de la langue et deuxièmement, aider l'utilisateur à mémoriser les mots ou les structures fondamentales d'une langue. Seul le premier point est développé.
Deux approches du mot recherché sont mises en place, l'une permettant l'accès par la forme, l'autre l'accès par le sens.
L'accès par la forme peut se faire de deux façons, tout comme opère la mémoire humaine, par la graphie ou le son.
L'accès par le sens se fonde sur une structure du dictionnaire mental analogue à une ontologie. C'est-à-dire que dans ce vaste réseau où les mots sont des nœuds et les liens entre eux des relations, retrouver un mot consisterait à entrer dans ce réseau et suivre les liens. Le système devrait construire pour retrouver un mot cible, un réseau lexical ayant comme noyau le mot le plus proche que l'on retrouve et comme satellites des mots ayant un rapport avec ce dernier (dans un processus récursif où les satellites pourront devenir des noyaux). L'application ressemblerait à :

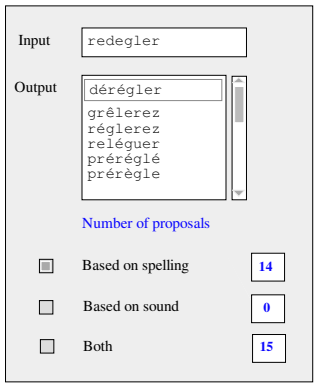

Figure 1 : *Recherche basée sur l'écrit*

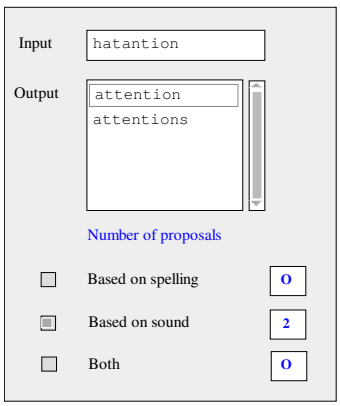

Figure 2 : *Recherche basée sur le son*

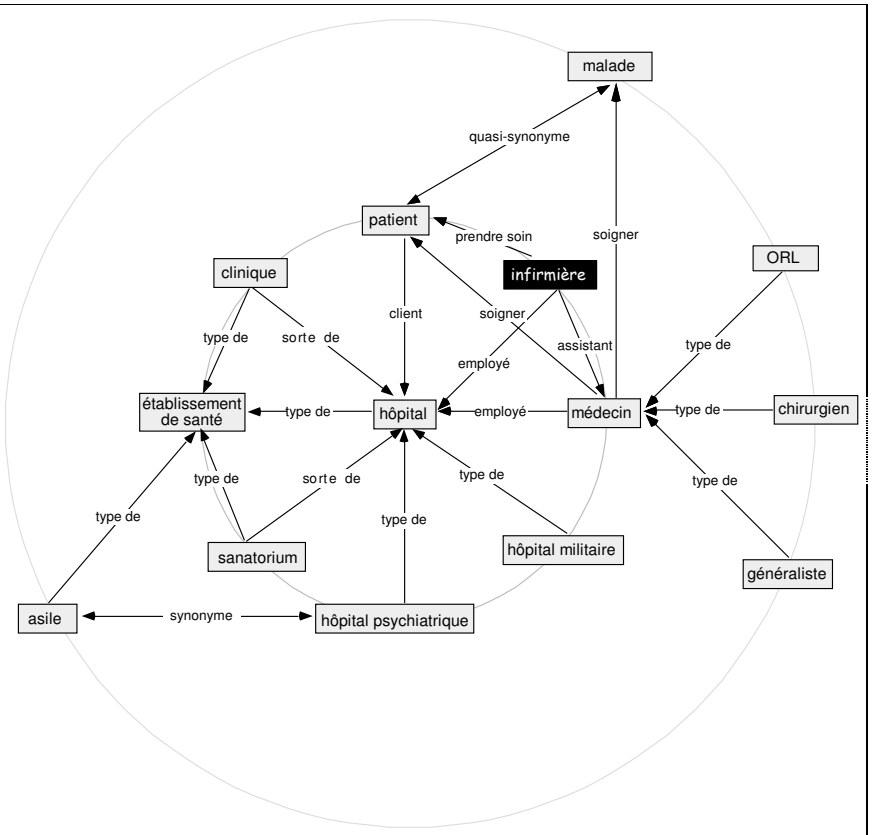

Figure 3 : *Recherche par le sens*

[ZOCKFOUR]

Le but de ce stage est de fournir les relations nécessaires à ce genre de réseau, donc de développer l'accès par le sens.
Pour cela plusieurs directions peuvent être étudiées. Il a été décidé d'observer le développement possible des relations sémantiques dites « syntagmatiques », principalement par la réutilisation d'outils automatiques de création de domaines, et l'analyse de corpus.

A la base, ce corpus devrait nous conduire à définir si à l'intérieur d'un domaine (structuré ou non) il existe des liens spécifiques entre les mots. Si des relations « typables » existent, alors il est intéressant de voir si elles sont pertinentes pour l'accès lexical, s'il est important de les inclure et de les hiérarchiser avec les relations paradigmatiques existantes (qui sont bien définies) ou syntagmatiques (à mettre en place), ou si au contraire elles sont contingentes, et qu'un système construisant des domaines automatiquement sans relations typées est suffisant.

Nous explorerons ces différentes pistes et leurs résultats, après avoir situé les théories principales de l'accès au lexique en psychologie, les observations linguistiques sur les relations sémantiques et lexicales ainsi que les outils développés dans ces domaines en informatique.
L'analyse mise en place grâce entre autres à une interface adaptée aux besoins permettra de quantifier les résultats des différents systèmes et de donner les lignes directrices d'une recherche plus poussée sur les relations nécessaires à l'accès lexical ainsi que l'enrichissement de ressources de recherche textuelle.

I. Etat de l'Art

I.1. Psycholinguistique

La psycholinguistique n'est pas seulement née de la collaboration entre psychologues et linguistes. L'apport de la neurologie et des techniques scientifiques modernes, permettant de voir les processus du langage sans interférences, est primordial dans les découvertes de ces dernières années.
Après une présentation du TOT dans son étude toute psychologique, nous verrons les avancées que la psycholinguistique a permises en traitement de la langue ainsi que celles de la psychologie.

I.1.1. Le Tip-Of-the-Tongue (TOT), phénomène du mot sur le bout de la langue
[TOT]

I.1.1.1. Définition

L'étude du phénomène du Tip-of-the-tongue ou TOT (le mot qu'on a sur le bout de la langue) indique que l'information est organisée au sein de la mémoire à long-terme (entre autres, les mots retrouvés font partie de la réserve de connaissances générales). Ce type de mémoire est constitué de différents sous-types de mémoire.

La *mémoire à long terme* stocke les informations pendant une longue période et même pendant toute la vie. Les informations y sont organisées et régies par deux systèmes qui fonctionnent en relation permanente : La *mémoire procédurale* (l'apprentissage d'habiletés perceptivo-motrices) et la *mémoire déclarative* (évocation de souvenirs sous forme de mots) d'une part, la *mémoire épisodique* et la *mémoire sémantique d'autre part.*

Ces dernières nous intéressent particulièrement puisque la mémoire sémantique est le système par lequel l'individu stocke sa connaissance du monde. C'est une base de connaissances, un magasin d'informations que nous possédons tous et dont une grande partie nous est accessible rapidement et sans effort ; leur oubli n'existe pas ; il est considéré comme un défaut d'accessibilité. La mémoire sémantique concerne les concepts, le sens des mots et des symboles.
Il existe également une mémoire qui concerne la forme des mots, sa graphie, sa prononciation... c'est la *mémoire lexicale* (ex: "Coquelicot" est composé de quatre syllabes, commence par la lettre "C", se termine par le son /o/, etc. ...). La mémoire sémantique et la mémoire lexicale sont regroupées sous le terme de *mémoire verbale*.
La *mémoire épisodique*, elle, enregistre tous les événements biographiques d'un sujet. Elle est soumise aux vicissitudes d'interférence, de l'oubli, de la subjectivité, des variations de contexte, de la tonalité affective, de la fréquence. La récupération d'information en mémoire épisodique est l'accès à une information en mémoire à partir d'une information de la situation présente, par un rappel, un indice, une reconnaissance.

Le TOT n'est pas un oubli effectif. On ne perd pas l'information en mémoire. Toutes les expériences que l'on a sont encodées dans la mémoire. Une fois en mémoire, elles deviennent disponibles. Il est possible de récupérer ces informations dans certaines mémoires. Cet ensemble représente les mémoires accessibles.
Le TOT est donc un exemple d'échec de récupération d'information. L'information est en mémoire, elle est donc disponible, mais pas accessible. Le TOT est un défaut d'accessibilité.

Deux types de récupération sont distingués ; le rappel et la reconnaissance.

Dans le cas du rappel (recall) l'information est reproduite de la mémoire. La catégorisation, la structuration de l'information permet le rappel.

La reconnaissance (recognition) est liée à la présentation d'un stimulus qui évoque une expérience passée.

Ces deux types de récupération mettent en jeu des définitions de symboles perceptuels différentes qui divisent certains théoriciens. Cependant, dans une problématique d'accès au lexique, les deux types de récupération peuvent être combinés.

I.1.1.2. Description

Quand un individu est dans un état de « TOT », c'est-à-dire qu'il est capable de réagir à la définition du mot mais n'est pas capable de produire le mot cible. Il est capable de récupérer dans sa mémoire des mots :

- qui ont le même son que le mot cible ;
- qui commencent par la même lettre que le mot cible ;
- qui contiennent le même nombre de syllabes que le mot cible ;
- qui ont un sens proche du mot cible

L'observation du phénomène du TOT permet de prouver que le langage ne porte pas le sens en soi-même. C'est le débat philosophique du mot avant le sens ou vice-versa. Selon Fauconnier, le langage n'est qu'un moyen puissant de contrôler les constructions de conceptualisations. Les mots ne portent pas le sens de façon inhérente.
Si, comme le soutient Jorion, les concepts sont représentés par des mots, non par des perceptions (« percepts » selon Barsalou), et que le langage est donc si central dans l'activité de la pensée, comment les aphasiques pourraient perdre le langage sans perdre les connaissances, comment pourrait-on penser à quelque chose et ne pas connaître le mot qui y réfère ? Où serait le mot sur le bout de la langue ? [Barsalou]

Ce phénomène pose des questions sur plusieurs domaines. L'application prévue par M.Zock et J-P.Fournier s'appuie sur les observations faites sur le TOT, et il s'agirait d'améliorer en développant les relations entre concepts, l'accessibilité qui fait défaut dans un état de TOT ([ZockFour]).
Pour se faire, il faut observer les possibilités d'accès au sens de notre mémoire que l'on a cherché à modéliser par les représentations de concepts (ou de connaissances), en particulier au niveau de la langue grâce aux réseaux sémantiques.

L'observation du TOT et de phénomènes du même type (lapsus par exemple) par l'alliance de la psychologie et de la linguistique a permis de confirmer ou de découvrir divers fonctionnement de la mémoire, mais aussi d'apporter de nombreuses réponses aux questions posées en traitement du langage.

I.1.2. Connaissances sur le langage par la psycholinguistique

[Labelle]

Tout d'abord au niveau du fonctionnement cognitif général, la perspective de modularité du cerveau est confortée par l'observation de traitements spécialisés selon le type de signaux appréhendé.
Cette modularité est mise en valeur par les travaux de Damasio et coll. (1996) qui observent par comparaison entre des sujets sains et des sujets avec différents types de lésions, que les informations lexicales sont regroupées par classes conceptuelles dans le cerveau.

De même la psycholinguistique permet de tester grâce à des techniques scientifiques telles que la magnéto-encéphalographie, les diverses opérations cognitives requises par une tâche de dénomination d'un dessin (Levelt et coll., 1998). On sait alors qu'après reconnaissance visuelle du dessin, il y a successivement, récupération du concept lexical, puis de la forme phonologique du lexème, et enfin l'encodage phonologique et l'initialisation de l'articulation.
Ce genre d'observation va dans le même sens que les théories de profil d'apprentissage d'une personne dans la gestion mentale (voir [GesMen]). Elles sont aussi réutilisées dans le système de MM. Zock et Fournier.

Sur le thème de la production du langage, les études ont foisonné, principalement grâce à l'étude du lapsus par Fromkin (1971), les hésitations par Goldman-Eisler (1972) et le phénomène du mot sur le bout de la langue (TOT) par Brown & McNeill (1966).
Ces phénomènes de performance observés dans la production linguistique montrent souvent la réalité psychologique des unités linguistiques à plusieurs niveaux (phonèmes, syllabes, morphèmes, mots, syntagmes), et reflètent aussi les processus de production du langage. Fromkin est la première à avoir proposé un modèle des opérations cognitives impliquées dans la production du langage. Elle considère que la phrase est construite en cinq étapes agissant sur une première séquence qui est le contenu sémantique de la phrase, puis sur la structure syntaxique (noeuds lexicaux, ensembles de traits sémantiques), ensuite sur l'intonation phrastique, la sélection de lexèmes aux traits sémantiques appropriés et enfin la forme phonologique (avec les règles morphophonémiques). Le résultat est envoyé aux organes d'articulation.
Les modèles élaborés ensuite (par Garrett/Levelt, Dell et Reich, Stemberger par exemple) postulent un retour en arrière ou des interactions possibles entre différents niveaux de modules.
Ces retours selon eux peuvent expliquer certains lapsus comme le remplacement du mot anglais *present* 'présenter' par *prevent* 'prévenir', dû à l'activation du mot *present* transmise aux unités correspondant aux phonèmes de ce mot; et que ces dernières, en retour, transmettent leur activation aux mots auxquels elles sont associées; *prevent* se retrouve ainsi activé par le fait qu'il partage presque tous ses phonèmes avec *present*.

Les travaux en psycholinguistique sur les lapsus ou le mot sur le bout de la langue montrent aussi l'influence du contexte grammatical sur le nombre de mots avec lesquels le mot cible compétitionne. Dahan et coll. (2000) ont montré que lorsqu'un mot est précédé d'un article marqué pour le genre, seuls les mots compatibles avec le genre de l'article sont activés.
La plupart des travaux sur ces dérapages de production mettent en valeur un système modulaire de la production, tel qu'il serait mis en place dans un système informatique simple. C'est-à-dire que tous les sens d'un mot sont activés dès qu'un mot commence à être reconnu, et que les sens non compatibles avec le contexte sont éliminés dans une deuxième étape.

Pourtant, un des problèmes posés par l'accès lexical est celui de la levée des ambiguïtés lexicales dans le cas où seuls les sens compatibles avec le contexte sont sélectionnés dès le début du décodage. On suppose alors que des niveaux supérieurs de traitement interviennent dans des décisions prises à des niveaux inférieurs.
Les techniques dites d'amorçage mises en place valident cependant toujours la première théorie dérivée de Fromkin. La technique d'amorçage consiste à faire apparaître un mot (une amorce) avant d'y présenter un stimulus. Cette technique permet de déterminer si l'amorce facilite la reconnaissance du stimulus. En général, les résultats montrent que tous les sens sont sélectionnés dans une première étape du décodage puis que seul le sens induit par le contexte est actif en mémoire. Duffy, Morris et Rayner apporte une information supplémentaire par leurs tests de mouvements des yeux, que l'activation d'un des sens dépend de sa fréquence et du contexte.

La fréquence du mot a priori agit aussi sur le classement des mots au niveau morphologique. En fait deux conditions jouent ; la fréquence et la transparence.
Par exemple, les recherches montrent que les mots construits sémantiquement opaques sont listés dans le lexique et ne donnent pas lieu à un découpage morphologique. Par contre, en ce qui concerne les mots sémantiquement transparents, on considère un modèle mixte où deux voies d'accès parallèles sont possibles ; les mots très fréquents ont leur propre représentation lexicale tandis que les mots rares sont décomposés (Frauenfelder/Schreuder/Baayen).
Cependant, l'étude de langues différentes peut mener à des conclusions différentes. Ainsi, des études sur les verbes anglais permettent de considérer que la flexion régulière verbale est calculée, tandis que les formes irrégulières sont mémorisées.

D'une façon globale, ces théories soulignent le fonctionnement des outils que nous utiliserons dans notre étude (fréquence) et guident la mise en place de notre expérience, dans laquelle on utilisera un contexte pour situer notre recherche lexicale. Notre amorce sera constituée à la fois par ce contexte et les informations que nous définirons pour une observation comme nécessaire.
Notre expérience cherche en fait à typer l'amorce nécessaire pour que notre système accède à un mot perdu.
Plusieurs théories en psychologie peuvent nous guider encore, comme celles mettant en exergue le contexte.

I.1.3. Les théories fondées sur le fonctionnement de la mémoire humaine

Deux conceptions principales sont mises en compétition en ce qui concerne les symboles perceptuels. Dans des systèmes du type Embodied theory (Glenberg) des symboles de type modal sont mis en jeu, alors que dans des théories du sens du type « High Dimensional », ce sont des symboles amodaux.

En effet, [Barsalou] redéfinit deux types de symboles perceptuels. Un symbole perceptuel est modal et analogique selon lui. Modal signifie qu'un symbole est représenté dans les mêmes systèmes que les états de perception qui les ont conçus. Du fait, ces symboles perceptuels sont aussi analogiques c'est-à-dire que la structure d'un symbole perceptuel correspond plus ou moins à l'état de perception qui l'a produit. Le système neuronal qui représente le symbole représente aussi l'état.

Si un symbole est dit amodal, il est aussi arbitraire. C'est le cas dans des systèmes symboliques, fonctionnels comme les théories « High dimensional ». Le symbole perceptuel est amodal dans la mesure où ce qui est stocké en mémoire (ou traité) l'est sous la forme de concepts qui n'entretiennent pas de relation avec ce qu'ils représentent. De plus la structure

interne et les états perceptuels n'utiliseront pas les mêmes systèmes de représentation, ni n'utiliseront les mêmes principes opérants. C'est le principe de la définition du signe linguistique.

Bien que ces deux types de théories donnent naissance à des applications approuvées, le problème de l'ancrage est souligné par des chercheurs en psychologie cognitive tels que Barsalou ou Glenberg, ce dernier proposant des solutions (Voir Annexes1).

Ainsi même si on peut avoir conscience du bien fondé d'une application proche du fonctionnement du sens chez l'Homme (Glenberg), si une application symbolique est suffisante dans une application, n'est-elle pas adéquate dans notre optique de travail?

I.1.3.1. High Dimensional Theories of meaning

Ce type de théorie est représenté principalement par HAL : Hyperspace Analogue to Language et LSA : Latent Semantic Analysis ([LSA] [HAL]), qui sont des théories linguistiques du sens fondées sur des bases mathématiques d'espaces de grande dimension. Elles se servent de symboles abstraits et arbitraires (le vecteur).

I.1.3.1.1. Principes

Burgess et Lund sont les principaux représentants du développement de ce modèle théorique du sens en 1997, l'Hyperspace Analogue to Language.
L'implémentation HAL est fondée sur les recherches en psycholinguistique de Burgess. Il apparaîtrait que les deux hémisphères du cerveau bien que très différents au niveau des taux d'activation du sens ne diffèreraient pas en ce qui concerne le niveau de représentation. Dans ce contexte psycholinguistique, au niveau de la reconnaissance du mot, a été développé une implémentation des asymétries cérébrales qui utilise l'HAL.
Ils considèrent que le sens d'un mot est dérivé d'une analyse dimensionnelle des mots en contexte ; le sens d'un mot est sa représentation vectorielle dans un espace construit par 140 000 co-occurrences mot-mot.
Dans la Latent Semantic Analysis de Landauer et Dumais (1997), le sens d'un mot est sa représentation vectorielle dans un espace d'environ 300 dimensions dérivé d'un espace à beaucoup d'autres dimensions. La dérivation des vecteurs sélectionne un « espace sémantique », c'est-à-dire un ensemble de contextes
Le but de HAL comme celui de LSA (ci-après) est d'offrir une théorie du sens adéquate.

I.1.3.1.2. Fonctionnement

HAL est une simulation informatique de la mémoire humaine. Il s'appuie sur un large corpus (environ 300 millions de mots) analysé en parties découpées grâce à une fenêtre de 10 mots se déplaçant sur une matrice de 70 000 lignes et colonnes de mots. La matrice permet de trouver les paires de mots en co-occurrence.

Chaque co-occurrence est alors affectée d'une valeur selon la proximité des deux mots de la paire. Ainsi pour un mot, les 70 000 éléments d'une ligne sont combinés au 70 000 éléments d'une colonne, donnant un vecteur de 140 000 éléments qui représente le sens.

Dans LSA, le fonctionnement global est assez proche. Un espace consiste en les 2 000 premiers caractères de chacun des 30 473 articles de la Grolier's Academic American Encyclopedia. Chacun des articles est placé dans une colonne de la matrice, tandis que les mots extraits des articles (environ 60 000) sont placés sur les lignes.

Les entrées de la matrice sont le nombre d'occurrences d'un mot dans un contexte (article). Ces entrées sont soumises à un traitement SVD (singular value decomposition) et à des transformations. L'analyse SVD permet l'extraction de 300/400 dimensions et des valeurs des mots dans celles-ci. De la sorte, chaque mot est représenté par un vecteur de 300 à 400 dimensions.

La notion de « similarité » dans HAL est en fait une similarité de contexte comme dans LSA, c'est-à-dire que les dix termes dans la fenêtre sont identiques pour deux termes. Une preuve de la productivité de ce modèle est l'utilisation validée du vecteur en tant que représentation du sens pour simuler la catégorisation.

Landauer et Dumais postulent que LSA peut permettre de représenter toutes les connaissances humaines. De plus, ils appliquent leur théorie à la compréhension de phrases et de discours. Une phrase sera représentée par la moyenne des vecteurs de mots qu'elle contient et la cohérence entre phrases est calculée par le cosinus de l'angle (dans un espace multidimensionnel) entre les vecteurs correspondant aux phrases successives. La moyenne de ces vecteurs LSA saisirait le sens principal d'un passage selon les auteurs.

I.1.4. Applications construites sur des fondements psycholinguistiques

HAL ou LSA sont des théories qui donnent lieu à des implémentations possibles. C'est-à-dire que pour prouver leur validité, Burgess et Lund ou Landauer et Dumais ont mis en place des applications-tests qui effectuent le TOEFL (Test Of English as Foreign Language) par exemple, ce qui prouve que les systèmes effectuent une catégorisation des mots en anglais aussi effective que celle d'un non natif. De même, LSA a été implémenté dans des modèles de machines apprenantes.

I.1.4.1. WordNet

L'application la plus connue à base de théories psycholinguistiques est sûrement Wordnet, developpée par l'université de Princeton, groupe de G.Miller [Miller].
Wordnet est une base de données constituée de connaissances sémantiques et organisée selon des principes théoriques issus de recherches en psycholinguistique et en informatique au sujet de la mémoire lexicale humaine.
Les thèses de création de Wordnet sont les résultats des travaux de Miller qui montre que la capacité de l'esprit humain est limitée. La mémoire à court-terme serait en effet limitée à environ sept unités.
Mais les limitations de cette mémoire sont améliorées grâce à trois principes facilitants. Tout d'abord, il est possible de faire une économie en utilisant des jugements relatifs plutôt que des jugements absolus. Ensuite, il est possible d'augmenter le nombre de dimensions dans lesquelles les stimuli peuvent se différencier, ou enfin, il est possible de modifier une tâche de façon qu'elle soit représentée par des séquences de plusieurs jugements absolus d'affilée.
De la sorte, partant des mots du corpus *Brown*, Miller construit sa base de données. Il la modèlera par une vision relationnelle (en opposition à la vision componentielle, atomisante issue de Katz et Fodor) et organisera les informations de ce thésaurus en « chunks » afin de repousser les limites d'un système organisé comme une mémoire humaine.

Dans Wordnet, le sens d'un mot est représenté par les termes qui lui sont « superordonnés », ainsi que ses traits distinctifs.

La psychologie a beaucoup apporté au développement des études sur le langage grâce à l'observation des performances d'un locuteur, que ce soit pour le développement de modèles de représentation ou celui d'outils du type WordNet.
C'est par les études en psychologie sur la catégorisation que se sont développés les modèles fondés sur le domaine tels que HAL ou LSA et en même temps de multiples applications informatiques comme nous le verrons lors de la présentation des outils utilisés pour notre sujet.

La catégorisation en linguistique peut être elle observée sous l'angle des concepts et de leur organisation, mais aussi des associations et relations lexicales.

I.2. Linguistique

L'étude des relations du lexique peut être effectuée à divers niveaux d'organisation ; du point de vue psychologique par exemple, avec l'observation de la mémoire et de ses faiblesses mais aussi d'un point de vue linguistique, puisque cette organisation transparaît dans la parole.
Différentes relations sont reconnues au niveau linguistique, suivant des dichotomies de base utilisées en sciences du langage, celle des deux axes syntagmatique et paradigmatique par exemple, ou encore celle des parties du discours. Ces relations permettent d'essayer d'organiser le lexique en termes de signifiés plutôt qu'en termes de signifiants. Après avoir présenté des relations plus propres aux noms, les relations paradigmatiques et syntagmatiques, nous verrons des relations propres aux verbes.

I.2.1. Relations paradigmatiques

Les relations de signification paradigmatiques permettent de distinguer les relations qu'une unité linguistique entretient avec des unités absentes et qui pourraient occuper sa place [Desgoutte]. Elles sont la réponse à « que puis-je mettre à la place du mot-clef ? ».
Elles représentent des liens entre des lexies reliées entre elles directement et sémantiquement.
Les relations paradigmatiques sont les relations les plus étudiées, c'est pourquoi nous ne ferons qu'un récapitulatif des relations les plus souvent évoquées [Girault] [Jenhani]. Les principales, utilisées dans la majorité des ontologies et thésaurus sont :

Hyponymie
Méronymie
Holonymie
Synonymie
Antonymie

Elles sont toutes définies ci-après, et exemplifiées en annexes (Annexes 2).

I.2.1.1. Hyponymie

L'hyponymie est la relation entre un terme spécifique et un terme générique, exprimée par les expressions «ISA / est-un / sorte de » ;
L'hypo/hyperonymie (ou relation ISA) est une relation sémantique entre deux signifiés : X est un hyponyme de Y si on peut dire que « X est une sorte de Y ».
L'hyponymie est transitive, antisymétrique et génère une structure sémantique hiérarchisée. Un hyponyme hérite de tous les traits du concept le plus générique et possède en plus des traits spécifiques qui le distinguent de son « père » et de ses « frères ».
Par exemple, un chien est un hyponyme de animal et animal est un hyperonyme de chien.
La relation « en particulier » est l'inverse de l'hyponymie.

I.2.1.2. Méronymie

La méronymie est une relation transitive et antisymétrique, qui organise le lexique mental exprimée par l'expression « HAS-A / partie de ».
X est un méronyme de Y si et seulement si « X est une partie de Y » ou « Y a un (des) X(s) » les descriptions sont conceptuellement acceptables.
La relation de méronymie est la relation de tout à partie. Le méronyme désigne la partie et l'holonyme désigne le tout.
A est un méronyme de B implique que A est une partie de B (ou B est fait de A).

Souvent la transitivité de la relation de méronymie est limitée. [WinstonChaffin] ont proposé six types de méronymie pour cela ils ont fixé trois critères : la fonctionnalité, la similarité et la divisibilité.
fonctionnalité + / - : les parties sont (/ ne sont pas) dans une position spatiale / temporelle spécifique par rapport aux autres qui supportent leur fonction au sein du tout.

similarité + / - : les partie sont (/ ne sont pas) similaires les unes aux autres et au tout auquel elles appartiennent.

divisibilité + / - : les parties peuvent (/ ne peuvent pas) être physiquement détachées du tout auquel elles appartiennent.

I.2.1.3. Holonymie

La réciproque de la méronymie est donc l'holonymie ; A est un holonyme de B implique que A est fait de B (A a/contient des B). Cette relation est exprimée par l'expression « composé de ».

I.2.1.4. Synonymie

La synonymie et l'antonymie sont des relations lexicales entre des signifiants, contrairement aux relations précédentes qui sont des relations entre signifiés.
Les vrais synonymes sont rares, on utilise une définition affaiblie relative au contexte : deux expressions X et Y sont synonymes dans un contexte linguistique C si la substitution de l'une par l'autre dans C n'altère pas la valeur de vérité.
La synonymie peut être plus large « synonymie + » ou plus étroite « synonymie - ». [Cruse] distingue trois types de synonymes :

Synonymes absolus,
Synonymes cognitifs,
Plesionymes.

I.2.1.5. Antonymie

L'antonyme d'un mot X est parfois non-X, mais tout comme la synonymie, autre relation lexicale entre des signifiants, la relation n'est pas manichéenne. Cette opposition particulière se répercute sur les relations de parenté par exemple (fils / fille, père / mère, roi / reine, acteur / actrice).
Il existe plusieurs sous classes d'antonymie [Cruse]:
La complémentarité : Ce sont des couples de mots dont la négation de l'un implique l'affirmation de l'autre mais on ne peut pas nier les deux en même temps.

L'antonymie scalaire : Ce sont des couples de mots dont la négation de l'un implique l'affirmation de l'autre mais on peut nier les deux en même temps et la gradation est possible.
La réciprocité : Ce sont des couples de mots réciproques.
L'inversion : Ce sont des couples de mots dont l'un signifie l'inverse de l'autre.

I.2.2. Relations syntagmatiques

Les relations de signification syntagmatiques sont le pendant des relations de signification paradigmatiques. Les relations syntagmatiques rendent compte en théorie de l'organisation temporelle (in præsentia) de l'énoncé. Elles permettent de distinguer les relations qu'une unité linguistique entretient avec d'autres unités présentes dans la chaîne du discours [Desgoutte]. Elles répondent à la question « Qu'est-ce qui va avec le mot-clef ? ».
Des relations de type syntagmatique sont entre autres des relations d'association, c'est-à-dire qui relient les mots qui sont activés lorsque l'on en évoque un autre. Ces relations permettent un amorçage sémantique ; l'accès à un mot est facilité s'il a été précédé par un autre mot qui lui est sémantiquement associé.
Le principe des collocations indique par exemple une organisation syntagmatique qui perdure dans la mémoire (voir TST).

Ces relations sont plus difficiles à traiter, elles varient énormément selon le sociolecte, l'idiolecte (dont les expériences personnelles) d'un individu, et même en fonction du contexte.
Selon [Girault], elles sont de plusieurs types :

I.2.2.1. Relations script (ou scénario)

Définition

Elles relient des mots appartenant à un même scénario. Ce type de relations peut être relatif à :
l'espace : « se trouve où », « lieu de », « se trouve avec » ;
le temps : « se trouve quand », « moment de » ;
les propriétés : caractéristique de / propriété ;
l'usage : l'utilisation d'un objet est souvent la partie centrale de la conception pour une personne de cet objet. « utilisé avec », « utilisé pour », « a pour objet » ;
la fabrication : « fait par », « auteur de », « cause de », « résultat de » ;

Exemple

Relation script	Script espace	Script temps	Script propriétés	Script usage	Script fabrication
Pain, croissant, café au lait (scénario petit déjeuner)	Bouteille, vin (vin, se trouve où, bouteille)	fleurs, printemps (se trouve quand, printemps	Voler, oiseau (caractéristique de, oiseau)	Marteau, clou (utilisé avec, clou)	Pain, boulanger (fait par, boulanger)

Fig.1 Relation_Script

I.2.2.2. Relations type analogique

Définition

La relation « analogique » peut être de type :
attribut : elle relie un mot et la valeur d'un de ses attributs ;
croyances : elle relie des mots dont l'un est un attribut que l'on prête à l'autre mais qui n'est pas toujours fondé ;
symbole : relie des mots dont l'un est le symbole de l'autre : « symbole de », « représenté par » ;
ressemblance : elle relie un mot à un autre qui a des caractéristiques semblables : « ressemble à ».

Exemple

Analogique attribut	Analogique croyance	Analogique symbole	Analogique ressemblance
canari / jaune ; nain / petit	rusé / renard, écossais / radin	Tour Eiffel, symbole de, France blanc, symbole de, pureté / paix	ballon de rugby / oeuf

Fig.2 Relation_analogie

I.2.2.3. Relations type lexical

Définition

Ce type de relations peut être une relation de :
collocation : elle relie des mots fréquemment associés au sein d'un énoncé ;
expression : elle relie des mots et des expressions dans lesquelles ils figurent ;
formule : elle relie des mots et un groupe nominal qui leur sont souvent associés ;
mots composés : elle relie des mots à des mots composés dont ils font partie ;
dérivés syntaxiques : relations entre deux mots qui ont le même radical mais qui appartiennent à des parties différentes du discours (au niveau syntaxique profond) [10] ;

Exemple

Collocation	Expressions	Formules	Mots composés	Dérivés syntaxiques
beurre / rance, pneu / crevé	mer / ce n'est pas la mer à boire	mer / bord de mer ; recette / recette de cuisine	père / beau-père ; mer / outre-mer	école / scolaire ; courir / course ; promesse / promettre

Fig.3 Relation_lexical

I.2.2.4. Relations liens personnels

Définition

Ce type de relations peut être de type :
possession : « appartient à », « possède »
affectif : associe des mots qui ont un lien affectif

épisodique : relie des mots qui font partie d'un événement marquant dans la vie d'un individu.

Exemple

Possession	Affectif	Épisodique
Médor, appartient à, voisin ; millionnaire, possède, argent	Nounours, Sam Roméo, Juliette	métro / agression ; voiture / accident

Fig.4 Relation_personnel

I.2.3. Relations nominales et verbales

Un autre type de relation de la sémantique lexicale apparaît lorsque qu'on observe les parties du discours. [Jenhani] a opté pour cette méthode.
En effet, les relations liées à la définition des propriétés des noms et à leurs aspects fonctionnels sont très différentes des relations allouées aux verbes par exemple à leurs relations thématiques [Jenhani].
Il est reconnu que les noms et les verbes doivent être traités indépendamment car ils n'ont pas le même type de relations. Par exemple, Bussone et Rossi (2000) ont déterminé les effets d'amorçage liés aux associations verbe-verbe et aux associations verbe-nom pour évaluer la structure du réseau sémantique des verbes. A l'aide d'une décision lexicale, ils ont découvert que le degré de proximité sémantique était plus élevé pour des associations verbe-nom. Les relations verbe-verbe (ex : acheter - vendre) n'étaient pas activées automatiquement. Ceci signifie que le lexique des verbes est structuré autour des noms comme il a été postulé dans le modèle SVETLAN'.

I.2.3.1. Le nom : relations structurantes et fonctionnelles

Ces types de relations sont équivalentes aux relations script de [Girault], mais d'un point de vue différents.

I.2.3.1.1. Relations structurantes

Définition

Ces relations permettent de définir la structure globale d'un nom en distinguant les propriétés intrinsèques et les propriétés extrinsèques.
Les propriétés intrinsèques d'un objet sont toutes les propriétés indispensables à l'existence de ce dernier comme par exemple, les propriétés physiques d'un objet (forme, couleur, poids, dimensions d'un objet), et les caractéristiques (état de l'objet, sa durée de vie, sa validité...).
Les propriétés extrinsèques d'un objet sont toutes les propriétés non indispensable à son existence ou qui peuvent changer au cours du temps comme par exemple, les propriétés d'identification (nom, nationalité, …), celles qui permettent de distinguer un objet d'un autre ou une personne d'une autre [Jenhani].

Exemple

Terme	Propriétés intrinsèques	Propriétés extrinsèques
Maison	propriétés physiques : dimensions, surface, forme état : neuve	propriétés d'identification : couleur

Fig.5 Relation_structurantes

I.2.3.1.2. Relations fonctionnelles

Définition

[Jenhani] décrit les relations fonctionnelles en se fondant sur le lexique génératif de Pustejovsky, principalement sur la composante de l'ensemble des mécanismes génératifs qui s'appliquent à la première composante (la représentation sémantique des mots) pour construire des expressions sémantiques correspondant à la signification des mots « en contexte ».
Ce qui est intéressant dans ce modèle, c'est principalement les deux rôles appartenant à la structure de « qualia ».
La structure de qualia permet de décrire la sémantique des noms en spécifiant entre autres, le rôle télique et le rôle agentif.
Le rôle télique marque le but et la fonction de l'objet (motivation de l'agent de l'action, ou fonction inhérente à l'activité), tandis que le rôle agentif spécifie ce qui est à l'origine de l'objet (facteurs liés à la création ou à l'origine de l'objet ; créateur, artefact, produit naturel, chaîne causale).
[Jenhani] y ajoute deux aspects, l'aspect causatif (causes provoquées par un objet.), et le but (cas où le but n'est pas le même que le télique de l'objet).

Exemple

Nom	Aspect agentif	Aspect télique	Aspect causatif	But
Livre	écrire	Lire Editer Publier relier	Connaître/ connaissance	Eduquer Informer
Voiture	construire	Transporter Conduire Assurer Entretenir	Déplacement	Transporter

Fig.6 Relation_fonctionnelle [Jenhani]

Ce point de vue structurant / fonctionnel est utile à la compréhension des relations, mais l'organisation de S.Girault nous semble suffisante et plus simple pour l'étiquetage de relations.

I.2.3.2. Le verbe : implication, troponymie, opposition et relations thématiques

Les verbes sont une catégorie syntaxique et lexicale du langage, clef dans la mise en place de la « charpente » relationnelle et sémantique de la phrase.
Il existe des relations propres aux verbes, telles que l'implication, la troponymie, ou l'opposition [Gardent].
Le langage contient beaucoup moins de verbes que de noms. Mais les verbes sont plus polysémiques que les noms. Les verbes peuvent changer de sens suivant le type de noms avec lesquels ils apparaissent, alors que les noms ont un sens plus stable en présence de différents verbes. Cette flexibilité des verbes rend leur analyse sémantique plus difficile. Les verbes les plus fréquents sont aussi souvent les plus polysémiques [Girault].

I.2.3.2.1. L'implication

Définition

C'est une relation unilatérale, sauf lorsque les verbes sont des synonymes, ils s'impliquent alors mutuellement.

Exemple
- jouer ➔ gagner / perdre

I.2.3.2.2. La troponymie

Définition

Les différents traits qui distinguent un verbe hyponyme de son « père » sont regroupés sous une relation que Fellbaum et Miller (1990) ont baptisée troponymie (du grec *tropos* manière de ou mode). La distinction sémantique entre deux verbes est différente des traits qui distinguent deux noms dans une relation d'hyponymie.
La relation de troponymie entre deux verbes peut être exprimée par la formule :
faire V1 est faire V2 d'une manière particulière.

Exemple
- ronfler / dormir
- boiter / marcher

I.2.3.2.3. Relation d'opposition entre verbes (ou relation converse)

Définition

La relation d'opposition est très variée pour les verbes. Elle peut être représentée par des :
paires opposées (ou converses) : par exemple les verbes de mouvement par les directions qu'ils indiquent
paires référant à une même activité : paires exposant des points de vue de différents participants sur une même activité. La forte association lexicale est probablement due au co-usage fréquent.
verbes d'état ou de changement d'état antonymes : ce sont des verbes d'état qui peuvent être exprimés en termes d'attributs.

Exemple

Converse	Activité identique	Verbes de changement d'état
Monter / descendre	Donner / prendre	Allonger / raccourcir

Fig.7 Relation_opposition_verbale

I.2.3.2.4. Les relations thématiques

Définition

Seuls les verbes possèdent des schémas thématiques, puisque ce schéma permet de « relier un verbe à ses arguments par l'intermédiaire de relations étiquetées par des rôles thématiques » [Jenhani].

[Jenhani] distingue six rôles thématiques :
Les acteurs : font l'action, ce sont des agents volitifs ou non.
Les thèmes : subissent l'action.
Les expérienceurs : sont utilisés avec certains cas de verbes comme les verbes psychologiques, verbes de causalité, …
La localisation : désigne différents types de lieux (spatiaux, temporels)
Le moyen : ce sont les moyens utilisés pour l'accomplissement de l'action.
Le but : désigne le but final (objectif) de l'action.

Exemple

[Jenhani]

- Verbes d'activités concrètes

1. Verbes de communication

Exemples :
Appeler, représenter, parler, répondre, écouter, dresser, téléphoner.
Schéma thématique :

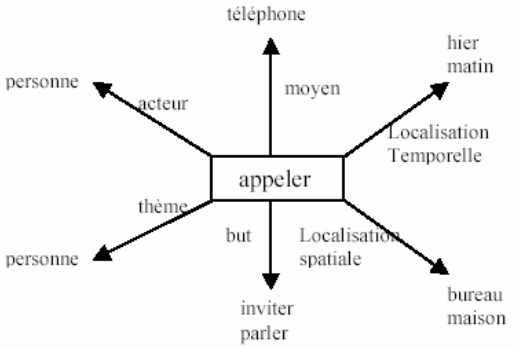

Fig.8 schéma_actvité_concrète

- **Verbes d'activités abstraites**

1. Verbes exprimant la causalité

Exemples :
Causer, entraîner, induire, résulter.
Schéma thématique :

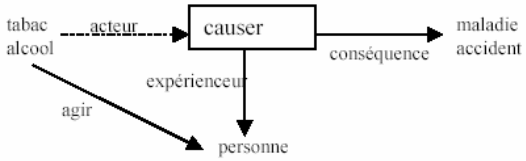

Fig.9 schéma_activité_abstraite

Les verbes sont classés tout d'abord en deux catégories principales, puis en sous-parties telles que :

Verbes d'activités concrètes	Verbes d'activités abstraites
Verbes de communication Verbes de possession Verbes de consommation :	Verbes exprimant la causalité Verbes psychologiques Verbes d'état

Verbes reliés à l'aspect et l'action Verbes de changement Verbes de compétition Verbes de création et de destruction Verbes de mouvement Verbes de contact Verbes d'interactions sociales Soins du corps, vie et mort	Verbes de cognition Verbes de perception

Fig.10 Verbe_activité

Après cet inventaire de relations nominales, verbales, paradigmatiques ou syntagmatiques, nous pouvons observer une théorie dans laquelle les relations syntagmatiques ont été développées.

I.2.4. La Théorie Sens – Texte et les fonctions lexicales

I.2.4.1. Fondements de la théorie

Les fondements de la Théorie Sens – Texte (TST) de Mel'čuk remonte à 1965 (Žolkovskij et Mel'čuk). La TST est selon [Polguère] une théorie linguistique visant la description de la correspondance Sens ⇔ Texte, au moyen de la construction de modèles formels. Le principe est de faire correspondre à un sens une multitude de représentations textuelles.

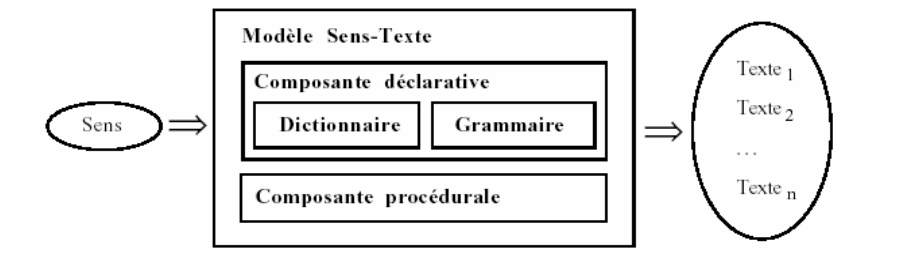

Fig.11 Structure_fonctionnelle_modèle Sens-Texte [Polguère]

Contrairement aux théories développées sur la même période, la TST porte le lexique au centre de l'organisation de la langue et non la syntaxe, ce qui nous paraît plus naturel vu l'appréhension classique de la langue par les locuteurs (position d'apprentissage par exemple). Les méthodes de la TST sont mises en place avec une optique de génération de la langue plus que d'analyse.
De plus, cette théorie ne prétend pas élucider la réalité psychique de la langue, bien qu'elle cherche à refléter les mêmes correspondances entre sens et texte que celles qu'établissent les locuteurs, d'où il nous paraît intéressant de faire intervenir cette théorie en compétition avec les théories psychologiques de l'organisation du lexique, voire de la tester en parallèle avec un locuteur.

La structure théorique de la TST est fondée sur cinq piliers [Polguère] :
1. La TST rend compte de l'association que tout locuteur d'une langue L est capable de faire entre un sens donné de L et l'ensemble des énoncés paraphrastiques de L exprimant ce sens.
La théorie imagine la langue comme une machine virtuelle permettant de traduire des Sens en énoncés (Textes) et vice versa.

2. La TST est universelle, c'est-à-dire qu'elle repose sur des principes généraux pouvant s'appliquer à toutes les langues.
3. La TST est linguistique en ce sens qu'elle permet, à partir des principes généraux sur lesquels elle repose, de construire des modèles linguistiques de description de phénomènes internes (structures lexicales et grammaticales) de la langue, spécifiques pour chaque langue humaine.
4. La TST permet de construire des modèles calculables (constitués d'un lexique, d'une grammaire et d'un ensemble de procédures pour activer ces deux composantes) permettant une utilisation par des applications informatiques.
5. La TST est formelle. C'est-à-dire qu'elle utilise des langages formels pour représenter les énoncés linguistiques et encoder les règles de manipulation des représentations linguistiques Elle se distingue des autres approches formelles, par sa richesse et sa relative complexité dans les formalismes utilisés.

Du point de vue de la méthode, la TST opère sur les niveaux canoniques définis en linguistique ; la sémantique, la syntaxe, la morphologie, la phonétique. Chaque niveau possède une représentation profonde et de surface de ses informations.
Afin de faire fonctionner ces représentations formelles, la TST a besoin d'une terminologie linguistique fondée sur un système logique et structuré de concepts.

I.2.4.2. Fonctions lexicales

I.2.4.2.1. Définition

Une partie de ce système conceptuel est représentée par le Dictionnaire Explicatif et Combinatoire (DEC) et ainsi les fonctions lexicales (FL). Le DEC est au centre de la TST, puisque représente le composant lexicographique de cette théorie fondée sur le sens/lexique.
Les FLs ont rapport à la partie non pas Explicative du DEC (définition analytique du sens de l'unité lexicale en termes plus proches des primitifs sémantiques) mais des définitions Combinatoires (lexicales, non les définitions de combinatoire syntaxique).
Les FLs sont un outil de modélisation des phénomènes collocationnels [Polguère]. Elles sont de deux types, reprenant la dichotomie utilisée précédemment pour les relations lexicales en général, c'est-à-dire paradigmatiques ou syntagmatiques.
Selon [Polguère], les fonctions lexicales paradigmatiques connectent la lexie en cause aux autres lexies de la langue qui y sont directement reliées sémantiquement (les synonymes, antonymes, conversifs, etc.), quant aux fonctions lexicales syntagmatiques, elles décrivent la combinatoire lexicale proprement dite (collocations contrôlées par la lexie).

I.2.4.2.2. Les FLs standard simples

Les FLs sont combinables entre elles. On en compte cinquante-six simples dans [Mel'čuk], mais une bonne soixantaine dans [inWanner] (voir annexe). Elles sont organisées comme suit :
Des FLs paradigmatiques [inWanner] :

1. **Syn** [Lat. *synonymum*] ➔ synonyme.
2. **Conv$_{kij}$** [Lat. *conversivum*] ➔ conversif.
3. **Contr** [Lat. *contrarium*] ➔ contrastif.
4. **Figur** [Lat. *figuraliter*] ➔ figuratif.
5. **S$_0$** ➔ Substantivation.

6. $\mathbf{A_i}$ ➔ propriété d'un actant $\mathbf{A_1}$ (participe mode actif) / $\mathbf{A_2}$(participe mode passif).

Des FLs syntagmatiques :

FLs Nominales :

7. **Centr** [Lat. *centrum*] ➔ centre.

FLs Adjectivales/Adverbiales

8. **Magn** [Lat. *magnus*] ➔ degré important de…, beaucoup, intense.

FLs Prépositionnelles

9. $\mathbf{Loc^{temp,in}}$ ➔ locatif temporel.

FLs Verbales

10. **Pred** [Lat. **praedicatum*] ➔ verbe avec sens de la copule « être » toujours en combinaison avec d'autres FLs.

Verbes support

11. $\mathbf{Oper_i}$ ➔ FL support .Verbalisation du Nom prédicatif (rôle syntaxique COD[1]).

Verbes phasiques

12. **Incep** [Lat. *incipere*] ➔ commencer.

Verbes causatifs

13. **Liqu** [Lat. **liquidare*] ➔ empêcher qqch en cours.

Verbes de réalisation

14. $\mathbf{Real_{0/i}}$ [Lat. *realis*] ➔ réalisation (Oper).

Divers

15. **Involv** [Lat. *involvere*] ➔ implication.
16. **Obstr** [Lat. *obstruere*] ➔ fonctionnement difficile.
17. **Stop** [Lat. **stuppare*] ➔ arrêt de fonctionnement.

Ceci est une présentation globale des fonctions lexicales, elles sont toutes répertoriées en annexe. Les fonctions nous intéressant seront décrites selon le besoin.

I.2.4.3. Application et Implémentations de la TST

I.2.4.3.1. DiCo

Le Dictionnaire de Combinatoire (DiCo) est un projet du groupe OLST (l'Observatoire de linguistique Sens-Texte) avec I.Mel'čuk et A.Polguère. C'est une implémentation du DECFC (Dictionnaire explicatif et combinatoire du français contemporain).
Le DiCo et le LAF (Lexique actif du français) consistent en l'élaboration d'un dictionnaire de dérivations et de collocations formalisé et fondé sur une base de données lexicale formelle et automatisée. DiCo vise les lexies intéressantes pour les fonctions lexicales et représente une nomenclature d'environ 2500 vocables.
Chacune des lexies est décrite selon deux axes : les dérivations sémantiques qu'elle entretient avec d'autres lexies de la langue et les collocations qu'elle contrôle. Le modèle fonctionnel utilisé dans DiCo pour décrire les propriétés de combinatoire lexicale des éléments est celui proposé par la TST (Fonctions lexicales) [Grizolle].

Le DiCo est organisé en 7 zones :

1. nom du vocable,

[1] COD / I : Complément d'Objet Direct / Indirect ; SG : Sujet Grammatical.

2. partie du discours et numéro de lexie,
3. caractéristiques grammaticales et les marques d'usage,
4. caractéristiques sémantiques,
5. nota bene,
6. tableau de régime,
7. fonctions lexicales (non-exhaustif), exemples et phraséologie

EdiCo est l'outil de DiCo; il permet d'utiliser DiCo, c'est-à-dire d'éditer les articles, faire des recherches sur la base du dictionnaire, mais aussi vérifier la cohérence intra- et interarticles.
Le LAF, lui, est un produit grand public dérivé entièrement du DiCo. Il a un objectif d'enseignement principalement.

I.2.4.3.2. Dictionnaire bilingue

C'est un projet de lexicographie computationnelle qui a donné lieu à la transformation du dictionnaire bilingue *Collins-Robert* en base de données enrichie sémantiquement.
[Fontenelle] a extrait les informations métalinguistiques et collocationnelles de la typographie des articles afin de développer les ressources lexicales informatisées.
T.Fontenelle a codé plus ou moins automatiquement, en utilisant des fonctions lexicales environ 70 000 paires d'items linguistiques.
Il a aussi étudié les possibilités d'exploitation de la base de données et les limitations ou les imprécisions de la TST. Par exemple, on apprend que les FLs sont insuffisantes pour ce genre de description et qu'il a dû créer de nouvelles fonctions lexicales (comme une fonction de base, la fonction partie-tout, considérée comme étant simplement une relation sémantique dans la théorie, mais qui permet cependant de raffiner l'information ; ou une fonction *« Telic »* pour indiquer la fonction entre un nom argument et son verbe typique qui lui est associé, ex : Telic (key) : open).
D'une façon générale, Fontenelle en conclut que la TST doit être approfondie et plaide pour une intensification de ce type de recherche [Clas].

I.2.5. Les relations dans EWN

Les relations sémantiques d'EuroWordNet sont naturellement celles que l'on trouve dans WordNet (WN), mais pas seulement.
La liste qui suit énumère les relations sémantiques disponibles dans WordNet. Ces relations, quand elles concernent des concepts, des synsets, sont dites relations sémantiques, mais si l'on a affaire à une relation entre deux mots c'est une relation lexicale (précisé dans le tableau). Seul les relations plus originales seront décrites [Jenhani].

Relation	Relation inverse
Synonymie	
Antonymie (complémentarité ; relation lexicale)	
Hyperonymie	**Hyponymie**
Méronymie	**Métonymie**
Implication	
Causalité	
Valeur: relation liant un concept-1 (adjectif)	**A pour valeur**: relation liant un concept-1 à

qui est un état possible pour un concept-2 (ex : pauvre / condition financière)	ses valeurs (adjectifs) possibles (taille / grand).
Voir aussi: relation entre des concepts ayant une certaine affinité (froid / gelé)	
Similaire à lie un synset périphérique au synset central (moite / humide).	
Dérivé de: indique une dérivation morphologique entre le concept cible (adjectif) et le concept origine (froidement / froid).	

Fig.12 Relations_WN

EWN possède cependant des relations en plus, telles que les labels.
Le label, représentant un domaine particulier, permet de lier des concepts de natures différentes. Les labels sont similaires aux domaines créés automatiquement par l'outil ROSA. Les labels servent à la désambiguïsation sémantique.

De plus, EWN propose des relations entre noms et verbes, contrairement à WN où chaque classe grammaticale forme un réseau indépendant. On trouve ainsi de :
la synonymie verbe-nom ;
la synonymie nom-verbe ;
l'hyperonymie nom vers verbe ;
l'hyponymie verbe vers nom.

Les verbes sont organisés comme expliqué précédemment (voir « les relations thématiques »).

A défaut de pouvoir utiliser tous ces liens lexicaux et sémantiques, les applications informatiques traitant le langage ont été développées avec une base plus proche de la linguistique de corpus et des méthodes statistiques.

I.3. Informatique

Plusieurs outils traitent le langage avec des liens lexicaux implicites. Certains de ces outils nous seront utiles dans notre analyse, principalement les trois suivants que nous présentons maintenant.

I.3.1. ROSA

Cette application a été développée au LIMSI par O.Ferret et B.Grau en 1998.
Il s'agit grâce à elle, de segmenter des textes en thèmes ainsi que de construire des représentations de ces thèmes de manière incrémentale. Ces représentations de thèmes sont appelées des domaines sémantiques.
Les domaines sont en fait une agrégation de segments de discours. C'est un ensemble de mots lemmatisés et de leur poids. Le poids est fonction de leur importance dans le thème décrit.
Le processus se fonde sur un large corpus (Los Angeles Times, dépêches AFP). Le traitement par contre ne se fera que sur les mots pleins lemmatisés.
ROSA est constituée de deux modules : SEGCOHLEX et SEGAPSITH. Le premier permet la segmentation thématique et la délimitation de segments de discours (une situation), alors que le second représente l'implémentation de la méthode et forme des descriptions spécifiques.

I.3.1.1. SEGmentation thématique par utilisation de la COHésion LEXicale

Ce module utilise un réseau de collocations pour être en mesure d'effectuer une segmentation. Les relations sémantico-pragmatiques entre les mots qui sont les liens de ce réseau, permettent de capturer la notion de cohésion lexicale (par l'information mutuelle entre deux mots), puisque le calcul de la cohésion se fait sur la somme des poids des liens du réseau reliant le mot et ceux qui appartiennent à son environnement. Un mot et ses relations sont observés par le passage d'une fenêtre d'environ 20 mots. [Chalendar01] explique que « l'hypothèse suivie est que les mots d'un segment de texte portant sur le même sujet sont fortement liés dans le réseau de cooccurences et induisent une valeur de cohésion élevée. ».
De là, une baisse de la cohésion permettra de délimiter un segment.
La valeur de cohésion se fonde aussi sur les mots mêmes du réseau reliés par un poids suffisant à d'autres mots de la fenêtre. Ces mots sont dits « inférés » et sont ajoutés à la représentation du texte [Chalendar01].
Les zones où la cohésion calculée est élevée sont candidates à être intégrées en mémoire dans la représentation d'une situation, formant une unité thématique lexicale (UTL).

I.3.1.2. SEGmentation et APprentissage de SIgnatures THématiques

Ce module ne se fonde plus sur du texte, mais sur la sortie du module précédent, c'est-à-dire les UTLs pour construire incrémentalement des représentations de thèmes par agrégations des UTLs similaires [Chalendar01], des domaines thématiques.

L'apprentissage d'une description complète consiste en l'agrégation de toutes les UTLs similaires dans un domaine. A chaque agrégation, de l'information nouvelle est apportée et les mots récurrents voient leur poids renforcé. Le poids est calculé d'après le nombre d'occurrences du mot dans ces UTLs.
Les domaines après avoir subi un certain nombre d'agrégations (environ dix) deviennent « stables » (selon O.Ferret), c'est-à-dire que les mots aux poids les plus élevés ne changent plus.
L'ensemble des domaines thématiques du système forme la mémoire thématique. Cette mémoire n'est pas qu'un résultat, puisqu'elle est remise en jeu par l'utilisation de domaines stables dans la segmentation.
Ces domaines non structurés dans SEGAPSITH sont utilisés par SVETLAN'. SVETLAN' intègre le système ROSA.

I.3.1.3. Architecture de ROSA

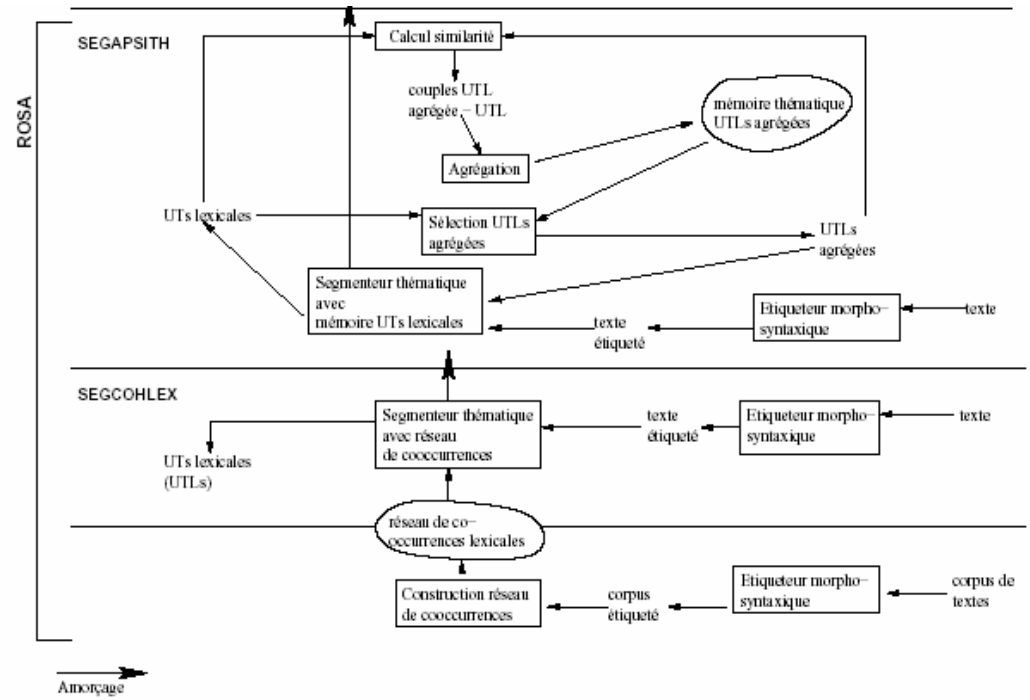

Fig.13 ROSA

I.3.2. SVETLAN'

SVETLAN' est une application développée par G. De Chalendar et B. Grau en 2001. Le but de cette application est de structurer des domaines sémantico-pragmatiques sur le français et l'anglais.
D'une certaine façon, sa base est le système ROSA de Ferret couplé à un analyseur syntaxique.
Il fonctionne d'ailleurs à base d'un corpus issu de ROSA, constitué de domaines sémantiques avec les unités thématiques (UT) qui les forment.
Le premier traitement sur le corpus est une analyse syntaxique (par un analyseur externe) qui permet d'obtenir des unités thématiques structurées (UTS). Les UTSs sont des triplets comportant un verbe, un nom constituant la tête d'un argument et son rôle syntaxique (relation syntaxique instanciée) [Chalendar02].

Les UTSs d'un même domaine sont ensuite agrégées de manière à former un domaine structuré.
L'agrégation consiste à regrouper les noms jouant un même rôle auprès du même verbe, définissant une classe sémantique. Pour agréger des UTSs avec un domaine structuré il y a tout d'abord agrégation des relations syntaxiques identiques liées au même verbe puis ajout de nouvelles relations syntaxiques (consistant à ajouter une nouvelle relation et son mot argument à un verbe existant ou à ajouter un nouveau verbe avec son argument).
Les noms ne sont pas pondérés à l'intérieur des classes mais conservent leur poids au niveau du domaine.
En fait, SVETLAN' délimite des sous-classes dans un domaine et les associe aux verbes qu'elles permettent ainsi de définir. C'est une approche distributionnelle.
Ainsi, l'ensemble des mots constituant une classe sont ceux qui jouent un même rôle par rapport à un verbe dans un contexte similaire, la similarité de contexte étant définie par une similarité lexicale calculée sur la totalité du domaine.
Bussone et Rossi ont déterminé les effets d'amorçage liés aux associations verbe-verbe et aux associations verbe-nom pour évaluer la structure du réseau sémantique des verbes. A l'aide d'une décision lexicale, ils ont découvert que le degré de proximité sémantique était plus élevé pour des associations verbe-nom. Les relations verbe-verbe (comme, acheter - vendre) n'étaient pas activées automatiquement. Ceci signifie que le lexique des verbes est structuré autour des noms [RESSEM].

Le but à terme pour SVETLAN' est de fournir une base de connaissances lexicales utilisable dans des applications de recherche d'information ou de désambiguïsation sémantique.

I.3.2.1. Architecture de SVETLAN'

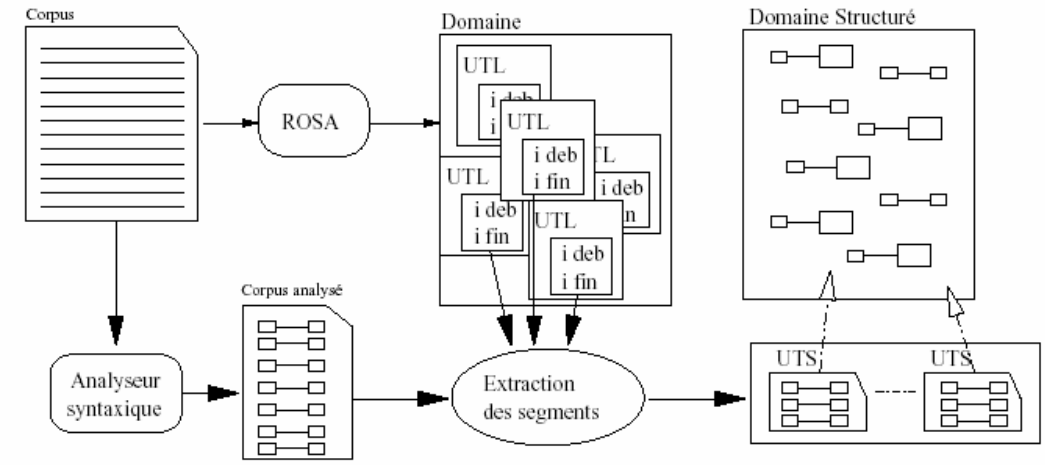

FIG. 6.9.: Schéma de l'apprentissage d'un Domaine Structuré dans SVETLAN' par transposition d'un Domaine Thématique

Fig.14 SVETLAN'

I.3.3. PROMETHEE

Le système de Morin, PROMETHEE, est un outil d'aide au repérage des relations entre les entités du domaine qui interviennent dans la constitution d'une représentation du domaine [Morin99]. Le modèle acquiert des schémas lexico-syntaxiques caractéristiques d'une relation sémantique d'une manière incrémentale, par l'analyse d'un corpus technique et exploite ceux-ci pour extraire des relations sémantiques entre termes.

PROMETHEE est constitué de trois modules :
1. Un analyseur de surface (mise en forme du corpus)
2. Un extracteur de schémas lexico-syntaxiques (d'une liste de termes conceptuellement liés et d'un corpus pré-traité, acquisition de schémas lexico-syntaxiques caractéristiques d'une relation sémantique)
3. Un extracteur de couples de termes conceptuellement liés (de schémas lexico-syntaxiques, extraction de relations entre termes).

Par ces relations entre termes, il est possible d'effectuer une représentation hiérarchique des couples de termes. Il s'agit ensuite d'inférer de nouveaux liens entre termes polylexicaux.

I.3.3.1. Architecture de PROMETHEE

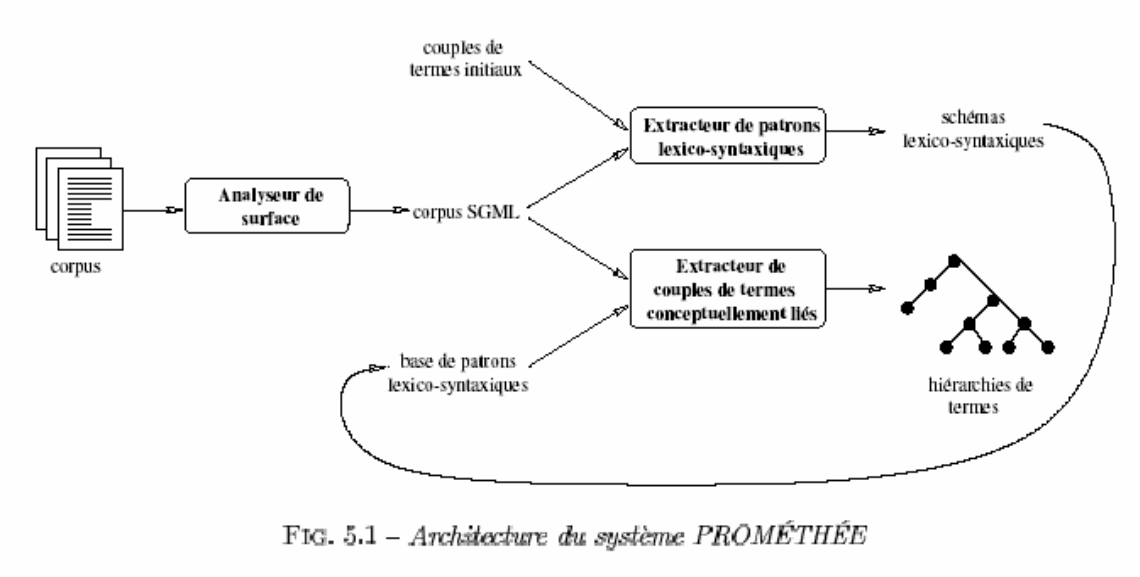

FIG. 5.1 – *Architecture du système PROMÉTHÉE*

Fig.15 Prométhée

Ces différents systèmes peuvent servir aussi bien dans la pratique, qu'au niveau de la méthodologie. Ainsi, ROSA et SVETLAN' seront la base pratique de l'observation, tandis que PROMÉTHÉE aide à la méthodologie.

Cet état de l'Art pose les bases théoriques nécessaires à la suite de l'observation. Il n'est pas exhaustif et ne représente qu'une partie des techniques appropriées à ce type de recherches. Une analyse plus poussée demanderait d'autres bases théoriques, plus profondes.
Globalement, nous voyons que les relations commencent à être bien décrites au niveau lexical et toujours en analyse aux niveaux sémantique et psychologique. Cependant, leur utilisation concrète, au sein d'applications informatiques par exemple, n'est que rare. À l'aide des applications ROSA et SVETLAN', ne contenant pas ou peu de liens typés explicitement, et EWN, base de données ne comportant que des lexèmes typés, nous allons tenter de construire une analyse sur ces différentes relations.

Bibliographie

Bibliographie Psycholinguistique

Bibliographie Mémoire :
[Mem] http://www.prevention.ch/lamemoire.htm
[Mem] http://tecfa.unige.ch/~lydia/staf_11/

Bibliographie Gestion mentale :
[GesMen] http://www.gestionmentale.com/pas01b.htm
[GesMen] http://epreau.chez.tiscali.fr/gm/presentation.htm
[GesMen]http://www.univ-reims.fr/Labos/LERI/membre/bittar/Motivation/html-rapport/Imagin.Gallien/node2.html#SECTION00011000000000000000

Bibliographie Tip-Of-the-Tongue :
[Labelle] Marie Labelle, *Trente ans de psycholinguistique,* Revue québécoise de linguistique 2001, vol. 30, no. 1
[Ben] Bennett L. Schwartz, *Illusory Tip-of-the-tongue States* in MEMORY, 1998, *6* (6), 623-642
[TOT] http://www.webschooling.com/1000809220408.html
[TOT] http://www.it.bton.ac.uk/staff/rng/teaching/notes/Memory.html
[ZockFour] Michael Zock et Jean-Pierre Fournier, "*How can computers help the writer/speaker experiencing the Tip of the Tongue Problem?"* (2001). in Proc. of RANLP, Tzigov Chark, Bulgarie, pp. 300-302

Bibliographie réseaux sémantiques :
[ResSem] http://www.mi2s.u-bordeaux2.fr/~gruselle/dea/essaisTC1/Bresson-DePaepe/

Bibliographie HAL :
[GlenRob]
[HAL] http://locutus.ucr.edu/~curt/resint97.html

Bibliographie LSA :
[GlenRob]
[LSA] http://lsa.colorado.edu/

Bibliographie Embodied Cognition/ Indexical Hypothesis :
[Barsalou] Lawrence W. Barsalou, P*erceptual Symbol Systems* in Behavioral and Brain Sciences (1999) 22, 577–660
[Glenberg] Arthur M. Glenberg, *What memory is for* in Behavioral And Brain Sciences (1997) 20, 1–55
[GlenRob] Arthur M.Glenberg et David A.Robertson, *Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning* in Journal of Memory and Language 43, 379-401 (2000).

Bibliographie Wordnet :
[Miller]http://www.cs.oswego.edu/~blue/hx/courses/cogsci1/s2001/section05/subsection9/main.html

Bibliographie Linguistique

Bibliographie Relations paradigmatiques et syntagmatiques :
[Cruse] D.Cruse, *Lexical Semantics*, Cambridge University Press, 1986.
[Desgoutte] Jean-Paul Desgoutte, *Multimédia, les mutations du texte*, éd. Thierry Lancien, Cahiers du français contemporain n° 6, E.N.S. Editions, Paris 2000.
[Girault] Stéphanie Girault, *Rapport de stage de DEA de sciences cognitives* Université Paris Sud, mai-septembre 1993.
[Jenhani] Olfa Jenhani *Rapport ARC INRIA GeNI*, 2003.

[WinstonChaffin] Chaffin, Hermann, Winston, *A taxonomy of part-whole relations*, Cognitive Science, 11, 1987.

Bibliographie Relations Structurantes et Fonctionnelles Noms/Verbes :

[Jenhani] Olfa Jenhani *Rapport ARC INRIA GeNI*, 2003.
[Gardent] Claire Gardent, *Wordnet, Fellbaum & Miller,* http://www.loria.fr/~gardent/teaching/semLex/wdnet4.pdf

Bibliographie TST :

I. Mel'cuk, A. Clas, A. Polguère (1995) *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve: Duculot, 1995, 256 pages.
[Polguère] Alain Polguère, *Théorie Sens-Texte*, Dialangue, Vol. 8-9, Université du Québec à Chicoutimi.
[Polguère02] Alain Polguère, *Modélisation des liens lexicaux au moyen des fonctions lexicales*, TALN 2002, Nancy, juin 2002.
[Mel'čuk] Igor Mel'čuk, André Clas, Alain Polguère, Introduction à la lexicologie explicative et combinatoire, ed.Duculot, 1995.
[inWanner] I.Mel'čuk, *Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon*, Lexical Functions in Lexicography and Natural Language Processing, Ed. by Leo Wanner, Universität Stuttgart, 1996.
[Mel'čuk97] I. Mel'cuk *Vers une linguistique Sens-Texte. Leçon inaugurale*. Paris: Collège de France, 1997
[Grizolle] Grizolle Bénédicte, *Classification des fonctions lexicales non standard du DiCo. Lexies étiquetées animal et artefact*. Rapport de stage, Département de linguistique et de traduction, Université de Montréal, 2003.
François Lareau*, A practical guide to writing DiCo entries, 2003 ?*
vulab.ias.unu.edu/papillon/papillon-docs/ GuideToDiCo-01_FLareau.pdf
[Fontenelle] Fontenelle, Thierry, *Turning a Bilingual Dictionary into a Lexical-Semantic*, Lexicographica, Series Maior 79, Tübingen, Niemeyer, 328 p.1997.
Fontenelle Thierry, *Fonctions lexicales et sémantique lexicale dans les dictionnaires informatisés,* présentation Jussieu, ppt. mai 2003.
[Clas] A.Clas, compte-rendu,
http://www.erudit.org/revue/meta/1999/v44/n3/002307ar.html

Bibliographie Relations EWN :

[Jenhani] Olfa Jenhani *Rapport ARC INRIA GeNI*, 2003.
[EWN] manuel EWN version date.

Bibliographie Informatique

Bibliographie ROSA :

[Chalendar01] Gaël De Chalendar, *SVETLAN', un système de structuration du lexique guidé par la détermination automatique du contexte thématique*, Thèse de doctorat, Univ. Paris XI, 2001.

Bibliographie SVETLAN' :

[Chalendar02] Gaël De Chalendar, Brigitte Grau, *Structuration de domaines sémantiques*, IC 2002.
[RESSEM]

Bibliographie Morin :

[Morin99] Emmanuel Morin, *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques,* Thèse de doctorat. Univ. Nantes, 1999.
http://www.sciences.univ-nantes.fr/info/perso/permanents/morin/promethee/promethee.html

Bibliographie divers

[ZOCKFOUR] M. Zock & J.P. Fournier, *How can computers help the writer/speaker experiencing the Tip of the Tongue Problem?* Proc. of RANLP, Tzigov Chark, Bulgarie, pp. 300-302, 2001.
[BQRDOC] http://www.limsi.fr/Individu/jacquemi/BQR99/projFormatCommun/projFormatCommun.html
[W3C] http://www.w3.org/XML/
[Perl] http://www.perl.org/
[MorTALInt] Nicolas Hernandez http://www.limsi.fr/cgi-bin/analyse.pl
[MorTAL] Christian Jacquemin http://www.limsi.fr/Individu/jacquemi/MORTAL/
[AFP] http://www.afp.fr/francais/home/